Cloud Computing : Manajemen dan Perencanaan Kapasitas

by Riko Herwanto

Submission date: 16-Nov-2020 11:04AM (UTC+0700)
Submission ID: 1447288589
File name: CLOUD_COMPUTING_-_Manajemen_dan_Perencanaan_Kapasitas.pdf (2.93M)
Word count: 32292
Character count: 224858

CLOUD COMPUTING

Manajemen dan Perencanaan Kapasitas

Komputasi awan *(cloud computing)* adalah teknologi yang menjadikan internet sebagai pusat pengelolaan data dan aplikasi, dengan pengguna komputer diberikan hak akses *(login)*. Awan *(cloud)* adalah metafora dari internet, sebagaimana awan yang sering digambarkan di diagram jaringan komputer. Sebagaimana awan dalam diagram jaringan komputer, awan *(cloud)* dalam *Cloud Computing* juga merupakan abstraksi dari infrastruktur kompleks yang disembunyikannya. Ia adalah suatu metode komputasi, yang kapabilitas terkait teknologi informasinya disajikan sebagai suatu layanan *(as a service)*, sehingga pengguna dapat mengaksesnya melalui internet (di dalam awan) tanpa mengetahui apa yang ada di dalamnya, ahli dengannya, atau memiliki kendali terhadap infrastruktur teknologi yang membantunya.

Buku ini bertujuan supaya pembaca dapat mengetahui dan memahami semua tentang kapasitas media penyimpanan online dengan jelas.

Di dalam buku ini akan dibahas segala hal mengenai manajemen dan perencanaan kapasitas dalam *cloud computing*, antara lain:

- karakteristik *cloud*
- bentuk dan pemodelan *cloud computing*
- layanan *cloud computing*
- manajemen kapasitas *cloud*
- mengelola kapasitas *cloud*
- perencanaan kapasitas
- pengukuran kapasitas
- desain manajemen kapasitas
- dan lain sebagainya

Penerbit ANDI
Jl. Beo 38-40 Yogyakarta
Telp (0274) 561881 Fax (0274) 588282
e-mail : penerbitan@andipublisher.com
website: www.andipublisher.com

COMPUTING & INTERNET
ISBN: 978-623-01-03815-5
Harga di Pulau Jawa: Rp57.000,00

CLOUD COMPUTING Manajemen dan Perencanaan Kapasitas

Riko Herwanto | Onno W. Purbo | RZ. Abd. Aziz

Judul: **CLOUD COMPUTING : Manajemen dan Perencanaan Kapasitas**

Penulis: **Riko Herwanto**
Onno W. Purbo
RZ. Abd. Aziz

Institusi: **Institut Informatika dan Bisnis Darmajaya**
Bandar Lampung

Kata Pengantar

Bismillahirahmanirrahiim
Assalamu'alaikum Warrahmatullahi Wabbarakatuh,

Allhamdulillah kami panjatkan kehadirat Allah SWT, karena dengan karunia-Nya sehingga melalui rahmat-Nya yang tiada terkira terselesaikan buku dengan judul "Cloud Computing : Merencanakan Kapasitas" , buku bertujuan supaya pembaca dapat mengetahui dan memahami secara jelas mengenai semua tentang kapasitas media penyimpanan online.
Ucapan terimakasih kepada kepada seluruh pihak yang telah membantu dalam menyelesaikan buku ini.
Kami sadar dengan banyaknya keterbatasan yang kami miliki, buku ini jauh dari sempurna. Masih butuh sentuhan tangan tangan yang lebih expert dalam mengembangkannya. Kami mengharapkan input dari semua masyarakat, karena buku ini adalah sedikit sumbangsih kami. Demikian buku ini kami susun, semoga dapat bermanfaat bagi penyusun khususnya dan pembaca pada umumnya. Saran dan kritik yang membangun sangat kami harapkan demi kesempurnaan buku ini.

Wassalamu'alaikum Warrahmatullahi Wabarakatuh.

Banndar Lampung, Januari 2019
Penulis

Tentang Penulis

Riko Herwanto pernah menjabat sebagai Koordinator Laboratorium STMIK Darmajaya, saat ini adalah mahasiswa Magister Teknik Informatika, Institut Informatika dan Bisnis Darmajaya,

Onno W. Purbo. memperoleh gelar Ph.D bidang Electrical Engineering dari University of Waterloo, Canada, adalah seorang copy left, educator dan ICT evangelist. Dia sudah mempublikasikan 50+ buku, termasuk free ICT ebook untuk sekolah tahun 2008. Beberapa buku terakhirnya adalah "Perjuangan Menyebarkan Internet", 2016; "Buku Pegangan Internet untuk Desa", 2016; and "Internet-TCP/IP: Konsep Dan Implementasi", 2018. dia memimpin sambungan pertama Internet di Institut Teknologi Bandung, tahun 1993-2000, dan menggunakannya untuk membuat jaringan Internet pendidikan yang pertama di Indonesia. Dia membebaskan frekuensi WiFi, memperkenalkan RT/RW-net, antenna Wajanbolic dan jaringan selular OpenBTS. Dia mempimpin jaringan telepon pertama di atas Internet, VoIP Merdeka, yang kemudian hari dikenal sebagai VoIP Rakyat berbasis SIP dan menggunakan kode area +62520 dan +62521. Dia saat ini aktif memperkenalkan e-Learning, dan menjalankan server e-Learnig yang free di http://lms.onnocenter.or.id/moodle/ dengan lebih dari 30,000 mahasiswa.

RZ. Abd. Aziz memperoleh gelar Ph.D bidang Information Science and Technology dari Osaka University, Japan, adalah tenaga pengajar pada Institut Informatika dan Bisnis Darmajaya. Dia tertarik dalam bidang kajian tentang Manajemen Kualitas, Knowledge Management, Optimasi dan Simulasi Sistem. Saat ini telah menulis beberapa buku dalam bidang Manajemen Kualitas dan Kewirausahaan."

Daftar Isi

BAB SATU

TENTANG CLOUD

Bab ini membahas konsep dasar komputasi cloud, teknologi cloud, dan bahan-bahannya. Sebelum secara mendalam ke dalam perencanaan kapasitas, penting untuk memiliki pemahaman yang jelas tentang definisi teknologi. Karakteristik, penyebaran, dan model layanan komputasi cloud telah disederhanakan untuk mengatur konteks untuk bab-bab berikutnya. Selain dasar-dasar cloud, bab ini membahas dampak komputasi cloud pada perusahaan. Bab ini juga membuat pembaca memahami bagaimana peran perusahaan IT berubah ketika solusi cloud dipertimbangkan. Di bawah lingkup komputasi cloud, lanskap IT tradisional mencari transformasi untuk mendukung aplikasi bisnis secara efisien dan efektif.

Cloud Computing

Komputasi awan (cloud computing) adalah teknologi yang menjadikan internet sebagai pusat pengelolaan data dan aplikasi, di mana pengguna komputer diberikan hak akses (login). wan (cloud) adalah metafora dari internet, sebagaimana awan yang sering digambarkan di diagram jaringan komputer. Sebagaimana awan dalam diagram jaringan komputer tersebut, awan (cloud) dalam Cloud Computing juga merupakan abstraksi dari infrastruktur kompleks yang disembunyikannya. Ia adalah suatu metoda komputasi di mana kapabilitas terkait teknologi informasi disajikan sebagai suatu layanan (as a service), sehingga pengguna dapat mengaksesnya lewat Internet ("di dalam awan") tanpa mengetahui apa yang ada didalamnya, ahli dengannya, atau memiliki kendali terhadap infrastruktur teknologi yang membantunya. Menurut sebuah makalah tahun 2008 yang dipublikasi *Institute of Electrical and Electronics Engineers,* IEEE, Internet Computing "Cloud Computing adalah suatu paradigma di mana informasi secara permanen tersimpan di server di internet dan tersimpan secara sementara di komputer pengguna (client) termasuk di dalamnya adalah desktop, komputer tablet, notebook, komputer tembok, handheld, sensor-sensor, monitor dan lain-lain."

Gambar 1-1 . *Konseptual Komputasi Awan*

Komputasi awan adalah suatu konsep umum yang mencakup SaaS, Web 2.0, dan tren teknologi terbaru lain yang dikenal luas, dengan tema umum berupa ketergantungan terhadap Internet untuk memberikan kebutuhan komputasi pengguna. Sebagai contoh, Google Apps menyediakan aplikasi bisnis umum secara daring yang diakses melalui suatu penjelajah web dengan perangkat lunak dan data yang tersimpan di server. Komputasi awan saat ini merupakan trend teknologi terbaru, dan contoh bentuk pengembangan dari teknologi Cloud Computing ini adalah iCloud

Cloud Computing adalah evolusi selanjutnya dari internet. Cloud Computing merupakan penyedia atau hal-hal yang berkaitan dari tenaga komputasi hingga infrastruktur komputasi, aplikasi-aplikasi, proses bisnis hingga kolaborasi yang muncul sebagai layanan yang dapat diakses pada saat dibutuhkan kapanpun dan dimanapun.

Karakteristik Cloud

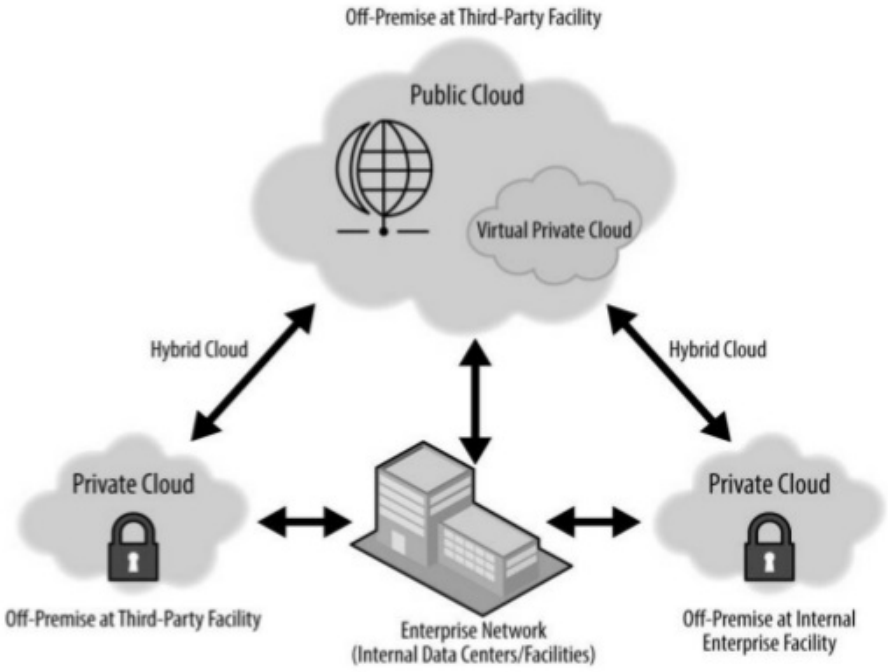
Cloud Computing terdiri atas enam buah karakteristik utama yaitu:

- *On Demand Self Service* (pelayanan mandiri diri sendiri saat diperlukan)
  Pengguna dapat memesan dan mengelola layanan tanpa interaksi manusia dengan penyedia layanan, misalnya dengan mengguna-kan, sebuah portal web dan manajemen antarmuka. Pengadaan dan perlengkapan layanan serta sumber daya yang terkait terjadi secara otomatis pada penyedia
- *Broad Network Access* (akses jaringan yang besar)
  Layanan yang tersedia terhubung melalui jaringan besar (broadband network), terutama untuk dapat diakses secara memadai melalui jaringan internet, baik menggunakan thin client, thick clien, ataupun media lain seperti smartphone.
- *Resource Pooling* (resource yang menyatu)
  Penyedia layanan cloud memberikan layanan melalui sumberdaya yang dikelompokkan di satu atau berbagai lokasi data center yang terdiri dari sejumlah server dengan mekanisme multi-tenant. Mekanisme multi tenant ini memungkinkan sejumlah sumberdaya komputasi digunakan bersama-sama oleh sejumlah user, dimana sumberdaya tersebut baik yang berbentuk fisik maupun virtual, dapat dialokasikan secara dinamis untuk kebutuhan pengguna/pelanggan sesuai permintaan. Dengan demikian, pelanggan tidak perlu tahu bagaimana dan darimana permintaan akan sumberdaya komputasinya dipenuhi oleh penyedia layanan. Yang penting, semua permintaan dapat terpenuhi. Sumberdaya komputasi ini meliputi media penyimpanan, memory, processor, pita jaringan, mesin virtual.
- *Rapid Elasticity* (kapasitas yang elastis dan cepat)
  Kapasitas komputasi yang disediakan dapat secara elastis dan cepat disediakan, baik itu dalam bentuk penambahan atau pengurangan kapasitas yang diperlukan.
- *Measured Service* (layanan yang terukur)
  Sumber daya cloud yang tersedia harus dapat diatur dan dioptimasi penggunaannya, dengan suatu sistem pengukuran yang dapat mengukur penggunaan dari setiap sumberdaya komputasi yang digunakan (penyimpanan,memory,processor,lebar pita, dan aktivitas user, dan lainnya). Dengan demikian, jumlah sumber daya yang digunakan dapat secara transparan diukur yang akan menjadi dasar bagi user untuk membayar biaya penggunaan layanan.

Gambar 1-2 . Konseptual Karakteristik Komputasi Awan

Bentuk dan Permodelan

Menurut *National Institute of Standard and Technology* (*NIST*), model pengembangan Cloud Computing digambarkan sebagai berikut:



Gambar 1-3 . Konseptual Model Pengembangan Komputasi Awan

- **Private cloud**
  Infrastruktur cloud yang disediakan secara khusus untuk digunakan oleh satu organisasi yang terdiri dari beberapa unit bisnis. Private Cloud dimiliki, dikelola dan dioperasikan oleh organisasi, pihak ketiga, atau kombinasi keduanya, dan dapat berada pada suatu tempat yang sama ataupun berbeda.

- **Community cloud**
  Infrastruktur cloud yang disediakan secara khusus untuk digunakan oleh komunitas yang spesifik dari organisasi-organisasi yang memiliki kepentingan bersama. Community Cloud dimiliki, dikelola dan di operasikan oleh satu atau lebih organisasi dalam komunitas tersebut, pihak ketiga, atau kombinasi keduanya, dan dapat berada pada suatu tempat yang sama ataupun berbeda.

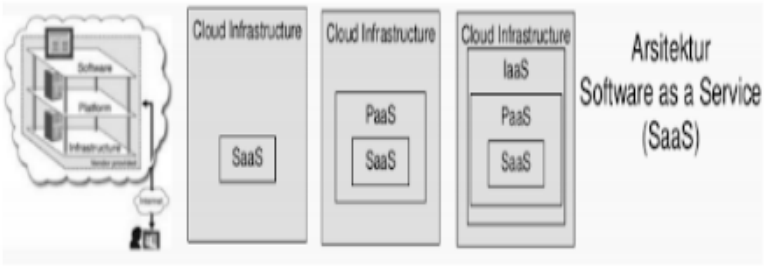
- **Public cloud**
  Infrastruktur yang disediakan secara terbuka untuk digunakan oleh masyarakat umum. Public Cloud dimiliki, dikelola dan dioperasikan oleh perusahaan, akademis, atau organisasi pemerintah, atau kombinasi dari semuanya. Public cloud berada pada tempat yang ditentukan penyedia layanan cloud.

- **Hybrid cloud**
  Infrastruktur cloud yang terdiri dari dua atau lebih infrastruktur cloud yang berbeda (private, community atau public) yang tetap unik, namun terikat pada standar atau paten teknologi yang memungkinkan portabilitas pada data dan aplikasi.

Layanan Cloud Computing

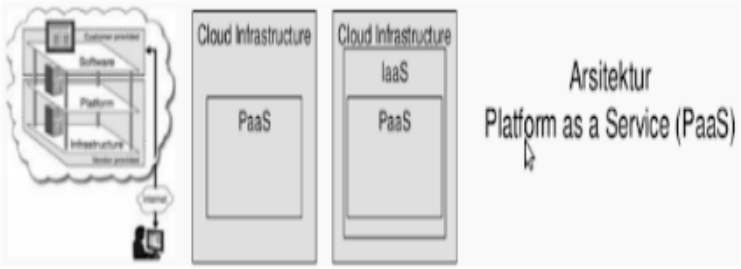
Secara umum terdapat tiga bentuk layanan Cloud Computing, yaiu sebagai berikut:

a. **Software as a Service (SaaS)**



Software as a Service adalah layanan cloud computing dimana pengguna bisa menggunakan secara langsung aplikasi yang telah disediakan. Pengguna yang menggunakan layanan SaaS hanya membutuhkan aplikasi yang dapat menghubungkan pengguna ke aplikasi yang ada internet. Beberapa contoh layanan SaaS yang popular adalah gmail, google+ dan google apps. SaaS dapat memiliki banyak keuntungan salah satunya adalah pengguna tidak perlu membeli lisensi untuk mengakses aplikasi tersebut, sebuah laporan yang diterbitkan oleh Microsoft Corporation menekankan salah satu manfaat terbesar dari penggunaan layanan SaaS adalah investasi awal lebih murah pada perangkat lunak dan perangkat keras.

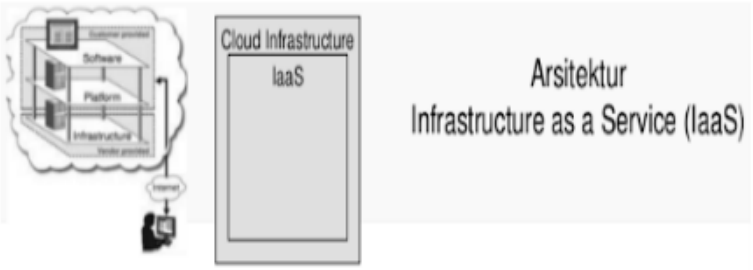
b. Platform as a Service (PaaS)



Platform as a Service adalah layanan yang menyediakan computing platform. Biasanya berupa desain aplikasi, proses percobaan dan deployment serta hosting. Layanan ini memperbolehkan pengguna membangun sebuah aplikasi dari virtualisasi perangkat keras, redundasi data, dan availability (ketersediaan) yang tinggi. Setelah pembangunan selesai, aplikasi dapat dikirim pengguna melalui internet. Keuntungan layanan ini adalah pengguna bisa lebih fokus pada pembangunan aplikasi tanpa memikirkan tentang

pemeliharaan dari computing platform. Google AppEngine, Microsoft Azura, SalesForce.com adalah contoh dari layanan.

c. Infrastructure as a Service (IaaS)



Infastructure as a Service adalah layanan cloud computing yang menyediakan infrastruktur dan perangkat keras seperti server, media penyimpanan, bandwidth, virtualisasi dan konfigurasi lain yang memungkinkan utilitas bagi pengguna. Keuntungan dari layanan IaaS adalah pengguna tidak perlu membeli komputer fisik sehingga lebih menghemat biaya, konfigurasi komputer virtual juga bisa diubah sesuai kebutuhan. Misalkan saat media penyimpanan hampir penuh, media penyimpanan bisa ditambah dengan segara. Perusahaan yang menyediakan layanan IaaS adalah Amazon EC2, TelkomCloud dan BizNetCloud.

Komponen Cloud Computing

Terdapat beberapa komponen yang digunakan dalam menerapkan Cloud Computing, yaitu sebagai berikut:

a. Cloud Clients

Cloud Clients adalah seperangkat komputer ataupun perangkat lunak yang didesain secara khusus untuk penggunaan layanan berbasis Cloud Computing.
Mobile: Windows Mobile, Symbian, dan lain-lain.
Thin Client: Windows Terminal Service, CherryPal, dll.
Thick Client: Internet Explorer, FireFox, Chrome, dll.

b. Cloud Services

Cloud Services adalah produk, layanan dan solusi yang dipakai dan disampaikan secara real-time melalui media Internet. Contoh yang paling popular adalah web service.

Identitas: OpenID, OAuth, dan lain2.
Integration: Amazon Simple Queue Service.
Payments: PayPal, Google Checkout.
Mapping: Google Maps, Yahoo! Maps.

c. Cloud Applications

Cloud Applications memanfaatkan Cloud Computing dalam arsitektur software. Sehingga pengguna tidak perlu menginstall dan menjalankan aplikasi dengan menggunakan komputer.

Peer-to-peer: BitTorrent, SETI, dan lain-lain.
Web Application: Facebook.
SaaS: Google Apps, SalesForce.Com, dan lain-lain.

d. Cloud Platform

Cloud Platform merupakan layanan berupa platform komputasi yang berisi hardware dan software-software infrasktruktur. Biasanya mempunyai aplikasi bisnis tertentu dan menggunakan layanan PaaS sebagai infrastruktur aplikasi bisnisnya.

Web Application Frameworks: Python Django, Rubyon Rails, .NET 2. Web Hosting.
Propietary: Force.Com

e. Cloud Storage

Cloud Storage melibatkan proses penyampaian penyimpanan data sebagai sebuah layanan.

Database: Google Big Table, Amazon SimpleDB.
Network Attached Storage: Nirvanix CloudNAS, MobileMe iDisk.

f. Cloud Infrastructure

Cloud Infrastructure merupakan penyampaian infrastruktur komputasi sebagai sebuah layanan.
Grid Computing: Sun Grid.
Full Virtualization: GoGrid, Skytap.
Compute: Amazon Elastic Compute Cloud.

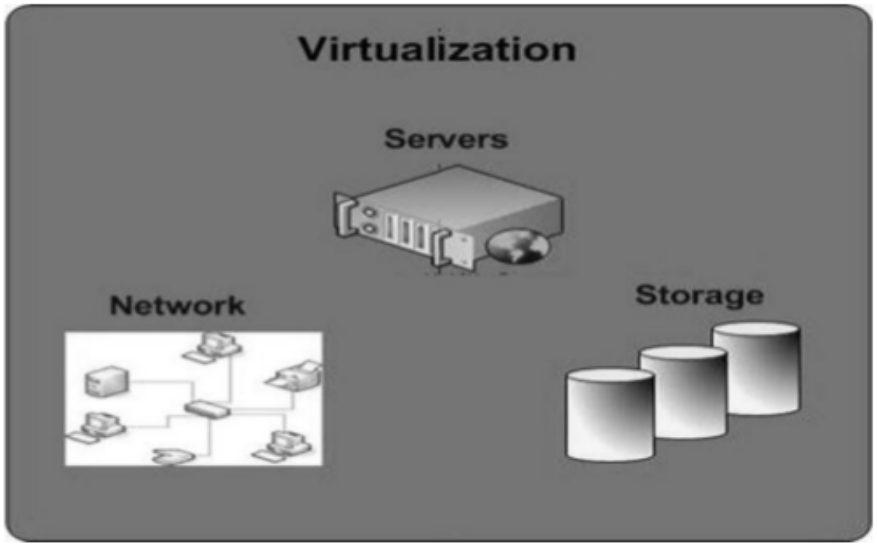
Teknologi Penggerak Cloud

Virtualisasi: Mesin Komputasi Cloud

Teknologi virtualisasi dan kecepatan yang sangat tinggi di mana teknologi prosesor komputer telah berkembang telah memungkinkan lingkungan komputasi awan menjadi layak dan bermanfaat bagi pelanggan.

Peningkatan kapasitas CPU, dimana ada beberapa core dan beberapa soket dalam satu server, menyediakan kapasitas CPU dan memori yang cukup untuk menjalankan beberapa gambar sistem operasi pada satu server.

Virtualisasi, dalam ilmu komputer, adalah pembuatan versi virtual (bukan aktual) versi perangkat atau layanan, seperti platform perangkat keras, OS, perangkat penyimpanan, atau sumber daya jaringan (Gambar 1-4). Virtualisasi adalah seni mengiris perangkat keras TI menjadi partisi dengan menerapkan teknologi virtualisasi atau hypervisor di atas perangkat keras TI dan mengubah infrastruktur fisik menjadi server virtual, penyimpanan virtual, jaringan virtual, dll.



Gambar 1-4 . *Komponen virtualisasi*

Virtualisasi dapat dilihat sebagai bagian dari tren keseluruhan dalam TI perusahaan yang mencakup komputasi otonom (skenario di mana lingkungan TI mampu mengelola dirinya sendiri berdasarkan aktivitas yang dirasakan) dan komputasi utilitas (di mana pemrosesan dan daya komputer dilihat sebagai utilitas yang digunakan klien dapat membayar hanya jika diperlukan.) Tujuan virtualisasi yang biasa adalah untuk memusatkan tugas-tugas administratif sambil meningkatkan skalabilitas dan beban kerja.

Cloud computing pada dasarnya menggunakan teknologi virtualisasi untuk berbagi server tunggal di beberapa gambar OS, yang mungkin berasal dari beberapa pelanggan. Virtualisasi adalah komponen kunci dari cloud, tetapi ruang lingkup komputasi awan jauh lebih dari virtualisasi. Menjaga teknologi virtualisasi di hati, kemampuan untuk menyebarkan dan skala infrastruktur dengan cepat dan terprogram, sesuai permintaan, dengan basis bayar-sesuai-Anda — itulah yang benar-benar mendefinisikan cloud, dan itu sulit, jika bukan tidak mungkin, untuk mencapai menggunakan virtualisasi tradisional saja.

Mesin virtual

Mesin virtual (VM) adalah jenis aplikasi komputer yang digunakan untuk membuat lingkungan virtual. Dengan kata lain, perangkat lunak mensimulasikan

lingkungan lain. Penciptaan lingkungan virtual ini disebut sebagai virtualisasi. Virtualisasi memungkinkan pengguna untuk melihat infrastruktur jaringan melalui proses agregasi. Virtualisasi juga dapat digunakan untuk menjalankan beberapa sistem operasi pada saat yang bersamaan.

Ada beberapa jenis mesin virtual. Paling umum, istilah ini digunakan untuk merujuk ke mesin virtual yang menciptakan dan menjalankan mesin virtual, juga dikenal sebagai hypervisor atau monitor mesin virtual (VMM). Perangkat lunak jenis ini memungkinkan untuk melakukan beberapa eksekusi pada satu komputer. Pada gilirannya, masing-masing eksekusi ini dapat menjalankan OS. Ini memungkinkan satu komponen perangkat keras digunakan untuk menjalankan berbagai sistem operasi dan berbagai aplikasi, yang dapat digunakan oleh banyak pelanggan cloud.

Menggunakan mesin virtual memungkinkan pengguna memiliki mesin yang tampaknya pribadi dengan fungsional penuh, perangkat keras yang diemulasikan yang terpisah dari pengguna lain. Perangkat lunak mesin virtual juga memungkinkan pengguna untuk boot dan me-restart mesin mereka dengan cepat, karena tugas seperti inisialisasi perangkat keras tidak diperlukan.

Mesin virtual juga dapat merujuk ke perangkat lunak aplikasi. Dengan perangkat lunak ini, aplikasi diisolasi dari komputer yang digunakan. Perangkat lunak VM ini dimaksudkan untuk digunakan pada sejumlah platform komputer. Ini membuatnya tidak perlu membuat versi terpisah dari perangkat lunak yang sama untuk OS dan komputer yang berbeda. Mesin virtual Java adalah contoh yang sangat terkenal dari mesin virtual aplikasi.

Mesin virtual juga bisa menjadi lingkungan virtual, yang juga dikenal sebagai server pribadi virtual. Lingkungan virtual jenis ini digunakan untuk menjalankan program di tingkat pengguna. Oleh karena itu, ini hanya digunakan untuk aplikasi dan bukan untuk driver atau kernel OS.

Mesin virtual juga dapat berupa sekelompok komputer yang bekerja sama untuk membuat mesin yang lebih kuat. Dalam jenis mesin ini, perangkat lunak memungkinkan satu lingkungan diterapkan di beberapa komputer. Ini membuatnya tampak bagi pengguna akhir seolah-olah dia menggunakan komputer tunggal, padahal sebenarnya ada banyak komputer di tempat kerja.

Server Virtual

Server virtual adalah mesin virtual yang menyediakan fungsionalitas seperti halnya server fisik. Server virtual dapat ditemukan di mana saja dan bahkan dapat dibagikan oleh banyak pemilik.

Jaringan Virtual

Jaringan virtual adalah kumpulan node virtual yang terhubung langsung dengan tautan virtual dan berdasarkan atas sumber daya fisik yang mendasarinya. Node virtual dan fisik berbicara satu sama lain melalui protokol yang umumnya lapisan 3 / protokol lapisan jaringan. Dengan kata lain, jaringan virtual adalah jaringan besar yang dibentuk oleh kombinasi kelompok jaringan yang saling berhubungan. Virtualisasi jaringan adalah teknologi di balik jaringan virtual.

Penyimpanan Virtual

Penyimpanan virtual mengambil kombinasi media penyimpanan (seperti disk, kaset, dll.) dan menggabungkannya ke dalam satu kolam penyimpanan, yang kemudian disediakan sesuai kebutuhan sebagai ruang virtual. Virtualisasi penyimpanan memungkinkan penggunaan dan pemanfaatan sumber daya yang hemat biaya. Penyimpanan virtual diakses dengan memetakan alamat virtual ke alamat fisik / nyata.

Virtual Firewall

Sebuah firewall virtualadalah perangkat lunak yang mengatur dan mengontrol komunikasi antara mesin virtual dalam lingkungan virtual. Firewall virtual memeriksa paket dan menggunakan aturan kebijakan keamanan untuk memblokir komunikasi yang tidak disetujui antara VM. Seiring dengan paket penyaringan, firewall virtual juga dapat membantu dalam menyediakan pemantauan komunikasi virtual antara VMs.

Load Balancer

Di lingkungan cloud, penyeimbang bebanmendistribusikan lalu lintas jaringan di sejumlah server virtual. Load balancer digunakan untuk menangani situasi konkurensi dan untuk memastikan bahwa kapasitas sumber daya dimanfaatkan secara optimal.

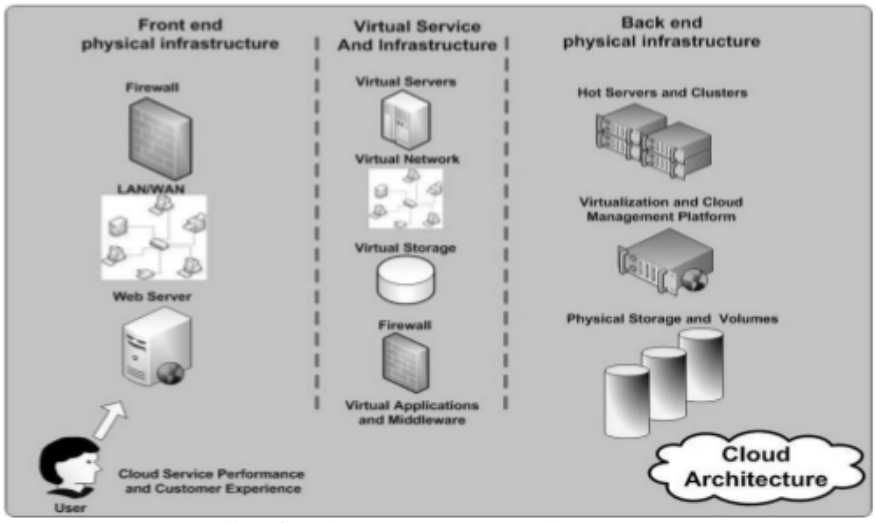
Load balancer digunakan untuk meningkatkan kinerja aplikasi secara keseluruhan dengan menyediakan server yang paling mampu melakukan tugas yang diminta. Ada beberapa teknik seperti round robin, connection at least, dan sebagainya dimana load balancer memutuskan server mana yang harus menangani lalu lintas atau tugas tertentu.

Aplikasi Virtual dan Middleware

Aplikasi virtuals dikerahkan di atas mesin virtual. Mesin virtual adalah gambar virtual yang disediakan sebagai katalog bagi pengguna untuk dipilih dan dibuat dari penyimpanan. Penyebaran aplikasi virtual lebih sederhana dan lebih cepat karena komponen yang terinstal dan dikonfigurasikan sebelumnya. Mesin virtual dapat digunakan untuk meng-host middleware, yang memungkinkan koneksi aplikasi dengan platform, jaringan, dan komponen lainnya.

Lapisan Arsitektur Cloud

Gambar 1-5 menggambarkan lapisan arsitektur cloud.



Gambar 1-5 *. Lapisan arsitektur awan*

Seperti yang ditunjukkan Gambar 1-4 , lapisan komputasi awan mungkin memiliki jaringan, firewall, dan server yang terletak di luar lingkungan cloud dan yang terhubung ke infrastruktur cloud virtual di tengah. Infrastruktur virtualisasi di-host di

server dan penyimpanan, dan dikelola oleh virtualisasi dan platform manajemen cloud.

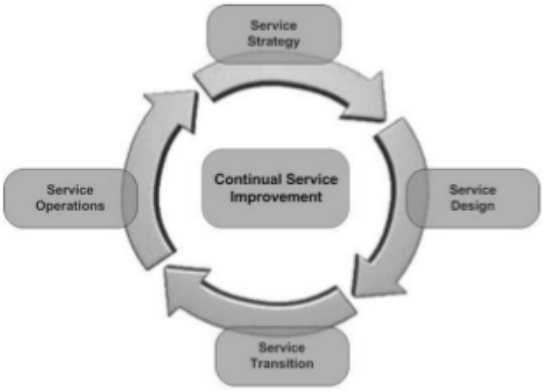
BAB DUA

MANAJEMEN KAPASITAS CLOUD

Ikhtisar Information Technology Infrastructure Library

Sebelum memahami proses manajemen kapasitas, mari kita lihat *Information Technology Infrastructure Library* (ITIL) Kerangka dari mana proses ini berasal.

Seperti yang disebutkan pada Gambar 2-1 , kerangka kerja *Information Technology Infrastructure Library* versi 3 terdiri dari lima dokumen yang menjelaskan fase-fase yang menggambarkan hampir semua proyek TI. Fase-fase ini disebut siklus hidup manajemen layanan. Judul dari fase ini adalah



***Gambar 2-1** . ITIL v3*

- Strategi layanan
- Desain layanan
- Transisi layanan
- Operasi layanan
- Perbaikan layanan berkelanjutan

strategi pelayanan Volume berfokus pada membangun tata kelola dan kebijakan di seluruh upaya manajemen layanan TI. Ini mungkin termasuk menetapkan proses keuangan, proses manajemen permintaan, dan proses manajemen portofolio layanan. Volume strategi layanan membantu perusahaan untuk menetapkan pedoman terkait kapasitas di mana kapasitas layanan dirancang.

layanan desain Volume khusus mendefinisikan proses manajemen kapasitas. Desain layanan adalah fase kunci dari setiap siklus hidup layanan dan jelas diperlukan untuk memiliki proses manajemen kapasitas yang terdefinisi dengan baik untuk mendukung layanan. Volume desain layanan juga menjelaskan elemen kunci lainnya yang merupakan bagian dari perancangan layanan TI, seperti manajemen ketersediaan, manajemen kontinuitas layanan, manajemen tingkat layanan, dan manajemen keamanan informasi.

layanan transisi dokumen berfokus pada pelaksanaan layanan baru atau memastikan pensiun dari layanan yang ada. Ini dicapai melalui manajemen perubahan dan proses manajemen rilis. Masukan dalam bentuk asumsi kapasitas

dipertimbangkan ketika memperkenalkan layanan baru. Juga, selama masa pensiun, kapasitas infrastruktur yang dibebaskan direklamasi untuk digunakan di masa depan. ***operasi layanan* fase** (atau volume) berkaitan dengan manajemen harian dari layanan yang beroperasi. Proses seperti manajemen acara, manajemen insiden, dan manajemen masalah diperkenalkan di sini. Kegagalan manajemen kapasitas akhirnya menjadi insiden, dan manajemen kapasitas yang tidak memadai sering disebut sebagai faktor yang berkontribusi atau bahkan akar masalah TI. Peningkatan ***layanan berkelanjutan*** fase memastikan bahwa layanan terus ditingkatkan dan dioptimalkan. Ini dicapai melalui pelaporan layanan, pengukuran, dan proses perbaikan. Pada fase ini, metrik dan pelaporan layanan memainkan peran penting dan dijelaskan untuk semua tindakan korektif. Masukan desain kapasitas dan tindakan perbaikan dari proses manajemen kapasitas diperhitungkan untuk keseluruhan proses dan tindakan peningkatan layanan.
Proses Peningkatan Layanan Berkelanjutan Sepanjang Siklus Hidup

Proses Peningkatan Layanan Berkelanjutan Sepanjang Siklus Hidup

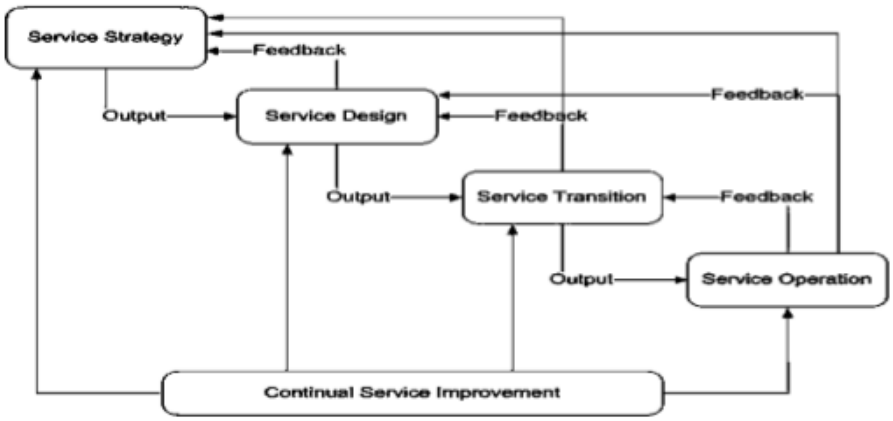
Setiap fase siklus hidup akan memberikan output ke fase siklus hidup berikutnya. Konsep yang sama ini berlaku untuk *Customer Satisfaction Index* (CSI, peningkatan layanan berkelanjutan) proses. Sebuah organisasi dapat menemukan peluang peningkatan di seluruh siklus hidup layanan. Organisasi TI perlu menunggu hingga proses layanan atau manajemen layanan dialihkan ke area operasi untuk mulai mengidentifikasi peluang peningkatan.

Agar efektif, proses CSI membutuhkan umpan balik yang terbuka dan jujur dari staf TI. Menyegmentasikan pembekalan atau ulasan ke dalam aktivitas individu yang lebih kecilselesai dalam setiap fase siklus hidup layanan dan menangkap pelajaran yang dipetik dalam fase itu membuat kebanyakan data lebih mudah dikelola. Mengumpulkan informasi ini adalah awal yang positif untuk memfasilitasi perbaikan di masa depan

Proses *Customer Satisfaction Index* (CSI, peningkatan layanan berkelanjutan) akan menggunakan secara ekstensif metode dan praktik yang ditemukan dalam banyak proses ITIL, seperti manajemen masalah, manajemen

ketersediaan, dan manajemen kapasitas yang digunakan di seluruh siklus hidup suatu layanan. Penggunaan output, dalam bentuk aliran, matriks, statistik, atau laporan analisis, akan memberikan wawasan yang berharga ke dalam desain dan pengoperasian layanan. Informasi ini, dikombinasikan dengan persyaratan bisnis baru, spesifikasi teknologi, kemampuan TI, anggaran, tren, dan kemungkinan persyaratan legislatif dan peraturan eksternal akan sangat penting bagi proses peningkatan layanan berkelanjutan untuk menentukan apa yang perlu ditingkatkan, memprioritaskan, dan menyarankan perbaikan, jika diperlukan.

Proses CSI sendiri tidak akan dapat mencapai hasil yang diinginkan. Oleh karena itu penting untuk meningkatkan aktivitas dan inisiatif peningkatan layanan berkelanjutan pada setiap fase siklus hidup layanan. Ini ditunjukkan pada Gambar 2-2 .



***Gambar 2-2** . Perbaikan layanan berkelanjutan lintas fase*

Mekanisme Peningkatan Umpan Balik Perbaikan Layanan

Proses CSI harus memastikan bahwa proses *Information Technology Service Management*, (ITSM, Manajemen Layanan Teknologi Informasi) dikembangkan dan digunakan untuk mendukung pendekatan manajemen layanan end-to-end kepada pelanggan bisnis. Sangat penting untuk mengembangkan strategi peningkatan berkelanjutan yang berkesinambungan untuk setiap proses serta layanannya**.**

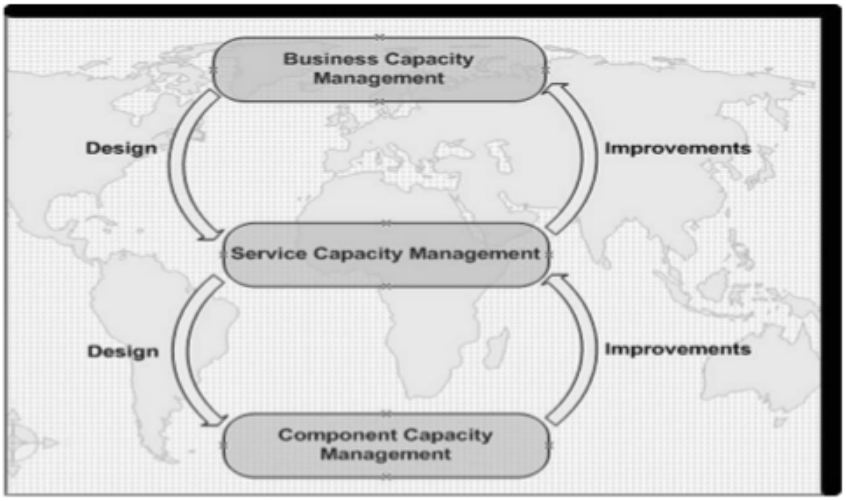
Ikhtisar Manajemen Kapasitas

Tujuan dari proses manajemen kapasitas adalah untuk memastikan bahwa kapasitas TI yang dapat dibebani biaya, di semua bidang TI, selalu ada dan disesuaikan dengan kebutuhan bisnis saat ini dan masa depan yang disepakati, pada waktu yang tepat manajemen kapasitas adalah salah satu proses desain layanan sebagaimana didefinisikan oleh *Information Technology Infrastructure Library* v3. Desain layanan adalah fase siklus hidup layanan TI yang mengubah strategi layanan menjadi layanan TI yang dapat diimplementasikan. Fase desain layanan meliputi prinsip-prinsip seperti manajemen ketersediaan layanan, manajemen kapasitas, manajemen kontinuitas layanan TI, dan manajemen keamanan. Untuk layanan berbasis cloud apa pun, manajemen kapasitas memainkan peran penting dalam memastikan pemanfaatan sumber daya, kinerja, dan efektivitas biaya secara optimal. Tujuan utama dari manajemen kapasitas adalah untuk menjaga kapasitas sumber daya yang optimal dan hemat biaya. Sumber daya ini mungkin berupa fasilitas, perangkat keras, perangkat lunak, atau sumber daya manusia. Sebagai tambahan,

Proses manajemen kapasitas bekerja dengan area proses manajemen layanan TI lainnya seperti manajemen keuangan, manajemen permintaan, dan manajemen portofolio layanan untuk memastikan bahwa kinerja layanan tetap terjaga dan Anda menghindari kehabisan sumber daya.

Dari perspektif manajemen layanan, proses manajemen kapasitas memiliki tiga pandangan yang saling terkait (Gambar 2-3):

- Manajemen kapasitas bisnis
- Manajemen kapasitas layanan
- Manajemen kapasitas komponen



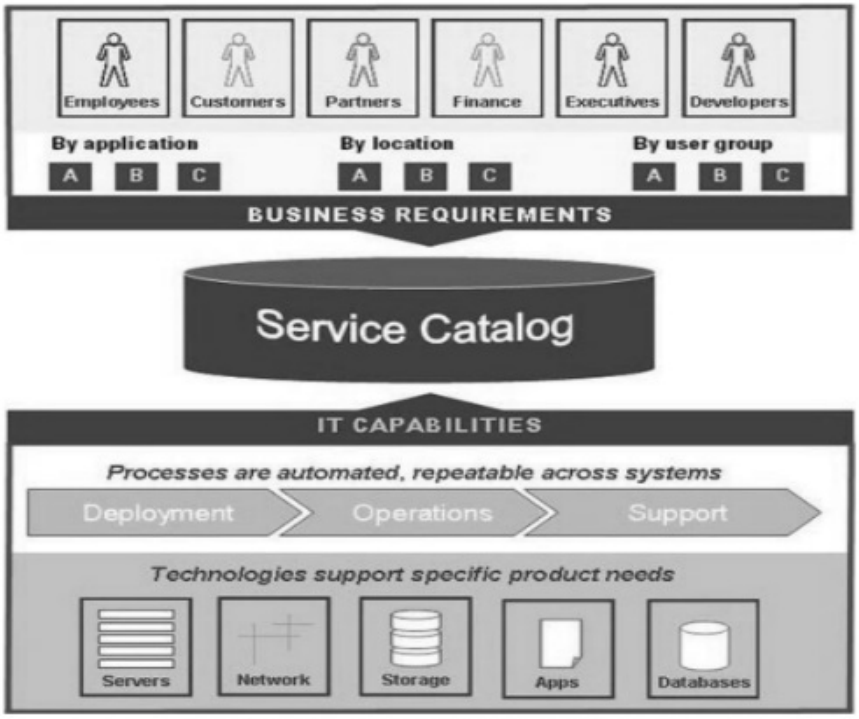
Gambar 2-3 . *Lapisan manajemen kapasitas*

Pembedaan ini dibuat untuk mencapai tujuan manajemen kapasitas pada tingkat yang berbeda. Area proses manajemen kapasitas harus memenuhi berbagai lapisan manajemen kapasitas. Misalnya, menentukan subproses persyaratan kapasitas keseluruhan atau prosedur dalam proses manajemen kapasitas yang berfokus pada manajemen kapasitas bisnis, daripada dua lapisan kapasitas lainnya (kapasitas layanan dan kapasitas komponen).

Kami akan membahas prosedur ini secara rinci dalam bab-bab selanjutnya. Ketiga lapisan kapasitas ini memiliki interaksi dinamis satu sama lain. Sebagai contoh, input desain kapasitas diterjemahkan ke bawah hirarki dari persyaratan bisnis ke persyaratan layanan untuk persyaratan tingkat komponen. Di sisi lain, input peningkatan (optimalisasi biaya dan peningkatan kinerja) mengalir dari tingkat komponen hingga ke tingkat bisnis. Manajemen kapasitas mengambil pendekatan siklus dimana kebutuhan bisnis mengalir dari atas ke bawah dan kinerja dan umpan balik optimasi bergerak dari bawah ke atas.

Penyedia cloud juga dapat mencari kebutuhan kapasitas bisnis dari survei pelanggan dan laporan optimalisasi kinerja untuk meningkatkan kinerja, fitur, dan cakupan layanan bisnis. Selain itu, input ini juga dapat ditangkap dengan menganalisis permintaan layanan yang dapat berupa permintaan untuk fitur baru atau kinerja yang lebih tinggi (lebih banyak RAM, CPU, dll.). Penyedia cloud dapat memperbarui katalog layanannya berdasarkan permintaan layanan berkala dan memastikan pelanggan dilayani dengan layanan yang tepat pada waktu yang tepat. Gambar 4-

4 menampilkan bagaimana bisnis dan persyaratan TI berkumpul untuk membuat katalog layanan.



Gambar 2-4 . *Persyaratan katalog layanan*

Aktivitas Manajemen Kapasitas

Manajemen kapasitas melibatkan kegiatan proaktif dan reaktif. Berikut ini daftar beberapa kegiatan proaktif :

- Mengambil tindakan pada masalah kinerja sebelum terjadi.
- Memperkirakan kebutuhan kapasitas masa depan dengan analisis tren dan pemanfaatan.
- Pemodelan dan tren perubahan yang diprediksi dalam layanan TI, dan mengidentifikasi perubahan yang perlu dilakukan untuk layanan.
- Memastikan bahwa peningkatan dianggarkan, direncanakan, dan diimplementasikan sebelum SLA dan target layanan dilanggar atau masalah kinerja muncul.
- Tuning dan mengoptimalkan kinerja layanan dan komponen.

Kegiatan reaktif termasuk

- Meninjau kinerja saat ini dari kedua layanan dan komponen.
- Bereaksi dan membantu dengan masalah kinerja tertentu.
- Menanggapi semua peristiwa ambang terkait kapasitas dan menghasut tindakan korektif.

A Balancing Act

Manajemen kapasitas pada dasarnya adalah tindakan penyeimbang yang memastikan bahwa kapasitas dan kinerja layanan TI dalam suatu organisasi digunakan dengan cara yang paling efektif dan tepat waktu. Tindakan ini termasuk:

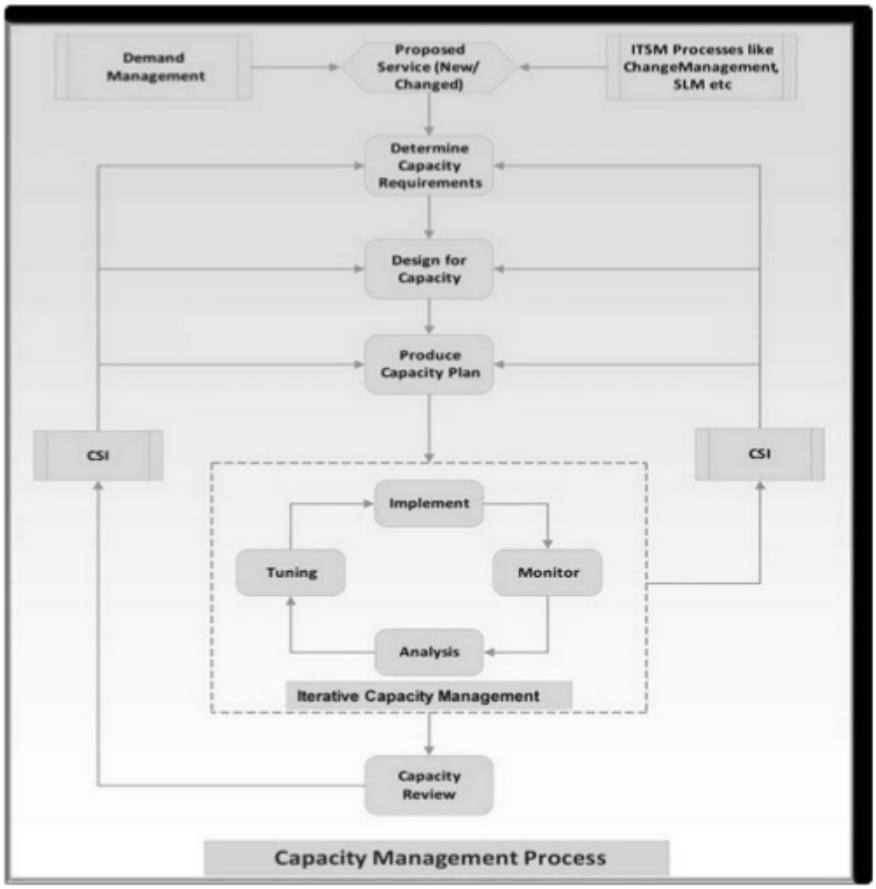
- Menyeimbangkan biaya terhadap sumber daya yang dibutuhkan.
- Menyeimbangkan penawaran terhadap permintaan.

Penyeimbangan ini menjadi parameter kunci yang diperlukan untuk berhasil menjalankan layanan cloud. Dengan demikian, penyedia cloud diharuskan memiliki kapasitas yang cukup untuk memenuhi beragam kebutuhan sesuai permintaan dari pelanggan sambil memastikan bahwa pemborosan sumber daya diminimalkan dan penyediaan kapasitas yang berlebihan dijaga agar tetap minimum. Ini adalah tantangan teknologi dan analisis yang harus dihadapi setiap penyedia cloud untuk menyediakan layanan dalam lingkungan yang kompetitif. Jika penyedia cloud kehabisan kapasitas, penyedia menjalankan risiko kehilangan pelanggan atau lebih buruk, dan jika penyedia cloud menyediakan kapasitas ekstra besar, biaya layanan akan naik dan penyedia menjalankan risiko tidak kompetitif dalam harga.

Manajemen Kapasitas: Ruang Lingkup dan Cakupan

Gambar 2-5 menggambarkan pandangan proses manajemen kapasitas yang sebagian besar perusahaan ikuti sambil merancang kapasitas layanan mereka. Dalam model tradisional Proses manajemen kapasitas mengambil masukan dari area proses manajemen layanan lainnya seperti manajemen permintaan untuk pola dan prakiraan permintaan, manajemen tingkat layanan untuk data terkait kinerja layanan yang disepakati, manajemen perubahan untuk permintaan perubahan layanan, dll. Berdasarkan input ini, kapasitas persyaratan dalam hal ruang pusat data, sumber daya virtual seperti server, bandwidth jaringan, dan persyaratan penyimpanan

termasuk infrastruktur yang mendasarinya ditentukan. Semua komponen yang diperlukan untuk mendukung layanan bisnis dikumpulkan sebagai langkah pertama. Seperti disebutkan, Gambar 2-5 memperlihatkan proses manajemen kapasitas. Proses ini dimulai dengan mengumpulkan persyaratan terkait kapasitas dari berbagai proses Manajemen Layanan Teknologi Informasi, (ITSM, *Information Technology Service Management*) lainnya, merancang kapasitas, dan merumuskan desain kapasitas ke dalam rencana kapasitas formal. Rencana kapasitas ini digunakan untuk menyimpan semua informasi terkait kapasitas selama implementasi manajemen iteratif.



Gambar 2-5 . *Proses manajemen kapasitas dalam model tradisional*

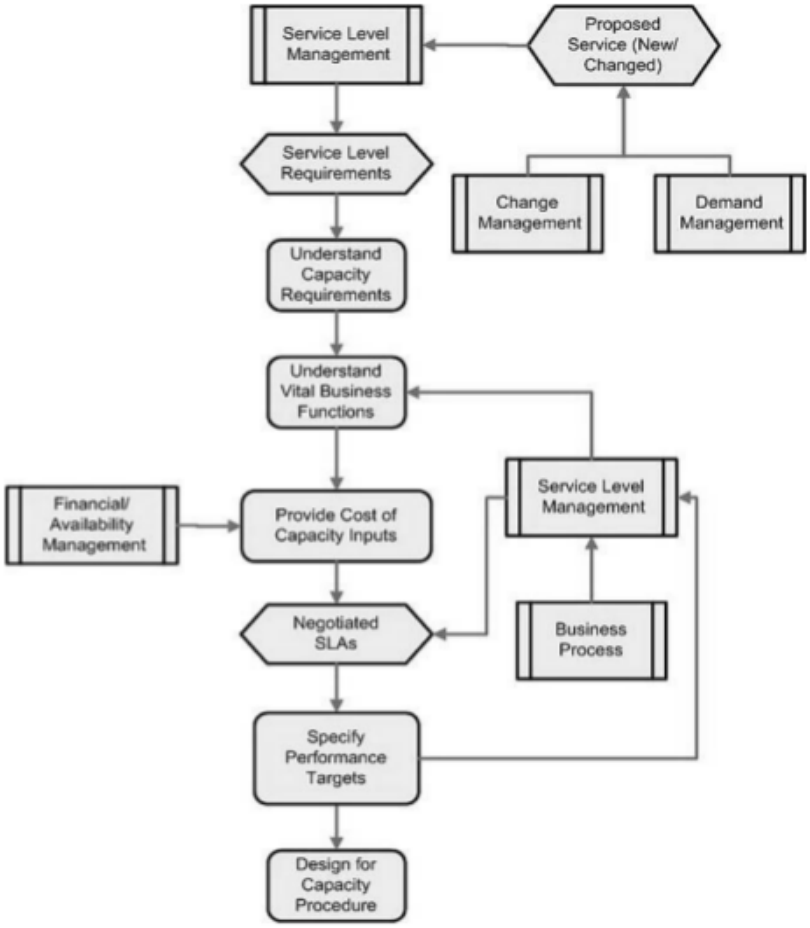
Prosedur Manajemen Kapasitas dalam Model Tradisional

Pada bagian ini, kami akan menjelaskan proses manajemen kapasitas yang terperinci.

Tentukan Persyaratan Kapasitas:

- *menentukan kebutuhan kapasitas* langkah mengambil masukan dari manajemen permintaan, manajemen tingkat layanan, dan mengubah dan melepaskan proses manajemen.
- *kebutuhan kapasitas* Langkah memastikan bahwa kebutuhan bisnis pengguna dan tingkat layanan ditangkap dijabarkan ke dalam kebutuhan kapasitas.

Gambar 2-6 menyoroti prosedur yang terlibat dalam menentukan persyaratan kapasitas untuk perencanaan kapasitas. Dalam hubungannya dengan manajemen keuangan, manajemen kapasitas memberikan perkiraan biaya untuk menggunakan persyaratan terkait kapasitas yang ditentukan dan *Service Level Agreement* (SLA) terkait. Untuk tujuan ini, *Service Level Agreement* (SLA) diterjemahkan ke dalam target kinerja spesifik yang akan didukung oleh manajemen kapasitas. Ini juga berfungsi sebagai dasar untuk negosiasi perjanjian tingkat layanan. Prosedur ini akan dijelaskan secara rinci dalam bab-bab selanjutnya.



Gambar 2-6 . *Menentukan persyaratan kapasitas*

Desain untuk Kapasitas

Setelah ini, arsitektur desain untuk kebutuhan kapasitas yang diharapkan dikembangkan. Tim desain kapasitas memformulasikan desain infrastruktur TI (terdiri dari semua komponen) yang diperlukan untuk memenuhi target kinerja tertentu. Prosedur ini sangat penting dan menentukan keberhasilan atau kegagalan proses manajemen kapasitas.

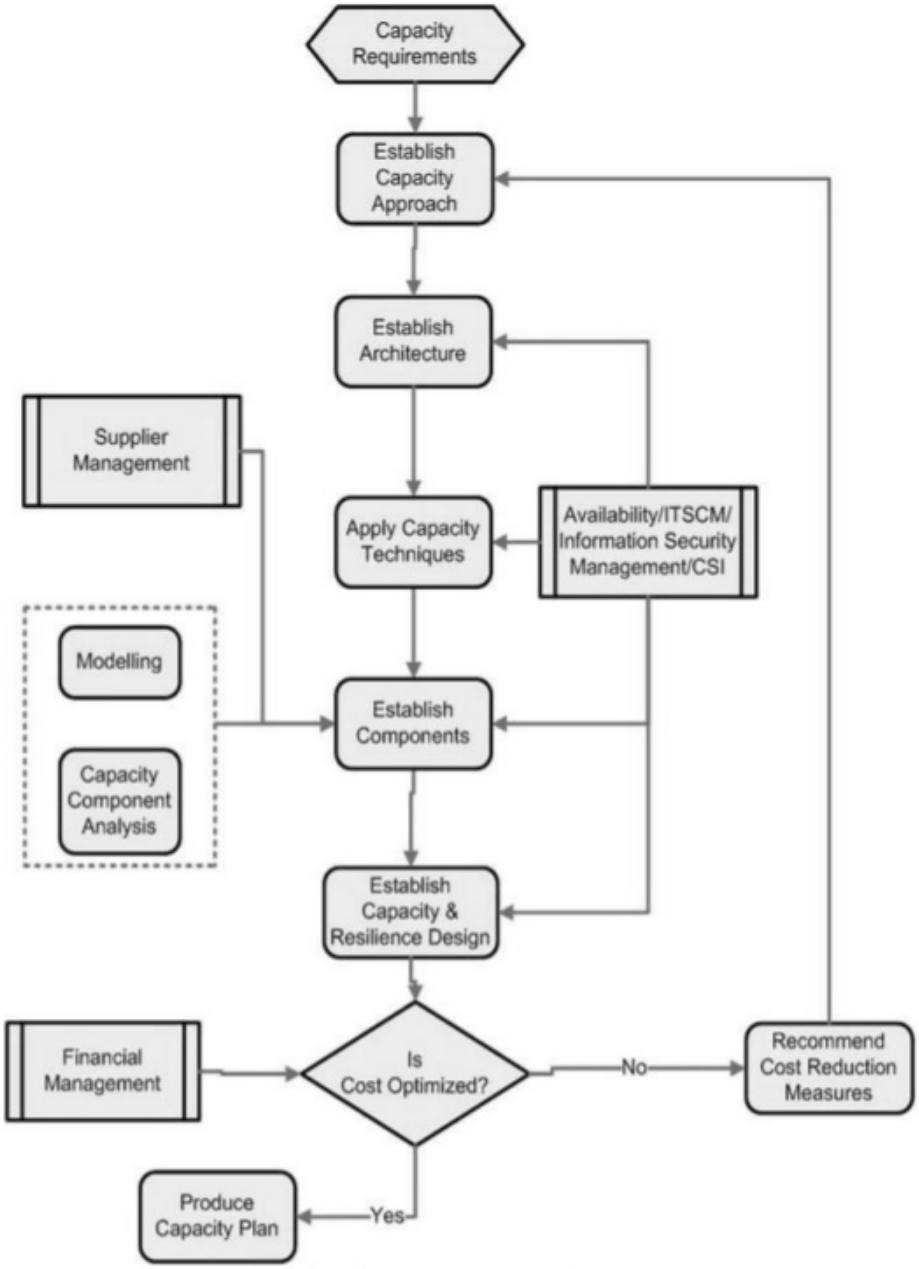
Berdasarkan target kinerja, rencana tingkat tinggi untuk memenuhi target dipilih. Misalnya, fokus biaya yang kaku dalam target kinerja mungkin memerlukan solusi tepat waktu sementara persyaratan terkait kesinambungan layanan dan target

kinerja yang kaku mungkin memerlukan solusi yang mencakup ketersediaan kapasitas margin.

Selama desain kapasitas, satu atau lebih teknik, seperti analisis dampak kegagalan komponen dan manajemen risiko, digunakan untuk mengoptimalkan desain kapasitas. Berbagai algoritma statistik juga dapat digunakan untuk menetapkan tingkat kapasitas komponen untuk penyimpanan, jaringan, dan server. Desain ketahanan juga dibangun atas dasar kebutuhan pemulihan bencana. Sebagai contoh, persyaratan kapasitas dapat digandakan dalam skenario di mana situs pemulihan bencana lain (DR, *Disaster Recovery*) akan dipertimbangkan.

Pertimbangan desain seperti penggunaan klaster digunakan untuk desain kapasitas. Selain itu, spesifikasi tingkat komponen seperti make, model, nama vendor, dan konfigurasi dilakukan untuk mematuhi target kinerja layanan. Tim pemasok / manajemen vendor juga dipanggil untuk memberikan masukan berharga mereka untuk mendukung kegiatan ini. Oleh karena itu desain kapasitas yang efisien diselesaikan berdasarkan aspek seperti ketahanan layanan, persyaratan keamanan, spesifikasi kinerja, persyaratan ketersediaan, dan sebagainya.

Dalam hubungannya dengan kebijakan manajemen keuangan perusahaan, pemeriksaan dilakukan untuk langkah-langkah pengurangan biaya yang tersedia tanpa mengorbankan target kinerja. Gambar 2-7 menetapkan *desain untuk* prosedur kapasitas untuk manajemen kapasitas.



***Gambar 2-7** . Desain untuk kapasitas*

Prosedur Manajemen Kapasitas dalam Model Cloud

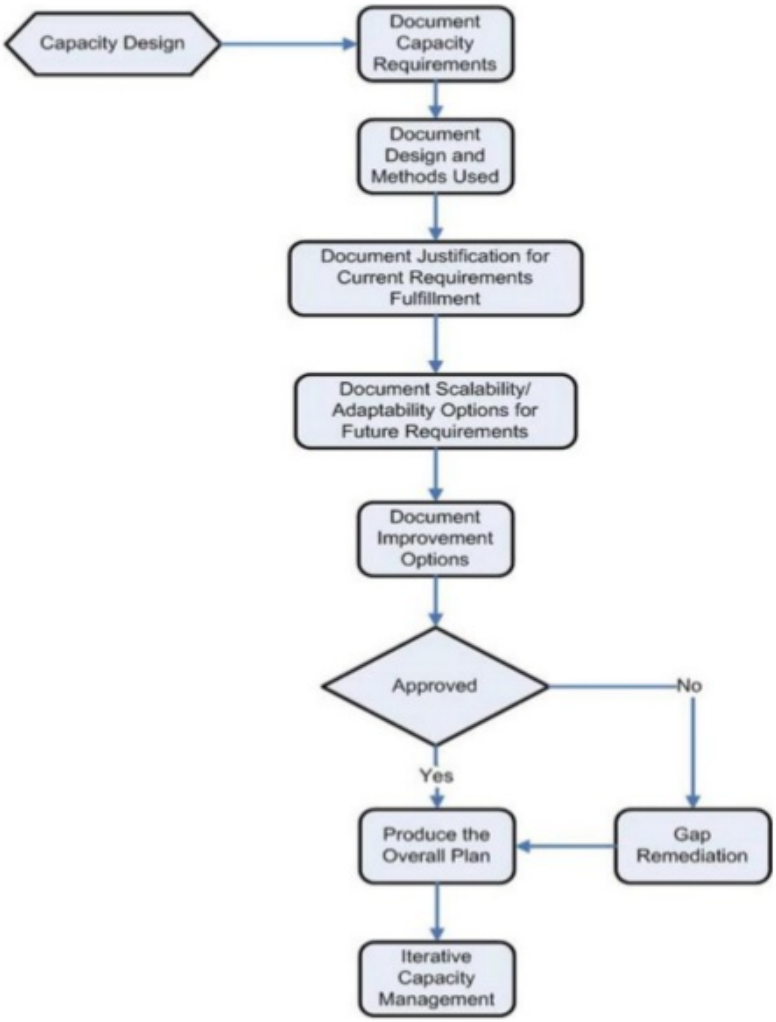
Dalam skenario cloud, desain untuk kapasitas mengambil pendekatan yang berbeda. Salah satu fokus adalah bagaimana cloud mempengaruhi pendekatan tradisional. Untuk saat ini kita akan melihat perubahan dalam model komputasi cloud dari perspektif konsumen layanan melalui hal-hal berikut:

- Merancang aplikasi untuk menggunakan kapasitas minimum dan idealnya untuk skala daripada skala.
- Merancang aplikasi untuk menyediakan metrik pada penggunaan sehingga instance baru dari aplikasi dapat diinisialisasi tergantung pada beban kerja dari perspektif aplikasi.
- Rancang aplikasi yang tidak bergantung pada memori pada satu server atau instance untuk hal-hal seperti status sesi.
- Rancang aplikasi yang memperhatikan batas operasi input / output penyimpanan per detik (*I Input Output Per Seconds* , IOPS) untuk beberapa penyedia cloud.
- Rancang aplikasi di mana database juga dapat bekerja pada beberapa sistem menggunakan replika baca dan dengan demikian memberikan skalabilitas.
- Aplikasi desain mengingat pemulihan bencana multi-regional untuk menjaga aspek pemulihan bencana. Berbagai penyedia cloud memiliki opsi berbeda yang tersedia untuk pemulihan bencana.
- Desain untuk pencadangan dan pemulihan. Cara kerja backup di lingkungan penyedia cloud tertentu berbeda. Ketika mendesain untuk kapasitas, masalah ini harus diingat.
- Desain untuk penggunaan jaringan pada penyedia cloud publik. Biaya jaringan dapat menjadi signifikan ketika memanfaatkan layanan cloud, dan aspek ini membutuhkan perhatian.

Menghasilkan Rencana Kapasitas

Berdasarkan kebutuhan kapasitas dan spesifikasi desain, rencana kapasitas formal dikembangkan untuk menangani semua aspek yang terkait dengan manajemen kapasitas; ini didokumentasikan dan dibagikan di antara semua pemangku kepentingan. Rencana kapasitas adalah sumber informasi yang terkait dengan semua persyaratan kapasitas seperti skalabilitas, kemampuan beradaptasi, arsitektur kapasitas komponen, dll. Rencana kapasitas harus dipelihara dan ditinjau

secara berkala untuk memastikan bahwa informasi selalu terkini sesuai dengan kapasitas saat ini dan masa depan. Persyaratan. Gambar 2-8 menguraikan kegiatan yang terlibat dalam menghasilkan rencana kapasitas.



***Gambar 2-8** . Menghasilkan rencana kapasitas*

Manajemen Kapasitas Iteratif untuk Layanan Langsung

Setelah desain dan rencana kapasitas diselesaikan untuk layanan baru, ia bergerak ke dalam produksi, dan di sini manajemen kapasitas mengambil pendekatan

yang berbeda sama sekali. Kita dapat menyebut manajemen kapasitas iteratif ini , dan ini melibatkan penerapan rencana kapasitas, pemantauan kinerja layanan, menganalisis laporan pemanfaatan, dan menyetel kapasitas untuk peningkatan berkelanjutan dan menangani fluktuasi kinerja. Manajemen kapasitas di sini lebih terfokus pada pengoptimalan kapasitas dan manajemen kinerja.

Empat fase manajemen kapasitas berkelanjutan / berulang dapat diringkas sebagai; Pelaksanaan, Pemantauan, Analisis, Tuning

Pelaksanaan

Tujuan dari fase ini adalah untuk mengadopsi pendekatan terstruktur terhadap pelaksanaan rencana kapasitas dengan mengidentifikasi semua persyaratan pelaksanaan dan bidang dampak. Setiap penambahan kapasitas / penghapusan / konfigurasi ulang dievaluasi dalam hal dampaknya pada operasi layanan saat ini. Setiap persyaratan dalam hal konfigurasi teknis, pengadaan perangkat keras dan perangkat lunak, biaya pelaksanaan, persyaratan perizinan, dan keterampilan yang diperlukan untuk pelaksanaan diidentifikasi untuk mencapai tujuan yang diinginkan dari usulan pelaksanaan proyek. Persyaratan pemantauan dan pelaporan juga diidentifikasi dan dirumuskan di sini. Setelah ini, rencana kegiatan rinci yang akan dilakukan dibuat; urutan, hubungan, upaya dan waktu yang dibutuhkan dikembangkan untuk implementasi proyek yang diusulkan.

Memantau Rencana

Fase ini difokuskan pada pengukuran dan pelaporan aspek kapasitas terkait kinerja. Semua kegiatan dalam fase ini sebaiknya dilakukan oleh mekanisme otomatis. Kapasitas dan pemantauan kinerja secara real-time dilakukan dalam proses manajemen acara atau alat pemantauan. Proses manajemen acara menggunakan masukan dari manajemen kapasitas untuk merancang sistem manajemen acara (ambang batas, parameter, dll.) Yang membentuk basis pemantauan real-time. Pemantauan aspek-aspek kapasitas yang teridentifikasi selama periode kinerja yang ditetapkan tercakup dalam ruang lingkup fase ini.

Analisis

Tujuan dari fase ini adalah untuk menerjemahkan data dari pemantauan kapasitas menjadi informasi yang dapat digunakan untuk menarik kesimpulan untuk menyarankan perbaikan atau mengidentifikasi masalah. Tergantung pada jenis pemodelan yang sedang dilakukan, data kinerja dari pemantauan diterjemahkan ke dalam bentuk yang dapat digunakan sebagai input ke model.

Data yang dilaporkan kepada pemilik kapasitas fungsional cenderung dan dianalisis menggunakan alat statistik. Laporan-laporan ini kemudian digunakan untuk pengambilan keputusan pada penyetelan kapasitas dan tindakan pengoptimalan. Ini mungkin termasuk laporan tentang ambang batas, kurangnya kapasitas yang konsisten, konfigurasi salah, pola pemuatan yang tidak biasa, dll.

Tuning

Fase tuning dalam manajemen kapasitas berkaitan dengan identifikasi solusi untuk pemenuhan tujuan manajemen kapasitas.

Profil beban kerja dari berbagai layanan dibuat; ini termasuk komponen-komponen yang menyediakan layanan dan beban kerja yang komponen-komponen ini servis. Profil beban kerja memberikan wawasan tentang bagaimana layanan dan berbagai komponen di bawahnya bekerja dalam berbagai skenario.

Laporan analisis, yang merupakan keluaran dari fase analisis, digunakan dalam fase tuning untuk menyesuaikan pemantauan kapasitas. Tindakan untuk penyetelan dapat menambahkan parameter baru untuk pemantauan, penambahan laporan baru, dan mengubah ambang batas untuk layanan atau komponen.

Semua perubahan mengikuti proses manajemen perubahan dan diimplementasikan setelah persetujuan.

Tinjauan Kapasitas

Terakhir, tinjauan manajemen kapasitas dilakukan di mana laporan kapasitas dibuat dan disajikan kepada semua pemangku kepentingan di tingkat bisnis, layanan, dan komponen. Semua temuan dan tindakan pasca-review kembali dimasukkan ke dalam prosedur manajemen kapasitas untuk perbaikan berkelanjutan.

Ulasan mingguan / bulanan dapat terjadi di mana peninjauan kinerja untuk aspek kapasitas dilakukan. Nilai aktual parameter kinerja seperti ini dibandingkan dengan nilai target dan target:

- Parameter kapasitas komponen seperti disk, server, penggunaan memori, dll.
- Parameter kapasitas layanan seperti waktu kerja layanan e-mail, kinerja, dll.
- Parameter kapasitas bisnis seperti peningkatan / penurunan jumlah pengguna bisnis, dll.
- Memproses indikator kinerja utama
- Integrasi proses
- Item tindakan periodik sebelumnya
- Data dari proses lain yang memengaruhi kapasitas (yaitu Insiden / masalah / data perubahan)

Penyimpangan dibahas dan analisis dilakukan.

Berdasarkan tinjauan kapasitas, item pembaruan yang dihasilkan untuk rencana kapasitas didokumentasikan. Rencana kapasitas diperbarui ketika semua poin yang didokumentasikan mendapatkan persetujuan yang diperlukan.

Seperti yang telah dibahas sebelumnya, lapisan manajemen kapasitas mencakup manajemen kapasitas sumber daya atau tingkat komponen, manajemen tingkat layanan, dan bisnis-tingkat kapasitas manajemen untuk semua layanan dan server. Manajemen kapasitas menangani kapasitas TI di semua tiga lapisan di mana kebutuhan kapasitas terpenuhi dari atas ke bawah.

Persyaratan kapasitas berorientasi bisnis pertama kali diterjemahkan ke kebutuhan yang terkait dengan layanan. Selanjutnya, parameter terkait layanan dipetakan dengan lapisan komponen yang mendasari yang mendukung layanan TI.

MENETAPKAN SASARAN UNTUK KAPASITAS

Perencanaan kapasitas melibatkan banyak asumsi yang terkait dengan mengapa membutuhkan kapasitas. Beberapa asumsi ini jelas, sedangkan yang lain tidak. Sebagai contoh, jika Anda tidak tahu bahwa Anda harus melayani halaman dalam waktu kurang dari tiga detik, Anda akan memiliki waktu yang sulit menentukan berapa banyak server akan diperlukan untuk memenuhi persyaratan itu. Lebih penting lagi, akan lebih sulit untuk menentukan berapa banyak server yang perlu Anda tambahkan ketika lalu lintas data bertambah

BAB TIGA

MENGELOLA KAPASITAS CLOUD

Bagian ini memperkenalkan perencanaan kapasitas dan manajemen pada berbagai lapisan: bisnis, layanan dan kapasitas komponen. Jika kita mencoba untuk menempatkan manfaat dari cloud computingdalam istilah yang lebih sederhana, itu akan menjadi "pemanfaatan sumber daya yang dioptimalkan dan penghematan biaya"; yang bertepatan dengan tujuan dan sasaran manajemen kapasitas. Dalam bab ini penekanan telah diberikan kepada apa yang merupakan proses manajemen kapasitas dan bagaimana berbagai pemangku kepentingan cloud melihat manajemen kapasitas. Bab ini juga menjelaskan bahwa bagaimana proses manajemen kapasitas dapat diterapkan pada komputasi awan untuk menghasilkan ekonomi dan optimisasi infrastruktur .

Manajemen Kapasitas dalam Komputasi Awan

Disini penekanan telah diberikan pada implementasi prosedur dan aktivitas manajemen kapasitas yang secara khusus dirancang untuk lingkungan cloud. Menerapkan manajemen kapasitas dapat menjadi rumit dan mahal karena infrastruktur yang heterogen dan kompleks serta toolset yang terkait. Sekarang mari kita lihat bagaimana proses manajemen kapasitas direncanakan dan diterapkan dalam lingkungan berbasis cloud.

Dalam lingkungan cloud, penyedia layanan harus merencanakan untuk mengelola kapasitas pusat data mereka dan memastikan tingkat kinerja layanan dan kontinuitas tertinggi.

Masukan untuk menentukan persyaratan kapasitas diambil dari alat yang memantau kinerja sumber daya, tren bisnis, proses manajemen layanan cloud lainnya seperti manajemen permintaan layanan, manajemen tingkat layanan, manajemen portofolio layanan, dan manajemen perubahan.

Awan memperkenalkan aspek multi penyewaan dan infrastruktur bersama yang disewakan atau disewa daripada dibeli dan, seperti yang telah kita bahas sebelumnya, itu mengubah cara manajemen kapasitas dilakukan.

Antarmuka proses yang didefinisikan ulang akan diperlukan untuk perusahaan yang ingin menggunakan pembayaran karena Anda pergi model dan model elastisitas yang ditawarkan di awan, untuk secara efektif mengelola kapasitas cloud.
Untuk konsumen layanan, hubungan yang lebih erat dengan manajemen keuangan perusahaan akan menjadi kunci untuk memahami biaya yang terkait dengan berbagai opsi publik, swasta, hibrida, dll. dan menggunakan informasi ini untuk menilai mana yang paling sesuai dengan kebutuhan bisnis. Penentuan biaya dan ukuran lingkungan dengan benar akan sangat penting dalam memastikan bahwa menggunakan cloud benar-benar membayar kembali bisnis seperti yang diharapkan.
Mari kita pertimbangkan skenario kehidupan nyata di mana bisnis memerlukan beberapa server high-end untuk menjadi tuan rumah beberapa aplikasi selama satu tahun. Pertimbangkan juga skenario pemanfaatan server di mana server dapat digunakan pada tingkat rendah, sedang, dan tinggi.
Untuk menjaga agar tetap sederhana ini, ambil satu server dan bandingkan biaya dari hosting DC konvensional versus hosting cloud publik.

Skenario A (Gambar 3-1) memberikan biaya hosting dari server high-end. Untuk menjaga sederhana ini kita tidak menghitung elemen biaya lain dalam lingkungan hosting DC seperti ruang, daya listrik, dan rak.

Skenario B (Gambar 3-2) menunjukkan biaya yang dibutuhkan untuk menghosting server yang sama pada cloud publik. Elemen biaya termasuk biaya dimuka sekaligus dan hosting tahunan.
Mempertimbangkan di atas dua skenario, ekonomi biaya dapat dengan mudah dilihat pada Gambar 3-3 . Cloud hosting umumnya jauh lebih ekonomis daripada hosting DC untuk beban kerja variabel. Angka-angka biaya ini dalam dolar AS.

Skenario A

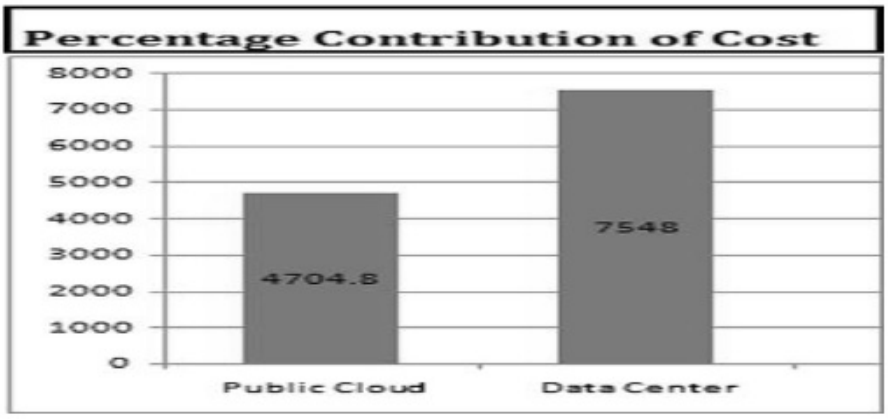
Vendor Name: Data Center	
DC Hosting Cost Components 1 Large Instance (7.5 GB, 4 ECU, 850 GB HDD)	
Essential Components	**Cost $ (yearly)**
Space	
Power	
Rack	
Compute	7548
Network	
Bandwidth	
Storage	
Set up	
Maintenance	
Top Up	
Support	
OS	
Security	
Monitoring	
Total Cost	7548

Gambar 3-1. *Model Berbasis Server Tradisional*

Skenario B

Provider Name: Public Cloud	
Cloud provider cost elements 1 Large Instance (7.5 GB, 4 ECU, 850 GB HDD)	
Essential Components	Cost $ (yearly)
Option1: Enter Overall cost	
Consolidated Cost (Source: Website)	4204.8
Option 2: Enter component-wise break up	
CPU	
Memory	
Storage	
I/O Performance	
NW Bandwidth	
OS	
Monitoring	
Upfront Cost	500
Total Cost	4704.8

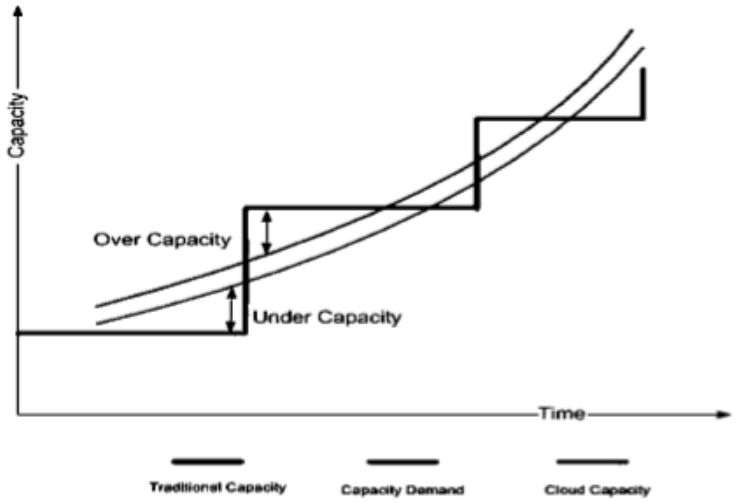
Gambar 3-2 . *Model Server Berbasis Cloud*



Gambar 3-3 *. Pembagian biaya . Hosting Cloud vs . DC*

Kurva Pemanfaatan Kapasitas

Grafik pada Gambar 3-4 menggambarkan kapasitas versus pemanfaatan.



Gambar 3-4 . *Kapasitas versus pemanfaatan*

Kurva ini menunjukkan tema inti di sekitar layanan cloud versus konsumsi layanan sebenarnya.

Penting untuk memahami bahwa ekonomi cloud dapat dipengaruhi oleh penyediaan sumber daya cloud yang berlebihan dan berlebihan. Dari perspektif perusahaan, setelah perangkat keras dibeli dan dibayar dalam skenario permintaan yang menurun maka akan ada kelebihan kapasitas yang akan sia-sia. Juga organisasi yang membuat keputusan dimuka untuk membeli perangkat keras mengambil risiko dari skenario penurunan permintaan.

Dalam model cloud, risiko teknologi menjadi usang atau permintaan bisnis turun diambil oleh penyedia cloud. Dari perspektif manajemen keuangan organisasi membuat pembayaran dimuka untuk perangkat keras dan dengan demikian biaya modal perlu menjadi faktor untuk kehidupan perangkat keras.

Jadi untuk dijumlahkan:

Biaya dalam model tradisional akan menjadi ringkasan dari berikut:

- Biaya fasilitas pusat data termasuk hosting, daya, dll.
- Perangkat keras server
- Peralatan jaringan
- Perangkat penyimpanan

- Komponen perangkat lunak dan lisensi
- Perawatan tahunan untuk perangkat keras dan perangkat lunak
- Implementasi atau penyediaan / implementasi
- Bandwidth jaringan
- Operasi dan pemeliharaan
- Sistem operasi
- Perangkat lunak virtualisasi
- Perangkat lunak pemantauan dan manajemen
- Sumber daya operasi
- Biaya modal

Biaya di atas perlu diperhitungkan untuk masa pakai perangkat keras yang bergantung pada kebijakan pembaruan perangkat keras organisasi dan biasanya sekitar 4 tahun.

Dari model cloud Perspektif biaya dapat sebagai berikut:

- Biaya dimuka untuk pemesanan atau pemesanan contoh
- Biaya bayar per penggunaan komputasi, penyimpanan dan sumber daya jaringan
- Biaya transaksi (IOPS / GET / PUT). Ini adalah biaya yang tidak berlaku dalam model pusat data pribadi dan karenanya organisasi perlu memperhatikan biaya per transaksi.
- Biaya bandwidth jaringan (ini dapat jauh lebih tinggi dalam skenario awan daripada skenario tradisional)
- Biaya migrasi ke lingkungan cloud
- Alat pemantauan dan manajemen
- Beban operasi
- Lisensi perangkat lunak
- Biaya migrasi lisensi (jika berlaku)
- Biaya pemeliharaan tahunan yang berlaku
- Biaya modal

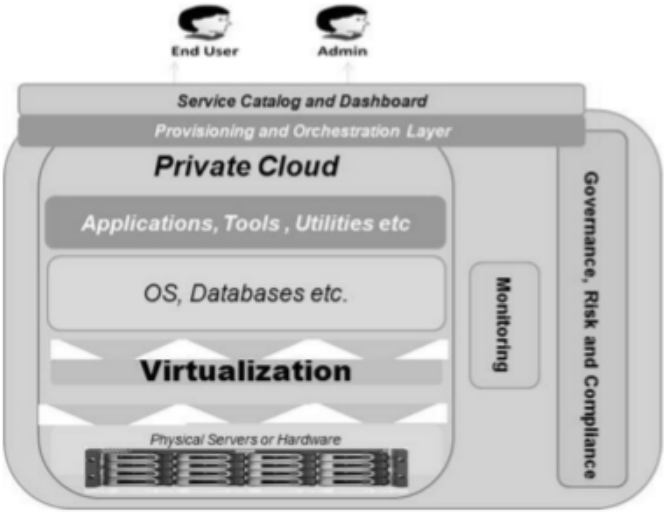
Terlepas dari hal di atas, risiko teknologi semakin usang dan risiko perampingan adalah aspek penting yang perlu ditelaah untuk melakukan perbandingan awan dengan opsi tradisional.
Proyeksi biaya dari model cloud atas siklus penyegaran perangkat keras akan memberikan perbandingan biaya antara cloud dan lingkungan tradisional.
Namun model di atas bersifat simplistik dan kompleksitas berikut perlu ditangani:

- Pengurangan biaya oleh penyedia cloud selama periode tersebut. Ini adalah aspek penting karena penyedia cloud dikenal untuk mengurangi biaya setiap beberapa bulan dan ini dapat menjadi pengurangan yang signifikan.
- Biaya sunk hardware dan software yang sudah diperoleh.
- Biaya perangkat keras atau perangkat lunak yang akan tetap berada di pusat data mengirim migrasi ke cloud.
- Biaya layanan bersama yang diperlukan dalam awan hibrida. Bagian apa ini akan dibagikan ke infrastruktur cloud.
- Perubahan kebijakan penyegaran perangkat keras. Apa yang terjadi jika sebuah organisasi memutuskan untuk mengubah siklus penyegaran perangkat keras dari 4 tahun menjadi 5 tahun.

Pandangan Konvensional vs. Cloud tentang Manajemen Kapasitas

Gambar 3-5 di bawah ini menggambarkan arsitektur / model cloud pribadi tingkat tinggi yang dibangun di atas infrastruktur perangkat keras, virtualisasi dan otomatisasi / lapisan orkestrasi.
Sebagaimana dibahas sebelumnya, Private Cloud adalah salah satu model penyebaran cloud yang disukai di antara UKM dan perusahaan besar. Biasanya penyedia layanan cloud untuk perusahaan-perusahaan ini melakukan pembangunan cloud pribadi mereka berdasarkan kesiapan aplikasi dan anggaran di tangan untuk memindahkan aplikasi ke cloud. Cloud enablement termasuk membangun dan mengkonfigurasi infrastruktur perangkat keras, menerapkan teknologi virtualisasi dan lapisan orkestrasi / otomatisasi.
Mungkin ada skenario di mana pelanggan sudah memiliki perangkat keras dan kadang-kadang lingkungan yang tervirtualisasi; dalam situasi ini penyedia cloud mengambil alih dari kondisi infrastruktur perangkat keras saat ini dan mengubah lingkungan menjadi pengaturan berbasis cloud. Hal ini dilakukan oleh virtualisasi infrastruktur dan mengatur lingkungan dan membawa karakteristik cloud computing

seperti pengumpulan sumber daya, skalabilitas, pengukuran, penagihan, chargeback, dan akses jaringan.



***Gambar 3-5** . Model Referensi Private Cloud*

Sebagaimana dibahas, virtualisasi adalah enabler kunci untuk memastikan pemanfaatan sumber daya yang dioptimalkan dari penyedia cloud. Komponen lingkungan cloud dapat berupa fasilitas pusat data, sasis, perangkat keras, mesin virtual, perangkat jaringan, bandwidth, cakram virtual, wilayah geografis, zona yang tidak memiliki isolasi, penyimpanan arsip, dan sebagainya.
Dalam lingkungan cloud, manajemen kapasitas harus dapat mengatasi masalah berikut:

- Kemudahan penyediaan kapasitas yang mengarah ke kelebihan penyediaan dan masalah seperti VM sprawl.
- Portal swalayan menyulitkan untuk memperkirakan permintaan kapasitas.
- Mekanisme chargeback yang tidak efisien dan kompleks karena hosting multi penyewa di lingkungan berbagi sumber daya.
- Infrastruktur dinamis menyebabkan inefisiensi alokasi manual dan rentan terhadap kesalahan.

Untuk melakukan perencanaan yang efektif, penyedia layanan dan pelanggan dapat bekerja berdampingan dan merencanakan kapasitas masa depan. Penyedia Cloud dapat memberikan kinerja dan tren penggunaan kapasitas melalui toolset pemantauan. Pelanggan Cloud dapat berbagi permintaan bisnis dan rencana ekspansi. Ini membantu mencapai persyaratan kapasitas yang tepat yang dapat memenuhi persyaratan kinerja dan mendukung bisnis. Penyedia Cloud perlu mengadopsi pendekatan holistik dan berorientasi layanan pada manajemen kapasitas yang mencakupmetrik teknis dan non-teknis, fokus yang kuat pada hal-hal berikut:

- Mengoptimalkan penempatan beban kerja di infrastruktur
- Manajemen ambang batas kapasitas dinamis
- Pemantauan kinerja aplikasi
- Analisis waktu nyata untuk alokasi kapasitas proaktif / prediktif

Sekarang mari kita coba memahami bagaimana prosedur kapasitas konvensional disesuaikan dengan model berbasis cloud. Manajemen kapasitas dalam lingkungan tradisional dapat dipandang sebagai pendekatan yang agak pesimiskarena ada fokus untuk menyediakan kapasitas setinggi mungkin untuk mendukung aplikasi sehingga mereka dapat dijalankan pada tingkat yang diinginkan pada jam sibuk. Di luar jam sibuk, sumber daya yang diperoleh tidak aktif dan terus menimbulkan biaya yang terkait dengan ruang pusat data, pendinginan, daya, dll. Ini termasuk penyediaan kapasitas dalam hal server, penyimpanan, CPU dan bandwidth jaringan yang cukup dan sumber daya ini sebagian besar kurang dimanfaatkan. Server diperoleh sesuai kebutuhan kapasitas.

CONTOH:
Misalnya perusahaan yang mencari CPU “x” darimenghitung waktu "t" terikat untuk mendapatkan kapasitas komputasi lebih dari "x" CPU terlepas dari pemanfaatan pada waktu tertentu. Masalah serupa muncul dengan kebutuhan jaringan dan penyimpanan.

Dengan demikian, ada pemborosan kapasitas dan kurangnya pemanfaatan dalam kapasitas cara tradisional dialokasikan. Yang konvensional pendekatan terhadap manajemen kapasitas jelas tidak sepenuhnya berlaku di lingkungan berbasis cloud

karena ini dapat mendatangkan inefisiensi biaya dan tantangan operasional yang ditimbulkan oleh lingkungan infrastruktur virtual dinamis dan beberapa model penyebaran cloud.
Dalam lingkungan modern dan sangat virtual / cloud, perusahaan TI perlu mengingat kembali strategi manajemen kapasitasnya yang kini melampaui elemen infrastruktur dasar seperti server, penyimpanan, dan jaringan. Ada kebutuhan untuk fokus pada kinerja aplikasi dan integrasi yang mendalam dengan tumpukan manajemen. Solusi kapasitas usia baru harus dapat menawarkan pandangan menyeluruh dan menyeluruh dari sumber daya infrastruktur seperti penyimpanan, jaringan, server, mesin virtual, dll. Selain itu, solusi ini harus menyajikan pandangan selam yang mendalam ke dalam sumber daya infrastruktur individu. Metrik kapasitas dan solusi kapasitas yang dihasilkan harus mampu mensimulasikan berbagai skenario penggunaan, memperkirakan konsumsi kapasitas, melakukan alokasi kapasitas proaktif, dll. Solusi kapasitas modern juga harus memiliki kemampuan untuk mematuhi aturan kepatuhan dan peraturan setiap kali keputusan penempatan beban kerja dibuat.
Di era baru komputasi awan, fokus dapat bergeser untuk menyediakan unit kapasitas terkecil yang memungkinkan untuk mendukung aplikasi. Desain dari aplikasi itu sendiri dilakukan sedemikian rupa sehingga sumber daya minimum digunakan.

Dalam perencanaan kapasitas dan desain, kami melihat unit kapasitas terendah yang tersedia dari penyedia cloud daripada desain untuk aplikasi yang ditingkatkan. Kapasitas sekecil mungkin akan membawa ekonomi biaya dan fleksibilitas kepada konsumen. Misalnya dalam menawarkan layanan berbasis cloud seperti IaaS, seseorang dapat mengkonfigurasi katalog layanan yang mencakup unit server dengan spesifikasi sekecil RAM 2 GB, 2 CPU, 40 GB Hardisk (HDD, *Hard Disk Drive*.) Keuntungan dari memiliki unit kapasitas kecil adalah dengan mendatangkan efisiensi biaya dan memastikan bahwa sumber daya yang optimal disediakan untuk menjalankan beban kerja. Selain itu, sesuai dengan persyaratan pelanggan, seseorang dapat mengkonfigurasi katalog layanan IaaS dengan spesifikasi pilihan yang dapat menjangkau minimal hingga konfigurasi maksimum yang mungkin. Bersamaan dengan ini, model pay as you go dapat menghemat banyak uang dengan membantu perusahaan mengurangi biaya kepemilikan total (*Total Cost of Ownership*,TCO) mereka.

CONTOH

Di area ini adalah database sebagai penawaran layanan. Di sini penyedia cloud dapat menyediakan layanan basis data berdasarkan pada membaca dan menulis bahwa aplikasi pelanggan mungkin perlu. Jadi untuk memulai dengan, pelanggan dapat membeli X membaca dan Y menulis dan sebagai aplikasi berskala ke lebih banyak pengguna dan penggunaan pelanggan secara dinamis dapat meningkatkan membaca dan menulis. Pelanggan hanya membayar untuk membaca dan menulis yang dialokasikan oleh penyedia cloud. Lapisan orkestrasi dan otomatisasi penyedia cloud mengatur penyediaan infrastruktur yang dibutuhkan dan meningkatkan basis data untuk memenuhi kebutuhan pelanggan.

Bandingkan skenario di atas dengan perencanaan kapasitas dalam model tradisional. Dalam model tradisional, input untuk kapasitas database akan dikumpulkan terlebih dahulu dengan cara yang mirip dengan contoh berikut:

Pengelolaan permintaan:

- Pengelolaan permintaan akan memberikan data berikut.
- Aplikasi ini akan skala dari 10 pengguna di lingkungan pengembangan untuk 1000 pengguna selama pengujian stres.
- Aplikasi ini akan skala dari 0 pengguna puncak di luar jam ke 100 pengguna puncak pada hari kerja dan 1000 pengguna selama akhir bulan.

Tingkat Layanan :

- Ketersediaan Layanan: 99,5%.
- Kinerja: Waktu respons hingga pengguna akhir kurang dari 5 detik per transaksi.

Sekarang dalam skenario perencanaan kapasitas tradisional, persyaratan perangkat keras akan didasarkan pada beban kerja puncak yaitu 1000 pengguna karena tingkat layanan tidak menentukan atau menerima tingkat layanan yang lebih rendah untuk penggunaan puncak. Jadi kapasitas yang 10 kali dari kapasitas rata-rata harus disediakan untuk, atau ditetapkan, berdasarkan permintaan bisnis dan tingkat layanan.

Dalam lingkungan cloud, unit terkecil yang mungkin untuk kapasitas dikurangi dari tumpukan perangkat keras penuh ke server virtual fleksibel yang dapat digunakan sesuai model utilitas dan akan dikenakan biaya hanya ketika digunakan.

Selain itu, jika aplikasi menghabiskan lebih banyak sumber daya, komputasi awan menyediakan infrastruktur terukur yang dapat disediakan dalam hitungan menit dibandingkan dengan minggu dalam lingkungan tradisional. Ini akan memastikan tidak hanya bahwa kinerja layanan baru atau yang berubah memenuhi target yang diharapkan, tetapi juga bahwa semua layanan yang ada terus memenuhi semua target mereka.

Pemahaman rinci tentang kebutuhan bisnis dan driver dan lagi bagaimana ini akan berhubungan dengan layanan dan infrastruktur sangat penting dalam lingkungan cloud, dan untuk tingkat yang lebih rendah, setiap proyek virtualisasi berskala besar. Mencapai tingkat kematangan dan integrasi ini menghadirkan tantangan yang cukup besar bagi tim manajemen kapasitas, tetapi jika tercapai, akan menguntungkan kedua bisnis dan meningkatkan profil manajemen kapasitas yang tak terukur.

Jika satu terurai model cloud, menjadi jelas bahwa ada banyak variasi pada tema dan variasi tertentu memberikan lebih banyak nilai awal daripada yang lain. Sebagai contoh: Awan dapat memberikan IaaS mentah, PaaS tingkat yang lebih tinggi (termasuk database paket terprogram dan tumpukan middleware), dan bahkan SaaS lengkap (yang akrab bagi pengguna otomatisasi tenaga penjualan atau alat produktivitas kantor melalui Internet.)

Atas dasar pertimbangansseperti tingkat layanan, beban kerja dan perilaku aplikasi, prosesor, analisis I / O, dll. ada beberapa implikasi dari model layanan cloud pada proses manajemen kapasitas. Sekarang, mari kita coba memahami apa saja berbagai lapisan manajemen kapasitas yang harus dipertimbangkan untuk perencanaan kapasitas keseluruhan.

Manajemen Kapasitas Bisnis di Cloud

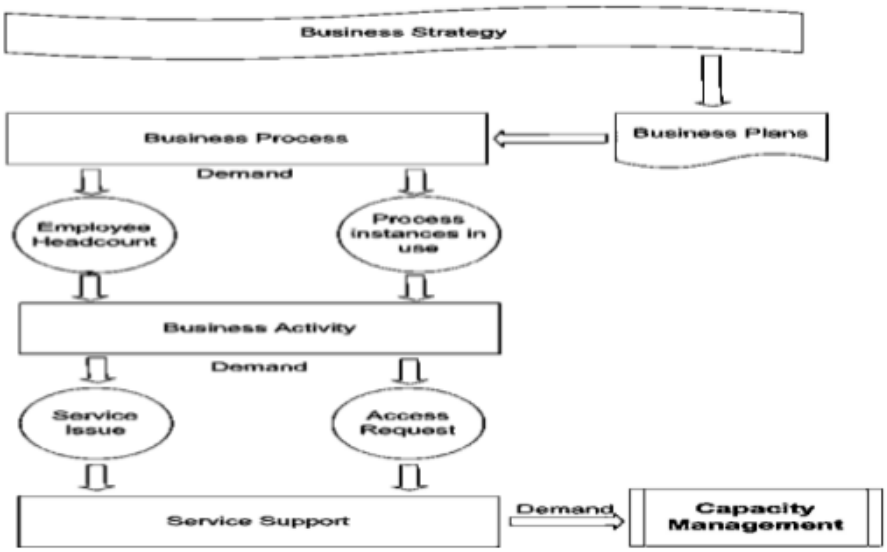
Target manajemen kapasitas bisnis adalah untuk memastikan kelangsungan bisnis yang menguntungkan.

Aplikasi yang sangat penting untuk bisnis dipetakan dengan layanan yang mendukung bisnis, tujuan bisnis dan perencanaan kapasitas untuk memastikan bisnis disediakan dengan kapasitas dan tingkat layanan yang diperlukan.

Manajemen kapasitas pada tingkat ini harus memastikan metrik kinerja aplikasi di tempat dan ini disetel untuk kebutuhan proses bisnis. Tingkat pemanfaatan kapasitas yang terkait dengan kinerja tingkat aplikasi harus ditentukan, dipantau dan divalidasi secara tepat waktu.

Fokus dari aspek inimanajemen kapasitas adalah untuk menerjemahkan kebutuhan bisnis dan rencana ke dalam persyaratan untuk layanan dan infrastruktur TI, memastikan bahwa persyaratan bisnis masa depan untuk layanan TI dikuantifikasi, dirancang, direncanakan, dan diimplementasikan secara tepat waktu. Manajemen kapasitas bisnis adalah pandangan yang relatif jangka panjang dari manajemen kapasitas karena ini mungkin melibatkan analisis beban kerja bisnis seperti penggunaan IT yang luas dan sesuai mendefinisikan tingkat layanan untuk mengakomodasi permintaan.

Gambar 5-6 menampilkan bagaimana aktivitas permintaan diformulasikan menjadi persyaratan kapasitas. Strategi bisnis, bersama dengan manajemen keuangan, membentuk rencana dan proses bisnis. Atas dasar kegiatan dan proses bisnis, input permintaan dimasukkan ke dalam proses manajemen kapasitas.



***Gambar 3-6** . Manajemen permintaan berdasarkan aktivitas*

Ketika penyedia cloud *(pencipta layanan dan agregator layanan)* mulai menawarkan model SaaS, mereka menambahkan pemantauan pengalaman pengguna akhir untuk melacak waktu respon transaksi. Maksud dari ini adalah untuk menangkap persyaratan bisnis dalam hal penggunaan layanan dan relevansi bisnis. Dengan kata lain, manajemen kapasitas bisnis mencari masukan bisnis untuk mengantisipasi kebutuhan bisnis dan mencari informasi permintaan bisnis secara proaktif dari jaminan layanan bisnis yang ada dan ukuran utilitas.

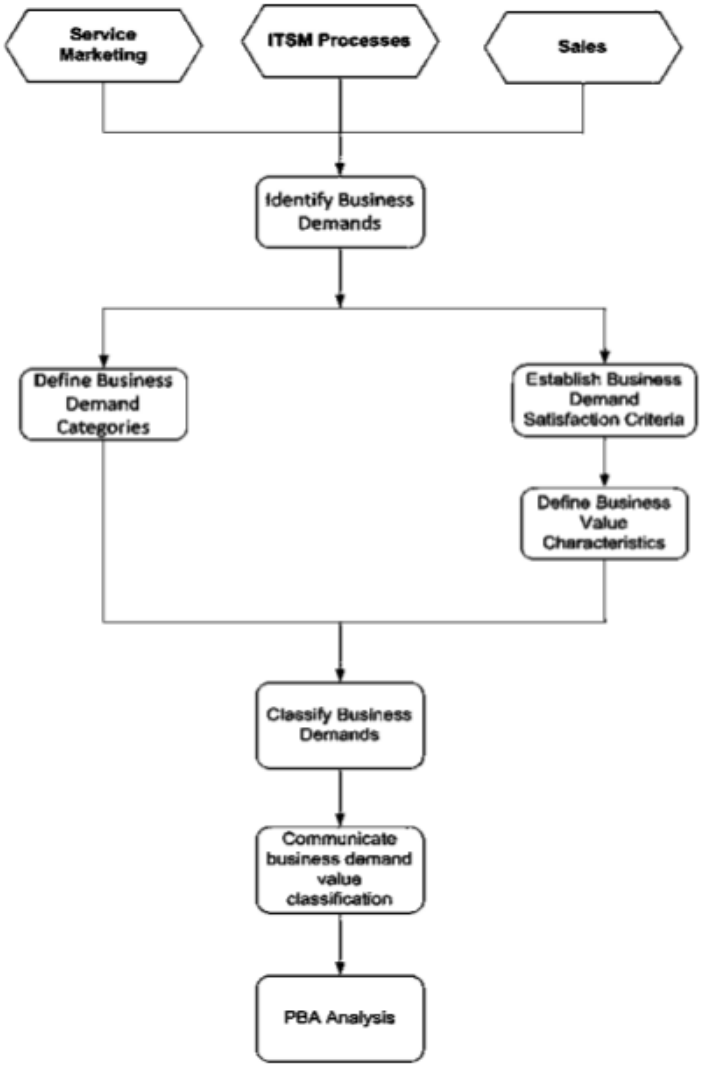
Jaminan layanan memastikan bahwa suatu layanan layak digunakan sementara utilitas memastikan kesesuaian dengan tujuan layanan.
Penyedia cloud dapat mencari persyaratan kapasitas bisnis dari sejumlah sumber termasuk survei, tren penggunaan, skor CSAT dan laporan optimalisasi kinerja untuk meningkatkan kinerja, fitur, dan cakupan layanan bisnis.
Data ini juga dapat ditangkap dengan menganalisis jenis permintaan layanan yang akan menentukan siapa yang memesan apa dan berapa frekuensi pesanan.
Masukan lain dapat berupa permintaan untuk peningkatan layanan dari pelanggan.
Analisis kompetitif dari produk yang kompetitif juga merupakan sumber input penting untuk permintaan bisnis.
Persyaratan masa depandari permintaan bisnis berasal dari menganalisis manajemen permintaan dan portofolio layanan. Analisis ini harus menawarkan detail tentang proses dan persyaratan layanan baru, perubahan, perbaikan, dan juga pertumbuhan dalam layanan yang ada.
Portofolio layanan terutama terdiri dari tiga jenis layanan: pipa layanan, layanan yang ada (dari katalog layanan) dan layanan pensiunan. Ketiga jenis kategori layanan ini memiliki implikasi kapasitas dalam hal penyediaan infrastruktur kapasitas yang dioptimalkan untuk layanan yang sedang dalam penyaluran, menyediakan kapasitas untuk mendukung layanan yang ada dalam katalog layanan, dan membebaskan kapasitas yang dialokasikan dari layanan pensiunan.
Manajemen kapasitas bisnis memastikan bahwa permintaan untuk pasokan kapasitas seimbang. Jika pasokan dan permintaan tidak seimbang, ini akan berdampak langsung pada biaya penyampaian layanan. Untuk melakukan manajemen kapasitas yang efektif, organisasi perlu memantau infrastruktur aplikasi, pengalaman pengguna akhir, dan pemanfaatan infrastruktur dari waktu ke waktu untuk mengukur bahwa kapasitas yang cukup ada untuk memenuhi persyaratan tingkat layanan yang disepakati. Misalnya jika tujuan bisnis adalah memiliki kolaborasi dan komunikasi yang efisien dalam organisasi, layanan yang mendasarinya dapat berupa email, obrolan, portal dan wiki dll. Masing-masing layanan ini dapat bergantung pada berbagai komponen seperti jaringan dan penyimpanan. Jadi perencanaan kapasitas perlu dilakukan di semua tingkatan untuk memastikan keseluruhan bisnis berjalan dengan biaya yang dapat dibenarkan.
Kegiatan manajemen kapasitas bisnis utama akan menjadi tren, peramalan, pembuatan prototipe, ukuran dan mencari masukan yang sedang berlangsung untuk memprediksi kebutuhan bisnis di masa depan.

Mari kita diskusikan bagaimana manajemen kapasitas bisnis memainkan peran di ketiga pemain cloud yang telah kita diskusikan sebelumnya:

Penyedia Layanan Cloud

Penyedia layanan cloud dalam bisnis menyediakan kapasitas cloud kepada konsumennya. Cloud adalah layanan yang ditawarkan oleh penyedia layanan cloud. Manajemen keuangan aspek layanan penawaran dipertimbangkan untuk keseluruhan perencanaan bisnis, kapasitas yang akan ditawarkan di berbagai lokasi dan untuk berbagai item portofolio layanan.Gambar 3-7 menjelaskan prosedur umum untuk menangkap tren bisnis. Penyedia Cloud dapat menggunakan tren ini untuk melakukan pola analisis aktivitas bisnis.

Pola aktivitas bisnis (PBA): Pola aktivitas bisnis digunakan untuk membantu penyedia layanan TI memahami dan merencanakan berbagai tingkat aktivitas bisnis. Sebagai bagian dari proses manajemen permintaan, konsep pola aktivitas bisnis adalah sumber utama informasi mengenai permintaan layanan yang diantisipasi. Kegiatan ini dapat berupa peningkatan layanan, ekspansi atau modifikasi lainnya. Misalnya, pelanggan dapat memperkenalkan layanan di geografi baru, meningkatkan fitur layanan dll. Gambar 5-8 menguraikan pendekatan sederhana untuk melakukan analisis PBA.



Gambar 3-7 . *Nilai dan Klasifikasikan kebutuhan Bisnis*

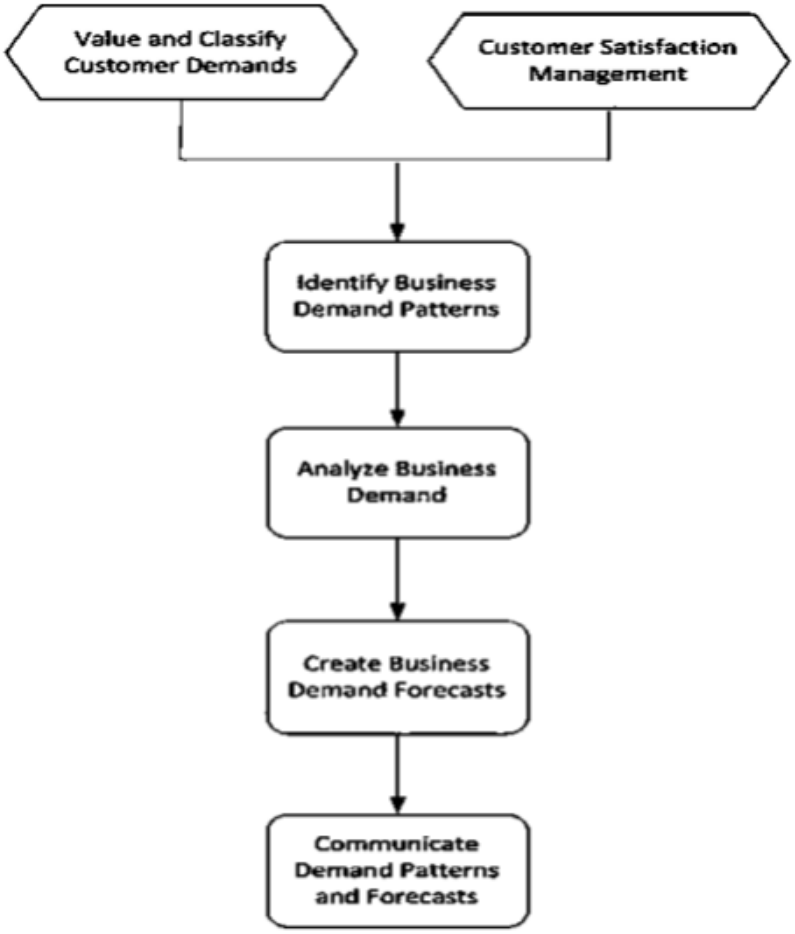
Pola aktivitas bisnis untuk konsumen cloud seperti yang dijelaskan dalam contoh di atas akan berarti variabilitas permintaan untuk menghitung sumber daya berdasarkan musim, hari dalam bulan, bulan tertentu, dll.

Setiap bisnis melewati fluktuasi musiman. Sebagai contoh, pengecer melihat peningkatan aktivitas bisnis selama musim liburan di mana permintaan meningkat berlipat karena pelanggan buru-buru berbelanja selama periode ini.

Peluncuran produk baru oleh perusahaan konsumen juga dapat menghasilkan minat yang besar dalam situs web dan produksebuah perusahaan. Ada kejadian di mana peluncuran produk tersebut telah mengakibatkan situs web perusahaan kewalahan dengan lalu lintas dan macet.

Pola analisis aktivitas bisnis di lingkungan cloud harus memberikan informasi dasar berikut:

- Item katalog layanan yang diperintahkan oleh pelanggan
- Frekuensi dan pola permintaan dari penggunaan kapasitas
- Lokasi barang katalog
- Pemanfaatan di tingkat lokasi
- Pemanfaatan di tingkat item katalog



Gambar 3-8 . *Analisis PBA*

Frekuensi dan permintaan item katalog di berbagai lokasi (pusat data) memberikan masukan dasar untuk analisis pola aktivitas bisnis. Data ini dimasukkan ke dalam peramalan dan alat analisis untuk menghasilkan kapasitas yang dibutuhkan untuk berbagai layanan dan komponen. Pola aktivitas bisnisuntuk penyedia cloud adalah tantangan yang rumit. Penyedia awan menggunakan alat analisis khusus yang khusus untuk mengukur pola aktivitas bisnis. Untuk penyedia cloud, mungkin ada variabilitas permintaan yang sangat besar karena sifat bisnis dan penyedia cloud harus memenuhi variasi permintaan ini.

Hal ini dicapai dengan campuran berbagai opsi kepada konsumen termasuk harga diskon untuk komitmen jangka panjang, memberikan contoh tempat kepada pelanggan dengan harga lebih murah untuk menggunakan kapasitas menganggur.

Untuk konsumen cloud yang menangani pola aktivitas bisnis menjadi lebih mudah karena sekarang kapasitas dapat disediakan sesuai permintaan saat dibutuhkan dan dirilis selama periode lean.

Kemudian meramalkan proyeksi, tren, dan pemodelan digunakan untuk memproyeksikan PBA untuk memberikan rincian tentang pola pesanan di masa mendatang dan penggunaan infrastruktur cloud. Lihat bab selanjutnya untuk membaca lebih lanjut tentang tren dan perkiraan.

Portofolio Layanan: Portofolio layanan manajemen adalah proses yang mengatur siklus hidup layanan. Layanan ini dapat mencakup layanan yang akan diluncurkan di pasar, layanan langsung yang tercantum dalam katalog layanan dan layanan yang tidak lagi tersedia. Di sini layanan pensiunan, layanan pipa dan layanan langsung dalam portofolio layanan dianalisis dan disediakan sebagai masukan kapasitas untuk komputasi awan. Masukan ini disediakan oleh tim manajemen portofolio layanan ke penyedia layanan cloud atau manajer dan arsitek kapasitas cloud. Penyedia cloud akan memutuskan layanan apa yang ditawarkan dan layanan baru apa yang akan ditawarkan kepada pelanggan. Konsumen cloud, di sisi lain, dapat memilih model cloud untuk memberikan layanan tertentu yang sedang dalam jalur pipa atau saat ini ditawarkan.

Tingkat Layanan untuk Ketersediaan dan Kinerja: Tergantung pada tingkat layanan yang diinginkan, persyaratan kapasitas dan kinerja layanan komputasi awan dapat bervariasi. Sebagai contoh, komitmen ketersediaan yang lebih tinggi akan memerlukan replikasi data dan penggunaan beberapa mesin atau ketersediaan mesin suku cadang yang lebih tinggi. Mesin-mesin cadangan ini dapat disediakan dalam kasus-kasus downtime infrastruktur. Tingkat kinerja layanan yang lebih tinggi akan membutuhkan sumber daya akhir yang lebih tinggi dalam hal komputasi, penyimpanan, dan jaringan.

Data kinerja dan tingkat kinerja saat ini merupakan indikator penting untuk apa yang perlu diubah dalam penawaran layanan baru atau bagaimana layanan yang ada perlu diubah. Pemantauan layanan cloud secara teratur memberikan masukan tentang pola aktivitas bisnis, tingkat layanan, ketersediaan, dan kapasitas, serta bagaimana hal ini memengaruhi manajemen kapasitas untuk layanan cloud. Akhirnya atas dasar semua masukan agregat dari proses yang disebutkan di atas seperti manajemen permintaan,

manajemen portofolio layanan, manajemen keuangan dll, penyedia cloud dapat mulai merancang kapasitas layanan cloud atau CSP (paket layanan inti) layanan utama seperti IaaS dan SLP (paket tingkat layanan)yang mengandung nilai tambah tambahan seperti manajemen, dukungan dan pemeliharaan. Dalam bab-bab selanjutnya kita akan melihat lebih dekat prosedur untuk menerapkan hal yang sama. Konsumen cloud dapat memilih opsi penerapan berdasarkan ketersediaan dan persyaratan kinerja. Sebagai contoh, penyedia cloud menyediakan opsi penerapan aplikasi di beberapa zona ketersediaan kesalahan toleran dan juga opsi penerapan aplikasi di berbagai wilayah geografis. Opsi-opsi ini menyediakan ketersediaan lebih tinggi untuk aplikasi.

Pelanggan Layanan Cloud

Konsumen layanan di sisi lain terutama akan difokuskan pada pengurangan TCO dengan memanfaatkan model Opex dan manfaat potensial dari membayar per penggunaan. Fungsi kapasitas bisnis untuk konsumen layanan akan lebih condong terhadap pengurangan biaya dengan mengevaluasi biaya yang terkait dengan berbagai penyedia layanan cloud.Selain itu, konsumen layanan masih perlu melakukan fungsi-fungsi tertentuuntuk perencanaan kapasitas seperti peramalan bisnis, perencanaan keuangan, membuat pola analisis aktivitas bisnis, memperkirakan permintaan, negosiasi tingkat layanan dan aplikasi dan rekayasa ulang proses. Semua fungsi ini membantu konsumen dalam menyiapkan pendekatan fit terbaik di mana komputasi awan dikhawatirkan. Dengan komputasi awan, konsumen layanan dapat benar-benar melihat opsi yang dapat disediakan cloud sesuai permintaan untuk kapasitas yang dapat dibenarkan biaya.Perencanaan kapasitas konsumen cloud terkena dampak karena model cloud dan bukannya memiliki kapasitas investasi dimuka, kapasitasnya bersumber berdasarkan model bayar per penggunaan di cloud.Berbagai pilihan yang tersedia di cloud dan harga yang selalu berubah dari penyedia cloud membuat perhitungan TCO dalam model cloud merupakan latihan yang ekstensif.Pelanggan layanan cloud, saat mengembangkan aplikasi baru, fokus pada pembuatan kasus dasar pemanfaatan aplikasi yaitu jumlah pengguna atau jumlah transaksi. Mereka sekarang dapat merencanakan kapasitas secara bertahap karena cloud dapat bekerja dalam model skala di mana aplikasi dapat bekerja pada beberapa mesin secara bersamaan dan menambah atau mengurangi jumlah node yang berfungsi.Di Perangkat Lunak

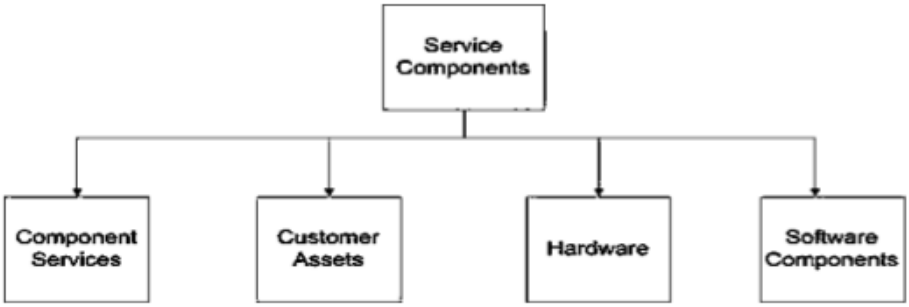
sebagai model Layanan konsumen sekarang dapat sumber aplikasi di bayar per transaksi atau bayar per pengguna per bulan jenis model.Ini memungkinkan konsumen untuk dengan mudah memetakan permintaan bisnis terhadap kapasitas dan biaya. Karena model pembayaran per pengguna memberi Anda kapasitas sesuai permintaan, hal yang sama dapat dikaitkan dengan jumlah pengguna yang diantisipasi untuk memanfaatkan layanan dalam periode tertentu berdasarkan permintaan bisnis.Tidak ada penalti bagi konsumen cloud jika ramalan permintaan menjadi salah karena tidak ada kapasitas yang tidak terpakai yang dibeli untuk mengantisipasi permintaan. Sekarang pengadaan lebih sederhana dan lebih cepat dan dilakukan ketika permintaan benar-benar muncul daripada dilakukan berdasarkan perkiraan.Belanja modal rendah karena kapasitas tidak dibeli di muka. Namun, Opex mungkin lebih tinggi di mana aspek perencanaan keuangan seperti biaya modal, laba atas investasi, total biaya kepemilikan, dll. Harus diterapkan pada model Opex untuk membandingkan berbagai opsi yang tersedia dari penyedia / agregator cloud.

Manajemen Kapasitas Layanan Cloud

Suatu perusahaan dapat mengkonsumsi beberapa jenisawan atau layanan cloud dan juga mempertahankan layanan di dalam perusahaan hingga tingkat tertentu sehingga mereka masih harus memperkirakan kapasitas layanan. Namun, kapasitas komponen untuk komponen permintaan dilakukan oleh penyedia cloud.

Fokus dari manajemen kapasitas layanan adalah untuk mengidentifikasi dan memahami layanan cloud, penggunaan sumber daya mereka, pola kerja, puncak dan palung, untuk memastikan bahwa layanan memenuhi target SLA mereka, sehingga untuk memastikan bahwa layanan memenuhi kebutuhan yang diperlukan.

Dalam model cloud, layanan yang disediakan agregator mungkin berasal dari beberapa vendor cloud dan dengan demikian agregator cloud mengumpulkan berbagai layanan dan menyediakan layanan gabungan ke konsumen akhir. Gambar 3-9 secara luas menjelaskan berbagai komponen layanan. Dalam kepemilikan skenario berbasis cloud dan kontrol komponen-komponen ini dapat bervariasi seperti yang dijelaskan sebelumnya.



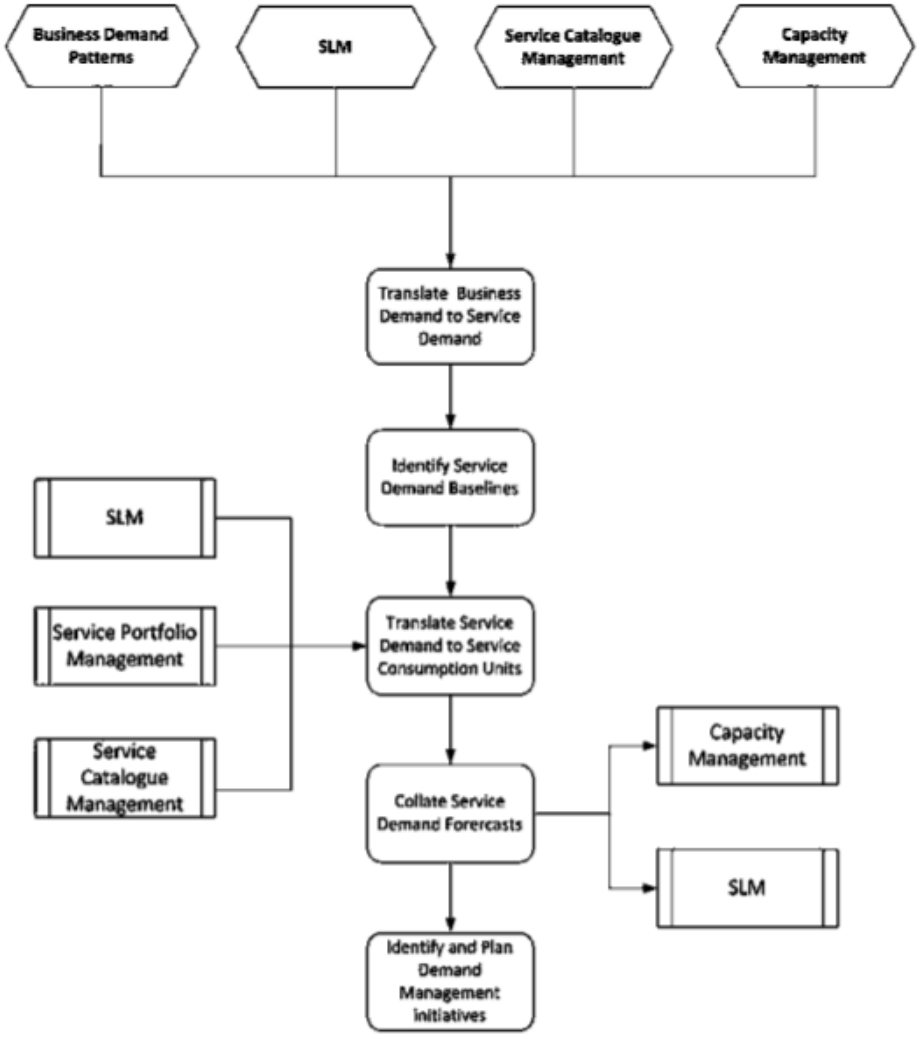
Gambar 3-9 . *Komponen Layanan*

Pelanggan departemen TI malah dapat bertindak sebagai agregator dengan menggabungkan berbagai layanan yang ditawarkan oleh penyedia cloud. Konsumen cloud memanfaatkan layanan TI yang disediakan oleh penyedia cloud dan menciptakan layanan tingkat yang lebih tinggi.

Sebagai contoh, di infrastruktursebagai model layanan, konsumen cloud tidak peduli dengan perangkat keras dan virtualisasi yang mendasarinya dan hanya menggunakannya sebagai layanan yang tersedia di cloud. Sebagai contoh, mungkin menciptakan layanan tingkat yang lebih tinggi seperti database sebagai layanan dengan menyediakan database yang dikelola di atas IaaS ke departemen atau pengguna.

Jadi tingkat di mana kontrol dan manajemen dilakukan dalam model cloud berbeda dari tingkat di mana kapasitas layanan dapat dilakukan di lingkungan TI tradisional.

Gambar 3-10 menjelaskan prosedur untuk meramalkan permintaan layanan. Meskipun ini murni permintaan fungsi manajemen tetapi input dari kapasitas layanan digunakan untuk menuntut perhitungan dan ramalan permintaan dimasukkan ke dalam proses manajemen kapasitas.



Gambar 3-10 . *Permintaan Layanan Prakiraan*

Penyedia Layanan Cloud

Layanan yang disediakan oleh penyedia cloud dapat berupa salah satu dari rasa (IaaS, PaaS, SaaS) dan penyedia cloud harus memastikan bahwa kapasitas layanan tersedia untuk digunakan oleh konsumen cloud. Penyedia cloud mengukur parameter kunci berikut untuk mengawasi kapasitas layanan cloud yang ditawarkan dan

mengambil tindakan yang tepat untuk memastikan perjanjian tingkat layanan (atau SLA) dipertahankan dan layanan tersedia untuk konsumen layanan:

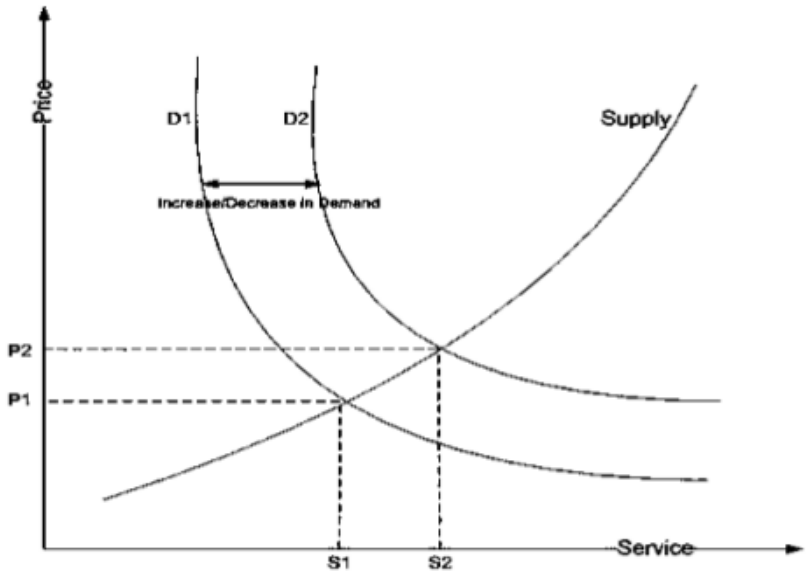
- Pengumpulan data dan ambang batas
- Analisis penggunaan saat ini dan masa depan layanan
- Roll up komponen individu yang merupakan layanan cloud
- Pemantauan SLA
- Beban kerja di berbagai lokasi / pusat data
- Tindakan proaktif dan reaktif untuk meningkatkan kapasitas.

Agregator awan akan melakukan serangkaian aktivitas serupa untuk kapasitas layanan karena layanan mereka adalah agregasi layanan yang disediakan oleh penyedia cloud.

Cloud Consumer

Konsumen cloudmenggunakan layanan dari berbagai penyedia cloud dan agregator dan menggabungkan layanan ini untuk menyediakan aplikasi dan utilitas bagi pengguna. Konsumen cloud tidak akan terganggu dengan bagaimana kapasitas layanan dari penyedia cloud ini akan dikelola. Sebaliknya, mereka terutama akan memantau SLA dari penyedia dan agregator cloud. Mereka juga akan memastikan bahwa layanan apa pun yang mereka sediakan di atas awan dimonitor untuk kapasitas layanan dan tindakan yang tepat diambil untuk mengelola kapasitas dengan biaya yang efektif.

Gambar 311 menguraikan peningkatan biaya karena penggunaan layanan dan peningkatan permintaan. P1 dan P2 adalah harga layanan. S1, S2 dan D1, D2 menjadi penawaran dan permintaan layanan.



Gambar 3-11 . *Permintaan Layanan dan Kurva Penawaran*

Layanan TI dapat dipantau melalui berbagai pendekatan pemantauan dan alat-alat yang menggunakan transaksi simulasi atau mengendus data pada jaringan untuk memantau pandangan pengguna akhir dari layanan. Selain ini ada beberapa toolsets tersedia yang menangkap pengalaman layanan pengguna akhir dan menciptakan laporan kepuasan pelanggan.

Di zaman sekarang yang semakin kompleks dan sifat lingkungan virtualisasi yang sangat terdistribusi, komputasi awan, desktop virtual, dan aplikasi web 2.0, ada kebutuhan yang kuat untuk menangkap, meninjau, dan meningkatkan pengalaman pengguna akhir. Ada berbagai teknologi manajemen kinerja aplikasi / layanan yang memantau tiga komponen pengalaman pengguna utama seperti kinerja aplikasi / layanan, kinerja infrastruktur, dan produktivitas pengguna. Toolset ini menggunakan agregasi, analisis, dan korelasi waktu nyata dari semua metrik ini dan menampilkan hasilnya di dasbor.

Sebagai contoh, konektivitas jaringan ke layanan cloud mungkin menjadi tanggung jawab konsumen cloud. Dan ini harus dimonitor untuk penggunaan dan bandwidth bersama dengan penerapan teknik yang disebutkan di atas sehingga konsumen cloud dapat mengidentifikasi masalah dan meningkatkan bandwidth jaringan untuk mengakses aplikasi / sumber daya yang tersedia di cloud.

Manajemen Kapasitas Komponen Cloud

Fokus manajemen kapasitas komponen adalah untuk mengidentifikasi dan memahami persyaratan sumber daya, tingkat kinerja, dan tren pemanfaatan masing-masing komponen individudalam lingkungan cloud. Komponen-komponen ini seperti server, komponen keamanan, komponen jaringan, penyimpanan, perangkat lunak ketika digabungkan dari layanan seperti email. Data kinerjadari komponen dicatat, dianalisis, dan dilaporkan untuk perencanaan sumber daya dan manajemen kapasitas komponen. Mekanisme threshold otomatis dan mekanisme peringatan membantu dalam mengelola semua komponen, untuk memastikan bahwa situasi di mana target layanan dilanggar atau terancam oleh penggunaan atau kinerja komponen diidentifikasi dengan cepat, dan tindakan yang hemat biaya untuk mengurangi atau menghindari dampak potensial mereka diimplementasikan.

Di lapisan komponen, solusi pemantauan harus secara otomatis menentukan waktu respons pengguna akhir normal dan mengirim peringatan untuk memperingatkan potensi kegagalan tingkat layanan, meningkatkan kemampuan penyedia cloud untuk menemukan akar penyebab insiden dengan menunjukkan dengan tepat bagian dari aplikasi yang menyebabkan kesalahan. Sebagai contoh, penyedia cloud dapat menggunakan alat pemantauan untuk manajemen kinerja dan analisis dampak untuk menentukan bagian-bagian dari aplikasi yang akan terpengaruh bahkan jika ada perubahan dalam skema database.

Mengumpulkan dan menganalisis data yang relevan, manajemen kapasitas komponen membantu Anda merencanakan kebutuhan yang akan datang dan mengoptimalkan pemanfaatan sumber daya, sehingga terus menyeimbangkan biaya terhadap kapasitas dan pasokan terhadap permintaan.

Penyedia Layanan Cloud

Penyedia layanan cloud memonitor komponen yang mendasari seperti CPU, Memori, Penyimpanan, OS, dll dari layanan cloud sementara agregator layanan cloud atau pelanggan cloud akan memantau tingkat layanan dan persyaratan terkait kinerja dari layanan cloud. Penyedia cloud juga membutuhkan data ini untuk tujuan penagihan.

Penyedia layanan cloudharus memonitor setiap komponen dan layanan melalui alat-alat yang sangat otomatis dan menggunakan otomatisasi untuk auto-tune atau auto-

correct. Tidak mungkin bagi penyedia cloud untuk mengelola tanpa alat pemantauan dan resolusi otomatis.
Penyedia awan memantau hal-hal berikut:

- Ketersediaan setiap layanan
- Kinerja setiap layanan
- Ketersediaan komponen
- Kinerja komponen
- Ketersediaan integrasi
- Kinerja integrasi

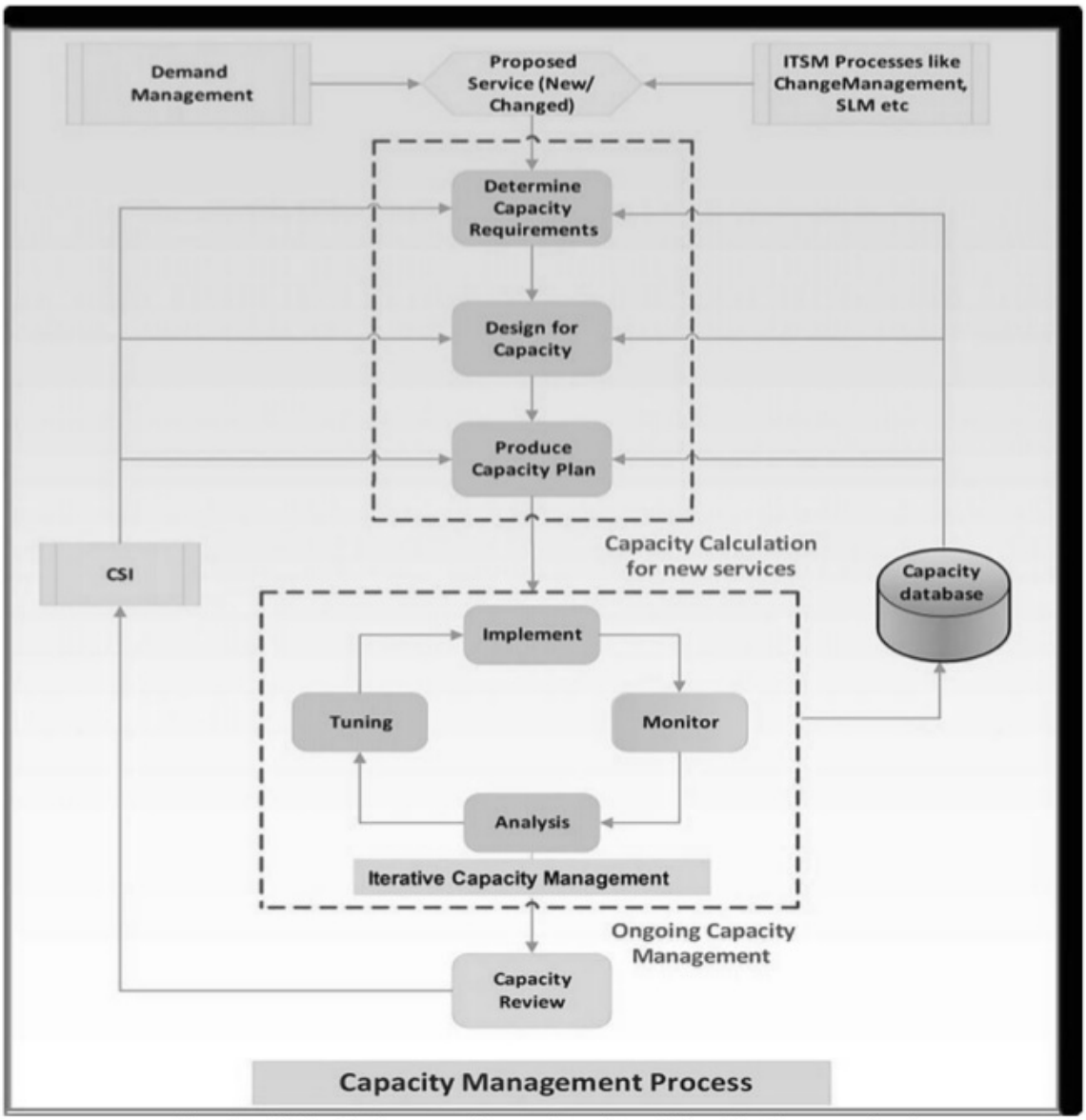
Pemantauan dilakukan dari dalam pusat data dan juga dari lokasi luar untuk memantau dampak kinerja atau tidak tersedianya karena masalah jaringan dari lokasi lain.

Cloud Consumer

Konsumen cloud juga akan memantau kapasitas sumber daya / komponen untuk membuat keputusan tentang perluasan kapasitas dan juga untuk membuat keputusan untuk meningkatkan infrastruktur.
Contoh ini akan menjadi pemantauan server untuk CPU dan RAM. Jika server mencapai kapasitas, konsumen cloud dapat mengonfigurasi otomatisasi untuk memutar mesin baru yang mulai mengambil beban kerja.
Beberapa penyedia cloud menyediakan data pemantauan komponen dasar kepada konsumen cloud untuk memungkinkan mereka menggunakan penggunaan sumber daya dan laporan kinerja untuk memutuskan tindakan perluasan sumber daya dan penyetelan kinerja aplikasi.
Sekarang mari kita menyelam ke dalam sub proses atau prosedur manajemen kapasitas yang spesifik untuk lingkungan cloud. Prosedur-prosedur ini dapat dikategorikan secara luas menjadi:

- **Perhitungan kapasitas untuk layanan baru**
- **Pengelolaan kapasitas berkelanjutan**

Gambar 3-12 menampilkan proses manajemen kapasitas cloud. Dalam sisa bab kita akan menyelam jauh ke dalam setiap area prosedur.



Gambar 3-12 . *Fokus manajemen kapasitas dalam lingkungan cloud*

BAB EMPAT

PERENCANAAN KAPASITAS

Perencanaan Kapasitas

Dalam setiap model penyebaran cloud, baik cloud pribadi atau publik, prinsip-prinsip dasar manajemen kapasitas akan tetap sama dan akan terus melayani tujuan utama mencapai kapasitas yang dapat dibenarkan biaya untuk memenuhi kebutuhan bisnis saat ini dan di masa depan.

Faktor keberhasilan yang penting untuk menentukan kapasitas layanan tetap menjadi data input(tingkat layanan, penggunaan saat ini, dan permintaan di masa mendatang) dari mana rencana kapasitas akan dibuat. Ingat, rencana kapasitas adalah hasil dari proses perencanaan kapasitas.

Untuk perencanaan kapasitas yang efisien, penting untuk mengkategorikan layanan ke dalam unit yang terukur. Beban kerja dapat dianalisis untuk menentukan persyaratan tingkat layanan. Tingkat kapasitas saat ini harus diukur sehingga perkiraan dapat dibuat berdasarkan kesenjangan antara kapasitas saat ini dan yang diharapkan.

Masukan dari manajemen untuk perencanaan kapasitas dapat mencakup hal-hal berikut:

- Pertumbuhan yang diharapkan dalam bisnis
- Persyaratan untuk menerapkan aplikasi baru
- Akuisisi terencana atau divestasi
- Rencana keuangan
- Permintaan untuk konsolidasi sumber daya TI

Perencanaan kapasitas adalah area utama di mana konsep seperti pemanfaatan sumber daya dan biaya kapasitas yang dapat dipertanggungjawabkan didasarkan. Dalam buku ini, akan bahasi perencanaan kapasitas dari perspektif penyedia layanan. Dalam setiap layanan bisnis berbasis TI, masukan pertama untuk kapasitas

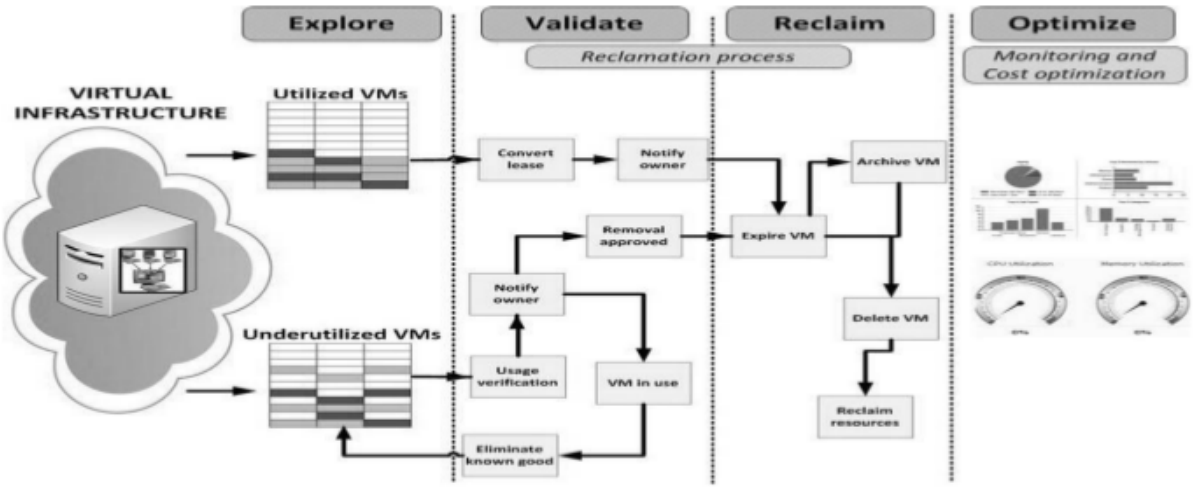
perencanaan akan datang dari permintaan untuk layanan itu dan data yang terkait mungkin berasal dari proses manajemen permintaan.
Proses manajemen permintaan dapat menggunakan alat riset pasar dan teknik lain untuk mengantisipasi permintaan untuk layanan TI tertentu. Data permintaan ini dimasukkan ke dalam prosedur manajemen kapasitasuntuk datang dengan rencana kapasitas untuk memenuhi permintaan di masa mendatang. Masukan lain untuk kapasitas perencanaan dapat berasal dari profil penggunaan layanan. Penggunaan layanan TI dapat diambil dari komponen atau lapisan sumber daya, yang dalam lingkungan cloud mungkin termasuk penggunaan CPU virtual, penyimpanan, jaringan, dan komponen terkait lainnya.

Manajemen Kapasitas di Awan

Manajemen kapasitas dianggap sebagai proses desain layananketika layanan baru diluncurkan atau layanan yang ada dimodifikasi. Setelah perumusan strategi layanan, yang mempertimbangkan permintaan layanan, portofolio layanan saat ini, dan perencanaan keuangan terkait TI, layanan ini dirancang bersama dengan fitur layanan yang diperlukan, tingkat layanan, proses manajemen perubahan, kebutuhan bisnis yang selalu berubah. , dll. Perencanaan kapasitas disesuaikan dengan kebutuhan bisnis ini. Dengan demikian, dengan benar merancang kapasitas layanan menjadi faktor penentu keberhasilan untuk peluncuran layanan TI dan penyelarasan bisnisnya.
Dalam lingkungan cloud, proses manajemen kapasitas adalah pembeda utama untuk memastikan keberhasilan dalam menjalankan proses bisnis. Ada situasi di mana kemudahan server atau penyebaran infrastruktur disalahartikan sebagai penyebaran infrastruktur yang murah. Kemudahan penyebaran tanpa tata kelola dan mekanisme kontrol di tempat dapat menghasilkan *Virtual Sprawl* . adalah hasil dari alokasi kapasitas berlebihan dan merupakan salah satu area nyeri di mana perusahaan gagal memanfaatkan manfaat sesungguhnya dari model cloud. Perencanaan kapasitas bertujuan untuk memberikan mitigasi terhadap situasi semacam itu dengan memastikan tingkat kapasitas optimal sudah ada.
Dalam model cloud, pemesanan kapasitas komputasi dibuat lebih sederhana dan lebih cepat. Namun, ini dapat mengakibatkan mesin yang tidak digunakan yang dibeli dan tidak aktif. Adalah penting bahwa kontrol yang tepat ada untuk menemukan dan menghilangkan sprawl VM sehingga orang atau departemen tidak

membeli kapasitas komputasi ketika tidak diperlukan dan kapasitas komputasi dilepaskan kembali ketika tidak diperlukan.

Gambar 6-1 menampilkan proses untuk mengelola VM sprawl melalui kerangka reklamasi sumber daya. Proses-proses ini dimulai dengan identifikasi efisiensi beban kerja, validasi yang sama, reklamasi sumber daya yang kurang termanfaatkan, dan pemantauan lingkungan untuk sumber daya dan optimalisasi biaya lebih lanjut. Tanpa proses pelaporan dan kontrol untuk mengendalikan hal di atas, adalah mungkin bagi pelanggan di lingkungan cloud untuk kehilangan uang karena kapasitas komputasi yang tidak digunakan tetapi diperoleh.



Gambar 4-1 . *Proses reklamasi sumber daya*

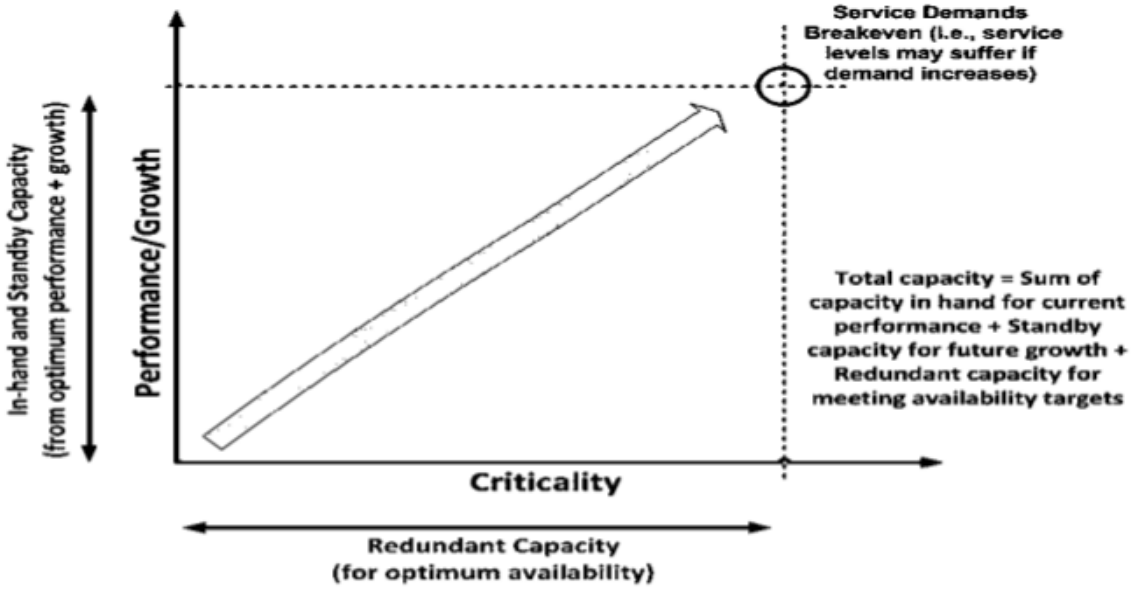
Dalam lingkungan cloud, proses manajemen kapasitas juga akan difokuskan pada pemahaman detail dari sisi level komponenbisnis (semua infrastruktur virtual yang mendasarinya seperti server, bandwidth jaringan, penyimpanan, dan load balancer). Untuk mencapai manfaat yang diinginkan dari implementasi cloud, prosesnya harus mencakup semua aspek layanan dan infrastruktur yang mendasarinya, pemantauan kinerja aplikasi, penempatan beban kerja, dan sebagainya.

Perusahaan yang memanfaatkan solusi cloud atau berencana untuk beralih ke model cloud harus memikirkan kembali strategi alokasi kapasitas mereka dengan berfokus pada kesadaran kinerja aplikasi dan mengintegrasikan dengan tumpukan manajemen. Solusi cloud harus menawarkan hal-hal berikut:

- Tampilan sumber daya infrastruktur end-to-end
- Pandangan ke depan pada konsumsi kapasitas menggunakan teknik pemodelan dan simulasi
- Alokasi kapasitas proaktif
- Kemampuan untuk mendefinisikan aturan teknis, bisnis, dan kepatuhan untuk penempatan beban kerja

Selain beradaptasi dengan strategi manajemen kapasitas baru, perusahaan Anda juga harus memilih dan mengimplementasikan toolset manajemen kapasitas yang dibutuhkan. Toolset ini tidak hanya memantau infrastruktur TI, tetapi juga perilaku aplikasi.

Suatu perusahaankemudian mencari aplikasi dan data yang memfasilitasi proses bisnis dan menjalankan proses tersebut dan menyimpan data penting. Kapasitas perangkat keras memainkan peran penting. Bisnis, berdasarkan sifatnya, mencari pertumbuhan, dan ini harus didukung oleh kinerja infrastruktur untuk memfasilitasi aplikasi bisnis yang terus berkembang. Kapasitas harus mampu menyesuaikan diri dengan cerdas sesuai dengan kekritisan yang mungkin timbul karena dinamika bisnis, musim, dan puncak yang tiba-tiba - namun tetap harus efektif biaya. Gambar 6-2menunjukkan skema penanganan kapasitas-permintaan dasar dari perspektif penyedia layanan. Komponen utama dalam perencanaan kapasitas layanan adalah kekritisan layanan atau aplikasi bisnis, persyaratan kinerja terkait, dan potensi bisnis yang tak terduga.



Gambar 4-2 . *Permintaan rapat manajemen kapasitas*

Atas dasar parameter ini, perencanaan kapasitas perlu disetel. Beragam definisi kapasitasyang dapat dipertimbangkan mungkin *kapasitas yang digunakan* , yang digunakan untuk mendukung tingkat layanan yang ada, dan *kapasitas siaga,* yang diperlukan untuk memenuhi kebutuhan terkait kapasitas dari puncak dan palung langsung, dan juga untuk setiap kegagalan yang mungkin terjadi di kapasitas digunakan. *Kapasitas redundan*dapat dianggap sebagai persediaan kapasitas keseluruhan, yang dipertahankan sesuai dengan perencanaan kelangsungan bisnis, permintaan layanan, manajemen portofolio layanan, dan input jangka panjang lainnya untuk tujuan perencanaan kapasitas strategis. Kapasitas redundanselalu bertindak sebagai bantalan untuk kebutuhan kapasitas standby. Kapasitas redundan adalah apa yang tersedia untuk ekspansi masa depan dari beban kerja saat ini. Untuk penyedia layanan cloud, keseluruhan manajemen kapasitas dan perencanaan akan berputar di sekitar mencari keseimbangan optimal di sekitar komponen-komponen ini — dengan kapasitas total untuk menyediakan layanan cloud menjadi jumlah semua komponen: kapasitas yang digunakan, kapasitas siaga, dan kapasitas yang berlebihan.

Kekritisan bisnis, pentingnya layanan untuk bisnis, harus ditafsirkan dengan baik untuk melihat kejadian tak terduga yang mengarah pada pemadaman

layanan; tindakan proaktif seperti redundansi, cadangan, pemulihan, dan failover harus berdiri teguh terhadap setiap tantangan downtime. Prosedur manajemen kapasitasharus dapat dengan jelas merumuskan persyaratan waktu kerja dalam hal tingkatan pusat data yang diorganisir oleh kritikalitas bisnis, persyaratan kinerja, dan rencana pertumbuhan. Selain itu, prosedur manajemen kapasitas harus mengawasi permintaan dan rencana pertumbuhan bisnis untuk memastikan bahwa permintaan dipenuhi oleh kapasitas siaga dan mampu mendukung tingkat layanan yang berkelanjutan.
Persyaratan layanan total dapat didorong dari persyaratan kinerja, kebutuhan pertumbuhan di masa depan, dan kekritisan bisnis dari layanan.

Persyaratan Kinerja

Persyaratan kinerja dapat mencakup persyaratan kapasitas saat ini seperti CPU, koneksi jaringan, saluran masukan – keluaran (Input/Output, I / O), dll. yang mungkin perlu dilakukan aplikasi pada tingkat yang diinginkan dan untuk memenuhi harapan pengguna. Ini dapat dikumpulkan dari persyaratan konfigurasi asli dari penyebaran aplikasi, pengujian kinerja, dan pemantauan kinerja aplikasi.Dalam model cloud, beberapa penyedia clouds dapat memberikan jaminan kinerja pada penyimpanan IOPS. Fitur-fitur ini dapat digunakan untuk memberikan jaminan kinerja untuk aplikasi. Persyaratan kinerja beban kerja dapat menentukan pilihan penyedia cloud berdasarkan jaminan kinerja atau berdasarkan arsitektur layanan cloud tertentu.

Kritisitas Bisnis

Kekritisan bisnismenentukan pentingnya aplikasi atau layanan untuk bisnis. Aplikasi dapat berupa aplikasi utama yang menjalankan bisnis atau aplikasi yang mendukung bisnis seperti email. Dampak apa pun pada aplikasi dapat memiliki dampak bisnis yang besar juga. Informasi kekritisan bisnis dapat dikumpulkan dari analisis dampak bisnis dan / atau rencana pemulihan bencana.Kekritisan bisnis adalah titik keputusan utama dalam migrasi ke lingkungan klien dan jenis penyedia cloud yang dapat dipilih oleh perusahaan. Tingkat layanan yang ditawarkan oleh penyedia cloud dan arsitektur awan publik yang mendasari dapat menjadi poin keputusan utama untuk migrasi cloud.Suatu perusahaan dapat memutuskan untuk

menyimpan aplikasi bisnis penting dalam cloud pribadi di dalam rumah untuk mendukung ketersediaan dan kebutuhan kinerja dari aplikasi tersebut.

Pertumbuhan Masa Depan

Ini menentukan pertumbuhan yang diharapkan dalam permintaan layanan dan selama periode tertentu, katakanlah, tiga tahun ke depan. Informasi diberikan dari rencana perkiraan berdasarkan pemanfaatan historis dan rencana bisnis yang akan datang seperti ekspansi pasar dan faktor lainnya.

Pendekatan manajemen kapasitas untuk penyedia layanan cloud harus memastikan bahwa kebutuhan bisnis pelanggan didukung oleh kapasitas optimal di semua lapisan dengan cara yang hemat biaya. Penyedia layanan cloud harus menyediakan dua opsi solusi kapasitas.

Pilihan pertama dapat dimulai dengan perencanaan dari awal menggunakan pengalaman sebelumnya, analisis ruang datacenter saat ini, data permintaan, analisis kebutuhan kapasitas pelanggan, tingkat layanan terjamin, ketersediaan dan kinerja kebutuhan layanan, dll. Hal ini dilakukan ketika bisnis baru layanan sudah dekat dan penyedia layanan harus mulai dari awal (yaitu, perencanaan ruang datacenter, chasis, pendinginan dan kebutuhan daya, perangkat keras, perangkat lunak dan alat-alat dan teknologi lainnya termasuk virtualisasi). Ini adalah prasyarat untuk merancang dan merencanakan kapasitas setiap kali layanan baru sedang dipertimbangkan.

Pilihan lainnya adalah manajemen kapasitas untuk layanan yang ada di mana pemantauan kinerja memainkan peran penting. Seiring dengan ini, analisis kinerja dilakukan untuk mempertahankan tingkat kinerja dan mencari tindakan peningkatan layanan berkelanjutan melalui manajemen ambang batas dan penyetelan.

Oleh karena itu, kita dapat meringkas tujuan manajemen kapasitass di lingkungan cloud sebagai berikut:

- Pengurangan pemborosan sumber daya
- Pemanfaatan sumber daya yang efisien
- Mendukung pemantauan dan manajemen tingkat layanan
- Manajemen beban kerja
- Memprakirakan pertumbuhan infrastruktur
- Mengontrol VM sprawl
- Alokasi sumber daya secara otomatis dalam kegagalan

- Menangani tuntutan yang tidak terduga dan musiman secara efektif
- Memastikan ekonomi cloud di lingkungan multi-cloud seperti cloud publik dan cloud hybrid

Menentukan Persyaratan Kapasitas untuk Layanan Baru

Dibagian ini menjelaskan bagaimana manajemen kapasitas dilakukan untuk layanan baru dan persyaratan kapasitas apa yang harus dikumpulkan untuk desain kapasitas berikutnya. Berbagai prosedur telah ditetapkan, memungkinkan penyedia layanan untuk menentukan persyaratan kapasitas untuk layanan baru. Prosedur ini termasuk persyaratan kapasitas pengumpulan melalui antisipasi permintaan, identifikasi fungsi bisnis yang penting, memahami implikasi biaya, mengumpulkan persyaratan terkait kinerja, dll. Penekanan telah diberikan pada menetapkan perjanjian tingkat layanan multi-vendor (Service Level Agreement, SLA), perjanjian tingkat operasional (*Operational Level Agreements*, OLAs), dan Underpinning Contracts (UCs) sehingga persyaratan kapasitas terpenuhi sebagaimana ditentukan. Persyaratan ini, setelah dikumpulkan, membentuk blok bangunan untuk desain kapasitas.

Perhitungan Kapasitas untuk Layanan Baru

Manajemen kapasitas terlibat dalam desain layanan baru atau yang berubah dan membuat rekomendasi untuk pengadaan infrastruktur cloud, di mana kapasitas dan / atau kinerja merupakan faktor. Keputusan-keputusan ini difasilitasi oleh skala algoritma dan perencanaan kapasitas. Manajemen permintaan juga memainkan peran penting dalam mencari tahu kebutuhan kapasitas infrastruktur yang tak terduga dengan memahami pola penggunaan yang secara resmi dikenal sebagai Pola Kegiatan Bisnis (PBA). Informasi permintaan harus lebih lanjut menelusuri elemen infrastruktur tingkat komponen layanan seperti CPU, memori, jaringan, dan penyimpanan untuk memastikan bahwa desain kapasitas dan penyediaan yang mencukupi ada untuk mendukung proses bisnis di masa depan.

Tentukan Persyaratan Kapasitas

Prosedurnya untuk menentukan persyaratan kapasitas adalah sebagai berikut:

- Memahami persyaratan kapasitas dan fungsi bisnis yang penting
- Penggabungan permintaan kapasitas
- Berikan biaya input kapasitas
- Tentukan target kinerja

Gambar 4-3 menjelaskan prosedur dan kegiatan yang harus dilakukan untuk mengumpulkan persyaratan kapasitas yang selaras dengan bisnis. Proses ITSM lainnya seperti manajemen permintaan, manajemen tingkat layanan, manajemen perubahan, dan manajemen keuangan menyediakan informasi yang diperlukan untuk pengumpulan kebutuhan kapasitas.

Gambar 4-3 . *Tentukan persyaratan kapasitas*

Persyaratan ini harus dipahami dengan baik dan — yang lebih penting — ditafsirkan dengan baik oleh proses manajemen kapasitas.

Memahami Persyaratan Kapasitas dan Fungsi Bisnis Vital

Ini adalah informasi yang terutama digunakan untuk kapasitas TI, ketersediaan layanan, dan kontinuitas layanan TI. Unsur-unsur bisnis penting dari proses bisnis yang didukung oleh layanan TI dipahami oleh manajemen kapasitas untuk pertimbangan khusus dalam persyaratan desain yang mendukung kinerja yang diinginkan. Sebagai contoh, klien dapat memilih salah satu model penyebaran cloud sesuai kebutuhan bisnis mereka.

Layanan bisnis dan keuangan (BFS)pelanggan mungkin ingin menjaga staf inti dan aplikasi pada cloud pribadi dan staf pihak ketiga dan kontrak lainnya pada cloud publik, termasuk layanan seperti e-mail dan pesan. Perencanaan kapasitas harus mampu merumuskan dan menangani kebutuhan bisnis tersebut secara efisien dan efektif.

Untuk layanan bisnis baru apa pun, penyedia layanan harus menyelami pemahaman dan memutuskan tingkat layanan yang sesuai. Tingkat layanan harus dipetakan dengan aplikasi dan persyaratan infrastruktur. Berbagai kasus bisnis harus disiapkan untuk mendapatkan tampilan kapasitas jangka panjang. Berbagai parameter kapasitas seperti kapasitas yang digunakan, kapasitas siaga, dan kapasitas yang berlebihan harus dipertimbangkan untuk mengumpulkan semua kemungkinan kapasitas yang dibutuhkan untuk semua fungsi bisnis penting. Berdasarkan kapasitas dan persyaratan redundansi untuk mendukung proses bisnis, manajemen kapasitas harus memastikan bahwa semua komponen kapasitas seperti konfigurasi pusat data tier-bijaksana, ruang, perangkat keras, dan alat-alat di tempat untuk melayani kekritisan aplikasi bisnis dan kebutuhan kinerja.

Berdasarkan fungsi bisnis penting, penyedia layanan cloud mungkin berpikir untuk menyediakan kapasitas yang diperlukan untuk mengakomodasi berbagai tingkatan pusat data (yaitu, Tingkat 1, 2, 3, atau 4). Kenaikan tingkat tier sebagai ketersediaan meningkat dan menurun sesuai dengan perkiraan waktu henti. Peringkat peringkat pusat data dapat membantu penyedia dalam merencanakan kapasitas bisnis dan kebutuhan kinerja. Tingkatan data center ini dapat diklasifikasikan sebagai berikut:

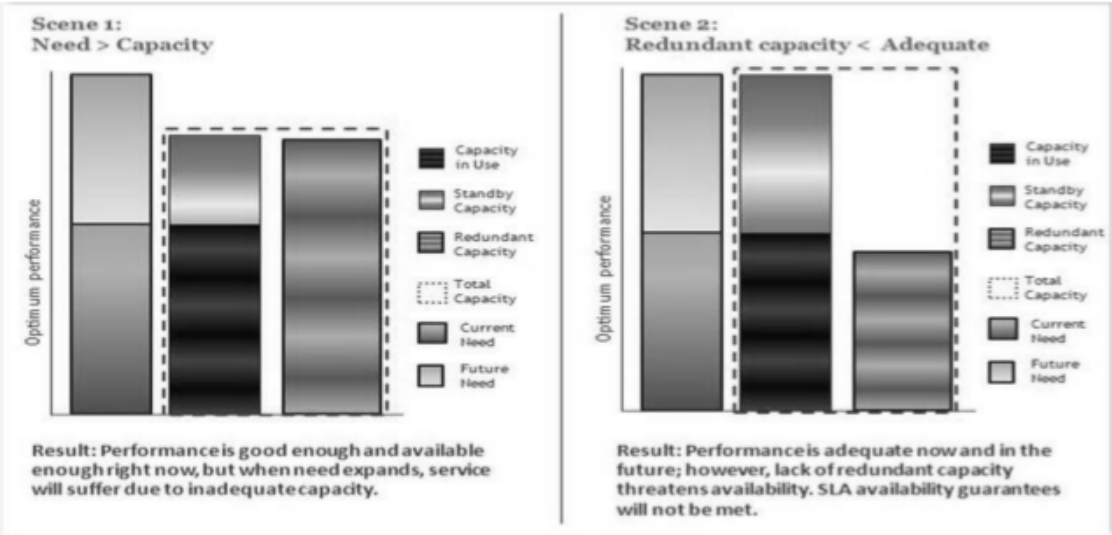
- Tingkat 1: Infrastruktur dasar situs dengan komponen kapasitas tidak berlebihan
- Tingkat 2: Komponen kapasitas infrastruktur situs yang redundan
- Tingkat 3: Infrastruktur situs yang dapat dipertahankan secara bersamaan

- Tingkat 4: infrastruktur situs Fault tolerant

Standar pusat data dasar dan referensi konfigurasi harus diingat ketika memutuskan pada tingkatan pusat data. Sebagai contoh, campuran konfigurasi komponen pusat data (DC) dari Tingkat 2 dan 3 akan menghasilkan standar konfigurasi Tingkat 2 DC. Selain itu, pendekatan kapasitas untuk pusat data dapat mempertimbangkan parameter lain, seperti ruang dan tata letak DC, standar pemasangan kabel, daya, pendinginan, tingkatan seperti yang dibahas di sini, dan pertimbangan lingkungan dan peraturan lainnya.

Persyaratan kapasitas di semua tingkatan (yaitu bisnis, layanan, dan komponen) harus didiskusikan dan disepakati, dan tanda formal harus diberikan. Kriteria kinerja dalam hal kapasitas (fisik dan server virtual, basis data, middleware, penyimpanan, jaringan, fasilitas, dll.) Dan tingkat layanan yang berasal dari bisnis dan pengguna TI harus dipahami dan diterjemahkan dalam hal tingkat pemanfaatan.,

Seperti yang digambarkan pada Gambar 4-4 , kapasitas yang berlebihan perlu direncanakan untuk memenuhi prediksi pertumbuhan. Jika kapasitas yang berlebihan tidak direncanakan, layanan TI mungkin tidak dapat memberikan layanan karena kurangnya kapasitas.



Gambar 4-4 . *Persyaratan kapasitas dan fungsi bisnis yang penting*

Cloud computing menyediakan jalan keluar karena perusahaan tidak perlu menyimpan kapasitas yang berlebihan tetapi dapat memperoleh kapasitas sesuai permintaan dari penyedia cloud. Namun, dari perspektif penyedia cloud, kapasitas redundan harus tersedia untuk melayani beban kerja variabel dan permintaan variabel dari pelanggan.
Oleh karena itu, persyaratan kapasitas totalmerupakan agregasi untuk memenuhi sasaran kinerja saat ini, kebutuhan masa depan, dan tujuan ketersediaan / pemulihan semua aplikasi dan layanan. Kebutuhan masa depan ditutupi oleh kapasitas siaga di atas apa yang saat ini sedang digunakan.
Ketersediaan / pemulihan biasanya diaktifkan melalui redundansi. Redundansi ini dapat berupa hal-hal berikut:

- Redundansi komponen
- Sumber daya penuh redundansi (server redundan, array penyimpanan, switch, UPS (power supply yang tidak pernah terputus), pendinginan)
- Redundansi data (snapshot data, cermin, salinan cadangan)

Jika permintaan melebihi kapasitas yang tersedia untuk pertumbuhan yang direncanakan dan / atau tidak meninggalkan kapasitas yang berlebihan, tingkat layanan akan dikompromikan. Kapasitas fisik tambahan perlu ditambahkan atau beban kerja lain harus dikeluarkan dari kolam kapasitas yang tersedia.

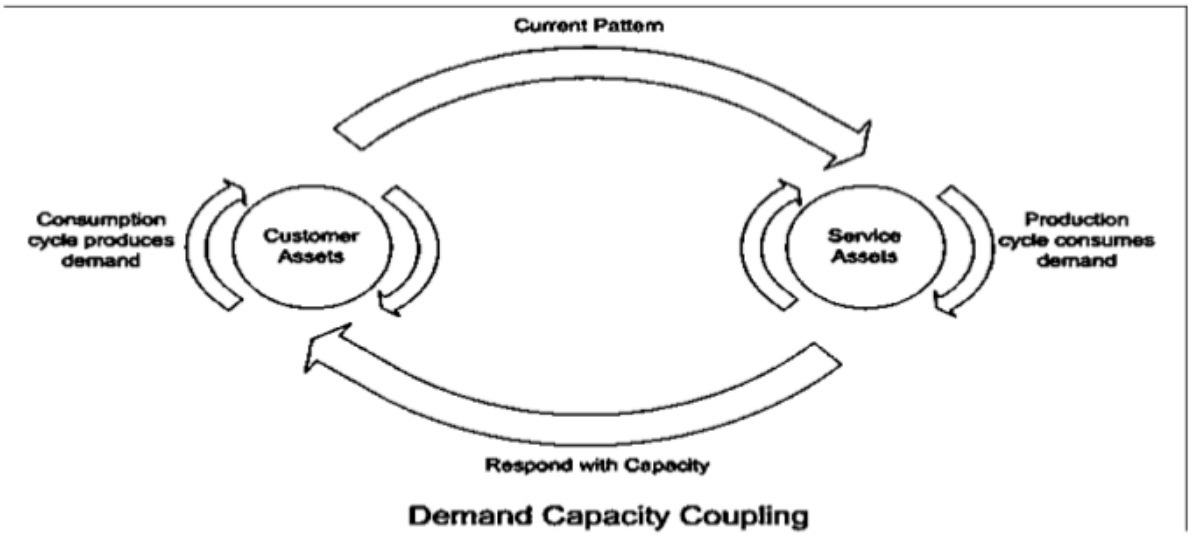
Memahami Persyaratan Disaster Recovery

Penting juga untuk mempertimbangkan persyaratan pemulihan bencana (DR) sambil merencanakan kapasitas. Situs DR akan memerlukan kapasitas untuk memastikan bahwa dalam kasus kegagalan, sistem penting yang ditandai untuk dipindahkan ke situs DR dapat melakukan seperti yang didefinisikan dalam SLA. Situs DR dapat pasif dan hanya dibesarkan jika terjadi bencana yang dinyatakan di situs utama. Atau, situs DR dapat menjadi situs aktif. Dalam hal ini, situs DR memiliki contoh aktif berjalan dan digunakan, dan dua situs digunakan sebagai failover untuk satu sama lain. Perencanaan kapasitas harus menyediakan untuk pemantauan dan pengelolaan kapasitas di situs DR dengan cara yang sama dengan itu untuk situs utama. Penyedia cloud dapat menyediakan beberapa pusat data dan opsi kepada pelanggan untuk memanfaatkan banyak pusat data untuk menyediakan

kemampuan pemulihan bencana kepada pelanggan. Pelanggan yang memiliki cloud pribadi atau penyebaran tradisional di rumah dapat memanfaatkan penyedia cloud untuk DR.

Permintaan kapasitas

Kapasitas untuk layanan harus didasarkan pada ramalan permintaan dan pola aktivitas bisnis. Kapasitas dikendalikan oleh permintaan. Artinya, siklus konsumsimengkonsumsi permintaan dan siklus produksi menghasilkan permintaan (lihat Gambar 4-5).



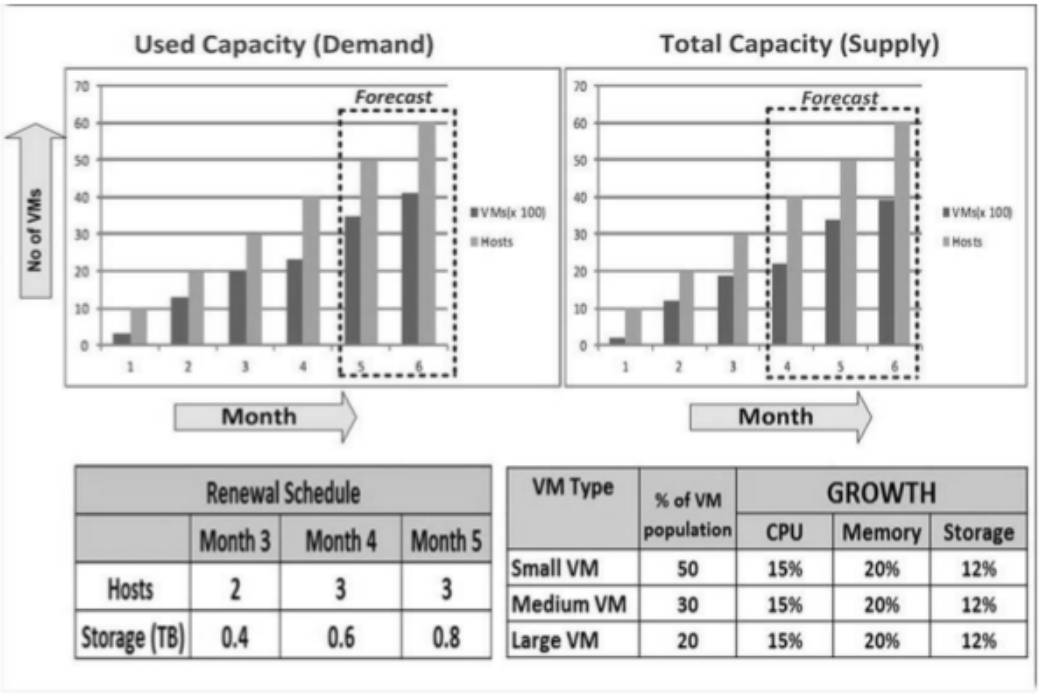
Gambar 4-5 *. Kapasitas dan permintaan*

Siklus konsumsi menghasilkan permintaan untuk lebih banyak sumber daya. Sebagai contoh, peningkatan pengguna layanan e-mail akan menghasilkan peningkatan permintaan kapasitas, yang dikonsumsi oleh siklus produksi. Dengan demikian, permintaan dari pelanggan atau pengguna mendorong permintaan kapasitas dari siklus produksi back-end untuk memenuhi permintaan tersebut.

Dalam skenario penyedia cloud, permintaan berasal dari pengguna layanan cloud. Dalam lingkungan perusahaan, perusahaan dapat memanfaatkan penyediaan sumber daya sesuai permintaan untuk melayani kapasitas permintaan yang lebih baik

dengan peningkatan kelincahan. Proses manajemen permintaan menyediakan masukan untuk Pola Kegiatan Bisnis (PBA) untuk membantu penyedia cloud agar lebih memahami persyaratan terkait kapasitas untuk memenuhi permintaan layanan di masa mendatang secara efektif. Pola permintaan ini dapat diformulasikan dalam hal kebutuhan untuk infrastruktur cloud yang akan ditambahkan selama periode waktu tertentu. Suatu pola analisis aktivitas bisnis harus dilakukan oleh orang yang mengelola permintaan dan kapasitas secara kolektif. Informasi seperti indikator kinerja utama (KPI),Skor CSAT, dan survei pelanggan dapat digunakan untuk melakukan analisis ini. Perencanaan kapasitas bisnis dapat dimulai di sini berdasarkan analisis PBA, pemodelan aktivitas bisnis, perkiraan, dan data permintaan. Pola-pola ini dikodifikasikan, dan profil pengguna dibuat. Ada berbagai toolsset dengan algoritma built-in dan metode statistik untuk melakukan analisis tersebut. Masukan pengelolaan permintaan harus dapat membantu perencana kapasitas untuk secara proaktif mengantisipasi penggunaan kapasitas yang terkait dengan layanan. Ini akan memastikan bahwa keseimbangan antara siaga dan kapasitas saat ini dipertahankan dan tingkat layanan didukung oleh manajemen kapasitas tanpa kejutan yang tidak diinginkan.

Gambar 4-6. menyediakan contoh yang mencakup perkiraan VM baru dengan campuran berbagai jenis mesin virtual. Ramalan ini kemudian digunakan oleh proses perencanaan kapasitas untuk merencanakan penambahan kapasitas dan garis waktu ketika kapasitas baru ini perlu ditambahkan.



Renewal Schedule			
	Month 3	Month 4	Month 5
Hosts	2	3	3
Storage (TB)	0.4	0.6	0.8

VM Type	% of VM population	GROWTH		
		CPU	Memory	Storage
Small VM	50	15%	20%	12%
Medium VM	30	15%	20%	12%
Large VM	20	15%	20%	12%

Gambar 4-6 *. Peramalan infrastruktur virtual dan perencanaan sumber daya*

Pemantauan Permintaan

Pemantauan permintaan adalah tentang memahami seberapa baik Anda saat ini mendukung tuntutan kapasitas pengguna Anda. Memantau tingkat permintaan kapasitas dapat memberi tahu Anda berapa banyak pengguna yang membutuhkan sumber daya dan berapa banyak sumber daya yang dikonsumsi setiap aktivitas bisnis. Ini juga memberikan informasi berharga tentang tingkat kapasitas sumber daya saat ini sehingga TI akan tahu cara mendukung layanan baru saat diperkenalkan ke lingkungan, serta bagaimana layanan baru akan memengaruhi SLA saat ini. Dalam hal ini, penting untuk memahami kapasitas efektif dari suatu sumber daya.

Ketika perkiraan kebutuhan kapasitas di masa mendatang diatasi melalui Capex untuk mengantisipasi permintaan di masa mendatang, kapasitas penyedia meningkat, dan ada kelebihan kapasitas untuk periode waktu ketika permintaan meningkat tetapi

belum menghabiskan semua kapasitas yang disediakan. Permintaan meningkat dengan kapasitas yang tersedia dalam skenario permintaan yang meningkat dan akan menghasilkan kapasitas yang kurang. Dalam skenario permintaan yang menurun, akan ada kelebihan kapasitas.

Triknya adalah menyeimbangkan permintaan dan penawaran melalui perubahan harga dan upaya lain termasuk layanan baru dan geografi baru untuk menangani skenario permintaan yang menurun. Kemampuan penyedia cloud untuk mengantisipasi naik dan turunnya permintaan adalah kunci untuk menjadi penyedia cloud yang sukses.

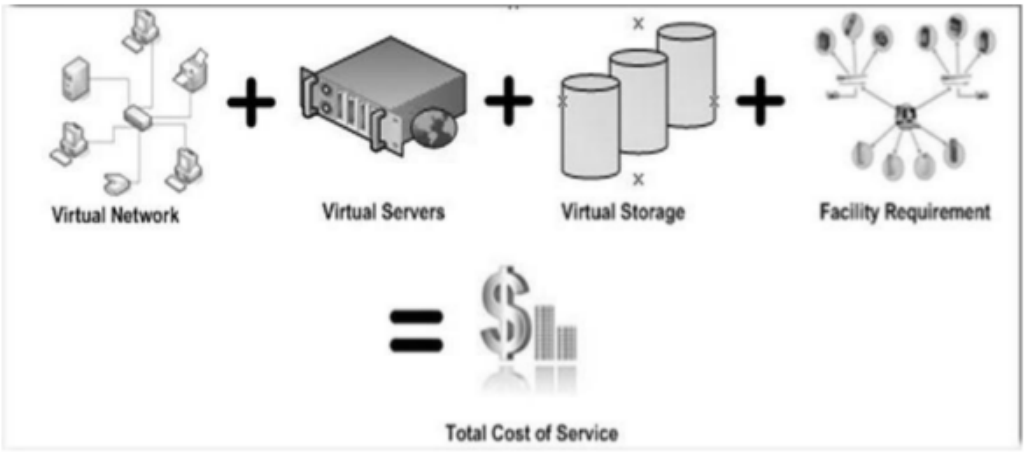
Dalam kasus di bawah kapasitas, konsumen cloud mungkin tidak dapat menyediakan sumber daya atau kinerja SLA akan menderita, yang akan mengakibatkan ketidakpuasan pelanggan dan kerugian finansial karena kehilangan kesempatan untuk menjual kapasitas tambahan.

Dengan demikian, keseimbangan antara kapasitas yang tersedia, kapasitas siaga, dan kapasitas yang berlebihan harus disesuaikan dengan baik melalui analisis dan alat canggih.

Menyediakan Biaya Input Kapasitas

Dalam hubungannya dengan anggaran keuangan dan rencana, manajemen kapasitas harus menyediakan biaya untuk memenuhi permintaan tingkat layanan yang ditentukan. Ini membentuk dasar negosiasi SLA antara penyedia layanan cloud (pencipta / aggregator) dan pelanggan. Ini bisa menjadi aktivitas berulang hingga SLA dinegosiasikan. Untuk cloud mana pun sumber daya dapat digunakan dalam mode swalayan, model biaya, metering, dan prosedur tolak bayar untuk sumber daya harus tersedia. Ada beberapa fitur utama saat menerapkan tolak bayar. Tagihan balik dan showback harus didasarkan pada penggunaan sumber daya yang dialokasikan atau aktual. Model biaya berbasis alokasi akan membebankan pelanggan pada penggunaan berdasarkan durasi, dan model berbasis penggunaan akan mengenakan biaya kepada pelanggan atas dasar penggunaan aktual. Perizinan, kekuasaan, tipe dan ruang data center, dan biaya lain yang terkait dengan alokasi pelanggan harus diperhitungkan untuk mencerminkan total biaya yang dikenakan setiap penyewa. Pelaporan tagihan balik perlu diintegrasikan ke dalam sistem keuangan atau sistem anggaran lainnya

Dengan demikian biaya layanan didasarkan pada komponen individu dari layanan (seperti biaya fasilitas, biaya jaringan, biaya server virtual, dan biaya penyimpanan virtual). Penyedia cloud biasanya menyediakan tagihan terpadu dan dapat menelusuri ke masing-masing komponen biaya (lihat Gambar 4-7).



Gambar 4-7 *. Biaya input kapasitas*

Konsumen cloud dapat mendasarkan keputusan pengadaan cloud pada tagihan keseluruhan serta komponen individu yang disediakan oleh penyedia cloud. Sebagai contoh, jika penggunaan jaringan aplikasi sangat tinggi, konsumen cloud dapat memutuskan untuk menggunakannya dari pusat data di-tempat daripada penyedia cloud.

Menentukan Target Kinerja

SLA yang dinegosiasikan diterjemahkan ke dalam target kinerja spesifiks yang didukung oleh manajemen. Ini membentuk dasar negosiasi OLA dan UCs untuk pemenuhan SLA. SLA antara penyedia layanan cloud dan pelanggan harus melakukan hal berikut:

- Tentukan layanan apa yang akan disediakan.
- Tetapkan cara di mana layanan (bagaimana) akan disediakan.
- Menetapkan standar kualitas yang ingin dicapai.
- Tentukan kriteria pengukuran.

- Tetapkan kriteria pelaporan.
- Tetapkan biaya pengiriman.

Dalam lingkungan multi-penyedia, kompleksitas SLA meningkat. Karena setiap penyedia hanya menyediakan bagian layanan, SLA, OLA, dan UCs harus dibuat untuk mencakup skenario yang kompleks.

Alat pengukuran untuk perhitungan tingkat layanan diperlukan untuk menghitung skenario tingkat layanan yang kompleks. Pemantauan perlu dilakukan untuk menentukan di mana layanan gagal sehingga tingkat layanan yang tepat dihitung secara tepat.

Penyedia Layanan Cloud

SLA dengan penyedia layanansebagian besar akan distandarisasi karena tingkat layanan dan target ketersediaan mereka distandardisasi. Penyedia layanan terkadang dapat menyesuaikan SLA mereka berdasarkan volume layanan yang diberikan kepada agregator layanan atau konsumen. Manajemen SLA untuk penyedia layanan adalah tugas yang relatif sederhana dibandingkan dengan agregator layanan cloud. Penyedia layanan juga mungkin memiliki OLA dalam organisasi.

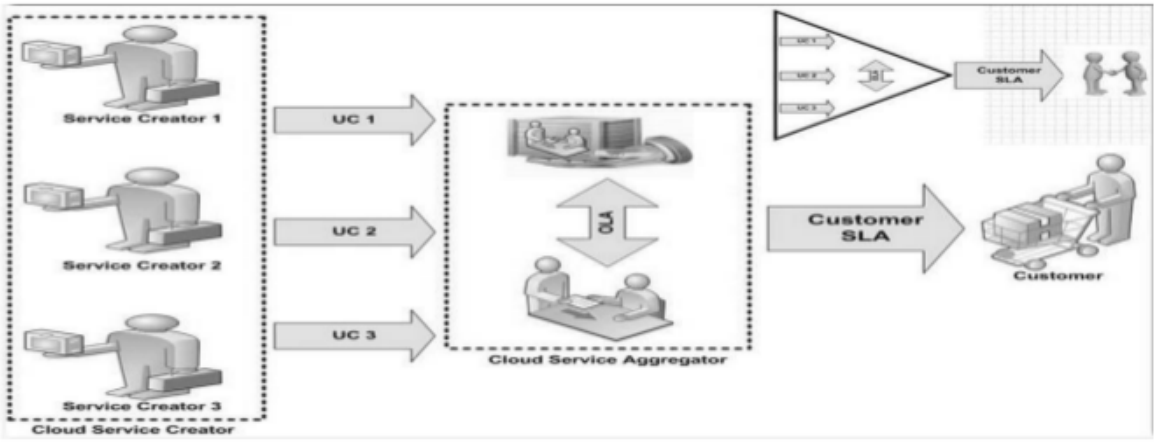
Agregator Layanan Cloud

Manajemen tingkat layanan untuk agregator layananadalah proses yang rumit karena keterlibatan semua jenis perjanjian, seperti SLA dengan pelanggan dan OLA dalam entitas organisasi. Semua SLA dan lingkungan multi-vendor berkontribusi pada SLA pelanggan. Agregator layanan cloud juga mungkin perlu dilibatkan dalam mengelola UCs dengan pembuat layanan lainnya.

Konsumen

SLA antara penyedia layanan cloud dan pelanggan terutama harus mempertimbangkan kebutuhan dan layanan bisnis pelanggan yang mendukungnya. Ketika penyedia layanan cloud memastikan bahwa persyaratan aplikasi dapat dilayani sesuai kebutuhan, negosiasi pada klausul SLA, penalti, kebijakan sengketa, dll. Harus ditetapkan.

Untuk konsumen cloud, tantangannya adalah penyedia cloud dan agregator cloud telah mempublikasikan SLA dan yang diperbaiki. Konsumen tidak memiliki fleksibilitas untuk menyempurnakan atau mengubah SLA sesuai kebutuhan mereka karena mereka membeli layanan standar dari penyedia cloud atau agregator.
Konsumen cloud harus memperhitungkan kompleksitas dan cara untuk memantau SLA penyedia cloud untuk menyediakan layanan yang diperlukan kepada organisasi internal dan melacak penalti dan biaya layanan.
Gambar 4-8 menggambarkan perjanjian yang ada dalam rantai pengiriman cloud. Dalam rantai nilai cloud, ada penyangga, hubungan perjanjian layanan UP antara agregator layanan cloud dan pencipta layanan, perjanjian tingkat operasional, OLA, dalam organisasi agregasi dan perjanjian tingkat layanan, dan SLA dengan pelanggan.



Gambar 4-8 . *Perjanjian layanan cloud*

Penyedia terikat oleh kontrak untuk memberikan tingkat layanan tertentu. Banyak layanan yang ditentukan dalam kontrak mereka memiliki dampak langsung atau tidak langsung pada tindakan peserta lain dalam rantai pengiriman layanan. Jika tindakan tersebut tidak dikelola secara memadai, maka penyedia layanan dapat berisiko gagal bayar pada tingkat layanan dan dengan demikian kontrak.
Titik penting dari manajemen kapasitas di lingkungan virtual adalah untuk mendapatkan pegangan pada tren kinerja. Memahami tingkat pertumbuhan komponen tertentu dan apa yang akan terjadi jika tren berlanjut sangat

penting. Selain itu, dengan melacak nilai historis, penyedia cloud dapat menentukan akar penyebab lebih tepat dalam suatu insiden.

BAB LIMA

PENGUKURAN KAPASITAS

Sebagian besar sistem operasi dilengkapi dengan beberapa utilitas bawaan dasar yang dapat mengukur berbagai metrik kinerja dan konsumsi sumber daya. Sebagian besar utilitas ini biasanya menyediakan cara untuk merekam hasil, juga. Misalnya, di Linux, perintah berikut inibiasanya digunakan:

uptime

Anda menggunakan ini untuk melihat rata-rata beban, yang pada gilirannya menunjukkan jumlah tugas (proses) yang diantrikan untuk dijalankan.

dmesg

Anda menggunakan ini untuk melihat 10 pesan sistem terakhir, jika ada, dan mencari kesalahan yang dapat menyebabkan masalah kinerja.

vmstat 1

Ini memberikan ringkasan statistik server utama — seperti proses yang berjalan pada CPU dan menunggu giliran, membebaskan memori dalam kilobyte, pertukaran, dan pertukaran — setiap detik.

mpstat –P ALL 1

Ini memberikan perincian waktu CPU per CPU setiap detik.

pistat 1

Ini memberikan ringkasan per-proses secara bergulir.

iostat –xz 1

Anda menggunakan ini untuk memahami perangkat blok (disk), baik beban kerja yang diterapkan dan kinerja yang dihasilkan.

free –m

Anda dapat menggunakan ini untuk melihat jumlah memori bebas yang tersedia dan ukuran buffer dan file cache.

sar –n DEV 1

Gunakan ini untuk melihat throughput antarmuka
jaringan: rxkB/sdan txkB/s, sebagai ukuran beban kerja dan juga untuk memeriksa apakah batas telah tercapai.

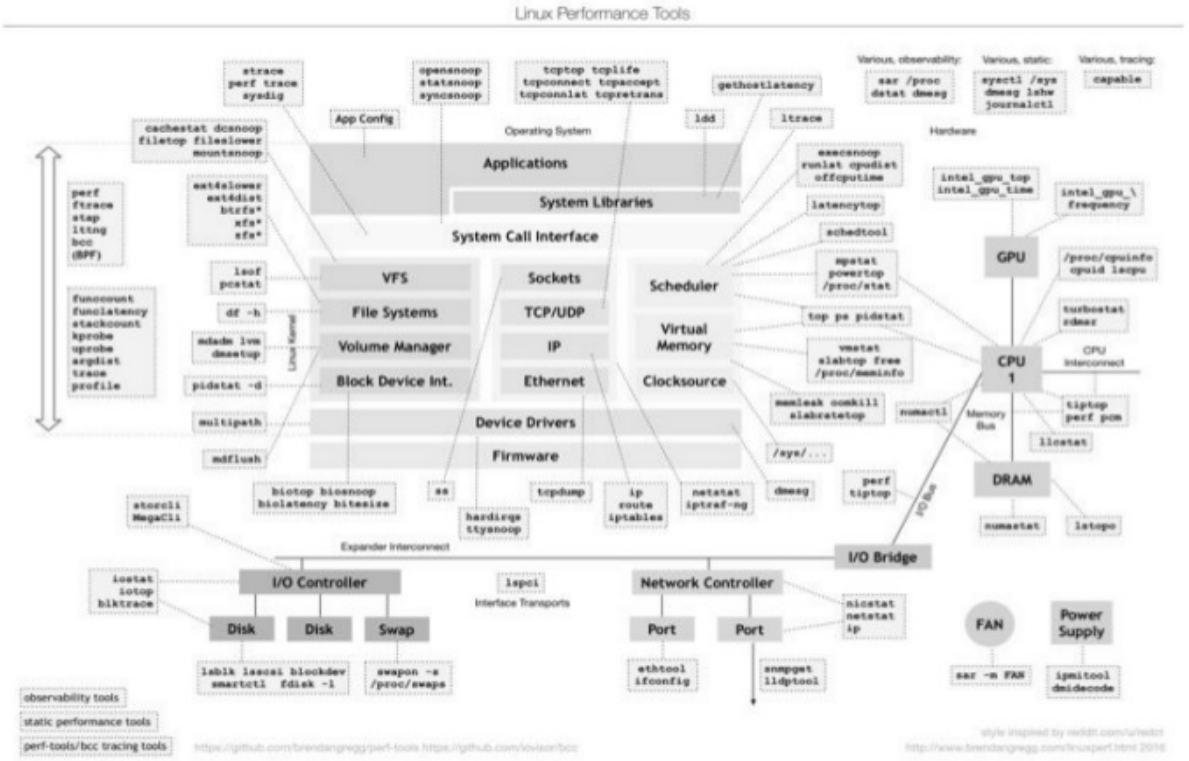
sar –n TCP,ETCP 1

Anda dapat menggunakan ini untuk melihat beban server dalam hal jumlah koneksi TCP lokal / jarak jauh yang dimulai per detik dan jumlah TCP mentransmisikan kembali per detik.

top

Gunakan ini untuk melihat ringkasan sebagian besar metrik yang diekspos oleh perintah yang disebutkan sebelumnya.

Alat-alat lain seperti netstat, tcpstat, dan ncstatyang umum digunakan. Gambar 5-1 menyajikan ikhtisar dari berbagai alat yang tersedia untuk Linux.



Gambar 5-1. *Lanskap alat pemantauan tersedia di Linux*
sumber: http://www.brendangregg.com/Perf/linux_perf_tools_full.png)

Selain alat yang ditunjukkan pada Gambar 3-1, ada alat yang lebih canggih dan memungkinkan pemantauan pada tingkat kernel. Sebagai contoh, Anda dapat menggunakan SystemTap untuk mengekstrak, menyaring, dan meringkas data sehingga Anda dapat mendiagnosa kinerja yang kompleks atau masalah fungsional dari sistem Linux. Dengan menggunakan skrip SystemTap, Anda dapat menamai acara dan memberi mereka penangan. Setiap kali peristiwa tertentu terjadi, kernel Linux menjalankan handler seolah-olah itu adalah subrutin cepat dan kemudian dilanjutkan. Ada banyak jenis acara lain seperti memasukkan atau keluar dari fungsi, pengatur waktu berakhir, atau seluruh sesi SystemTap mulai atau berhenti. Handler adalah serangkaian pernyataan bahasa skrip yang menentukan pekerjaan yang harus dilakukan setiap kali peristiwa terjadi. Pekerjaan ini biasanya termasuk mengekstraksi data dari konteks acara, menyimpan data ke dalam variabel internal, atau mencetak hasil.

Serupa dengan SystemTap (stap), Anda juga dapat menggunakan ktap — yang didasarkan pada bytecode, sehingga tidak bergantung pada GNU Compiler Collection (GCC) dan tidak perlu menyusun modul kernel untuk setiap skrip — untuk pelacakan dinamis kernel Linux.Dalam nada yang sama, Linux ditingkatkan BPF (Berkeley Packet Filter) juga memiliki kemampuan penelusuran mentah, dan Anda dapat menggunakannya untuk melakukan analisis kustom dengan melampirkan BPF bytecode dengan tracing dinamis kernel Linux (kprobe), tracing dinamis tingkat pengguna (uprobes), pelacakan statis kernel (tracepoints), dan profil acara. eBPF dijelaskan oleh Ingo Molnár sebagai berikut:

Salah satu fitur yang lebih menarik dalam siklus ini adalah kemampuan untuk melampirkan program-program eBPF (user-defined, bytecode sandbox yang dijalankan oleh kernel) ke kprobes. Ini memungkinkan instrumentasi yang ditentukan pengguna pada imej kernel langsung yang tidak pernah bisa crash, menggantung atau mengganggu kernel secara negatif.

Tidak seperti pelacak built-in lainnya di Linux, eBPF dapat meringkas data dalam konteks kernel dan hanya memancarkan ringkasan yang Anda minati ke tingkat pengguna — misalnya, histogram latensi dan filesystem I / O. Anda dapat menggunakan eBPF dalam berbagai konteks yang luas, seperti Mitigasi Perangkat Lunak Jaringan (SDN), Mitigasi Penyangkalan Terdistribusi (DDoS), dan deteksi intrusi.

Diskusi sejauh ini telah berada di tingkat tuan rumah. Namun, pemantauan tingkat host tidak cukup dalam lingkungan yang tervirtualisasi atau

terbungkus. Untuk tujuan ini, baik mesin virtual (VMs) dan kontainer datang bersama dengan alat untuk mengekspos CPU, memori, I / O, dan metrik jaringan berdasarkan per-VM / kontainer. Selain metrik standar, metrik khusus kontainer seperti CPU throttling juga diekspos. Untuk Misalnya, Docker memunculkan jumlah pelambatan kali diberlakukan untuk masing-masing wadah, dan total waktu setiap kontainer dikompresi. Demikian pula, Docker menghadapkan metrik yang disebut memori kontainer gagal, yang meningkat setiap kali alokasi memori gagal, yaitu, setiap kali batas memori preset dipukul. Jadi, lonjakan dalam metrik ini menunjukkan bahwa satu atau lebih kontainer membutuhkan lebih banyak memori daripada yang dialokasikan. Jika proses dalam penampung berakhir karena kesalahan ini, Anda juga mungkin melihat peristiwa di luar memori dari Docker.

Alat sumber terbuka paling mudah diunduh dan dijalankan di hampir semua sistem modern. Untuk perencanaan kapasitas, alat ukur harus menyediakan, paling tidak, cara mudah untuk melakukan hal berikut:

- Catat dan simpan data dari waktu ke waktu — mempertahankan riwayat adalah kunci untuk banyak alasan seperti peramalan dan analisis tren
- Buat metrik khusus
- Bandingkan metrik dari berbagai sumber seperti RRD , Hadoop, OpenTSDB, dan seterusnya
- Impor dan ekspor metrik seperti, misalnya, CSV, JSON, dan sebagainya

Selama Anda memilih alat yang dapat memuaskan kriteria yang disebutkan di atas, Anda tidak perlu menghabiskan banyak waktu untuk merenungkan mana yang akan digunakan. Yang lebih penting adalah menentukan metrik mana yang harus diukur dan metrik mana yang memberikan perhatian khusus.

Aplikasi Pemantauan

Sisa bab ini digunakan contoh untuk mendemonstrasikan beberapa teknik pemantauan penting yang perlu Anda ketahui dan lakukan.

Pengukuran Tingkat Aplikasi

Seperti disebutkan sebelumnya, statistik server hanya menggambarkan sebagian dari gambar kapasitas. Anda juga seharusnya mengukur dan merekam metrik tingkat lebih tinggi khusus untuk aplikasi — tidak spesifik untuk satu server, tetapi ke seluruh sistem.Penggunaan CPU dan disk pada server web tidak menceritakan kisah lengkap tentang apa yang terjadi pada setiap permintaan web, dan aliran permintaan web dapat melibatkan banyak perangkat keras. Contoh metrik tingkat aplikasi mencakup jumlah Tweet / mnt, jumlah Foto yang diunggah / mnt (Instagram), jumlah Pesan / mnt (WhatsApp), jumlah Aliran serentak / mnt (Netflix). Lebih lanjut, metrik tingkat aplikasi sering dikumpulkan dengan perincian yang berbeda — oleh detik, per menit, harian, mingguan, bulanan, atau tahunan — tergantung pada use case.

Kembali di Flickr, metrik tingkat aplikasi dikumpulkan pada basis harian dan kumulatif. Beberapa metrik dapat diambil dari database, seperti jumlah foto yang diunggah. Lainnya berasal dari menggabungkan beberapa statistik server, seperti total ruang disk yang dikonsumsi di mesin yang berbeda. Teknik pengumpulan data bisa sesederhana menjalankan skrip dari tugas cron dan memasukkan hasilnya ke dalam database sendiri untuk penambangan di masa mendatang. Beberapa metrik yang dilacak termasuk yang berikut:

- Foto yang diunggah (harian, kumulatif)
- Foto diunggah per jam
- Ukuran foto rata-rata (harian, kumulatif)
- Memproses waktu untuk memisahkan foto berdasarkan ukurannya yang berbeda (setiap jam)
- Pendaftaran pengguna (harian, kumulatif)
- Pendaftaran akun Pro (harian, kumulatif)
- Jumlah foto yang ditandai (harian, kumulatif)
- Lalu lintas API (kunci API digunakan, permintaan dibuat per detik, per kunci)
- Jumlah tag unik (harian, kumulatif)
- Jumlah foto yang diberi geotag (harian, kumulatif)

Metrik keuangan tertentu seperti pembayaran yang diterima (yang berada di luar cakupan buku ini) juga dilacak. Untuk aplikasi apa pun, ini adalah latihan yang baik untuk meluangkan waktu untuk menghubungkan bisnis dan data keuangan dengan sistem dan metrik aplikasi yang dilacak.

Misalnya, perhitungan Total Biaya Kepemilikan (TCO) tidak akan lengkap tanpa indikasi berapa banyak metrik sistem dan aplikasi ini membebani bisnis.Bayangkan kemampuan untuk mengkaitkan biaya nyata untuk melayani satu halaman web suatu aplikasi. Memiliki perhitungan ini tidak hanya akan menempatkan arsitektur ke dalam konteks yang berbeda dari operasi web (metrik bisnis, bukan ketersediaan, atau metrik kinerja), tetapi mereka juga dapat memberikan konteks bagi manajemen atas yang lebih terdidik keuangan dan non-teknis yang mungkin memiliki akses ke alat.

Kami tidak dapat terlalu menekankan nilai yang melekat pada identifikasi dan pelacakan metrik aplikasi. Upaya ini akan diberi imbalan dengan menjiwai statistik sistem dengan konteks di luar kesehatan server, dan akan membantu memandu ramalan. Selama proses pengadaan, perhitungan TCO akan terbukti tidak ternilai, seperti yang akan kita lihat nanti.

Sekarang kita telah membahas dasar-dasar pengukuran kapasitas, mari kita lihat pengukuran Anda, sebagai pengelola situs web yang berpotensi berkembang pesat, kemungkinan besar ingin memberikan perhatian khusus. Kami membahas elemen umum infrastruktur web dan pertimbangan daftar untuk mengukur kapasitas mereka dan menetapkan batas atas mereka. Kami juga menyediakan beberapa contoh yang diambil dari perencanaan kapasitas Flickr sendiri untuk menambah relevansi yang lebih besar. Contoh-contoh ini dirancang untuk menggambarkan metrik berguna yang mungkin ingin Anda lacak juga. Mereka tidak dimaksudkan untuk menyarankan arsitektur atau implementasi Flickr akan sesuai dengan lingkungan setiap aplikasi.

Kapasitas penyimpanan

Topik penyimpanan data sangat luas. Untuk tujuan kami, kami akan fokus hanya pada segmen penyimpanan yang secara langsung memengaruhi perencanaan kapasitas untuk situs web dengan volume data tinggi.

Salah satu analogi penyimpanan yang paling efektif adalah segelas air. Analogi menggabungkan batas hingga (ukuran gelas) dengan variabel (jumlah air yang dapat

dimasukkan ke dalam dan dibawa keluar dari kaca pada waktu tertentu). Ini membantu seseorang untuk memvisualisasikan dua faktor utama yang perlu dipertimbangkan ketika memilih di mana dan bagaimana cara menyimpan data:

- Kapasitas maksimum media penyimpanan
- Tingkat di mana data dapat diakses

Secara tradisional, sebagian besar operasi web berkaitan dengan pertimbangan pertama — ukuran kaca. Namun, sebagian besar vendor penyimpanan komersial telah menyelaraskan keluarga produk mereka dengan mempertimbangkan kedua pertimbangan tersebut. Dalam banyak kasus, ada dua opsi dalam hal hard disk drive (HDD):

- Disk yang besar, lambat, murah — biasanya menggunakan protokol ATA / SATA
- Lebih kecil, cepat, mahal disk-SCSI (Serial Computer System Interface) dan SAS (Serial Attached SCSI)

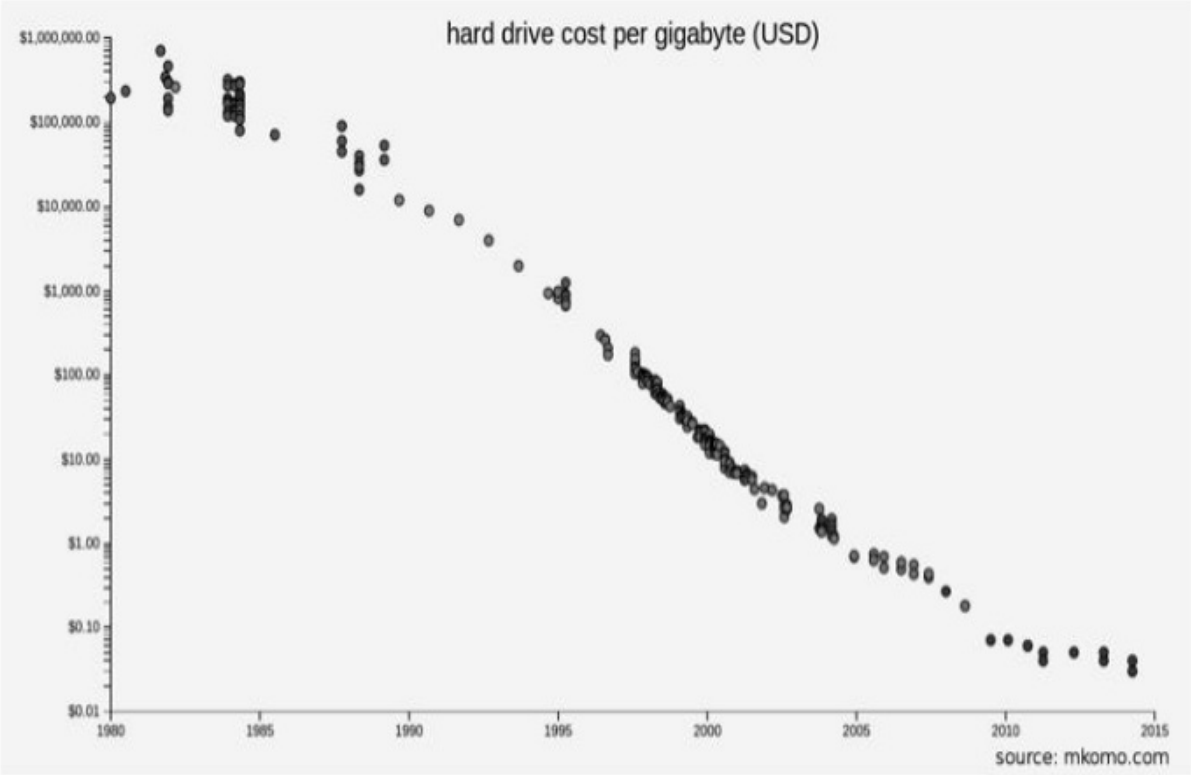
Pada 8 Januari 2017, harga HDD Seagate 4TB, 7200 RPM, 6 Gbps, 128 MB cache adalah $ 169 dan $ 180 untuk versi SATA dan SAS, masing-masing. Perhatikan bahwa kinerja acak atau transaksional (IOPS) dari HDDS didominasi oleh waktu akses, yang pada gilirannya ditentukan oleh latensi rotasi dan waktu pencarian. Performa antarmuka hampir tidak berdampak pada IOPS. Selain itu, kecepatan antarmuka tidak memiliki efek terukur pada kinerja berkelanjutan. Metrik berikut biasanya digunakan untuk membandingkan kinerja dari berbagai HDD:

- Kecepatan transfer berkelanjutan
- Latensi rata-rata
- Kekuatan operasi
- Daya idle
- Cache buffer — ukuran dan jenis
- Mean Time Between Failures (MTBF)

Memilih Hardisk yg tepat selalu bermuara pada kapasitas, kinerja, dan metrik konsumsi daya, tetapi tidak selalu dalam urutan itu. Misalnya, data massal dan beban

kerja arsip membutuhkan kinerja pejalan kaki tetapi kapasitas yang berlebihan. Konsumsi daya dan kapasitas sering menjadi fokus utama dalam segmen ini, dan kinerja jatuh ke tempat ketiga yang jauh.

Meskipun bidang penyimpanan data telah matang, masih banyak teknologi yang muncul — dan mungkin mengganggu — yang harus Anda waspadai. Popularitasnya hard *disk solid state* (SSD) - ini kira-kira enam kali lebih mahal daripada Hardisk mereka; namun, mereka memiliki kecepatan menyalin / menulis file yang jauh lebih cepat daripada Hardisk — dan skema penyimpanan hierarkis yang menggabungkannya mungkin segera menjadi norma karena biaya penyimpanan terus menurun (seperti yang diilustrasikan pada Gambar 5-2) dan I / Kecepatan penyimpanan O tetap datar dalam beberapa tahun terakhir. Untuk beban kerja transaksional, solusi penyimpanan semua SSD kemungkinan memiliki modal dan biaya operasional keseluruhan yang lebih rendah daripada yang dibuat dari 15.000 RPM Hardisk karena pengurangan total slot yang diperlukan untuk mencapai kinerja transaksi yang diberikan. Selain itu, SSD memiliki pengurangan daya yang sangat kecil dibandingkan dengan drive yang berputar untuk sejumlah transaksi tertentu.



Gambar 5-2. Tren biaya hard drive per gigabyte

MEMPERKIRAKAN KAPASITAS PENYIMPANAN

Menghitung Kapasitas Disk yang Dapat Digunakan

Untuk menghitung berapa banyak data yang dapat dipegang oleh sistem Database harus menghitung kapasitas disk yang dapat digunakan per host segmen dan kemudian mengalikannya dengan jumlah host segmen dalam susunan Database. Mulai dengan kapasitas mentah dari disk fisik pada host segmen yang tersedia untuk penyimpanan data (raw_capacity), yaitu:

disk_size * number_of_disks

untuk overhead pemformatan sistem file (kira-kira 10 persen) dan tingkat RAID yang dgunakan. Misalnya, jika menggunakan RAID-10, perhitungannya adalah:

(raw_capacity * 0.9) / 2 = formatted_disk_space

Untuk kinerja yang optimal, jangan sepenuhnya mengisi disk ke kapasitas, tetapi jalankan pada 70% atau lebih rendah. Jadi dengan mengingat hal ini, hitung ruang disk yang dapat digunakan sebagai berikut:

formatted_disk_space * 0.7 = usable_disk_space

Setelah memformat array disk RAID dan memperhitungkan kapasitas maksimum yang direkomendasikan (usable_disk_space), perlu menghitung berapa banyak penyimpanan yang sebenarnya tersedia untuk data pengguna (U). Jika menggunakan Mirror Database untuk redundansi data, ini akan menggandakan ukuran data pengguna Anda (2 * U). Database juga membutuhkan beberapa ruang dicadangkan sebagai area kerja untuk kueri aktif. Ruang kerja harus sekitar sepertiga ukuran data pengguna Anda (ruang kerja = U / 3):

Dengan Mirror: **(2 * U) + U / 3 = usable_disk_space**

Tanpa Mirror: **U + U / 3 = usable_disk_space**

Panduan untuk ruang file sementara dan ruang data pengguna mengasumsikan beban kerja analitik yang khas. Beban kerja atau beban kerja yang sangat bersamaan dengan kueri yang membutuhkan ruang sementara yang sangat besar dapat dimanfaatkan untuk memesan area kerja yang lebih besar. Biasanya, throughput sistem secara keseluruhan dapat ditingkatkan sementara mengurangi penggunaan area kerja melalui manajemen beban kerja yang tepat. Selain itu, ruang sementara dan ruang pengguna dapat diisolasi satu sama lain dengan menentukan bahwa mereka berada di ruang tabel yang berbeda.

Menghitung Ukuran Data Pengguna

Seperti halnya semua basis data, ukuran data mentah Anda akan sedikit lebih besar setelah dimuat ke dalam basis data. Rata-rata, data mentah akan menjadi sekitar 1,4 kali lebih besar pada disk setelah dimuat ke dalam database, tetapi bisa lebih kecil atau lebih besar tergantung pada tipe data yang Anda gunakan, jenis penyimpanan tabel, kompresi dalam database, dan sebagainya.

- Halaman Overhead - Ketika data Anda dimuat ke dalam Database, itu dibagi menjadi halaman 32KB masing-masing. Setiap halaman memiliki 20 byte overhead halaman.
- Baris Overhead - Dalam tabel penyimpanan 'tumpukan' reguler, setiap baris data memiliki 24 byte dari overhead baris. Tabel penyimpanan 'tambahan-dioptimalkan' hanya memiliki 4 byte overhead baris.
- Atribut Overhead - Untuk nilai data itu sendiri, ukuran yang terkait dengan setiap nilai atribut tergantung pada tipe data yang dipilih. Sebagai aturan umum, Anda ingin menggunakan tipe data terkecil yang mungkin untuk menyimpan data Anda (dengan asumsi Anda mengetahui nilai yang mungkin akan dimiliki kolom).
- Indeks - Dalam Database, indeks didistribusikan di seluruh host segmen seperti data tabel. Jenis indeks default dalam Database adalah B-tree. Karena ukuran indeks bergantung pada jumlah nilai unik dalam indeks dan data yang akan disisipkan, perkiraan ukuran indeks yang pasti tidak mungkin dilakukan. Namun, Anda dapat memperkirakan ukuran indeks menggunakan rumus ini.

B-tree: unique_values * (data_type_size + 24 byte)

Bitmap: (unique_values * number_of_rows * 1 bit * compression_ratio / 8) + (unique_values * 32)

Menghitung Kebutuhan Bandwidth

Bandwidth adalah jumlah data yang bisa ditransfer dari dan ke website kita setiap bulannya. Ini termasuk download dan upload via HTTP maupun FTP. Besarnya bandwidth juga tergantung jumlah visitor dan jumlah tampilan halaman (page view) website kita. Langsung saja kita hitung, misalkan rata-rata ukuran halaman adalah 500 KB, pengunjung per hari 1000 orang dan rata-rata page view per visitor adalah 3. Berapa bandwidth yang dibutuhkan selama satu bulan ?

Visitor = 1000 / day = 30k / month
BW = 500 x 3 x 30,000 = 45,000,000 KB atau sekitar 45 GB per Bulan.

Perencana kapasitas sering dihadapkan dengan keputusan yang sulit untuk dibuat, karena banyak pekerjaan mereka membutuhkan keseimbangan antara kinerja aplikasi dan efisiensi infrastruktur. Tanggung jawab perencana kapasitas adalah untuk memahami kapan lebih banyak perangkat keras diperlukan untuk menjamin kinerja aplikasi, sementara pada saat yang sama, menghindari perangkat keras yang terbuang. Pendekatan tradisional melibatkan mencari tahu kelebihan kapasitas saat ini, kemudian mencoba mencocokkannya dengan pertumbuhan di masa depan. Ini biasanya dilakukan terhadap metrik utama seperti CPU dan memori.
Rumus sederhana di bawah ini membantu menggambarkan pendekatan ini:

$$Headroom = \frac{(Total\ Capacity * Desired\ Utilization(\%)) - Used\ Capacity}{Average\ Utilization}$$

Misalnya, jika mencoba memecahkan ruang kepala memori dalam kluster, akan diambil langkah-langkah berikut:

- Cluster memiliki 5 host; setiap host memiliki 128 GB memori. Jadi, total kapasitas = 640 GB.
- Pemanfaatan memori host rata-rata di cluster adalah 80 GB. Jadi, kapasitas yang digunakan = 400 GB.
- Pemanfaatan yang diinginkan dari memori adalah 70%.
- VM rata-rata menggunakan 0,5 GB memori (utilisasi rata-rata = 0,5).
- *((640 * 0,7) - 400) / 0,5* = 96 Mesin Virtual = Tersedia Headroom

Jelas ini adalah formula yang disederhanakan, dan ada ketergantungan lain yang perlu dipertimbangkan, seperti berapa lama perangkat keras mengambil ke rak, kendala lisensi, permintaan musiman, dll. Apapun, tidak peduli seberapa canggih perhitungannya, Anda akhirnya mengerti bahwa kluster dapat menerima sejumlah VM tertentu dari ukuran template tertentu, pada titik waktu tertentu. Data itu kemudian dihitung secara terpisah terhadap pertumbuhan masa depan yang diharapkan dari lingkungan. Ini biasanya dilakukan dalam mode "proses batch" terhadap siklus pembelian tahunan atau kuartalan.
Masalah dengan pendekatan ini adalah bahwa ini adalah perhitungan statis yang membuat asumsi BANYAK, dan tidak berubah seiring dengan perubahan dalam pemanfaatan lingkungan. Sebagaimana dibahas dalam posting sebelumnya yang merinci keterbatasan kapasitas penghitungan dalam ruang hampa pada lembar excel , tidak mungkin untuk menjamin ada sumber daya yang cukup untuk masing-masing VM pada masing-masing host yang menggunakan pendekatan ini. Untuk memahami kapasitas yang tersedia, Anda harus secara cerdas membuat keputusan penempatan dan penentuan ukuran yang akan mempertimbangkan semua sumber daya di lingkungan. Penempatan yang lebih baik menyebabkan kepadatan yang lebih baik yang memungkinkan pembelian perangkat keras lebih sedikit

Lebih lanjut akan dijabarkan pada Lampiran B

BAB ENAM

DESAIN MANAJEMEN KAPASITAS

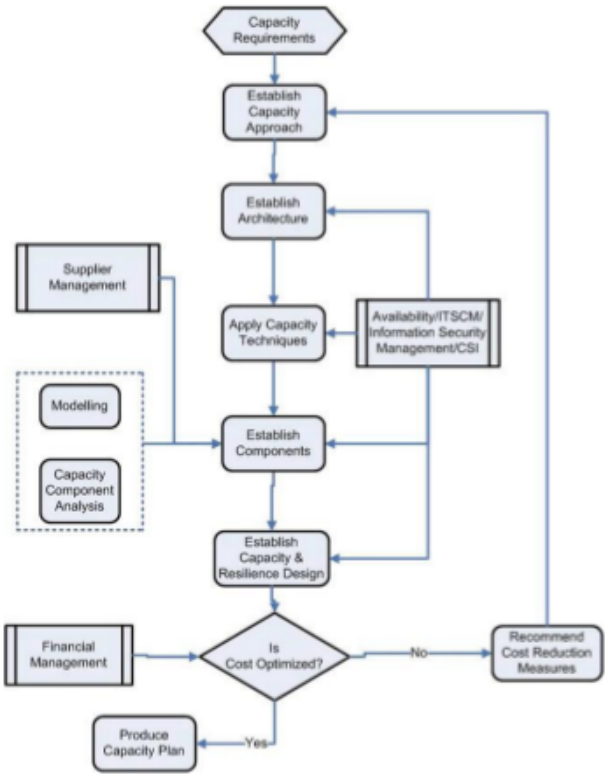
Bagian ini menguraikan tentang bagaimana manajemen kapasitas dirancang sesuai dengan persyaratan kapasitas yang dikumpulkan. Penekanan telah diberikan pada bagaimana memilih pendekatan kapasitas yang tepat dengan memilih infrastruktur mekanisme penempatan beban kerja terbaik, mendefinisikan arsitektur dan komponen manajemen kapasitas, dan langkah-langkah pengoptimalan biaya. Ini memastikan pemanfaatan sumber daya yang maksimal dan kinerja optimal dengan cara yang dapat dibenarkan biaya. Salah satu faktor kunci keberhasilan dalam manajemen kapasitas adalah membangun keseluruhan arsitektur kapasitas. Bagian ini menjelaskan arsitektur dalam berbagai lapisan yang memenuhi persyaratan terkait bisnis dan kinerja dari perspektif manajemen kapasitas. Sangat penting untuk layanan baru apa pun yang mendesain kapasitas untuk juga memperhitungkan tingkat layanan manajemen, persyaratan keamanan, persyaratan ketersediaan,

Desain untuk Kapasitas

Prosedur untuk mendesain kapasitas adalah untuk

- Bentuk pendekatan kapasitas.
- Menetapkan arsitektur.
- Terapkan teknik kapasitas.
- Menetapkan komponen dan memeriksa pengoptimalan biaya.

Prosedur ini untuk manajemen kapasitas (ditunjukkan pada Gambar 6-1) terpicu setelah persyaratan kapasitas dikumpulkan. Setelah persyaratan kapasitas telah dikumpulkan, pendekatan manajemen kapasitas keseluruhan untuk perencanaan kapasitas dan manajemen ditetapkan. Ini mungkin termasuk merancang arsitektur dan model manajemen kapasitas. Teknik kapasitas digunakan untuk menetapkan komponen kapasitas, atau lapisan, seperti data, analisis, dan lapisan presentasi.



Gambar 6-1 . *Kegiatan desain manajemen kapasitas*

Menetapkan Pendekatan Kapasitas

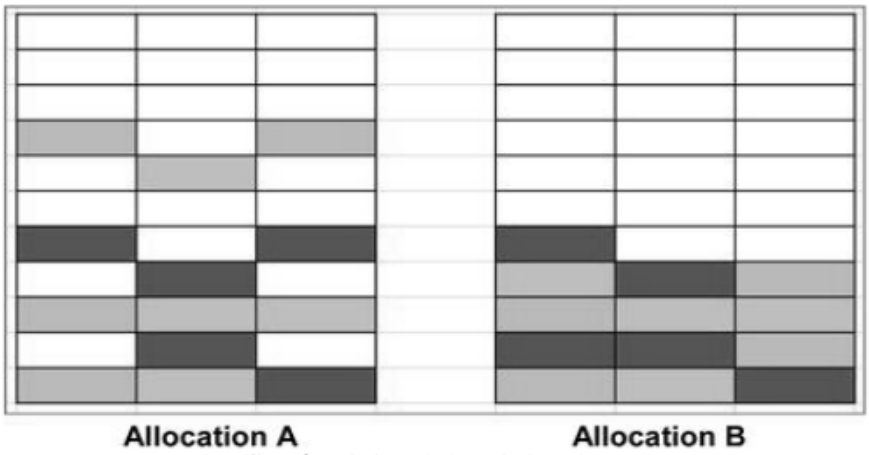
Pendekatan kapasitas harus memastikan pemanfaatan infrastruktur cloud maksimum dengan menyelenggarakan instance klien seefisien mungkin. Dalam lingkungan virtual di mana penyediaan infrastruktur mudah, mungkin ada masalah kurangnya penyediaan dan kelebihan penyediaan. Jadi ada kebutuhan untuk pendekatan kapasitas yang efisien yang dapat menemukan keseimbangan antara keduanya.

Pendekatan kapasitas harus difokuskan terutama pada penempatan beban kerja dan alokasi sumber daya. Kapasitas yang terfragmentasi dapat menyebabkan inefisiensi dan bahkan menggandakan infrastruktur yang diperlukan untuk menampung beban

kerja. Prosedur kapasitas harus ditentukan untuk memastikan alokasi cerdas mesin virtual seperti yang dipersyaratkan oleh aplikasi.

Alat-alat manajemen kapasitas dapat dipertimbangkan untuk memungkinkan penyedia layanan menentukan aturan teknis, bisnis, dan kepatuhan untuk penempatan beban kerja. Aturan-aturan penempatan beban kerja ini dikonfigurasi dalam toolset manajemen siklus hidup cloud di lapisan manajemen. Mesin aturan harus dievaluasi untuk menjamin kesehatan dan keakuratan solusi manajemen kapasitas.

Seperti yang ditunjukkan pada Gambar 6-2 , pendekatan terbaik untuk menempatkan beban kerja atau mesin virtual pada infrastruktur virtual harus digunakan untuk memastikan sumber daya yang ada dikonsumsi secara efisien. Seperti yang ditunjukkan dalam Alokasi A, tambalan yang tidak digunakan berpotensi menandakan kapasitas yang tidak digunakan, dan ini akan menyebabkan ketidakefisienan. Di sisi lain, Alokasi B memastikan penggunaan sumber daya sebaik mungkin.



Gambar 8-2 *. Alokasi beban kerja*

Beban kerja adalah klasifikasi logis dari pekerjaan yang dilakukan dalam infrastruktur virtual. Beban kerja dapat diklasifikasikan oleh siapa yang melakukan pekerjaan, pekerjaan apa yang sedang dikerjakan, dan bagaimana pekerjaan dilakukan. Misalnya, penyedia layanan harus dapat mengklasifikasikan beban kerja sesuai dengan fungsi bisnis seperti penjualan, pemasaran, atau keuangan. Beban

kerja yang relevan dengan bisnis juga berguna ketika tiba waktunya untuk merencanakan masa depan.

Aplikasi bisnis harus dianalisis untuk kebutuhan penggunaan infrastruktur. Bisa ada dua aplikasi bisnis yang sama pentingnya untuk bisnis tetapi mengkonsumsi sumber daya secara berbeda. Manajemen kapasitas harus dapat menetapkan kebutuhan tersebut dan dengan demikian mengelompokkan tingkat layanan yang terkait dengan aplikasi sehingga infrastruktur disediakan dengan tepat (seperti kapasitas disk, kapasitas komputasi, memori, dan kebutuhan bandwidth jaringan). Langkah-langkah ini akan membantu dalam perencanaan kebutuhan sistem masa depan.

Gambar 8-3 menggambarkan bagaimana perencanaan untuk beban kerja sangat penting untuk pemanfaatan sumber daya yang efisien. Teknik seperti estimasi, pemodelan, dan pengujian beban dapat digunakan untuk perencanaan beban kerja yang efisien. Rencana Beban Kerja A, ketika ditempatkan dalam infrastruktur, jelas bukan pola beban kerja terbaik untuk infrastruktur yang tersedia sedangkan Rencana Beban Kerja B sesuai untuk infrastruktur yang tersedia, yang mengarah ke efisiensi dan skema pemanfaatan sumber daya yang tepat. Ada beberapa alat yang tersedia yang berspesialisasi dalam perencanaan beban kerja untuk membantu perencana kapasitas dalam menetapkan pendekatan terbaik untuk kapasitas.

Gambar 8-3 . *Perencanaan beban kerja*

Berdasarkan kinerja, target tingkat layanan, dan fungsi vital bisnis, pendekatan tingkat tinggi untuk memenuhi target dipilih. Misalnya, fokus biaya yang kaku dalam target kinerja akan menyerukan solusi just-in-time, sedangkan kesinambungan layanan pada target kinerja yang kaku mungkin memerlukan solusi

untuk memanfaatkan ketersediaan kapasitas margin. Suatu pendekatan harus diputuskan untuk menyediakan laporan beban kerja real-time dan historis yang dapat digunakan untuk optimasi sumber daya dan diagnosis masalah on-the-fly. Laporan beban kerja real-time harus mengidentifikasi kapasitas server yang kurang dimanfaatkan atau berlebihan, yang dapat digunakan untuk mengoptimalkan distribusi beban kerja di semua perangkat keras yang tersedia, serta untuk membantu mencegah pembelian perangkat keras yang tidak perlu.

Saat menyiapkan pendekatan untuk kapasitas, penyedia layanan harus menilai beban kerja melalui penemuan dan inventarisasi aset TI karena ini adalah penyebab utama di balik kemacetan kapasitas. Ini harus dilakukan fungsi-, lokasi-, dan lingkungan-bijaksana. Algoritme statistik harus digunakan untuk menghitung pertumbuhan beban kerja dan metrik terkait lainnya. Laporan-laporan ini juga dapat membantu untuk mendiagnosis masalah, terutama dalam situasi di mana keterbatasan kapasitas tidak diketahui atau di mana load balancing dalam peternakan server atau kelompok tidak berfungsi sebagaimana mestinya. Mereka mengidentifikasi kemacetan dan membutuhkan kapasitas tambahan untuk mendukung pertumbuhan beban kerja yang diharapkan atau dikehendaki sambil menghormati ambang pemanfaatan sumber daya dan waktu respons. Pendekatan kapasitas juga harus memperhatikan manajemen kapasitas cadangan dan memastikan defragmentasi dan teknik lainnya diterapkan untuk memastikan penempatan beban kerja yang tepat di infrastruktur. Metode alokasi kapasitas proaktif harus dipertimbangkan berdasarkan pemantauan dan analisi peristiwa infrastruktur.

Penting bagi pengguna cloud untuk mensimulasikan beban kerja dalam lingkungan cloud sebelum pergerakan sebenarnya dari aplikasi ke cloud. Di cloud, sumber daya dibagi di antara pelanggan, dan di beberapa lingkungan cloud, arsitektur komponen yang mendasarinya berbeda dari yang ditemukan di lingkungan TI tradisional, jadi penting bahwa konsumen cloud menguji dan menyimulasikan aplikasi sebelum gerakan aktual ke awan.

Aspek lain selama perencanaan kapasitas berkaitan dengan jenis kapasitas untuk membeli. Berikut ini adalah opsi yang tersedia untuk konsumen cloud.

- **Sewa untuk periode tertentu:** Opsi ini memberikan jaminan kapasitas kepada konsumen untuk jangka waktu tertentu. Biayanya lebih kecil

daripada opsi menyewa kapasitas sesuai permintaan karena konsumen mengambil kumpulan kapasitas untuk jangka waktu yang lebih lama.

- **Kapasitas sesuai permintaan** : Opsi ini menyediakan kapasitas per jam; namun, tidak ada jaminan ketersediaan kapasitas baru dalam model ini di pusat data atau lokasi tertentu karena ini didorong oleh pasar.

- **Pasar spot:** Opsi ini memungkinkan pengguna cloud untuk membeli kapasitas yang disediakan di pasar spot. Harga di sini jauh lebih rendah dari kapasitas sesuai permintaan. Opsi ini dapat digunakan dalam kasus di mana aplikasi memanfaatkan jumlah komputasi yang tinggi untuk pemrosesan dan dapat menyimpan hasil untuk pemrosesan selanjutnya. Dalam model ini, kapasitas dapat diambil dari bawah Anda jika seseorang mengalahkan Anda, sehingga hanya cocok untuk skenario yang dapat memproses dan menyimpan hasil sering dan memulai kembali dari tempat mereka tinggalkan, seperti mesin analisis penemuan obat, mesin pemrosesan offline, dll.

Jadi, konsumen cloud harus mengambil pendekatan holistik terhadap kebutuhan kapasitas di cloud dan tidak pergi dengan asumsi bahwa kapasitas tidak terbatas akan tersedia jika diperlukan.
Strategi berlapis dengan kombinasi sewa jangka panjang, kapasitas sesuai permintaan, dan pembelian pasar spot akan memberikan ROI maksimum dan memastikan bahwa aplikasi penting memiliki kapasitas yang cukup pada waktu yang dibutuhkan.
Strategi cloud hibrida juga dapat dimanfaatkan untuk mengurangi risiko di mana masalah kapasitas atau ketersediaan pada satu penyedia cloud tidak berdampak pada aplikasi penting.

Membangun Arsitektur

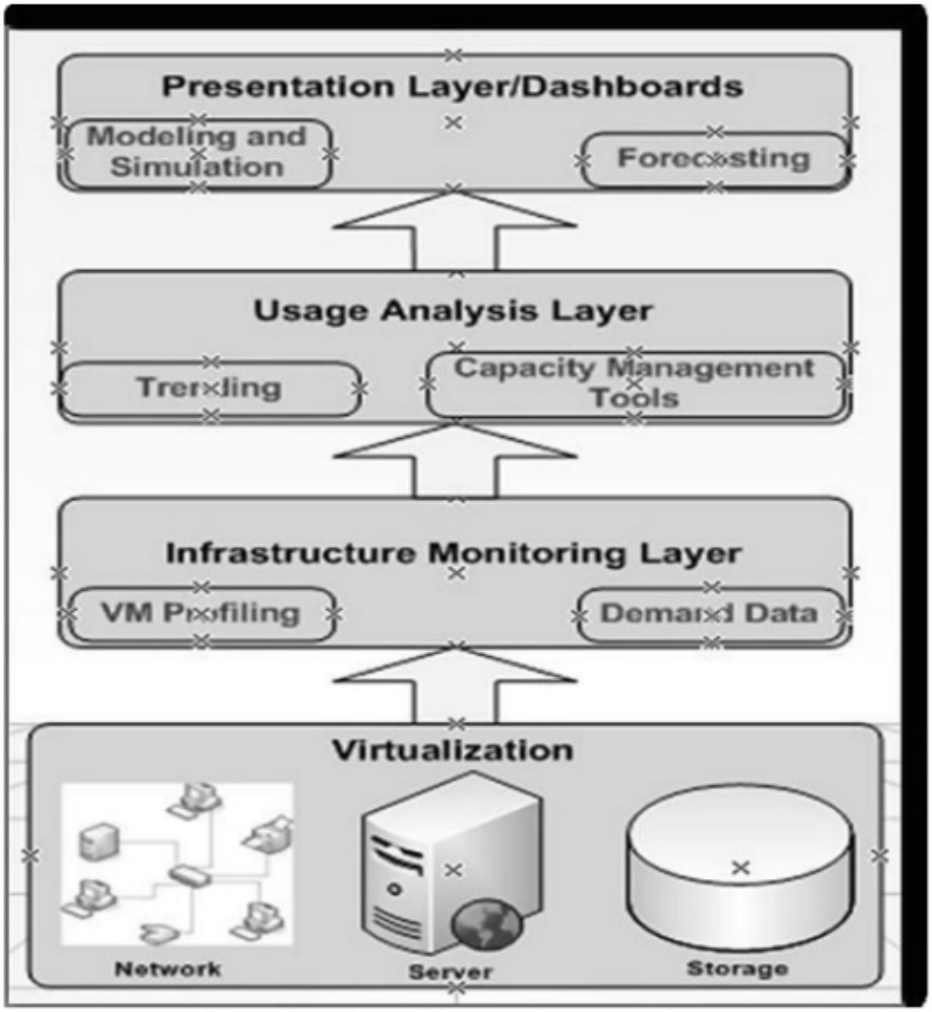
Arsitektur kapasitas memiliki implikasi besar pada pemenuhan target manajemen kapasitas. Di sini, arsitektur yang ada dipertimbangkan untuk layanan yang diubah dan modifikasi atau proposal baru yang diperlukan ditetapkan. Kegiatan

ini dilakukan bersama dengan manajemen tingkat layanan, persyaratan keamanan, dan persyaratan ketersediaan.

Selain arsitektur kapasitas, aplikasi harus dikembangkan untuk mengkonsumsi unit kapasitas serendah mungkin, tidak seperti aplikasi konvensional yang dibangun dengan infrastruktur khusus yang mengarah ke pemanfaatan sumber daya yang buruk.

Skalabilitas aplikasi juga merupakan faktor penting ketika infrastruktur cloud sedang beraksi. Skalabilitas suatu aplikasi juga memungkinkan aplikasi untuk ditingkatkan untuk mengakomodasi pertumbuhan. Aplikasi harus mampu berbicara secara cerdas ke infrastruktur yang mendasarinya ketika kebutuhan akan lebih banyak sumber daya seperti basis data muncul. Arsitektur aplikasi harus didukung oleh kedua jenis metode penskalaan, vertikal dan horizontal. Teknik scaling akan dibahas di bagian selanjutnya.

Arsitektur kapasitas dapat diatur untuk mengakomodasi kebutuhan bisnis yang spesifik. Mari kita lihat pada tampilan arsitektur kapasitas lapisan dasar dengan mendiskusikan Gambar 6-4 .



Gambar 6-4 *. Menetapkan arsitektur*

- Tingkat Virtualisasi
  Di bagian bawahnya terdapat infrastruktur virtual yang mendukung aplikasi dan data bisnis. Tier ini terdiri dari infrastruktur virtual, yang mendukung layanan cloud.

- Tingkat Pemantauan Infrastruktur
  Tingkat pemantauan infrastruktur adalah lapisan yang menyediakan pemantauan melalui alat otomatis untuk tingkat virtualisasi. Tingkat ini mengumpulkan data peristiwa dan kinerja untuk tingkat virtualisasi.

- Tingkat Analisis Penggunaan
  Tingkat analisis penggunaan mengambil data pemantauan sebagai input dan memberikan analisis pada data tersebut. Lapisan ini menjalankan analisis statistik untuk mengubah data pemantauan menjadi peta panas kapasitas, perkiraan kapasitas, dan laporan analitik lainnya. Jadi, tier ini menyediakan data analisis yang diperlukan untuk penyedia cloud untuk membuat keputusan tentang kapasitas.

- Tingkat Presentasi / Dasbor
  Tingkat presentasi / dasbor memberikan pandangan untuk data analitis yang dihasilkan dalam tingkat analisis penggunaan. Lapisan ini memungkinkan manajer kapasitas untuk menggunakan teknik peramalan untuk menghasilkan analisis prediktif. Lapisan ini juga memungkinkan manajer kapasitas untuk melakukan pemodelan dan simulasi untuk memahami dampak dari setiap perubahan pada kapasitas.

Saat ini ada banyak alat yang berada pada lapisan yang berbeda dan mampu mengumpulkan, menghubungkan, dan menyajikan data kapasitas untuk membantu mengidentifikasi sebagian besar sumber daya yang kurang dimanfaatkan dan terlalu banyak bekerja. Selain itu, alat-alat ini, pada lapisan peramalan dan penggunaan analisis, bahkan mampu menentukan faktor-faktor yang mendorong beban kerja. Alat belajar mandiri ini secara otomatis membuat keputusan tentang penentuan garis dasar otomatis, pengaturan ambang batas, pengaturan alarm, dll., Dan memungkinkan pengguna untuk memanfaatkan sumber daya sebaik mungkin dan meningkatkan produktivitas staf TI.
Arsitektur kapasitas harus memanfaatkan kemampuan belajar mandiri, pengaturan ambang otomatis, visibilitas top-down, dan perencanaan kapasitas proaktif ketika merancang solusi manajemen kapasitas. Merancang untuk kapasitas juga menentukan integrasi dan aliran data antara toolsset di setiap tier.

Arsitektur kapasitas harus secara jelas menentukan bagaimana komponen yang berbeda berinteraksi dengan lapisan manajemen kapasitas. Ada berbagai sumber pengumpulan data seperti alat pemantauan peristiwa, layanan dan alat manajemen konfigurasi, alat migrasi, alat kinerja aplikasi, manajemen aset, dan alat penemuan. Alat-alat ini harus dapat diintegrasikan dengan alat-alat manajemen kapasitas untuk analisis, pelaporan, peramalan, rencana kapasitas generasi, dan seterusnya.

Tingkatan Penyimpanan

Sistem penyimpanan berjenjang menawarkan berbagai jenis kapasitas penyimpanan, mulai dari drive SATA (Serial ATA) berkecepatan rendah hingga hard disk solid state high-end. Perangkat lunak ini dapat secara otomatis memindahkan data dari satu tier ke yang lain berdasarkan aturan yang ditetapkan.Biasanya data untuk penggunaan waktu nyata yang sering diakses ditempatkan pada penyimpanan akhir yang lebih tinggi, yang lebih cepat tetapi lebih mahal, dan data yang jarang diakses dipindahkan ke disk yang lebih lambat dan murah.Seiring dengan ini, de-duplikasi data dapat mengurangi data aktual dengan menghapus beberapa salinan dari data yang sama dan merebut kembali kapasitas penyimpanan. Kompresi data dapat lebih mengurangi keseluruhan kapasitas untuk kapasitas. Teknologi ini, bila digunakan secara efektif, dapat menurunkan biaya penyimpanan.

Menerapkan Teknik Kapasitas

Teknik seperti analisis dampak kegagalan komponen, pemetaan ketergantungan aplikasi, dan manajemen risiko digunakan untuk mengoptimalkan desain kapasitas. Penyedia layanan cloud membutuhkan pemahaman yang baik tentang infrastruktur kapasitasnya sendiri dan kemampuan untuk memprediksi kapasitas dan fluktuasi besar dalam beban permintaan. Awalnya, prediksi mungkin didasarkan pada perkiraan kasar "aturan praktis" dan tren linier, tetapi pada akhirnya mereka harus didasarkan pada metode ilmiah dan penerapan teknologi pemodelan prediktif yang didukung oleh algoritma antrian teori yang terbukti.Harus ada simulasi konsolidasi atau kegiatan virtualisasi yang dihasilkan dengan mengidentifikasi kandidat dan target terbaik dan penempatan yang optimal

(misalnya, sesuai dengan beban kerja yang kompatibel) berkenaan dengan kriteria teknis, geografis, bisnis, dan kepatuhan. Simulasi harus dapat menggambarkan bagaimana skala layanan dari lingkungan pengujian ke lingkungan tingkat produksi dengan menggunakan hasil pengujian beban. Simulasi perubahan infrastruktur (misalnya, skala horizontal atau vertikal atau kegagalan) dan skenario perubahan bisnis harus dilakukan (misalnya, tren bisnis dan rencana pemasaran).

Menetapkan Komponen dan Memeriksa Pengoptimalan Biaya

Atas dasar kegiatan desain sebelumnya dilakukan, pertimbangan seperti penggunaan klaster, dll ditetapkan. Spesifikasi tingkat komponen dalam hal spesifikasi mesin virtual dan konfigurasi dilakukan sejalan dengan target kinerja layanan. Database kapasitas harus diintegrasikan dengan database infrastruktur TI pusat (disebut database manajemen konfigurasi, atau CMDB) untuk membangun hubungan infrastruktur komponen dengan layanan dalam tindakan.Pemeriksaan akhir dalam hubungannya dengan perencanaan keuangan dilakukan untuk mencari langkah-langkah pengurangan biaya yang tersedia tanpa mengorbankan target kinerja. Jika ada langkah-langkah untuk optimalisasi biaya yang diidentifikasi, mereka dimasukkan. Dengan menggunakan pendekatan "virtualisasi kecuali jika sebaliknya", aplikasi baru dan yang diperbarui harus dinilai untuk menghosting pada tingkat yang tervirtualisasi. Pembaruan dapat mencakup kebutuhan untuk tingkat kinerja dan kapasitas baru. Aplikasi legacy pada perangkat keras akhir-hidup juga harus dievaluasi untuk migrasi ke tingkat virtual. Penyedia layanan harus membandingkan opsi perangkat keras yang berbeda sehubungan dengan tolok ukur standar dan kustom dan kriteria lain seperti biaya keseluruhan atau kepatuhan dengan, misalnya, pedoman Go Green. Komponen untuk menyediakan kapasitas dapat ditetapkan sesuai kebutuhan aplikasi: kekritisan, pertumbuhan, dan kapasitas. Kebutuhan terkait kapasitas dapat diformulasikan dalam hal kebutuhan komputasi. Kekritisan dapat diungkapkan, dan tingkat toleransi kesalahan dan pertumbuhan dapat dinyatakan sebagai kebutuhan masa depan. Berdasarkan kapasitas aplikasi, penyimpanan pertumbuhan dan kekritisan, persyaratan jaringan dan server diformulasikan. Ini digabungkan dengan persyaratan fasilitas.Semua komponen ini dirangkum dalam istilah keuangan, dan penyedia layanan dapat menyediakan model layanan yang berbeda sesuai biaya terkait, yang mungkin adalah emas, perak, dan perunggu. Model biaya emas mungkin termasuk Capex yang lebih tinggi tetapi dapat

memastikan kinerja dan keandalan yang tinggi. Beban kerja aplikasi mission-critical atau aplikasi pencarian kapasitas tinggi dapat ditempatkan dalam model biaya emas. Demikian pula beban kerja seperti server produksi dapat ditempatkan dalam model biaya perak, dan pengujian, pengembangan, penyebaran murah dan cepat mungkin sesuai dengan model biaya perunggu. Terserah kebutuhan bisnis TI sejalan dengan perencanaan keuangan untuk memilih tingkat kapasitas infrastruktur yang dicari untuk mengembangkan bisnis. Emas, perak,Untuk menilai dampak dari beban kerja baru pada kapasitas, penilaian persyaratan yang hati-hati diperlukan. Penilaian infrastruktur dapat dipraktekkan untuk mendapatkan pandangan infrastruktur pada tingkat kemampuan luas industri. Karena cloud menawarkan berbagai opsi untuk membeli kapasitas, pengoptimalan biaya harus mempertimbangkan aspek seperti

- Berapa banyak kapasitas untuk membeli di muka untuk jangka waktu yang lebih lama karena ini memberikan harga diskon.
- Berapa banyak kapasitas on-demand yang akan digunakan.
- Cara memanfaatkan instance tempat untuk menurunkan biaya.
- Opsi untuk SaaS dan manfaat keseluruhan, dan pada titik waktu mana opsi SaaS akan menjadi lebih murah atau lebih mahal untuk dijalankan dibandingkan dengan berjalan di cloud pribadi atau IaaS.

Akibatnya, konsumen kini memiliki tugas yang lebih kompleks di tangan karena berbagai model cloud, vendor, dan berbagai opsi harga yang tersedia.

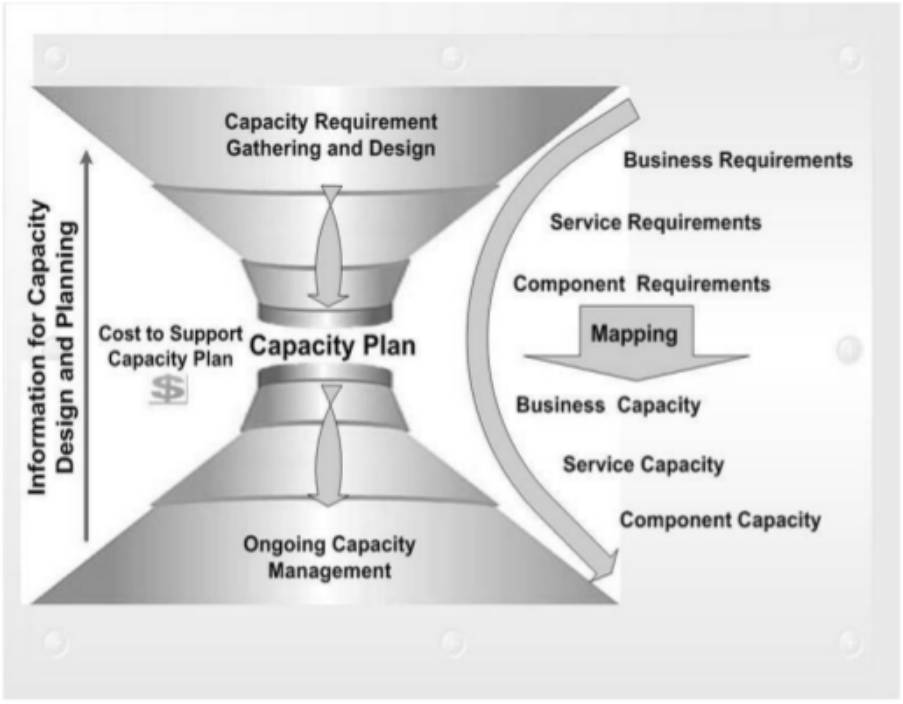
RENCANA KAPASITAS

Setelah persyaratan kapasitas dikumpulkan dan diselesaikan, desain untuk kapasitas dikembangkan, yang terutama mencakup arsitektur kapasitas. Bab ini menjelaskan bagaimana rencana kapasitas dibentuk berdasarkan desain kapasitas. Rencana kapasitas membentuk dokumen referensi yang diedarkan di antara semua pemangku kepentingan yang terlibat dalam penerapan prosedur manajemen kapasitas. Persyaratan kapasitas di semua tingkatan (bisnis, layanan, dan komponen) didokumentasikan dalam rencana kapasitas.Langkah-langkah yang harus dipertimbangkan untuk perencanaan kapasitas yang efektif dalam lingkungan virtual / cloud dibahas dalam bab ini. Setelah itu, bagaimana rencana kapasitas harus dibentuk dan bahan apa yang diperlukan untuk membuat rencana kapasitas

dijelaskan.Ketika meluncurkan layanan baru di cloud, rencana kapasitas adalah salah satu artefak utama yang memastikan bahwa manfaat dari model komputasi cloud dihasilkan secara efektif. Setelah rencana kapasitas diproduksi, prosedur manajemen kapasitass dilaksanakan sesuai dengan rencana. Rencana kapasitas adalah hasil dari prosedur yang dibahas dalam bab-bab sebelumnya, seperti pengumpulan kebutuhan kapasitas dan pembuatan desain kapasitas. Rencana ini diterapkan untuk menjaga layanan tetap hidup dengan cara yang paling optimal dan biayanya dapat dibenarkan. Jika suatu layanan sudah hidup, rencana kapasitas merupakan metrik kinerja dan pengoptimalan terkait dan informasi yang menjaga persyaratan kapasitas dioptimalkan dan disetel.

Menghasilkan Rencana Kapasitas

Gambar 6-5 menguraikan input rencana kapasitas dan bagaimana rencana kapasitas menyediakan keluaran untuk manajemen kapasitas berkelanjutan. Bab ini akan mencantumkan berbagai kegiatan yang diperlukan untuk menghasilkan rencana kapasitas. Pembaca dapat menemukan rencana kapasitas sampel dalam Apendiks buku ini.



Gambar 6-5 *. Rencana kapasitas*

Manajemen kapasitas dimulai dengan menerjemahkan aktivitas bisnis ke dalam persyaratan layanan untuk organisasi, dan kemudian bekerja melalui persyaratan komponen yang diperlukan untuk memberikan tingkat layanan yang ditentukan.
Untuk perencanaan kapasitas yang efektif, kebutuhan bisnis pertama kali diterjemahkan ke dalam aplikasikebutuhan dan kemudian persyaratan kinerja dan kapasitas yang terkait aplikasi ini diformulasikan ke dalam kebutuhan kapasitas infrastruktur. Persyaratan kapasitas ini dilaksanakan melalui desain dan metode kapasitass. Tingkat atau paket layanan yang berbeda dapat ditentukan oleh penyedia layanan untuk menyediakan layanan dan tingkat kinerja yang diinginkan. Sebagai contoh, tingkatan layanan Emas, Perunggu, dan Perak akan memberikan tingkat kapasitas yang berlebihan dan dioptimalkan untuk mendukung persyaratan bisnis tertentu. Keputusan ini dibuat menggunakan tingkat layanan dan rincian keuangan yang diperlukan untuk mendukung tingkat layanan yang diperlukan dengan memastikan bahwa kapasitas yang tepat diberikan pada waktu yang tepat dengan cara yang paling hemat biaya.

Setelah persyaratan kapasitas, desain, dan arsitektur tersedia, rencana kapasitas harus diproduksi oleh manajer infrastruktur menggunakan informasi yang jelas, ringkas, relevan untuk memungkinkan mereka mengalokasikan sumber daya dan modal di mana akan memberikan manfaat bisnis yang paling banyak.
Mari pertama-tama menguraikan langkah-langkah yang harus dipertimbangkan untuk perencanaan kapasitas yang efektif di lingkungan cloud.

- **Mengidentifikasi Profil Mesin Virtual**
  Ada berbagai cara pengelompokan atau standardisasi profil, yang disediakan sebagai template untuk penyediaan mesin virtual. Profil ini akan terdiri dari sumber daya komputasi yang disediakan pada awalnya untuk mesin virtual bersama dengan hal-hal lain seperti versi sistem operasi, aplikasi, atau paket yang sudah diinstal sebelumnya pada template. Penting untuk memasukkan aspek perencanaan kapasitas dalam template. Sebagai contoh, ada persyaratan minimum untuk tipe tertentu dari OS dan basis data, jadi berdasarkan penggunaan aplikasi, kapasitas template itu dapat diputuskan. Contoh berbeda dari template yang sama dapat meningkatkan kapasitas untuk mengakomodasi skenario penggunaan yang skala.Pada akhir latihan ini kita akan memiliki daftar profil dengan semua rincian yang diperlukan tersedia.

- **Mengidentifikasi Profil Server Host**
  Server host di mana mesin virtual ditempatkan dapat dikelompokkan berdasarkan atribut penting seperti kapasitas komputasi yang tersedia dan penggunaan yang dimaksudkan. Sebagai contoh, host dapat dibedakan berdasarkan jenis layanan yang mereka tawarkan, seperti server basis data, server web, atau server middleware.Server juga dapat dikonfigurasi untuk berada di kolam yang melayani tujuan tertentu dan memiliki keamanan dan firewall yang dikonfigurasi dengan sesuai. Sebagai contoh, Anda dapat membuat kumpulan server web dan kumpulan server basis data lainnya dan memiliki dua kolam yang dipisahkan oleh firewall. Konfigurasi jaringan dari server web memungkinkan mereka mengakses dari luar jaringan internal, sementara konfigurasi jaringan dari server basis data membuatnya hanya dapat diakses secara internal.

- **Mengidentifikasi Penyimpanan**
  Penyimpanan profil termasuk jenis disk penyimpanan berdasarkan kinerja dan sistem penyimpanan berdasarkan konektivitas ke kolam host, seperti yang dijelaskan di atas. Dengan demikian, penyimpanan terhubung ke sekelompok host berdasarkan persyaratan beban kerja yang ditempatkan pada host tersebut. Sebagai contoh, penyimpanan pada beban kerja database mungkin memiliki kinerja yang lebih tinggi dan dapat memanfaatkan disk yang lebih cepat untuk memberikan kinerja yang lebih baik, sedangkan untuk server web mungkin tidak memiliki tingkat kinerja yang sama. Profil penyimpanan akan menyertakan Konfigurasi RAID, Konfigurasi LUN, konfigurasi jaringan, dan jenis penyimpanan seperti SATA, SAS, dan Fiber Optics.

- **Mengembangkan Ambang Batas dan Peringatan**
  Ambang dan waspada definisiIni dibuat untuk berbagai komponen dan layanan termasuk jaringan, penyimpanan, CPU, dan memori. Ini melibatkan mendefinisikan berbagai metrik kinerja dan ambang batas untuk metrik ini. Ambang batas mungkin berlipat ganda dan dapat memicu berbagai status, seperti peringatan atau kritis . Katakanlah CPU host digunakan hingga 90 persen selama satu jam terakhir. Ini dapat memicu peringatan peringatan sehingga tindakan yang tepat dapat diambil oleh administrator untuk menghindari penurunan kinerja.

- **Pengisian Sumber Daya**
  Dalam lingkungan komputasi awan yang sangat dinamis, pengisian sumber dayaadalah kegiatan yang sedang berlangsung dan tantangan besar. Karena pola penggunaan dan kegiatan bisnis mengumpulkan banyak pelanggan, itu menjadi tantangan bagi penyedia cloud untuk menentukan bagaimana pengisian sumber daya akan terjadi. Tidak memiliki sumber daya yang cukup untuk memenuhi permintaan pelanggan akan mengakibatkan hilangnya pendapatan dan profitabilitas ditambah kehilangan pelanggan karena mereka dapat pergi ke tempat lain untuk mendapatkan kapasitas. Dengan demikian, penyedia cloud biasanya memiliki kapasitas ekstra untuk menyediakan fitur elastisitas dan komputasi sesuai permintaan kepada

pelanggannya.Penyedia cloud harus memiliki sistem yang sangat otomatis dan integrasi back-end dengan penyedia untuk menyediakan kapasitas secara otomatis dan untuk mendapatkan kapasitas melalui pemesanan berdasarkan analisis statistik dan otomatis dari data pemanfaatan.

Manajemen Permintaan dan Peramalan

Seperti yang dijelaskan di atas, penting bagi penyedia cloud untuk memiliki sistem yang sangat otomatis dan kaya untuk melakukan manajemen dan peramalan permintaan. Algoritma dan sistem rumit dikembangkan untuk menyediakan manajemen permintaan dan peramalan permintaan. Analisis berikut diperlukan untuk penyedia cloud untuk mengantisipasi dan memperkirakan permintaan:

- Permintaan selama waktu hari
- Permintaan selama hari dalam seminggu
- Permintaan selama hari dalam sebulan
- Variasi musiman
- Hari-hari khusus (libur nasional)
- Permintaan dari zona waktu yang berbeda
- Permintaan untuk lokasi / pusat data yang berbeda
- Kumpulan sumber daya

Sumber daya komputasi dan penyimpanan dikelompokkan di bawah kumpulan sumber daya sehingga alokasi mesin virtual yang membutuhkan sumber daya yang dicirikan oleh kumpulan sumber daya ini dapat terjadi pada kumpulan sumber daya ini. Ini pada dasarnya adalah pengelompokan sumber daya awan berdasarkan atribut tertentu seperti jenis beban kerja. Sebagai contoh, kumpulan sumber daya dapat dibuat untuk lingkungan pengembangan yang memiliki host, penyimpanan, dll. Dialokasikan, dan semua mesin pengembangan akan dialokasikan sumber daya dari kumpulan ini.

Membuat Rencana Kapasitas

Sekarang mari kita bahas pendekatan untuk membuat rencana kapasitas. Prosedur atau kegiatan ini dapat bervariasi tergantung pada nomenklatur dan kebijakan

perusahaan tetapi kegiatan ini dapat dirujuk kapan pun rencana kapasitas disiapkan. Prosedur untuk menghasilkan rencana kapasitas adalah sebagai berikut:

- Dokumen Persyaratan Kapasitas.
  Penyedia layanan cloud harus menetapkan peran dan tanggung jawab untuk memfasilitasi pengumpulan informasi untuk perhitungan kapasitas, dan informasi ini didokumentasikan dalam template standar yang disebut rencana kapasitas . Rencana ini harus dikelola dan dipelihara melalui pembaruan rutin, validasi informasi, pemeriksaan kontrol dokumen, dan manajemen versi.
  Persyaratan kapasitas diterjemahkan dari bisnis ke layanan ke tingkat komponen, dan terkait SLAS seperti target kinerja bisnis, layanan, dan parameter komponen didokumentasikan dalam rencana kapasitas. SLA bisnis seperti ketersediaan layanan, kinerja, dan pemantauan jelas disebutkan dalam rencana kapasitas.
  Layanan SLA harus diterjemahkan dari bisnis ke layanan. Misalnya, menyediakan SLA bisnis ketersediaan layanan harus didukung oleh SLA dukungan layanan seperti waktu respons dan waktu resolusi. Komponen SLA dapat mencakup OLA dengan pihak ketiga. Multi-vendor SLA dan OLA dibahas dalam bab-bab sebelumnya.
  Perencana kapasitas perlu secara jelas menetapkan pemetaan ketergantungan antara persyaratan bisnis seperti ketersediaan layanan atau kinerja dan persyaratan dukungan layanan. Demikian pula, pemetaan antara persyaratan layanan dan konsumsi sumber daya harus dihitung dan didokumentasikan. Profil pengguna dan matriks tugas harus dihitung dan didokumentasikan. Peramalan berdasarkan tren konsumsi layanan dilakukan dan didokumentasikan dalam rencana kapasitas untuk menangani lonjakan permintaan di masa mendatang. Penyedia layanan cloud dapat menggunakan algoritme skala untuk memahami persyaratan bisnis guna menyediakan kapasitas optimal untuk opsi naik / turun skala, yang biaya dapat dibenarkan dan mendukung tingkat layanan yang diinginkan. Pemodelan kapasitas dapat digunakan untuk meramalkan perilaku infrastruktur menggunakan data permintaan, keuangan, virtual, operasional, perangkat lunak, dan vendor.

Hal-hal yang perlu diperhatikan saat mengumpulkan persyaratan kapasitas:

- Siapa pelanggannya?
- Nilai apa yang dikirimkan?
- Fitur apa yang dicari pelanggan dengan implikasi pada manajemen kapasitas, seperti skalabilitas, cloud bursting, mobilitas aplikasi, ketersediaan tinggi, toleransi kesalahan, beban kerja produksi, dan SLA yang disepakati?

Memahami kompetensi kapasitas internal dan klien target Anda akan membantu Anda memprediksi kebutuhan kapasitas pelanggan secara lebih akurat. Melakukan ukuran kapasitas ideal masih merupakan tantangan, tetapi penyedia layanan cloud harus mengadopsi penggunaan praktik terbaik, model, dan perangkat lunak. Probabilitas keberhasilan untuk manajemen kapasitas dalam layanan cloud akan dimaksimalkan. Memiliki alat yang memberikan visibilitas dan kontrol ke dalam cloud dan memungkinkan penyedia cloud untuk mengukur dan melaporkan penggunaan kapasitas yang sebenarnya per pelanggan akan mengatur kelas layanan cloud terpisah dan memastikan keuntungan maksimal.

Ada alat perencanaan kapasitas untuk memetakan ketergantungan aplikasi pada infrastruktur, untuk menggambarkan kekritisan setiap aplikasi dan ambang batas SLA dan mengatur aplikasi sesuai tingkat layanan. Penyedia cloud harus sangat logis ketika memetakan tingkat layanan dalam infrastruktur yang ada. Konsep yang muncul seperti belajar mandiri, manajemen hasil, dan optimalisasi kapasitasmembantu penyedia cloud untuk menetapkan kebutuhan kapasitas yang dioptimalkan. Jika tingkatan layanan yang berbeda bergantung pada infrastruktur yang sama, penyedia layanan dapat mempertimbangkan restrukturisasi untuk menyelaraskan kapasitas dengan kekritisan setelah kebutuhan bisnis diidentifikasi dan didokumentasikan. Persyaratan manajemen kapasitas ini didukung oleh teknik dan peralatan manajemen kapasitas. Lihat bagian selanjutnya untuk teknik implementasi.

Pada tingkat komponen, rencana kapasitas harus mencakup informasi tentang CPU, memori, input atau output, jaringan, penyimpanan, dan kebutuhan penggunaan. Penyedia layanan harus memastikan bahwa

informasi SLA-, aplikasi-, infrastruktur-, dan informasi terkait vendor dikumpulkan dan didokumentasikan serta berkontribusi pada keseluruhan rencana kapasitas. Selain itu analisis beban kerja, analisis pemanfaatan, analisis waktu tanggapan, dll harus diperhitungkan.

Penting untuk memastikan bahwa perencanaan kapasitas dilakukan untuk mengatur permintaan tingkat komponen dengan menggunakan faktor pengimbang beban puncak ke rata-rata untuk mengidentifikasi mereka CI (item konfigurasi) yang mungkin menjadi kandidat untuk menyeimbangkan. CI dengan penggunaan sangat tinggi adalah kandidat yang baik untuk keseimbangan permintaan. Membandingkan beban rata-rata CI dengan beban jam puncaknya mengidentifikasi kandidat tersebut dan membuat metrik untuk keseimbangan beban puncak-ke-rata-rata dan memastikan bahwa ini akan dilacak sebagai bagian dari perencanaan kapasitas.

- Desain dan metode dokumen yang digunakan.

  Metode dan teknikdigunakan untuk tiba di desain untuk mendukung persyaratan kapasitas harus disebutkan dalam rencana kapasitas. Desain kapasitas ini harus didasarkan pada prinsip-prinsip dasar arsitektur data center dan pembenaran biaya. Ini termasuk teknik dan metodologi yang mendukung kinerja, skalabilitas, dan ketersediaan di lingkungan cloud seperti load balancing, pengelompokan, dan alokasi sumber daya. Profil populasi mesin virtual harus dirancang untuk mendukung aplikasi bisnis yang berjalan secara optimal.

  Aspek pelaporan harus mencakup mesin virtual berukuran besar dan di bawah ukuran. Mesin virtual berukuran lebih besar menggunakan kapasitas yang lebih sedikit dibandingkan dengan apa yang telah dialokasikan ke mesin virtual. Mesin yang berukuran di bawahnya kekurangan kapasitas komputasi dan menghadapi kendala kapasitas yang dapat menghambat kinerja aplikasi dari aplikasi yang berjalan pada mesin virtual tersebut. Pelaporan kondisi ini akan menghasilkan penyediaan kapasitas yang paling dibutuhkan. Teknik untuk implementasi desain kapasitas akan dibahas secara rinci di bagian manajemen kapasitas yang sedang berlangsung.

Desain akhirnya menyimpulkan, termasuk arsitektur, ukuran desain, dan rincian kapasitas, juga harus disebutkan. Pemetaan tingkat bisnis-ke-layanan-ke-komponen harus dilakukan dan dengan SLA kinerja yang diperlukan di semua tingkatan. Sebagaimana dibahas dalam bagian sebelumnya, SLA multi-vendor lingkungan, OLA, dan UCs harus didefinisikan dengan baik, dan desain manajemen kapasitas dan metode harus dapat membenarkan hal yang sama.

Pembenaran untuk bagaimana desain yang diusulkan akan digunakan mendukung persyaratan bisnis dan pengguna dalam hal SLA, dengan pemetaan rinci dan perhitungan pendukung yang didokumentasikan dalam rencana kapasitas. OLA pihak ketiga dan UC harus dipertimbangkan ketika merancang arsitektur manajemen kapasitas sehingga perhitungan kapasitas untuk mendukung SLA adalah

- Hasilkan keseluruhan rencana.

Setelah semua bahan dari rencana kapasitas siap dan diputuskan, kegiatan pengumpulan informasi dilakukan untuk menyusun informasi dari berbagai pemegang saham. Sebelum tahap pengumpulan data dari rencana manajemen kapasitas, harus memiliki template yang terorganisir atau format lain untuk mengumpulkan data (termasuk bisnis / layanan / SLA komponen, aplikasi, infrastruktur, informasi terkait vendor, tingkat layanan, dan sebagainya). Ini harus disebutkan dengan jelas dalam rencana kapasitas.
Bagian penting dari rencana kapasitas mencakup hal-hal berikut:

- **SLA:** Bagian ini mencakup tingkat layanan yang disepakati yang dilakukan penyedia cloud kepada pelanggan.
- **Aplikasi:** Ini harus menyertakan data mengenai masing-masing aplikasi organisasi. Juga harus dipastikan bahwa semua aplikasi telah diperhitungkan dan bahwa informasi mengenai setiap aplikasi sudah benar.
- **Infrastruktur Saat Ini:** Bagian ini mencakup data mengenai aset virtual dan fisik pusat data. Ini juga harus memastikan bahwa semua perangkat telah diperhitungkan dan bahwa informasi mengenai setiap perangkat sudah benar.

- **Skenario Tugas Pengguna:** Skenario penggunaan(Juga disebut use cases) menentukan urutan tugas yang dilakukan pengguna dan interaksi mereka dengan fungsi solusi untuk membantu mereka melakukan tugas-tugas ini. Bagian dari rencana kapasitas ini harus menentukan skenario yang dikerjakan oleh pengguna di masing-masing area fungsional. Mengidentifikasi dan menjelaskan skenario penggunaan memberikan detail yang memungkinkan estimasi beban kapasitas dan faktor lainnya.
- **Matriks Beban Tugas:** Matriks tugas-beban menjelaskan berbagai jenis beban yang digunakan setiap skenario penggunaan pada sistem.
- **Pemantauan dan Metrik:** Pemantauan dan metrikBagian ini menjelaskan metode pemantauan, teknik, dan alat yang akan digunakan untuk mengevaluasi solusi dan kinerja komponen serta menyediakan metrik untuk merencanakan intervensi. Informasi ini harus disediakan untuk setiap komponen utama pada tingkat solusi.
- **Prakiraan:** Rencana bisnis harus menyediakan manajemen kapasitas dengan rincian layanan baru yang direncanakan dan pertumbuhan atau kontraksi dalam penggunaan layanan yang ada. Subbagian ini harus melaporkan layanan baru dan matinya sistem warisan.
- **Tingkat Layanan:** Bertanggung jawab untuk membuat paket layananatau tingkat layanan yang didasarkan pada sekumpulan kriteria aplikasi-kritis yang telah ditetapkan sebelumnya, seperti tingkatan layanan Gold untuk aplikasi yang paling penting.
- **Informasi Pemasok:** Bertanggung jawab untuk mengumpulkan informasi kontak dari vendor dan kontak internal sehingga mereka tersedia selama krisis kapasitas.

Setelah rencana kapasitas disusun, itu diedarkan di antara semua pemangku kepentingan dalam format yang tidak ambigu dan dapat diakses yang menyajikan komponen, layanan, dan pandangan bisnis untuk implementasi dalam manajemen kapasitas yang sedang berlangsung. Setelah rencana kapasitas dirumuskan, toolset manajemen kapasitas dikerahkan dan semua parameter yang berhubungan dengan kapasitas seperti ambang batas, peringatan, pemberitahuan, dan tingkat layanan dikonfigurasikan.

Pada tahap ini, proses manajemen kapasitas mengambil giliran baru dan fokus bergerak dari perencanaan kapasitas layanan baru ke manajemen kapasitas berkelanjutan.

INDEKS

CAPEX, ***Capital Expenditure*** atau ***Capital Expense*** adalah alokasi yang direncanakan (dalam anggaran) untuk melakukan pembelian/perbaikan/penggantian segala sesuatu yang dikategorikan sebagai aset perusahaan.

CPU adalah (***Central Processing Unit/Processor; CPU***), merujuk kepada perangkat keras komputer yang memahami dan melaksanakan perintah dan data dari perangkat lunak. Istilah lain, pemroses/prosesor (processor), sering digunakan untuk menyebut CPU**.**

CSI*, Customer Satisfaction Index* adalah tingkatan kepuasan pengguna atas suatu layanan.

CSV, ***Comma Separated Values*** adalah suatu format data dalam basis data di mana setiap record dipisahkan dengan tanda koma (,) atau titik koma (;). Selain sederhana, format ini dapat dibuka dengan berbagai text-editor seperti Notepad, Wordpad, bahkan MS Excel.

Daemon adalah proses yang didesain agar proses tidak mendapatkan intervensi dari user. respon terhadap request tersebut, berdasarkan tipe dari request.

Database (Basisdata) adalah kumpulan data yang disimpan secara sistematis di dalam komputer yang dapat diolah atau dimanipulasi menggunakan perangkat lunak (program aplikasi) untuk menghasilkan informasi.

DDoS, *Distributed Denial-of-Service attacks* adalah jenis serangan terhadap sebuah komputer atau server di dalam jaringan internet dengan cara menghabiskan sumber (resource) yang dimiliki oleh komputer tersebut sampai komputer tersebut tidak dapat menjalankan fungsinya dengan benar sehingga secara tidak langsung mencegah pengguna lain untuk memperoleh akses layanan dari komputer yang diserang tersebut.

DR, ***Disaster Recovery,*** Disaste*r* (bencana) didefinisikan sebagai kejadian yang waktu terjadinya tidak dapat diprediksi dan bersifat sangat merusak. Pengertian ini mengidentifikasikan sebuah kejadian yang tiba-tiba, tidak diharapkan, bersifat sangat merusak, dan kurang perencanaan. Bencana terjadi dengan frekuensi yang tidak menentu dan akibat yang ditimbulkannya meningkat bagi mereka yang tidak mempersiapkan diri terhadap kemungkinan-kemungkinan timbulnya bencana

Firewall adalah sistem keamanan jaringan komputer yang digunakan untuk melindungi komputer dari beberapa jenis serangan dari komputer luar. Definisi Firewall adalah sebuah sistem yang didesain untuk mencegah akses yang tidak sah ke atau dari jaringan pribadi (*Private Network*).

Hadoop, *adalah framework atau platform open source berbasis Java di bawah lisensi Apache untuk support aplikasi yang jalan pada Big Data. Hadoop*

menggunakan teknologi Google MapReduce dan Google File System (GFS) sebagai fondasinya.

Hosting (disebut juga ***Web Hosting*** / sewa hosting) adalah penyewaan tempat untuk menampung data-data yang diperlukan oleh sebuah website dan sehingga dapat diakses lewat Internet. Data disini dapat berupa file, gambar, email, aplikasi/program/script dan database.

IaaS adalah ***Infastructure as a Service***, layanan cloud computing yang menyediakan infrastruktur dan perangkat keras seperti server, media penyimpanan, bandwidth, virtualisasi dan konfigurasi lain yang memungkinkan utilitas bagi pengguna.

iCloud adalah merupakan layanan komputasi awan terbaru yang dipublikasikan oleh Apple Inc. dalam acara *Apple Worldwide Developers Conference* (WWDC) yang diadakan tanggal 6 Juni 2011 di San Fransisco.[iCloud memungkinkan para penggunanya untuk mensinkronisasi data seperti foto, musik, dan dokumen ke dalam iPhone, iPad, iPod Touch, Mac dan komputer secara otomatis pada waktu yang bersamaan. Sehingga pengguna dapat mengaksesnya di mana saja dan kapan saja tanpa perlu mem-back up data secara manual**.**

IEEE adalah organisasi internasional, beranggotakan para insinyur, dengan tujuan untuk pengembangan teknologi untuk meningkatkan harkat kemanusiaan. Sebelumnya IEEE memiliki kepanjangan yang dalam Indonesia berarti Institut Insinyur Listrik dan Elektronik (***Institute of Electrical and Electronics Engineers***).

IOPS*, Input Output PerSeconds* adalah input/output per detik yang merupakan angka penting dari kemampuan masing-masing SSD. Angka IOPS yang memiliki kapasitas lebih besar maka performa dari SSD tersebut akan semakin baik. Latency adalah kecepatan dalam IO Task

IT adalah Teknologi Informasi (TI), atau dalam bahasa Inggris dikenal dengan istilah ***Information Technology*** (IT) adalah istilah umum untuk teknologi apa pun

yang membantu manusia dalam membuat, mengubah, menyimpan, mengomunikasikan dan/atau menyebarkan informasi.

ITIL, ***Information Technology Infrastructure Library*** (diterjemahkan Pustaka Infrastruktur Teknologi Informasi), adalah suatu rangkaian konsep dan teknik pengelolaan infrastruktur, pengembangan, serta operasi teknologi informasi (TI).

ITSM (***Information Technology Service Management***, Manajemen Layanan Teknologi Informasi) adalah suatu metode pengelolaan sistem teknologi informasi (TI) yang secara filosofis terpusat pada perspektif konsumen layanan TI terhadap bisnis perusahaan.

JSON, *JavaScript Object Notation* adalah format pertukaran data yang ringan, mudah dibaca dan ditulis oleh manusia, serta mudah diterjemahkan dan dibuat (generate) oleh komputer. Format ini dibuat berdasarkan bagian dari Bahasa Pemprograman JavaScript, Standar ECMA-262 Edisi ke-3 - Desember 1999.

NIST adalah ***National Institute of Standards and Technology***, (Badan Nasional Standar dan Teknologi Amerika Serikat) yang dulunya dikenal sebagai ***The National Bureau of Standards*** - NBS (Biro Standar Nasional) adalah sebuah badan non-regulator dari bagian Administrasi Teknologi dari Departemen Perdagangan Amerika Serikat. Misi dari badan ini adalah untk membuat dan mendorong pengukuran, standar, dan teknologi untuk meningkatkan produktivitas, mendukung perdagangan, dan memperbaiki kualitas hidup semua orang**.**

Node adalah setiap komputer, printer atau periferal yang terhubung dalam jaringan. Sebuah jaringan komputer sekurang-kurangnya terdiri dari dua unit komputer atau lebih, dapat berjumlah puluhan komputer, ribuan atau bahkan jutaan node yang saling terhubung satu sama lain.

OLA, ***Operational Level Agreement*** adalah kontrak yang menentukan bagaimana berbagai kelompok TI dalam perusahaan berencana memberikan layanan atau rangkaian layanan. OLA dirancang untuk mengatasi dan memecahkan masalah TI

dengan menetapkan seperangkat kriteria tertentu dan menentukan rangkaian layanan TI tertentu yang menjadi tanggung jawab masing-masing departemen.

OpenTDSP,

OPEX, ***Operating Expenditure*** adalah alokasi yang direncanakan dalam budget untuk melakukan operasi perusahaan secara normal. Dengan kata lain operating expenditure (biaya operasi) digunakan untuk menjaga kelangsungan aset dan menjamin aktivitas perusahaan yang direncanakan berlangsung dengan baik. Karena sifatnya biaya sehari-hari maka biaya operasi tidak meliput pajak pendapatan, depresiasi, dan biaya financing (bunga pinjaman).

OS adalah Sistem operasi (***Operating System***) adalah perangkat lunak sistem yang mengatur sumber daya dari perangkat keras dan perangkat lunak, serta sebagai daemon untuk program komputer. Tanpa sistem operasi, pengguna tidak dapat menjalankan program aplikasi pada komputer mereka, kecuali program booting.

PaaS adalah ***Platform as a Service***, kategori layanan komputasi awan yang menyediakan platform yang memungkinkan pelanggan untuk mengembangkan, menjalankan, dan mengelola aplikasi tanpa kompleksitas membangun dan memelihara infrastruktur yang biasanya terkait dengan pengembangan dan peluncuran aplikasi.

PBA, ***Patterns of Business Activity*** adalah profil beban kerja yang menggambarkan permintaan untuk layanan tertentu. PBAs adalah alat penting yang digunakan oleh Manajemen Permintaan untuk mengantisipasi dan mempengaruhi permintaan layanan**.**

RAM adalah berasal dari singkatan ***Random Access Memory***, RAM yaitu suatu memori tempat penyimpanan data sementara, ketika saat komputer dijalankan dan dapat diakses secara acak (random). Fungsi RAM adalah mempercepat pemeprosesan data pada PC atau computer.

ROI, ***Return on Investment*** adalah rasio laba bersih terhadap biaya. ROI biasanya merupakan pengukuran yang paling penting bagi pengiklan karena pengukuran ini didasarkan pada sasaran iklan tertentu dan menunjukkan pengaruh yang nyata dari upaya periklanan terhadap bisnis.

RPM,

SaaS adalah (***Software as a Service*** atau perangkat lunak berbentuk layanan) adalah suatu model penyampaian aplikasi perangkat lunak oleh suatu vendor perangkat lunak yang mengembangkan aplikasi web yang diinangi dan dioperasikan (baik secara mandiri maupun melalui pihak ketiga) untuk digunakan oleh pelanggannya melalui Internet**.**

SATA, (*Serial Advanced Technology Attachment*) adalah pada komputer yang didesain biasanya untuk mentransfer data antara motherboard dan media penyimpanan data, seperti hard disk dan optical drive di dalam komputer.

SDN, ***Software Defined Network*** adalah istilah yang merujuk pada konsep/paradigma baru dalam mendisain, mengelola dan mengimplementasikan jaringan, terutama untuk mendukung kebutuhan dan inovasi di bidang ini yg semakin lama semakin kompleks. Konsep dasar SDN adalah dengan melakukan pemisahan eksplisit antara control dan forwarding plane, serta kemudian melakukan abstraksi sistem dan meng-isolasi kompleksitas yg ada pada komponen atau sub-sistem dengan mendefinisikan antar-muka (interface) yg standard.

SLA, ***Service Level Agreement*** adalah kontrak dari penyedia layanan dengan kita sebagai pengguna yang memberikan jaminan tingkat pelayanan yang dapat diharapkan.

SSD, Solid State Drive adalah salah satu media penyimpanan utama selain hard disk. Di dalam sebuah komputer, SSD dan hard disk berfungsi sebagai media untuk menaruh semua data. Booting sebuah komputer juga menggunakan kedua benda ini.

TCO adalah singkatan dari ***total cost of ownership*** (biaya total kepemilikan) adalah jenis perhitungan yang dirancang untuk membantu konsumen dan manajer perusahaan menilai biaya dan manfaat yang terkait dengan pembelian komponen TI secara langsung dan tidak langsung

UKM, Usaha Kecil Menengah adalah salah satu motor penggerak perekonomian di negara kita. Bahkan menurut informasi yang saya baca di berbagai media informasi, Usaha mikro, kecil, dan menengah (UMKM) merupakan 'tulang punggung' perekonomian di Indonesia.

Virtual Sprawl, adalah fenomena yang terjadi ketika jumlah mesin virtual yang terhubung ke jaringan melebihi kemampuan jaringan. Didefinisikan sebagai sejumlah besar mesin virtual di jaringan tanpa manajemen atau kontrol TI yang tepat. Misalnya, Anda mungkin memiliki beberapa departemen yang memiliki server mulai membuat mesin virtual tanpa prosedur atau kontrol yang tepat dari rilis mesin virtual ini.

VM adalah ***Virtual Machine*** (Mesin Virtualisasi), implementasi perangkat lunak dari sebuah mesin (misalnya komputer) yang mengeksekusi program-program seperti mesin fisik. Mesin virtual dipisahkan menjadi dua kategori utama, didasarkan pada penggunaan dan tingkat korespondensi untuk setiap mesin nyata.

Daftar Pustaka

BACAAN

- T. Ruotsale et al. (2015). Interactive Intent Modeling: Information Discovery Beyond Search.
- J. C. Corbett et al. (2013). Spanner: Google's Globally Distributed Database.
- A. Gupta et al. (2016). Mesa: A Geo-Replicated Online Data Warehouse for Google's Advertising System.
- D. E. Eisenbud et al. (2016). Maglev: A Fast and Reliable Software Network Load Balancer.
- DA Menascé et al. Performance by Design: Computer Capacity Planning with Example. NJ Gunther. Guerrilla Capacity Planning.
- R. Cammarota et al. (2014). Pruning Hardware Evaluation Space through Similarity Analysis Correlation-driven Applications.
- K. Matthias and SP Kane. Docker: Up & Running: Delivery of Reliable Container in Production.
- Mouat. Using Docker: Developing and Deploying Software with Containers.
- K. Hightower. Kubernetes: Ride and Walk.
- EM Goldratt and J. Cox. Purpose: Continuous Improvement Process.
- G. Kim et al. Phoenix Project: Novel About IT, DevOps, and Help Your Business Win a Kindle Edition.
- EW Dijkstra. (1965). Solution to problems in concurrent programming control.
- L. Lamport. (1974). A new solution to Dijkstra's concurrent programming problems.
- GL Peterson and MJ Fischer. (1977). Economical solution to the problem of critical parts in a distributed system (Extended Abstract).
- M. Blasgen et al. (1977). The Convoy Phenomenon.
- Katseff cellphone. (1978). New solution to critical part problems.

- L. Lamport. (1986). Mutual Exclusion Problems: Part I - Interprocess Communication Theory.
- L. Lamport. (1986). Mutual Exclusion Problems: Part II - Statements and Solutions.
- T. Ruotsale et al. (2015). Interactive Intent Modeling: Discovery of Information Outside Search.
- JC Corbett et al. (2013). Spanner: Google's Global Distributed Database.
- Gupta et al. (2016). Mesa: A Geo-Replica Online Data Warehouse for the Google Ad System.
- DE Eisenbud et al. (2016). Maglev: Fast and Reliable Load Balancing Network Software.
- JM Anderson et al. (1997). Continuous profiling: where do all cycles go?
- G. Ren et al. (2010). Google-Wide Profiling: Sustainable Profile Infrastructure for Data Centers.
- J. Dai et al. (2011). HiTune: data-based performance analysis for big cloud data.
- M. Kambadur et al. (2012). Measure interference between direct data center applications.
- Ú. Erlingsson et al. (2012). Fay: Flexible Tracing from Kernels to Cluster.
- B. Gregg. (2012). Methodically Thinking about Performance.
- M. Schwarzkopf et al. (2013). Omega: flexible and scalable schedulers for large computing clusters.
- C. Wang et al. (2013). Solving performance problems in the data center: annotated bibliography?
- S. Mahlke et al. (2013). Sampling instrumentation for profile data center applications.
- Verma et al. (2015). Large-scale cluster management on Google with Borg.
- S. Kanev et al. (2015). Creating a warehouse-scale computer profile.
- Burns et al. (2016). Borg, Omega, and Kubernetes: Lessons learned from three container management systems over a decade.
- W. Hassanein. (2016). Understand and improve JVM GC work on a data center scale
- Y. Zhang et al. (2016). Collection of data-based backup and storage cycles on a large scale data center.

- GE Moore. (1965). Collect More Components into Integrated Circuits.
- MA Cusumano and DB Yoffie. (2016). Extrapolation from Moore's Law.
- PE Denning and TG Lewis. (2017). Exponential Law of Computational Growth.

SUMBER DAYA

- "Benchmarking Cassandra Scalability on AWS—Over a million writes per second." (2011) http://techblog.netflix.com/2011/11/benchmarking-cassandra-scalability-on.html.
- "Mobile vs Desktop: 13 Essential User Behaviors." (2016) http://bit.ly/mobile-vs-desktop-13.
- "Keywords Are Dead! Long Live User Intent!" (2013) http://bit.ly/keywords-are-dead.
- "Measuring Perceived Performance." (2016) http://bit.ly/measuring-perceived.
- "A Practical Guide to SLAs." (2016) http://bit.ly/sla-practical-guide.
- "The Very Real Performance Impact on Revenue." (2017) http://blog.catchpoint.com/2017/01/06/performance-impact-revenue-real/.
- "Performance Impact of Third Party Components." (2016) http://blog.catchpoint.com/2016/09/23/third-party-performance-impact/.
- "Speed Index." https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index.
- "Above the Fold Time: Measuring Web Page Performance Visually." (2011) http://bit.ly/above-the-fold-time.
- "Hero Image Custom Metrics." (2015) http://bit.ly/hero-image.
- "Critical Metric: Critical Resources." (2016) http://bit.ly/crit-met-crit-res.
- "Cloud Environment Will Encourage IT Infrastructure Spending to Spend All Regional Markets in 2016, According to IDC." (2016) http://bit.ly/idc-cloud-env.
- "Cisco Visual Networking Index: Forecast and Methodology, 2016-2021." (2017) http://bit.ly/cisco-vis-net.
- M. Costigan. (2016). Risk-based Capacity Planning.
- "Risk-based Capacity Planning." (2016) http://ubm.io/2h85HKL.

- "You Now Have a Shorter Attention Than Carp." (2017) http://ti.me/2wnVOQv.
- L. Ridley. (2014). People exchange devices 21 times per hour, OMD said.
- B. Koley. (2014). Network Defined Software on Scale.
- "Scaling to exabytes and so on." (2016) https://blogs.dropbox.com/tech/2016/03/magic-pocket-infrastructure/.
- "Risk Mitigation through Capacity Planning to achieve Competitive Advantage." (2013) http://india.cgnglobal.com/node/47.
- "Epic Story of the Dropbox Output from the Amazon Cloud Empire." (2016) https://www.wired.com/2016/03/epic-story-dropboxs-exodus-amazon-cloud-empire/.
- "Speed Matters for Google Web Search." (2009) http://services.google.com/fh/files/blogs/google_delayexp.pdf.
- "Cedexis Announces Impacts, Linking Website Performance With Online Business Results." (2015) http://www.cedexis.com/blog/cedexis-announces-impact-connects-website-performance-to-online-business-results /
- "How Time of Loading Affects Your Bottom Line." (2011) https://blog.kissmetrics.com/loading-time/.
- " Speed Is a Killer - Why Reducing Time to Load Pages Can Drastically Increase Conversions. " (2011) https://blog.kissmetrics.com/speed-is-a-killer/.
- • "Why Web Performance Is Important: Is Your Site Dispelling Customers?" (2010) http://bit.ly/why-web-perf.
- • "Why You Need a Serious Website." (2013) http://www.copyblogger.com/website-speed-matters/.
- "Monitor and Improve Web Performance Using RUM Data Visualization." (2014) http://bit.ly/mon-improve-web-perf.
- "The Importance of Speed of Website Loading & 3 Top Factors that Limit Web Site Speed." (2014) http://bit.ly/importance-load-speed.
- • "Seven Thumb Rules for Website Experiments." (2014) http://stanford.io/2wsXzKJ.
- • "SEO 101: How important is Site Speed in 2014?" (2014) http://www.searchenginejournal.com/seo-101-important-site-speed-2014/111924/.

- "User Preferences and Search Engine Latency." Http://bit.ly/user-pref-search.
- "Flash Sale Engineering." (2016) https://www.usenix.org/conference/srecon16europe/program/presentation/stolarsky.
- "How do micro services break the company's monolith." (2016) http://www.appstechnews.com/news/2016/nov/16/micro-services-breaking-down-monolith/.
- "Solving Monolithic Software: A Case for Microservices vs. Independent System. "(2016) http://bit.ly/breaking-down-monolith.
- "Break through the Monolithic API into Micro Services in Uber." (2016) https://www.infoq.com/news/2016/07/uber-microservices.
- "What is your headroom?" (2016) http://akamai.me/2x5a3eR.
- 1 "Make data analytics work for you - rather than vice versa" (2016) http://bit.ly/making-analytics-work
- 2 Potential root causes can be considered, but not limited to, poor selection of algorithms / data structures or poor implementations.
- "A Working Theory-of-Monitoring." (2013) https://www.usenix.org/conference/lisa13/working-theory-monitoring .
- "EBPF and system performance." (2017) https://www.oreilly.com/ideas/ebpf-and-systems-performance.
- "runtime metrics." Https://docs.docker.com/engine/admin/runmetrics.
- "Linux Performance Analysis in 60,000 Milliseconds". (2015) http://www.brendangregg.com/Articles/Netflix_Linux_Perf_Analysis_60s.pdf.
- "Linux performance." Http://www.brendangregg.com/linuxperf.html.
- "UNIX Load Average Part 1: How It Works." (2010) http://www.teamquest.com/files/9214/2049/9761/ldavg1.pdf.
- "UNIX Load Average Part 2: Not Your Average Average." (2010) http://www.teamquest.com/import/pdfs/whitepaper/ldavg2.pdf.
- "Understanding load rates and stretching factors." (2007) http://www.linux-magazine.com/content/download/62593/485442/Load_Average.pdf.
- "System tutorial." (2015) https://sourceware.org/systemtap/tutorial.pdf.
- "Go to GitHub repo." Https://github.com/ktap/ktap.

- “ Thought about the Time-series Database." (2015) http://jmoiron.net/blog/thoughts-on-timeseries-databases.
- "MEAN is great, but then you grow up." (2014) https://rclayton.silvrback.com/means-great-but-then-you-grow-up. “Load Balancing Methods & Algorithms.” Http://www.peplink.com/technology/load-balancing-algorithms/ .
- “SSD vs HDD.” Http://www.storagereview.com/ssd_vs_hdd .
- "The sad situation of server use and the upcoming post-hypervisor era." (2013) http://bit.ly/sorry-state-server.
- "How one startup hopes to complete server underutilization." (2015) http://bit.ly/solve-server-under.
- "Cache replacement policy - Reference." Http://bit.ly/cache-replacement-policies.
- 1 There are several commercial services for RUM, such as Catchpoint, Keynote, Soasta
- “Benchmarking Cassandra Scalability on AWS—Over a million writes per second.” (2011) http://techblog.netflix.com/2011/11/benchmarking-cassandra-scalability-on.html.
- “Mobile vs Desktop: 13 Essential User Behaviors.” (2016) http://bit.ly/mobile-vs-desktop-13.
- “Keywords Are Dead! Long Live User Intent!” (2013) http://bit.ly/keywords-are-dead.
- “Measuring Perceived Performance.” (2016) http://bit.ly/measuring-perceived.
- “A Practical Guide to SLAs.” (2016) http://bit.ly/sla-practical-guide.
- “The Very Real Performance Impact on Revenue.” (2017) http://blog.catchpoint.com/2017/01/06/performance-impact-revenue-real/.
- “Performance Impact of Third Party Components.” (2016) http://blog.catchpoint.com/2016/09/23/third-party-performance-impact/.
- “Speed Index.” https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index.
- “Above the Fold Time: Measuring Web Page Performance Visually.” (2011) http://bit.ly/above-the-fold-time.
- “Hero Image Custom Metrics.” (2015) http://bit.ly/hero-image.

- “Critical Metric: Critical Resources.” (2016) http://bit.ly/crit-met-crit-res.
- “Benchmarking Cassandra Scalability on AWS—Over a million writes per second.” (2011) http://techblog.netflix.com/2011/11/benchmarking-cassandra-scalability-on.html.
- “Mobile vs Desktop: 13 Essential User Behaviors.” (2016) http://bit.ly/mobile-vs-desktop-13.
- “Keywords Are Dead! Long Live User Intent!” (2013) http://bit.ly/keywords-are-dead.
- “Measuring Perceived Performance.” (2016) http://bit.ly/measuring-perceived.
- “A Practical Guide to SLAs.” (2016) http://bit.ly/sla-practical-guide.
- “The Very Real Performance Impact on Revenue.” (2017) http://blog.catchpoint.com/2017/01/06/performance-impact-revenue-real/.
- “Performance Impact of Third Party Components.” (2016) http://blog.catchpoint.com/2016/09/23/third-party-performance-impact/.
- “Speed Index.” https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index.
- “Above the Fold Time: Measuring Web Page Performance Visually.” (2011) http://bit.ly/above-the-fold-time.
- “Hero Image Custom Metrics.” (2015) http://bit.ly/hero-image.
- “Critical Metric: Critical Resources.” (2016) http://bit.ly/crit-met-crit-res.

KINERJA

JARINGAN

- Singh et al. (2015). Jupiter Rising: A Decade of Topology Clos and Centralized Control on Google Datacenter.
- Guo et al. (2015). Pingmesh: Large Scale System for Data Center Network Latency Measurement and Analysis.
- YE Sung et al. (2016). Robotron: Top-down Network Management on the Scale of Facebook.
- R. Govindan et al. (2016). Evolve or Die: Principles of High Availability Design Taken from the Googles Network Infrastructure.

- P. Tammana et al. (2016). Simplifies the datacenter network debugger with PathDump.
- Y. Geng et al. (2016). Juggler: a powerful network of practical reordering for data centers.
- K. He et al. (2016). AC / DC TCP: Virtual Congestation Control Enforcement for Datacenter Networks.
- T. Chen, X. Gao and G. Chen. (2016). Features, hardware, and data center network architecture.
- WM Mellette et al. (2016). P-FatTree: A multi-channel datacenter network topology.

LOAD BALANCER

- DE Eisenbud et al. (2016). Maglev: fast and reliable software network load balancer.

PENYIMPANAN

- C. Ruemmler and J. Wilkes. (1993). UNIX Disk Access Pattern.
- C. Ruemmler and J. Wilkes. (1994). Introduction to disk drive modeling.
- D. Anderson et al. (2003). More Than Interfaces - SCSI vs. ATA.
- D. Anderson. (2003). You Don't Know Jack About Disks.
- WW Hsu and AJ Smith. (2004). Performance impact of I / O optimization and disk repair.
- J. Elerath. (2007). Hard Disk Drives, The Good, Bad and Ugly.
- MK McKusick. (2012). Disk from the File System Perspective: Disk lying. And the controller who runs it is a partner in crime.
- M. Cornwell. (2012). Anatomy of a Solid-state Drive.

DATABASE DAN CACHING

- PJ Denning. (1968). Model work arrangements for program behavior.
- PJ Denning. (1980). Past and Present Working Sets.
- H.-T. Chou and DJ DeWitt. (1985). Evaluation of Buffer Management Strategies for Relational Database Systems.
- EJ O'Neil et al. (1993). Substitute Algorithms for LRU-K Pages for Database Disk Buffers.

- R. Nishtala et al. (2013). Memcache Scaling on Facebook.
- S. Podlipnig and L. Boszormenyi. (2003). Web Cache Replacement Strategy Survey.

BELI ATAU SEWA

- RF Vancil. (1961). Lease or Borrow: New Analysis Method.
- KD Ripley. (1962). Leasing: a means of financing a business that should not be ignored.
- RW Johnson. (1972). Analysis of Lease-or-Buy Decisions.
- WL Sartoris and Paul Hospital. (1973). Rent Evaluation: Other Capital Budgeting Decisions.
- GB Harwood and RH Hermanson. (1976). Lease-or-Buy Decision.
- Sykes. (1976). Lease-Buy Decision - A Survey of Current Practices in 202 Companies.
- Jack E. Gaumnitz and Allen Ford. (1978). Lease or Sell Decision.
- ACC Herst. (1984). Rent or Buy: Theory and Practice.
- BH Nunnally, JR and DA Plath. (1989). Leasing Versus Loans: Evaluating Alternative Consumer Credit Forms.
- FJ Fabozzi. (2008). Lease versus Borrow-to-Buy Analysis.
- E. Walker. (2009). Real Cost of CPU Clock.
- E. Walker et al. (2010). For Rent or Not for Rent from a Storage Cloud.
- Time Series Forecasting
- GEP Box, et al. (2015). Time-Series Analysis: Forecasting and Control (5th edition).
- PJ Brockwell and RA Davis. (2002). Introduction to Time Series and Forecasting (edition 2).

FITTING KURVA

- G. Wabba. (1990). Spline Model for Observation Data.
- S. Arlinghaus. (1994). Practical Handbook from Curve Fitting.
- GD Garson. (2012). Curve Fitting and Non-linear Regression.

SUMBER DAYA

- “Moore's Law Is Dead. Now what? "(2016) http://bit.ly/moores-law-dead.

- L. Muehlhauser. (2014). Exponential and non-exponential trends in information technology.

LAMPIRAN A

KERANGKA RENCANA KAPASITAS

Berikut ini adalah ikhtisar bagian-demi-bagian tentang apa yang harus dimasukkan dalam rencana kapasitas. Jadi ini pada dasarnya adalah template untuk rencana kapasitas.

Rencana kapasitas digunakan untuk mengelola sumber daya yang diperlukan untuk memberikan layanan TI. Rencana ini berisi skenario untuk prediksi yang berbeda dari permintaan bisnis dan opsi biaya untuk memberikan target tingkat layanan yang disepakati.

Tujuan dari rencana kapasitas yang dijelaskan dan templated di bawah ini adalah untuk membantu merumuskan strategi untuk menilai dan mengelola kinerja komponen infrastruktur. Informasi yang dihimpun dalam rencana kapasitas memfasilitasi keputusan perencanaan kapasitas. Keputusan ini mungkin termasuk akuisisi infrastruktur tambahan, perubahan konfigurasi, dan keputusan peningkatan untuk memenuhi persyaratan bisnis secara efisien.

Perencanaan untuk kapasitas memastikan bahwa persyaratan bisnissecara efisien dan efektif dipenuhi oleh infrastruktur dan elemen penerapan solusi. Ini menyediakan manajemen dengan hal-hal berikut:
Pemahaman yang jelas tentang kapasitas sumber daya saat ini untuk mendukung solusi bisnis
Penilaian kapasitas manajemen kapasitas saat ini
Daftar sumber daya yang akan dimasukkan atau ditingkatkan untuk memenuhi permintaan bisnis di masa depan

KONTROL BERKAS

Dalam rencana kapasitas, perlu untuk membuat sistem kerja kontrol dokumen. Seiring berjalannya rencana, tahapan rencana berjalan melalui perubahan dan begitu juga dengan dokumen rencana. Tidak mengherankan bahwa ada istilah-istilah yang biasa digunakan untuk berbagai tahap rencana kapasitas yang dilalui. Status dokumen meliputi tahap-tahap berikut:
Draf: Ini adalah dokumen untuk ditinjau dan kemungkinan memiliki perubahan signifikan.
Baseline Kerja: Dokumen telah mencapai akhir fase peninjauan awal dan dapat digunakan sebagai dasar untuk desain teknis, tetapi diharapkan memiliki perubahan lebih lanjut. Dokumen ini akan mengalami perubahan yang dilacak sejak draf terakhir.
Calon Dasar: Dokumen ini siap untuk diterbitkan terakhir dan hanya diharapkan memiliki pembaruan kecil lebih lanjut.
Baseline: Dokumen ini diterbitkan dan diperkirakan tidak akan berubah. Dokumen ini akan mengalami perubahan yang dilacak sejak baseline kerja.
Sekarang mari kita lihat materi rencana kapasitas dan apa yang harus dimasukkan di bawah setiap bagian. Harap dicatat bahwa rencana ini didasarkan pada praktik terbaik sebagaimana ditentukan oleh penyedia teknologi terkemuka dan kerangka kerja seperti ITIL. Anda pasti dapat memiliki rencana kapasitas organisasi-spesifik di lingkungan TI yang dinamis dan selalu berubah ini. Sisa dari apendiks ini adalah template rencana kapasitas aktual.

RINGKASAN BISNIS PLAN

Sebagian besar rencana kapasitas, dengan kebutuhan, mengandung detail teknis yang tidak menarik bagi semua pembaca rencana. Bagian ringkasan eksekutif

harus menyoroti masalah utama, opsi, rekomendasi, dan biaya. Ini juga harus berisi poin-poin utama dari masing-masing bagian dari rencana utama. Tujuan utama dari dokumen ini adalah untuk memprovokasi keputusan investasi untuk area yang tepat dan untuk menghindari atau menunda keputusan investasi jika diperlukan. Oleh karena itu, bagian ini harus memotong langsung ke masalah bisnis utama yang harus ditangani dan keputusan yang diperlukan — mengingat bahwa keputusan untuk tidak mengambil tindakan masih merupakan keputusan kunci.

Hindari detail teknologi yang tidak penting untuk keputusan. Nyatakan masalah yang memiliki dampak bisnis paling besar, opsi biaya, dan rekomendasi yang dapat dibenarkan secara singkat.

Keseluruhan kapasitas dan rencana kinerja cenderung mengandung banyak investasi yang disarankan dan analisis keuangan yang kompleks, sehingga ringkasan eksekutif dari keseluruhan kapasitas infrastruktur dan rencana kinerja dapat dipisahkan menjadi dokumen terpisah dan dapat berfokus pada keuangan dan bisnis berbasis-pembinaan yang spesifik.

Rencana kapasitas mendukung sasaran optimal, dan efektif biaya, penyediaan sumber daya dan layanan organisasi dengan mencocokkannya dengan tuntutan bisnis. Rencana kapasitas mencerminkan kebutuhan bisnis saat ini dan masa depan. Ini membantu mengidentifikasi dan mengurangi inefisiensi yang terkait dengan sumber daya yang kurang dimanfaatkan atau permintaan pelanggan yang tidak terpenuhi dan untuk menyediakan tingkat layanan yang memuaskan dengan cara yang hemat biaya. Rencana ini membantu memastikan bahwa semua komponen infrastruktur mampu melakukan semua fungsi yang dibutuhkan, dan bahwa komponen-komponen tersebut akan bekerja seefisien mungkin dan dapat mengakomodasi pertumbuhan yang wajar tanpa terlalu boros.

Bagian ini harus berisi informasi ringkasan tentang hal-hal berikut:

Ruang lingkup rencana (apakah ini rencana tahunan, enam bulan, atau bergulir bulanan)

Komponen, layanan, fasilitas, sumber daya, dan keterampilan dalam lingkup rencana ini

TINGKAT KAPASITAS SAAT INI

Kinerja saat ini disampaikan, termasuk pencapaian tingkat layanan dan informasi tentang insiden kinerja yang dicatat

Ringkasan insiden yang disebabkan oleh kurangnya kapasitas

Pandangan tentang kapan insiden layanan atau dampak keuangan dipertimbangkan karena kelebihan atau kekurangan kapasitas
Perubahan dalam infrastruktur, lingkungan bisnis, rencana, dan prakiraan sejak edisi terakhir dari rencana

RUANG LINGKUP DAN KERANGKA ACUAN RENCANA

Bagian rencana ini harus mempertimbangkan tujuan, sasaran, dan hasil, ditambah masalah orang, proses, alat, dan teknik yang lebih luas. Ini harus mencakup semua sumber daya TI.
Misalnya, rencana kapasitas harus mencakup perincian berikut:

- Manajemen kapasitas bisnis (menerjemahkan kebutuhan bisnis dan rencana menjadi persyaratan untuk layanan dan infrastruktur TI)
- Manajemen kapasitas layanan (manajemen, kontrol, dan prediksi kinerja end-to-end dan kapasitas penggunaan layanan dan beban kerja IT operasional langsung)
- Manajemen kapasitas komponen (manajemen, kontrol, dan prediksi kinerja, pemanfaatan, dan kapasitas komponen teknologi IT individu)
- Bagian ini harus secara eksplisit menyebutkan elemen-elemen di atas yang disertakan dan yang tidak termasuk.

METODE YANG DIGUNAKAN

Manajemen kapasitas sangat tergantung pada informasi yang diberikan oleh proses lain. Bagian ini harus menyatakan sumber informasi, alat yang digunakan untuk mengumpulkan dan menganalisis informasi, dan metode yang digunakan untuk memodelkan dampak pada infrastruktur dan kinerja layanan. Hal ini mungkin termasuk pemantauan data dari komponen infrastruktur, alat kinerja aplikasi, perkiraan bisnis (termasuk dampak ekonomi makro), perkiraan beban kerja, teknik pemodelan, dan output dari alat pemodelan layanan.
Ini harus mencakup data kinerja, ketersediaan, dan tingkat layanan yang biasanya dihasilkan oleh alat pemantauan yang ada yang menyelidiki komponen infrastruktur dan pelaksanaan aplikasi, yang menghasilkan perkiraan beban kerja rinci dan statistik pada pengiriman pengguna akhir. Organisasi yang lebih matang juga akan mencakup prakiraan bisnis (termasuk dampak ekonomi makro), teknik pemodelan yang digunakan, dan output dari alat pemodelan layanan.

APLIKASI

Ini harus menyertakan data mengenai masing-masing aplikasi organisasi dan mendokumentasikannya. Juga harus dipastikan bahwa semua aplikasi telah diperhitungkan dan informasi mengenai setiap aplikasi sudah benar.

Infrastruktur

Bagian ini mencakup data mengenai aset virtual dan fisik pusat data dan mendokumentasikannya di dalam buku kerja ini. Semua perangkat harus dipertanggungjawabkan dan informasi mengenai setiap perangkat harus benar. Bagian ini menerjemahkan permintaan layanan yang diperkirakan menjadi pemanfaatan sumber daya infrastruktur (seperti prosesor, memori, penyimpanan, lisensi, jaringan, pusat data, dan daya). Praktik terbaik adalah menggunakan model kapasitas layanan untuk membuat perhitungan ini. Investasi dalam alat-alat kapasitas dapat sangat efektif dalam meningkatkan kualitas analisis permintaan sumber daya dan dampak visual dari materi.

Mengumpulkan data yang diperlukan untuk menulis bagian ini akan memerlukan bekerja erat dengan infrastruktur dan memonitor staf teknis. Kerja sama yang erat dalam upaya ini di antara berbagai tim teknis akan sering menghasilkan wawasan tentang bagaimana memanfaatkan investasi yang ada dengan lebih baik untuk mendapatkan data yang lebih baik dan kinerja yang lebih baik.

Bagian ini harus menunjukkan pemanfaatan baru-baru ini di semua menara infrastruktur dan juga memperkirakan permintaan sumber daya setidaknya 12 bulan ke depan. Setiap tanggal kritis di mana kendala infrastruktur akan menghasilkan insiden kinerja atau pemadaman harus disorot. Bagian opsi dan rekomendasi harus berisi saran untuk menghindari dampak bisnis negatif.

SKENARIO TUGAS PENGGUNA

Skenario penggunaan (juga disebut use cases) menentukan urutan tugas yang dilakukan pengguna dan interaksi mereka dengan fungsi solusi untuk membantu mereka melakukan tugas. Bagian dari rencana kapasitas ini harus mendefinisikan skenario yang dikerjakan oleh pengguna di bidang fungsional masing-masing. Mengidentifikasi dan menjelaskan skenario penggunaan memberikan detail yang memungkinkan estimasi beban kapasitas dan faktor lainnya.

MATRIKS BEBAN TUGAS

Matriks beban-tugas menggambarkan berbagai jenis beban yang digunakan setiap skenario penggunaan pada sistem. Matriks ini mengukur konfigurasi server dan klien sesuai dengan skenario perencanaan kapasitas.

PEMANTAUAN DAN METRIK

Bagian pemantauan dan metrik menjelaskan metode pemantauan, teknik, dan alat yang akan digunakan untuk mengevaluasi solusi dan kinerja komponen, dan menyediakan metrik untuk merencanakan intervensi. Informasi ini harus disediakan untuk setiap komponen utama pada tingkat solusi.

PERMINTAAN DAN PRAKIRAAN LAYANAN

Rencana bisnis harus menyediakan manajemen kapasitas dengan rincian layanan baru yang direncanakan dan pertumbuhan atau kontraksi dalam penggunaan layanan yang ada. Subbagian ini harus melaporkan layanan baru dan matinya sistem warisan. Bagian ini harus profil layanan TI yang disediakan dalam hal akrab dengan manajer layanan dan pemimpin bisnis (seperti puncak transaksi layanan, rata-rata dan total, jumlah rekening diproses, dan sebagainya). Profil dan prakiraan harus disediakan untuk layanan baru dan yang sudah ada, termasuk setiap rencana untuk pensiun layanan. Perkiraan jangka pendek, jangka menengah, dan jangka panjang harus dimasukkan berdasarkan informasi terbaik yang tersedia dari rencana bisnis, promosi, dan jadwal kegiatan. Layanan bisnis yang penting dan berdampak tinggi harus diprofilkan secara individual, sementara layanan yang kurang kritis dapat dikumpulkan,

TINGKAT LAYANAN

Tingkatan ini digunakan untuk memastikan bahwa tingkatan layanan yang tepat seperti Emas, Perunggu, dan Perak didasarkan pada kriteria kritikalitas aplikasi yang telah ditetapkan sebelumnya.

PIHAK KETIGA

Bagian ini mengumpulkan semua informasi kontak dari vendor dan kontak internal sehingga tersedia bagi semua pemangku kepentingan. Setelah disusun, rencana kapasitas disirkulasikan di antara semua pemangku kepentingan dalam format yang tidak ambigu dan dapat diakses yang menyajikan komponen, layanan,

dan pandangan bisnis untuk implementasi dalam manajemen kapasitas yang berkelanjutan.

PROFIL PENGGUNA

Profil pengguna menggambarkan pengguna solusi yang diusulkan dan karakteristik pengguna penting tertentu seperti frekuensi penggunaan solusi dan kompetensi dalam menggunakan solusi. Pengguna dapat diidentifikasi dalam kelompok (atau kelas), biasanya dinyatakan dalam hal bidang fungsional mereka. Pengguna teknologi informasi termasuk help desk, administrasi database, dll. Pengguna bisnis termasuk akuntansi, pergudangan, pengadaan, dll. Menggambarkan pengguna dan karakteristik penting mereka membantu dalam membentuk skenario yang relevan dengan kapasitas.

SKENARIO PENGGUNAAN

Skenario penggunaan (juga disebut use cases) menentukan urutan tugas yang dilakukan pengguna dan interaksi mereka dengan fungsi solusi untuk membantu mereka melakukan tugas. Bagian ini harus menentukan skenario yang dilakukan oleh pengguna di setiap area fungsional. Mengidentifikasi dan menjelaskan skenario penggunaan memberikan detail yang memungkinkan estimasi beban kapasitas dan faktor lainnya.

Skenario 1 << Deskripsi skenario >>

Bagian ini menjelaskan skenario dan karakteristiknya yang memberikan masukan untuk memperkirakan beban, pertumbuhan, dan dampak.

Skenario 2 << Deskripsi skenario >>

Matriks beban tugas

Bagian ini menjelaskan komponen muatan beban apa yang harus dilakukan untuk menjalankan tugas pengguna. Tabel A-1 mengukur untuk konfigurasi server dan klien yang sesuai dengan skenario perencanaan kapasitas.

Tabel A-1 . Kapasitas Beban untuk Server dan Konfigurasi Klien

Scenario Load	User Scenario 1	User Scenario 2	User Scenario 3	User Scenario 4	User Scenario 5
Storage					
Software					
CPU					
Memory					
I/O					
Others					

PERTUMBUHAN YANG DIHARAPKAN

Tabel pertumbuhan yang diharapkan (Tabel A-2) menggambarkan pola pertumbuhan untuk skenario sebagai fungsi waktu.

Tabel A-2 . Pertumbuhan yang Diharapkan

Tipe Kapasitas	Analisis Kapasitas Saat Ini	Pertumbuhan dan Rekomendasi yang Direncanakan / Diharapkan
Jelaskan skenario kapasitas yang dianalisis. Masukkan detail tentang persyaratan kapasitas saat ini dan masa mendatang.	Jelaskan kapasitas yang tersedia saat ini.	Jelaskan bagaimana harapan pertumbuhan masa depan telah diidentifikasi dan dianalisis.Buat garis besar rekomendasi untuk mengelola dan menangani pertumbuhan

Tipe Kapasitas	Analisis Kapasitas Saat Ini	Pertumbuhan dan Rekomendasi yang Direncanakan / Diharapkan
		yang diharapkan ini.

ASUMSI BUATAN

Kesalahan pemodelan kurang umum daripada kegagalan dalam asumsi tentang pendorong dan prakiraan bisnis. Nyatakan semua asumsi bisnis, ekonomi, dan teknis yang dibuat dalam produksi rencana. Ini cukup sulit, karena asumsi yang paling berbahaya adalah yang tidak kita sadari — saksikan kesulitan saat ini di zona Euro. Anda harus seketat mungkin dalam menyatakan asumsi yang mendukung rencana dan keputusan yang ditetapkan.

MEMINTA CADANGAN SISTEM

Bagian ini mendefinisikan kapasitas cadangan semua komponen sistem yang dibutuhkan oleh solusi, termasuk yang berikut:

- Jaringan
- Server
- Klien
- Aplikasi

KAPASITAS KOMPONEN

Bagian ini mengidentifikasi komponen solusi (manusia, peralatan, perangkat lunak, fasilitas, dll.) Dan menentukan kapasitas arus komponen. Anda dapat merekam informasi ini dalam tabel yang menetapkan pemecahan berbagai komponen fungsional solusi dan mencatat parameter pengukuran yang relevan untuk batas kapasitas yang diketahui sistem. Tabel A-3 menetapkan pemecahan berbagai

komponen fungsional dari solusi sistem dan mencatat parameter pengukuran yang relevan untuk batas kapasitas yang diketahui untuk sistem.

Tabel A-3 . Komponen dan Batas Kapasitas yang Diketahui

Functional areas	Components	Configuration	Measurement	Capacity	Comments
Virtual Architecture					
Protocols and Transport					
Operating Systems					
Application Software					
Others					

HAMBATAN

Bagian ini menjelaskan area apa pun dari solusi sistem yang dapat mewakili kemacetan fungsional.

Strategi Pertumbuhan dan Intervensi

Bagian ini menjelaskan cara solusi diproyeksikan untuk menambah kapasitas tambahan, seperti berikut:

Inkremental vs penggantian

Skala horizontal vs. vertikal

Paralel vs. hub berbicara

Teknologi baru

RINGKASAN LAYANAN

Bagian ringkasan layanan harus menyertakan subbagian berikut.

Ketentuan Layanan Saat Ini dan Terakhir

Untuk setiap layanan yang dikirimkan, berikan profil layanan. Ini harus mencakup tingkat throughput dan pemanfaatan sumber daya yang dihasilkan (seperti memori, ruang penyimpanan, kecepatan transfer, penggunaan prosesor, dan penggunaan jaringan). Tren jangka pendek, menengah, dan panjang harus disajikan di sini.

PRAKIRAAN LAYANAN

Rencana bisnis harus menyediakan manajemen kapasitas dengan rincian layanan baru yang direncanakan dan pertumbuhan atau kontraksi dalam penggunaan layanan yang ada. Subbagian ini harus melaporkan layanan baru dan matinya sistem warisan.

RINGKASAN TEMUAN

Jika berlaku, jelaskan pola pertumbuhan kapasitas historis. Jelaskan bagaimana kebutuhan kapasitas yang diharapkan di masa depan telah diidentifikasi dan dianalisis. Buat garis besar rekomendasi untuk mengelola dan menangani pertumbuhan yang diharapkan.

Masukkan tabel / ilustrasi, atau berikan referensi ke tempat penyimpanannya, yang menunjukkan rekomendasi berbeda untuk mengatasi masing-masing skenario kapasitas yang digambarkan di atas. Contoh di bawah ini akan bervariasi dari proyek ke proyek.

Jelaskan bagaimana pertumbuhan yang diharapkan akan dipantau dan dikelola. Tabel A-4 adalah contoh dasar dari tabel yang dapat digunakan untuk menggambarkan satu pendekatan untuk memantau dan mengelola kapasitas masa depan. Pendekatan yang digunakan untuk menggambarkan persyaratan ini mungkin berbeda dari proyek ke proyek.

Tabel A-4 . Prediksi Pertumbuhan

Area/Item Monitored	Capacity Requirement(s)	% Increase NeededPer (Time Period)	Capacity Threshold(s)	Threshold Response Strategy (Action to Be Take Upon Reaching Threshold(s))
< Server>	<Enter capacity requirements and measures>	<Enter projected increases over intervals of time>	<Enter acceptable capacity threshold(s)>	<Enter response strategies to varying threshold limits. *Threshold* is defined as the level at which an event or change occurs>

RINGKASAN SUMBER DAYA

Subbagian ini berkonsentrasi pada penggunaan sumber daya yang dihasilkan oleh layanan. Ini melaporkan, sekali lagi, pada tren jangka pendek, menengah, dan panjang dalam penggunaan sumber daya, dipecah oleh platform perangkat keras. Informasi ini telah dikumpulkan dan dianalisis oleh subproses dari manajemen kapasitas layanan dan manajemen kapasitas komponen, sehingga harus tersedia.

OPSI UNTUK PENINGKATAN LAYANAN

Berdasarkan hasil dari bagian sebelumnya, bagian ini menguraikan opsi-opsi yang mungkin untuk meningkatkan efektivitas dan efisiensi penyampaian layanan. Ini

dapat berisi opsi untuk menggabungkan berbagai layanan pada satu prosesor, meningkatkan jaringan untuk memanfaatkan kemajuan teknologi, menyetel penggunaan sumber daya atau kinerja layanan, menulis ulang sistem warisan, membeli perangkat keras atau perangkat lunak baru, dll.

PRAKIRAAN BIAYA

Biaya yang terkait dengan opsi ini harus didokumentasikan di sini. Selain itu, biaya saat ini dan perkiraan penyediaan layanan TI harus dimasukkan. Dalam praktiknya, manajemen kapasitas memperoleh banyak informasi ini dari proses manajemen keuangan dan rencana keuangan TI.

Pemantauan dan Metrik

Bagian ini menjelaskan berbagai komponen yang memerlukan pemantauan untuk manajemen kinerja. Misalnya, server virtual dapat menjadi komponen dan utilisasi CPU-nya dapat menjadi salah satu metrik. Bagian ini menjelaskan metode pemantauan, teknik, dan alat yang akan digunakan untuk mengevaluasi solusi dan kinerja komponen, dan menyediakan metrik untuk merencanakan intervensi. Informasi ini harus disediakan untuk setiap komponen utama pada tingkat solusi.

KOMPONEN 1 << PELAYANAN >>
KOMPONEN 2 << PENYIMPANAN >>
KOMPONEN 3 << KOMPONEN >>

AMBANG BATAS UNTUK INTERVENSI

Bagian ini mengidentifikasi, menjelaskan, dan mengkuantifikasi ambang batas untuk memicu intervensi dengan mengubah konfigurasi komponen.

REKOMENDASI

Bagian terakhir dari rencana harus berisi ringkasan rekomendasi yang dibuat dalam rencana sebelumnya dan status mereka (misalnya, ditolak, direncanakan, diimplementasikan) dan setiap varian dari rencana itu. Setiap rekomendasi baru dalam iterasi rencana ini harus dibuat di sini (yaitu, yang mana dari opsi yang disebutkan dalam rencana itu lebih disukai). Ini juga harus mencakup implikasi jika rencana, dan rekomendasinya, tidak dilaksanakan.

Rekomendasi harus dikuantifikasi dalam hal berikut:

Manfaat bisnis yang diharapkan
Dampak potensial dari melaksanakan rekomendasi
Risiko yang terlibat
Sumber daya diperlukan
Biaya, baik yang disiapkan maupun yang sedang berlangsung
Sejarah Dokumen

Bagian ini menentukan versi terbaru dokumen dan komentar terkait seperti yang ditunjukkan pada Tabel A-5 .

Tabel A-5 . Spreadsheet Riwayat Dokumen

Table A-5. The Document History Spreadsheet

Date	Author	Version	Status	Description	Sections Affected

LAMPIRAN B

STUDI KASUS IMPLEMENTASI KAPASITAS

PENDAHULUAN DAN RUANG LINGKUP

Sekarang mari kita coba untuk menerapkan konsep yang telah Anda pelajari melalui studi kasus dan melihat bagaimana proses perencanaan kapasitas yang kuat membantu mencapai pengurangan biaya dan kepatuhan terhadap SLA.

Untuk studi kasus ini, kami akan menyederhanakan sedikit, mencakup pengaturan e-niaga online dari pengecer online fiktif KedaiKU, dan mengesampingkan aplikasi lain.

KedaiKU adalah peritel teknologi multinasional, produk hiburan, dan layanan dengan komitmen untuk pertumbuhan dan inovasi. Keluarga merek dan kemitraan KedaiKU secara kolektif menghasilkan lebih dari $ 2 miliar dalam pendapatan tahunan. Merek KedaiKU tersedia untuk pelanggan melalui lokasi ritel, pusat panggilan, dan situs web, solusi di rumah, pengiriman produk, dan kegiatan masyarakat. Tempat-tempat ini didukung oleh strategi ekspansi yang agresif di seluruh dunia. Saat ini, KedaiKU tertarik pada skenario yang akan memungkinkan pertumbuhan di masa depan, adopsi cloud, optimasi infrastruktur TI, dan mengurangi total biaya kepemilikan (TCO). Inventaris infrastruktur TI milik KedaiKU mencakup sekitar 2.000 perangkat pengguna akhir, jaringan, dan pusat data. KedaiKU's IT dijalankan pada 200 server high-end.

VISI TEKNOLOGI

Dalam hubungannya dengan visi bisnis dan strategi ekspansi ini, strategi adopsi skalabilitas infrastruktur TI dari KedaiKU adalah kunci keberhasilan TI-nya. Perencanaan kapasitas dan manajemen, bersama dengan adopsi cloud, adalah area proses yang mereka harapkan akan memungkinkan TI yang efisien dan responsif untuk mencapai pengurangan biaya dan skalabilitas. Dua tujuan utama adalah

- Memanfaatkan metode baru dan teknologi terobosan
- Menyelaraskan IT dengan bisnis dan memastikan ekspansi global

PEMANFAATAN INFRASTRUKTUR SAAT INI DAN IKHTISAR BIAYA

Dengan meningkatnya skala operasi bisnis dan strategi ekspansi secara keseluruhan, ada persyaratan segera untuk skalabilitas TI. Karena peningkatan penjualan, pencitraan merek, dan pemasaran, bisnis ini menembus segmen pasar baru dan pelanggan. Juga, input permintaan bisnis yang diantisipasi menyarankan penggunaan kapasitas masa depan yang terkait dengan layanan yang ditawarkan. Selain itu, perhatian segera diperlukan untuk menangani lonjakan musiman yang naik sebesar 10 kali penggunaan normal.

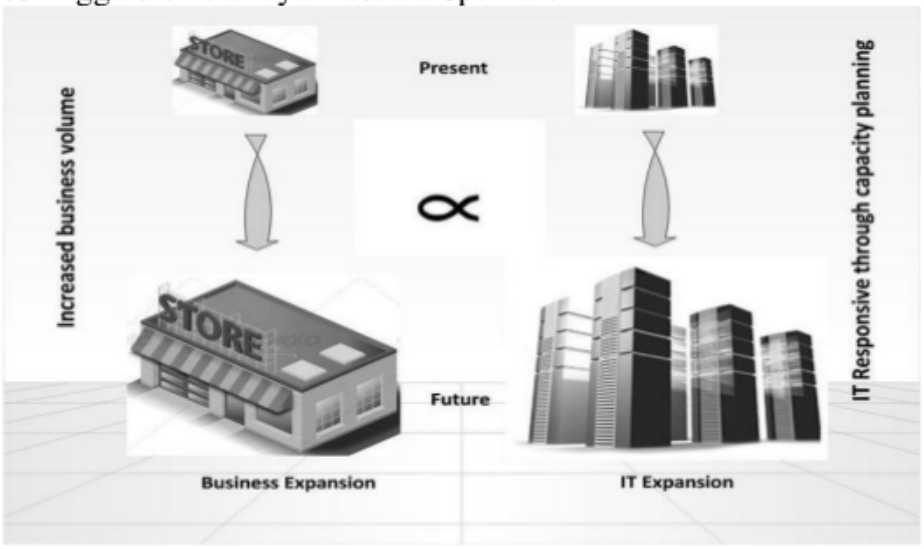
Dalam beberapa tahun ke depan, KedaiKU mengharapkan bisnis untuk tumbuh secara eksponensial dan ini didukung oleh pemasaran yang agresif dan kampanye branding secara global. TI karena itu harus merencanakan dan mengelola persyaratan bisnis menggunakan kapasitas TI yang optimal. Persyaratan kapasitas bisnis ditentukan berdasarkan hal berikut: Inventaris server KedaiKU saat ini terdiri dari 200 server, terutama Xeon 1270 quad core CPU, 8 GB RAM, 750 SATA HDD, 3,2 GHz yang mendukung lebih dari 10.000 pengguna bersamaan pada beban puncak. Server ini berjalan di pusat data penyedia hosting.

Dalam analisis awal, titik data utama yang keluar adalah tingkat pemanfaatan server. Tingkat utilisasi server rata-rata saat ini adalah 55%, sementara utilisasi beban puncak sekitar 95%, memastikan kinerja dan waktu kerja. Namun, dengan meningkatnya permintaan dan musim liburan semakin dekat, KedaiKU pasti perlu meningkatkan kapasitas untuk memenuhi peningkatan permintaan. KedaiKU dihadapkan pada pilihan untuk membeli kapasitas atau pindah ke penyedia cloud untuk skala melalui pendekatan cloud hybrid yang dapat memenuhi tantangan berikut:

- Persyaratan kapasitas total harus memenuhi sasaran kinerja saat ini, kebutuhan masa depan, dan tujuan ketersediaan / pemulihan semua aplikasi dan layanan.
- Pemeriksaan masa depan akan dilakukan dengan menyiapkan kapasitas siaga selain apa yang saat ini digunakan.
- Ketersediaan / pemulihan biasanya diaktifkan melalui redundansi dan tingkatan pusat data. Redundansi ini bisa:

    - Redundansi komponen
    - Sumber daya penuh redundansi
    - Redundansi data
- Jika permintaan melebihi kapasitas yang tersedia untuk pertumbuhan yang direncanakan dan / atau tidak meninggalkan kapasitas yang berlebihan, maka akan membahayakan tingkat layanan.
- Kapasitas fisik tambahan perlu ditambahkan atau beban kerja harus dikeluarkan dari kolam kapasitas yang tersedia.

Gambar B-1 menggambarkan roadmap kapasitas TI saat ini dan masa depan sehingga bisnis dilayani secara optimal.



***Gambar B-1** . IT / ekspansi bisnis*

IT IMPERATIVES

- IT imperatif jangka panjang termasuk
- Efektivitas biaya dalam ekspansi TI untuk memenuhi permintaan bisnis di masa depan
- Pemahaman yang jelas tentang tuntutan pada layanan dan rencana masa depan untuk pertumbuhan atau penyusutan beban kerja
- Perencanaan kapasitas sesuai dengan tujuan bisnis secara keseluruhan
- Komitmen untuk memenuhi atau melampaui SLA yang disetujui
- Tingkat pemanfaatan infrastruktur yang dioptimalkan
- Analisis kinerja data pengukuran, termasuk analisis dampak dari layanan baru pada kapasitas saat ini
- Memantau sumber daya dan kinerja sistem, pemanfaatan sistem, batas kapasitas, dan kebutuhan kapasitas yang diharapkan, dan merekam informasi tersebut dalam sistem informasi manajemen kapasitas (CMIS).
- Penyetelan kinerja kegiatan untuk memastikan penggunaan infrastruktur yang paling efisien
- Metode formal untuk proyeksi kapasitas tepat waktu untuk dimasukkan dalam proses perencanaan anggaran tahunan KedaiKU

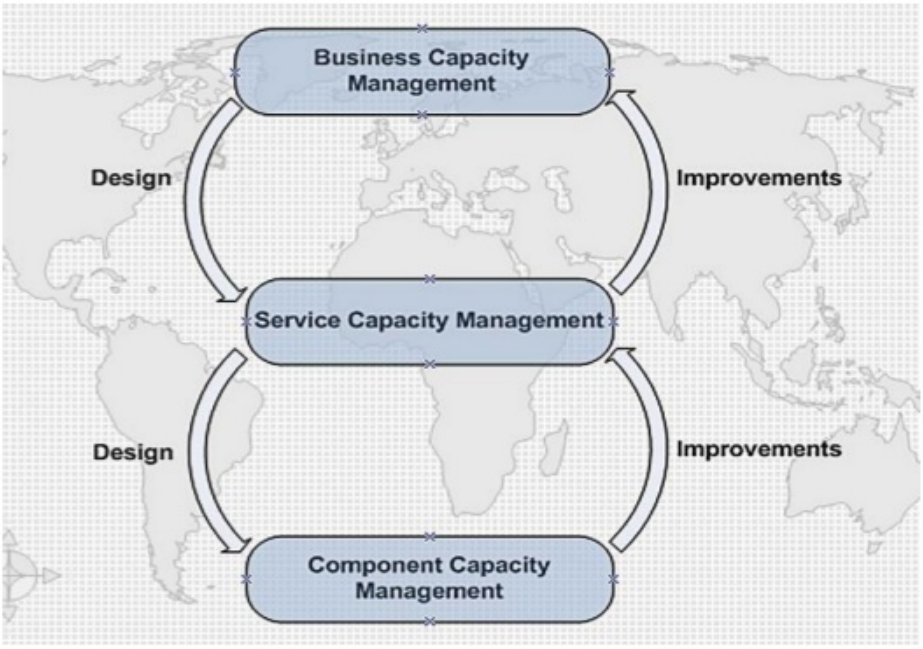
KEBUTUHAN PERENCANAAN KAPASITAS

Kegiatan perencanaan kapasitas memastikan bahwa kapasitas layanan TI dan infrastruktur TI mampu memberikan tingkat layanan yang disepakati dengan cara yang hemat biaya. Semua sumber daya harus dipertimbangkan untuk memenuhi ramalan permintaan, yang didasarkan pada persyaratan bisnis.

KedaiKU menghargai peran dan pentingnya proses manajemen kapasitas yang efisien untuk ekspansi bisnis yang mulus. Ini memastikan bahwa peningkatan kapasitas diimplementasikan sebelum batas keamanan tingkat layanan kapasitas dilanggar. Dengan peran ini dalam perusahaan, TI juga mencari transformasi dari pemasok ke agregator layanan. Sebagai agregator layanan, TI melayani bisnis dengan solusi TI terbaik dan hemat biaya. Dengan demikian, transformasi TI KedaiKU adalah tulang punggung dari keseluruhan strategi TI yang mendukung bisnis.

Ini membutuhkan peninjauan persyaratan kapasitas sebagai bagian dari siklus perencanaan bisnis normal dan manajemen permintaan yang efektif.

Semua persyaratan terkait kapasitas mencari terjemahan dari bisnis ke layanan ke tingkat komponen (Gambar B-2). Persyaratan tingkat bisnis seperti peningkatan penjualan, musim, permintaan lonjakan, dll. Perlu ditangkap dan implikasi untuk kapasitas layanan perlu diturunkan dari mereka. Demikian juga, persyaratan kapasitas komponen perlu diformulasikan dan disusun.



***Gambar B-2** . Lapisan manajemen kapasitas yang sudah familiar*

MENERAPKAN MANAJEMEN KAPASITAS

Sub bagian berikut menguraikan langkah-langkah untuk menerapkan manajemen kapasitas.

LANGKAH 1: TENTUKAN PERSYARATAN KAPASITAS DARI TOOLSET PEMANTAUAN YANG ADA

Alat pemantauan kinerja saat ini dan tinjauan kapasitas menyediakan informasi tentang pemanfaatan saat ini dan persyaratan pengoptimalan. Persyaratan ini dikumpulkan dari alat pemantauan elemen melalui alarm, laporan ambang batas, kejadian, insiden, dll. Toolset lain yang berada di atas sistem pemantauan dasar memberikan informasi yang berarti mengenai kapasitas.

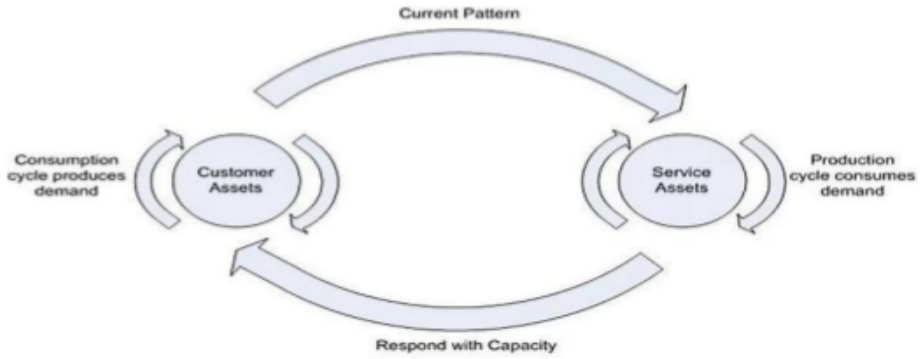
Kebutuhan kapasitas untuk semua sistem baru diambil dari toolset yang ada di tempat untuk menentukan komputer yang diperlukan dan sumber daya jaringan yang diperlukan, ukuran sistem baru tersebut, dengan mempertimbangkan penggunaan perangkat keras, persyaratan untuk ketahanan, kinerja tingkat layanan, dan biaya.
Setelah menjalankan prosedur penemuan di CMDB dan menganalisanya, snapshot informasi dalam Tabel B-1 diambil pada lingkungan saat ini.

Tabel B-1 . Snapshot Lingkungan

Users	Servers	CPU Cores/Server	Avg. Utilization	Memory/Server
10000	200	4	55%	8

PERMINTAAN KAPASITAS KOPLING

Manajemen permintaan memperkirakan kebutuhan masa depan, dan terjemahan ini menyiratkan peningkatan 10% dalam server setiap tahun. Ini dilakukan dengan menggunakan pola analisis aktivitas bisnis dan metode manajemen permintaan lainnya. Gambar B-3 menjelaskan hubungan antara kapasitas dan permintaan. Permintaan membantu dalam merumuskan kebutuhan kapasitas masa depan.



Gambar B-3 . *Permintaan / kapasitas kopling*

- Analisis permintaan termasuk
- Permintaan selama waktu hari / minggu / bulan / tahun
- Variasi musiman
- Hari-hari khusus (musim liburan)

- Permintaan dari zona waktu yang berbeda
- Permintaan untuk lokasi / pusat data yang berbeda

Masukan untuk menentukan jumlah pengguna telah diambil dari ujung depan KedaiKU *'s* portal web. Infrastruktur TI milik KedaiKU saat ini rata-rata melayani 10.000 pengguna sekaligus, mengelola beban kerja beragam pengguna yang memukul portal web, dan mempertahankan waktu respons.
Manajemen permintaan yang efisien dari KedaiKU menangani semua ramalan layanan di masa mendatang. Variasi musiman yang tinggi selama festival memiliki implikasi langsung pada IT KedaiKU. Untuk mengelola permintaan ini, 200 server dikerahkan dalam model hosting server. Selama masa puncak, seperti yang disebutkan sebelumnya, penggunaan server meningkat hingga 95% selama satu bulan. Sisa tahun, kisaran pemanfaatan server adalah 45-55%. Rata-rata, ada peningkatan 10% dalam inventaris server dari tahun ke tahun untuk memenuhi peningkatan permintaan tahunan secara teratur.

KedaiKU mengharapkan skenario permintaan berikut:

- Musim liburan selama 1 bulan: 10x pengguna di situs web
- Pertumbuhan lalu lintas tahunan di situs web: 10% tahun ke tahun pertumbuhan

Angka-angka ini diperoleh menggunakan manajemen permintaan dan proses manajemen kapasitas yang kuat bersama dengan implementasi alat dan analisis data.

Oleh karena itu, berikut ini disimpulkan:

- Kapasitas komputasi yang dibutuhkan untuk 1 bulan sibuk: 2000 server (x10 kali)
- Hitung peningkatan kapasitas dalam server 1: 20 tahun (10% peningkatan)
- Hitung peningkatan kapasitas pada server 2: 22 tahun (10% peningkatan)

BIAYA KAPASITAS

Sekarang, untuk mempertahankan kinerja dan mengakomodasi permintaan, persediaan server KedaiKU membutuhkan kapasitas 10x selama musim permintaan tinggi.

Anggaran keuangan mendorong belanja TI pada peningkatan kapasitas. Anggaran keuangan bersama dengan tuntutan masa depan merupakan faktor penentu untuk kebutuhan kapasitas masa depan. KedaiKU sedang mengevaluasi opsi baik memilih untuk infrastruktur khusus atau beralih ke model cloud hibrida.
KedaiKU harus menghitung kapasitas dari tahun ke tahun untuk peningkatan biaya menggunakan salah satu dari dua opsi.

Opsi 1: Biaya infrastruktur khusus
Opsi 2: Biaya penyedia cloud

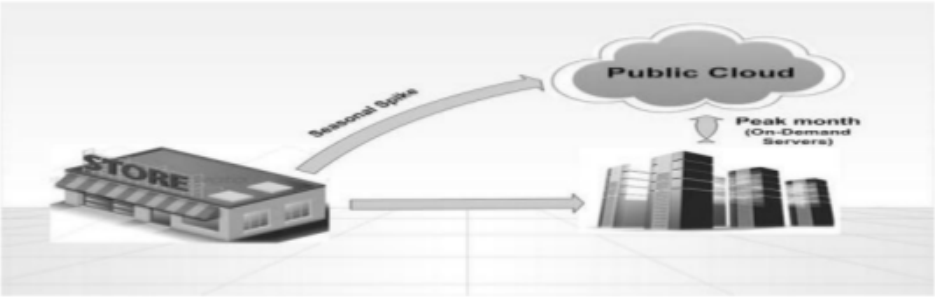
OPSI 1: INFRASTRUKTUR KHUSUS
Biaya yang dikeluarkan dalam infrastruktur khusus mencakup elemen-elemen berikut:

- Ruang pusat data
- Kekuasaan
- Rak
- Menghitung
- Jaringan
- Bandwidth
- Penyimpanan
- Mempersiapkan
- Pemeliharaan dan manajemen

Lingkungan yang disebutkan di atas termasuk menjalankan 1800 server dengan berbagai tingkat pemanfaatan sebagaimana disebutkan sebelumnya untuk mengakomodasi permintaan. Server dengan konfigurasi Xeon 1270 quad core, 8GB RAM, 750 SATA HDD, 3,2 GHz prosesor dari penyedia hosting saat ini datang ke $ 13.000.000 setiap tahun untuk 1.800 server. Elemen biaya tambahan yang dikeluarkan termasuk manajemen, pemantauan, biaya pemeliharaan, dll.

OPSI 2: PENYEDIA CLOUD
Pada penyedia cloud mengusulkan model cloud hybrid, untuk bulan-bulan sibuk, persyaratan server tambahan 1800 dipenuhi dari cloud publik (yaitu, AWS).

Dalam skenario ini, seperti yang ditunjukkan pada Gambar B-4 , KedaiKU sedang mempertimbangkan penggunaan 200 server yang ada dan tambahan 1.800 dari cloud publik, dan biaya ditampilkan pada Tabel B-2 .



Gambar B-4 . *Skalabilitas saat puncak*

Tabel B-2 . Snapshot Lingkungan

Users	Servers	CPU/User	Utilization	Memory/Server	Cloud cost $
90000	1800	4	95%	8	4,000,000

Sekarang, termasuk manajemen cloud, pemantauan, ketersediaan tinggi, 99,9 SLA, dan otomatisasi lainnya, total biaya server berbasis cloud ternyata menjadi $ 4.000.000.

Hasil dari analisis biaya ini diumpankan ke proses manajemen keuangan untuk menetapkan anggaran tahunan, dan berdasarkan kelayakan finansial, perencanaan dan penganggaran, dana dialokasikan untuk solusi terbaik.

TARGET KINERJA

Target kinerja adalah target yang ditetapkan melalui pengujian otomatis dan alat pemantauan tingkat layanan. Alat-alat ini secara otomatis mengumpulkan persyaratan terkait kinerja untuk menetapkan ambang baru secara dinamis. Ini membantu dalam menetapkan tolok ukur kinerja baru dan SLA. Alat-alat ini membantu dalam menentukan langkah-langkah dan persyaratan dalam hal waktu respons dan tindakan terkait kinerja lainnya.

LANGKAH 2: DESAIN UNTUK KAPASITAS

Pendekatan hybrid cloud menghasilkan pemanfaatan kapasitas yang ada dan memanfaatkan cloud untuk beban kerja puncak. Pertimbangan berikut dibuat saat membangun pendekatan kapasitas:

- Penghapusan kapasitas yang terfragmentasi dapat menyebabkan inefisiensi yang tinggi dan bahkan menggandakan infrastruktur yang diperlukan untuk menampung beban kerja.
- Untuk memastikan alokasi cerdas mesin virtual.
- Untuk mempertimbangkan toolset yang mampu mendefinisikan aturan teknis, bisnis, dan kepatuhan untuk penempatan beban kerja. Aturan-aturan penempatan beban kerja ini dikonfigurasi dalam toolset manajemen siklus hidup cloud di lapisan manajemen.
- Rule engine digunakan untuk menjamin kesehatan dan ketepatan solusi manajemen kapasitas.
- Mengikuti prinsip arsitektur yang tangguh, lincah, dan terukur.
- Manajemen kapasitas mampu menetapkan kebutuhan terkait aplikasi seperti multi-tenancy, persyaratan tingkat layanan terkait, persyaratan infrastruktur, persyaratan skalabilitas, ruang disk, kapasitas komputasi, kebutuhan bandwidth dan bandwidth jaringan, dll. Langkah-langkah ini membantu dalam merencanakan kebutuhan sistem masa depan dan arsitektur lingkungan.

MENETAPKAN ARSITEKTUR KAPASITAS

Berikut ini adalah pertimbangan untuk membangun arsitektur kapasitas KedaiKU. Panduan arsitektur ini terutama dipertimbangkan untuk memenuhi kebutuhan bisnis dinamis KedaiKU.

- Skalabilitas suatu aplikasi memungkinkan untuk ditingkatkan untuk mengakomodasi pertumbuhan. Arsitektur berbasis lapisan untuk manajemen kapasitas harus ditetapkan dengan serangkaian protokol yang ditetapkan di antara berbagai lapisan. Di bagian bawah arsitektur berlapis ini adalah lapisan virtualisasi. Ini bisa berupa server, jaringan, atau lapisan virtualisasi penyimpanan.
- Alat digunakan untuk memantau lapisan virtualisasi dan kelompok virtual yang mendasari, peternakan, dan mesin termasuk server, jaringan, dan penyimpanan.

- Pada lapisan ini, profil mesin virtual dapat dilakukan, yang membantu dalam prosedur yang sedang tren. Ini adalah lapisan dasar dan data permintaan, dalam hal tingkat penggunaan dan pemanfaatan, dan diambil dari lingkungan virtual.
- Detail penggunaan seperti CPU, memori, ruang disk, disk I / O bandwidth, dan bandwidth I / O jaringan diperhitungkan oleh lapisan analisis penggunaan yang berada di atas lapisan pemantauan.
- Tren kapasitas masa depan dihitung berdasarkan analisis statistik.
- Pada lapisan ini, model kapasitas meramalkan perilaku infrastruktur menggunakan permintaan, keuangan, operasional, kinerja aplikasi, dan data terkait vendor. Ukuran aplikasi memperkirakan kebutuhan sumber daya untuk mendukung aplikasi untuk memastikan bahwa SLA terpenuhi. Ukuran aplikasi juga membantu dalam mengidentifikasi konsumsi sumber daya dan implikasi biaya untuk aplikasi baru atau yang diubah dan efeknya pada aplikasi terkait lainnya.

Terapkan Teknik Kapasitas

Teknik-teknik ini dapat digunakan untuk memodelkan dan mengoptimalkan desain kapasitas.

- Sebagian besar teknik kapasitas didasarkan pada metode ilmiah dan aplikasi teknologi pemodelan prediktif didukung oleh algoritma antrian teori terbukti.
- Simulasi kegiatan virtualisasi harus dilakukan untuk mengidentifikasi kandidat, target, dan penempatan optimal terbaik (misalnya, sesuai dengan beban kerja yang kompatibel) dengan memperhatikan kriteria teknis, geografis, bisnis, dan kepatuhan.
- Simulasi dapat menggambarkan bagaimana skala layanan dari lingkungan pengujian ke lingkungan tingkat produksi dengan menggunakan hasil pengujian beban.
- Simulasi perubahan infrastruktur (misalnya, skala horizontal atau vertikal atau kegagalan) dan skenario perubahan bisnis dilakukan (misalnya, tren bisnis dan rencana pemasaran.)

LANGKAH 3: MENGHASILKAN RENCANA KAPASITAS

Rencana kapasitas, berdasarkan permintaan bisnis yang ada dan di masa depan, dikembangkan dan didistribusikan. Ini termasuk informasi tentang

- SLA
- Spesifikasi aplikasi
- Infrastruktur saat ini
- Skenario tugas pengguna
- Matriks beban tugas
- Pemantauan dan metrik
- Prakiraan
- Tingkat layanan, dll.

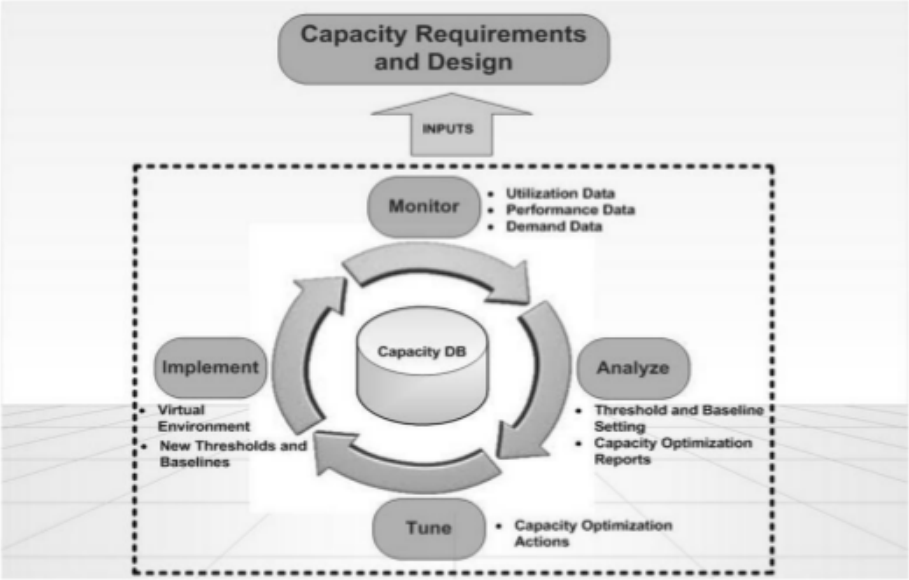
LANGKAH 4: PENGELOLAAN KAPASITAS BERKELANJUTAN

Selama operasi, data (seperti pemanfaatan server, dll.) Diambil dari berbagai alat yang mendasari dan sumber data terkait. Data ini menyediakan manajemen kapasitas dengan laporan kesehatan proses dan indikator kinerja utama lainnya. Sumber informasi ini kemudian dapat digunakan untuk menyediakan portal pelaporan terpadu untuk membantu dalam pemantauan kapasitas dan perencanaan untuk cloud, layanan, dan komponen yang mendasarinya.

Solusi kapasitas saat ini juga untuk pemantauan end-to-end dan analisis lingkungan virtual. Pemantauan dan analisis penggunaan infrastruktur oleh pelanggan adalah dua bidang utama dari manajemen kapasitas iteratif, diikuti dengan penyetelan dan implementasi. Ini adalah siklus pengoptimalan kapasitas yang sedang berlangsung.

Alat Manajemen Kapasitas Berkelanjutan yang Sebenarnya

Penyedia cloud mengimplementasikan rangkaian manajemen operasi VMware vCenter berdasarkan praktik terbaik kapasitas untuk operasi yang sedang berlangsung dan manajemen kapasitas. Seperti yang ditunjukkan pada Gambar B-5 , alat ini memastikan semua aktivitas manajemen kapasitas yang sedang berlangsung dilakukan secara efisien. Sekarang, dengan perangkat analisis dan peramalan yang tepat, manajemen kapasitas berkelanjutan terdiri dari empat kegiatan pengoptimalan utama.



Gambar B-5 . *Kerangka manajemen kapasitas berkelanjutan*

- Implementasi: Menerapkan kapasitas dan mendukung teknik manajemen virtualisasi seperti DRS, pengelompokan, balon memori, penskalaan, dll.
- Pemantauan: Pemantauan pemanfaatan dan kinerja
- Analisis: Menganalisis data kapasitas untuk tren dan perkiraan
- Penyelarasan: Tindakan pengoptimalan untuk meningkatkan pemanfaatan dan kinerja sumber daya

Kegiatan ini memberikan informasi historis dasar dan pemicu yang diperlukan untuk semua kegiatan dan proses lain dalam manajemen kapasitas. Monitor didirikan pada semua komponen dan untuk masing-masing layanan.

Data dianalisis menggunakan sistem pakar untuk membandingkan tingkat penggunaan terhadap ambang batas sedapat mungkin. Hasil analisis dimasukkan dalam laporan, dan rekomendasi dibuat sebagaimana mestinya.

Mekanisme kontrol disiapkan untuk bertindak berdasarkan rekomendasi. Ini dapat berupa layanan penyeimbang, menyeimbangkan beban kerja, mengubah tingkat konkurensi, dan menambah atau menghapus sumber daya.

Semua informasi yang terkumpul selama kegiatan ini disimpan dalam database manajemen kapasitas dan siklus (implementasikan, pantau, analisa, penyetelan) kemudian mulai lagi, pantau setiap perubahan yang dilakukan untuk memastikan

mereka memiliki efek yang menguntungkan dan mengumpulkan lebih banyak data untuk tindakan di masa mendatang. .
Penyedia solusi cloud mengimplementasikan platform VMware untuk virtualisasi dan kapasitas suite manajemen mendukung fitur-fitur berikut untuk manajemen lingkungan yang efisien dan efektif, optimalisasi, pelacakan, pemanfaatan sumber daya, gerakan beban kerja, dll.:

- Scaling
- Clustering
- Penyeimbang beban
- Balon memori
- Swapping
- Penjadwalan sumber daya terdistribusi
- Orkestrasi awan
- Pengelolaan awan hybrid

Untuk pemantauan, analisis, dan penyetelan berkelanjutan, Manajer Operasi vCenter mengumpulkan, menganalisis, dan menyajikan semua persyaratan kinerja dan pemantauan yang terkait. Ini ditunjukkan pada Gambar B-6 . Ini memastikan operasi otomatis dan penggunaan manajemen analitik dan pendekatan terpadu untuk kinerja, kapasitas, dan manajemen konfigurasi. Solusi VMware menyediakan manajemen, kontrol, dan visibilitas yang efektif atas lingkungan cloud KedaiKU, memungkinkan kecerdasan yang dapat ditindaklanjuti dengan lebih baik secara proaktif.

Gambar B-6 . *Contoh grafik Manajer Operasi VMware*

LANGKAH 5: TINJAUAN KAPASITAS

Semua laporan kapasitas dalam CDB diekspor ke area proses lain dalam format dan tampilan yang disesuaikan. Berdasarkan laporan konsolidasi dan basis data kapasitas yang diperbarui, pengambilan keputusan strategis didukung. Laporan termasuk

- Kapasitas keseluruhan tersedia
- Kapasitas keseluruhan digunakan
- Kapasitas keseluruhan tidak terpakai
- Laporan perkiraan untuk kapasitas cloud tersedia / digunakan
- Prakiraan untuk permintaan
- Pusat data / laporan lokasi
- Prakiraan yang kurang dimanfaatkan / terlalu banyak digunakan
- Yield per unit cloud
- Penggunaan kapasitas oleh jenis mesin virtual

KESIMPULAN

Dengan menerapkan solusi manajemen kapasitas dan adopsi cloud, KedaiKU menyadari manfaat langsung dan terukur berikut ini:

- Pengurangan biaya hingga 65% membantu KedaiKU dengan penghematan biaya yang besar
- Meningkatkan kelincahan TI melalui pengaktifan cloud
- Pemanfaatan sumber daya yang optimal dengan menerapkan praktik terbaik dan peralatan manajemen kapasitas
- Peningkatan SLA melalui akses layanan sesuai permintaan
- Mengurangi biaya pemeliharaan dan persediaan tercatat

Manfaat lain seperti peningkatan produktivitas staf, peningkatan efisiensi, dan berkurangnya waktu respons direalisasikan pada waktunya dengan menerapkan proses manajemen kapasitas cloud yang efisien.

LAMPIRAN C

ALAT KAPASITAS

PENGUKURAN, MONITORING, DAN MANAJEMEN INFORMASI ALAT DAN PANDUAN RENCANA KAPASITAS ANDA. DI LAMPIRAN INI ,telah disusun daftar beberapa alat dan utilitas yang lebih populer untuk referensi Anda. Kami menggunakan banyak alat ini di Flickr, dan beberapa di antaranya adalah perangkat lunak sumber terbuka yang setara yang telah ditulis dalam Yahoo! untuk mencapai tujuan yang sama.

PEMANTAUAN

Beberapa alat berikut memiliki kemampuan memperingatkan, beberapa dari mereka lebih fokus pada grafik dan koleksi, dan beberapa memiliki keduanya.

SISTEM PENGUMPULAN METRIK DAN PEMBERITAHUAN PERISTIWA

- *Ganglia, http://ganglia.info*
  Lahir dari komunitas HPC, Ganglia memiliki komunitas pengguna dan pengembang yang sangat aktif. Kami menggunakan Ganglia secara ekstensif di Flickr, seperti halnya Wikipedia dan situs jejaring sosial berskala besar lainnya.
- *Nagios, http://nagios.org*
  Nagios di Yahoo! Untuk memonitor layanan di ribuan mesin.
- Cacti, http://cacti.net
- Zabbix, http://zabbix.com
- Hyperic HQ, http://hyperic.com
- Munin, http://munin.projects.linpro.no/
- ZenOSS, http://www.zenoss.com/
- OpenNMS, http://opennms.org
- *GroundWork, http://www.groundworkopensource.com/*

- GroundWork adalah hibrida dari Nagios dan Ganglia.
- Monit, http://www.tildeslash.com/monit
- *Reconnoiter, https://labs.omniti.com/trac/reconnoiter*

PENGUKURAN AD HOC DAN ALAT GRAPHING

- *RRDTool, http://oss.oetiker.ch/rrdtool/*
  Alat penyimpanan grafik dan metrik.
- *Collectd, http://collectd.org/*
  Scalable daemon pengumpulan statistik sistem. Menggunakan multicast, seperti Ganglia.
- *Rrd2csv, http://www.opennms.org/index.php?title=Rrd2csv*
  RRD ke konverter csv.
- *Dstat, http://dag.wieers.com/home-made/dstat/*
  Alat statistik sistem, modular.
- *GraphClick, http://www.arizona-software.ch/graphclick/*
  Digitizer yang membangun data dari gambar grafik — berguna saat Anda memiliki gambar tetapi bukan data mentah.

INSTALASI OS OTOMATIS

- *SystemImager, http://wiki.systemimager.org/*
  SystemImager berasal dari komunitas HPC dan digunakan untuk menginstal cluster komputer ribu-node. Digunakan oleh banyak operasi web skala besar juga. Pekerjaan yang menarik telah dilakukan untuk menggunakan bittorrent sebagai mekanisme transfer.
- *FAI, http://www.informatik.uni-koeln.de/fai*
  Sebuah auto-Debianalat instalasi dengan komunitas yang sehat.
- KickStart, http://fedoraproject.org/wiki/Anaconda/Kickstart/
- *http://cobbler.et.redhat.com*
  proyek yang relatif baru dari RedHat, mendukung RedHat, Fedora, dan CentOs.

MANAJEMEN KONFIGURASI

- Puppet*, http://reductivelabs.com/trac/puppet*

Cepat menjadi alat konfigurasi yang sangat populer, Wayang memiliki beberapa pengembang yang sangat bersemangat dan komunitas pengguna yang sangat terlibat. Ditulis dalam Ruby.

- *Cfengine, http://www.cfengine.org/*
  Ditulis dalam bahasa C, sudah ada selama bertahun-tahun dan memiliki basis instalasi dan komunitas aktif yang besar.
- Bcfg2, http://trac.mcs.anl.gov/projects/bcfg2
- Lcfg (sistem konfigurasi Unix berskala besar), http://www.lcfg.org/

MANAJEMEN CLUSTER

- *Capistrano, http://www.capify.org/*
  Ditulis dalam Ruby, Capistrano menjadi populer di lingkungan Rails.
- Dsh, http://freshmeat.net/projects/dsh/
- Fabric, http://savannah.nongnu.org/projects/fab
- *Func, https://fedorahosted.org/func/*
  Func adalah Fedora Unified Network Controller, dan dapat menggantikan perintah ad-hoc cluster-wide ssh dengan arsitektur client / server yang diotentikasi.
- XCat, http://xcat.sourceforge.net/

MANAJEMEN PERSEDIAAN

- *iClassify, https://wiki.hjksolutions.com/display/IC/Home*
  iClassify adalah sistem manajemen aset yang relatif baru, yang mendukung pendaftaran otomatis dan menyediakan kait untuk Puppet dan Capistrano.
- Inventaris OCS NG, http://www.ocsinventory-ng.org/

ANALISIS TREND DAN FITTING KURVA

- *Fityk, http://www.unipress.waw.pl/fityk/*
  GUI yang sangat bagus dan baris perintah alat pas kurva.
- *SciPy, http://www.scipy.org*
  Ilmiah dan analisis alat dan pustaka untuk Python, termasuk beberapa rutin pencocokan kurva.
- *R, http://www.r-project.org*
  Paket komputasi statistik, termasuk utilitas kurva-pas.

BUKU TENTANG TEORI ANTRIAN DAN MATEMATIKA PERENCANAAN KAPASITAS

- Gunther, Neil. *Perencanaan Kapasitas Gerilya* (Springer, 2006)
- Menascé, Daniel A. dan Virgilio AF Almeida. *Perencanaan Kapasitas Untuk Layanan Web: Metrik, Model, dan Metode* (Prentice Hall, 2001)
- Menascé, Daniel A. dkk. *Kinerja Berdasarkan Desain* (Prentice Hall, 2004)
- Menascé, Daniel A. dan Virgilio AF Almeida. *Scaling untuk E-bisnis* (Prentice Hall, 2000)

Overview Buku

Gelombang Cloud Computing mengubah cara organisasi menggunakan teknologi informasi, menyebabkan Manajer TI akan ingin menerapkan praktik terbaik manajemen kapasitas tradisional untuk adopsi dalam lingkungan komputasi cloud di mana banyak pemangku kepentingan, abstraksi tingkat tinggi, konsolidasi, virtualisasi infrastruktur TI, model biaya dinamis, dll.. Buku ini mengambil pendekatan pragmatis untuk menerapkan praktik terbaik untuk manajemen kapasitas dan yang akan membantu penyedia layanan cloud dalam merancang dan menerapkan proses manajemen kapasitas dengan cara yang paling hemat biaya.

Buku ini memberikan penekanan signifikan pada model layanan cloud dan rantai nilai di mana manajemen kapasitas dan perencanaan memiliki implikasi yang berbeda untuk berbagai pemangku kepentingan seperti pembuat layanan cloud, agregator layanan cloud, dan konsumen layanan cloud. Dengan demikian, buku ini bermaksud untuk mengajak penonton dari dasar-dasar Cloud Computingdan virtualisasi untuk memahami berbagai model Cloud Computing — dan bagaimana mengubah dan menerapkan proses perencanaan kapasitas sambil mengingat model baru.

Buku ini membahas proses manajemen kapasitas dalam dua bagian. Bagian pertama berfokus pada perencanaan kapasitas untuk layanan baru, dan bagian kedua mencakup pemantauan kinerja atau manajemen kapasitas untuk layanan berkelanjutan dalam lingkungan cloud. Bisnis yang berjalan pada model penyebaran seperti cloud harus mampu mengatasi lonjakan permintaan sehingga kapasitas infrastruktur yang hemat biaya disediakan dari waktu ke waktu untuk mendukung operasi bisnis dan untuk menjamin kelangsungan bisnis dan hasil bisnis yang diinginkan. Setelah membaca buku ini, kami berharap Anda akan mendapatkan pemahaman yang jelas tentang kegunaan potensi cloud dan model yang mungkin masuk akal bagi Anda dalam upaya Anda untuk mengelola kebutuhan kapasitas Anda secara efektif.

Cloud Computing : Manajemen dan Perencanaan Kapasitas

ORIGINALITY REPORT

3%	3%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.nurhidayat.id Internet Source	2%
2	khansadhiyasavira.wordpress.com Internet Source	1%

Exclude quotes	On	Exclude matches	< 300 words
Exclude bibliography	On		