BIG DATA ANALYSIS

MODULE – 1

1. **How does the Hadoop MapReduce Data flow work for a word count program? Give an example. (08 Marks)**

- MapReduce is the processing engine of Hadoop that processes and computes large amounts of data
- It has 2 components – Map and Reduce.

Properties

- In MapReduce, Data flow is in one direction (map to reduce)
- The input data are not changed.
- The mapper and reducer data flow can be implemented in any number of ways to provide better performance

The Hadoop MapReduce Data flow for a word count is done by Apache Hadoop parallel MapReduce data flow

Map reduce algorithm for word count can be described as:

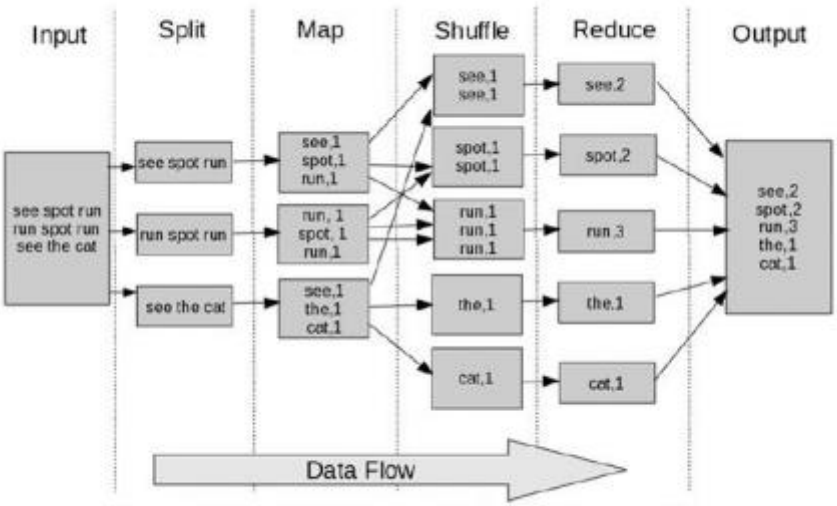


Figure 5.1 Apache Hadoop parallel MapReduce data flow

The basic steps are

1. Input Splits:

- HDFS distributes and replicates data over multiple servers.
- The default data chunk or block size is 64MB.

- The data is also replicated on multiple machines. These data slices are physical boundaries determined by HDFS
- The input splits used by MapReduce are logical boundaries based on the input data.

2. Map Step:

- The user provides the specific mapping process.
- MapReduce will try to execute the mapper on the machines where the block resides.
- The least busy node with the data will be chosen because file is replicated in HDFS.
- If all nodes holding the data are too busy, MapReduce will try to pick a node that is closest to the node that hosts the data block.

3. Combiner Step:

- The combiner step is optional
- We can provide an optimization or pre-reduction as part of the map stage.
- Here, key–value pairs are combined before the next stage.

4. Shuffle Step:

- Before the parallel reduction stage can complete, all similar keys must be combined and counted by the same reducer process.
- Results of the map stage is collected by key–value pairs and shuffled to the same reducer process.
- If only a single reducer process is used, the shuffle stage is not needed.

5. Reduce Step:

- The final step is the reduction.
- In this stage, the data reduction is performed as per the programmer's design.
- The results are written to HDFS.
- Each reducer will write an output file.

2. Briefly explain HDFS Name Node Federation, NFS Gateway, Snapshots, Checkpoint and Backups (08 Marks)

<u>Name Node Federation</u>

- It is an important feature of HDFS.

- Older versions of HDFS provided a single namespace for the entire cluster managed by a single Name Node. So, the resources of a single Name Node determined the size of the namespace. Federation addresses this limitation by adding support for multiple Name Nodes/namespaces to the HDFS file system.

Key benefits:

- **Namespace scalability**. HDFS cluster storage scales horizontally without placing a burden on the NameNode.
- **Better performance.** the file system read/write operations throughout by separating the total namespace.
- **System isolation**. Multiple NameNodes enable different categories of applications to be distinguished, and users can be isolated to different name spaces

NFS Gateway

- The HDFS NFS Gateway supports NFSv3 and enables HDFS to be mounted as part of the client's local file system.
- Users can browse the HDFS file system through their local file systems that provide an NFSv3 client compatible operating system.

Features

- Users can easily download/upload files from/to the HDFS file system to/from their local file system.
- Users can stream data directly to HDFS through the mount point. Appending to a file is supported, but random write capability is not supported.

Snapshots

The important features of snapshots are:

- Snapshots can be taken of a sub-tree of the file system or the entire files system.
- Snapshots can be used for data backup, protection against user errors, and disaster recovery.
- Snapshot creation is instantaneous.

- Blocks on the DataNodes are not copied because the snapshot files record the block list and the file size. There is no data copying, although it appears to the user that there are duplicate files.
- Snapshots do not adversely affect regular HDFS operations.

Checkpoint

- NameNode stores the metadata of the HDFS files system in a file called fs_image.
- File systems modifications are written to an edits log file, and at startup the NameNode merges the edits into a new fs_image.
- CheckpointNode periodically fetches edits from the NameNode, merges them, and returns an updated fs_image to the NameNode.

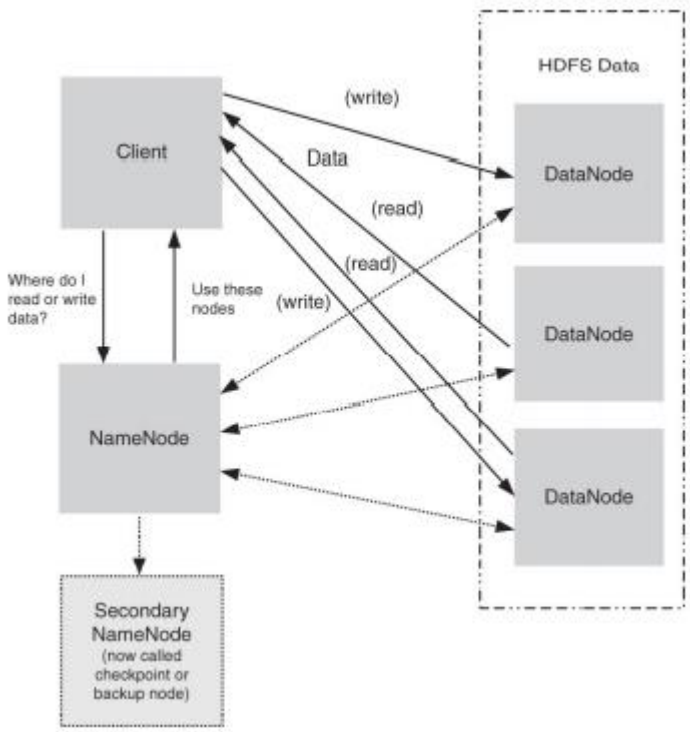
3. **What do you understand by HDFS? Explain its components with a neat diagram (10 Marks) +2**

- The Hadoop Distributed File System (HDFS) was designed for Big Data processing.
- HDFS is designed for data streaming where large amounts of data are read from disk in bulk.
- The HDFS block size is typically 64MB or 128MB.
- The large block and file sizes make it more efficient to reread data from HDFS than to try to cache the data.
- The most important aspect of HDF is its data locality.
- A principal design aspect of Hadoop MapReduce is the emphasis on moving the computation to the data rather than moving the data to the computation.

Components

- The design of HDFS is based on two types of nodes:
  - NameNode
  - Multiple DataNodes.
- In a basic design, a single NameNode manages all the metadata needed to store and retrieve the actual data from the DataNodes. No data is actually stored on the NameNode.

- But for a minimal Hadoop installation, there needs to be a single NameNode daemon and a single DataNode daemon running on at least one machine



- The design is a master/slave architecture in which the master (NameNode) manages the file system namespace and regulates access to files by clients.
- File system namespace operations such as opening, closing, and renaming files and directories are all managed by the NameNode.
- The NameNode also determines the mapping of blocks to DataNodes and handles DataNode failures.
- The NameNode manages block creation, deletion, and replication
- The slaves (DataNodes) are responsible for serving read and write requests from the file system to the clients.

4. **Bring out the components of HDFS block replication with an example. (06 Marks) +1**

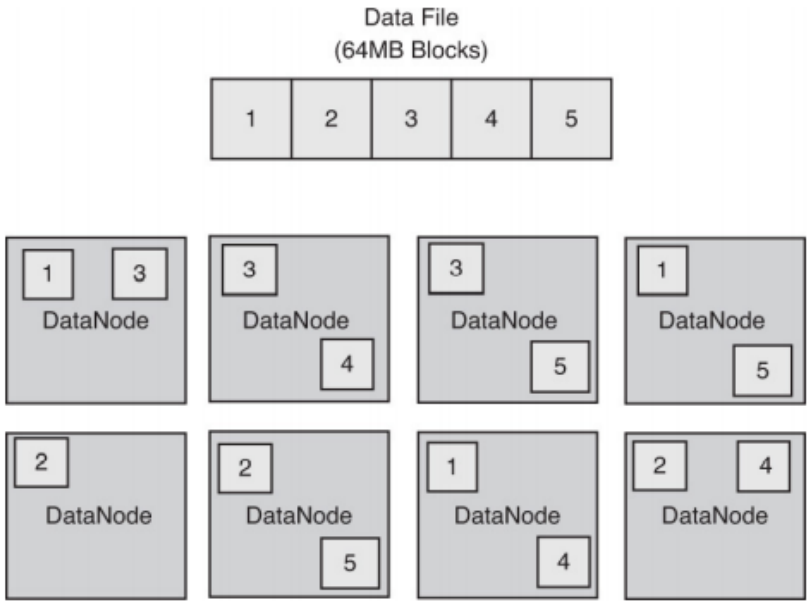


Figure 3.2 HDFS block replication example

- When HDFS writes a file, it is replicated across the cluster.
- The amount of replication is based on the value of dfs.replication in the hdfs-site.xml file.
- If several machines must be involved in the serving of a file, then a file could be rendered unavailable by the loss of any one of those machines. HDFS combats this problem by replicating each block across a number of machines.
- For Hadoop clusters containing >8 DataNodes, the replication value is usually set to 3.
- In a Hadoop cluster of ≤8 DataNodes but >1 DataNode, a replication factor of 2 is enough.
- For a single machine the replication factor is set to 1.
- The HDFS default block size is often 64MB and typical operating system block size is 4KB to 8KB. If 20KB file is written to HDFS, it will create a block that is approximately 20KB in size. If a file of size 80MB is written to HDFS, a 64MB block and a 16MB block will be created.
- HDFS blocks are not exactly the same as the data splits used by the MapReduce process.
- The HDFS blocks are based on size, while the splits are based on a logical partitioning of the data.

5. **Demonstrate various system roles and components of HDFS, with neat diagram (08 Marks) +1**

(Same answer of question 4 with diagram and the following points)

Thus, the various roles in HDFS are:

- HDFS uses a master/slave model designed for large file reading/streaming.
- The NameNode is a metadata server or "data traffic cop."
- HDFS provides a single namespace that is managed by the NameNode.
- Data is redundantly stored on DataNodes; there is no data on the NameNode.
- The SecondaryNameNode performs checkpoints of NameNode file system's state but is not a failover node.

(The following points are optional after the diagram)

- When a client writes data, it first communicates with the NameNode and requests to create a file.
- The NameNode determines how many blocks are needed and provides the client with the DataNodes that will store the data.
- As part of the storage process, the data blocks are replicated after they are written to the assigned node.
- Depending on how many nodes are in the cluster, the NameNode will attempt to write replicas of the data blocks on nodes that are in other separate racks. If there is only one rack, then the replicated blocks are written to other servers in the same rack.
- After the DataNode acknowledges that the file block replication is complete, the client closes the file and informs the NameNode that the operation is complete.
- The NameNode does not write any data directly to the DataNodes. It gives the client a limited amount of time to complete the operation. If it does not complete in the time period, the operation is canceled.

6. Illustrate any 8 HDFS commands and briefly explain. (08 Marks) +1

1. List Files in HDFS

- To list the files in the root HDFS directory, enter the following command:
  $ hdfsdfs -ls /
- To list files in your home directory, enter the following command:
  $ hdfsdfs –ls

2. Make a Directory in HDFS

- To make a directory in HDFS, use the following command. As with the –lscommand, when no path is supplied, the user's home directory is used

  $ hdfsdfs -mkdir stuff

3. Copy Files to HDFS

- To copy a file from your current local directory into HDFS, use the following command.
- If a full path is not supplied, your home directory is assumed. In this case, the file test is placed in the directory stuff that was created previously.

  $ hdfsdfs -put test stuff
- The file transfer can be confirmed by using the -ls command:

  $ hdfsdfs -ls stuff

Found 1 items

-rw-r--r-- 2 hdfshdfs 12857 2015-05-29 13:12 stuff/test

4. Copy Files from HDFS

- Files can be copied back to your local file system using the following command.
- In this case, the file we copied into HDFS, test, will be copied back to the current local directory with the name test-local.

  $ hdfsdfs -get stuff/test test-local

5. Copy Files within HDFS

- The following command will copy a file in HDFS:

  $ hdfsdfs -cp stuff/test test.hdfs

6. Delete a File within HDFS

- The following command will delete the HDFS file test.dhfsthat was created previously:

  $ hdfsdfs -rmtest.hdfs

Moved: 'hdfs://limulus:8020/user/hdfs/stuff/test' to trash at:hdfs://limulus:8020/user/hdfs/.Trash/Current

Note: When the fs.trash.interval option is set to a non-zero value in coresite.xml, all deleted files are moved to the user's .Trash directory. This can be avoided by including the -skipTrashoption.

$ hdfsdfs -rm –skip Trash stuff/test

Deleted stuff/test

7. Delete a Directory in HDFS

- The following command will delete the HDFS directory stuff and all its contents:

  $ hdfsdfs -rm -r -skipTrash stuff

Deleted stuff

8. Get an HDFS Status Report

- Regular users can get an abbreviated HDFS status report using the following command.
- Those with HDFS administrator privileges will get a full (and potentially long) report.
- Also, this command uses dfs admin instead of dfs to invoke administrative commands.
- The status report is similar to the data presented in the HDFS web GUI .

  $ hdfsdfsadmin –report

7. Demonstrate Apache Hadoop Parallel Map Reduce data flow with neat diagram for word count. (10 Marks) +3

(same as question 1)

8. Explain Terasort benchmark run steps. (06 Marks)

The terasort benchmark sorts a specified amount of randomly generated data. This benchmark provides combined testing of the HDFS and MapReduce layers of a Hadoop cluster. A full terasortbenchmark run consists of the following three steps:

1. Generating the input data via teragen program.

1. Run teragen to generate rows of random data to sort

```
/opt/hadoop-2.6.0/share/hadoop/mapreduce/
$ find / -name "hadoop-mapreduce-examples*.jar" -print
$ export HADOOP_EXAMPLES=/usr/hdp/2.2.4.2-2/hadoop-mapreduce
$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples.jar
```

2. Running the actual terasort benchmark on the input data.

2. Run terasort to sort the database

```
$ yarn jar $HADOOP EXAMPLES/hadoop-mapreduce-examples.jar
terasort /user/hdfs/TeraGen-50GB /user/hdfs/TeraSort-50GB
```

3. Validating the sorted output data via the teravalidate program

3. Run teravalidate to validate the sort.

```
$ yarn jar $HADOOP EXAMPLES/hadoop-mapreduce-examples.jar
teravalidate /user/hdfs/TeraSort-50GB /user/hdfs/TeraValid-50GB
```

- In general, each row is 100 bytes long; so, the total amount of data written is 100 times the number of rows specified as part of the benchmark (i.e., to write 100GB of data, use 1 billion rows). The input and output directories need to be specified in HDFS.
- To report results, the time for the actual sort (terasort) is measured and the benchmark rate in megabytes/second (MB/s) is calculated.
- For best performance, the actual terasort benchmark should be run with a replication factor of 1.
- The default number of terasort reducer tasks is set to 1.
- Increasing the number of reducers often helps with benchmark performance.

9. **Write the java code for MAP and REDUCE of word count problem. Describe the steps of compiling and removing the mapreduce program.**

Java code for Mapper

```
public class WordCount {
        public static class TokenizerMapper
                extends Mapper<Object, Text, Text, IntWritable>{
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();
        public void map(Object key, Text value, Context context) throws IOException,
        InterruptedException {
                StringTokenizer itr = new
                StringTokenizer(value.toString());
                while (itr.hasMoreTokens()) {
                        word.set(itr.nextToken());
                        context.write(word, one);
                        }
                }
        }

public static class IntSumReducer

        extends Reducer<Text,IntWritable,Text,IntWritable> {
```

```
private IntWritable result = new IntWritable();

public void reduce(Text key, Iterable<IntWritable> values, Context context) throws
IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
                sum += val.get();
                }
        result.set(sum);
        context.write(key, result);
        }
}
```

1. Make a local wordcount_classes directory.

```
$ mkdir wordcount_classes
```

2. Compile the WordCount.java program using the 'hadoop classpath' command to include all the available Hadoop class paths.

```
$ javac -cp `hadoop classpath` -d wordcount_classes
WordCount.java
```

3. The jar file can be created using the following command:

```
$ jar -cvf wordcount.jar -C wordcount_classes/
```

4. To run the example, create an input directory in HDFS and place a text file in the new directory. For this example, we will use the war-and-peace.txt file

```
$ hdfs dfs -mkdir war-and-peace-input
$ hdfs dfs -put war-and-peace.txt war-and-peace-input
```

5. Run the WordCount application using the following command:

```
$ hadoop jar wordcount.jar WordCount war-and-peace-input war-and-peace-output
```

10. Write code for simple mapper script and simple reducer script. +1

Listing 5.1 **Simple Mapper Script**

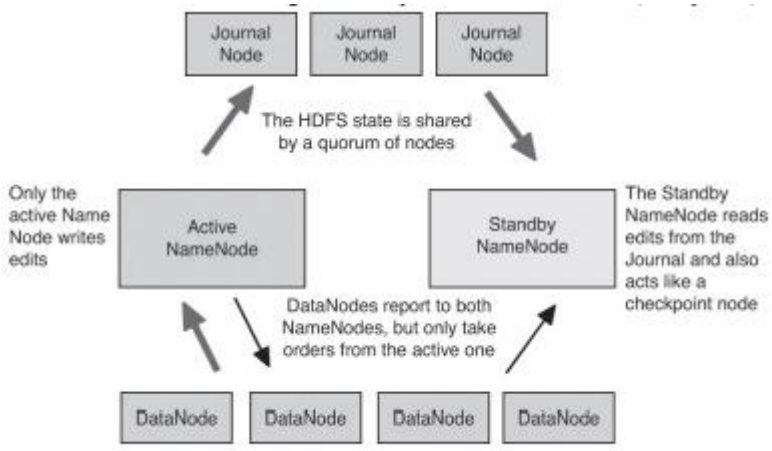
```
#!/bin/bash
while read line ; do
  for token in $line; do
  if [ "$token" = "Kutuzov" ] ; then
    echo "Kutuzov,1"
  elif [ "$token" = "Petersburg" ] ; then
    echo "Petersburg,1"
  fi
  done
done
```

Listing 5.2 **Simple Reducer Script**

```
#!/bin/bash
kcount=0
pcount=0
while read line ; do
  if [ "$line" = "Kutuzov,1" ] ; then
   let kcount=kcount+1
  elif [ "$line" = "Petersburg,1" ] ; then
   let pcount=pcount+1
  fi
done
echo "Kutuzov,$kcount"
echo "Petersburg,$pcount"
```

11. Explain the following roles in HDFS deployment with a diagram (i) High availability (ii) Name Node function (08 Marks)

(i) High availability



- A High Availability Hadoop cluster has two (or more) separate NameNode machines.
- Each machine is configured with exactly the same software.
- One of the NameNode machines is in the Active state, and the other is in the Standby state.
- Like a single NameNode cluster, the Active NameNode is responsible for all client HDFS operations in the cluster.

- The Standby NameNode maintains enough state to provide a fast failover (if required)

(ii). Name Node function

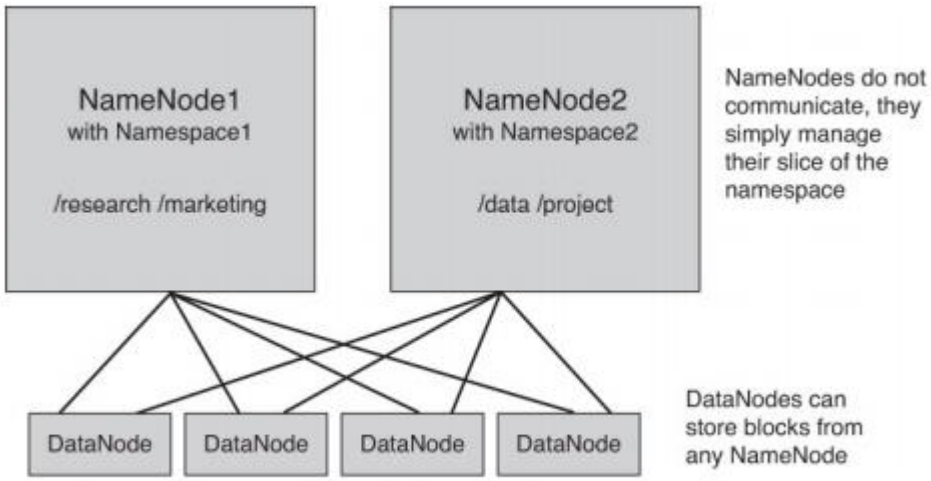


Figure 3.4 HDFS NameNode Federation example

(same as in question 2)

BIG DATA ANALYSIS

MODULE – 2

1. Explain the two step Apache Sqoop data import and export method (08 Marks) +4

1. Import method

The data import is done in two steps.

- In the first step, Sqoop examines the database to gather the necessary metadata for the data to be imported.
- The second step is a map-only (no reduce step) Hadoop job that Sqoop submits to the cluster.

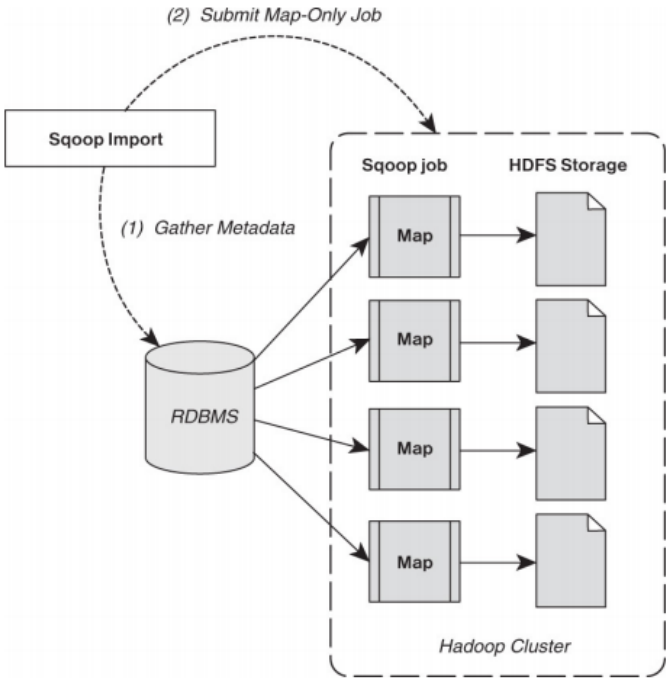


Figure 7.1 Two-step Apache Sqoop data import method (Adapted from Apache Sqoop Documentation)

- The imported data is saved in an HDFS directory.
- Sqoop will use the database name for the directory, or the user can specify any alternative directory where the files should be populated.
- By default, these files contain comma-delimited fields, with new lines separating different records.
- We can easily override the format in which data are copied over by explicitly specifying the field separator and record terminator characters.
- Once placed in HDFS, the data is ready for processing

2. Export method

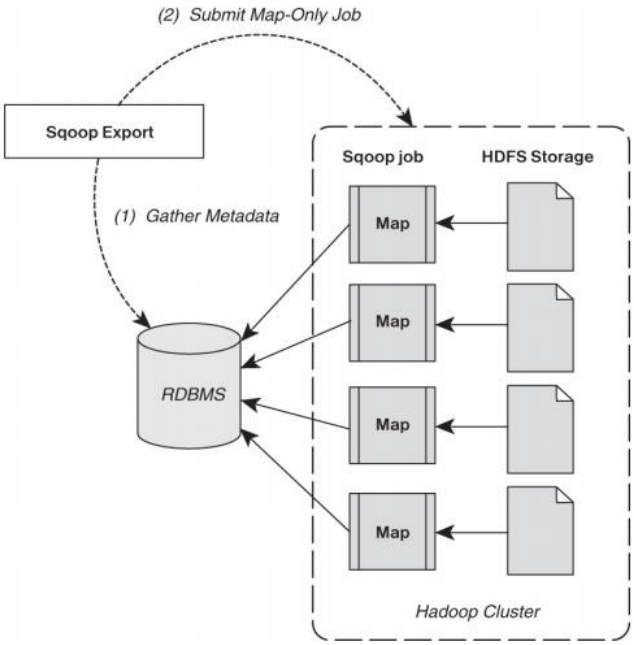
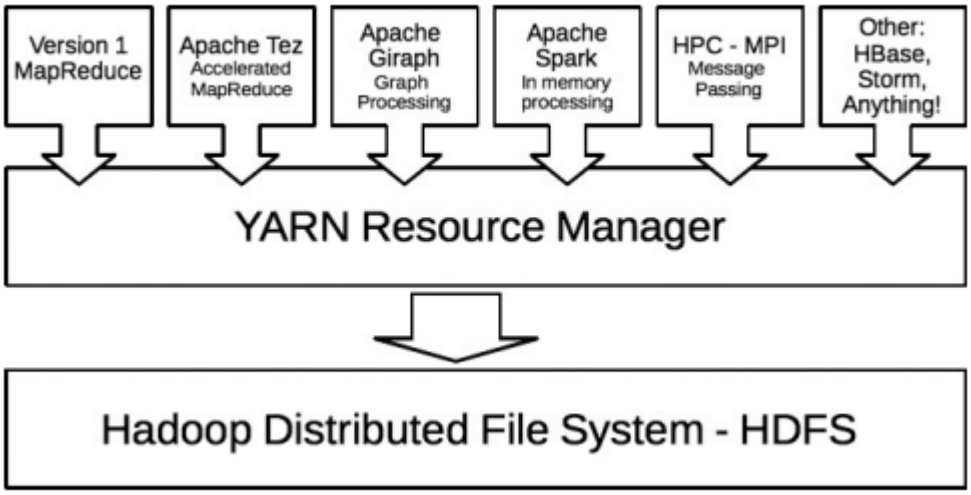


Figure 7.2 Two-step Sqoop data export method (Adapted from Apache Sqoop Documentation)

- The export is done in two steps
- As in the import process, the first step is to examine the database for metadata.
- The export step again uses a map-only Hadoop job to write the data to the database.
- Sqoop divides the input data set into splits, then uses individual map tasks to push the splits to the database.
- Again, this process assumes the map tasks have access to the database

2. With a neat diagram explain YARN Application frameworks (08 Marks) +1



- YARN presents a resource management platform

- It provides services such as scheduling, fault monitoring, data locality, and more to MapReduce and other frameworks.

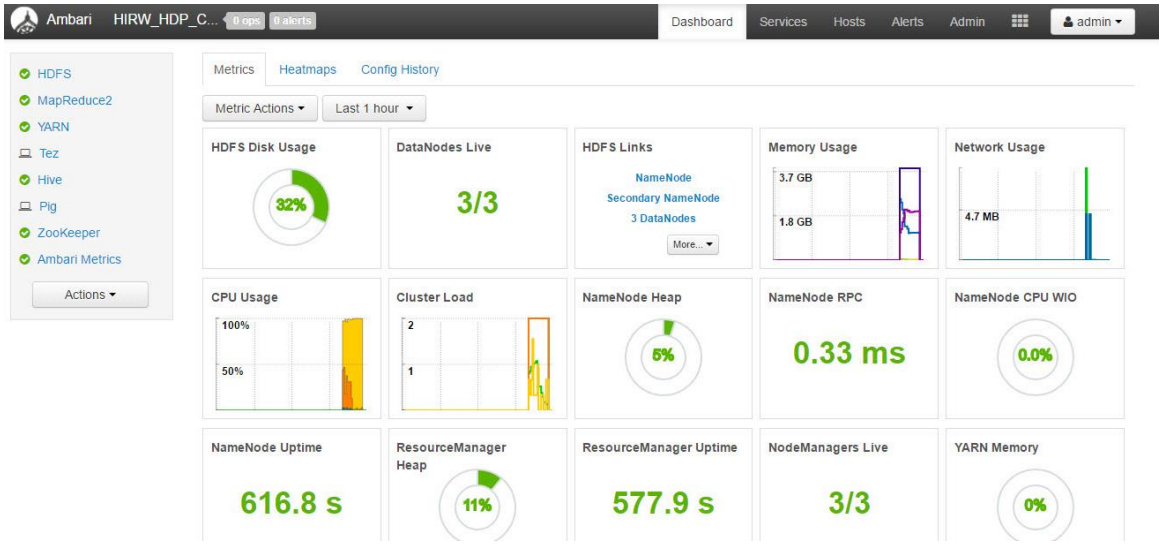
A brief survey of emerging open-source YARN application frameworks that are being developed to run under YARN are:

1. **Distributed-Shell**
   - Distributed-Shell is an example application included with the Hadoop core components.
   - It demonstrates how to write applications on top of YARN.
   - It provides a simple method for running shell commands and scripts in containers in parallel on a Hadoop YARN cluster
2. **Hadoop MapReduce**
   - MapReduce was the first YARN framework and drove many of YARN's requirements.
   - It is integrated tightly with the rest of the Hadoop ecosystem projects, such as Apache Pig, Apache Hive, and Apache Oozie
3. **Apache Tez**
   - Many Hadoop jobs involve the execution of a complex directed acyclic graph (DAG) of tasks using separate MapReduce stages. Apache Tez generalizes this process.
   - It enables these tasks to be spread across stages so that they can be run as a single, all-encompassing job.
   - Tez can be used as a MapReduce replacement for projects such as Apache Hive and Apache Pig.
4. **Apache Giraph**
   - Apache Giraph is an iterative graph processing system built for high scalability.
   - Facebook, Twitter, and LinkedIn use it to create social graphs of users.
   - The native Giraph implementation under YARN provides the user with an iterative processing model that is not directly available with MapReduce.
5. **Hoya: HBase on YARN**
   - The Hoya project creates dynamic and elastic Apache HBase clusters on top of YARN

- Hoya also asks YARN for the number of containers matching the number of HBase region servers it needs.

6. **Dryad on YARN**
   - Microsoft's Dryad provides a DAG as the abstraction of execution flow
   - This framework is ported to run natively on YARN and is fully compatible with its non-YARN version.
7. **Apache Spark**
   - Spark was initially developed for applications in which keeping data in memory improves performance, such as iterative algorithms, and it is common in machine learning, and interactive data mining
   - Spark holds intermediate results in memory, rather than writing them to disk.
   - Spark supports more than just MapReduce functions;
8. **Apache Storm**
   - This framework is designed to process unbounded streams of data in real time.
   - It can be used in any programming language.
   - The basic Storm use-cases include real-time analytics, online machine learning, continuous computation, distributed RPC (remote procedure calls), ETL (extract, transform, and load) etc.

3. **Explain Apache Ambari dashboard view of a Hadoop cluster (08 Marks)**



- The Dashboard view provides small status widgets for many of the services running on the cluster.
- The actual services are listed on the left-side vertical menu.

- The Dashboard view also includes a heatmap view of the cluster.
- Cluster heatmaps physically map selected metrics across the cluster.
- When you click the Heatmaps tab, a heatmap for the cluster will be displayed.
- To select the metric used for the heatmap, choose the desired option from the Select Metric pull-down menu.
- Configuration history is the final tab in the dashboard window.
- This view provides a list of configuration changes made to the cluster.
- Ambari enables configurations to be sorted by Service, Configuration Group, Data, and Author.
- To find the specific configuration settings, click the service name

4. How basic Hadoop YARN administration is carried out? Explain (08 Marks) +1

- YARN has several built-in administrative features and commands.

Decommissioning YARN Nodes

- If a NodeManager host/node needs to be removed from the cluster, it should be decommissioned first.
- If the node is responding, we can easily decommission it from the Ambari web UI.
- Go to the Hosts view, click on the host, and select Decommission from the pull-down menu next to the NodeManager component.
- Note: The host may also be acting as a HDFS DataNode.
- Use the Ambari Hosts view to decommission the HDFS host in a similar fashion.

YARN WebProxy

- The Web Application Proxy is a separate proxy server in YARN that addresses security issues with the cluster web interface on ApplicationMasters.
- By default, the proxy runs as part of the Resource Manager itself, but it can be configured to run in a stand-alone mode by adding the configuration property yarn.web-proxy.address to yarn-site.xml.
- In stand-alone mode, yarn.web-proxy.principal and yarn.web-proxy.keytab control the Kerberos principal name and the corresponding keytab, respectively, for use in secure mode.

- These elements can be added to the yarn-site.xml if required.

Using the JobHistoryServer

- The removal of the JobTracker and migration of MapReduce from a system to an application-level framework requires the creation of a place to store MapReduce job history.
- The JobHistoryServer provides all YARN MapReduce applications with a central location.
- The settings for the JobHistoryServer can be found in the mapred-site.xml file.

Managing YARN Jobs

- YARN jobs can be managed using the yarn application command.
- The following options, including -kill, -list, and -status, are available to the administrator with this command.
- MapReduce jobs can also be controlled with the 'mapred job' command

5. Discuss the different views supported by Apache Ambari (06 Marks)

1. Dashboard View

- The Dashboard view provides small status widgets for many of the services running on the cluster.
- We can move, edit, remove, or add these widgets
- The actual services are listed on the left-side vertical menu

2. Services View

- The Services menu provides a detailed look at each service running on the cluster.
- Similar to the Dashboard view, the currently installed services are listed on the left-side menu.
- To select a service, click the service name in the menu.
- When applicable, each service will have its own Summary, Alerts and Health Monitoring, and Service Metrics windows
- The Alerts and Health Checks window provides the latest status of the service and its component systems.
- Several important real-time service metrics are displayed as widgets at the bottom of the screen.

- As on the dashboard, these widgets can be expanded to display a more detailed view.

3. Hosts View

- Selecting the Hosts menu item provides the host name, IP address, number of cores, memory, disk usage, current load average, and Hadoop components in tabular form.
- To display the Hadoop components installed on each host, click the links in the rightmost columns.
- We can also add new hosts by using the Actions pull-down menu.
- The remaining options in the Actions pull-down menu provide control over the various service components running on the hosts.

4. Admin View

- The Administration (Admin) view provides three options.
  - First displays a list of installed software. This Repositories listing generally reflects the version of Hortonworks Data Platform (HDP) used during the installation process.
  - Second, The Service Accounts option lists the service accounts added when the system was installed. These accounts are used to run various services and tests for Ambari.
  - The third option, Security, sets the security on the cluster. A fully secured Hadoop cluster is important in many instances and should be explored if a secure environment is needed.

5. Views View

- Views allows us to extend and customize Ambari to meet your specific needs.

6. Explain different HDFS administration features.

The following section covers some basic administration aspects of HDFS:

The NameNode User Interface

- Monitoring HDFS can be done in several ways. One of the more convenient ways to get a quick view of HDFS status is through the NameNode user interface.
- This web-based tool provides essential information about HDFS and offers the capability to browse the HDFS namespace and logs.

- The web-based UI can be started from within Ambari or from a web browser connected to the NameNode.
- In Ambari, simply select the HDFS service window and click on the Quick Links pull-down menu in the top middle of the page.
  Select NameNode UI. A new browser tab will open with the UI

Perform an FSCK on HDFS

- To check the health of HDFS, we can issue the hdfs fsck <path> (file system check) command.
- The entire HDFS namespace can be checked, or a subdirectory can be entered as an argument to the command.

Balancing HDFS

- Based on usage patterns and DataNode availability, the number of data blocks across the DataNodes may become unbalanced.
- To avoid over-utilized DataNodes, the HDFS balancer tool rebalances data blocks across the available DataNodes.
- Data blocks are moved from over utilized to under-utilized nodes to within a certain percent threshold.
- Rebalancing can be done when new DataNodes are added or when a DataNode is removed from service.
- This step does not create more space in HDFS, but improves efficiency

HDFS Safe Mode

- When the NameNode starts, it loads the file system state from the fsimage and then applies the edits log file.
- It then waits for DataNodes to report their blocks.
- During this time, the NameNode stays in a read-only Safe Mode.
- The NameNode leaves Safe Mode automatically after the DataNodes have reported that most file system blocks are available

Decommissioning HDFS Nodes

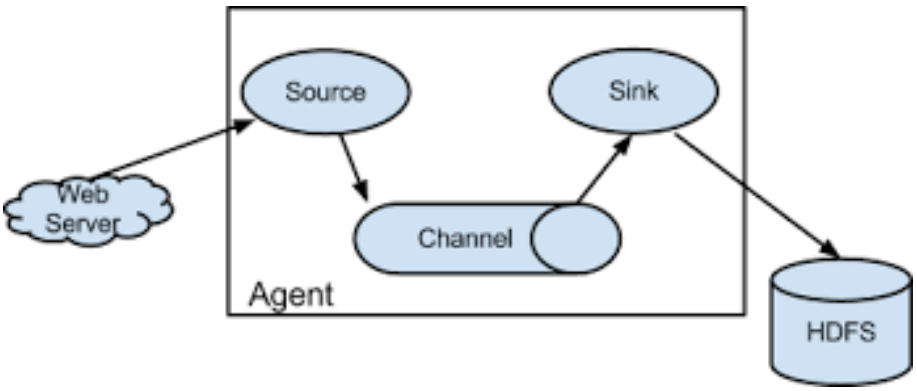
- If we have to remove a DataNode host/node from the cluster, we should decommission it first.
- If the node is responding, it can be easily decommissioned from the Ambari web UI.
- Simply go to the Hosts view, click on the host, and selected Decommission from the pull-down menu next to the DataNode component.

- Use the Ambari Hosts view to decommission the YARN host in a similar way

7. **List and explain Apache HBase Capabilities**
    - Apache HBase is an open source, distributed, versioned, nonrelational database modeled after Google's Bigtable.
    - Like Bigtable, HBase leverages the distributed data storage provided by the underlying distributed file systems spread across commodity servers.
    - Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

    Some of the more important features include the following capabilities:

    - Linear and modular scalability
    - Strictly consistent reads and writes
    - Automatic and configurable sharding of tables
    - Automatic failover support between RegionServers
    - Convenient base classes for backing Hadoop MapReduce jobs with Apache HBase tables
    - Easy-to-use Java API for client access
8. **Illustrate Flume agent and pipeline of Apache to acquire data streams with neat diagram**
    - Apache Flume is an independent agent designed to collect, transport, and store data into HDFS.
    - Often data transport involves a number of Flume agents that may traverse a series of machines and locations.
    - Flume is often used for log files, social media-generated data, email messages, and just about any continuous data source.



A Flume agent must have all three of these components defined.

a. **Source.** The source component receives data and sends it to a channel. It can send the data to more than one channel. The input data can be from a real-time source or another Flume agent.

b. **Channel.** A channel is a data queue that forwards the source data to the sink destination. It can be thought of as a buffer that manages input (source) and output (sink) flow rates.

c. **Sink.** The sink delivers data to destination such as HDFS, a local file, or another Flume agent.

Sqoop agents may be placed in a pipeline, possibly to traverse several machines or domains. This configuration is normally used when data are collected on one machine and sent to another machine that has access to HDFS.

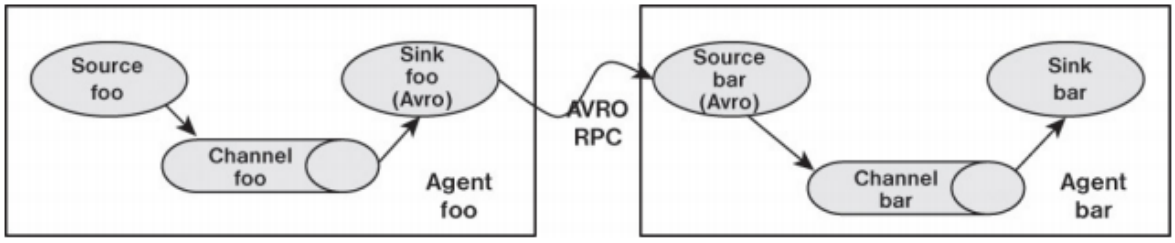


Figure 7.4 Pipeline created by connecting Flume agents (Adapted from Apache Flume Sqoop Documentation)

- In a Flume pipeline, the sink from one agent is connected to the source of another.
- The data transfer format normally used by Flume, which is called Apache Avro, provides several useful features.
- Avro is a data serialization/deserialization system that uses a compact binary format.
- The schema is sent as part of the data exchange and is defined using JSON (JavaScript Object Notation).
- Avro also uses remote procedure calls (RPCs) to send data i.e., an Avro sink will contact an Avro source to send data.

9. **List and explain XML configuration used for Hadoop configuration accomplishment.**

Hadoop has two main areas of administration: the YARN resource manager and the HDFS file system. Other application frameworks (e.g., the MapReduce framework) and tools have their own management files. Hadoop configuration is accomplished through the use of XML configuration files. The basic files and their function are as follows:

- `core-default.xml`: System-wide properties
- `hdfs-default.xml`: Hadoop Distributed File System properties

- `mapred-default.xml`: Properties for the YARN MapReduce framework
- `yarn-default.xml`: YARN properties

10. Explain with a neat diagram, the Apache Oozie work flow for Hadoop architecture

- Oozie is a workflow director system designed to run and manage multiple related Apache Hadoop jobs. For instance, complete data input and analysis may require several discrete Hadoop
- Oozie is designed to construct and manage these workflows. Oozie is not a substitute for the YARN scheduler.
- Oozie workflow jobs are represented as directed acyclic graphs (DAGs) of actions. (DAGs are basically graphs that cannot have directed loops.)

Three types of Oozie jobs are permitted:

- Workflow—a specified sequence of Hadoop jobs with outcome-based decision points and control dependency. Progress from one action to another cannot happen until the first action is complete
- Coordinator—a scheduled workflow job that can run at various time intervals or when data become available.
- Bundle—a higher-level Oozie abstraction that will batch a set of coordinator jobs

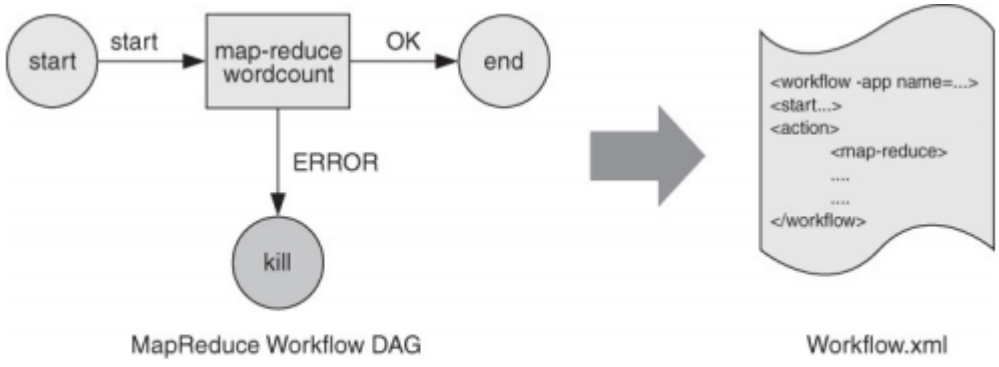


Figure 7.6 A simple Oozie DAG workflow (Adapted from Apache Oozie Documentation)

Oozie workflow definitions are written in hPDL (an XML Process Definition Language). This workflow contains several types of nodes:

- **Control flow nodes** define the beginning and the end of a workflow. They include start, end, and optional fail nodes.
- **Action nodes** are where the actual processing tasks are defined. When an action node finishes, the remote systems notify Oozie and the next node in the workflow is executed. Action nodes can also include HDFS commands.

- **Fork/join nodes** enable parallel execution of tasks in the workflow. The fork node enables two or more tasks to run at the same time. A join node represents a rendezvous point that must wait until all forked tasks complete.
- **Control flow nodes** enable decisions to be made about the previous task. Control decisions are based on the results of the previous action (e.g., file size or file existence). Decision nodes are essentially switch-case statements that use JSP EL (Java Server Pages—Expression Language) that evaluate to either true or false.

11. How do you run Map Reduce and Message passing Interface (MPI) on YARN architecture?

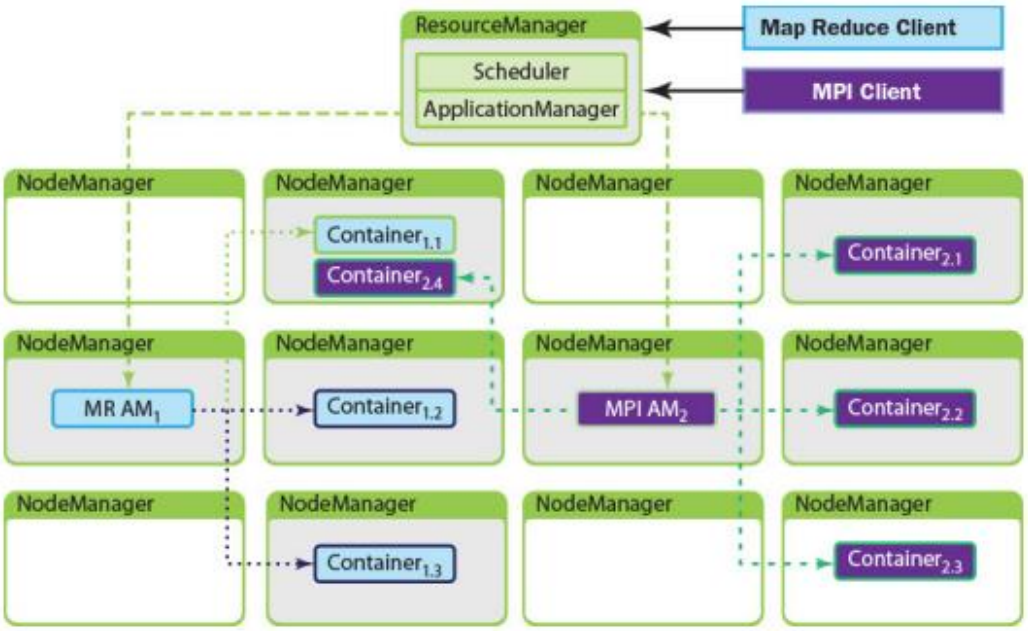


Figure 8.1 YARN architecture with two clients (MapReduce and MPI). The darker client (MPI AM_2) is running an MPI application, and the lighter client (MR AM_1) is running a MapReduce application. (From Arun C. Murthy, et al., *Apache Hadoop™ YARN*, copyright © 2014, p. 45. Reprinted and electronically reproduced by permission of Pearson Education, Inc., New York, NY.)

- MapReduce runs as a YARN application. Consequently, it may be necessary to adjust some of the mapred-site.xml properties as they relate to the map and reduce containers.

The following properties are used to set some Java arguments and memory size for both the map and reduce containers:

- mapred.child.java.opts - provides a larger or smaller heap size for child JVMs of maps e.g., --Xmx2048m.
- mapreduce.map.memory.mb provides a larger or smaller resource limit for maps (default = 1536MB).

- mapreduce.reduce.memory.mb provides a larger heap size for child JVMs of maps (default = 3072MB).
- mapreduce.reduce.java.opts provides a larger or smaller heap size for child reducers
- The Message Passing Interface (MPI) is widely used in high-performance computing (HPC).
- MPI is primarily a set of optimized message-passing library calls for C, C++, and Fortran that operate over popular server interconnects such as Ethernet and InfiniBand.
- Because users have full control over their YARN containers, there is no reason why MPI applications cannot run within a Hadoop cluster.
- Currently, an alpha version of MPICH2 is available for YARN that can be used to run MPI applications.

12. What do you understand by YARN distributed shell?

- The Hadoop YARN project includes the Distributed-Shell application, which is an example of a Hadoop non-MapReduce application built on top of YARN.
- Distributed-Shell is a simple mechanism for running shell commands and scripts in containers on multiple nodes in a Hadoop cluster.
- This application is not meant to be a production administration tool, but rather a demonstration of the non-MapReduce capability that can be implemented on top of YARN.
- There are multiple mature implementations of a distributed shell that administrators typically use to manage a cluster of machines.
- Distributed-Shell can be used as a starting point for exploring and building Hadoop YARN applications.
- The Hadoop YARN Distributed-Shell is a simple demonstration of a non-MapReduce program.
- It can be used to learn about how YARN operates and launches jobs across the cluster.
- The structure and operation of YARN programs are designed to provide a highly scalable and flexible method to create Hadoop applications.

13. What is the significance of Apache pig in Hadoop context? Describe the main components and working of Apache pig with a suitable example

- Apache Pig is a high-level language that enables programmers to write complex MapReduce transformations using a simple scripting language.
- Pig Latin (the actual language) defines a set of transformations on a data set such as aggregate, join, and sort.
- Pig is often used to extract, transform, and load (ETL) data pipelines, quick research on raw data, and iterative data processing.

Apache Pig has several usage modes.

- The first is a local mode in which all processing is done on the local machine. The non-local (cluster) modes are MapReduce and Tez. These modes execute the job on the cluster using either the MapReduce engine or the optimized Tez engine.
- There are also interactive and batch modes available; they enable Pig applications to be developed locally in interactive modes, using small amounts of data, and then run at scale on the cluster in a production mode.

Pig Example Walk-Through

For this example, the following software environment is assumed. Other environments should work in a similar fashion.

OS: Linux

Platform: RHEL 6.6

Hortonworks HDP 2.2 with Hadoop version: 2.6

Pig version: 0.14.0

In this simple example, Pig is used to extract user names from the `/etc/passwd` file. To begin the example, copy the `passwd` file to a working directory for local Pig operation:

```
$ cp /etc/passwd .
```

Next, copy the data file into HDFS for Hadoop MapReduce operation:

```
$ hdfs dfs -put passwd passwd
```

In the following example of local Pig operation, all processing is done on the local machine (Hadoop is not used). First, the interactive command line is started:

```
$ pig -x local
```

If Pig starts correctly, we can see a grunt> prompt. we can also see a bunch of INFO messages. Next, enter the following commands to load the passwd file and then grab the user name and dump it to the terminal. Note that Pig commands must end with a semicolon (;).

```
grunt> A = load 'passwd' using PigStorage(':');

grunt> B = foreach A generate $0 as id;

grunt> dump B;
```

The processing will start and a list of user names will be printed to the screen. To exit the interactive session, enter the command quit.

```
$ grunt> quit
```

To use Hadoop MapReduce, start Pig as follows:

```
$ pig -x mapreduc
```

14. **With neat diagrams, explain the Oozie DAG workflow and the types of nodes in the workflow.**
    (same as question 10)
15. **What is Apache Flume? Describe the features, components and working of Apache Flume.**
    (same as question 8)

BIG DATA ANALYSIS

MODULE – 3

1. **List any ten different Business Intelligence applications and explain them in brief (08 Marks) +1**

1. **Education**

Student Enrollment (Recruitment and Retention):

- Marketing to new potential students requires schools to develop profiles of the students that are most likely to attend.
- Schools can develop models of what kinds of students are attracted to the school, and then reach out to those students.
- The students at risk of not returning can be flagged, and corrective measures can be taken in time.

Course Offerings

- Schools can use the class enrolment data to develop models of which new courses are likely to be more popular with students.
- This can help increase class size, reduce costs, and improve student satisfaction.

Fund-Raising from Alumni and Other Donors

- Schools can develop predictive models of the alumni that are most likely to pledge financial support to the school.
- Schools can create a profile for alumni more likely to pledge donations to the school.
- This can lead to a reduction in the cost of mailings and other forms of outreach to alumni

2. Retail

Optimize Inventory Levels at Different Locations

- Predicting sales trends dynamically help retailers move inventory to where it is most in demand.

- Retail organizations can provide their suppliers with real time information about sales of their items, so the suppliers can deliver their product to the right locations and minimize stock-outs.

Improve Store Layout and Sales Promotions

- A market basket analysis can develop predictive models of the products often sold together.
- Promotional discounted product bundles can be created to push a nonselling item along with a set of products that sell well together.

Optimize Logistics for Seasonal Effects

- Understanding the products that are in season in which market can help retailers dynamically manage prices to ensure their inventory is sold during the season.

Minimize Losses due to Limited Shelf Life

- Perishable goods offer challenges in terms of disposing off the inventory in time. By tracking sales trends, the perishable products at risk of not selling before the sell-by date, can be suitably discounted and promoted

3. Banking

Automate the Loan Application Process

- Decision models can be generated from past data that predict the likelihood of a loan proving successful.
- These can be inserted in business processes to automate the financial loan approval process.

Detect Fraudulent Transactions

- Billions of financial transactions happen around the world every day.
- Exception-seeking models can identify patterns of fraudulent transactions.

Maximize Customer Value (Cross-selling, Up-selling)

- A checking account customer in good standing could be offered home, auto, or educational loans on more favorable terms than other customers, and thus, the value generated from that customer could be increased.

Optimize Cash Reserves with Forecasting

- Banks have to maintain certain liquidity to meet the needs of depositors who may like to withdraw money.
- Using past data and trend analysis, banks can forecast how much to keep and invest the rest to earn interest

4. Public sector

Law Enforcement

- Social behavior is a lot more patterned and predictable than one would imagine.
- Internet chatter can be analyzed to learn about and prevent any evil designs.

Scientific Research

- Any large collection of research data is suitable to being mined for patterns and insights.
- Protein folding (microbiology), nuclear reaction analysis (sub-atomic physics), disease control (public health) are some examples where data mining can yield powerful new insights.

5. Telecom

6. Manufacturing

7. Insurance

8. Financial Services

9. Healthcare and wellness

10. Customer Relationship Management

2. **With a neat diagram explain Data warehousing architecture. (08 Marks) +1**

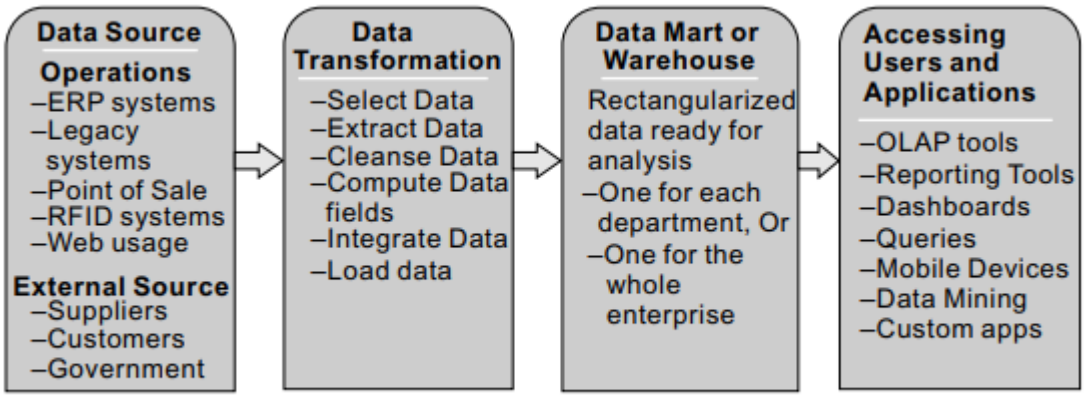


FIGURE 3.1 Data Warehousing Architecture

DW has four key elements:

- 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.

Data sources

DWs are created from structured data sources. Unstructured data, such as text data, would need to be structured before inserted into DW.

1. Operations data include 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 DW.

2. External syndicated data, such as weather or economic activity data, could also be added to DW, as needed, to provide good contextual information to decision makers.

Data transformation

The heart of a useful DW is the processes to populate the DW with good quality data. This is called the extract-transform-load (ETL) cycle.

Data Access

Data from DW could be accessed for many purposes, through many devices. Data from DW is used to provide data for mining purposes. Parts of the data would be extracted, and then combined with other relevant data, for data mining.

3. **How do you evaluate data mining results, explain with confusion matrix? (08 Marks)**

   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

Confusion Matrix		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	False Positive (FP)

FIGURE 4.1 Confusion Matrix

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

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

All classification techniques have a predictive accuracy associated with a predictive model. The highest value can be 100 percent. In practice, predictive models with more than 70 percent accuracy can be considered usable in business domains, depending upon the nature of the business

4. **Explain with a neat diagram different types of graphs (08 Marks)**

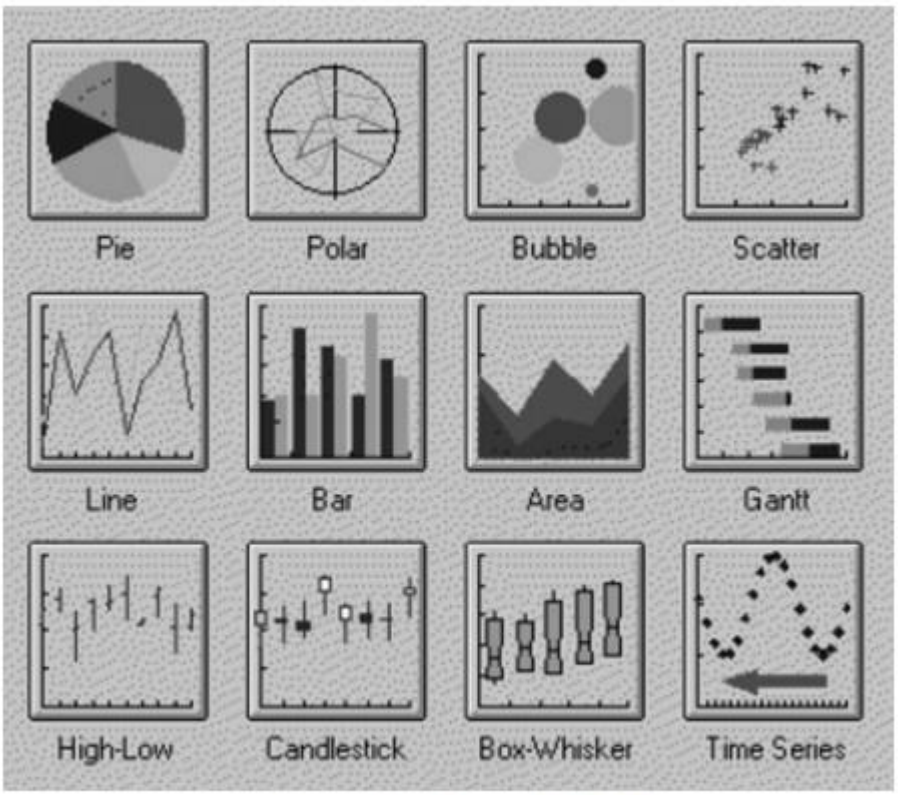


FIGURE 5.1 Types of Graphs

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 the line graphs of all the variables.
2. **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.
3. **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.
4. **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.
5. **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.
6. **Bubble Graph** This is an interesting way of displaying multiple dimensions in one chart. It is a variant of the scatter plot with many data points marked in two dimensions.
7. **Dial** 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.
8. **Geographical Data Maps** These are particularly useful to denote statistics.

5. **What is Business Intelligence? List the different BI applications and explain in detail any 5 applications (10 Marks) +1**

   Business intelligence is a broad set of information technology (IT) solutions that includes tools for gathering, analyzing, and reporting information to the users about performance of the organization and its environment. These IT solutions are among the most highly prioritized solutions for investment

   (continue question 1)

6. **Explain with diagram CRISP-DM data mining cycle (08 Marks)+2**

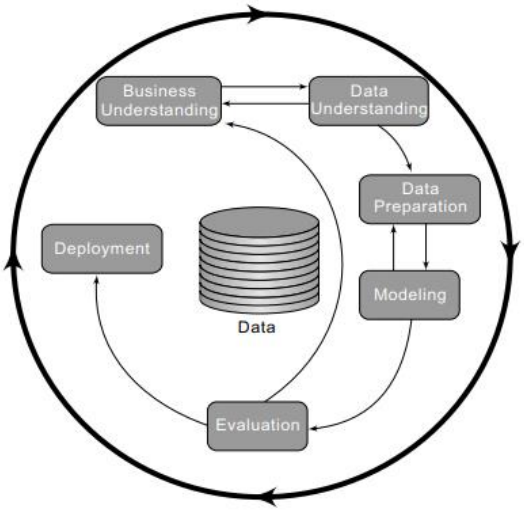


FIGURE 4.3 CRISP-DM Data Mining Cycle

Cross-Industry Standard Process for Data Mining (CRISP-DM) has 6 essential steps.

1. **Business Understanding -** selecting a data mining project is like any other project, in which it should show strong payoffs if the project is successful. There should be strong executive support for the data mining project, which means that the project aligns well with the business strategy.
2. **Data Understanding** 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 understands the domain and the data. Data cleaning can take 60-70 percent of the time in a data mining project.
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 evaluate and improve the model's predictive accuracy with more test data. When the accuracy has reached some satisfactory level, then the model should be deployed
6. **Dissemination and Rollout -** It is important that the data mining solution is presented to the key stakeholders, and is deployed in the organization. Otherwise, the project will be a waste of time and a setback for establishing and supporting a data-based decision-process culture in the organization.

7. **Describe the common data mining mistakes (04 Marks)**

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. Getting the right answer to an irrelevant question could be interesting, but it would be pointless from a business perspective. A good goal would be one that would deliver a good ROI to the organization.

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. There may be insufficient knowledge about the data, or metadata. Examine the data with a critical eye and do not naively believe everything you are told about the data.

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. This can come from being sloppy about keeping track of the data mining procedure and results. Not leaving sufficient time for data acquisition, selection and preparation can lead to data quality issues, and GIGO. Similarly not providing enough time for testing the model, training the users and deploying the system can make the project a failure.

Mistake #4 Insufficient Business Knowledge

- Without a deep understanding of the business domain, the results would be gibberish and meaningless. Don't make erroneous assumptions, courtesy of experts. Don't rule out anything when observing data analysis results. Don't ignore suspicious (good or bad) findings and quickly move on. Be open to surprises. Even when insights emerge at one level, it is important to slice and dice the data at other levels to see if more powerful insights can be extracted.

8. **List and describe the various charts used for data visualization (04 Marks)**

(same as Question 4)

9. **Justify the importance of Business Intelligence tools in: (12 Marks)**

i. Education ii. Retail iii. Banking iv. Public sector

(same as Question 1)

10. Describe any 4-design consideration of Data Warehousing (04 Marks)

Subject oriented

- To be effective, a DW should be designed around a subject domain, i.e., to help solve a certain category of problems.

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.

Time-variant (time series)

- The data in a DW should grow at daily or other chosen intervals. This allows latest comparisons over time.

Nonvolatile

- DW should be persistent, that is, it should not be created on the fly from the operations databases.
- Thus, a DW is consistently available for analysis, across the organization and over time.

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 reducing the number of variables or dimensions of the data to make it more meaningful for the decision makers.

11. Discuss the key steps in Data Mining Process, with neat diagram (08 Marks)

(same as question 6)

12. Describe objectives of graphical excellence in data visualization (08 Marks)

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.

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.

Avoid Distorting What the Data Have to Say

Statistics can be used to hide the truth. In the name of simplifying, some crucial context could be removed leading to distorted communication.

Make Large Datasets Coherent

By giving shape to data, visualizations can help bring the data together to tell a comprehensive story.

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.

Reveal the Data at Several Levels of Detail

Graphs lead to insights, which raise further curiosity, and thus presentations help get to the root cause.

Serve a Reasonably Clear Purpose Informing or decision-making.

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

13. **Explain the star schema of Data warehousing with an example (06 Marks)**

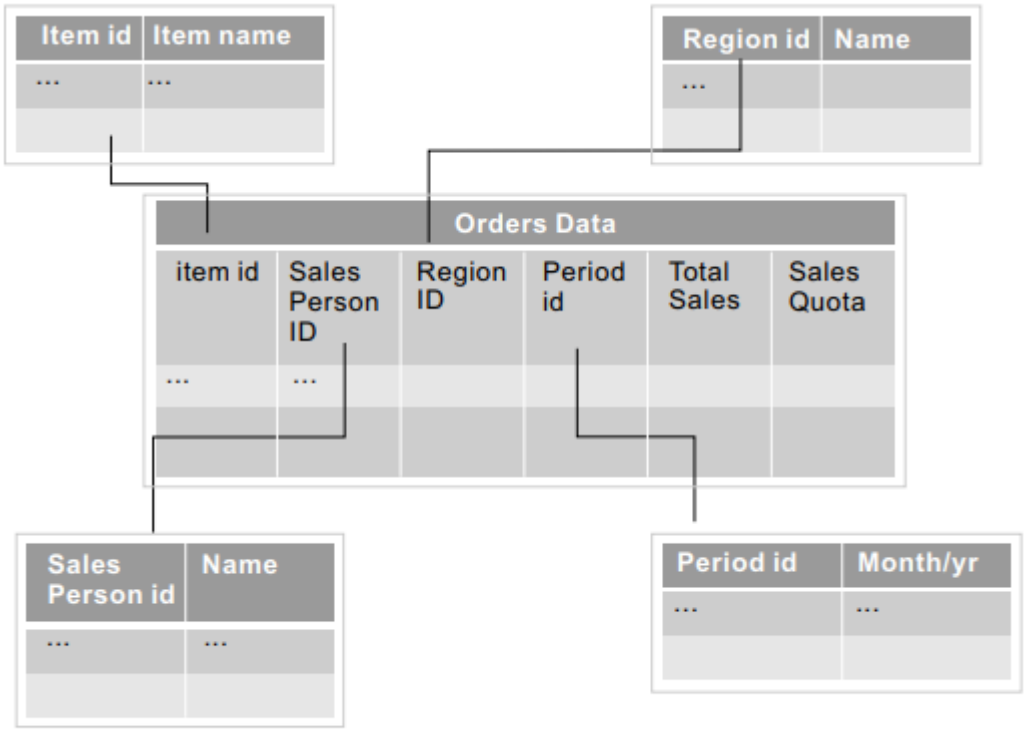


FIGURE 3.2 Star Schema Architecture for DW

- Star schema is the preferred data architecture for most DWs.
- There is a central fact table that provides most of the information of interest.
- There are lookup tables that provide detailed values for codes used in the central table.

- For example, the central table may use digits to represent a sales person. The lookup table will help provide the name for that sales person code.
- The difference between a star and snowflake is that in the snowflake, the lookup tables can have their own further lookup tables.

14. What is confusion matrix (02 Marks)

Confusion Matrix		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	False Positive (FP)

FIGURE 4.1 Confusion Matrix

When a true positive data point is positive, that is a correct prediction, called a true positive (TP). When a true negative data point is classified as negative, that is a true negative (TN). When a true-positive data point is classified by the model as negative, that is an incorrect prediction, called a false negative (FN). When a true-negative data point is classified as positive, that is classified as a false positive (FP).
This is called the confusion matrix.

15. What do you understand by the term data visualization? How is it important in Big Data Analytics? (05 Marks)

- 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 consumers' needs, nature of 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.

16. Differentiate between Data Mining and Data Warehousing. (03 Marks) +1

Data Warehousing	Data Mining
A data warehouse is database system which is designed for analytical analysis instead of transactional work.	Data mining is the process of analyzing data patterns.
Data is stored periodically.	Data is analyzed regularly.
Data warehousing is the process of extracting and storing data to allow easier reporting.	Data mining is the use of pattern recognition logic to identify patterns
Data warehousing is solely carried out by engineers.	Data mining is carried by business users with the help of engineers.
Data warehousing is the process of pooling all relevant data together.	Data mining is considered as a process of extracting data from large data sets

17. Draw the flow of BIDM cycle. Explain the strategic and operational decisions

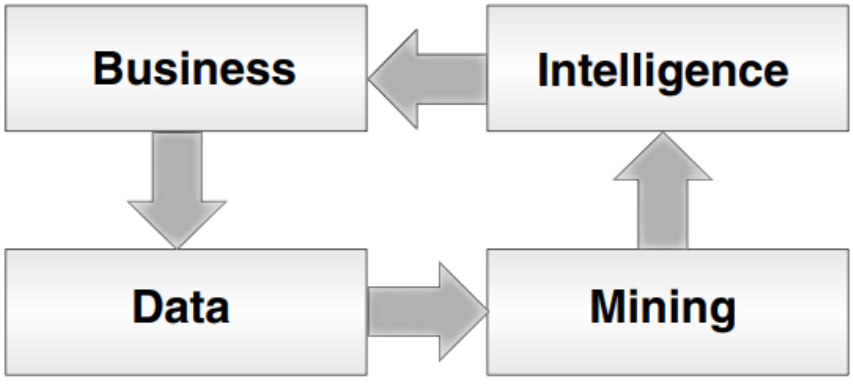


FIGURE 1.1 Business Intelligence and Data Mining (BIDM) Cycle

- Strategic decisions are those that impact the direction of the company.
- The decision to reach out to a new customer set would be a strategic decision.

- In strategic decision-making, the goal itself may or may not be clear, and the same is true for the path to reach the goal.
- The consequences of the decision would be apparent some time later.
- Thus, one is constantly scanning for new possibilities and new paths to achieve the goals.
- BI can help with what-if analysis of many possible scenarios.
- BI can also help create new ideas based on new patterns found from data mining.

- Operational decisions are more routine and tactical decisions, focused on developing greater efficiency.
- Updating an old website with new features will be an operational decision.
- Operational decisions can be made more efficient using an analysis of past data.
- A classification system can be created and modeled using the data of past instances to develop a good model of the domain.
- This model can help improve operational decisions in the future.
- BI can help automate operation level decision-making and improve efficiency by making millions of microlevel operational decisions in a model-driven way.

18. Describe any 8 considerations for a data warehouse and explain the key elements with a diagrammatic representation

Considerations

(Question 10)

Not normalized

DW often uses a star schema, which is a rectangular central table, surrounded by some lookup tables. The single table view significantly enhances speed of queries.

Metadata

Many of the variables in the database are computed from other variables in the operational database. Every element in the DW should be sufficiently well-defined.

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 a 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.

Key elements (Question 2)

BIG DATA ANALYSIS

MODULE – 4

1. **What is a splitting variable? Describe the criteria for choosing a splitting variable. (04 Marks)**
    - At every node, a set of possible split points is identified for every predictor variable.
    - The algorithm calculates the improvement in purity of the data that would be created by each split point of each variable.
    - The split with the greatest improvement is chosen to partition the data and create child nodes. The variable which is used to split is called as splitting variable.

    Splitting criteria

    - Algorithms use different measures like least errors, information gain, Gini's coefficient etc., to compute the splitting variable that provides the most benefit.
    - Information gain is a mathematical construct to compute the reduction in information entropy from a prior state to the next state that takes some information as given.
    - The greater the reduction in entropy, the better it is.
    - The Gini coefficient is a statistical concept that measures the inequality among values of a frequency distribution.
    - The lower the Gini's coefficient, the better it is.
2. **List some of the advantages and disadvantages of Regression model (04 Marks) +1**

    Advantages
    - Regression models are easy to understand as they are built upon basic statistical principles such as correlation and least square error.
    - Regression models provide simple algebraic equations that are easy to understand and use.
    - Regression models can match and beat the predictive power of other modeling techniques.
    - Regression models can include all the variables that one wants to include in the model.

Disadvantages

- Regression models cannot cover for poor data quality issues.
- Regression models do not automatically take care of nonlinearity
- Regression models can be unwieldy and unreliable if a large number of variables are included in the model.
- Regression models suffer from collinearity problems.

3. Create a decision tree for the following data set (08 Marks)

Age	Job	House	Credit	Loan Approved
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

Age	Job	House	Credit	Loan Approved
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	Loan Approved
Young	False	False	Good	???

4. Explain the design principles of an Artificial Neural Network (08 Marks) +1

1.

- A neuron is the basic processing unit of the network.
- It receives inputs from its previous neurons.
- After receiving input, it does nonlinear weighted computation and forms the output. This output is given as input to the next neuron.
- x is input, w is weights and y is output.

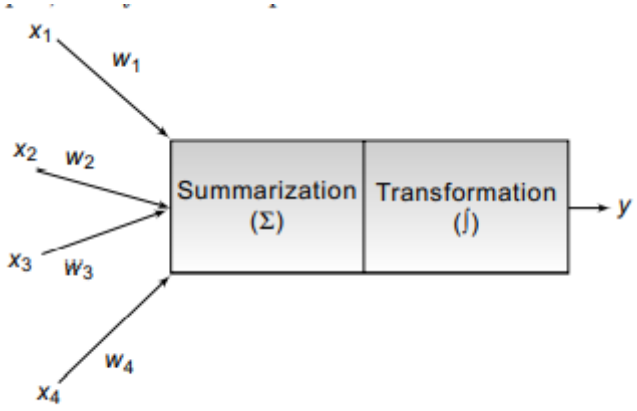


FIGURE 8.2 Model for a Single Artificial Neuron

2.

- A neural network is a multilayer model.
- There will be at least one input neuron, one output neuron and one processing neuron. This is a single-stage computational unit. This is used for simple tasks
- ANN have multi layers of processing elements in sequence
- Layers of neurons can work in sequence or parallel

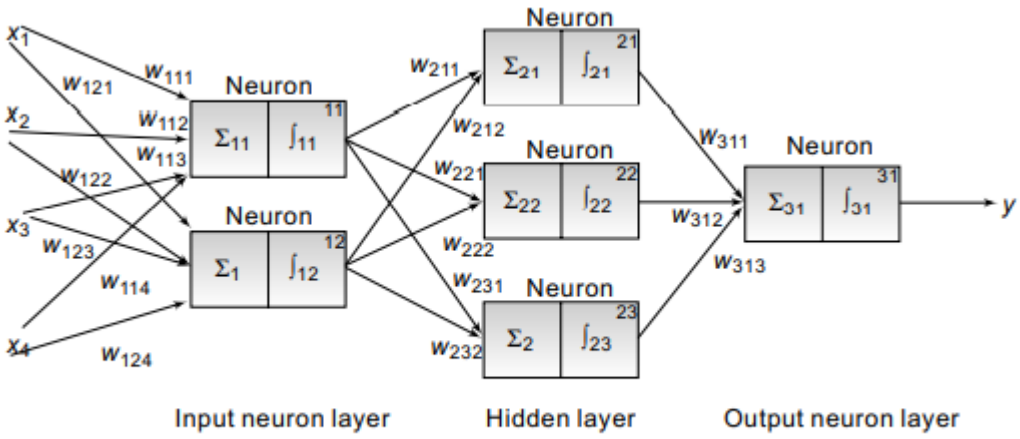


FIGURE 8.3 Model for a Multilayer ANN

3.

- The processing logic of each neuron can assign different weights to input streams
- The processing logic, intermediate weight and processing function works for the whole system for solving a problem collectively
- Neural networks are considered to be opaque and a black-box system

4.

- The neural network can be trained by making similar decisions again and again 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 making a decision as they handle more and more decisions

5. **How does the Apriori Algorithm work? Apply the same for the following example. (08 Marks)**

T_{ID}	List of Item-IDs
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

Assume the support count = 2.

6. Explain the different steps for constructing the decision tree for the following example (08 Marks) +2

OUTLOOK	TEMP	HUMIDITY	WINDY	PLAY
SUNNY	HOT	HIGH	FALSE	NO
SUNNY	HOT	HIGH	TRUE	NO
OVERCAST	HOT	HIGH	FALSE	YES
RAINY	MILD	HIGH	FALSE	YES
RAINY	COOL	NORMAL	FALSE	YES
RAINY	COOL	NORMAL	TRUE	NO
OVERCAST	COOL	NORMAL	TRUE	YES
SUNNY	MILD	HIGH	FALSE	NO
SUNNY	COOL	NORMAL	FALSE	YES
RAINY	MILD	NORMAL	FALSE	YES
SUNNY	MILD	NORMAL	TRUE	YES
OVERCAST	MILD	HIGH	TRUE	YES
OVERCAST	HOT	NORMAL	FALSE	YES
RAINY	MILD	HIGH	TRUE	NO

Table 7(a)

1 of 2

The decision problem is

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	Normal	True	? ?

7. Write the steps involved in developing an artificial neural network (05 Marks)

The steps required to build an ANN are as follows

1. Gather data and divide into training data and test data. The training data needs to be further divided into training data and validation data.
2. Select the network architecture, such as Feedforward network.
3. Select the algorithm, such as Multi-Layer Perception.
4. Set network parameters.
5. Train the ANN with training data.
6. Validate the model with validation data.
7. Freeze the weights and other parameters.
8. Test the trained network with test data.
9. Deploy the ANN when it achieves good predictive accuracy.

8. Describe the advantages of using ANN

The advantages of using ANN are:

■ ANNs impose very little restrictions on their use. ANN can deal with highly nonlinear relationships on their own, without much work from the user or analyst.

They help find practical data-driven solutions where algorithmic solutions are nonexistent or are too complicated.

■ There is no need to program neural networks as they learn from examples. They get better with use, without much programming effort.

■ They can handle a variety of problem types, including classification, clustering, associations, etc.

■ ANNs are tolerant of data quality issues and they do not restrict the data to follow strict normality and/or independence assumptions.

■ They can handle both numerical and categorical variables.

■ ANNs can be much faster than other techniques.

■ They usually provide better results (prediction and/or clustering) compared to statistical counterparts, once they have been trained enough.

(extra)

The key disadvantages arise from the fact that they are not easy to interpret or explain or compute.

■ They are deemed to be black-box solutions, lacking explainability. Thus they are difficult to communicate about, except through the strength of their results.

■ Optimal design of ANN is still an art. It requires expertise and extensive experimentation.

■ It could be difficult to handle a large number of variables (especially the rich nominal attributes).

■ It takes large datasets to train an ANN.

9. **For the following example describe the steps of forming association rules using Apriori algorithm +2**

S.No.	TRANSACTION LIST			
1	MILK	EGG	BREAD	BUTTER
2	MILK	BUTTER	EGG	KETCHUP
3	BREAD	BUTTER	KETCHUP	
4	MILK	BREAD	BUTTER	
5	BREAD	BUTTER	COOKIES	
6	MILK	BREAD	BUTTER	COOKIES
7	MILK	COOKIES		
8	MILK	BREAD	BUTTER	
9	BREAD	BUTTER	EGG	COOKIES
10	MILK	BUTTER	BREAD	
11	MILK	BREAD	BUTTER	
12	MILK	BREAD	COOKIES	KETCHUP

10. **List the steps required to build Artificial Neural Networks +1**

(same as Question 7)

11. Explain K-means algorithm with steps

- K-means is the most popular clustering algorithm.
- It iteratively computes the clusters and their centroids.
- It is a top-down approach to clustering.
- Starting with a given number of K clusters example K=3, that means 3 random centroids will be created as starting points of the centers

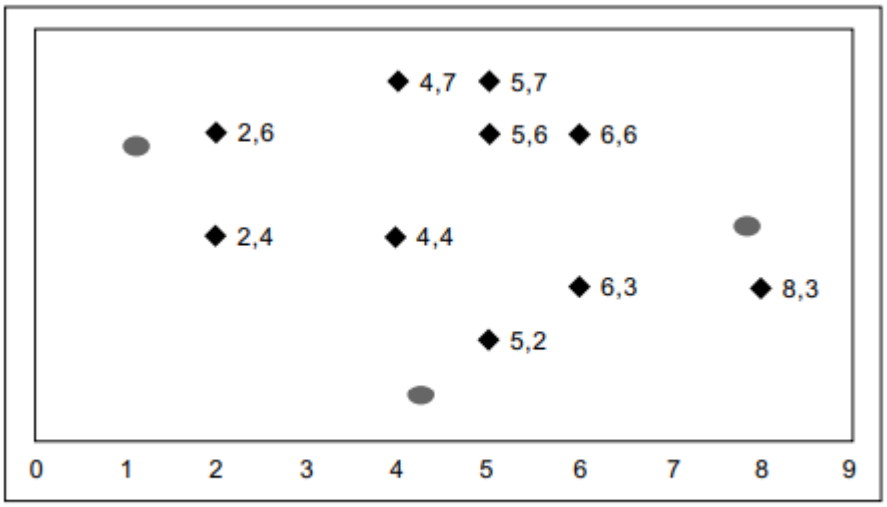


FIGURE 9.5 Randomly Assigning Three Centroids for Three Data Clusters

Step 1:

- For a data point, distance values will be from each of the three centroids.
- The data point will be assigned to the cluster with the shortest distance to the centroid.
- All data points will be assigned to one data point or the other.
- The arrows from each data element show the centroid that the point is assigned to.

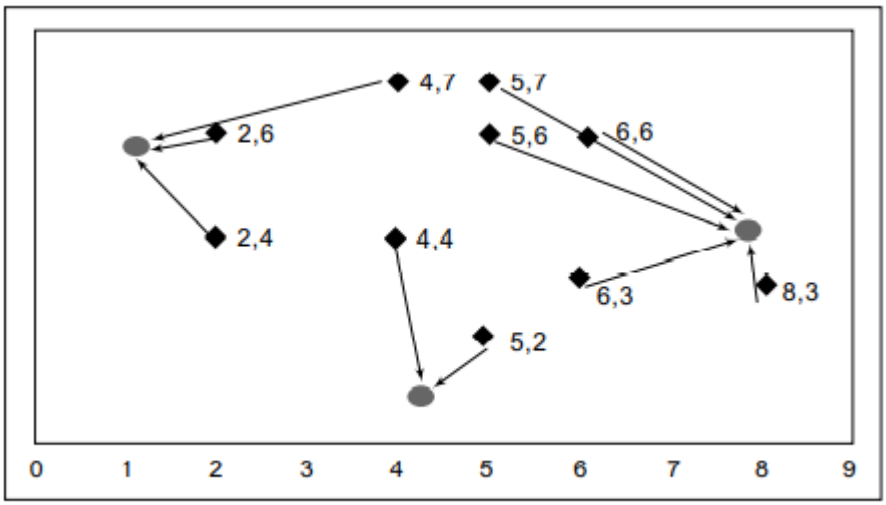


FIGURE 9.6 Assigning Data Points to Closest Centroid

Step 2:

- The centroid for each cluster will now be recalculated such that it is closest to all the data points allocated to that cluster.
- The dashed arrows show the centroids being moved from their old (shaded) values to the revised new values

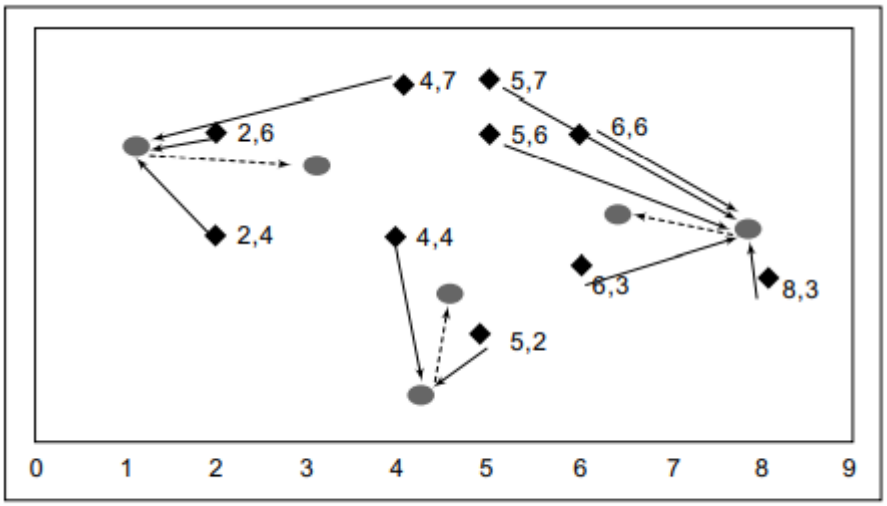


FIGURE 9.7 Recomputing Centroids for Each Cluster

Step 3:

- Once again, data points are assigned to the three centroids closest to it.
- The new centroids will be computed from the data points in the cluster until finally, the centroids stabilize in their locations.
- These are the three clusters computed by this algorithm.

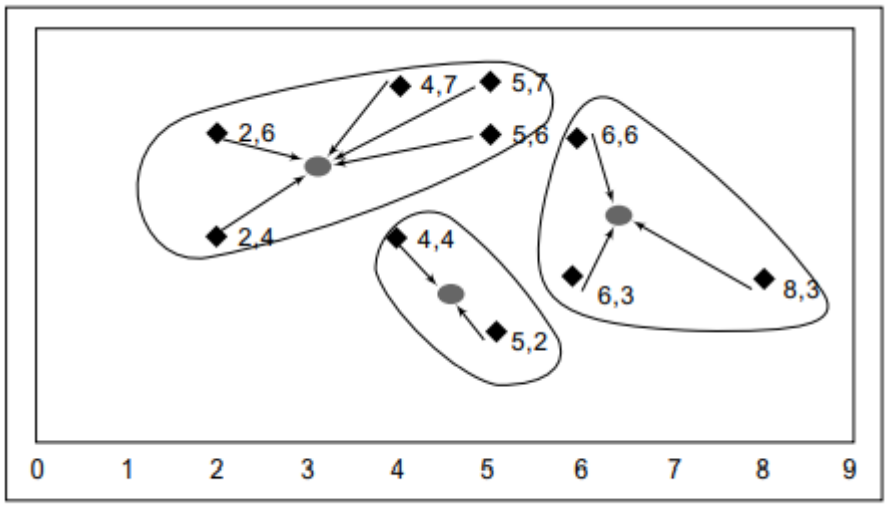


FIGURE 9.9 Recomputing Centroids for Each Cluster till Clusters Stabilize

12. Differentiate between C4.5, CART and CHAID decision tree algorithms

DecisionTree	C4.5	CART	CHAID
Full name	Iterative Dichotomizer (ID3)	Classification and Regression Trees	Chi-square Automatic Interaction Detector
Basic algorithm	Hunt's algorithm	Hunt's algorithm	Adjusted significance testing
Developer	Ross Quinlan	Bremman	Gordon Kass
When developed	1986	1984	1980
Type of trees	Classification	Classification and Regression tree	Classification and Regression tree
Serial implementation	Tree growth and Tree pruning	Tree growth and Tree pruning	Tree growth and Tree pruning
Type of data	Discrete and Continuous; Incomplete data	Discrete and Continuous	Non-normal data also accepted
Type of splits	Multi-way splits	Binary splits only; clever surrogate splits to reduce tree depth	Multiway splits as default
Splitting criteria	Information gain	Gini's coefficient, and others	Chi-square test
Pruning criteria	Clever bottom-up technique avoids over-fitting	Remove weakest links first	Trees can become very large
Implementation	Publicly available	Publicly available in most packages	Popular in market research for segmentation

MODULE - 4

Q3. Create a decision tree for the following dataset

Age	Job	House	Credit	Loan Approved
Young	False	No	Fair	No
Young	False	No	Good	No
Young	True	No	Good	Yes
Young	T	Yes	Fair	Yes
Young	F	No	Fair	No
Middle	F	No	Fair	No
Middle	F	No	Good	No
Middle	T	Yes	Good	Yes
Middle	F	Yes	Excellent	Yes
Middle	F	Yes	Excellent	Yes
Old	F	Yes	Excellent	Yes
Old	F	Yes	Good	Yes
Old	T	No	Good	Yes
Old	T	No	Excellent	Yes
Old	F	No	Fair	No

Then solve the following problem using the model

Age	Job	House	Credit	Loan Approved
Young	False	False	Good	???

Solution

① Attribute Age

Young — No - 3, Yes - 2 ; middle — No - 2, Yes - 3 ; old — No - 1, Yes - 4

Attribute	Rules	Error	Total Error
Age	Young - No	2/5	5/15
	Middle - Yes	2/5	
	old - Yes	1/5	

② Attribute Job

False — Yes - 4, No - 6 ; True — Yes - 5, No - 0

Attribute	Rules	Error	Total Error
Job	False - No	4/10	4/15
	True - Yes	0/5	

(3) Attribute : House

House No — Yes - 3
NO - 6

House Yes — Yes - 6
NO - 0

Attribute	Rules	Error	Total Error
House	NO - NO	3/9	3/15
	Yes - Yes	0/6	

(4) Attribute - Credit

Fair — Yes - 1
NO - 4

Good — Yes - 4
NO - 2

Excellent — Yes - 4
NO - 0

Attribute	Rules	Error	Total Error
Credit	Fair - No	1/5	3/15
	Good - Yes	2/6	
	Excellent - Yes	0/4	

least Total Error is seen in House and Credit, So we should compare 0 errors. In House it is 0/6 and in credit it is 0/4. Hence House is having least credit

Decision Tree will become.

Yes

NO

Loan=Yes (6)

House	Age	Job	Credit	loan
Yes	Y	T	F	Yes
	M	T	G	
	M	F	E	
	M	F	E	
	O	F	E	
	O	F	G	

House = No

loan=Yes (3)

loan = No (6)

House	Age	Job	Credit	loan
No	Y	F	F	N
	Y	F	G	N
	Y	T	G	Y
	Y	F	F	N
	M	F	F	N
	M	F	G	N
	O	T	G	Y
	O	T	E	Y
	O	F	F	N

House - NO — Yes - 3/9
NO - 6/9

Attribute	Rules	Error	Total Error
House	No → No	3/9	3/9

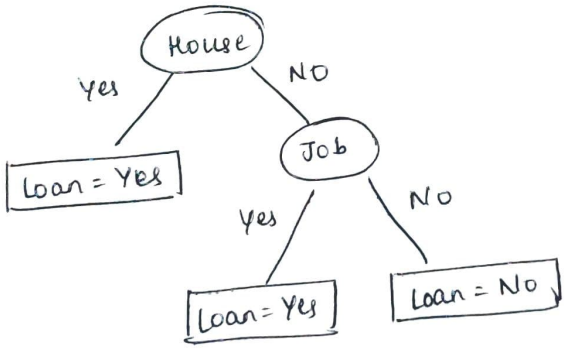
Attribute	Rules	Error	Total Error
Job	False → No	0/6	0/9
	True → Yes	0/3	
Credit	Fair → No	0/4	2/9
	Good → Yes	2/4	
	Excellent → Yes	0/1	
Age	Young → No	1/4	2/9
	Middle → No	0/2	
	Old → Yes	1/3	

Job — False — Yes = 0, No = 6
- True — Yes = 3, No = 0

Credit — Fair — Y = 0, N = 4
Good — Y = 2, N = 2
Excellent — Y = 1, N = 0

Age — Y — Y = 1, N = 3
M — Y = 0, N = 2
O — Y = 2, N = 1

Job has the least error. So, Right side of decision tree is split as



Hence the given test case

Age	Job	House	Credit	Loan Approved
Young	false	False	Good	???

Loan Approved = NO //.

Steps for decision Tree

#1 :- Take each attribute separately and list the classification in it [Ex:- Age has Young, Middle, Old]

#2 :- For each classification, see the outcome of it,
[Ex:- Age - Young $\begin{cases} \text{Yes} = 2 \\ \text{No} = 3 \end{cases}$]

#3 :- Take the classification which has highest number
[Ex:- Age - Young - No]

#4 :- calculate error for the classification
[Ex:- Age - Young - No = $\frac{3}{5}$ ∴ Error is $\frac{2}{5}$]

#5 :- Repeat for all classifications and find total error
[Ex:- Error of Age - Young-No = 2/5, Middle - Yes = $\frac{2}{5}$; Old - Yes = $\frac{1}{5}$

Total error = $\frac{2+2+1}{5+5+5} = \frac{5}{15}$]

#6 :- Repeat above process for all attributes.

#7 :- Take the attribute which has the least total error
If there are many attributes with least error, look at zero error which has highest denominator
Ex:- House has 0/6 and credit has 0/4 so House has least error.

#8 :- Fix the attribute as splitting variable and continue for other attributes.

#9 :- Repeat the step till both sides of decision tree have zero error.

#10 :- Classify the test case by seeing the splitting in decision tree.

Decision Tree problem

Q6.

Attribute - Outlook

Sunny - Yes = 2, No = 3 ; Overcast - Yes = 4, No = 0 ; Rainy - Yes = 3, No = 2

Attribute	Rules	Error	Total Error
Outlook	Sunny = No	2/5	4/14
	Overcast = Yes	0/4	
	Rainy = Yes	2/5	

Attribute = Temperature

Hot - Y = 2, N = 2

Cool - Y = 3, N = 1

Mild - Y = 4, N = 2

Attribute	Rules	Error	Total Error
Temperature	Hot - No	2/4	5/14
	Cool - Yes	1/4	
	Mild - Yes	2/6	

Attribute = Humidity

High - Y = 3, N = 4

Normal - Y = 6, N = 1

Attribute	Rules	Error	Total Error
Humidity	High - No	3/7	4/14
	Normal - Yes	1/7	

Attribute = Windy

False - Y = 6, N = 2

True - Y = 3, N = 3

Attribute	Rules	Error	Total Error
Windy	False - Yes	2/8	5/14
	True - No	3/6	

Here outlook and Humidity have least total error, but outlook has 1 zero error. so outlook is chosen

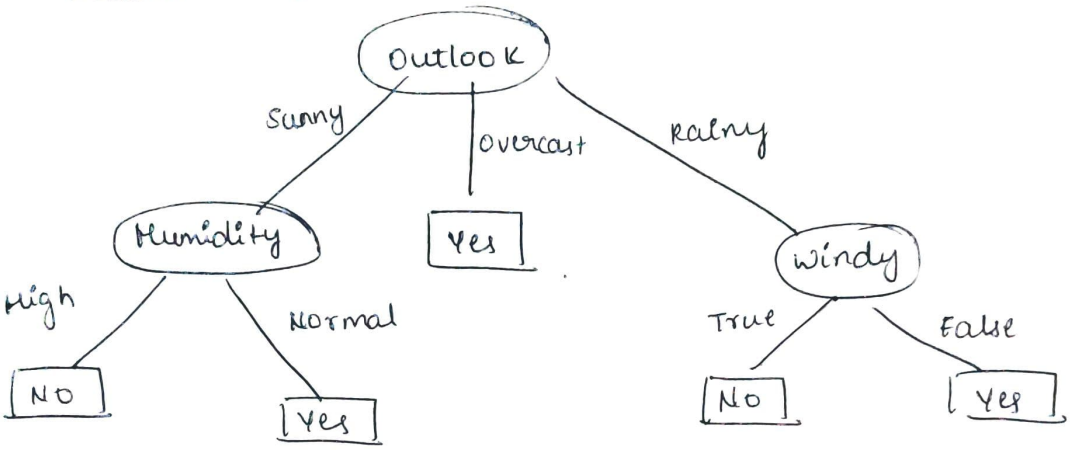
Outlook → Sunny

Attribute	Rules	Error	Total Error
Temperature	Hot → No	0/2	1/5
	Mild → No	1/2	
	cool → Yes	0/1	
Humidity	High - No	0/3	0/5
	Normal - Yes	0/2	
Windy	False - No	1/3	2/5
	True - Yes	1/2	

Outlook – Rainy

Attribute	Rules	Error	Total Error
Temperature	Mild → Yes	1/3	2/5
	Cool → Yes	1/2	
Humidity	High → No	1/2	2/5
	Normal → Yes	1/3	
Windy	False → Yes	0/3	0/5
	True → No	0/2	

Decision Tree



Hence,

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	Normal	True	~~No~~ Yes //

9.9 Given

Sl No	Transaction list				
1	Milk	Egg	Bread	Butter	
2	Milk	Butter	egg	Ketchup	
3	Bread	Butter	Ketchup		
4	Milk	Bread	Butter		
5	Bread	Butter	cookies		
6	Milk	Bread	Butter	cookies	
7	Milk	cookies			
8	Milk	Bread	Butter		
9	Bread	Butter	Egg	cookies	
10	Milk	Butter	Bread		
11	Milk	Bread	Butter		
12	Milk	Bread	cookies	Ketchup	

support level = 33%

confidence level = 50%

$\frac{33}{100} \times 12 = 3.96 \approx 4$

For 1 item sets

1-itemset	Frequency	
Milk	9	✓
Bread	10	✓
Butter	10	✓
Egg	3	
Ketchup	3	
cookies	5	✓

Frequent 1-item sets	Frequency
Milk	9
Bread	10
Butter	10
cookies	5

Now using frequent 1-item sets

2-item sets	Frequency	
Milk, Bread	7	✓
Milk, Butter	7	✓
Milk, cookies	3	
Bread, Butter	9	✓
Butter, cookies	3	
Bread, cookies	4	✓

Frequent 2-item sets	Frequency
Milk, Bread	7
Milk, Butter	7
Bread, Butter	9
Bread, cookies	4

Similarly

3-item sets	Frequency	
Milk, Bread, Butter	6	✓
Milk, Bread, cookies	1	
Bread, Butter, cookies	3	

Frequent 3-itemset	Frequency
Milk, Bread, Butter	6

Association Rules

Highest level itemset

Milk, Bread, Butter - 6

1) (Bread Butter) → Milk

Support level = $\frac{6}{12} = 50\%$

confidence level = $\frac{\text{Bread, Butter, Milk}}{\text{Bread, Butter}} = \frac{6}{9} = 67\%$

} valid

2) (Milk, Bread) → Butter

$S = \frac{6}{12} = 50\%$

$C = \frac{6}{7} = 86\%$

} valid.

3) (Milk, Butter) → Bread

$S = \frac{6}{12} = 50\%$

$C = \frac{6}{7} = 86\%$

} valid

4) Milk → Bread

$S = \frac{7}{12} = 58\%$

$C = \frac{7}{9} = 78\%$

} valid

Steps for Apriori Algorithm

#1 From given support level & total number of instances, calculate minimum number of instances

Ex:- 33% → $\frac{33}{100} \times 12 = 3.96 \approx 4$

#2 Calculate frequency for 1-itemset

#3 Choose items which have frequency greater or equal to minimum number of instances.

#4. Using the selected items, calculate frequency for 2-itemsets.

#5 Repeat the process till no frequency is greater or equal to minimum number of instances.

#6. To create association rule, choose the highest itemset

Ex:- Here we have 3-itemset - Milk, Bread, Butter = 6

#7 ~~Calculate~~ Take combination of the itemset

#8 Calculate support level for each combination

Ex:- (Bread, Butter) → Milk

$$\text{Support level} = \frac{\text{Number of instances of (Bread, Butter, Milk)}}{\text{Total number of instances}} = \frac{6}{12}$$

#9. Calculate confidence level for each combination

Ex:- (Bread, Butter) → Milk

$$\text{Confidence level} = \frac{\text{Number of instances of (Bread, Butter, Milk)}}{\text{Number of instances of (Bread, Butter)}} = \frac{6}{9}$$

#10. If both values are greater than given value, it is valid.

#11. Repeat for all combinations.

Q. 5. Given

T_{ID}	List of items
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

No of unique items = 5

Total no of transactions = 9

Support count = 2

Support level = $\frac{2}{9}$ = 22.22%

1) 1-itemset

1-itemset	Frequency
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

⇒

Frequent 1-itemset	Frequency
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

2)

2-itemset	Frequency
I_1, I_2	4 ✓
I_1, I_3	4 ✓
I_1, I_4	1
I_1, I_5	2 ✓
I_2, I_3	4 ✓
I_2, I_4	2 ✓
I_2, I_5	2 ✓
I_3, I_4	0
I_3, I_5	1
I_4, I_5	0

⇒

Frequent 2-itemset	Frequency
I_1, I_2	4
I_1, I_3	4
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2

3)

3-itemset	Frequency
I_1, I_2, I_3	2 ✓
I_1, I_2, I_5	2 ✓
I_1, I_3, I_5	1
I_2, I_3, I_5	0
I_2, I_3, I_4	1
I_2, I_4, I_5	0

⇒

Frequent 3-itemset	Frequency
I_1, I_2, I_3	2
I_1, I_2, I_5	2

Rule generation

1) $[I_1] \rightarrow [I_2\ I_3]$

$$S = \frac{2}{9} = 22.22\%$$

$$C = \frac{2}{6} = 33.33\%$$

2) $[I_1] \rightarrow [I_2\ I_5]$

$$S = \frac{2}{9} = 22.22\%$$

$$C = \frac{2}{6} = 33.33\%$$

BIG DATA ANALYSIS

MODULE – 5

1. **What is Naïve Bayes technique? Explain its model (05 Marks)**

   - Naïve-Bayes (NB) technique is a supervised learning technique that uses probability-theory-based analysis.
   - It is a machine-learning technique that computes the probabilities of an instance belonging to each one of many target classes, given the prior probabilities of classification using individual factors.

   Naïve Bayes Model

   - Naïve-Bayes is a conditional probability model for classification purposes.
   - The goal is to find a way to predict the class variable (Y) using a vector of independent variables (X) - finding the function f: X →Y.
   - In probability terms, the goal is to find P(Y|X), i.e., the probability of Y belonging to a certain class X.
   - Y is generally assumed to be a categorical variable with two or more discrete values.
   - Given an instance to be classified, represented by a vector x = (x1, ..., xn) it represents 'n' features (independent variables), the Naïve-Bayes model assigns probabilities of belonging to any of the K classes. The class K with the highest posterior probability is the label assigned to the instance.
   - The posterior probability (of a Class K) is calculated as a function of prior probabilities and current likelihood value, as shown in the equation below

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

$P(C_k|\boldsymbol{x})$ is the posterior probability of class K, given predictor X.
$P(C_k)$ is the prior probability of class K.
$P(\boldsymbol{x})$ is the prior probability of predictor.
$P(\boldsymbol{x}|C_k)$ is the current likelihood of predictor given class.

2. **What is Support Vector Machine? Explain its model (08 Marks) +2**

   - Support Vector Machine (SVM) is a mathematically rigorous, machine learning technique to build a linear binary classifier.
   - It creates a hyperplane in a high-dimensional space that can accurately slice a dataset into two segments according to the desired objective.

   SVM Model

- Suppose there is a labeled set of points classified into two classes. The goal is to find the best classifier between the points of the two types.

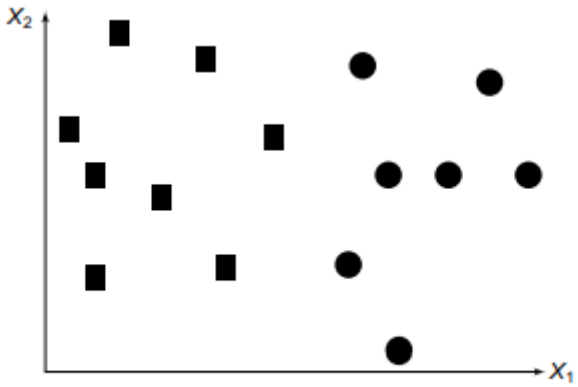


FIGURE 13.1 Data Points for Classification

- SVM takes the widest street (a vector) approach to define the two classes and thus finds the hyperplane that has the widest margin, i.e., largest distance to the nearest training data points of either class.

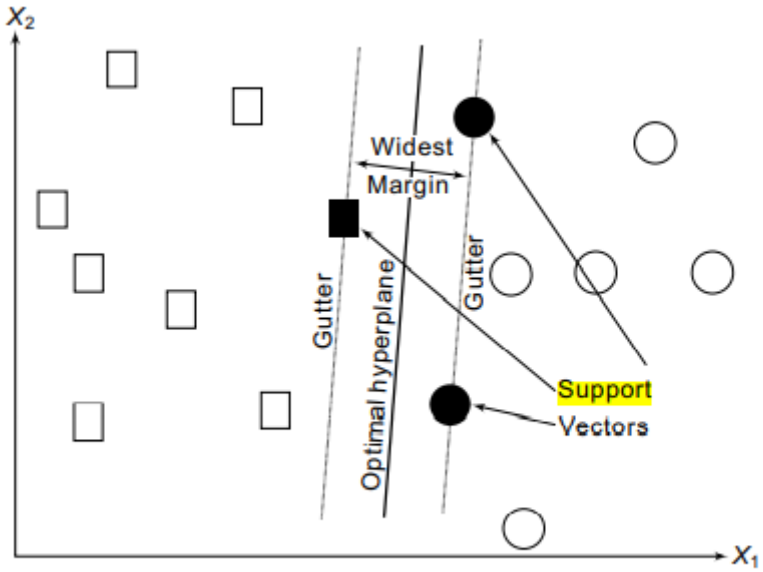


FIGURE 13.2 Support Vector Machine Classifier

- In the above figure, the hard line is the optimal hyperplane. The dotted lines are the gutters on the sides of the two classes. The gap between the gutters is the maximum or widest margin.
- The classifier (hyperplane) is defined by only those points that fall on the gutters on both sides. These points are called the support vectors (shown in their bold). The rest of the data points in their class are irrelevant for defining the classifier

- Suppose that the training data of n points is (X1, y1), ….., (Xi, yi). there are two classes represented as 1 and –1.
- Assuming that the data is indeed linearly separable, the classifier hyperplane is defined as a set of points that satisfy the equation:
  $\boldsymbol{W} . \boldsymbol{X} + b = 0$, where $\boldsymbol{W}$ is the normal vector to the hyperplane.
- The hard margins can be defined by the following hyperplanes
  $\boldsymbol{W} . \boldsymbol{X} + b = 1$ and $\boldsymbol{W} . \boldsymbol{X} + b = -1$
  The width of the hard margin is $(2/|\boldsymbol{W}|)$
- For all points not on the hyperplane, they will be safely in their own class. Thus, the y values will have to be either greater than 1 (for point in class 1) or less than –1 (for points in class –1).
- The SVM algorithm finds the weights vector ($\boldsymbol{W}$) for the features, such that there is a widest margin between the two categories.

3. Mention the 3-step process of Text Mining (03 Marks)

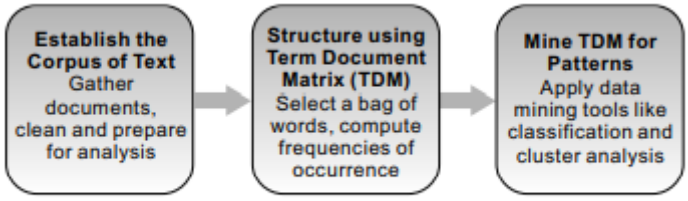


FIGURE 11.1 Text Mining Architecture

Text mining is a semi-automated process. Text data needs to be gathered, structured, and then mined, in a 3-step process:

1. The text and documents are first gathered into a corpus and organized.

2. The corpus is then analyzed for structure. The result is a matrix mapping important terms to source documents.

3. The structured data is then analyzed for word structures, sequences, and frequency.

4. Explain briefly the three different types of web mining (06 Marks) +2

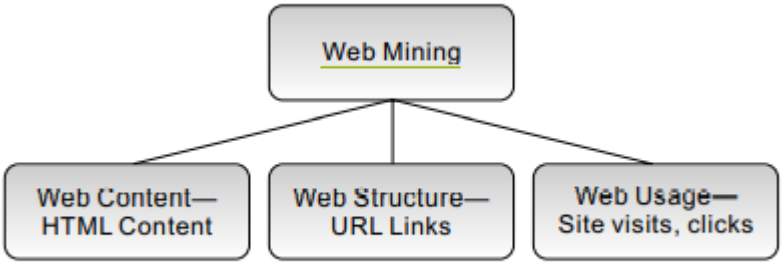


FIGURE 14.1 Web Mining Structure

1. WEB CONTENT MINING

- A website is designed in the form of pages with a distinct URL
- These pages are managed using specialized software systems called Content Management Systems.
- The websites keep a record of all requests received for its page/URLs, including the requester information using 'cookies'.
- The log of these requests could be analyzed to gauge the popularity of those pages among different segments of the population.
- The pages on a website themselves could be analyzed for quality of content that attracts most users.
- Thus, the unwanted or unpopular pages could be weeded out or they can be transformed with different content and style.

2. 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
    - Authorities

3. WEB USAGE MINING

- The goal of web usage mining is to extract useful information and patterns from data generated through web page visits and transactions.
- The activity data comes from data stored in server access logs, referrer logs, agent logs, and client-side cookies.
- The user characteristics and usage profiles are also gathered directly or indirectly through syndicated data.
- Further, metadata such as page attributes, content attributes, and usage data are also gathered.

5. Compute the rank values for the nodes for the following network shown in fig. which is the highest rank node. Solve the same with eight iterations (10 Marks) +2

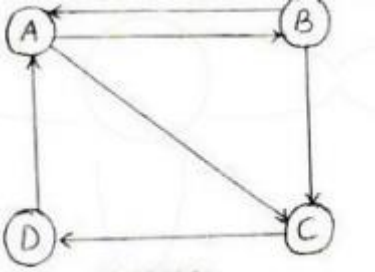


Fig.Q10(b)

6. **Describe the difference between text mining and data mining (06 Marks) +1**

Table 11.2 Comparing Text Mining and Data Mining

Dimension	Text Mining	Data Mining
Nature of Data	Unstructured data: words, phrases, sentences	Numbers, alphabetical and logical values
Language Used	Many languages and dialects used in the world; many languages are extinct, new documents are discovered	Similar numerical systems across the world
Clarity and Precision	Sentences can be ambiguous; sentiment may contradict the words	Numbers are precise
Consistency	Different parts of the text can contradict each other	Different parts of data can be inconsistent, thus requiring statistical significance analysis
Sentiment	Text may present a clear and consistent or mixed sentiment, across a continuum. Spoken words add further sentiment	Not applicable
Quality	Spelling errors. Differing values of proper nouns such as names. Varying quality of language translation	Issues with missing values, outliers, etc.
Nature of Analysis	Keyword based search; co-existence of themes; sentiment mining;	A full wide range of statistical and machine learning analysis for relationships and differences

7. **Explain Naïve bayes model to classify the text data into right class using following dataset (06 Marks)**

Training Set	Document ID	Keyword in the document	Class = h (Healthy)
	1	Love Happy Joy Joy Love	Yes
	2	Happy Love KICK JOY Happy	Yes
	3	Love Move Joy Good	Yes
	4	Love Happy Joy Pain Love	Yes
	5	Joy Love Pain Kick pain	No
	6	Pain Pain Love Kick	No
Test data	7	Love Pain Joy Love Kick	?

Table Q9 (c)

8. **Discuss the application and practical considerations of social network analysis (08 marks)**

Applications of SNA

Self-awareness

- Visualizing his/her social network can help a person organize their relationships and support network.

Communities

- Social Network Analysis can help identification, construction, and strengthening of networks within communities to build wellness, comfort and resilience.
- Analysis of joint authoring relationships and citations help identify subnetworks of specializations of knowledge in an academic field.

Marketing

- There is a popular network insight that any two people are related to each other through at most seven degrees of links.
- Organizations can use this insight to reach out with their message to large number of people and also to listen actively to opinion leaders as ways to understand their customers' needs and behaviors.

Public Health

- Awareness of networks can help identify the paths that certain diseases take to spread.
- Public health professionals can isolate and contain diseases before they expand to other networks.

PRACTICAL CONSIDERATIONS

Network Size

- Most SNA research is done using small networks.
- Collecting data about large networks can be very challenging. This is because the number of links is the order of the square of the number of nodes.
- Thus, in a network of 1000 nodes there are potentially 1 million possible pairs of links.

Gathering Data

- Electronic communication records (emails, chats, etc.) can be harnessed to gather social network data more easily.
- Data on the nature and quality of relationships need to be collected using survey documents.
- Capturing and cleansing and organizing the data can take a lot of time and effort, just like in a typical data analytics project.

Computation and Visualization

- Modeling large networks can be computationally challenging and visualizing them also would require special skills.
- Big data analytical tools may be needed to compute large networks.

Dynamic Networks

- Relationships between nodes in a social network can be fluid.
- They can change in strength and functional nature.
- The network should be modeled frequently to see the dynamics of the network

9. **Explain Text mining architecture and term document matrix (08 Marks) +1**

(same as question 3 with additional points)

TERM DOCUMENT MATRIX

Table 11.1 Term-Document Matrix

Term Document Matrix					
Document/Terms	**Investment**	**Profit**	**Happy**	**Success**	**...**
Doc 1	10	4	3	4	
Doc 2	7	2	2		
Doc 3			2	6	
Doc 4	1	5	3		
Doc 5		6		2	
Doc 6	4		2		
...					

- TDM is the heart of the structuring process.
- Free flowing text can be transformed into numeric data in a TDM, which can then be mined using regular data mining techniques
- There are several efficient techniques for identifying key terms from a text. This approach measures the frequencies of select important terms occurring in each document.
- This creates a t x d Term-by-Document Matrix (TDM), where t is the number of terms and d is the number of documents
- Creating a TDM requires making choices of which terms to include.
- The terms chosen should reflect the stated purpose of the text mining exercise.
- The list of terms should be as extensive as needed, but should not include unnecessary stuff that will serve to confuse the analysis or slow the computation.

10. Explain Web Usage Mining architecture

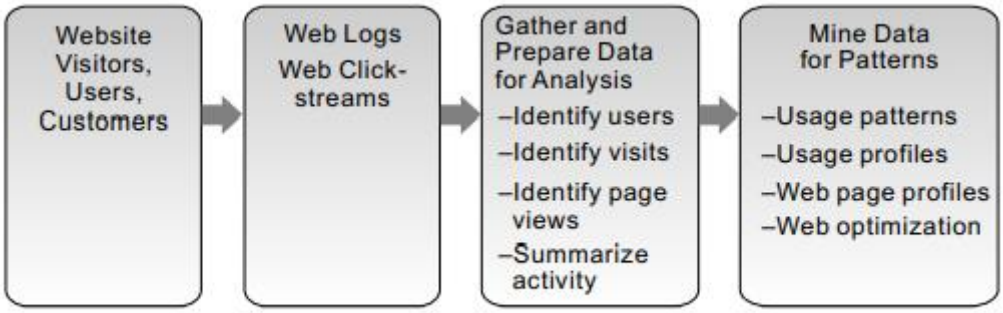


FIGURE 14.2 Web Usage Mining Architecture

- The web content could be analyzed at multiple levels.
- The server-side analysis would show the relative popularity of the web pages accessed. Those websites could be hubs and authorities.
- The client-side analysis would focus on the usage pattern or the actual content consumed and created by users.

  (a) Usage pattern could be analyzed using 'clickstream' analysis, i.e., analyzing web activity for patterns of sequence of clicks, and the location and duration of visits on websites.

  Clickstream analysis can be useful for web activity analysis, software testing, market research, and analyzing employee productivity.

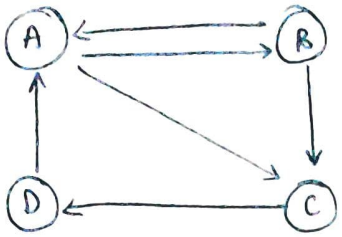
  (b) Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques

  The text would be gathered and structured using the bag-of-words technique to build a term-document matrix.

  This matrix could then be mined using cluster analysis and association rules for patterns such as popular topics, user segmentation, and sentiment analysis.

MODULE 5

Q. 5



Let R_A, R_B, R_C, R_D be the rank for each node

1. Node A gets all influence from D & half from B

$$R_A = 0.5 R_b + R_d$$

Similarly

$$R_b = 0.5 R_a$$

$$R_C = 0.5 R_a + 0.5 R_b$$

$$R_d = R_C$$

Equations are formed into influence matrix

	R_a	R_b	R_c	R_d
R_a	0	0.5	0	1
R_b	0.5	0	0	0
R_c	0.5	0.5	0	0
R_d	0	0	1	0

Initial rank value is assigned as $1/n = 1/4 = 0.25$

$\therefore$ $R_a = R_b = R_c = R_d = 0.25$

Iteration 1

Variable	Initial value	Equation	Equation value	Iteration 1
R_a	0.25	$0.5R_b + R_d$	$0.5 \times 0.25 + 0.25$	0.375
R_b	0.25	$0.5 R_a$	0.5×0.25	0.125
R_c	0.25	$0.5 R_a + 0.5 R_b$	$0.5 \times 0.25 + 0.5 \times 0.25$	0.250
R_d	0.25	R_c	0.25	0.250

Similarly continue for other iterations

variable	Inital value	Equation value	Iteration 2
R_a	0.375	$0.5 \times 0.125 + 0.25$	0.3125
R_b	0.125	0.5×0.375	0.1875
R_c	0.250	$0.5 \times 0.375 + 0.5 \times 0.125$	0.25
R_d	0.250	0.25	0.25

Similarly for 8 iterations

Variable	Initial value	Iteration 3
R_a	0.3125	0.3437
R_b	0.1875	0.1562
R_c	0.25	0.25
R_d	0.25	0.25

Variable	Initial value	Iteration 4
R_a	0.3437	0.328125
R_b	0.1562	0.171875
R_c	0.25	0.25
R_d	0.25	0.25

Variable	Initial value	Iteration 5
R_a	0.328125	0.336
R_b	0.171875	0.164
R_c	0.25	0.25
R_d	0.25	0.25

Variable	Initial value	Iteration 6
R_a	0.336	0.332
R_b	0.164	0.168
R_c	0.25	0.25
R_d	0.25	0.25

variable	Initial value	Iteration 7
R_a	0.332	0.334
R_b	0.168	0.166
R_c	0.25	0.25
R_d	0.25	0.25

Variable	Initial value	Iteration 8
R_a	0.334	0.333
R_b	0.166	0.167
R_c	0.25	0.25
R_d	0.25	0.25

The highest ~~node~~ rank is for A → Most important node

The lowest rank is for B → least important node

Q.7 Naive Bayes model

Given Training Set	Document ID	Keywords	Class = h (healthy)
	1	Love Happy Joy Joy love	Yes
	2	Happy love kick Joy Happy	Yes
	3	Love move Joy Good	Yes
	4	Love Happy Joy Pain love	Yes
	5	Joy love Pain kick Pain	No
	6	Pain Pain love kick	No
Testing Data	7	love Pain Joy love kick	?

Prior probability $P(h) = 4/6 = 2/3$ – Yes

$P(\sim h) = 2/6 = 1/3$ – No

Conditional probability for each term

Class h - Yes	Class ~h - No
$P(\text{Love}/h) = 5/19$	$P(\text{Love}/\sim h) = 2/9$
$P(\text{Pain}/h) = 1/19$	$P(\text{Pain}/\sim h) = 4/9$
$P(\text{Joy}/h) = 5/19$	$P(\text{Joy}/\sim h) = 1/9$
$P(\text{Kick}/h) = 1/19$	$P(\text{Kick}/\sim h) = 2/9$

Hence $P(h \mid d7) = P(h) * [P(\text{Love}/h)^2 * P(\text{Pain}/h) * P(\text{Joy}/h) * P(\text{Kick}/h)]$

$= \frac{2}{3} \times \left[\left(\frac{5}{19}\right)^2 * \left(\frac{1}{19}\right) * \left(\frac{5}{19}\right) * \left(\frac{1}{19}\right)\right] = 0.0000067$

$P(\sim h \mid d7) = P(\sim h) * [P(\text{Love}/\sim h)^2 * P(\text{Pain}/\sim h) * P(\text{Joy}/\sim h) * P(\text{Kick}/\sim h)]$

$= \frac{1}{3} \times \left[\left(\frac{2}{9}\right)^2 \times \left(\frac{4}{9}\right) * \left(\frac{1}{9}\right) * \left(\frac{2}{9}\right)\right] = 0.00018$

Hence since $P(\sim h \mid d7) > P(h \mid d7)$

The test document is 'not h'

Steps for Rank value

#1. Assign Ranks for each node [Ex:- $R_A, R_B \dots$]

#2 Get the equations by looking at diagram

#3 Calculate initial value for rank which is $1/n$

#4 Using initial value & equations, calculate value of rank for one iteration

#5 Repeat #4 for given number of iterations

#6 The highest rank is the most important node
The lowest rank is least important node.

Steps for Naive bayes

#1 Calculate prior probability for outcome

#2 ~~cond~~ Calculate conditional probability for each instance - both h & $\sim h$

Ex:- There are 19 words in Yes and out of those love is 5 $\therefore P(\text{love}|h) = 5/19$

#3 For given test case use Naive Bayes theorem to calculate for both h & $\sim h$

Ex:- Test case - Love Pain Joy Love kick

$$P(h|\text{test case}) = \underbrace{P(h)}_{\text{prior probability}} \times \left[P(\text{love}|h) * P(\text{pain}|h) * \underbrace{P(\text{joy}|h)}_{\text{conditional probability}} * P(\text{love}|h) * P(\text{kick}|h) \right]$$

#4. The greater value will be the outcome of the test case.