BDA Question Bank for Module 1 and Module 2

1. List and explain any five essential Hadoop tools with their features.

Essential Hadoop Tools

Hadoop is an open source distributed processing framework which is at the center of a growing big data ecosystem. Used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications, Hadoop manages data processing and storage for big data applications and can handle various forms of structured and unstructured data.

1. Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size. We describe the architecture of HDFS and report on experience using HDFS to manage 40 petabytes of enterprise data at Yahoo.

Features:

a. Rack awareness allows consideration of a node’s physical location, when allocating storage and scheduling tasks

b. Minimal data motion. MapReduce moves compute processes to the data on HDFS and not the other way around. Processing tasks can occur on the physical node where the data resides. This significantly reduces the network I/O patterns and keeps most of the I/O on the local disk or within the same rack and provides very high aggregate read/write bandwidth.

c. Utilities diagnose the health of the files system and can rebalance the data on different nodes

d. Rollback allows system operators to bring back the previous version of HDFS after an upgrade, in case of human or system errors

e. Standby NameNode provides redundancy and supports high availability

f. Highly operable. Hadoop handles different types of cluster that might otherwise require operator intervention. This design allows a single operator to maintain a cluster of 1000s of nodes.

2. Hbase

HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java much like a typical MapReduce application. HBase does support writing applications in Avro, REST, and Thrift.

Features:

a. Linear and modular scalability.

b. Strictly consistent reads and writes.

c. Automatic and configurable sharding of tables

d. Automatic failover support between RegionServers.

e. Convenient base classes for backing Hadoop MapReduce jobs with Apache HBase tables.

f. Easy to use Java API for client access.

g. Block cache and Bloom Filters for real-time queries.

h. Query predicate push down via server side Filters

3. HIVE

The Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.Support for exporting metrics via the Hadoop metrics subsystem to files or Ganglia; or via JMX.

Features:

a. Indexing to provide acceleration, index type including compaction and Bitmap index as of 0.10, more index types are planned.

b. Different storage types such as plain text, RCFile, HBase, ORC, and others.

c. Metadata storage in an RDBMS, significantly reducing the time to perform semantic checks during query execution.

d. Operating on compressed data stored into Hadoop ecosystem, algorithm including gzip, bzip2, snappy, etc.

e. Built-in user defined functions (UDFs) to manipulate dates, strings, and other data-mining tools. Hive supports extending the UDF set to handle use-cases not supported by built-in functions.

f. SQL-like queries (Hive QL), which are implicitly converted into map-reduce jobs.

4. Sqoop

Sqoop is a tool designed to transfer data between Hadoop and relational databases. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.

Features:

a. Connecting to database server
b. Controlling parallelism
c. Controlling the import process
d. Import data to hive
e. Import data to Hbase

5. Pig

Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin

Features:

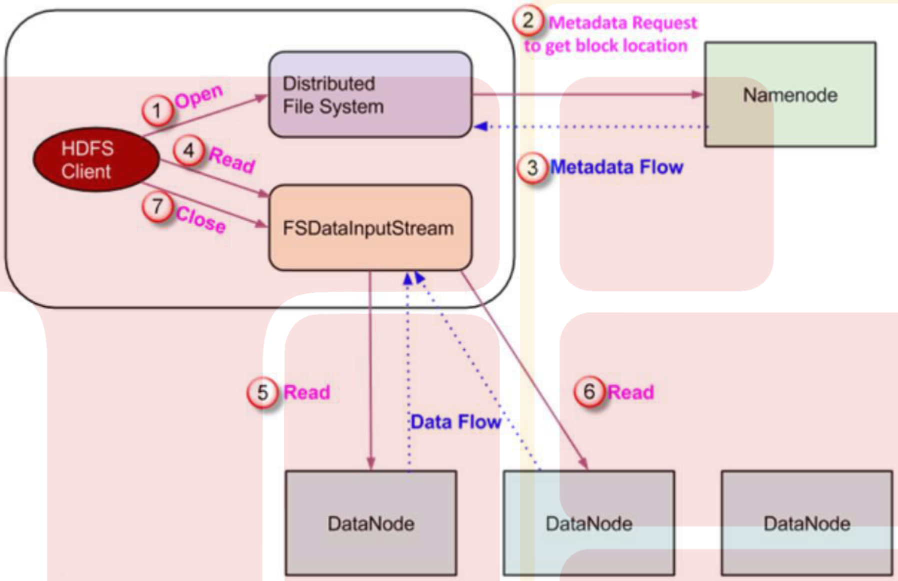
a. Ease of programming.
b. It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
c. Optimization opportunities.
d. The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
e. Extensibility. Users can create their own functions to do special-purpose processing.

2. Explain HDFS Read and write operations in detail.

HDFS is a distributed file system for storing very large data files, running on clusters of commodity hardware. It is fault tolerant, scalable, and extremely simple to expand. Hadoop comes bundled with **HDFS** (**Hadoop Distributed File Systems**). When data exceeds the capacity of storage on a single physical machine, it becomes essential to divide it across a number of separate machines. A file system that manages storage specific operations across a network of machines is called a distributed file system. HDFS is one such software.

Read Operation In HDFS

Data read request is served by HDFS, NameNode, and DataNode. Let's call the reader as a 'client'. Below diagram depicts file read operation in Hadoop.



1.A client initiates read request by calling **'open()'** method of FileSystem object; it is an object of type **DistributedFileSystem**.

2. This object connects to namenode using RPC and gets metadata information such as the locations of the blocks of the file. Please note that these addresses are of first few blocks of a file.
3. In response to this metadata request, addresses of the DataNodes having a copy of that block is returned back.

4. Once addresses of DataNodes are received, an object of type **FSDataInputStream** is returned to the client. **FSDataInputStream** contains **DFSInputStream** which takes care of interactions with DataNode and NameNode. In step 4 shown in the above diagram, a client invokes **'read()'** method which causes **DFSInputStream** to establish a connection with the first DataNode with the first block of a file.

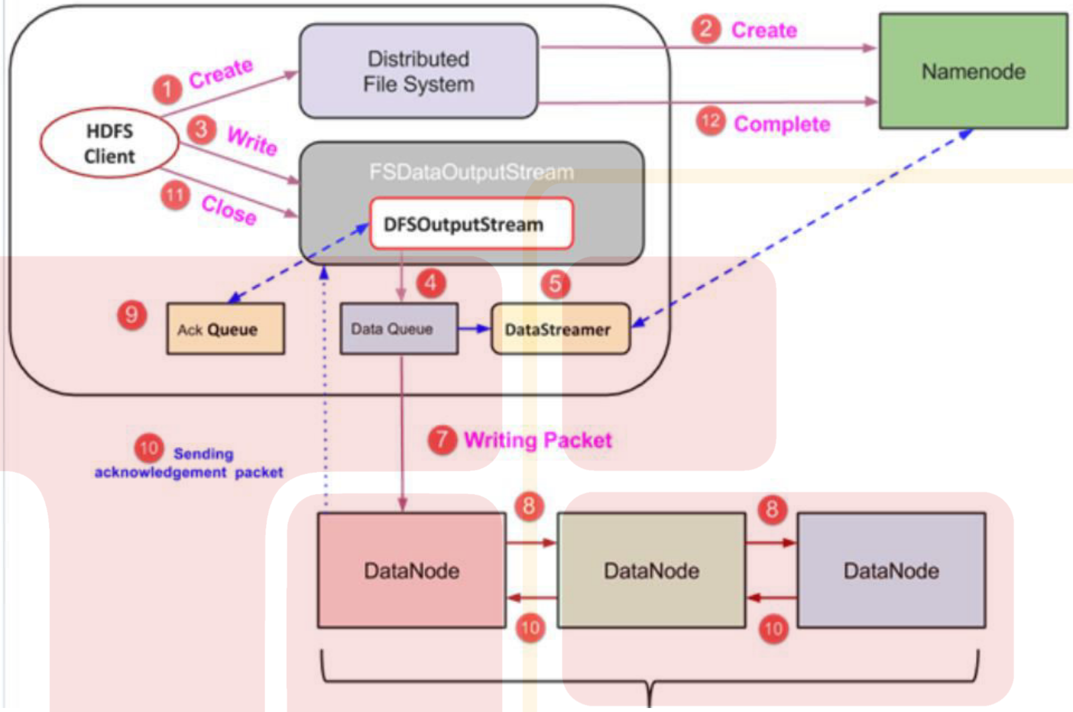
5. Data is read in the form of streams wherein client invokes **'read()'** method repeatedly. This process of **read()** operation continues till it reaches the end of block.

6. Once the end of a block is reached, DFSInputStream closes the connection and moves on to locate the next DataNode for the next block

7. Once a client has done with the reading, it calls **a close()** method.

Write Operation In HDFS

Lets understand how data is written into HDFS through files.



1. A client initiates write operation by calling 'create()' method of DistributedFileSystem object which creates a new file - Step no. 1 in the above diagram.
   2. DistributedFileSystem object connects to the NameNode using RPC call and initiates new file creation. However, this file creates operation does not associate any blocks with the file. It is the responsibility of NameNode to verify that the file (which is being created) does not exist alreadyand a client has correct permissions to create a new file. If a file already exists or client does nothave sufficient permission to create a new file, then **IOException** is thrown to the client.Otherwise, the operation succeeds and a new record for the file is created by the NameNode.
   3. Once a new record in NameNode is created, an object of type FSDataOutputStream is returned to the client. A client uses it to write data into the HDFS. Data write method is invoked (step 3 in the diagram).

   4. FSDataOutputStream contains DFSOutputStream object which looks after communication withDataNodes and NameNode. While the client continues writing data, **DFSOutputStream** continues creating packets with this data. These packets are enqueuedinto a queue which is called as **DataQueue**.

   5. There is one more component called **DataStreamer** which consumes this **DataQueue**. DataStreamer also asks NameNode for allocation of new blocks thereby picking desirable DataNodes to be used for replication.

6. Now, the process of replication starts by creating a pipeline using DataNodes. In our case, wehave chosen a replication level of 3 and hence there are 3 DataNodes in the pipeline.

7. The DataStreamer pours packets into the first DataNode in the pipeline.

8. Every DataNode in a pipeline stores packet received by it and forwards the same to the second DataNode in a pipeline.

9. Another queue, 'Ack Queue' is maintained by DFSOutputStream to store packets which are waiting for acknowledgment from DataNodes.

10. Once acknowledgment for a packet in the queue is received from all DataNodes in the pipeline, it is removed from the 'Ack Queue'. In the event of any DataNode failure, packets from this queueare used to reinitiate the operation.

11. After a client is done with the writing data, it calls a close() method (Step 9 in the diagram) Call toclose(), results into flushing remaining data packets to the pipeline followed by waiting for acknowledgment.

12. Once a final acknowledgment is received, NameNode is contacted to tell it that the file write operation is complete.

3. What is Map Reduce? Explain how it works with example.

What is MapReduce in Hadoop?

MapReduce is a programming model suitable for processing of huge data. Hadoop is capable of running MapReduce programs written in various languages: Java, Ruby, Python, and C++. MapReduce programs are parallel in nature, thus are very useful for performing large-scale data analysis using multiple machines in the cluster.

MapReduce programs work in two phases:

1. Map phase
2. Reduce phase.

An input to each phase is **key-value** pairs. In addition, every programmer needs to specify two functions: **map function** and **reduce function**.

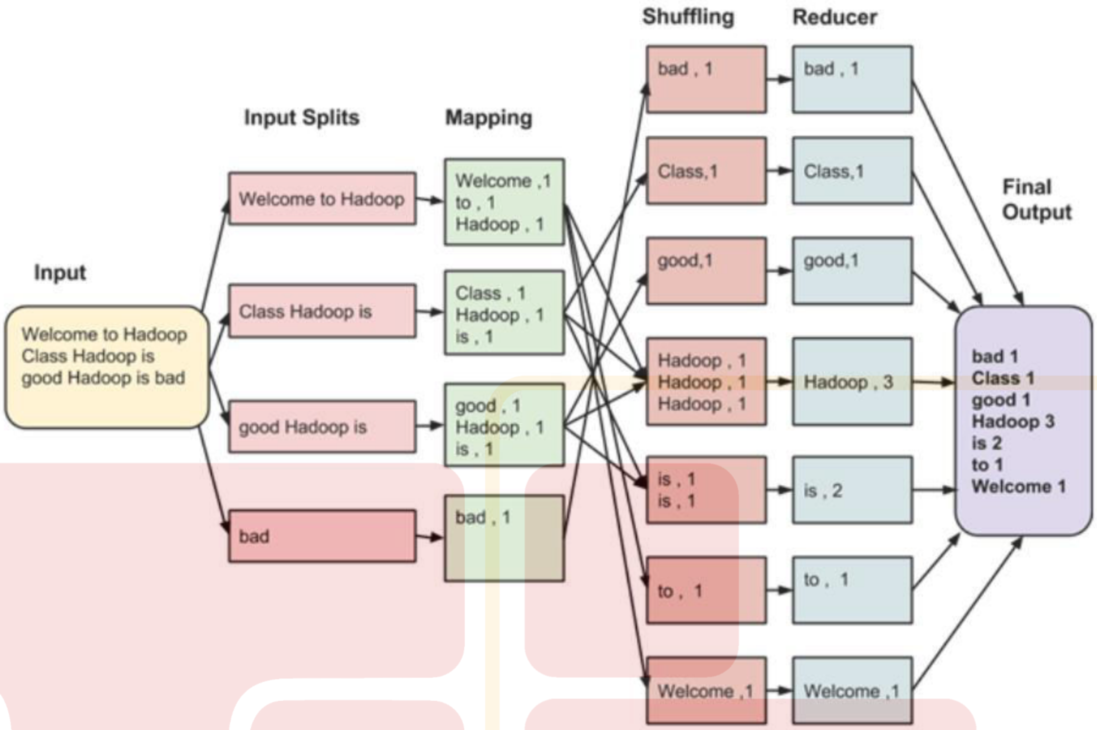
How MapReduce Works:-

The whole process goes through four phases of execution namely, splitting, mapping, shuffling, and reducing.

Let's understand this with an example –

Consider you have following input data for your Map Reduce Program

Welcome to Hadoop Class
Hadoop is good
Hadoop is bad



The final output of the MapReduce task is

bad	1
Class	1
good	1
Hadoop	3
is	2
to	1
Welcome	1

The data goes through the following phases

Input Splits:

An input to a MapReduce job is divided into fixed-size pieces called **input splits** Input split is a chunk of the input that is consumed by a single map

Mapping

This is the very first phase in the execution of map-reduce program. In this phase data in each split is passed to a mapping function to produce output values. In our example, a job of mapping phase is to count a number of occurrences of each word from input splits (more details about input-split is given below) and prepare a list in the form of <word, frequency>

Shuffling

This phase consumes the output of Mapping phase. Its task is to consolidate the relevant records from Mapping phase output. In our example, the same words are clubed together along with their respective frequency.

Reducing

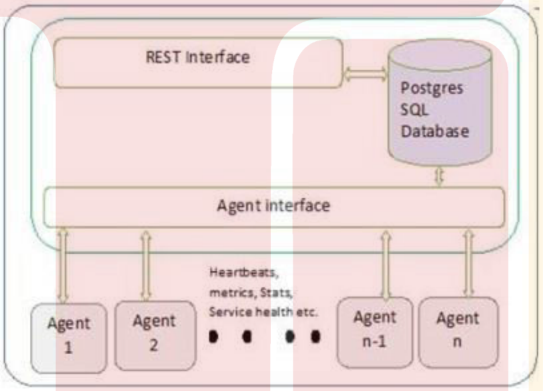
In this phase, output values from the Shuffling phase are aggregated. This phase combines values from Shuffling phase and returns a single output value. In short, this phase summarizes the complete dataset.

In our example, this phase aggregates the values from Shuffling phase i.e., calculates total occurrences of each word.

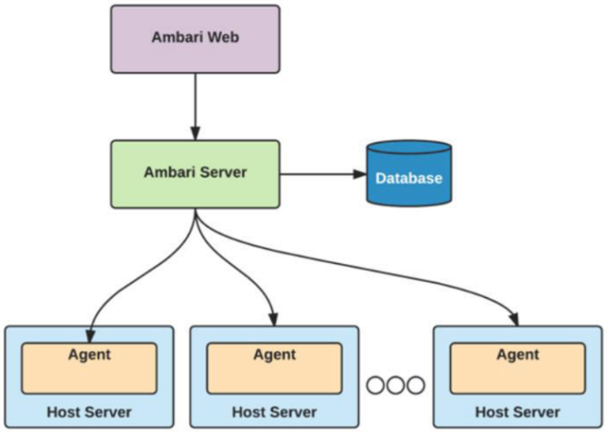
4. Explain architecture of Apache Ambari.

Apache Ambari architecture

Ambari provides intuitive and REST APIs that automate the operations in the Hadoop cluster. Its consistent and secure interface allows it to be fairly efficient in operational control. Its easy and user-friendly interface efficiently diagnoses the health of Hadoop cluster using an interactive dashboard.



To have a better understanding of how Ambari works, let’s look at the detailed architecture of Ambari, in the following diagram:



Apache Ambari follows a master/slave architecture where the master node instructs the slave nodes to perform certain actions and report back the state of every action. The master node is responsible for keeping track of the state of the infrastructure. To do this, the master node uses a database server, which can be configured during setup time.

These are the following applications in Apache Ambari, at the core:

- Ambari server
- The Ambari agent
- Ambari web UI
- Database

1. Ambari server

The entry point for all administrative activities on the master server is known as Ambari server. It is a shell script. Internally this script uses Python code, **ambari-server.py** and routes all the requests to it. Ambari server consists of several entry points that are available when passed different parameters to the Ambari-server program like:

- Daemon management
- Software upgrade
- Software setup
- LDAP (Lightweight Direct Access Protocol) /PAM (Pluggable Authentication Module) /Kerberos management
- Ambari backup and restore
- Miscellaneous options

2. Ambari Agent

The Ambari Agent runs on all the nodes that we want to manage with Ambari. This program periodically heartbeats to the master node. By using this agent, Ambari-server executes many of the tasks on the servers.

3. Ambari web interface

Ambari web interface is one of the powerful features of Ambari application. The web application is through the server of Ambari program which is running on the master host exposed on port 8080. You can access this application and this application is protected by authentication. Also, you can control and view all aspects of your Hadoop Cluster, once you log in to the web portal.

4. Database

Ambari supports multiple RDBMS (Relational Database Management Systems) to keep track of the state of the entire Hadoop infrastructure. Also, you can choose the database you want to use during the setup of the Ambari for the first time.

Ambari supports these following databases at the time of writing:

- PostgreSQL
- Oracle
- MySQL or MariaDB
- Embedded PostgreSQL

- Microsoft SQL Server
- SQL Anywhere
- Berkeley DB

This technology is preferred by the big data developers as it is quite handy and comes with a step-by-step guide allowing easy installation on the Hadoop cluster. Its pre-configured key operational metrics provide quick look into the health of Hadoop core, i.e., HDFS and MapReduce along with the additional components such as Hive, HBase, HCatalog, etc. Ambari sets up a centralized security system by incorporating Kerberos and Apache Ranger into the architecture. The RESTful APIs monitor the information as well as integrate the operational tools. Its user-friendliness and interactivity has brought it in the range of top ten open source technologies for Hadoop cluster.

5. Discuss the features and benefits of Apache Ambari.

Features of Apache Ambari

Following are some of features of Ambari. Read on to understand how the tool is expertly used in big data arena.
1. Platform independent – Apache Ambari runs in Windows, Mac and many other platforms as it architecturally supports any hardware and software systems. Other platforms where Ambari runs are Ubuntu, SLES, RHEL etc. Those components which are dependent on a platform like yum, rpm packages, debian packages ought to be plugged with well defined interfaces.
2.Pluggable component – Any current Ambari application can be customized. Any specific tools and technologies ought to be encapsulated by pluggable components. The goal of pluggability doesn't encompass standardization of inter-component.
3.Version management and upgrade – Ambari itself maintains versions and hence there is no need of external tools like Git. If any Ambari application is to be upgraded or even Ambari is to be upgraded then doing it fairly easy.
4.Extensibility – We can extend the functionality of existing Ambari applications by adding different view components.
5.Failure recovery – Assume you are working on an Ambari application and something wrong happens. Then the system should gracefully recover from it. If you are a Windows user you can relate well to this. You might have worked on word file and suddenly there is a power outage. After turning the system on there will be an autosaved version of the document when you run the MS word.
6.Security – The Ambari application comes with robust security and it can sync with LDAP over the active directory.

Benefits of using Apache Ambari

This is given with respect to Hortonworks Data Platform (HDP). Ambari eliminates the need for manual tasks used to watch over Hadoop operations. It gives a simple secure platform for provisioning, managing and monitoring HDP deployments. Ambari is an easy to use Hadoop management UI and is solidly backed by REST APIs.

It provides numerous benefits like:

1.Installation, configuration and management is way simplified

Ambari can efficiently create Hadoop clusters at scale. It wizard driven approach lets the configuration be automated as per the environment so that the performance is optimal. Master slave and client components are assigned to configuring services. It is also used to install, start and test the cluster. Configuration blueprints give recommendations to those seeking a hands-on approach. The blue print of an ideal cluster is stored. How it is provisioned is clearly traced. This is then used to automate the creation of successive clusters without any user interaction. Blueprints also preserve and ensure the application of best practices across different environments. Ambari also provides rolling upgrade feature where running clusters can be updated on the go with maintenance releases and feature bearing releases and therefore there is no unnecessary downtime. When there are large clusters involved then rolling updates are simply not possible in which case express updates are used. Here the downtime is there but is minimum as when the update is manual. Both rolling and express updates are free of manual updates.

2. Centralized security and application

The complexity of cluster security configuration and administration is greatly reduced by Ambari which is among the components of Hadoop ecosystem. The tool also helps with automated setup of advanced security constructs like Kerboros and Ranger.

3. Complete visibility to cluster health

Through this tool you can monitor your cluster’s health and availability. An easily customized web based dashboard has metrics that give status information for each service in the cluster like HDFS, YARN and HBase. The tool also helps with garnering and visualizing critical operational metrics for troubleshooting and analysis. Ambari predefines alerts which integrate with existing enterprise monitoring tools that monitor cluster components and hosts as per specified check intervals. Through the browser interface users can browse alerts for their clusters, search and filter alerts. They can also view and modify alert properties alert instances associated with that definition.

4.Metrics visualization and dashboarding

In this Apache Ambari tutorial you can know that it provides scalable low latency storage system for Hadoop component metrics. To pick the metrics of Hadoop which truly matter requires considerable expertise and understanding on how the components work with each other and with themselves. Grafana is a leading graph and dashboard builder which simplifies the metrics reviewing process. This is included with Ambari metrics along with HDP.

5.Extensibility and customization

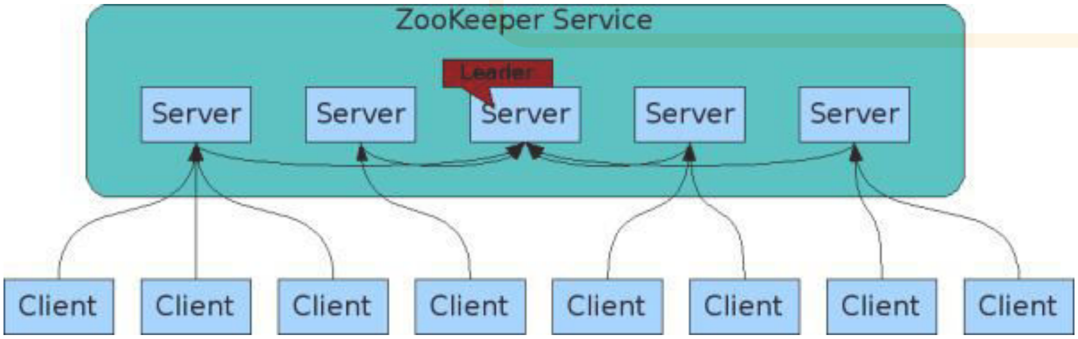
Ambari lets a developer to work on Hadoop gracefully in one's enterprise setup. Ambari leverages the large innovative community which improve upon the tool and it also eliminates vendor lock in. REST APIs along with Ambari Stacks and Views allows extensive flexibility for customization of HDP implementation.

Ambari Stacks wrap lifecycle control layer to rationalize operations over a broad set of services. This includes a consistent approach which the Ambari technology uses to manage different types of services like install, start, configure, status, stop. When provisioning, cluster install experience is rationalized across a set of services by Stacks technology. A natural extension point for operators is provided by the Stacks to plug in newly created services that can perform alongside Hadoop.

Third parties can plug in their views through Ambari views. A view is an application that is deployed into Ambari container where it offers UI capabilities to be plugged in to give out custom visualization, management and monitoring

6. Explain how Zookeeper works. How is Apache Ambari different from Zookeeper?

ZooKeeper, while being a coordination service for distributed systems, is a distributed application on its own. ZooKeeper follows a simple client-server model where clients are nodes (i.e., machines) that make use of the service, and servers are nodes that provide the service. A collection of ZooKeeper servers forms a ZooKeeper ensemble. Once a ZooKeeper ensemble starts after the leader election process, it will wait for the clients to connect. At any given time, one ZooKeeper client is connected to one ZooKeeper server. Each ZooKeeper server can handle a large number of client connections at the same time. Each client periodically sends pings to the ZooKeeper server it is connected to let it know that it is alive and connected. The ZooKeeper server in question responds with an acknowledgment of the ping, indicating the server is alive as well. When the client doesn't receive an acknowledgment from the server within the specified time, the client connects to another server in the ensemble, and the client session is transparently transferred over to the new ZooKeeper server.



ZooKeeper has a file system-like data model composed of znodes. Think of znodes (ZooKeeper data nodes) as files in a traditional UNIX-like system, except that they can have child nodes. Another way to look at them is as directories that can have data associated with themselves. Each of these directories is called a znode. The znode hierarchy is stored in memory within each of the

ZooKeeper servers. This allows for scalable and quick responses to reads from the clients. Each ZooKeeper server also maintains a transaction log on the disk, which logs all write requests. ZooKeeper server must sync transactions to disk before it returns a successful response. The default maximum size of data that can be stored in a znode is 1 MB. Zookeeper should only be used as a storage mechanism for the small amount of data required for providing reliability, availability, and coordination to your distributed application.

Features:

a. Fast. ZooKeeper is especially fast with workloads where reads to the data are more common than writes. The ideal read/write ratio is about 10:1.

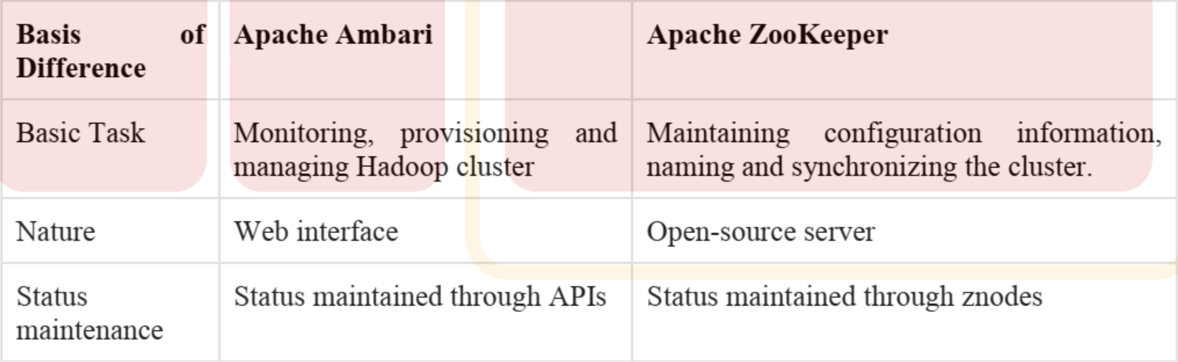
b. Reliable. ZooKeeper is replicated over a set of hosts (called an ensemble) and the servers are aware of each other. As long as a critical mass of servers is available, the ZooKeeper service will also be available. There is no single point of failure.

c. Simple. ZooKeeper maintain a standard hierarchical name space, similar to files and directories.

d. Ordered. The service maintains a record of all transactions, which can be used for higher-level abstractions, like synchronization primitives.

How is Ambari different from ZooKeeper?

This description may confuse you as Zookeeper performs the similar kind of tasks. But, there is a huge difference between the tasks performed by these two technologies if looked closely.



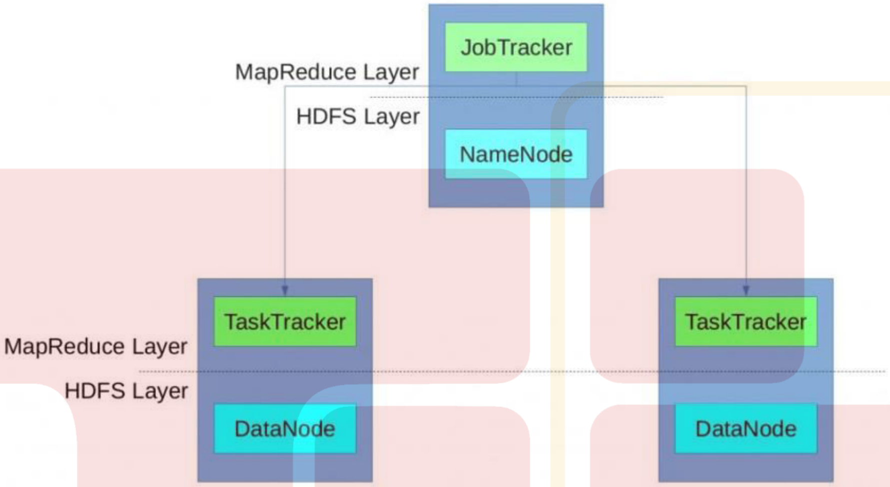
Basis of Difference	**Apache Ambari**	**Apache ZooKeeper**
Basic Task	Monitoring, provisioning and managing Hadoop cluster	Maintaining configuration information, naming and synchronizing the cluster.
Nature	Web interface	Open-source server
Status maintenance	Status maintained through APIs	Status maintained through znodes

7. Discuss YARN architecture in detail.

YARN Architecture

YARN (Yet Another Resource Negotiator) has been introduced to Hadoop with version 2.0 and solves a few issues with the resources scheduling of MapReduce in version 1.0. In order to understand the benefits of YARN, we have to review how resource scheduling worked in version 1.0.

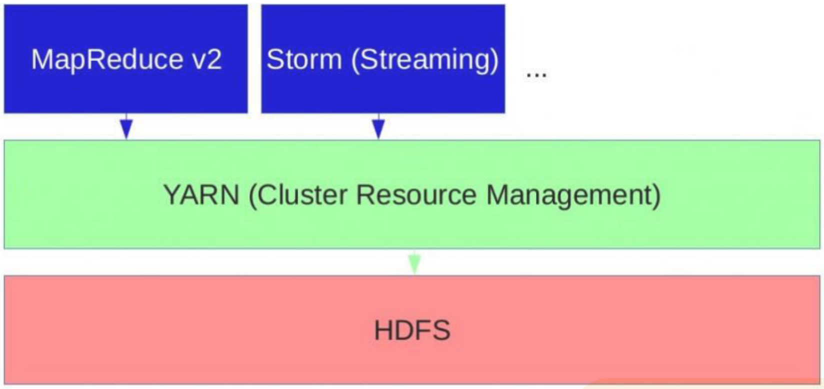
A MapReduce job is split by the framework into tasks (Map tasks, Reducer tasks) and each task is run on of the DataNode machines on the cluster. For the execution of tasks, each DataNode machine provided a predefined number of slots (map slots, reducers slots). The JobTracker was responsible for the reservation of execution slots for the different tasks of a job and monitored their execution. If the execution failed, it reserved another slot and re-started the task. It also cleaned up temporary resources and make the reserved slot available to other tasks.



The fact that there was only one JobTracker instance in Hadoop 1.0 led to the problem that the whole MapReduce execution could fail, if the the JobTracker fails (single point of failure). Beyond that, having only one instance of the JobTracker limits scalability (for very large clusters with thousands of nodes).

The concept of predefined map and reduce slots also caused resource problems in case all map slots are used while reduce slots are still available and vice versa. In general it was not possible to reuse the MapReduce infrastructure for other types of computation like real-time jobs. While MapReduce is a batch framework, applications that want to process large data sets stored in HDFS and immediately inform the user about results cannot be implemented with it. Beneath the fact that MapReduce 1.0 did not offer realtime provision of computation results, all other types of applications that want to perform computations on the HDFS data had to be implemented as Map and Reduce jobs, which was not always possible.

Hence Hadoop 2.0 introduced YARN as resource manager, which no longer uses slots to manage resources. Instead nodes have "resources" (like memory and CPU cores) which can be allocated by applications on a per request basis. This way MapReduce jobs can run together with non-MapReduce jobs in the same cluster.



The heart of YARN is the Resource Manager (RM) which runs on the master node and acts as a global resource scheduler. It also arbitrates resources between competing applications. In contrast to the Resource Manager, the Node Managers (NM) run on slave nodes and communicates with the RM. The NodeManager is responsible for creating containers in which the applications run, monitors their CPU and memory usage and reports them to the RM.

Each application has its own ApplicationMaster (AM) which runs within a container and negotiates resources with the RM and works with the NM to execute and monitor tasks. The MapReduce implementation of Hadoop 2.0 therefore ships with an AM (named MRAppMaster) that requests containers for the execution of the map tasks from the RM, receives the container IDs from the RM and then executes the map tasks within the provided containers. Once the map tasks have finished, it requests new containers for the execution of the reduce tasks and starts their execution on the provided containers.

If the execution of a task fails, it is restarted by the ApplicationMaster. Should the ApplicationMaster fail, the RM will attempt to the restart the whole application (up to two times per default). Therefore the ApplicationMaster can signal if it supports job recovery. In this case the ApplicationMaster receives the previous state from the RM and can only restart incomplete tasks. If a NodeManager fails, i.e the RM does not receive any heartbeats from it, it is removed from the list of active nodes and all its tasks are treated as failed. In contrast toversion 1.0 of Hadoop, the ResourceManager can be configured for High Availability.

8. Explain the benefits of Big data processing and features of Hadoop.

Benefits of Big Data Processing

Ability to process Big Data brings in multiple benefits, such as-

- **Businesses can utilize outside intelligence while taking decisions**

  Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.

- **Improved customer service**
  Traditional customer feedback systems are getting replaced by new systems designed with

Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

• **Early identification of risk to the product/services, if any**

• **Better operational efficiency**

Features of Hadoop

- **Suitable for Big Data Analysis**

As Big Data tends to be distributed and unstructured in nature, HADOOP clusters are best suited for analysis of Big Data. Since it is processing logic (not the actual data) that flows to the computing nodes, less network bandwidth is consumed. This concept is called as **data locality concept** which helps increase the efficiency of Hadoop based applications.

• Scalability

HADOOP clusters can easily be scaled to any extent by adding additional cluster nodes and thus allows for the growth of Big Data. Also, scaling does not require modifications to application logic.

• Fault Tolerance

HADOOP ecosystem has a provision to replicate the input data on to other cluster nodes. That way, in the event of a cluster node failure, data processing can still proceed by using data stored on another cluster node.

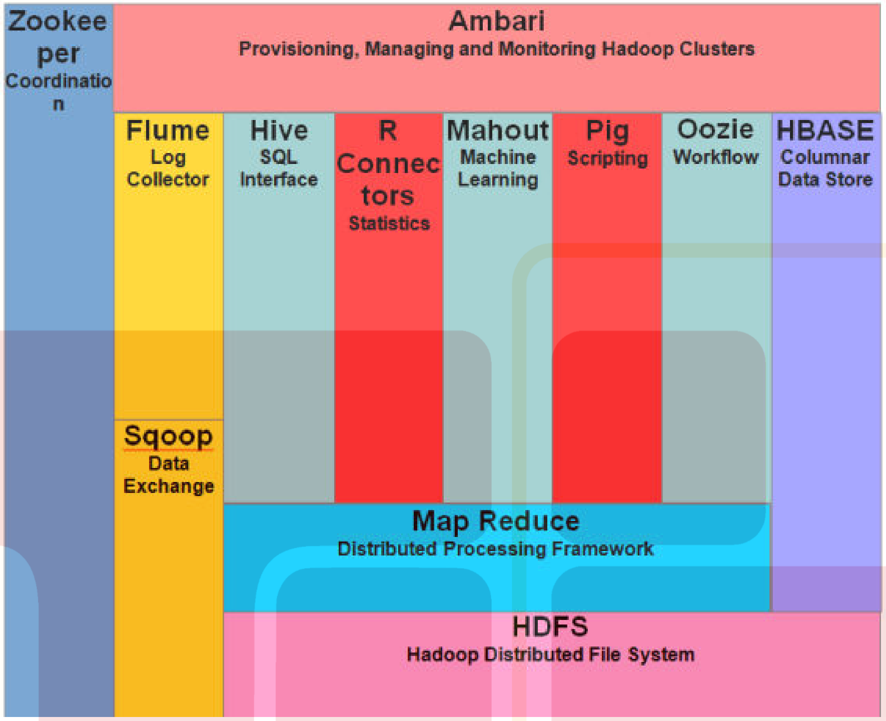
9. Explain Hadoop Architecture and its Ecosystem Components

Apache Hadoop is an open source software framework used to develop data processing applications which are executed in a distributed computing environment. Applications built using HADOOP are run on large data sets distributed across clusters of commodity computers. Commodity computers are cheap and widely available. These are mainly useful for achieving greater computational power at low cost.

Similar to data residing in a local file system of a personal computer system, in Hadoop, data resides in a distributed file system which is called as a **Hadoop Distributed File system**. The processing model is based on **'Data Locality'** concept wherein computational logic is sent to cluster nodes(server) containing data. This computational logic is nothing, but a compiled version of a program written in a high-level language such as Java. Such a program, processes data stored in Hadoop HDFS. Computer cluster consists of a set of multiple processing units (storage disk + processor) which are connected to each other and acts as a single system.

Hadoop EcoSystem and Components

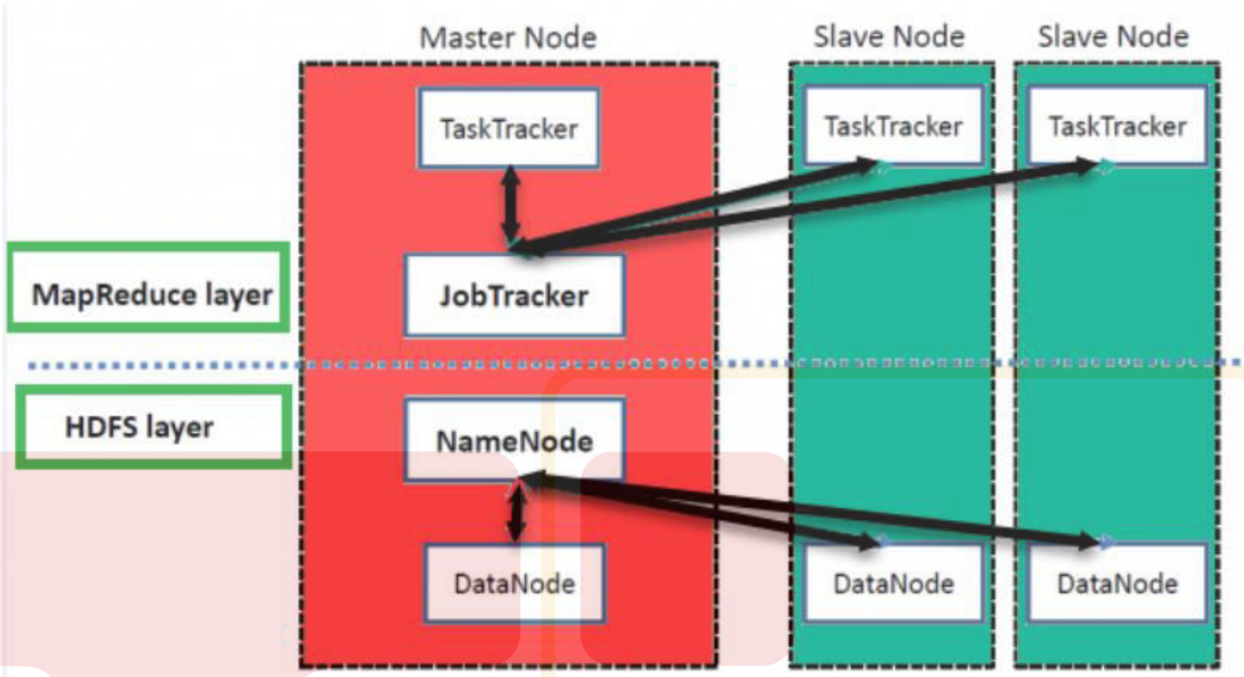
Below diagram shows various components in the Hadoop ecosystem



Apache Hadoop consists of two sub-projects –

1. **Hadoop MapReduce:** MapReduce is a computational model and software framework for writing applications which are run on Hadoop. These MapReduce programs are capable of processing enormous data in parallel on large clusters of computation nodes.

2. **HDFS** (**Hadoop Distributed File System**): HDFS takes care of the storage part of Hadoop applications. MapReduce applications consume data from HDFS. HDFS creates multiple replicasof data blocks and distributes them on compute nodes in a cluster. This distribution enablesreliable and extremely rapid computations. Although Hadoop is best known for MapReduce and its distributed file system- HDFS, the term is also used for a family of related projects that fall under the umbrella of distributed computing and large-scaledata processing. Other Hadoop-related projects at Apache include are **Hive, HBase, Mahout, Sqoop, Flume, and ZooKeeper.**



Hadoop has a Master-Slave Architecture for data storage and distributed data processing using MapReduce and HDFS methods.

NameNode:
NameNode represented every files and directory which is used in the namespace
DataNode:
DataNode helps you to manage the state of an HDFS node and allows you to interacts with the blocks
MasterNode:
The master node allows you to conduct parallel processing of data using Hadoop MapReduce.
Slave node:
The slave nodes are the additional machines in the Hadoop cluster which allows you to store data to conduct complex calculations. Moreover, all the slave node comes with Task Tracker and a DataNode. This allows you to synchronize the processes with the NameNode and Job Tracker respectively.In Hadoop, master or slave system can be set up in the cloud or on-premise.

10. Explain Hadoop installation steps.
Installation steps
Part 1
Step 1:

- sudo addgroup hadoop_
- sudo adduser –ingroup Hadoop_ hduser_
- Enter your password, name and other details.
- **NOTE:** There is a possibility of below-mentioned error in this setup and installation process.
- **"hduser is not in the sudoers file. This incident will be reported."**
- This error can be resolved by Login as a root user
- Execute the command
- Sudo adduser hduser_ sudo
- Re-login as hduser

Step 2) **Configure SSH**

In order to manage nodes in a cluster, Hadoop requires SSH access

- First, switch user, enter the following command

```
su -hduser_
```

- This command will create a new key.

```
ssh-keygen -t rsa -P “”
```

- Enable SSH access to local machine using this key.

```
cat $HOME/.ssh/id_rsa-pub >>$HOME/.ssh/authorized_keys
```

- Now test SSH setup by connecting to localhost as 'hduser' user.

```
ssh localhost
```

- **Note:** Please note, if you see below error in response to 'ssh localhost', then there is a possibility
that SSH is not available on this system-
- **To resolve this -** Purge SSH using, sudo apt-get purge openssh -server

It is good practice to purge before the start of installation

- Install SSH using the command- sudo apt-get install openssh-server

Step 3) Next step is to Download Hadoop

Select Stable

Select the tar.gz file (not the file with src)

- Once a download is complete, navigate to the directory containing the tar file

Enter, sudo tar xzf Hadoop-2.2.0.tar/gz

- Now, rename hadoop-2.2.0 as hadoop

```
Sudo mv Hadoop-2.2.0 hadoop
Sudo chown -R hduser_:hadoop_ Hadoop
```

Part 2) Configure Hadoop

Step 1) Modify **~/.bashrc** file

- Add following lines to end of file **~/.bashrc**
- #Set HADOOP_HOME
- export HADOOP_HOME=<Installation Directory of Hadoop>
- #Set JAVA_HOME
- export JAVA_HOME=<Installation Directory of Java>
- # Add bin/ directory of Hadoop to PATH

- export PATH=$PATH:$HADOOP_HOME/bin
- Now, source this environment configuration using below command

. ~/.bashrc

Step 2) Configurations related to HDFS

Set **JAVA_HOME** inside file **$HADOOP_HOME/etc/hadoop/hadoop-env.sh**

There are two parameters in **$HADOOP_HOME/etc/hadoop/core-site.xml** which need to be set-

1. 'hadoop.tmp.dir' - Used to specify a directory which will be used by Hadoop to store its data files.

2. 'fs.default.name' - This specifies the default file system.

To set these parameters, open core-site.xml

sudo gedit $HADOOP_HOME/etc/hadoop/core-site.xml

Copy below line in between tags <configuration></configuration>

```
<property>
<name>hadoop.tmp.dir</name>
<value>/app/hadoop/tmp</value>
<description>Parent directory for other temporary directories.</description>
</property>
<property>
<name>fs.defaultFS </name>
<value>hdfs://localhost:54310</value>
<description>The name of the default file system. </description>
</property>
```

- Navigate to the directory **$HADOOP_HOME/etc/Hadoop**
- Now, create the directory mentioned in core-site.xml
  sudo mkdir -p <Path of Directory used in above setting>
- Grant permissions to the directory
  sudo chown -R hduser_:Hadoop_ <Path of Directory created in above step>

sudo chmod 750 <Path of Directory created in above step>

Step 3) **Map Reduce Configuration**

- Before you begin with these configurations, lets set HADOOP_HOME path
- sudo gedit /etc/profile.d/hadoop.sh
- And Enter
  export HADOOP_HOME=/home/guru99/Downloads/Hadoop
- Next enter
  sudo chmod +x
  /etc/profile.d/hadoop.sh
- Exit the Terminal and restart again
  Type echo $HADOOP_HOME. To verify the path
- Now copy files

sudo cp $HADOOP_HOME/etc/hadoop/mapred-site.xml.template
$HADOOP_HOME/etc/hadoop/mapred-site.xml

- Open the **mapred-site.xml** file

```
sudo gedit $HADOOP_HOME/etc/hadoop/mapred-site.xml
```

- Add below lines of setting in between tags <configuration> and </configuration>

```
<property>
<name>mapreduce.jobtracker.address</name>
<value>localhost:54311</value>
<description>MapReduce job tracker runs at this host and port.
</description>
</property>
```

- Open **$HADOOP_HOME/etc/hadoop/hdfs-site.xml** as below,

```
sudo gedit $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

- Add below lines of setting between tags <configuration> and </configuration>

```
<property>
<name>dfs.replication</name>
<value>1</value>
<description>Default block replication.</description>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/hduser_/hdfs</value>
</property>
```

- Create a directory specified in above setting

```
sudo mkdir -p <Path of Directory used in above setting>
sudo mkdir -p /home/hduser_/hdfs
sudo chown -R hduser_:hadoop_ <Path of Directory created in above step>
sudo chown -R hduser_:hadoop_ /home/hduser_/hdfs
sudo chmod 750 <Path of Directory created in above step>
sudo chmod 750 /home/hduser_/hdfs
```

Step 4)

- Before we start Hadoop for the first time, format HDFS using below command

```
$HADOOP_HOME/bin/hdfs namenode -format
```

Step 5)

- Start Hadoop single node cluster using below command

```
$HADOOP_HOME/sbin/start-dfs.sh
$HADOOP_HOME/sbin/start-yarn.sh
```

- Using **'jps'** tool/command, verify whether all the Hadoop related processes are running or not.

If Hadoop has started successfully then an output of jps should show NameNode, NodeManager,
ResourceManager, SecondaryNameNode, DataNode.

Step 6) Stopping Hadoop

```
$HADOOP_HOME/sbin/stop-dfs.sh
$HADOOP_HOME/sbin/stop-yarn.sh
```

11. Explain JobTracker and Task Tracker in Hadoop.

JobTracker and TaskTracker are 2 essential process involved in MapReduce execution in MRv1 (or Hadoop version 1). Both processes are now deprecated in MRv2 (or Hadoop version 2) and replaced by Resource Manager, Application Master and Node Manager Daemons

Job Tracker –

1. JobTracker process runs on a separate node and not usually on a DataNode.
2. JobTracker is an essential Daemon for MapReduce execution in MRv1. It is replaced by ResourceManager/ApplicationMaster in MRv2.
3. JobTracker receives the requests for MapReduce execution from the client.
4. JobTracker talks to the NameNode to determine the location of the data.
5. JobTracker finds the best TaskTracker nodes to execute tasks based on the data locality (proximity of the data) and the available slots to execute a task on a given node.
6. JobTracker monitors the individual TaskTrackers and the submits back the overall status of the job back to the client.
7. JobTracker process is critical to the Hadoop cluster in terms of MapReduce execution.
8. When the JobTracker is down, HDFS will still be functional but the MapReduce execution can not be started and the existing MapReduce jobs will be halted.

TaskTracker –

1. TaskTracker runs on DataNode. Mostly on all DataNodes.
2. TaskTracker is replaced by Node Manager in MRv2.
3. Mapper and Reducer tasks are executed on DataNodes administered by TaskTrackers.
4. TaskTrackers will be assigned Mapper and Reducer tasks to execute by JobTracker.
5. TaskTracker will be in constant communication with the JobTracker signalling the progress of the task in execution.
6. TaskTracker failure is not considered fatal. When a TaskTracker becomes unresponsive, JobTracker will assign the task executed by the TaskTracker to another node.

12. List and explain limitations and solutions of Hadoop for Big Data Analytics

1. Issue with Small Files

Hadoop does not suit for small data. **(HDFS) Hadoop distributed file system** lacks the ability to efficiently support the random reading of small files because of its high capacity design.
Small files are the major problem in HDFS. A small file is significantly smaller than the **HDFS block** size (default 128MB). If we are storing these huge numbers of small files, HDFS can't handle this much of files, as HDFS is for working properly with a small number of large files for storing large data sets rather than a large number of small files. If there are too many small files, then the **NameNode** will get overload since it stores the namespace of HDFS.

Solution-

- Solution to this Drawback of Hadoop to deal with small file issue is simple. Just merge the small files to create bigger files and then copy bigger files to HDFS.
- The introduction of **HAR files** (Hadoop Archives) was for reducing the problem of lots of files putting pressure on the namenode's memory. By building a layered filesystem on the top of HDFS, HAR files works. Using the Hadoop archive command, HAR files are created, which runs a **MapReduce**job to pack the files being archived into a small number of HDFS files. Reading through files in a HAR is not more efficient than reading through files in HDFS. Since each HAR file access requires two index files read as well the data file to read, this makes it slower.
- **Sequence files** work very well in practice to overcome the 'small file problem', in which we use the filename as the key and the file contents as the value. By writing a program for files (100 KB), we can put them into a single Sequence file and then we can process them in a streaming fashion operating on the Sequence file. MapReduce can break the Sequence file into chunks and operate on each chunk independently because the Sequence file is splittable.
- Storing files in **HBase is a very common design pattern** to overcome small file problem with HDFS. We are not actually storing millions of small files into HBase, rather adding the binary content of the file to a cell.

2. Slow Processing Speed

In Hadoop, with a parallel and distributed algorithm, the MapReduce process large data sets. There are tasks that we need to perform: Map and Reduce and, MapReduce requires a lot of time to perform these tasks thereby increasing latency. Data is distributed and processed over the cluster in MapReduce which increases the time and reduces processing speed.

Solution-

As a Solution to this Limitation of Hadoop spark has overcome this issue, by in-memory processing of data. In-memory processing is faster as no time is spent in moving the data/processes in and out of the disk. Spark is 100 times faster than MapReduce as it processes everything in memory. We also Flink, as it processes faster than spark because of its streaming architecture and Flink gets instructions to process only the parts of the data that have actually changed, thus significantly increases the performance of the job.

3. Support for Batch Processing only

Hadoop supports batch processing only, it does not process streamed data, and hence overall performance is slower. The MapReduce framework of Hadoop does not leverage the memory of the **Hadoop cluster** to the maximum.

Solution-
To solve these limitations of Hadoop spark is used that improves the performance, but **Spark stream processing** is not as efficient as Flink as it uses micro-batch processing. Flink improves the overall performance as it provides single run-time for the streaming as well as batch processing. Flink uses native closed loop iteration operators which make **machine learning** and graph processing faster.

4. No Real-time Data Processing

Apache Hadoop is for batch processing, which means it takes a huge amount of data in input, process it and produces the result. Although batch processing is very efficient for processing a high volume of data, depending on the size of the data that processes and the computational power of the system, an output can delay significantly. Hadoop is not suitable for Real-time data processing.

Solution-

- **Apache Spark** supports stream processing. Stream processing involves continuous input and output of data. It emphasizes on the velocity of the data, and data processes within a small period of time. Learn more about **Spark Streaming APIs.**
- **Apache Flink** provides single run-time for the streaming as well as batch processing, so one common run-time is utilized for data streaming applications and batch processing applications. Flink is a stream processing system that is able to process row after row in real time.

5. No Delta Iteration

Hadoop is not so efficient for iterative processing, as Hadoop does not support cyclic data flow(i.e. a chain of stages in which each output of the previous stage is the input to the next stage).

Solution-
We can use Apache Spark to overcome this type of Limitations of Hadoop, as it accesses data from RAM instead of disk, which dramatically improves the performance of iterative algorithms that access the same dataset repeatedly. Spark iterates its data in batches. For iterative processing in Spark, we schedule and execute each iteration separately.

6. Latency

In Hadoop, MapReduce framework is comparatively slower, since it is for supporting different format, structure and huge volume of data. In **MapReduce**, Map takes a set of data and converts it into another set of data, where individual elements are broken down into **key-value pairs** and Reduce takes the output from the map as input and process further and MapReduce requires a lot of time to perform these tasks thereby increasing latency.

Solution-
Spark is used to reduce this limitation of Hadoop, Apache Spark is yet another batch system but it is relatively faster since it caches much of the input data on memory by **RDD(Resilient Distributed Dataset)** and keeps intermediate data in memory itself. Flink's data streaming achieves low latency and high throughput.FR

7. Not Easy to Use

In Hadoop, MapReduce developers need to hand code for each and every operation which makes it very difficult to work. MapReduce has no interactive mode, but adding one such as **hive** and pig makes working with MapReduce a little easier for adopters.

Solution-

To solve this Drawback of Hadoop, we can use the spark. Spark has interactive mode so that developers and users alike can have intermediate feedback for queries and other activities. Spark is easy to program as it has tons of high-level operators. We can easily use Flink as it also has high-level operators. This way spark can solve many limitations of Hadoop.

8. Security

Hadoop is challenging in managing the complex application. If the user doesn't know how to enable a platform who is managing the platform, your data can be a huge risk. At storage and network levels, Hadoop is missing encryption, which is a major point of concern. Hadoop supports **Kerberos authentication**, which is hard to manage.

HDFS **supports access control lists** (ACLs) and a traditional file permissions model. However, third-party vendors have enabled an organization to leverage**Active Directory Kerberos** and **LDAP** for authentication.

Solution-

Spark provides a security bonus to overcome these limitations of Hadoop. If we run the spark in HDFS, it can use HDFS ACLs and file-level permissions. Additionally, Spark can run on **YARN** giving it the capability of using Kerberos authentication.

9. No Abstraction

Hadoop does not have any type of abstraction so MapReduce developers need to hand code for each and every operation which makes it very difficult to work.

Solution-

To overcome these drawbacks of Hadoop, Spark is used in which we have **RDD abstraction for the batch.** Flink has Dataset abstraction.

10. Vulnerable by Nature

Hadoop is entirely written in **Java**, a language most widely used, hence java been most heavily exploited by cyber criminals and as a result, implicated in numerous security breaches.

11. No Caching

Hadoop is not efficient for caching. In Hadoop, MapReduce cannot cache the intermediate data in memory for a further requirement which diminishes the performance of Hadoop.

Solution-

Spark and Flink can overcome this limitation of Hadoop, as Spark and Flink cache data in memory for further iterations which enhance the overall performance.

12. Lengthy Line of Code

Hadoop has a 1,20,000 line of code, the number of lines produces the number of bugs and it will take more time to execute the program.

Solution-
Although, Spark and Flink are written in scala and java but the implementation is in Scala, so the number of lines of code is lesser than Hadoop. So it will also take less time to execute the program and solve the lengthy line of code limitations of Hadoop.

13. Uncertainty

Hadoop only ensures that the data job is complete, but it’s unable to guarantee when the job will be complete.

13. List and explain Hadoop Commands With Examples

1. version

Command Name: **version**
Command Usage: version
Example:

```
hadoop version
```

Description: Shows the version of hadoop installed.

2. mkdir

Command Name: **mkdir**
Command Usage: mkdir <path>
Example:

```
hdfs dfs -mkdir /user/dataflair/dir1
```

Description: This command takes the <path> as an argument and creates the directory.

3. ls

Command Name: **ls**
Command Usage: ls <path>
Example:

```
hdfs dfs -ls /user/dataflair
```

Description: This command displays the contents of the directory specified by <path>. It shows the name, permissions, owner, size and modification date of each entry.
Second Example:

```
hdfs dfs -ls -R /user
```

<insert image ls-R.png>
Description: This command behaves like ls but displays entries in all the sub-directories recursively

4. put

Command Name: **put**
Command Usage: put <localsrc> <dest>
Example:

```
hdfs dfs -put /home/sample.txt /user/dataflair/dir1
```

Description: This command copies the file in the local filesystem to the file in DFS.

5. copyFrom Local

Command Name: **copyFrom Local**
Command Usage: copyFrom Local <localsrc> <dest>
Example:

```
hdfs dfs -copyFromLocal /home/sample /user/dataflair/dir1
```

Description: This command is similar to put command. But the source should refer to local file.

6. get

Command Name:**get**
Command Usage: get <src> <localdest>
Example:

```
hdfs dfs -get /user/dataflair/dir1 /home
```

Description: This **Hadoop shell command** copies the file in HDFS identified by <src> to file in local file system identified by <localdest>
Second Example:

```
hdfs dfs -getmerge /user/dataflair/dir1/sample.txt /user/dataflair/dir2/sample2.txt /home/sample1.txt
```

Description: This **HDFS command** retrieves all files in the source path entered by the user in HDFS. And merges them into one single file created in the local file system identified by local destination.
Third Example:

```
hadoop fs –getfacl /user/dataflair/dir1
```

Fourth Example:

```
hadoop fs –getfacl -R /user/dataflair/dir1
```

Description: This Hadoop command shows the Access Control Lists (ACLs) of files and directories. This command displays default ACL if the directory contains the same.
Options : -R: It recursively displays a list of all the ACLs of all files and directories.
Fifth Example:

```
hadoop fs –getfattr –d /user/dataflair/dir1
```

Description: This HDFS command displays if there is any extended attribute names and values for the specified file or directory.

Options:-R: It lists the attributes for all files and directories recursively. -n name: It shows the named extended attribute value. -d: It shows all the extended attribute values associated with the pathname. -e encoding: Encodes values after extracting them. The valid coded forms that are “text”, “hex”, and “base64”. The values which are encoded as text strings gets enclosed with double quotes (” “). It uses prefix 0x for hexadecimal conversion. And 0s for all the values which gets coded as base64.

7. copyToLocal

Command Name: **copyToLocal**

Command Usage: copyToLocal <src> <localdest>

Example:

```
hdfs dfs -copyToLocal /user/dataflair/dir1 /home
```

Description: It is similar to get command. Only the difference is that in this the destination of copied file should refer to a local file.

8. cat

Command Name: **cat**

Command Usage: cat <file-name>

Example:

```
hdfs dfs -cat /user/dataflair/dir1/sample.txt
```

Description: This Hadoop shell command displays the contents of file on console or stdout.

9. mv

Command Name: **mv**

Command Usage: mv <src> <dest>

Example:

```
hdfs dfs -mv /user/dataflair/dir1/sample.txt /user/dataflair/dir2
```

Description: This Hadoop shell command moves the file from the specified source to destination within HDFS.

10. cp

Command Name: **cp**

Command Usage: cp <src> <dest>

Example:

```
hdfs dfs -cp /user/dataflair/dir2/sample.txt /user/dataflair/dir1
```

Description: This Hadoop shell command copies the file or directory from given source to destination within HDFS.

14. Discuss the Benefits of Hadoop Clusters

Here is a list of benefits provided by Clusters in Hadoop –

- Robustness
- Data disks failures, heartbeats and re-replication
- Cluster Rrbalancing
- Data integrity
- Metadata disk failure
- Snapshot

i. Robustness

The **main objective of Hadoop** is to store data reliably even in the event of failures. Various kind of failure is NameNode failure, DataNode failure, and network partition. DataNode periodically sends a heartbeat signal to NameNode. In network partition, a set of DataNodes gets disconnected with the NameNode. Thus NameNode does not receive any heartbeat from these DataNodes. It marks these DataNodes as dead. Also, Namenode does not forward any I/O request to them. The replication factor of the blocks stored in these DataNodes falls below their specified value. As a result, NameNode initiates replication of these blocks. In this way, NameNode recovers from the failure.

ii. Data Disks Failure, Heartbeats, and Re-replication

NameNode receives a heartbeat from each DataNode. NameNode may fail to receive heartbeat because of certain reasons like network partition. In this case, it marks these nodes as dead. This decreases the replication factor of the data present in the dead nodes. Hence NameNode initiates replication for these blocks thereby making the cluster fault tolerant.

iii. Cluster Rebalancing

The **HDFS architecture** automatically does cluster rebalancing. Suppose the free space in a DataNode falls below a threshold level. Then it automatically moves some data to another DataNode where enough space is available.

iv. Data Integrity

Hadoop **cluster** implements checksum on each block of the file. It does so to see if there is any corruption due to buggy software, faults in storage device etc. If it finds the block corrupted it seeks it from another DataNode that has a replica of the block.

v. Metadata Disk Failure

FSImage and Editlog are the central data structures of HDFS. Corruption of these files can stop the **functioning of HDFS**. For this reason, we can configure NameNode to maintain multiple copies of FSImage and EditLog. Updation of multiple copies of FSImage and EditLog can degrade the performance of Namespace operations. But it is fine as Hadoop deals more with the data-intensive application rather than metadata intensive operation.

vi. Snapshot

Snapshot is nothing but storing a copy of data at a particular instance of time. One of the usages of the snapshot is to rollback a failed HDFS instance to a good point in time. We can take Snapshots of the sub-tree of the file system or entire file system. Some of the uses of snapshots are disaster recovery, data backup, and protection against user error. We can take snapshots of any directory. Only the particular directory should be set as Snapshottable. The administrators can set any directory as snapshottable. We cannot rename or delete a snapshottable directory if there are snapshots in it. After removing all the snapshots from the directory, we can rename or delete it.

15. What are the Design Principles of Hadoop ?

Below are the design principles of Hadoop on which it works:
a) System shall manage and heal itself

- Automatically and transparently route around failure (Fault Tolerant)
- Speculatively execute redundant tasks if certain nodes are detected to be slow

b) Performance shall scale linearly

- Proportional change in capacity with resource change (Scalability)

c) Computation should move to data

- Lower latency, lower bandwidth (Data Locality)

d) Simple core, modular and extensible (Economical)

Question Bank for Big Data Analytics

Module 3

Q.1) Why should organizations invest in business intelligence solutions? What Are Potential Business Intelligence Problems?

A Business Intelligence (BI) solution helps in producing accurate reports by extracting data directly from your data source. With Business Intelligence solutions today eliminate the time-consuming task of consolidating data manually. Since BI tools can produce recent data, it allows managers to monitor businesses in real-time. A BI solution provides real-time reports directly to managers on-demand from any location. This helps to reduce the scope of error by providing managers with accurate data to make better decisions on what is happening now and to forecast for the future. BI solutions also focus on providing data security by using existing established security infrastructures to keep data private.

Potential Business Intelligence Problems

1. **User resistance** — Implementations can be dogged by cultural challenges.
2. **Irrelevant and poor quality data** — To get accurate insights, you must have standard data. Get your data in good working order before extracting and acting on insights.
3. **BI tools** — The core of BI is reporting, not process management. Be careful not to confuse business intelligence with business analytics.
4. **Companies don’t understand their business processes well enough** — Strive to understand *all* the activities that make up a particular business process before starting a BI project.

Q.2) Describe 2 BI tools used in your organization.

a) A spreadsheet tool, such as Microsoft Excel, can act as an easy but effective BI tool by itself. Data can be downloaded and stored in the spreadsheet, then analyzed to produce insights, then presented in the form of graphs and tables. This system offers limited automation using macros and other features. The analytical features include basic statistical and financial functions. Pivot tables help do sophisticated what-if analysis. Add-on modules can be installed to enable moderately sophisticated statistical analysis.

b)A dashboarding system, such as IBM Cognos or Tableau, can offer a sophisticated set of tools for gathering, analyzing, and presenting data. At the user end, modular dashboards can be designed and redesigned easily with a graphical user interface. The back-end data analytical capabilities include many statistical functions. The dashboards are linked to data warehouses at the back end to ensure that the tables and graphs and other elements of the dashboard are updated in real time

Q.3) What is the purpose of a data warehouse?

- A data warehouse (DW) is an organized collection of integrated, subject oriented databases designed to support decision support functions.
- The purpose of DW:
- It is organized at the right level of granularity to provide clean enterprise-wide data in a standardized format for reports, queries, and analysis.
- DW is physically and functionally separate from an operational and transactional database.
- Creating a DW for analysis and queries represents significant investment in time and effort. It has to be constantly kept up-to-date for it to be useful.
- DW offers many business and technical benefits. DW supports business reporting and data mining activities. It can facilitate distributed access to up-to-date business knowledge for departments and functions, thus improving business efficiency and customer service.
- DW can present a competitive advantage by facilitating decision making and helping reform business processes.
- DW enables a consolidated view of corporate data, all cleaned and organized. Thus, the entire organization can see an integrated view of itself.
- DW thus provides better and timely information. It simplifies data access and allows end users to perform extensive analysis. It enhances overall IT performance by not burdening the operational databases used by Enterprise Resource Planning (ERP) and other systems.

Q.4) What are the key elements of a data warehouse? Describe each one.

1. *Subject oriented*: To be effective, a DW should be designed around a subject domain, i.e. to help solve a certain category of problems.
2. *Integrated*: The DW should include data from many functions that can shed light on a particular subject area. Thus the organization can benefit from a comprehensive view of the subject area.
3. *Time-variant (time series):* The data in DW should grow at daily or other chosen intervals. That allows latest comparisons over time.
4. *Nonvolatile*: DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, DW is consistently available for analysis, across the organization and over time.
5. *Summarized*: DW contains rolled-up data at the right level for queries and analysis. The process of rolling up the data helps create consistent granularity for effective comparisons. It also helps reduces the number of variables or dimensions of the data to make them more meaningful for the decision makers.
6. *Not normalized*: DW often uses a star schema, which is a rectangular central table, surrounded by some look-up tables. The single table view significantly enhances speed of queries.
7. *Metadata*: Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in the DW should be sufficiently well-defined.
8. *Near Real-time and/or right-time (active*): DWs should be updated in near real-time in many high transaction volume industries, such as airlines. The cost of implementing and updating DW in real time could be discouraging though. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.

Q.5) What are the sources and types of data for a data warehouse?

Data Warehouses are created from structured data sources. Unstructured data such as text data would need to be structured before inserted into the DW.

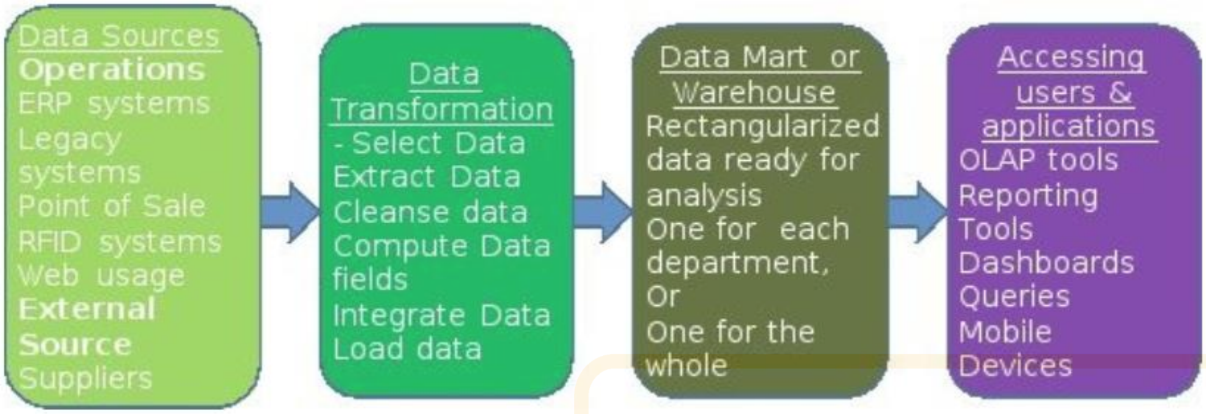
1. *Operations data:* This includes data from all business applications, including from ERPs systems that form the backbone of an organization's IT systems. The data to be extracted will depend upon the subject matter of the data warehouse. For example, for a sales/marketing data mart, only the data about customers, orders, customer service, and so on would be extracted.

2. *Specialized applications*: This includes applications such as Point of Sale (POS) terminals, and e-commerce applications, that also provide customer-facing data. Supplier data could come from Supply Chain Management systems. Planning and budget data should also be added as needed for making comparisons against targets.

3. *External syndicated data*: This includes publicly available data such as weather or economic activity data. It could also be added to the DW, as needed, to provide good contextual information to decision makers.

Q.6) Explain the architecture of Data Warehouse.

DW has four key elements (Figure 1). The first element is the data sources that provide the raw data. The second element is the process of transforming that data to meet the decision needs. The third element is the methods of regularly and accurately loading of that data into EDW or data marts. The fourth element is the data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users.



Q.7) What is data mining? What are supervised and unsupervised learning techniques?

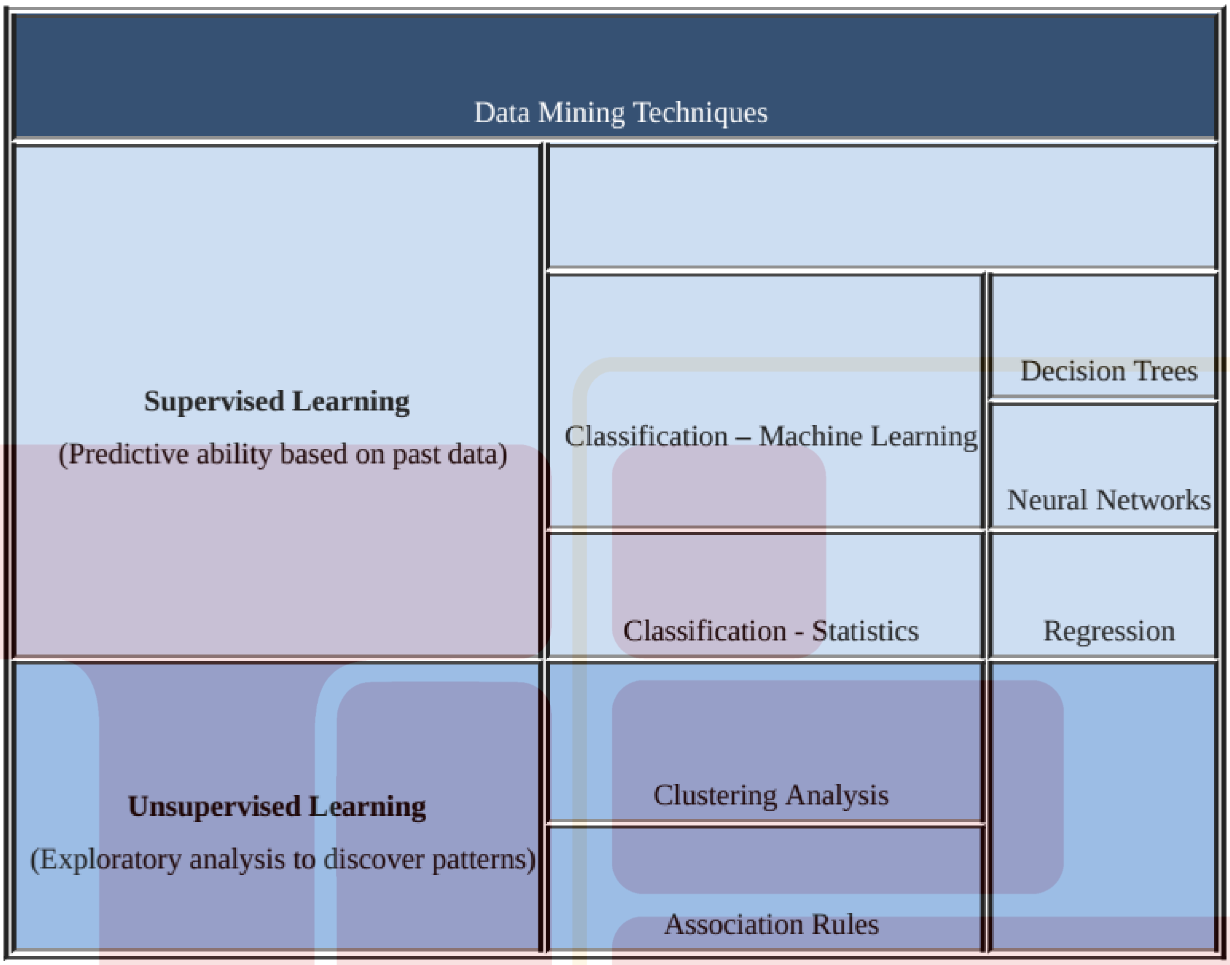
Data mining is the art and science of discovering knowledge, insights, and patterns in data. It is the act of extracting useful patterns from an organized collection of data. Patterns must be valid, novel, potentially useful, and understandable. The implicit assumption is that data about the past can reveal patterns of activity that can be projected into the future.

Data mining is a multidisciplinary field that borrows techniques from a variety of fields. It utilizes the knowledge of data quality and data organizing from the databases area. It draws modeling and analytical techniques from statistics and computer science (artificial intelligence) areas. It also draws the knowledge of decision-making from the field of business management.

The field of data mining emerged in the context of pattern recognition in defense, such as identifying a friend-or-foe on a battlefield. Like many other defense-inspired technologies, it has evolved to help gain a competitive advantage in business.

Supervised and Unsupervised Leaning

Data may be mined to help make more efficient decisions in the future. Or it may be used to explore the data to find interesting associative patterns. The right technique depends upon the kind of problem being solved.



Data Mining Techniques		
Supervised Learning (Predictive ability based on past data)	Classification – Machine Learning	Decision Trees
		Neural Networks
	Classification - Statistics	Regression
Unsupervised Learning (Exploratory analysis to discover patterns)	Clustering Analysis	
	Association Rules	

Q.8) Describe the key steps in the data mining process. Why is it important to follow these processes?

Key steps in the data mining process

1. Business Understanding: The first and most important step in data mining is asking the right business questions. A related important step is to be creative and open in proposing imaginative hypotheses for the solution.

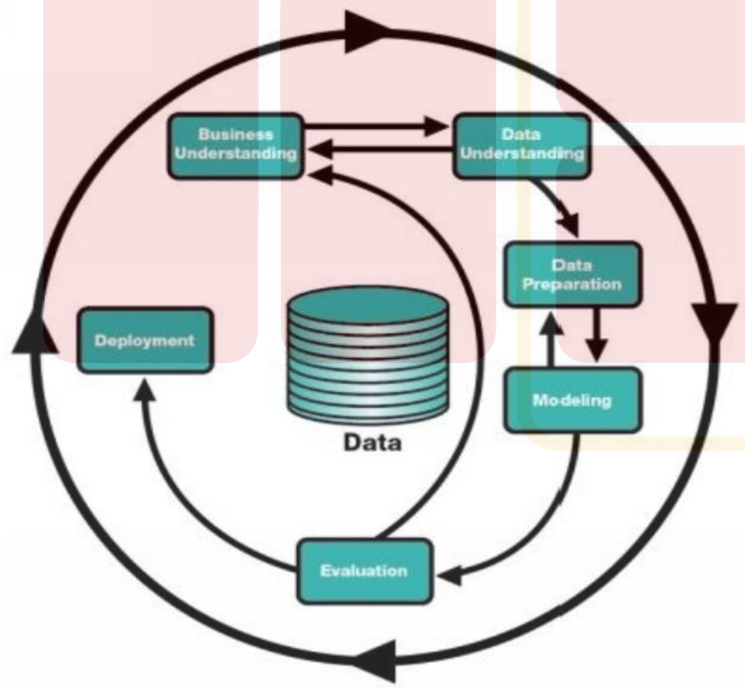
2. *Data Understanding*: A related important step is to understand the data available for mining. One needs to be imaginative in scouring for many elements of data through many sources in helping address the hypotheses to solve a problem. Without relevant data, the hypotheses cannot be tested.

3. *Data Preparation*: The data should be relevant, clean and of high quality. It's important to assemble a team that has a mix of technical and business skills, who understand the domain and the data. It helps to improve predictive accuracy.

4. *Modeling*: This is the actual task of running many algorithms using the available data to discover if the hypotheses are supported. Patience is required in continuously engaging with the data until the data yields some good insights.

5. *Model Evaluation*: One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques, and conducting many what-if scenarios, to build confidence in the solution

6. *Dissemination and rollout:* It is important that the data mining solution is presented to the key stakeholders and is deployed in the organization. The model should be eventually embedded in the organization's business processes.



Q.9) What is a confusion matrix?

There are two primary kinds of data mining processes: supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances.

Classification is the main category of supervised learning activity. There are many techniques for classification, decision trees being the most popular one. Each of these techniques can be implemented with many algorithms. A common metric for all of classification techniques is predictive accuracy.

Predictive Accuracy = (Correct Predictions) / Total Predictions

Suppose a data mining project has been initiated to develop a predictive model for cancer patients using a decision tree. Using a relevant set of variables and data instances, a decision tree model has been created. The model is then used to predict other data instances. When a true positive data point is positive, that is a correct prediction, called a true positive (TP). Similarly, when a true negative data point is classified as negative, that is a true negative (TN). On the other hand, when a true-positive data point is classified by the model as negative, that is an incorrect prediction, called a false negative (FN). Similarly, when a true-negative data point is classified as positive, that is classified as a false positive (FP). This is represented using the confusion matrix.

ConfusionMatrix		**True Class**	
		Positive	*Negative*
Predicted Class	Positive	**True Positive (TP)**	False Positive (FP)
Predicted class	Negative	False Negative (FN)	**True Negative (TN)**

Thus the predictive accuracy can be specified by the following formula.

Predictive Accuracy = (TP +TN) / (TP + TN + FP + FN).

Q.10) What are some of the most popular data mining techniques?

1. *Decision trees (DT)*

- Decision trees are easy to understand and easy to use, by analysts as well as executives. They also show a high predictive accuracy.
- DT selects the most relevant variables automatically out of all the available variables for decision making.
- DT are tolerant of data quality issues and do not require much data preparation from the users.
- Even non-linear relationships can be handled well by decision trees.

2. ***Regression*** is a most popular statistical data mining technique.

- The goal of regression is to derive a smooth well-defined curve to best the data.
- Regression analysis techniques, for example, can be used to model and predict the energy consumption as a function of daily temperature.
- Simply plotting the data may show a non-linear curve. Applying a non-linear regression equation will fit the data very well with high accuracy. Once such a regression model has been developed, the energy consumption on any future day can be predicted using this equation.
- The accuracy of the regression model depends entirely upon the dataset used and not at all on the algorithm or tools used.

*3. **Artificial Neural Networks*** **(ANN)**

- It is a sophisticated data mining technique from the Artificial Intelligence stream in Computer Science. It mimics the behavior of human neural structure.
- The neural network can be trained by making a decision over and over again with many data points. It will continue to learn by adjusting its internal computation and communication parameters based on feedback received on its previous decisions. The intermediate values passed within the layers of neurons may not make any intuitive sense to an observer.
- ANNs are popular because they are eventually able to reach a high predictive accuracy.
- ANNs are also relatively simple to implement and do not have any issues with data quality. However, ANNs require a lot of data to train it to develop good predictive ability.

4. Cluster Analysis

- It is an exploratory learning technique that helps in identifying a set of similar groups in the data.
- It is a technique used for automatic identification of natural groupings of things. Data instances that are similar to (or near) each other are categorized into one cluster, while data instances that are very different (or far away) from each other are categorized into separate clusters.
- There can be any number of clusters that could be produced by the data. The K-means technique is a popular technique and allows the user guidance in selecting the right number (K) of clusters from the data.

5. Association rules

- It is a popular data mining method in business, especially where selling is involved.
- Also known as market basket analysis, it helps in answering questions about cross-selling opportunities. This is the heart of the personalization engine used by ecommerce sites like Amazon.com and streaming movie sites like Netflix.com.
- The technique helps find interesting relationships (affinities) between variables (items or events). These are represented as rules of the form *X* ® *Y*, where *X* and *Y* are sets of data items.
- It is a form of unsupervised learning, it has no dependent variable; and there are no right or wrong answers. There are just stronger and weaker affinities.

Q.11) What are the major mistakes to be avoided when doing data mining?

Data mining is an exercise in extracting non-trivial useful patterns in the data. It requires a lot of preparation and patience to pursue the many leads that data may provide. Much domain knowledge, tools and skill is required to find such patterns. Here are some of the more common mistakes in doing data mining, and should be avoided.

Mistake #1: *Selecting the wrong problem for data mining*: Without the right goals or having no goals, data mining leads to a waste of time.

Mistake #2: Buried under mountains of data without clear metadata: It is more important to be engaged with the data, than to have lots of data. The relevant data required may be much less than initially thought.

Mistake #3: Disorganized data mining: Without clear goals, much time is wasted. Doing the same tests using the same mining algorithms repeatedly and blindly, without thinking about the next stage, without a plan, would lead to wasted time and energy.

Mistake #4: Insufficient business knowledge: Without a deep understanding of the business domain, the results would be gibberish and meaningless.

Mistake #5: Incompatibility of data mining tools and datasets. All the tools from data gathering, preparation, mining, and visualization, should work together. Use tools that can work with data from multiple sources in multiple industry standard formats.

Mistake #6: Looking only at aggregated results and not at individual records/predictions. It is possible that the right results at the aggregate level provide absurd conclusions at an individual record level. Diving into the data at the right angle can yield insights at many levels of data.

Mistake #7: Not measuring your results differently from the way your sponsor measures them. If the data mining team loses its sense of business objectives and beginning to mine data for its own sake, it will lose respect and executive support very quickly.

Q.12) What is data visualization?

Data Visualization is the art and science of making data easy to understand and consume, for the end user. Ideal visualization shows the right amount of data, in the right order, in the right visual form, to convey the high priority information. The right visualization requires an understanding of the consumer’s needs, nature of the data, and the many tools and techniques available to present data. The right visualization arises from a complete understanding of the totality of the situation. One should use visuals to tell a true, complete and fast-paced story.

Data visualization is the last step in the data life cycle. This is where the data is processed for presentation in an easy-to-consume manner to the right audience for the right purpose. The data should be converted into a language and format that is best preferred and understood by the consumer of data. The presentation should aim to highlight the insights from the data in an actionable manner. If the data is presented in too much detail, then the consumer of that data might lose interest and the insight.

Q.13) What are the data visualization techniques? When would you use tables or graphs?

1. Line graph. This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.
2. Scatter plot: This is another very basic and useful graphic form. It helps several the relationship between two variables. In the above case let, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points.
3. Bar graph: A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate.
4. Stacked Bar graphs: These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.
5. Histograms: These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.

6. Pie charts: These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.

7. Box charts: These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.

8. Bubble Graph: This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is a bubble (or a circle) ... the size of the circle and the color fill in the circle could represent two additional dimensions.

9. Dials: These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range. These ranges could be colored red, yellow and green to give an instant view of the data.

10. Geographical Data maps are particularly useful maps to denote statistics.

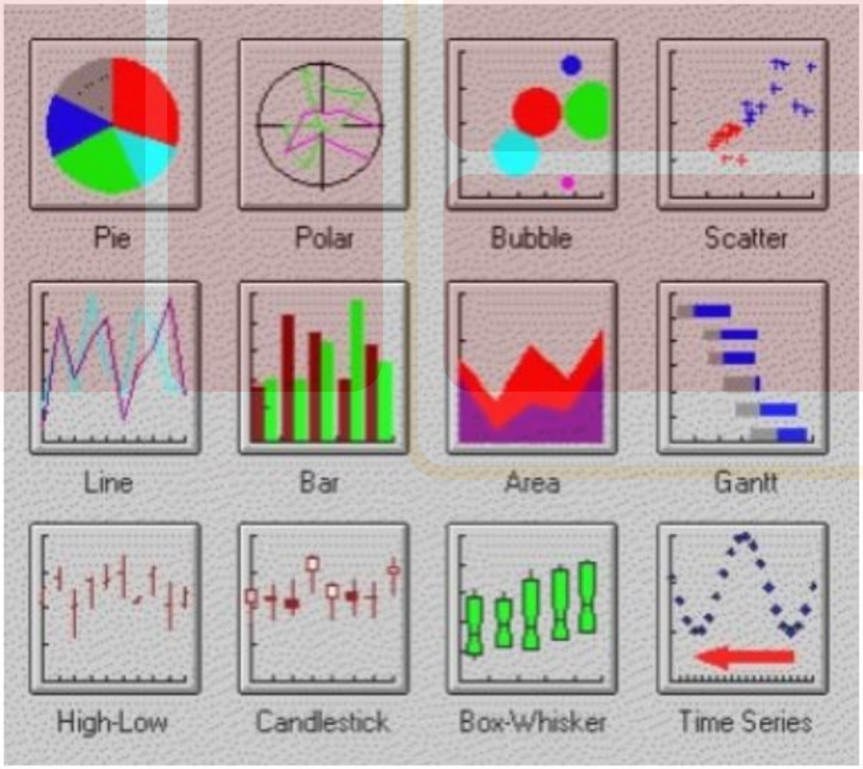


Fig.1 Different types of graphs

When to use tables and graphs for data visualization

Data can be presented in the form of rectangular *tables*, or it can be presented in colorful graphs of various types. "Small, non-comparative, highly-labeled data sets usually belong in tables" – (Ed Tufte, 2001, p 33). However, as the amount of data grows, graphs are preferable. Graphics help give shape to data.

Q.14) Describe some key steps in data visualization.

1. Show, and even reveal, the data: The data should tell a story, especially a story hidden in large masses of data. However, reveal the data in context, so the story is correctly told.

2. *Induce the viewer to think of the substance of the data*: The format of the graph should be so natural to the data, that it hides itself and lets data shine.

3. *Avoid distorting what the data have to say*: Statistics can be used to lie. In the name of simplifying, some crucial context could be removed leading to distorted communication.

4. *Make large data sets coherent*: By giving shape to data, visualizations can help bring the data together to tell a comprehensive story.

5. *Encourage the eyes to compare different pieces of data*: Organize the chart in ways the eyes would naturally move to derive insights from the graph.

6. *Reveal the data at several levels of detail*: Graphs leads to insights, which raise further curiosity, and thus presentations should help get to the root cause.

7. *Serve a reasonably clear purpose* – informing or decision-making.

8. *Closely integrate with the statistical and verbal descriptions of the dataset*: There should be no separation of charts and text in presentation. Each mode should tell a complete story. Intersperse text with the map/graphic to highlight the main insights.

Q.15) What are some key requirements for good visualization.

Key requirements for good visualization are :

1. *Fetch appropriate and correct data for analysis*. This requires some understanding of the domain of the client and what is important for the client. E.g. in a business setting, one may need to understand the many measure of profitability and productivity.

2. *Sort the data in the most appropriate manner*. It could be sorted by numerical variables, or alphabetically by name.

3. *Choose appropriate method to present the data*. The data could be presented as a table, or it could be presented as any of the graph types.

4. *The data set could be pruned* to include only the more significant elements. More data is not necessarily better, unless it makes the most significant impact on the situation.

5. *The visualization could show additional dimension for reference* such as the expectations or targets with which to compare the results.

6. *The numerical data may need to be binned into a few categories*. E.g. the orders per person were plotted as actual values, while the order sizes were binned into 4 categorical choices.

7. *High-level visualization could be backed by more detailed analysis.* For the most significant results, a drill-down may be required.

8. *There may be need to present additional textual information* to tell the whole story. For example, one may require notes to explain some extraordinary results.

Module 4

1. **What is a decision tree? Why is decision tree the most popular classification technique?**

   Decision trees are a simple way to guide one’s path to a decision. The decision may be a simple binary one, whether to approve a loan or not. Or it may be a complex multi-valued decision, as to what may be the diagnosis for a particular sickness. Decision trees are hierarchically branched structures that help one come to a decision based on asking certain questions in a particular sequence.

   Decision trees are one of the most widely used techniques for classification.

- A good decision tree should be short and ask only a few meaningful questions.
- They are very efficient to use, easy to explain, and their classification accuracy is competitive with other methods.
- Decision trees can generate knowledge from a few test instances that can then be applied to abroad population.
- Decision trees are used mostly to answer relatively simple binary decisions.

2. **What is a splitting variable? Describe three criteria for choosing splitting variable.**

   Splitting criteria

   1. Which variable to use for the first split? How should one determine the most important variable for the first branch, and subsequently, for each sub-tree? There are many measures like least errors, information gain, gini’s coefficient, etc.
   2. What values to use for the split? If the variables have continuous values such as for age or blood pressure, what value-ranges should be used to make bins?

3. How many branches should be allowed for each node? There could be binary trees, with just two branches at each node. Or there could be more branches allowed.

3. **What is pruning? What are pre-pruning and post-pruning? Why choose one over the other?**

Pruning: The tree could be trimmed to make it more balanced and more easily usable. The pruning is often done after the tree is constructed, to balance out the tree and improve usability. The symptoms of an overfitted tree are a tree too deep, with too many branches, some of which may reflect anomalies due to noise or outliers. Thus, the tree should be pruned. There are two approaches to avoid over-fitting.

Pre-pruning means to halt the tree construction early, when certain criteria are met. The downside is that it is difficult to decide what criteria to use for halting the construction, because we do not know what may happen subsequently, if we keep growing the tree.

Post-pruning:

Remove branches or sub-trees from a “fully grown” tree. This method is commonly used. C4.5 algorithm uses a statistical method to estimate the errors at each node for pruning. A validation set may be used for pruning as well.

How to choose one over the other

Pre-pruning that stop growing the tree earlier, before it perfectly classifies the training set.

Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree.

Practically, the second approach of post-pruning overfit trees is more successful because it is not easy to precisely estimate when to stop growing the tree.

The first method is the most common approach. In this approach, the available data are separated into two sets of examples: a training set, which is used to build the decision tree, and a validation set, which is used to evaluate the impact of pruning the tree.

4. **What are gini's coefficient and information gain?**

The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions. Information Gain multiplies the probability of the class times the log (base=2) of that class probability. Information Gain favors smaller partitions with many distinct values.

Algo / Split Criterion	**Description**	**Tree Type**
Gini Split / Gini Index	Favors larger partitions. Very simple to implement.	CART
Information Gain / Entropy	Favors partitions that have small counts but many distinct values.	ID3 / C4.5

Using Gini Split / Gini Index

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- Favors larger partitions.
- Uses squared proportion of classes.
- Perfectly classified, Gini Index would be zero.
- Evenly distributed would be 1 – (1/# Classes).
- You want a variable split that has a low Gini Index.
- The algorithm works as 1 – (P(class1)^2 + P(class2)^2 + … + P(classN)^2)
- The Gini index is used in the classic CART algorithm and is very easy to calculate.

Splitting with Information Gain and Entropy

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

- Favors splits with small counts but many unique values.
- Weights probability of class by log(base=2) of the class probability
- A smaller value of Entropy is better. That makes the difference between the parent node's entropy larger.

- Information Gain is the Entropy of the parent node minus the entropy of the child nodes.
- Entropy is calculated [P(class1)*log(P(class1),2) + P(class2)*log(P(class2),2) + … + P(classN)*log(P(classN),2)]

  The maximum value for Entropy depends on the number of classes.
- Two classes: Max entropy is 1.
- Four Classes: Max entropy is 2.
- Eight Classes: Max entropy is 3.
- 16 classes: Max entropy is 4.

5. **Create a decision tree for the following data set. The objective is to predict the class category. (loan approved or not)**

Age	**Job**	**House**	**Credit**	**LoanApproved**
Young	False	No	Fair	*No*
Young	False	No	Good	*No*
Young	True	No	Good	*Yes*
Young	True	Yes	Fair	*Yes*
Young	False	No	Fair	*No*
Middle	False	No	Fair	*No*
Middle	False	No	Good	*No*
Middle	True	Yes	Good	*Yes*
Middle	False	Yes	Excellent	*Yes*
Middle	False	Yes	Excellent	*Yes*
Old	False	Yes	Excellent	*Yes*
Old	False	Yes	Good	*Yes*
Old	True	No	Good	*Yes*
Old	True	No	Excellent	*Yes*
Old	False	No	Fair	*No*

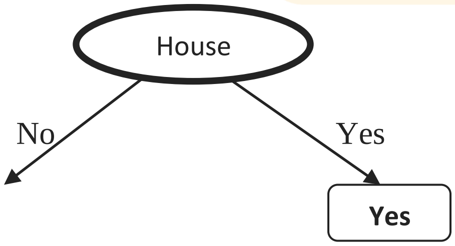
Then solve the following problem using the model.

Age	Job	House	Credit	LoanApproved
Young	False	No	Good	??

Solution :

Attributes	Rules	Error	Total Error
Age	Young →No	2/5	5/15
	Middle→Yes	2/5	
	Old→Yes	1/5	
Job	False→No	4/10	4/15
	True→Yes	0/5	
House	No→No	3/9	3/15
	Yes→Yes	0/6	
Credit	Fair→	1/5	3/15
	Good→	2/6	
	Excellent→	0/4	

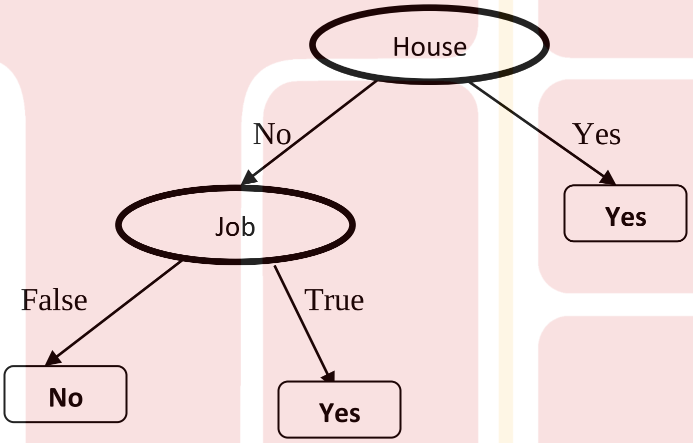
- To select the root node, we find the attribute which is having least number of error. But there is a tie between two attributes, House and Credit.
- Select attribute House as a root node as it has two branches as compared with Credit attribute.



- Now grow the tree for House=No

Attributes	**Rules**	**Error**	**Total Error**
Age	Young →No	1/4	2/9
	Middle→No	0/2	
	Old→Yes	1/3	
Job	False→No	0/6	0/9
	True→Yes	0/3	
Credit	Fair→No	0/4	2/9
	Good→Yes	2/4	
	Excellent→Yes	0/1	

- Job attribute is having least error than others



- For the following test data **Answer is No**

Age	**Job**	**House**	**Credit**	**LoanApproved**
Young	False	No	Good	***No***

6. **What is a regression model?**

Regression is a well-known statistical technique to model the predictive relationship between several independent variables (DVs) and one dependent variable. The objective is to find the best-fitting curve for a dependent variable in a multidimensional space, with each independent variable being a dimension. The curve could be a straight line, or it could be a nonlinear curve.

The quality of fit of the curve to the data can be measured by a coefficient of correlation (r), which is the square root of the amount of variance explained by the curve.

The key steps for regression are simple:

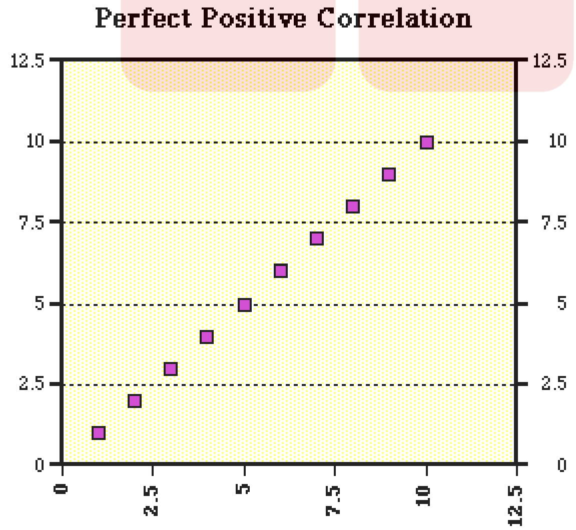
1. List all the variables available for making the model.
2. Establish a Dependent Variable (DV) of interest.
3. Examine visual (if possible) relationships between variables of interest.
4. Find a way to predict DV using the other variables.

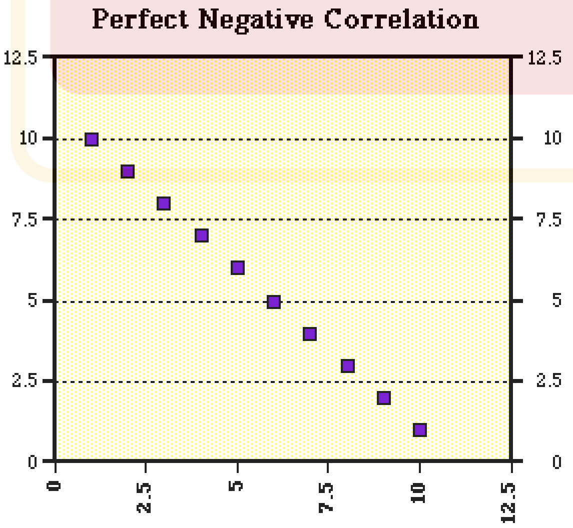
7. What is a scatter plot? How does it help?

Scatter plots are similar to line graphs in that they use horizontal and vertical axes to plot data points. However, they have a very specific purpose. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their **correlation** .

Scatter plots usually consist of a large body of data. The closer the data points come when plotted to making a straight line, the higher the correlation between the two variables, or the stronger the relationship.

If the data points make a straight line going from the origin out to high x- and y-values, then the variables are said to have a **positive correlation** . If the line goes from a high-value on the y-axis down to a high-value on the x-axis, the variables have a **negative correlation** .





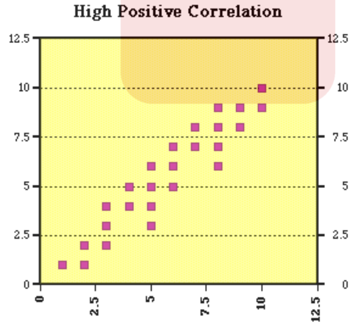
A perfect positive correlation is given the value of 1. A perfect negative correlation is given the value of -1. If there is absolutely no correlation present the value given is 0. The closer the number is to 1 or -1, the stronger the correlation, or the stronger the relationship between the variables. The closer the number is to 0, the weaker the correlation. So something that seems to kind of correlate in a positive direction might have a value of 0.67, whereas something with an extremely weak negative correlation might have the value -.21.

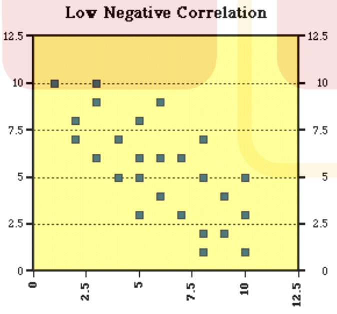
An example of a situation where you might find a perfect positive correlation, as we have in the graph on the left above, would be when you compare the total amount of money spent on tickets at the movie theater with the number of people who go. This means that every time that "x" number of people go, "y" amount of money is spent on tickets without variation.

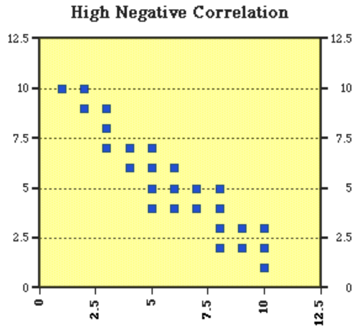
An example of a situation where you might find a perfect negative correlation, as in the graph on the right above, would be if you were comparing the amount of time it takes to reach a destination with the distance of a car (traveling at constant speed) from that destination.

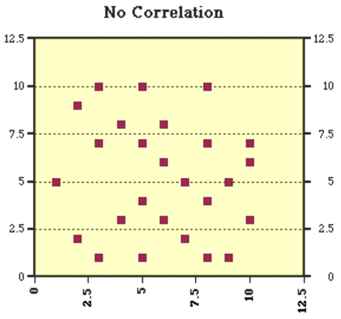
On the other hand, a situation where you might find a strong but not perfect positive correlation would be if you examined the number of hours students spent studying for an exam versus the grade received. This won't be a perfect correlation because two people could spend the same amount of time studying and get different grades. But in general the rule will hold true that as the amount of time studying increases so does the grade received.

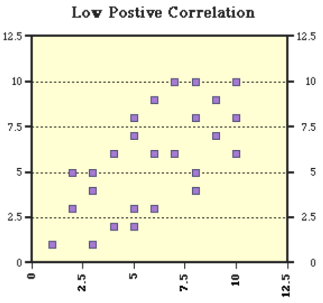
Let's take a look at some examples. The graphs that were shown above each had a perfect correlation, so their values were 1 and -1. The graphs below obviously do not have perfect correlations.











8. **Consider the following dataset.**

Student	Test_Marks	Grade
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

a) What linear regression equation best predicts statistics performance, based on math aptitude scores?

b) If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

c) How well does the regression equation fit the data?

Solution:

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two columns show deviations scores - the difference between the student's score and the average score on each test. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	(x_i-x)	(y_i-y)
1	95	85	17	8
2	85	95	7	18
3	80	70	2	-7
4	70	65	-8	-12
5	60	70	-18	-7
Sum	390	385		
Mean	78	77		

And for each student, we also need to compute the squares of the deviation scores (the last two columns in the table below).

Student	x_i	y_i	$(x_i-x)^2$	$(y_i-y)^2$

1	95	85	289	64
2	85	95	49	324
3	80	70	4	49
4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630
Mean	78	77		

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below. Notice that all of our inputs for the regression analysis come from the above three tables.

And finally, for each student, we need to compute the product of the deviation scores.

Student	**x_i**	**y_i**	**$(x_i-x)(y_i-y)$**
1	95	85	136
2	85	95	126
3	80	70	-14
4	70	65	96
5	60	70	126
Sum	390	385	470
Mean	78	77	

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below. Notice that all of our inputs for the regression analysis come from the above three tables.

First, we solve for the regression coefficient (b_1):

$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$

$b_1 = 470/730$

$b_1 = 0.644$

Once we know the value of the regression coefficient (b_1), we can solve for the regression slope (b_0):

$$b_0 = y - b_1 * x$$

$$b_0 = 77 - (0.644)(78)$$

$$b_0 = 26.768$$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

b) In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade ($\hat{y}$) would be:

$$\hat{y} = b_0 + b_1x$$

$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80$$

$$\hat{y} = 26.768 + 51.52 = 78.288$$

c) Whenever you use a regression equation, you should ask how well the equation fits the data. One way to assess fit is to check the coefficient of determination, which can be computed from the following formula.

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

Computations for the sample problem of this lesson are shown below. We begin by computing the standard deviation of x (σ_x):

$$\sigma_x = \text{sqrt} [\Sigma (x_i - x)^2 / N]$$

$$\sigma_x = \text{sqrt}(730/5) = \text{sqrt}(146) = 12.083$$

Next, we find the standard deviation of y, (σ_y):

$$\sigma_y = \text{sqrt} [\Sigma (y_i - y)^2 / N]$$

$$\sigma_y = \text{sqrt}(630/5) = \text{sqrt}(126) = 11.225$$

And finally, we compute the coefficient of determination (R^2):

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

$$R^2 = [(1/5) * 470 / (12.083 * 11.225)]^2$$

$$R^2 = (94 / 135.632)^2 = (0.693)^2 = 0.48$$

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

Using the data below, create a regression model to predict the Test2 from the Test1 score. Then predict the score for one who got a 46 in Test1. (Apply above method)

Test1	Test2
59	56
52	63
44	55
51	50
42	66
42	48
41	58
45	36
27	13
63	50
54	81
44	56
50	64
47	5

9. List Advantages and Disadvantages of Regression Models

Regression Models are very popular because they offer many advantages.

1. Regression models are easy to understand as they are built upon basic statistical principles such as correlation and least square error.

2. Regression models provide simple algebraic equations that are easy to understand and use.

3. The strength (or the goodness of fit) of the regression model is measured in terms of the correlation coefficients, and other related statistical parameters that are well understood.

4. Regression models can match and beat the predictive power of other modeling techniques.

5. Regression models can include all the variables that one wants to include in the model.

6. Regression modeling tools are pervasive. They are found in statistical packages as well as data mining packages. MS Excel spreadsheets can also provide simple regression modeling capabilities.

Regression models can however prove inadequate under many circumstances.

1. Regression models can not cover for poor data quality issues. If the data is not prepared well to remove missing values or is not well-behaved in terms of a normal distribution, the validity of the model suffers.

2. Regression models suffer from collinearity problems (meaning strong linear correlations among some independent variables).

3. Regression models can be unwieldy and unreliable if many variables are included in the model.

4. Regression models do not automatically take care of non-linearity. The user needs to imagine the kind of additional terms that might be needed to be added to the regression model to improve its fit.

5. Regression models work only with numeric data and not with categorical variables. There are ways to deal with categorical variables though by creating multiple new variables with a yes/no value.

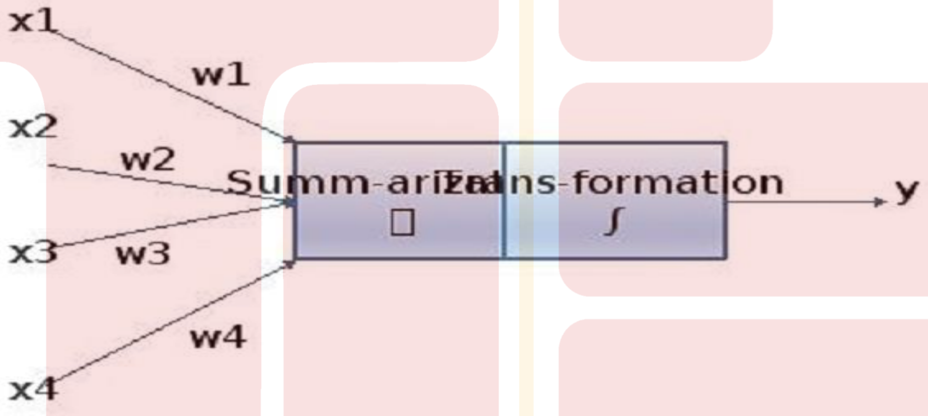
10. What is a neural network? How does it work?

Artificial Neural Networks (ANN) are inspired by the information processing model of the mind/brain. The human brain consists of billions of neurons that link with one another in an intricate pattern. Every neuron receives information from many other neurons, processes it, gets excited or not, and passes its state information to other neurons. Just like the brain is a multipurpose system, so also the ANNs are very versatile systems. They can be used for many kinds of pattern recognition and prediction. They are also used for classification, regression, clustering, association, and optimization activities. They are used in finance, marketing, manufacturing, operations, information systems applications, and so on.

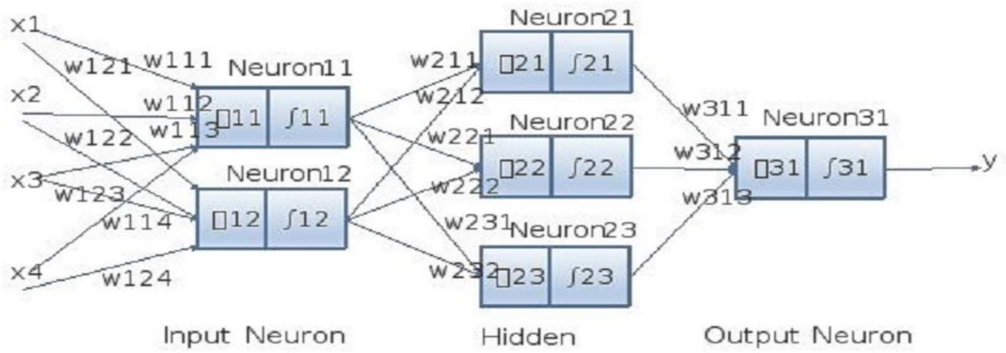
ANNs are composed of a large number of highly interconnected processing elements (neurons) working in a multi-layered structure that receive inputs, process the inputs, and produce an output. An ANN is designed for a specific application, such as pattern recognition or data classification, and trained through a learning process. Just like in biological systems, ANNs adjust the synaptic connections with each learning instance.

Design Principles of an Artificial Neural Network

1. A neuron is the basic processing unit of the network. The neuron (or processing element)receives inputs from its preceding neurons (or PEs),does some nonlinear weighted computation on the basis of those inputs, transforms the result into its output value, and then passes on the output to the next neuron in the network. X's are the inputs, w's are the weights for each input, and y is the output.



2. A Neural network is a multi-layered model. There is at least one input neuron, one output neuron, and at least one processing neuron. An ANN with just this basic structure would be a simple, single-stage computational unit. A simple task may be processed by just that one neuron and the result may be communicated soon. ANNs however, may have multiple layers of processing elements in sequence. There could be many neurons involved in a sequence depending upon the complexity of the predictive action. The layers of PEs could work in sequence, or they could work in parallel.



3. The processing logic of each neuron may assign different weights to the various incoming input streams. The processing logic may also use nonlinear transformation, such as a sigmoid function, from the processed values to the output value. This processing logic and the intermediate weight and processing functions are just what works for the system, in its objective of solving a problem collectively. Thus, neural networks are an opaque and a black-box system.

4. The neural network can be trained by making similar decisions repeatedly with many training cases. It will continue to learn by adjusting its internal computation and communication based on feedback about its previous decisions. Thus, the neural networks become better at planning as they handle more and more decisions. Depending upon the nature of the problem and the availability of good training data, at some point the neural network will learn enough and begin to match the predictive accuracy of a human expert. In many practical situations, the predictions of ANN, trained over a long period of time with many training data, have begun to decisively become more accurate than human experts. At that point ANN can begin to be seriously considered for deployment in real situations in real time.

11. **Compare a neural network with a decision tree.**

Decision tree vs neural network :

- Both finds non-linear solutions, and have interaction between independent variables.
- Decision trees are better when there is large set of categorical values in training data.
- Decision trees are better than NN, when the scenario demands an explanation over the decision.
- NN outperforms decision tree when there is sufficient training data.

12. **Describe three business applications in your industry where cluster analysis will be useful.**

Cluster analysis is used in almost every field where there is a large variety of transactions. It helps provide characterization, definition, and labels for populations. It can help identify natural groupings of customers, products, patients, and so on. It can also help identify outliers in a specific domain and thus decrease the size and complexity of problems. A prominent business application of cluster analysis is in market research. Customers are segmented into clusters based on their characteristics—wants and needs, geography, price sensitivity, and so on. Here are some examples of clustering:

1. *Market Segmentation*: Categorizing customers according to their similarities, for instance by their common wants and needs, and propensity to pay, can help with targeted marketing.

2. *Product portfolio*: People of similar sizes can be grouped together to make small, medium and large sizes for clothing items.

3. *Text Mining*: Clustering can help organize a given collection of text documents according to their content similarities into clusters of related topics.

13. Data about height and weight for a few volunteers is available. Create a set of clusters for the following data, to decide how many sizes of T-shirts should be ordered.

Data Sample

Height	**Weight**
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67

177	76

Step 1: Input

Dataset, Clustering Variables and Maximum Number of Clusters (K in Means Clustering)

In this dataset, only two variables –height and weight – are considered for clustering

Height	**Weight**
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Step 2: Initialize cluster centroid

In this example, value of K is considered as 2. Cluster centroids are initialized with first 2 observations.

Cluster	Initial Centroid	
	Height	**Weight**
K_1	185	72
K_2	170	56

Step 3: Calculate Euclidean Distance

Euclidean is one of the distance measures used on K Means algorithm. Euclidean distance between of a observation and initial cluster centroids 1 and 2 is calculated. Based on euclidean distance each observation is assigned to one of the clusters - based on minimum distance.

Euclidean Distance = $\sqrt{(X_H \ - H_1)^2 +}\,(X_W - W_1)^2$

Where

X_H: Observation value of variable Height

H_1: Centroid value of Cluster 1 for variable Height

X_W: Observation Value of variable Weight

W_1: Centroid value of cluster 1 for variable Weight

First two observations

Height	**Weight**
185	72
170	56

Now initial cluster centroids are :

Cluster	Updated Centroid	
	Height	**Weight**
K_1	185	72
K_2	170	56

Euclidean Distance Calculation from each of the clusters is calculated.

Euclidian Distance from Cluster 1	Euclidian Distance from Cluster 2	Assignment
$(185\text{-}185)^2\text{+}(72\text{-}72)^2$ =0	$(185\text{-}170)^2\text{+}(72\text{-}56)^2$ = 21.93	1
$(170\text{-}185)^2\text{+}(56\text{-}72)^2$ = 21.93	$(170\text{-}170)^2\text{+}(56\text{-}56)^2$ = 0	2

There is no change in Centroids as these two observations were only considered as initial centroids

Step 4: Move on to next observation and calculate Euclidean Distance

Height	Weight
168	60

Euclidean Distance from Cluster 1	Euclidean Distance from Cluster 2	Assignment
$(168-185)^2+(60-72)^2$ =20.808	$(168-185)^2+(60-72)^2$ = 4.472	2

Since distance is minimum from cluster 2, so the observation is assigned to cluster 2. Now revise Cluster Centroid – mean value Height and Weight as Custer Centroids. Addition is only to cluster 2, so centroid of cluster 2 will be updated

Updated cluster centroids

Cluster	Updated Centroid	
	Height	**Weight**
K=1	185	72
K=2	(170+168)/2 = 169	(56+60)/2 = 58

Step 5: Calculate Euclidean Distance for the next observation, assign next observation based on minimum Euclidean distance and update the cluster centroids.

Next Observation.

Height	Weight
179	68

Euclidean Distance Calculation and Assignment

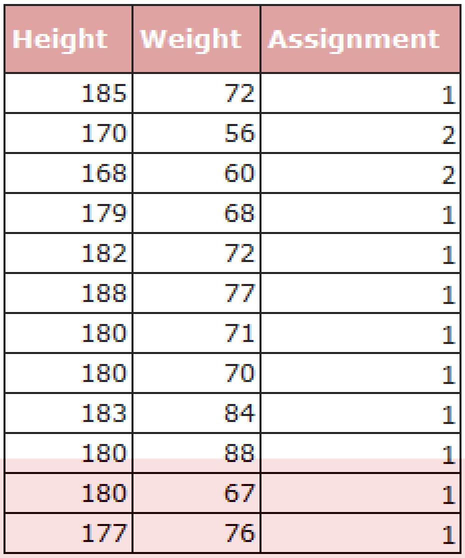
Euclidean Distance from Cluster 1	Euclidean Distance from Cluster 2	Assignment
7.211103	14.14214	1

Update Cluster Centroid

Cluster	Updated Centroid	
	Height	**Weight**
K=1	182	70.6667
K=2	169	58

Continue the steps until all observations are assigned

Final assignments



Height	Weight	Assignment
185	72	1
170	56	2
168	60	2
179	68	1
182	72	1
188	77	1
180	71	1
180	70	1
183	84	1
180	88	1
180	67	1
177	76	1

Cluster Centroids

Cluster	Updated Centroid	
	Height	**Weight**
K=1	182.8	72
K=2	169	58

14. What are association rules? How do they help?

In business environments a pattern or knowledge can be used for many purposes. In sales and marketing, it is used for cross-marketing and cross-selling, catalog design, e-commerce site design, online advertising optimization, product pricing, and sales/promotion configurations. This analysis can suggest not to put one item on sale at a time, and instead to create a bundle of products promoted as a package to sell other non-selling items. In retail environments, it can be used for store design. Strongly associated items can be kept close tougher for customer convenience. Or they could be placed far from each other so that the customer has to walk the aisles and by doing so is potentially exposed to other items. In medicine, this technique can be used for relationships between symptoms and illnesses; diagnosis and patient characteristics/treatments; genes and their functions; etc.

Representing Association Rules

A generic Association Rule is represented between a set X and Y: **XÞ Y [S%,C%]**

X, Y: products and/or services

X: Left-hand-side (LHS)

Y: Right-hand-side (RHS)

S: Support: how often **X** and **Y** go together in the dataset – i.e. P (**X** U **Y**)

C: Confidence: how often **Y** is found, given **X** – i.e. P (**Y** | **X**)

Example: {Hotel booking, Flight booking}Þ {Rental Car} [30%, 60%]

[Note: P (X) is the mathematical representation of the probability or chance of X occurring in the data set.}

Computation example:

Suppose there are 1000 transactions in a data set. There are 300 occurrences of X, and 150 occurrences of (X,Y) in the data set.

Support S for **XÞ Y** will be P(X U Y) = 150/1000 = 15%. Confidence for **XÞ Y** will be P (**Y** | **X**); or P (**X** U **Y**) / P (**X**) = 150/300 = 50%

15. List Advantages and Disadvantages of K-Means algorithm

Advantages of K-Means Algorithm

1. K-Means algorithm is simple, easy to understand and easy to implement.
2. It is also efficient, in that the time taken to cluster k-means, rises linearly with the number of data points.
3. No other clustering algorithm performs better than K-Means, in general.

There are a few disadvantages too:

1. The user needs to specify an initial value of K.
2. The process of finding the clusters may not converge.
3. It is not suitable for discovering clusters shapes that are not hyper ellipsoids (or hyper-spheres).

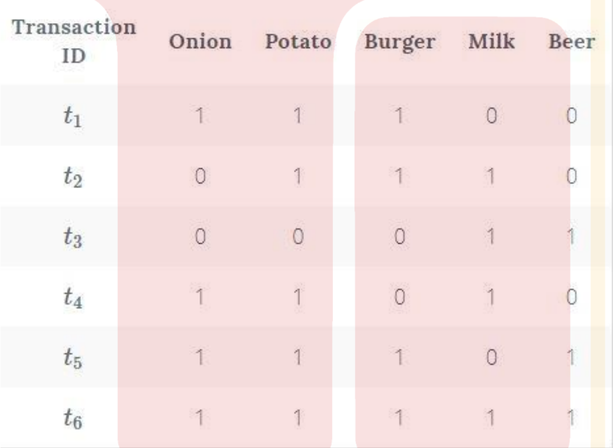
16. How does the Apriori algorithm works?

Apriori Algorithm

This is the most popular algorithm used for association rule mining. The objective is to find subsets that are common to at least a minimum number of the itemsets. A frequent itemset is an itemset whose support is greater than or equal to minimum support threshold. The Apriori property is a downward closure property, which means that any subsets of a frequent itemset are also frequent itemsets. Thus, if (A,B,C,D) is a frequent itemset, then any subset such as (A,B,C) or (B,D) are also frequent itemsets.

It uses a bottom-up approach; and the size of frequent subsets is gradually increased, from one-item subsets to two-item subsets, then three-item subsets, and so on. Groups of candidates at each level are tested against the data for minimum support.

Consider a supermarket scenario where the itemset is I = {Onion, Burger, Potato, Milk, Beer}. The database consists of six transactions where 1 represents the presence of the item and 0 the absence.



Transaction ID	Onion	Potato	Burger	Milk	Beer
t_1	1	1	1	0	0
t_2	0	1	1	1	0
t_3	0	0	0	1	1
t_4	1	1	0	1	0
t_5	1	1	1	0	1
t_6	1	1	1	1	1

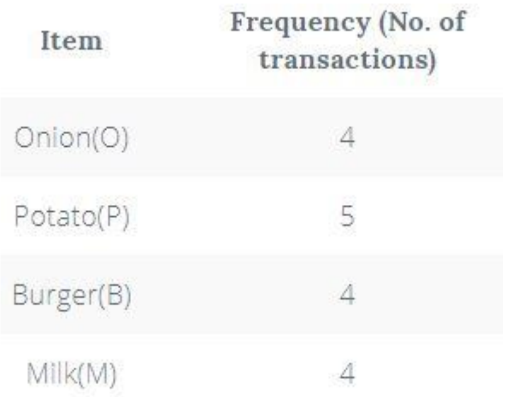
Simple Apriori Algorithm Example

The Apriori Algorithm makes the following assumptions.

- All subsets of a frequent itemset should be frequent.
- In the same way, the subsets of an infrequent itemset should be infrequent.
- Set a threshold support level. In our case, we shall fix it at 50%

Step 1

Create a frequency table of all the items that occur in all the transactions. Now, prune the frequency table to include only those items having a threshold support level over 50%. We arrive at this frequency table.

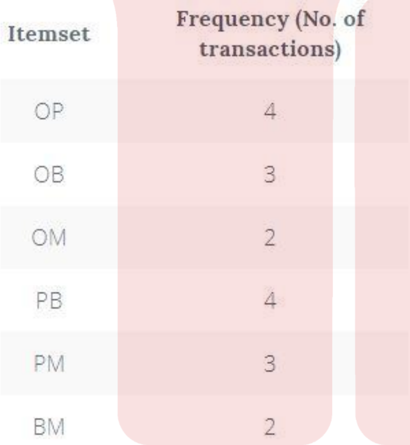


Item	Frequency (No. of transactions)
Onion(O)	4
Potato(P)	5
Burger(B)	4
Milk(M)	4

This table signifies the items frequently bought by the customers.

Step 2

Make pairs of items such as OP, OB, OM, PB, PM, BM. This frequency table is what you arrive at.



Itemset	Frequency (No. of transactions)
OP	4
OB	3
OM	2
PB	4
PM	3
BM	2

Step 3

Apply the same threshold support of 50% and consider the items that exceed 50% (in this case 3 and above).

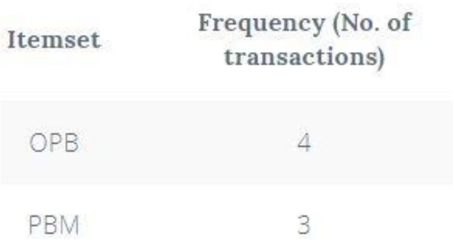
Thus, you are left with OP, OB, PB, and PM

Step 4

Look for a set of three items that the customers buy together. Thus we get this combination.

- OP and OB gives OPB
- PB and PM gives PBM

Step 5



Itemset	Frequency (No. of transactions)
OPB	4
PBM	3

Determine the frequency of these two itemsets. You get this frequency table.

If you apply the threshold assumption, you can deduce that the set of three items frequently purchased by the customers is OPB.

Module 5

1. **Why is text mining useful in the age of social media?**

Text is an important part of the growing data in the world. Social media technologies have enabled users to become producers of text and images and other kinds of information. Text mining can be applied to large-scale social media data for gathering preferences and measuring emotional sentiments. It can also be applied to societal, organizational and individual scales.

2. **What kinds of problems can be addressed using text mining?**

Text mining is a useful tool in the hands of chief knowledge officers to extract knowledge relevant to an organization. Text mining can be used across industry sectors and application areas, including decision support, sentiment analysis, fraud detection, survey analysis, and many more.

1. *Marketing*: The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.

a. Social personas are a clustering technique to develop customer segments of interest. Consumer input from social media sources, such as reviews, blogs, and tweets, contain numerous leading indicators that can be used towards anticipating and predicting consumer behavior.

b. A 'listening platform' is a text mining application, that in real time, gathers social media, blogs, and other textual feedback, and filters out the chatter to extract true consumer sentiment. The insights can lead to more effective product marketing and better customer service.

c. The customer call center conversations and records can be analyzed for patterns of customer complaints. Decision trees can organize this data to create decision choices that could help with product management activities and to become proactive in avoiding those complaints.

2. ***Business operations***: Many aspects of business functioning can be accurately gauged from analyzing text.

a. Social network analysis and text mining can be applied to emails, blogs, social media and other data to measure the emotional states and the mood of employee populations. Sentiment analysis can reveal early signs of employee dissatisfaction which can then be proactively managed.

b. Studying people as emotional investors and using text analysis of the social Internet to measure mass psychology can help in obtaining superior investment returns.

3. *Legal*: In legal applications, lawyers and paralegals can more easily search case histories and laws for relevant documents in a particular case to improve their chances of winning.

a. Text mining is also embedded in e-discovery platforms that help in minimizing risk in the process of sharing legally mandated documents.

b. Case histories, testimonies, and client meeting notes can reveal additional information, such as morbidities in a healthcare situation that can help better predict high-cost injuries and prevent costs.

4. Governance and Politics:

Governments can be overturned based on a tweet originating from a self-immolating fruit-vendor in Tunisia

a. Social network analysis and text mining of large-scale social media data can be used for measuring the emotional states and the mood of constituent populations. Micro-targeting constituents with specific messages gleaned from social media analysis can be a more efficient use of resources when fighting democratic elections.

b. In geopolitical security, internet chatter can be processed for realtime information and to connect the dots on any emerging threats.

c. In academic, research streams could be meta-analyzed for underlying research trends.

3. What kinds of sentiments can be found in the text?

As the amount of social media and other text data grows, there is need for efficient abstraction and categorization of meaningful information from the text.

The first level of analysis is identifying frequent words. This creates a bag of important words. Texts – documents or smaller messages – can then be ranked on how they match to a particular bag-of-words. However, there are challenges with this approach. For example, the words may be spelled a little differently. Or there may be different words with similar meanings.

The next level is at the level of identifying meaningful phrases from words. Thus ‘ice’ and ‘cream’ will be two different key words that often come together. However, there is a more meaningful phrase by combining the two words into ‘ice cream’. There might be similarly meaningful phrases like ‘Apple Pie’.

The next higher level is that of Topics. Multiple phrases could be combined into Topic area. Thus, the two phrases above could be put into a common basket, and this bucket could be called ‘Desserts’.

4. **Create a TDM with not more than 6 key terms. [Hint: Treat each comment as a document]**

Here are a few comments from customer service calls received by Liberty.

1. *I loved the design of the shirt. The size fitted me very well. However, the fabric seemed flimsy. I am calling to see if you can replace the shirt with a different one. Or please refund my money.*

2. *I was running late from work, and I stopped by to pick up some groceries. I did not like the way the manager closed the store while I was still shopping.*

3. *I stopped by to pick up flowers. The checkout line was very long. The manager was polite but did not open new cashiers. I got late for my appointment.*

4. *The manager promised that the product will be there, but when I went there the product was not there. The visit was a waste. The manager should have compensated me for my trouble.*

5. *When there was a problem with my catering order, the store manager promptly contacted me and quickly got the kinks out to send me replacement food immediately. There are very courteous.*

5. What is Web Mining? Explain its characteristics and three types of web mining.

Solution:

Web mining is the art and science of discovering patterns and insights from the World-wide web so as to improve it. The world-wide web is at the heart of the digital revolution. More data is posted on the web every day than was there on the whole web just 20 years ago. Billions of users are using it every day for a variety of purposes. The web is used for electronic commerce, business communication, and many other applications. Web mining analyzes data from the web and helps find insights that could optimize the web content and improve the user experience. Data for web mining is collected via Web crawlers, web logs, and other means.

Here are some characteristics of optimized websites:

1. Appearance: Aesthetic design. Well-formatted content, easy to scan and navigate. Good color contrasts.

2. Content: Well-planned information architecture with useful content. Fresh content. Search engine optimized. Links to other good sites.

3. Functionality: Accessible to all authorized users. Fast loading times. Usable forms. Mobile enabled. This type of content and its structure is of interest to ensure the web is easy to use. The analysis of web usage provides feedback on the web content, and also the consumer's browsing habits. This data can be of immense use for commercial advertising, and even for social engineering. The web could be analyzed for its structure as well as content. The usage pattern of web pages could also be analyzed.

Depending upon objectives, web mining can be divided into three different types:

1. Web usage mining

As a user clicks anywhere on a webpage or application, the action is recorded by many entities in many locations. The browser at the client machine will record the click, and the web server providing the content would also make a record of the pages served and the user activity on those pages. The entities between the client and the server, such as the router, proxy server, or ad server, too would record that click

2. Web content mining

A website is designed in the form of pages with a distinct URL (universal resource locator). A large website may contain thousands of pages. These pages and their content is managed using specialized software systems called Content Management Systems. Every page can have text, graphics, audio, video, forms, applications, and more kinds of content including user generated content.

3. Web structure mining

The Web works through a system of hyperlinks using the hypertext protocol (http). Any page can create a hyperlink to any other page, it can be linked to by another page. The intertwined or self-referral nature of web lends itself to some unique network analytical algorithms. The structure of Web pages could also be analyzed to examine the pattern of hyperlinks among pages. There are two basic strategic models for successful websites: Hubs and Authorities.

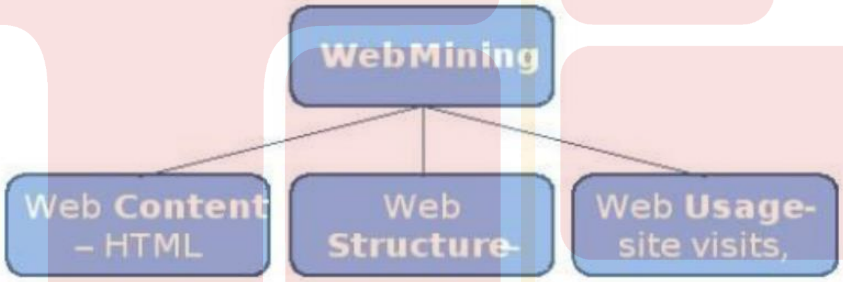


Figure: 1 Web Mining structure

6. What is Naïve Bayes technique? What do Naïve and Bayes stand for?

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Algorithm

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- *P*(*c*|*x*) is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- *P*(*c*) is the prior probability of *class*.
- *P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.

  P(*x*) is the prior probability of *predictor*.

7. List Advantages and disadvantages of Naïve bayes algorithm

- It is easy and fast to predict the class of the test data set. It also performs well in multi-class prediction.
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

  Disadvantages
- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- On the other side naive Bayes is also known as a bad estimator, so the probability outputs are not to be taken too seriously.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

8. What are the most popular applications of NB techniques?

Real-time Prediction: As Naive Bayes is super fast, it can be used for making predictions in real time.

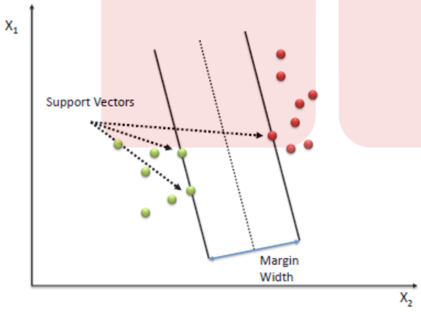
Multi-class Prediction: This algorithm can predict the posterior probability of multiple classes of the target variable.

Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers are mostly used in text classification (due to their better results in multi-class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

Recommendation System: Naive Bayes Classifier along with algorithms like Collaborative Filtering makes a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

9. What is Support Vector Machine? What are support vectors? Explain Kernel method.

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.

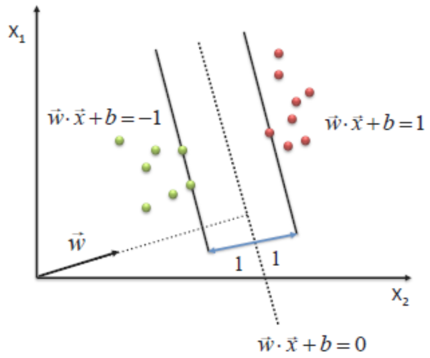


Algorithm

1. Define an optimal hyperplane: maximize margin
2. Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications.

Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.

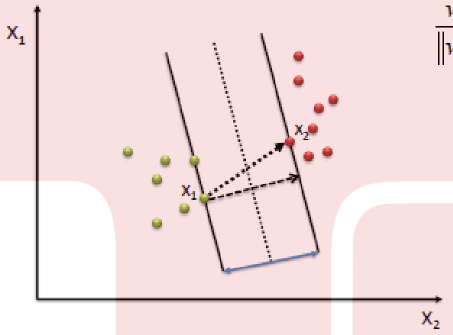
To define an optimal hyperplane we need to maximize the width of the margin (*w*).



X1
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\vec{w}$
1
1
X2
$\vec{w}\cdot\vec{x}+b=0$

$$\boxed{\max \frac{2}{\|w\|}}$$

$$\begin{aligned}&s.t.\\ &(w\cdot x+b)\geq 1, \forall x \text{ of class 1}\\ &(w\cdot x+b)\leq -1, \forall x \text{ of class 2}\end{aligned}$$



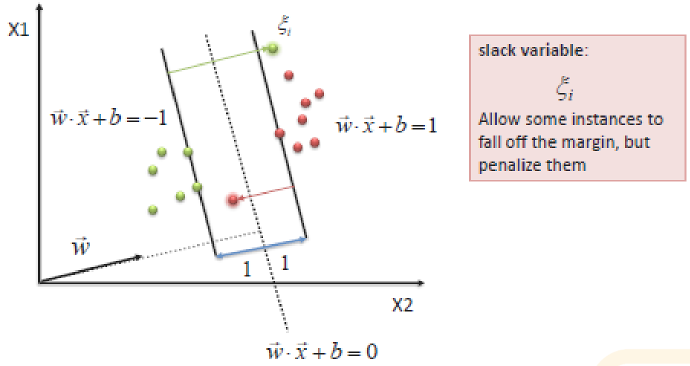
$$\frac{w}{\|w\|}\cdot(x_2-x_1)=\text{width}=\frac{2}{\|w\|}$$

$$\begin{aligned}&w\cdot x_2+b=1\\ &w\cdot x_1+b=-1\\ &w\cdot x_2+b-w\cdot x_1-b=1-(-1)\\ &w\cdot x_2-w\cdot x_1=2\\ &\frac{w}{\|w\|}(x_2-x_1)=\frac{2}{\|w\|}\end{aligned}$$

We find *w* and *b* by solving the following objective function using Quadratic Programming.

$$\min \frac{1}{2}\|w\|^2$$
$$s.t.\ y_i(w\cdot x_i+b)\geq 1,\ \ \forall x_i$$

The beauty of SVM is that if the data is linearly separable, there is a unique global minimum value. An ideal SVM analysis should produce a hyperplane that completely separates the vectors (cases) into two non-overlapping classes. However, perfect separation may not be possible, or it may result in a model with so many cases that the model does not classify correctly. In this situation SVM finds the hyperplane that maximizes the margin and minimizes the misclassifications.

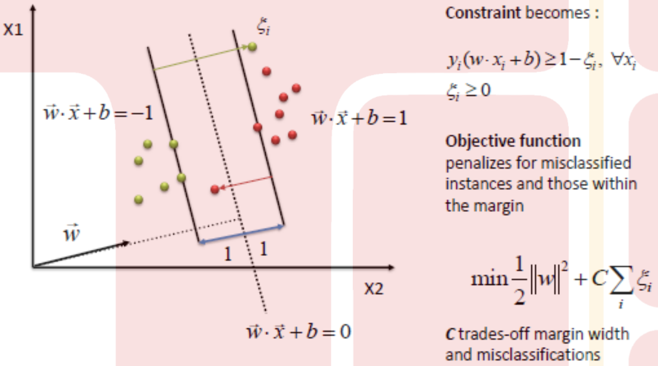

X1
ξ_i
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\vec{w}$
1
1
X2
$\vec{w}\cdot\vec{x}+b=0$

slack variable:

ξ_i

Allow some instances to fall off the margin, but penalize them

The algorithm tries to maintain the slack variable to zero while maximizing margin. However, it does not minimize the number of misclassifications (NP-complete problem) but the sum of distances from the margin hyperplanes.


X1
ξ_i
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\vec{w}$
1
1
X2
$\vec{w}\cdot\vec{x}+b=0$

Constraint becomes :

$y_i(w\cdot x_i+b)\ge 1-\xi_i,\ \forall x_i$

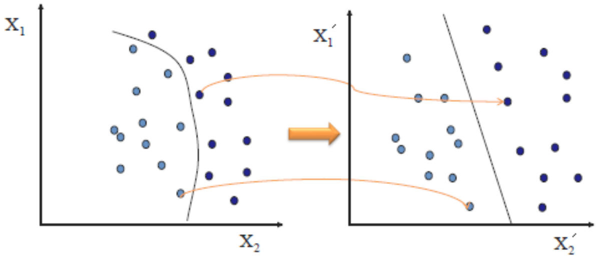
$\xi_i\ge 0$

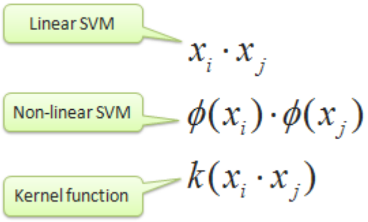
Objective function penalizes for misclassified instances and those within the margin

$$\min\frac{1}{2}\|w\|^2+C\sum_i\xi_i$$

C trades-off margin width and misclassifications

The simplest way to separate two groups of data is with a straight line (1 dimension), flat plane (2 dimensions) or an N-dimensional hyperplane. However, there are situations where a nonlinear region can separate the groups more efficiently. SVM handles this by using a kernel function (nonlinear) to map the data into a different space where a hyperplane (linear) cannot be used to do the separation. It means a non-linear function is learned by a linear learning machine in a high-dimensional feature space while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space. This is called *kernel trick* which means the kernel function transform the data into a higher dimensional feature space to make it possible to perform the linear separation.





Map data into new space, then take the inner product of the new vectors. The image of the inner product of the data is the inner product of the images of the data. Two kernel functions are shown below.

Polynomial

$$k(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i . \mathrm{x}_j)^d$$

Gaussian Radial Basis function

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(- \frac{\left\| \mathrm{x}_i - \mathrm{x}_j \right\|^2}{2\sigma^2} \right)$$

10. List advantages, disadvantages and applications of SVM

SVM Advantages

- SVM's are very good when we have no idea on the data.
- Works well with even unstructured and semi structured data like text, Images and trees.
- The kernel trick is real strength of SVM. With an appropriate kernel function, we can solve any complex problem.
- Unlike in neural networks, SVM is not solved for local optima.
- It scales relatively well to high dimensional data.
- SVM models have generalization in practice, the risk of over-fitting is less in SVM.
- SVM is always compared with ANN. When compared to ANN models, SVMs give better results.

SVM Disadvantages

- Choosing a "good" kernel function is not easy.
- Long training time for large datasets.
- Difficult to understand and interpret the final model, variable weights and individual impact.

- Since the final model is not so easy to see, we can not do small calibrations to the model hence its tough to incorporate our business logic.
- The SVM hyper parameters are Cost -C and gamma. It is not that easy to fine-tune these hyper-parameters. It is hard to visualize their impact

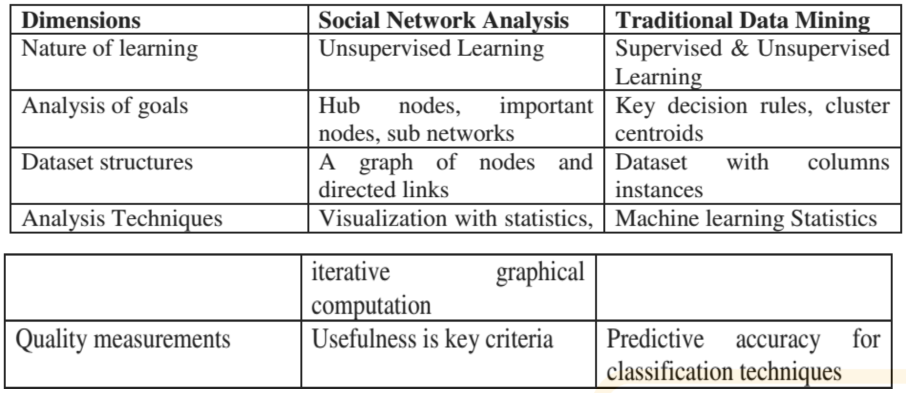
SVM Application

- Protein Structure Prediction
- Intrusion Detection
- Handwriting Recognition
- Detecting Steganography in digital images
- Breast Cancer Diagnosis
- Almost all the applications where ANN is used

11. What is social network analysis (SNA)? How is it different from other data mining techniques?

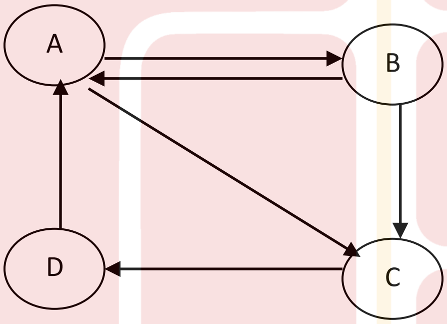
Social network analysis (**SNA**) is the process of investigating social structures using networks and graph theory. It characterizes networked structures in terms of *nodes* (individual actors, people, or things within the network) and the *ties*, *edges*, or *links* (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, information circulation, friendship and acquaintance networks, business networks, social networks, collaboration graphs. These networks are often visualized through *sociograms* in which nodes are represented as points and ties are represented as lines. These visualizations provide a means of qualitatively assessing networks by varying the visual representation of their nodes and edges to reflect attributes of interest.

a) Social Network Analysis and Traditional Data Mining



Dimensions	Social Network Analysis	Traditional Data Mining
Nature of learning	Unsupervised Learning	Supervised & Unsupervised Learning
Analysis of goals	Hub nodes, important nodes, sub networks	Key decision rules, cluster centroids
Dataset structures	A graph of nodes and directed links	Dataset with columns instances
Analysis Techniques	Visualization with statistics, iterative graphical computation	Machine learning Statistics
Quality measurements	Usefulness is key criteria	Predictive accuracy for classification techniques

12. Compute the Rank values for the nodes for the following network. Which the highest rank node after computation?



Solution :

a) Compute the Influence matrix (rank matrix)

- Assign the variables for influence value for each node, as Ra, Rb, Rc, Rd.
- There are two bound links from node A to nodes B and C. Thus, both B and C receives half of node A's influence. Similarly, there are two outbound links from node B to nodes C and A, So both C and A received half of node B's influence.

Ra = 0.5*Rb + Rd

Rb = 0.5*Ra

Rc =0.5*Ra + 0.5*Rb

Rd = Rc

	Ra	Rb	Rc	Rd
Ra	0	0.5	0	1.0
Rb	0.5	0	0	0
Rc	0.5	0.5	0	0
Rd	0	0	1.0	0

b) Set the initial set of rank values such as 1/n (n is number of nodes). As 4 nodes are there, initial rank values for all nodes are ¼ i.e 0.25

Variables	Initial Values
Ra	0.25
Rb	0.25
Rc	0.25
Rd	0.25

c) Compute the rank values for 1st iteration and then iteratively compute new rank values till they stabilized.

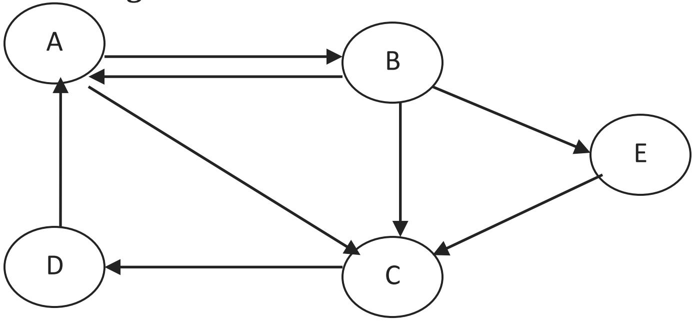
Variables	Initial Values	Iteration 1
Ra	0.25	0.375
Rb	0.25	0.125
Rc	0.25	0.250
Rd	0.25	0.250

Variables	Initial Values	Iteration 1	Iteration 2
Ra	0.25	0.375	0.3125
Rb	0.25	0.125	0.1875
Rc	0.25	0.250	0.250
Rd	0.25	0.250	0.250

Variables	Initial Values	Iteration 1	Iteration 2	--------	Iteration 8
Ra	0.25	0.375	0.3125	…….	0.333
Rb	0.25	0.125	0.1875	……	0.167
Rc	0.25	0.250	0.250	……	0.250
Rd	0.25	0.250	0.250	…….	0.250

The Final rank shows of node A is highest at 0.333

Exercise: Which is the highest rank node now?



a) Compute the Influence matrix (rank matrix)

- Assign the variables for influence value for each node, as Ra, Rb, Rc, Rd, Re

$$Ra = 1/3\ \ *Rb + Rd$$

$$Rb = ½ *Ra$$

$$Rc =1/2 *Ra + 1/3*Rb +Re$$

$$Rd = Rc$$

$$Re = 1/3*Rb$$

b) Set the initial set of rank values such as 1/n (n is number of nodes). As 4 nodes are there, initial rank values for all nodes are 1/5 i.e 0.2

Variables	**Initial Values**
Ra	0.2
Rb	0.2
Rc	0.2
Rd	0.2
Re	0.2

c) Compute the rank values for 1st iteration and then iteratively compute new rank values till they stabilized.

Variables	**Initial Values**	**Iteration 1**
Ra	0.2	0.267
Rb	0.2	0.1
Rc	0.2	0.367
Rd	0.2	0.367

Re	0.2	0.06

Variables	Initial Values	Iteration 1	Iteration 2
Ra	0.2	0.267	0.400
Rb	0.2	0.1	0.134
Rc	0.2	0.367	0.234
Rd	0.2	0.367	0.234
Re	0.2	0.067	0.033

Variables	Initial Values	Iteration 1	Iteration 2	Iteration 3
Ra	0.2	0.267	0.400	0.279
Rb	0.2	0.1	0.134	0.200
Rc	0.2	0.367	0.234	0.278
Rd	0.2	0.367	0.234	0.278
Re	0.2	0.067	0.033	0.045

Continue the iterations …. The Final rank shows of node A is highest