Introduction:

- Virtualization is a key feature that supports cloud computing.
- A cloud with a more virtualized infrastructure has higher resource utilization.

1. ADMINISTRATING CLOUDS

Various suppliers possess products structured for cloud computing management, such as OpenQRM, Managed Methods, VMware, and Cloud Kick, together with established names such as CA, BMC, IBM, Tivoli, and HP.

The main features are provided by the main cloud substructure managing products which are as follows:

1. Almost all of these assist in designing and furnishing new items and in eliminating useless items.
2. Almost all offer a common set of reports on status such as response time, used quota, uptime, etc.
3. Almost all of these assist distinct clouds forms (often stated as hybrid clouds).

Few suppliers present comprehensive means in handling metrics and managing provisioning in hybrid environments and they are Zeus, Morph, RightScale Kaavo, and Scalr.

Cloud suppliers offer certain options to meet the first and second standard, like CloudWatch by Web Services of Amazon.

HP company Open View, also known as Operating Manager, can administer cloud-based servers, but in the same way as administered by some other server.

- At present, the company BMC offers:
  1. Ways to **monitor and manage** the cloud
  2. **Cloud supervising** and **cloud service administration** for suppliers
  3. **New information** on cloud management
  4. **Licensing and pricing for tools** of cloud management

RightScale: RightScale is one of the main suppliers. RightScale is classified as follows:

1. Cloud management environment:

(a) Multi-cloud engine

(b) Cloud-ready server template and best practice deployment library

(c) Adjustable automation engine

2 Cooperative cloud tips from RightScale:

(a) Initiation of server in the cloud by means of Server Templates

(b) Organization of servers with RightScale deployments

(c) Scalable batch processing by way of RightGrid automation

- It is intended to turn a user during the **introductory process of migrating** the cloud by the means of **library and templates.**
- The environments are managed with the help of the **management environment tool**, specifically continuing manufactures and guaranting the **availability of resource**.

Kaavo: The product is used:

1. To handle demand variations through automatically adding up or eliminating resources.

2 For single-click of complicated multi-tier appliances in the cloud

3.For **ciphering preserved data** in the cloud

4.To **manage run-time environment** and **automation of workflows** without human interference

5. For run-time organization of **relevant infrastructure** within the cloud

- The essential product of Kaavo is known as **IMOD**.
- **Provisioning, changes and configuration** to the environment of the cloud has been **controlled by IMOD** and also in the **hybrid model across various suppliers**.

Zeus:

- The **durable and trustworthy** Web server made Zeus famous.
- To test the **availability, conventional load balancing tools** were used by it which **created or terminated supplementary insistence in the cloud**, hence supplying **tool provisioning**.

Scalr:

- It builds **vibrant clusters** similar to that of RightScale and Kaavo.
- It also assists in **upsizing and downsizing**, depending upon the **custom building of images**, **traffic demands,** and **snapshots** for every server or server type.

Morph:

- It refilling the **provisioning space and management** as an application.
- It is intended at the **business enterprise** looking to **install a private cloud.**
- It also assists **numerous virtual machines**.
- Morph provides **customized and stylish platforms** by **providing guidance at each and every step.**

CloudWatch:

- Amazon CloudWatch is a service that provides customers' facilities to **monitor applications, systems** and also provides you with easy solutions by giving you the **facility to collect metrics, logs, and events** in a real-time system about a **resource**.
- In CloudWatch for EC2, a central management console, It manages all **load balancing, monitoring, and dynamic provisioning** known as auto-scaling.
- Automatic scaling is generally used in cloud computing, where the **total amount of resources in a server is measured** in terms of the **number of active servers** and **scaling is done on the basis of load** at a particular instant of time.

2. CLOUD MANAGEMENT PRODUCTS

- The framework of cloud infrastructure is made up of the following components:
- Physical infrastructure
- Virtual infrastructure
- Applications and platform software
- Cloud infrastructure management and service creation tools

Physical Infrastructure:

- Physical infrastructure includes **physical IT resources** which comprise **physical network components such as switches, routers, physical adaptors, physical servers, and storage systems.**
- Physical servers are **linked with one another**, to the **storage systems** and to the **clients.**
- Cloud service providers may use **physical IT resources** from **one or more data centers** to provide services.
- The connectivity facilitates **data centers at different locations** to **work as a single unit** and **resources can also be shared** as per user requirement.
- This enables both migration of cloud services across data centers and provisioning of cloud services using resources from multiple data centers.

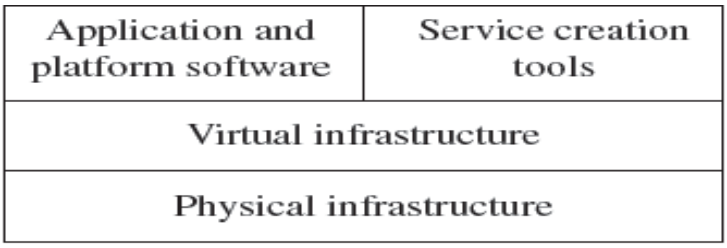

Fig. 9.1 Framework of cloud infrastructure

Virtual Infrastructure:

Virtual infrastructure consists of the following resources:

- Resource pools such as **CPU** pools, **memory** pools, **network bandwidth** pools and **storage** pools
- Identity pools such as **VLAN ID** pools, **VSAN ID** pools and **MAC address** pools
- Virtual IT resources consist of the following:

(a) VMs, virtual volumes, and virtual networks

(b) VM network components such as virtual switches and virtual NICs

- IT resources gain capacities such as CPU cycles, memory, and network bandwidth and storage space from the resource pools.

Applications and Platform Software: It includes a suite of software such as the following:

1. Business applications
2. Operating systems and database
3. Software required to build environments for running applications
4. Migration tools

- Applications and platform software are hosted on VMs to create Software as a Service (SaaS) and Platform as a Service (PaaS).
- For SaaS, **applications and platform software** are provided by cloud service providers.
- For PaaS, only the **platform software** is provided by the cloud service providers.

- In infrastructure as a service (IaaS), consumers upload both applications and **platform** software to the cloud.
- Cloud service providers supply **migration tools** to consumers, **enabling deployment of their applications and platform software** to the cloud

Cloud Infrastructure Management and Service Creation Tools:

- Cloud infrastructure management and service creation tools are responsible for **managing physical and virtual infrastructure**.
- They provide cloud services based on **consumer requests** and allow consumers to use the services.
- Cloud infrastructure management and service creation tools **automate consumer requests processing** and **creation of cloud services**.
- They also provide **administrators** a single management interface to **manage resources** distributed in multiple **virtualized data centers (VDCs)**
- Virtual Infrastructure Management software provides **interfaces** to **construct virtual infrastructure** underlying physical infrastructure.
- It enables **communication with tools**, such hypervisors and physical switch operating systems and also configuration of pools and virtual resources with the assistance of these tools.
- In a VDC, compute, storage and network are used. By using distinct virtual infrastructure management software, **resources are organized separately**.
- **Physical servers and networks are handled separately** using compute management software and networks appropriately

3. PROCESSES IN CLOUD SERVICE MANAGEMENT

- Cloud service administration comprises a set of organizational procedures that connect the distribution of cloud services to customers.
- Cloud service management procedures perform in the environment to make certain all the service tasks are carried out as committed.
- Cloud service suppliers should make use of appropriate service management procedures to supply cloud services.

Service benefit and configuration administration:

- Service asset and pattern administration **information regarding cloud infrastructure resources** as storage display, spare elements, physical servers.
- The information comprises manufacturer name, customer ID's(Cls), license status, inventory status, serial number, account of amendment, edition and position
- It also sustains information the **interrelationships among CIs** such as a VM to service, service to its customer, VDC to its site, a physical to a control transfer data to the server and a physical a VM hosted on the server.
- This guarantees **design objects are an**alyzed as incorporated elements

Service asset and design management sustains the information regarding configuration objects in one or more databases which are known as configurationmanagement database **(CMDB**).

- This database is used by every cloud service management procedure to **transact with troubles** or to **comprise alterations into cloud infrastructure and services.**

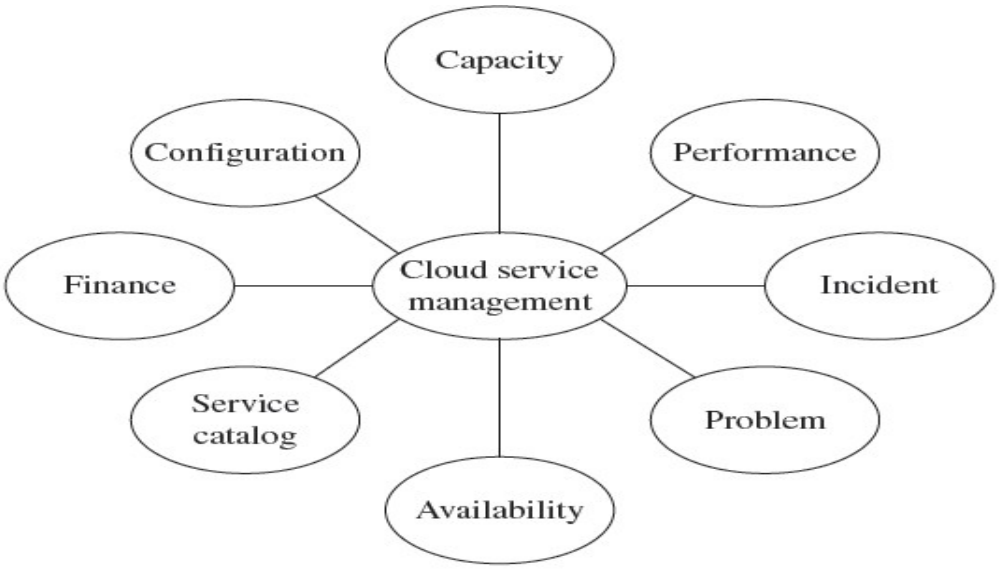

Fig. 9.2 Process administration involved in cloud service management

Capacity administration:

- Capacity administration deals with the management of various cloud computing products and services
- Capacity administration guarantees that a cloud infrastructure is capable of meeting the mandatory requirements for cloud services in a price efficient way
- Capacity administration is accountable for setting up future IT infrastructure necessities for cloud services.
- It collects information the past and present consumption of resources setup tendency on aptitude utilization

Performance administration:

- Performance administration involves monitoring, measuring, analyzing and improving performance of cloud infrastructure and services.
- Performance management monitors and measures performance such as response time and data transfer rate of the cloud infrastructure resources services.
- It analyses the performance statistics and identifies resources and services that are performing below expected level.
- Overconsumption of resources is a major cause for lack of performance

Incident administration:

- Incident administration arranges events on the basis of their **strictness and reinstates of cloud services** inside a decided time frame.

It attempts to recover cloud services as soon as possible by amending th malfunction or **fault** which instigated the event.

- Incident administration records the **event history** along with the particulars of the event indications, cloud infrastructure resources which create the cloud service, severity of the incident, time to determine the incident, depiction of the fault and the determination data.
- Incident administration might comprise **numerous support clusters to support fault tolerance.**

Problem administration:

- Problem administration **prevents events from recurring and reduces the unfavorable influence of events** which may not be avoided.
- It **lists troubles on the basis of their influence** to the company and **take remedial measures**.
- It **recognizes the basic reason of a trouble** and **prompts the most suitable solution** for the trouble.
- Problem administration also records the **trouble history** with particulars on the indications.
- The trouble history offers a chance **to study and control future problems.**

Availability administration:

- Availability administration makes certain that the **accessibility necessities of a cloud service** are continuously met.
- Availability administration assists when **server outcomes in the services failover** to a further accessible server in the group
- It guarantees that the corporation continuity procedures, **cloud infrastructure and devices are suitable to meet the necessary accessibility level.**
- It constantly **evaluates and scrutinizes the declared accessibility and the attained accessibility of cloud services** and recognizes the areas where the accessibility ought to be enhanced

Service catalog administration

- Service Catalogue Administration comprises **generating and sustaining a service catalog**
- It guarantees the information in the service catalog is the **latest and correct**
- It brings in **usefulness and is clear the service offering**s in catalog.
- It guarantees that the **service depiction is precious** and definite to customers

Financial administration

- Financial administration comprises estimating the **price of supplying a service.**
- The price comprises ongoing **operational expenditures** (OPEX) such as cooling, power, and facility cost, **capital expenditure (CAPEX)** such as attaining and installation prices of cloud infrastructure, and **management expenditure** like labour charge.

- It also scrutinizes and reports on consumption and distribution of resources by customers.

Compliance administration:

- Compliance administration makes certain that the cloud infrastructure resources, cloud services and service creation procedures fulfil the **strategies and authorized demands**.
- It guarantees that the compliance needs are **satisfied at the same time as provisioning cloud services and configuring cloud infrastructure.**
- It comprises the following: 1. Strategies and rules 2. External legal demands
- Strategies and rules applies to both **cloud service suppliers and customers**. Strategies and rules might be on **designing the finest practices.**
- External legal demands are **data secrecy laws** forced by distinct nations. These laws can identify geographical sites in order to **save consumer data** and **prohibit alteration or removal of data throughout maintenance period**.

4. CLOUD PROVIDERS AND TRADITIONAL IT SERVICES PROVIDERS

- Traditional IT providers **control** the **software, hardware, storage, networks** their customers.
- Traditional IT service provider handles the **whole setting** including **software licensing bills and infrastructure**
- To raise the demand for cloud-based services, **lower capital cost**s and **better flexibility** are required.

	Traditional IT services	Cloud services
Data	If there is a requirement for higher bandwidth or storage, you should upgrade your package. These packages are generally deployed on a monthly or yearly subscription.	According to your requirement, cloud services permit you to alter the bandwidth and storage. When you are not using bandwidth, you need not pay for it.
Performance	When any issues occur, you need to be worried	A backup facility is provided by cloud services that gets activated during issues in the main server. The data is retrieved as well as saved within a secure period.
Security	Less secure	More secure

Differences between Traditional IT Services and Cloud Services

Cloud providers always offer clouds apps with advanced features as given here:

- **Data** can be **stored locally**, even on offline mode
- Web browsers along with **custom-built appliance** installed on the Internet connect devices such **mobiles phones and desk**tops
- **Access** to a broad scale of **services** such as **application development platforms, on demand computing cycles** and **storage is available.**

5. HOW TO ACCESS THE CLOUD

For accessing the cloud, an Internet connection is required. Every provider offers a certain **function** so that **users can manage** their clouds according to their form. One can buy additional storage space at any time from the respective cloud provider according to technological requirements, thus the cloud provider will offer services on a pay-as-you-use method. The **tools required for acce**ssing cloud services and tools are **Platforms, Web Applications, Web Application Programming Interface and Web Browsers**.

Platforms:

- A platform refers to the way in which a cloud computing environment is supplied to you.
- The frame work of a web application involves maintaining the expansion of dynamic web applications, web services and websites.
- Static pages can also be made dynamic using interface (CGI). This allows external applications to interface with web servers.

Web Applications:

The decision of choosing a web application on the cloud depends on the type of providers available and services they offer. Google Apps is a set of applications which comprises the following:

- Start page for designing a customizable home page on a particular domain
- Gmail webmail services
- Google Talk instantaneous messaging and Voice Over IP
- Google Calendar and shared calendaring

A premium service known as **Google apps premier edition** is being provided by Google. This offer

1. **Ten gigabytes storage per user** -Around **100 times more storage space** than that of the standard commercial mailbox, purging the necessity of deleting emails repeatedly.
2. **APIs for business incorporation** -APIs for single sign-on, mail gateways and data migration permit businesses to further modify the service for unique situations.
3. **Almost 99.9** per cent uptime service level agreements for high accessibility of Gmail as well as Google crediting and monitoring consumers in case of failures of service levels.
4. **24/7 support** for serious problems comprising comprehensive business hours' telephonic support for managers.
5. **Optional advertising can be turned off,** but if needed, companies can decide to take in significant target-based ads of Google.
6. **Low fee** -Due to **affordable and low annual fee** per customer account per year, it is **easy to provide these applications to everybody** in the business.
7. Along with Google talk, Gmail, start page and Google Calendar, **all versions of Google apps** also involve other customized tools if required by the user.
8. **Google docs and spreadsheets** -It helps the **groups to work together on**

documents and spreadsheets without emailing documents back and forth.

9. **Gmail for mobile devices on BlackBerry** -Gmail, meant for mobile services, offers the same experience like conversation view, search, and synchronization of the desktop edition on BlackBerry handheld tools for Google apps' users.
10. **Application level control** -Permits **managers to adjust services** to business strategies like distribution of documents or calendars outside of the business.

Web Application Programming Interface:

- An API is a suite of **programming instructions and values** to access a web-based program.
- Different APIs are used by different cloud servers.
- **Web services and API are invisible** to your users as they access the cloud.

XML is not the only standard that makes APIs work. Other standards include the following:

1. **Simple object access protocol (SOAP)-**It **encodes XML messages** so that they can be **received and understood by any operating system** over any type **of network protocol**.
2. **Universal description, discovery, and integration (UDDI)** -It is an XML-based service of the directory that permits businesses to enroll themselves, **correlate with each other and collaborate using web services**.
3. **Web services description language (WSDL)** -It is the SOAP of UDDI WSDL is the **XML-based language** that businesses use to describe their services in **the UDDI creators**

Google Gadgets:

To search the emails, files, web history and chats, desktop search application, Google Gadgets, is used

The Google Gadgets API is composed of three languages:

1. XML-This language is generally used for **writing gadget specifications**. A gadget is just an XML file that can be easily found.
2. HTML- HTML is the markup language used to **develop static pages on the web**. XML is used to define **rules both in human readable language and machine readable form used with markup language and with other tags available in it that can be used to define text based format.**
3. JavaScript- It is the scripting language you can use to add dynamic behaviousr to your gadgets.

The Google data APIs provide you with a facility to read and write data with simplestandard protocol.

Some of the Google data APIs are given below:

- **Google apps APIs**
- **Blogger data API**
- **Google Calendar data API**
- **Google Contacts data API**
- **Google Finance Portfolio data API**
- **Google Notebook data API**
- **Google Spreadsheets data API**
- **Webmaster Tools data API**
- **Google base data API**
- **Google book search data API**
- **Google Code Search data API**
- **Google Documents List data API**
- **Google Health data API**
- **Picasa Web Albums data API**
- **YouTube data API**

Web Browsers:

- To **connect to the cloud**, you need an interface, which is a web browser.

Internet Explorer:

Internet Explorer 8 supports and brings a new look with enhanced capabilities to support everyday tasks such as **searching, browsing websites and printing**.

Internet Explorer 8 has been designed with the following representing modes:

1. The first one reflects the **execution of the current web standards** of Microsoft.
2. The second one reflects the **execution of web standards** of Microsoft at the time of launching of **Internet Explorer 7** in 2006
3. The third one is based on **rendering techniques of the web during the initial period.**

Mozilla Firefox:

- Firefox 3 was launched by Mozilla in June 2008 which was the main update to its **open-source, the famous free browser**
- Firefox 3 has been developed on the best of the Gecko 1.9 platform, resulting in a more **subjective easier-to-use and safer product.**
- **Less memory** is used by Firefox during its operation in comparison to its earlier issues.
- The **redesigned page interpretation and layout** depicts that users can see web pages two to three times faster than that of Firefox 2.
- Firefox has **increased the security** bar. The latest phishing protection and malware helps in **protection from worms, spyware, viruses and Trojans** to keep users secure on the web.
- Firefox allows users to confirm authenticity of the site.

Safari:

- Safari 3.1 is declared by Apple as the **fastest web browser** of the world for Window PCs and MAC
- It loads web pages **1.7 times faster than Firefox 2** and **1.9** times quicker when compared to **Internet Explorer**.
- It features a **natural browsing experience** along with managed amalgamated **"find"** displays number of matches **page, drag and drop bookmarks**, and a **built-in reader** to quickly scan the latest information and news.

Chrome:

- Chrome venture Google into market of **open-source browsers**.
- It is an influential platform which facilitates users to **collaborate with colleagues and friends via web applications** including email, listening music, watching videos and editing documents etc.
- Google Chrome created for the web today and tomorrow's applications

Chrome Cloud:

This is the most due to the power of V8 JavaScript engine and built in Google gear. API elements Gear are follows:

1. A **worker module** which offers **parallel implementation of code of JavaScript**
2. A **desktop module** which allows **web applications interrelate with desktop more logically**
3. A **database module** which can **save information locally**
4. A **geolocation module** which allows **web applications to discover the geological positions of the users**
5. A **local server module** which **serves and accumulates application resources** such as images, HTML, JavaScript and many more

6. MIGRATING TO CLOUD

- A corporation's aim of migrating to the cloud must be a clear perceptive of **what the cloud may** and **may not do it**.
- A well-formed cloud policy needs **thorough deliberation of every cloud service kind**, **installation module** and **access** alternative considered against specific application characteristics of the company.
- A full-grown immigration policy will be expected in various objects, but among these, three are the most vital— **organizational customs, IT governance, and virtualization.**

Advantages of Clouds:

- **Scalable** Nearly all cloud computing services are payable as pe a monthly charge. Therefore, if your business expands, you may order additional server space.
- **Nearly zero provisioning cost—**Along with the cloud, you need not establish a server and waste the man-hours it takes to get one, up and working.
- **Money can be saved for other means**— It lessens the expenses involved in setting up a novel infrastructure.
- **Fewer IT infrastructure staff to administer**— Along with cloud computing, you need not employ a technical group to administer and scrutinize the cloud.

Disadvantages of Clouds:

- **Scarcity of control**— Servers might not be directly administrated by you, but even now you are accountable for your data. Before deciding on a cloud service, you must inquire about the backup system, safety, and upholding policies. Decide what happens to your data if the company modifies hands or moves out of business, and understand that your company even possesses the information and may get it from the cloud whenever you need it.
- **Integrating systems**— It is impossible for everybody to shift a corporation totally to the cloud due to the systems it has constructed already.
- **Speed**— Along with the data transfer speed, cloud-based services sometimes do not communicate among themselves during low-speed connectivity.
- **Cost**— As there is no up-front assets' investment associated with cloud services, the technology is yet to have long-standing expenses' investments. The expenses linked with downtime, renewing of native computers to access the system, and unavoidable unpredicted faults are all drawbacks for drifting to the cloud.

Challenges in Migration to and from the Cloud:

- Successful cloud migrations need adequate resources and time.
- The three main dealers—Google, Microsoft, and Amazon,—control the cloud market, whereas most businesses use multiple dealers' apps instead of asking employees to follow one set of devices.
- With the lack of third-party support, controlling cloud apps might be a budget buster.
- Safety remains the biggest obstacle to cloud adoption.
- Although various businesses have made the jump to the cloud, almost one quarter is still uncertain to migrate—mainly because of concerns over safety.

Disaster Recovery

Disaster recovery is the practice of making a system capable of surviving unexpected or extraordinary failures. A disaster recovery plan, for example, will help your IT systems

Survive a fire in your data center that destroys all of the servers in that data center and the systems they support. Every organization should have a documented disaster recovery process and should test that process at least twice each year

7. DISASTER RECOVERY PLANNING

Disaster recovery deals with catastrophic failures that are extremely unlikely to occur during the lifetime of a system. If they are reasonably expected failures, they fall under the auspices of traditional availability planning. Although each single disaster is unexpected over the lifetime of a system, the possibility of some disaster occurring over time is reasonably nonzero.

"Disaster Recovery Planning is identifies an acceptable recovery state and develops processes and procedures to achieve the recovery state in the event of a disaster."

Defining a disaster recovery plan involves **two key metrics**:

Recovery Point Objective (RPO)

"The recovery point objective identifies how much data you are willing to lose in the event of a disaster**".** This value is typically specified in a number of hours or days of data. **For example**, if you determine that it is OK to lose 24 hours of data, you must make sure that the backups you'll use for your disaster recovery plan are never more than 24 hours old.

Recovery Time Objective (RTO)

"The recovery time objective identifies how much downtime is acceptable in the event of a disaster**"**. If your RTO is 24 hours, you are saying that up to 24 hours may elapse between the point when your system first goes offline and the point at which you are fully operational again.

Accomplishing that level of redundancy is expensive. It would also come with a nontrivial performance penalty. The cold reality for most businesses is likely that the cost of losing 24 hours of data is less than the cost of maintaining a zero downtime/zero loss of data infrastructure.

Determining an appropriate RPO and RTO is ultimately a financial calculation: at what point does the cost of data loss and downtime exceed the cost of a backup strategy that will prevent that level of data loss and downtime? The right answer is radically different for different businesses.

The final element of disaster recovery planning understands the catastrophic scenario. A good disaster recovery plan can describe that scenario so that all stakeholders can understand and accept the risk.

Recovery Point Objective (RPO)

Any software system should be able to attain an RPO between 24 hours for a simple disaster to one week for a significant disaster without incurring absurd costs. Losing 24 hours of banking transactions would never be acceptable, much less one week.

RPO is typically governed by the way in which you save and back up data:

• Weekly off-site backups will survive the loss of your data center with a week of data loss. Daily off-site backups are even better.

• Daily on-site backups will survive the loss of your production environment with a day of data loss plus replicating transactions during the recovery period after the loss of the system. Hourly on-site backups are even better.
• A NAS (Network Attached System) /SAN (Storage Area Network) will survive the loss of any individual server, except for instances of data corruption with no data loss.
• A clustered database will survive the loss of any individual data storage device or database node with no data loss.
• A clustered database across multiple data centers will survive the loss of any individual data center with no data loss.

The Recovery Time Objective(RTO)

In a traditional infrastructure, a rapid RTO is very expensive. It would have to have an agreement in place with another managed services provider to provide either a backup infrastructure or an SLA for setting up a replacement infrastructure in the event your provider goes out of business. Depending on the nature of that agreement, it could nearly double the costs of your IT infrastructure. The cloud—even over virtualized datacenters—alters the way you look at your RTO

8. DISASTERS IN THE CLOUD

Assuming unlimited budget and capabilities, three key things in disaster recovery planning:

1. Backups and data retention
2. Geographic redundancy
3. Organizational redundancy

It can take care of those three items, it's nearly certain meet most RPO and RTO needs.

In addition, if you're hosting provider is a less-proven organization, organizational redundancy may be more important than geographic redundancy. Fortunately, the structure of the Amazon cloud makes it very easy to take care of the first and second items. In addition, cloud computing in general makes the third item much easier.

Backup Management

Now it's time to take a step back from the technical details and examine the kinds of data you are planning to back up and how it all fits into your overall disaster recovery plan.

Table 6-1 illustrates the different kinds of data that web applications typically manage.

In disaster recovery, persistent data is generally the data of greatest concern. We can always rebuild the operating system, install all the software, and reconfigure it, but we have no way of manually rebuilding the persistent data.

Fixed data strategy

If you are fixated on the idea of backing up your machine images, you can download the images out of S3 and store them outside of the Amazon cloud. If S3 were to go down and incur data loss or corruption that had an impact on your AMIs, you would be able to upload the images from your off-site backups and reregister them.

Configuration Data strategy

A good backup strategy for configuration information comprises two levels. The first level can be either a regular filesystem dump to your cloud storage or a filesystem snapshot. For most applications, you can back up your configuration data once a day or even once a week and be fine. You should, however, think back to your Recovery Point Objective. If your configuration data changes twice a day and you have a two-hour RPO, you will need to back up your configuration data twice a day. If configuration data changes irregularly, it may be necessary to make hourly backups or specifically tie your backups to changes in application configuration.

Kind of data	Description
Fixed data	Fixed data, such as your operating system and common utilities, belong in your AMI. In the cloud, you don't back up your AMI, because it has no value beyond the cloud.*
Transient data	File caches and other data that can be lost completely without impacting the integrity of the system. Because your application state is not dependent on this data, don't back it up.
Configuration data	Runtimeconfiguration data necessary to make the system operate properly in a specific context. This data is not transient, since it must survive machine restarts. On the other hand, it should be easily reconfigured from a clean application install. This data should be backed up semi-regularly.
Persistent data	Your application state, including critical customer data such as purchase orders. It changes constantly and a database engine is the best tool for managing it. Your database engine should store its state to a block device, and you should be performing constant backups. Clustering and/or replication are also critical tools in managing the database.

An alternate approach is to check your application configuration into a source code repository outside of the cloud and leverage that repository for recovery from even minor losses. Whether you perform filesystem snapshots or simply zip up the filesystem that data will hibernate inside S3. Snapshots tend to be the most efficient and least intrusive mechanism for performing backups, but they are also the least portable

At some point, you do need to get that data out of the cloud so that you have off-site backups in a portable format. Here's what I recommend:

• Create regular—at a minimum, daily—snapshots of your configuration data.

• Create semi-regular—at least less than your RPO—filesystem archives in the form of ZIP or TAR files and move those archives into Amazon S3.

• On a semi-regular basis—again, at least less than your RPO—copy your filesystem archives out of the Amazon cloud into another cloud or physical hosting facility

Persistent data strategy (aka database backups)

Relational database to store customer information and other persistent data. After all, the purpose of a relational database is to maintain the consistency of complex transactional data. MySQL, like all database engines, provides several convenient tools for backups, but you must use them carefully to avoid data corruption.

With the configuration data, it is highly unlikely you will be making a backup in between the writing of two different files that must remain consistent or in the middle of writing out a file to the filesystem. With database storage, it is a near certainty that every time you try to copy those files, the database will be in the middle of doing something with them. As a result, you need to get clever with your database backup strategy.

The first line of defense is either multimaster replication or clustering. A multimaster database is one in which two master servers execute write transactions independently and replicate the transactions to the other master. A clustered database environment contains multiple servers that act as a single logical server. Under both scenarios, when one goes down, the system remains operational and consistent.

Instead, you can perform master-slave replication. Master-slave replication involves setting up a master server that handles your write operations and replicating transactions over to a slave server. Each time something happens on the master, it replicates to the slave. If a master can crash after a transaction has completed but before it has had time to replicate to the slave. To get around
this problem, I generally do the following:

- Set up a master with its data files stored on a block storage device.
- Set up a replication slave, storing its data files on a block storage device.
- Take regular snapshots of the master block storage device based on my RPO.
- Create regular database dumps of the slave database and store them in S3.
- Copy the database dumps on a semi-regular basis from S3 to a location outside the Amazon cloud.

Actually taking snapshots or creating database dumps for some database engines is actually very tricky in a runtime environment, especially if you want to do it hourly or even more frequently. The challenge in creating your backups for these database engines is the need to stop processing all transactions while the backup is taking place. To complicate the situation, database dumps can take a long time to complete. As a result, your applications will grind to a halt while you make any database dumps.

You need to freeze the database only for an instant to create your snapshot. The process follows these steps:

1. Lock the database.
2. Sync the filesystem (this procedure is filesystem-dependent).
3. Take a snapshot.
4. Unlock the database.

The whole process should take about one second.

On Amazon EC2, you will store your snapshots directly onto block storage. Unfortunately, the snapshots are not portable, so you can't use them for off-site storage. You therefore will need to do database dumps, no matter how much you would rather avoid doing them. Because of this need, I run my backups against a database slave. The slave can afford to be locked for a period of time while a database dump completes.

The steps for creating the database dump are:

1. Execute the database dump
2. When complete, encrypt the dump and break it into small, manageable chunks.
3. Move the dump over to S3.

Amazon S3 limits your file size to 5 GB. As a result, you probably need to break your database into chunks, and you should definitely encrypt it and anything else you send into Amazon S3. Now that you have a portable backup of your database server, you can copy that backup out of the Amazon cloud and be protected from the loss of your S3 backups.

Backup security

The harder part is securing your portable backups as you store them in S3 and move them off site. I typically use PGP-compatible encryption for my portable backups. You need to worry about two issues:

• Keeping your private decryption key out of the cloud.
• Keeping your private decryption key some place that it will never, ever get lost.

The cloud needs only your public encryption key so it can encrypt the portable backups.

You can't store your decryption key with your backups. Doing so will defeat the purpose of encrypting the backups in the first place. Because you will store your decryption key somewhere else, you run the risk of losing your decryption key independent of your backups.

The best approach? Keep two copies:

• One copy stored securely on a highly secure server in your internal network.
• One copy printed out on a piece of paper and stored in a safety deposit box nowhere near the same building in which you house your highly secure server.

If you are automating the recovery from portable backups, you will also need to keep a copy of the private decryption key on the server that orchestrates your automated recovery efforts.

Geographic Redundancy

The virtualization technologies behind the cloud simply make it a lot easier to automate those processes and have a relatively inexpensive mechanism for off-site backups.

Turning now to your Recovery Time Objective, the key is redundancy in infrastructure. If you can develop geographical redundancy, you can survive just about any physical disaster that might happen. With a physical infrastructure, geographical redundancy is expensive. In the cloud, however, it is relatively cheap.

The ability to bring your application up from the redundant location in a state that meets your Recovery Point Objective within a timeframe that meets your Recovery Time Objective. If you have a 2-hour RTO with a 24-hour RPO, geographical redundancy means that your second location can be operational within two hours of the complete loss of your primary location using data that is no older than 24 hours.

Amazon provides built-in geographic redundancy in the form of regions and availability zones. If you have your instances running in a given availability zone, you can get them started back up in another availability zone in the same region without any effort. If you have

Specific requirements around what constitutes geographic redundancy, Amazon's availability zones may not be enough—you may have to span regions.

Spanning availability zones

Just about everything in your Amazon infrastructure except block storage devices is available across all availability zones in a given region. Although there is a charge for network traffic that crosses availability zones, that charge is generally well worth the price for the leveraging ability to create redundancy across availability zones.

If you lose the entire availability zone B, nothing happens. The application continues to operate normally, although perhaps with degraded performance levels.

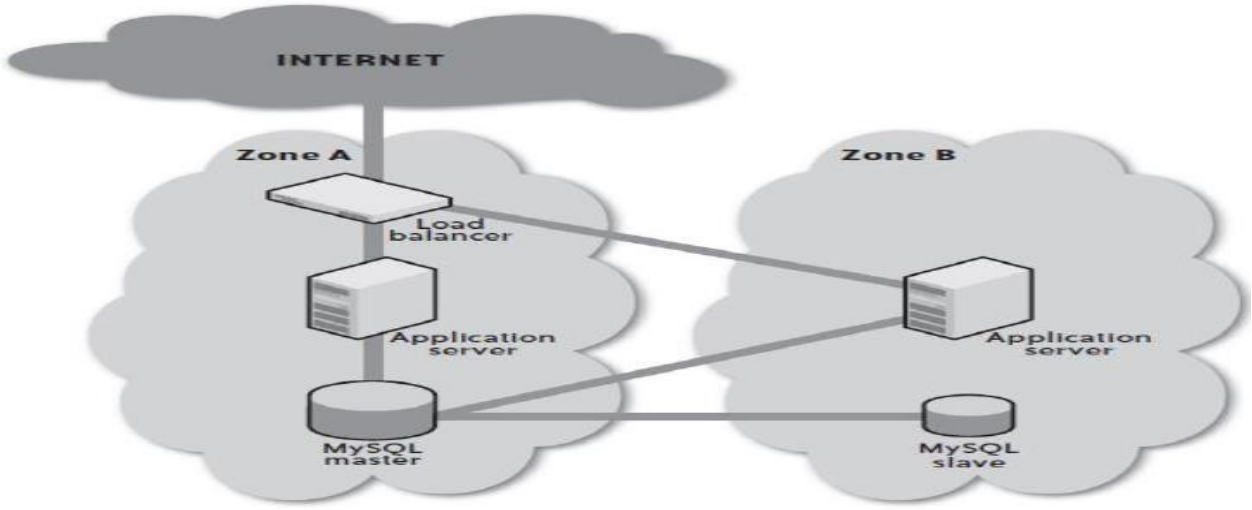

FIGURE . By spanning multiple availability zones, you can achieve geographic redundancy

If you lose availability zone A, you will need to bring up a new load balancer in availability zone B and promote the slave in that availability zone to master. The system can return to operation in a few minutes with little or no data loss. If the database server were clustered and you had a spare load balancer running silently in the background, you could reassign the IP address from the old load balancer to the spare and see only a few seconds of downtime with no data loss

.The Amazon SLA provides for a 99.95% uptime of at least two availability zones in each region. If you span multiple availability zones, you can actually exceed the Amazon SLA in regions that have more than two availability zones. The U.S. East Coast, for example, has three availability zones. As a result, you have only a 33% chance of any given failure of two availability zones being exactly the two zones you are using.

Operating across regions

Amazon supports two regions: us-east-1 (Eastern United States) and eu-west-1 (Western Europe). These regions share little or no meaningful infrastructure. The advantage of this structure is that your application can basically survive a nuclear attack on the U.S. or EU (but not on both!) if you operate across regions.

How you manage operations across regions on the nature of your web application and redundancy needs.

The issues you need to consider for simultaneous operation include:

DNS management

You can use round-robin DNS to work around the fact that IP addresses are not portable across regions, but you will end up sending European visitors to the U.S. and vice versa and lose half your traffic when one of the regions goes down. You can leverage a dynamic DNS system such as UltraDNS that will offer up the right DNS resolution based on source and availability.

Database management

Clustering across regions is likely not practical. You can also set up a master in one region with a slave in the other. Then you perform write operations against the master, but read against the slave for traffic from the region with the slave. Another option is to segment your database so that the European region has "European data" and the U.S. region has "American data." Each region also has a slave in the other region to act as a recovery point from the full loss of a region.

Regulatory issues

The EU does not allow the storage of certain data outside of the EU. As a result, legally may not be allowed to operate across regions, no matter what clever technical solutions you devise

Organizational Redundancy

Physical disasters are a relatively rare thing, but companies go out of business everywhere every day—even big companies like Amazon and Rackspace. Even if a company goes into bankruptcy restructuring, there's no telling what will happen to the hardware assets that run their cloud infrastructure. Your disaster recovery plan should therefore have contingencies that assume your cloud provider simply disappears from the face of the earth.

The best approach to organizational redundancy is to identify another cloud provider and establish a backup environment with that provider in the event your first provider fails.

The issues associated with organizational redundancy are similar to around operating across Amazon EC2 regions. In particular, you must consider all of the following concerns:

• Storing your portable backups at your secondary cloud provider.

• Creating machine images that can operate your applications in the secondary provider's Virtualized environment.

• Keeping the machine images up to date with respect to their counterparts with the primary provider.

• Not all cloud providers and managed service providers support the same operation systems or file systems. If your application is dependent on either, you need to make sure you select a cloud provider that can support your needs

9. DISASTER MANAGEMENT

To complete the disaster recovery scenario, you need to recognize when a disaster has happened and have the tools and processes in place to execute your recovery plan.

One of the coolest things about the cloud is that all of this can be automated.

Monitoring

Monitoring your cloud infrastructure is extremely important. You cannot replace a failing server or execute your disaster recovery plan if you don't know that there has been a failure. The trick, however, is that your monitoring systems cannot live in either your primary or secondary cloud provider's infrastructure. They must be independent of your clouds. The primary monitoring objective should be to figure out what is going to fail before it actually fails.

There are many other more mundane things that you should check on in a regular environment. In particular, you should be checking capacity issues such as disk usage, RAM, and CPU. In the end, however, you will need to monitor for failure at three levels:

- Through the provisioning API (for Amazon, the EC2 web services API)
- Through your own instance state monitoring tools
- Through your application health monitoring tools

Your cloud provider's provisioning API will tell you about the health of your instances, any volumes they are mounting, and the data centers in which they are operating. When you detect a failure at this level, it likely means something has gone wrong with the cloud itself. Before engaging in any disaster recovery, you will need to determine whether the outage is limited to one server or affects indeterminate servers, impacting an entire availability zone or an entire region.

Monitoring is not simply about checking for disasters; mostly it is checking on the mundane. With enStratus, I put a Python service on each server that checks for a variety of server health indicators—mostly related to capacity management. The service will notify the monitoring system if there is a problem with the server or its configuration and allow the monitoring system to take appropriate action. It also checks for the health of the applications running on the instance.

Load Balancer Recovery

One of the reasons companies pay absurd amounts of money for physical load balancers is to greatly reduce the likelihood of load balancer failure. Recovering a load balancer in the cloud, however, is lightning fast. As a result, the downside of a failure in your cloud-based load balancer is minor.

Recovering a load balancer is simply a matter of launching a new load balancer instance from the AMI and notifying it of the IP addresses of its application servers. You can further reduce any downtime by keeping a load balancer running in an alternative availability zone and then remapping your static IP address upon the failure of the main load balancer.

Application Server Recovery

If you are operating multiple application servers in multiple availability zones, your system as a whole will survive the failure of any one instance—or even an entire availability zone. You will still need to recover that server so that future failures don't affect your infrastructure.

The recovery of a failed application server is only slightly more complex than the recovery of a failed load balancer. Like the failed load balancer, you start up a new instance from the

Application server machine image. You then pass it configuration information, including where the database is. Once the server is operational, you must notify the load balancer of the existence of the new server (as well as deactivate its knowledge of the old one) so that the new server enters the load-balancing rotation

Database Recovery

Database recovery is the hardest part of disaster recovery in the cloud. Your disaster recovery algorithm has to identify where an uncorrupted copy of the database exists. This process may involve promoting slaves into masters, rearranging your backup management, and reconfiguring application servers.

The best solution is a clustered database that can survive the loss of an individual database server without the need to execute a complex recovery procedure. Absent clustering, the best recovery plan is one that simply launches a new database instance and mounts the still functional EC2 volume formerly in use by the failed instance. When an instance goes down, however, any number of related issues may also have an impact on that strategy:

- The database could be irreparably corrupted by whatever caused the instance to crash.
- The volume could have gone down with the instance.
- The instance's availability zone could be unavailable.
- You could find yourself unable to launch new instances in the volume's availability zone.

The following process will typically cover all levels of database failure:

1. Launch a replacement instance in the old instance's availability zone and mount its old volume.
2. If the launch fails but the volume is still running, snapshot the volume and launch a new instance in any zone, and then create a volume in that zone based on the snapshot.
3. If the volume from step 1 or the snapshots from step 2 are corrupt, you need to fall back to the replication slave and promote it to database master.
4. If the database slave is not running or is somehow corrupted, the next step is to launch a replacement volume from the most recent database snapshot.
5. If the snapshot is corrupt, go further back in time until you find a backup that is not corrupt.

Step 4 typically represents your worst-case scenario. If you get to 5, there is something wrong with the way you are doing backups.