UNIT 8: SCALABLE AND EMERGING INFORMATION SYSTEM TECHNIQUES

8.1. Techniques for voluminous data

For machine intelligence applications to work successfully, machines must perform reliably under variations of data and must be able to keep up with data streams. **Internet-Scale Pattern Recognition: New Techniques for Voluminous Data Sets and Data Clouds** covering from computational models that address performance and scalability to achieve higher levels of reliability. It explores different ways of implementing pattern recognition using machine intelligence.

The text draws on concepts from pattern recognition, parallel processing, distributed systems, and data networks. It describes fundamental research on the scalability and performance of pattern recognition, addressing issues with existing pattern recognition schemes for Internet-scale data deployment. It reviews numerous approaches and introduces possible solutions to the scalability problem. It offers an extendable template for Internet-scale pattern recognition applications as well as guidance on the programming of large networks of devices.

Features of voluminous data

- Covers the key technologies that contribute to Internet-scale pattern recognition, including distributed systems, parallel computing, and machine intelligence
- Outlines the underlying theory and principles of distributed pattern recognition
- Discusses one-shot learning and hierarchical approaches in distributed pattern recognition applications
- Includes examples of distributed models and parallel programming techniques—two forces driving the expansion of distributed applications in Internet-scale environments
- Shows how pattern recognition can be a scalable commodity for information processing

Possible techniques or model for voluminous data

1. **Infrastructure as a service (IAAS)**

In the most basic cloud-service model - and according to the IETF (Internet Engineering Task Force) - providers of IAAS offer computers – physical or (more often) virtual machines – and other resources. IAAS refers to online services that abstract user from the detail of infrastructure like physical computing resources, location, data partitioning, scaling, security, backup etc.

2. **Platform as a service (PAAS)**

A PAAS vendor offers a development environment to application developers. The provider typically develops toolkit and standards for development and channels for distribution and payment. In the PAAS models, cloud providers deliver a computing platform, typically including operating system, programming-language execution environment, database, and web server. Application developers can develop and run their software solutions on a cloud platform without the cost and complexity.

3. **Software as a service (SAAS)**

In the software as a service (SAAS) model, users gain access to application software and databases. Cloud providers manage the infrastructure and platforms that run the applications. SAAS is sometimes referred to as "on-demand software" and is usually priced on a pay-per-use basis or using a subscription fee.

In the SAAS model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. Cloud users do not manage the cloud infrastructure and platform where the application runs. This eliminates the need to install and run the application on the cloud user's own computers, which simplifies maintenance and support.

8.2 Cloud computing technologies and their types

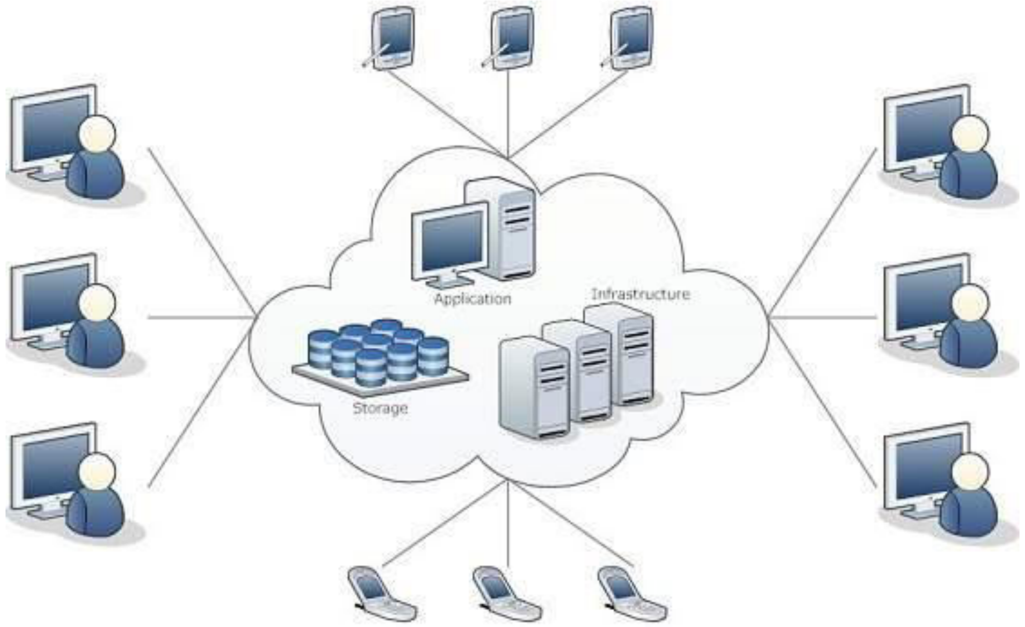
Cloud Computing provides us means of accessing the applications as utilities over the Internet. It allows us to create, configure, and customize the applications online.

What is Cloud?

The term **Cloud** refers to a **Network** or **Internet.** In other words, we can say that Cloud is something, which is present at remote location. Cloud can provide services over public and private networks, i.e., WAN, LAN or VPN. Applications such as e-mail, web conferencing, customer relationship management (CRM) execute on cloud.

What is Cloud Computing?

Cloud Computing refers to **manipulating, configuring,** and **accessing** the hardware and software resources remotely. It offers online data storage, infrastructure, and application.



Cloud computing offers **platform independency,** as the software is not required to be installed locally on the PC. Hence, the Cloud Computing is making our business applications **mobile** and **collaborative.**

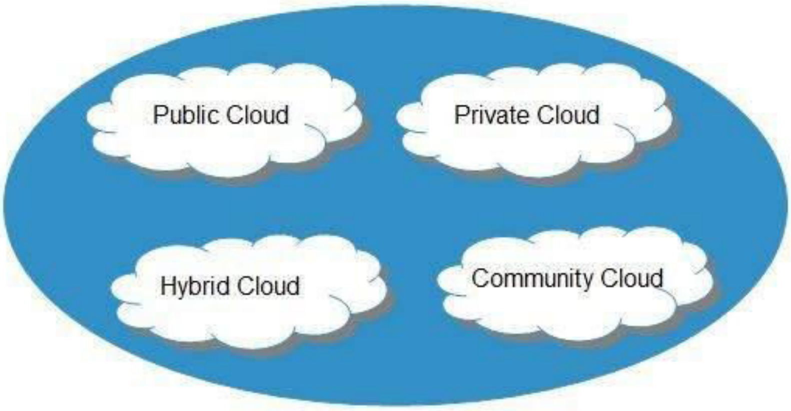
Basic Concepts

There are certain services and models working behind the scene making the cloud computing feasible and accessible to end users. Following are the working models for cloud computing:

- Deployment Models
- Service Models

Deployment Models

Deployment models define the type of access to the cloud, i.e., how the cloud is located? Cloud can have any of the four types of access: Public, Private, Hybrid, and Community.



1. **PUBLIC CLOUD**

The **public cloud** allows systems and services to be easily accessible to the general public. Public cloud may be less secure because of its openness.

2. **PRIVATE CLOUD**

The **private cloud** allows systems and services to be accessible within an organization. It is more secured because of its private nature.

3. **COMMUNITY CLOUD**

The **community cloud** allows systems and services to be accessible by a group of organizations.

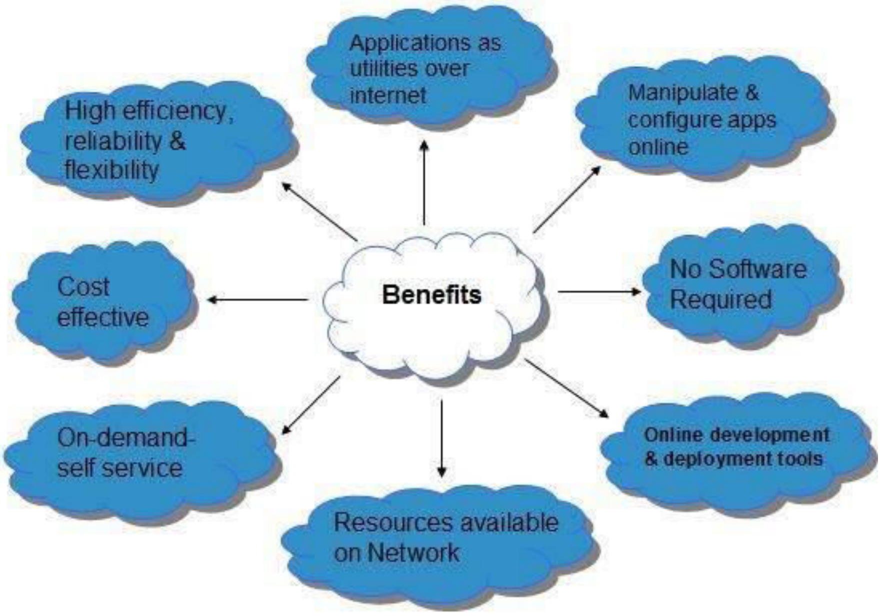
4. **HYBRID CLOUD**

The **hybrid cloud** is a mixture of public and private cloud, in which the critical activities are performed using private cloud while the non-critical activities are performed using public cloud.

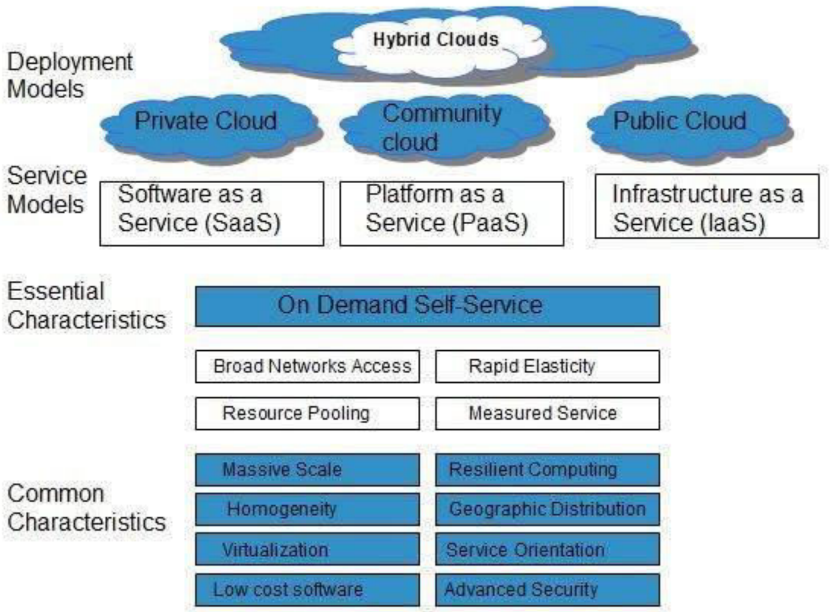
Benefits of Cloud Computing

Cloud Computing has numerous advantages. Some of them are listed below -

- One can access applications as utilities, over the Internet.
- One can manipulate and configure the applications online at any time.
- It does not require to install a software to access or manipulate cloud application.
- Cloud Computing offers online development and deployment tools.
- Cloud resources are available over the network in a manner that provide platform independent access to any type of clients.
- Cloud Computing offers **on-demand self-service.** The resources can be used without interaction with cloud service provider.
- Cloud Computing is highly cost effective because it operates at high efficiency with optimum utilization. It just requires an Internet connection
- Cloud Computing offers load balancing that makes it more reliable.


Applications as utilities over internet
High efficiency, reliability & flexibility
Manipulate & configure apps online
Cost effective
Benefits
No Software Required
On-demand-self service
Online development & deployment tools
Resources available on Network

Characteristics of Cloud Computing



1. **On Demand Self Service**

Cloud Computing allows the users to use web services and resources on demand. One can logon to a website at any time and use them.

2. **Broad Network Access**

Since cloud computing is completely web based, it can be accessed from anywhere and at any time.

3. **Resource Pooling**

Cloud computing allows multiple tenants to share a pool of resources. One can share single physical instance of hardware, database and basic infrastructure.

4. **Rapid Elasticity**

It is very easy to scale the resources vertically or horizontally at any time. Scaling of resources means the ability of resources to deal with increasing or decreasing demand.

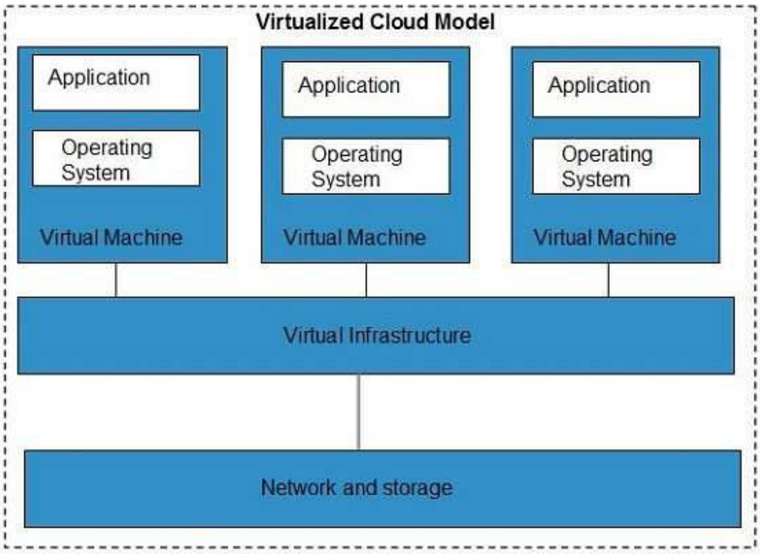
Cloud Computing Technologies and their types

There are certain technologies working behind the cloud computing platforms making cloud computing flexible, reliable, and usable. These technologies are listed below:

- Virtualization
- Service-Oriented Architecture (SOA)
- Grid Computing
- Utility Computing

➢ **Virtualization**

Virtualization is a technique, which allows sharing single physical instance of an application or resource among multiple organizations or tenants (customers). It does this by assigning a logical name to a physical resource and providing a pointer to that physical resource when demanded.

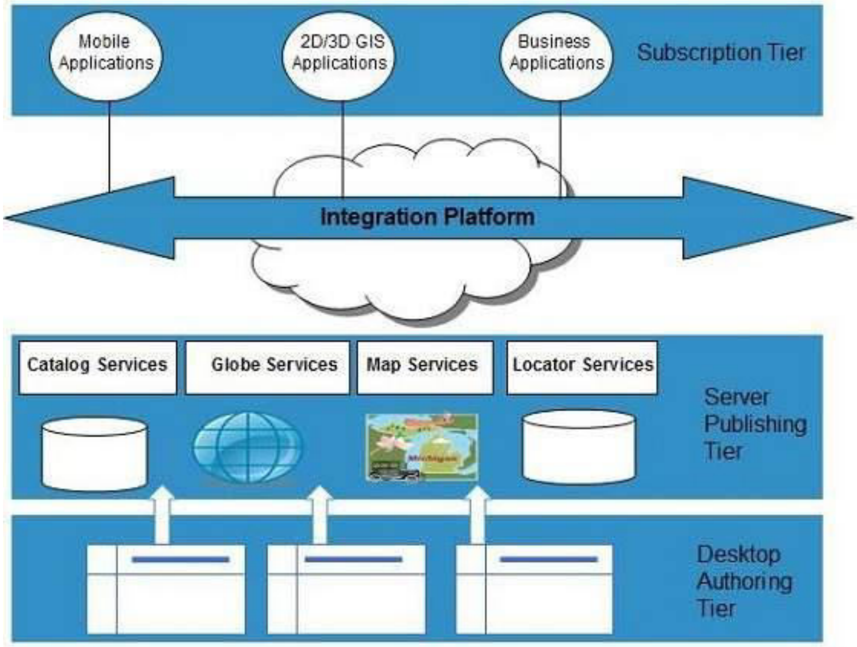


The **Multitenant** architecture offers **virtual isolation** among the multiple tenants. Hence, the organizations can use and customize their application as though they each have their instances running.

- **Service-Oriented Architecture (SOA)**

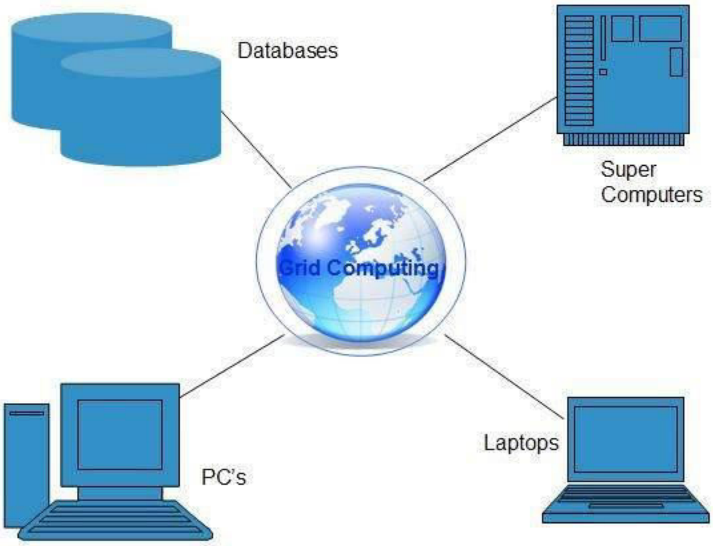
Service-Oriented Architecture helps to use applications as a service for other applications regardless the type of vendor, product or technology. Therefore, it is possible to exchange the data between applications of different vendors without additional programming or making changes to services.

The cloud computing service oriented architecture is shown in the diagram below.



- **Grid Computing**

Grid Computing refers to distributed computing, in which a group of computers from multiple locations are connected with each other to achieve a common objective. These computer resources are heterogeneous and geographically dispersed. Grid Computing breaks complex task into smaller pieces, which are distributed to CPUs that reside within the grid.



- **Utility Computing**

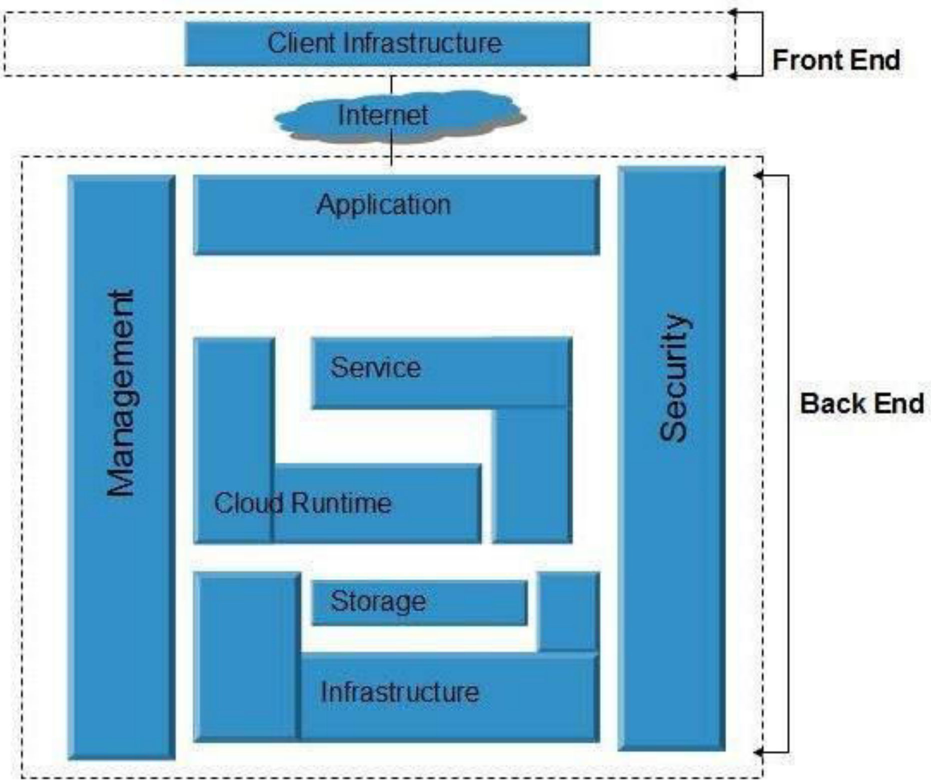
Utility computing is based on **Pay-per-Use model.** It offers computational resources on demand as a metered service. Cloud computing, grid computing, and managed IT services are based on the concept of utility computing

Architecture of cloud computing

Cloud Computing architecture comprises of many cloud components, which are loosely coupled. We can broadly divide the cloud architecture into two parts:

- Front End
- Back End

Each of the ends is connected through a network, usually Internet. The following diagram shows the graphical view of cloud computing architecture:



Front End

The **front end** refers to the client part of cloud computing system. It consists of interfaces and applications that are required to access the cloud computing platforms, Example - Web Browser.

Back End

The **back End** refers to the cloud itself. It consists of all the resources required to provide cloud computing services. It comprises of huge data storage, virtual machines, security mechanism, services, deployment models, servers, etc.

8.3 Map Reduce and Hadoop systems

8.3.1 Hadoop Systems

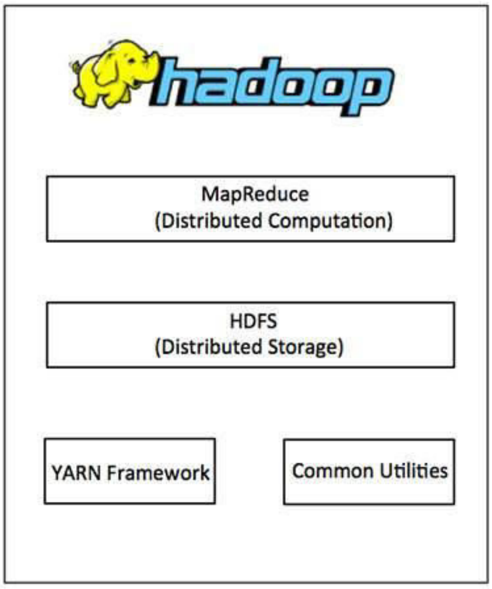
Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture

Hadoop framework includes following four modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries provide files system and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
- **Hadoop Distributed File System (HDFS):** A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.

We can use following diagram to depict these four components available in Hadoop framework.



How Does Hadoop Work?

Stage 1

A user/application can submit a job to the Hadoop (a hadoop job client) for required process by specifying the following items:

1. The location of the input and output files in the distributed file system.
2. The java classes in the form of jar file containing the implementation of map and reduce functions.
3. The job configuration by setting different parameters specific to the job.

Stage 2

The Hadoop job client then submits the job (jar/executable etc) and configuration to the Job Tracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client.

Stage 3

The Task Trackers on different nodes execute the task as per Map Reduce implementation and output of the reduce function is stored into the output files on the file system.

Advantages of Hadoop

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

8.3.2 Map Reduce

Map Reduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

The term Map Reduce actually refers to the following two different tasks that Hadoop programs perform:

- **The Map Task:** This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples (key/value pairs).
- **The Reduce Task:** This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

Typically both the input and the output are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

The Map Reduce framework consists of a single master **Job Tracker** and one slave **Task Tracker** per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks.

.

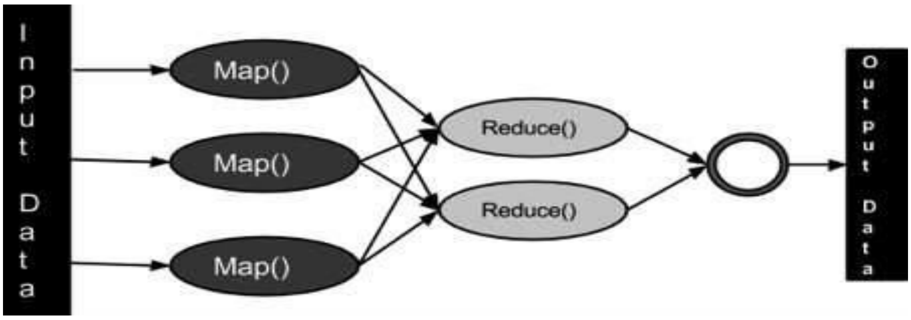
Advantage of Map Reduce

The major advantage of Map Reduce is that it is easy to scale data processing over multiple computing nodes. Under the Map Reduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial.

But, once we write an application in the Map Reduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the Map Reduce model.

The Algorithm for Map Reduce

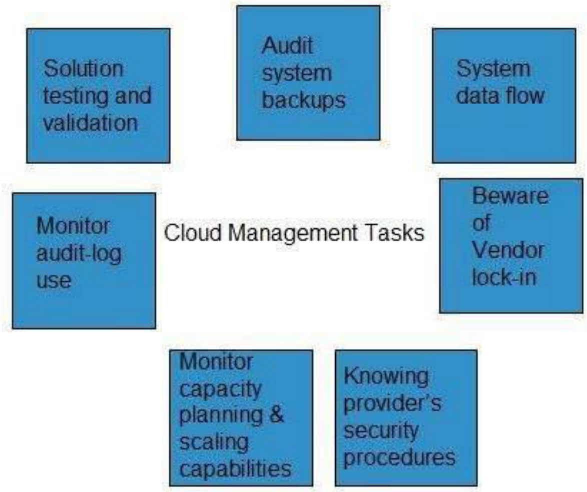
- Generally MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
  - **Map stage**: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
  - **Reduce stage**: This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.
  -



8.4 Data Management in the cloud

It is the responsibility of cloud provider to manage resources and their performance. Management of resources includes several aspects of cloud computing such as load balancing, performance, storage, backups, capacity, deployment, etc. The management is essential to access full functionality of resources in the cloud.

Cloud Management Tasks


Solution testing and validation
Audit system backups
System data flow
Monitor audit-log use
Cloud Management Tasks
Beware of Vendor lock-in
Monitor capacity planning & scaling capabilities
Knowing provider's security procedures

1. **Audit System Backups**

It is required to audit the backups timely to ensure restoring of randomly selected files of different users. Backups can be performed in following ways:

- Backing up files by the company, from on-site computers to the disks that reside within the cloud.
- Backing up files by the cloud provider.

2. **Data Flow of the System**

The managers are responsible to develop a diagram describing a detailed process flow. This process flow describes the movement of data belonging to an organization throughout the cloud solution.

3. **Vendor Lock-In Awareness and Solutions**

The managers must know the procedure to exit from services of a particular cloud provider. The procedures must be defined to enable the cloud managers to export data of an organization from their system to another cloud provider.

4. **Knowing Provider's Security Procedures**

The managers should know the security plans of the provider for the following services:

- Multitenant use
- E-commerce processing
- Employee screening
- Encryption policy

5. **Monitoring Capacity Planning and Scaling Capabilities**

The managers must know the capacity planning in order to ensure whether the cloud provider is meeting the future capacity requirement for his business or not.

The managers must manage the scaling capabilities in order to ensure services can be scaled up or down as per the user need.

6. **Monitor Audit Log Use**

In order to identify errors in the system, managers must audit the logs on a regular basis.

Solution Testing and Validation

When the cloud provider offers a solution, it is essential to test it in order to ensure that it gives the correct result and it is error-free. This is necessary for a system to be robust and reliable.

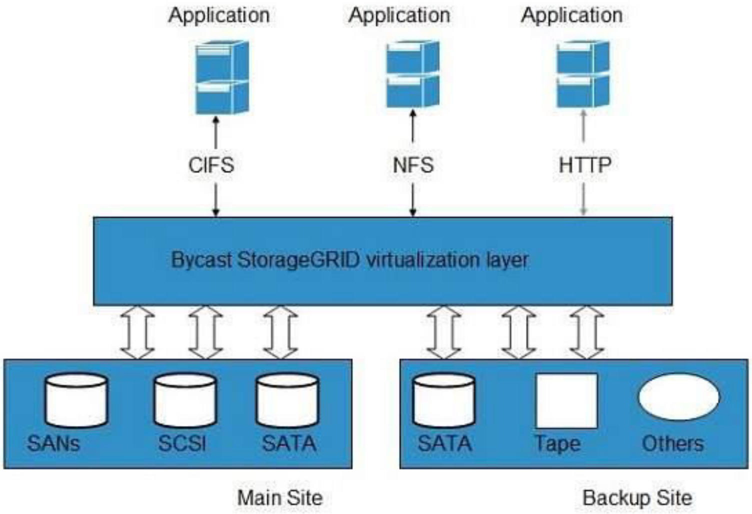
Managing of Data Storage

Cloud Storage is a service that allows saving data on offsite storage system managed by third-party and is made accessible by a **web services API.**

Storage Devices

Storage devices can be broadly classified into two categories:

- Block Storage Devices
- File Storage Devices



8.5 Information retrieval in the cloud

Information Retrieval

Information retrieval deals with the retrieval of information from a large number of text-based documents. Some of the database systems are not usually present in information retrieval systems because both handle different kinds of data. Examples of information retrieval system include −

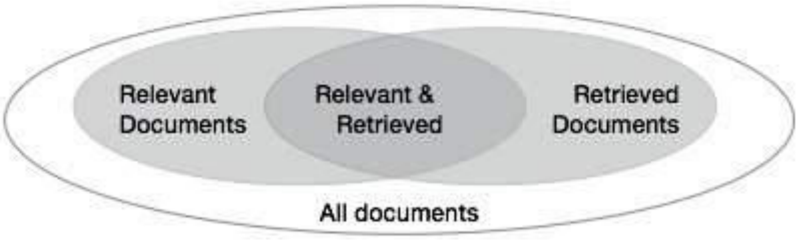
- Online Library catalogue system
- Online Document Management Systems
- Web Search Systems etc.

In such search problems, the user takes an initiative to pull relevant information out from a collection. This is appropriate when the user has ad-hoc information need, i.e., a short-term need. But if the user has a long-term information need, then the retrieval system can also take an initiative to push any newly arrived information item to the user.

This kind of access to information is called Information Filtering. And the corresponding systems are known as Filtering Systems or Recommender Systems.

Basic Measures for Text Retrieval

We need to check the accuracy of a system when it retrieves a number of documents on the basis of user's input. Let the set of documents relevant to a query be denoted as {Relevant} and the set of retrieved document as {Retrieved}. The set of documents that are relevant and retrieved can be denoted as {Relevant} ∩ {Retrieved}. This can be shown in the form of a Venn diagram as follows −



There are three fundamental measures for assessing the quality of text retrieval −

- Precision
- Recall
- F-score

Precision

Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as −

$$\text{Precision}= |\{\text{Relevant}\} \cap \{\text{Retrieved}\}| / |\{\text{Retrieved}\}|$$

Recall

Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as −

$$\text{Recall} = |\{\text{Relevant}\} \cap \{\text{Retrieved}\}| / |\{\text{Relevant}\}|$$

F-score

F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa. F-score is defined as harmonic mean of recall or precision as follows −

$$\text{F-score} = \text{recall} \times \text{precision} / (\text{recall} + \text{precision}) / 2$$

Information Retrieval Services

There exist several Information retrieval services offering easy access to information present on the internet. The following table gives a brief introduction to these services:

S.N.	Service Description
1	**File Transfer Protocol (FTP)** Enable the users to transfer files.
2	**Archie** It's updated database of public FTP sites and their content. It helps to search a file by its name.
3	**Gopher** Used to search, retrieve, and display documents on remote sites.
4	**Very Easy Rodent Oriented Net wide Index to Computer Achieved (VERONICA)** VERONICA is gopher based resource. It allows access to the information resource stored on gopher's servers.

8.6 Link Analysis in cloud setup

Pre-Installation Setup

Before installing Hadoop into Linux environment, we need to set up Linux using **ssh** (Secure Shell). Follow the steps mentioned below for setting up the Linux environment.

Creating a User

It is recommended to create a separate user for Hadoop to isolate the Hadoop file system from the UNIX file system. Follow the steps given below to create a user:

- Open root using the command "su".
- Create a user from the root account using the command **"useradd username"**.
- Now you can open an existing user account using the command **"su username"**.
- Open the Linux terminal and type the following commands to create a user.

```
$ su
password:
# useradd hadoop
# passwd hadoop
New passwd:
Retype new passwd
```

SSH Setup and Key Generation

SSH setup is required to perform different operations on a cluster such as starting, stopping, and distributed daemon shell operations. To authenticate different users of Hadoop, it is required to provide public/private key pair for a Hadoop user and share it with different users.

The following commands are used to generate a key value pair using SSH, copy the public keys form id_rsa.pub to authorized_keys, and provide owner, read and write permissions to authorized_keys file respectively.

```
$ ssh-keygen -t rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys
```

Verifying ssh

```
ssh localhost
```

Installing Java

Java is the main prerequisite for Hadoop and HBase. First of all, you should verify the existence of Java in your system using “java -version”. The syntax of Java version command is given below.

```
$ java -version
```

It should produce the following output.

```
java version "1.7.0_71"
Java(TM) SE Runtime Environment (build 1.7.0_71-b13)
Java HotSpot(TM) Client VM (build 25.0-b02, mixed mode)
```

Downloading Hadoop

```
hadoop version
```

It should produce the following output:

```
Hadoop 2.6.0
Compiled by jenkins on 2014-11-13T21:10Z
Compiled with protoc 2.5.0
From source with checksum 18e43357c8f927c0695f1e9522859d6a
This command was run using /home/hadoop/hadoop/share/hadoop/common/hadoopcommon-2.6.0.jar
```

Installing Hadoop

Install Hadoop in any of the required modes. Here, we are demonstrating HBase functionalities in pseudo-distributed mode, therefore install Hadoop in pseudo-distributed mode.

Follow the steps given below to install **Hadoop 2.4.1** on your system.

Step 1: Setting up Hadoop

You can set Hadoop environment variables by appending the following commands to **~/.bashrc** file.

```
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME

export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native

export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_INSTALL=$HADOOP_HOME
```

Now, apply all changes into the currently running system.

```
$ source ~/.bashrc
```

8.7 Case studies of voluminous data

"Service Sales in Telecom billing sector Data Analysis tools shattered myths & set the perspective right"

Overview

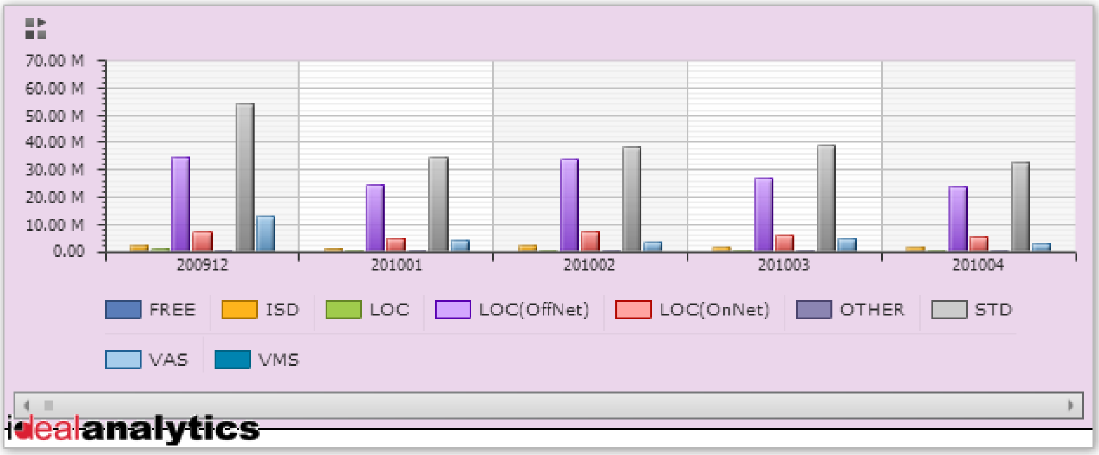
Telecom usage billing is a sale of service. Mobiles, Internet usage, broadband usage, along with various other communication services are billed to the customer depending on their usage. There are two broad-brush categories of billing – Pre-Paid & Post-Paid. Many types & categories of services are there within the two broad categories. Their usage is measured in seconds or pulses and customers pay based on that. Sales departments of various service providers invest in campaigns and advertisements and measure the sales they acquire from the customers. Meticulous accounting is the key to bill a customer and gain the confidence and loyalty of each customer.

The company in the Case study:

A giant service provider in telecom in India who has products in all possible wireless communication technology is the case we could reach out with the tool Ideal Analytics. Their business analysts analyzed the data and them and our team in Ideal Analytics Solutions (P) Ltd, were really amazed to find knowledge snippets that not only shattered our commonly acquired miss-conceptions but did really find the quantitative differential values within the different ranges, dimensions and other measures.

Experience of Ideal Analytics:
Ideal Analytics provided the analysts the tool and the know-how to go-about with the tool. The analysts of the company had the privilege to run & familiarize with various case studies, examples and to-n-fro discussions. They embarked on their newly gain knowledge and within weeks they came out with startling inferences- the inferences that were not only mathematically interesting but business wise very revealing and path breaking.

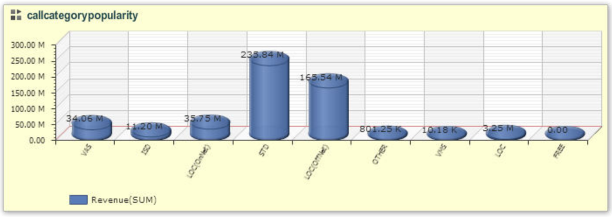
What startled us?
In this industry the overarching fact is the revenue earned. This single fact is qualified by many dimensions starting from Line-of-Business [Pre/Post paid], category of service, call-direction, even regions and many others. So the primary approach was to bring up a kind of comparison of the revenue earnings qualified through these dimensions.
Then we found out some general tests in terms of ratios and coefficients and did the next deep level of comparison.

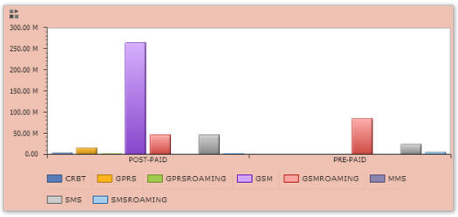


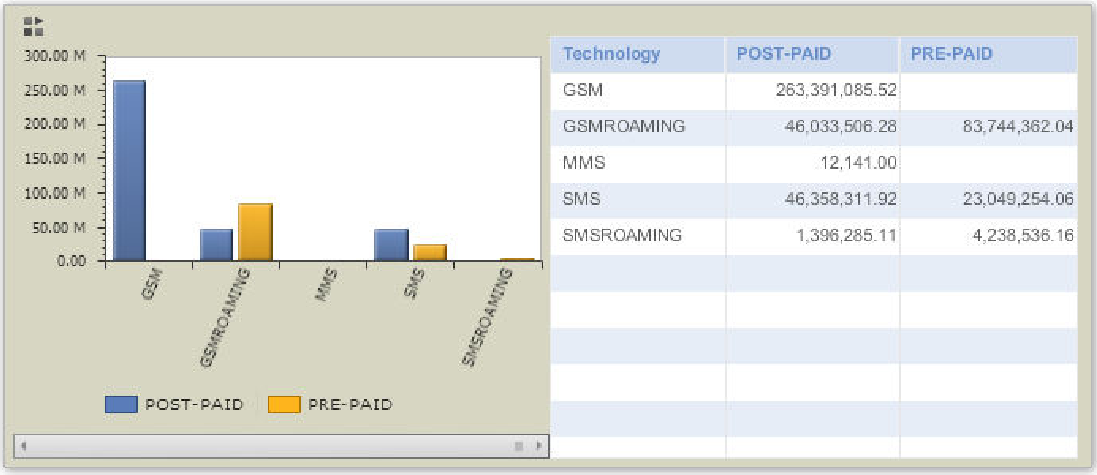
We then classified the results with respect to derived measures and found the variations from generally held misconceptions.

1. We found that even today Post-paid connections are more popular and more revenue giving than the Pre-paid ones.
2. We figured that only in case of GSM roaming the pre-paid ones triumph over the post-paid ones. This might tell us that people use Post-Paid connections for resident use and use a different pre-paid SIM when they are on the move, this way they keep a tab on their usage and cost. We have found that STD calls bring in more revenue than other types of calls.
3. We found that the niche and hi-tech services are not that revenue earning compared to the traditional and simple calls.
4. Our popular misconception was that the young subscribers actually are the most revenue givers- it is not; even fancy type services are far lower in the significance rank – at least for this company.

Through our tool we could compare between two facts i.e. measuring one fact when the other fact is reconfigured as a dimension. We found a complex curve that does not follow any regular polynomial expression. But we could easily extrapolate the last values in a small neighborhood and can predict the next value [in that small neighborhood]. This had been found to be more accurate than a statistically averaged out trend through normal correlation exercises. So, this kind of data depiction has provided more significant predictions, although in a proximate neighborhood assuring no fast and drastic changes take place. These are very important decision enabling points that actually changed the companies' orientation in service and re-focusing much beyond the grapevine gossip that we most often confuse as market-intelligence. Empirical data and their flexible analysis have given more wisdom than hearsay in Sales of services in Telecom industry.







Technology	POST-PAID	PRE-PAID
GSM	263,391,085.52	
GSMROAMING	46,033,506.28	83,744,362.04
MMS	12,141.00	
SMS	46,358,311.92	23,049,254.06
SMSROAMING	1,396,285.11	4,238,536.16

Question no 1: What types of data can be called as voluminous data in Telecom?
Question no 2: How can you access the data and what types of techniques can be used?
Question no 3: Is this possible to extract data from telecom server using cloud computing techniques?
Question no 4: How can you manage data in server?