USN □□□□□□□□□□

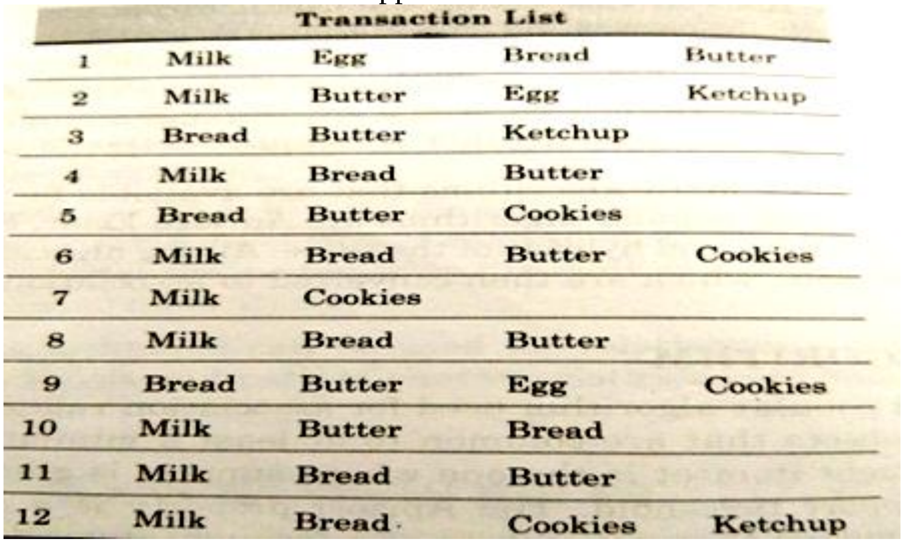
CELEBRATING 25 YEARS
CMRIT
CMR INSTITUTE OF TECHNOLOGY, BENGALURU.
ACCREDITED WITH A+ GRADE BY NAAC

Internal Assessment Test 2 – April 2019

Sub:	Big Data Analytics					Sub Code:	15CS82	Branch:	CSE
Date:	16/04/2019	Duration:	90 min's	Max Marks:	50	Sem / Sec:	8th A,B,C		OBE

Answer any FIVE FULL Questions

		MARKS	CO	RBT
1	Describe the design principles of ANN and list its business applications	[5+5]	CO4	L1
2 (a)	What is a cluster? Write the generic pseudo code for clustering.	[05]	CO4	L2
(b)	List the advantages and disadvantages of K – Means clustering.	[05]	CO4	L2
3	Describe K-means clustering with an example.	[10]	CO4	L3
4	For the dataset given below show the k-frequent itemsets for k=2, 3 and create association rules with 33% support level and 50% confidence.	[4+2+4]	CO4	L3



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

5	Describe the various components of HDFS with a neat diagram.	[10]	CO1	L2
6	Write a note on the following: a> HDFS Block replication b>HDFS safe mode c>Rack awareness	[4+3+3]	CO1	L2
7.	Describe basic steps in MapReduce parallel data flow with neat diagram.	[10]	CO1	L2

Course Outcomes		**Modules covered**	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO10**	**PO11**	**PO12**	**PSO1**	**PSO2**	**PSO3**	**PSO4**
CO 1	Master the concepts of HDFS and MapReduce framework	**1**	1	3	2	3	3	2	0	0	0	0	0	0	3	0	2	3
CO 2	Investigate Hadoop related tools for Big Data Analytics and perform basic Hadoop Administration	**1, 2**	0	3	2	2	3	2	0	0	0	0	0	0	3	0	2	3
CO 3	Recognize the role of Business Intelligence, Data warehousing and Visualization in decision making	**3**	1	2	1	2	1	2	0	0	0	0	0	0	1	0	0	2
CO 4	Infer the importance of core data mining techniques for data analytics	**3, 4**	2	3	3	3	1	1	0	0	0	0	0	0	1	0	0	2
CO 5	Analyze Data Mining Techniques	**3, 5**	2	2	3	3	1	3	0	0	0	0	0	0	1	0	2	3

CELEBRATING 25 YEARS
CMRIT
CMR INSTITUTE OF TECHNOLOGY, BENGALURU.
ACCREDITED WITH A+ GRADE BY NAAC

Scheme Of Evaluation

Internal Assessment Test 2 – April.2019

Sub:	Big Data Analytics							**Code:**	15CS82
Date:	16/04/2019	Duration:	90mins	Max Marks:	50	**Sem:**	VIII	**Branch:**	CSE

Note: Answer Any Five Questions

Question #		**Description**	**Marks Distribution**		**Max Marks**
1	a)	**Describe the design principles of ANN and list its business applications.** • design principles of ANN • business applications	5 M 5 M	10 M	10 M
2	a)	**What is a cluster? Write the generic pseudo code for clustering** • Define cluster • generic pseudo code for clustering	2 M 3 M	5 M	10 M
	b)	**List the advantages and disadvantages of K – Means clustering** • List the advantages of K – Means clustering • disadvantages of K – Means clustering	2.5 M 2.5 M	5 M	
3	a)	**Describe K-means clustering with an example.** • K-means clustering • example	5 M 5 M	10 M	10 M
4	a)	**For the dataset given below show the k-frequent itemsets for k=2, 3 and create association rules with 33% support level and 50% confidence.**	1 M		

		• create association rules with 33% support level and 50% confidence	10 M	10 M	10 M
5	a)	**Describe the various components of HDFS with a neat diagram.** • Description of the various components of HDFS • neat diagram	8 M 2 M	10 M	10 M
6	a)	**Write a note on the following:** a> **HDFS Block replication** **b> HDFS safe mode** **c> Rack awareness** • **HDFS Block replication** • **HDFS safe mode** • **Rack awareness**	4 M 3 M 3 M	10 M	10 M
7	b)	**Describe basic steps in MapReduce parallel data flow with neat diagram** • basic steps in MapReduce parallel data flow • neat diagram	8 M 2 M	10 M	10 M

Solution Of Evaluation

Internal Assessment Test 2 – April.2019

Sub:	Big Data Analytics							**Code:**	15CS82
Date:	16/04/2019	Duration:	90mins	Max Marks:	50	**Sem:**	VIII	**Branch:**	CSE

Q1. Describe the design principles of ANN and list its business applications.

Artificial Neural Networks (ANN) are inspired by the information processing model of the mind/brain. The human brain consists of billions of neurons that link with one another in an intricate pattern. Every neuron receives information from many other neuro ns, processes it, gets excited or not, and passes its state information to other neurons.

Just like the brain is a multipurpose system, so also the ANNs are very versatile systems. They can be used for many kinds of pattern recognition and prediction. They are also used for classification, regression, clustering, association, and optimization activities. They are used in finance, marketing, manufacturing, operations, information systems applications, and so on.

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

ANNs are like a black box trained into solving a particular type of problem, and they can develop high predictive powers. Their intermediate synaptic parameter values evolve as the system obtains feedback on its predictions, and thus an ANN learns from more training data.
Design Principles of an Artificial Neural Network

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

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

depending upon the complexity of the predictive action. The layers of PEs could work in sequence, or they could work in parallel.

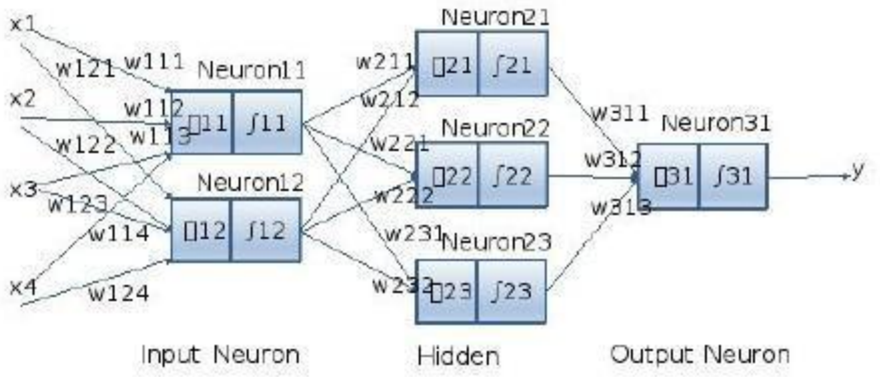


Fig : Model for a multi-layer ANN

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

4. The neural network can be trained by making similar decisions over and over 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. Depending upon the nature of the problem and the availability of good training data, at some point the neural network will learn enough and begin to match the predictive accuracy of a human expert. In many practical situations, the predictions of ANN, trained over a long period of time with a large number of training data, have begun to decisively become more accurate than human experts. At that point ANN can begin to be seriously considered for deployment in real situations in real time.

Q2 a> What is a cluster? Write the generic pseudo code for clustering

An operational definition of a cluster is that, given a representation of n objects, find K groups based on a measure of similarity, such that objects within the same group are alike but the objects in different groups are not alike.

However, the notion of similarity can be interpreted in many ways. Clusters can differ in terms of their shape, size, and density. Clusters are patterns, and there can be many kinds of patterns. Some clusters are the traditional types, such as data points hanging together. However, there are other clusters, such as all points representing the circumference of a circle. There may be concentric circles with points of different circles representing different clusters. The presence of noise in the data makes the detection of the clusters even more difficult.

Here is the generic pseudocode for clustering

1. Pick an arbitrary number of groups/segments to be created
2. Start with some initial randomly-chosen center values for groups
3. Classify instances to closest groups
4. Compute new values for the group centers
5. Repeat step 3 & 4 till groups converge
6. If clusters are not satisfactory, go to step 1 and pick a different number of groups/segments

Q2b> List the advantages and disadvantages of K – Means clustering

There are many advantages of K-Means Algorithm

1. K-Means algorithm is simple, easy to understand and easy to implement.

2. It is also efficient, in that the time taken to cluster k-means, rises linearly with the number of data points.

3. No other clustering algorithm performs better than K-Means, in general.

There are a few disadvantages too:

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

Q3. Describe K-means clustering with an example.

Cluster analysis is a machine-learning technique. The quality of a clustering result depends on the algorithm, the distance function, and the application. First, consider the distance function. Most cluster analysis methods use a distance measure to calculate the closeness between pairs of items. There are two major measures of distances: Euclidian distance (―as the crow flies‖ or straight line) is the most intuitive measure. The other popular measure is the Manhattan (rectilinear) distance, where one can go only in orthogonal directions. The Euclidian distance is the hypotenuse of a right triangle, while the Manhattan distance is the sum of the two legs of the right triangle.

In either case, the key objective of the clustering algorithm is the same:

- Inter-clusters distanceÞ maximized; and
- Intra-clusters distanceÞ minimized

There are many algorithms to produce clusters. There are top-down, hierarchical methods that start with creating a given number of best-fitting clusters. There are also bottom-up methods that begin with identifying naturally occurring clusters.

The most popular clustering algorithm is the K-means algorithm. It is a topdown, statistical technique, based on the method of minimizing the least squared distance from the center points of the clusters. Other techniques, such as neural networks, are also used for clustering. Comparing cluster algorithms is a difficult task as there is no single right number of clusters. However, the speed of the algorithm and its versatility in terms of different dataset are important criteria.

Q4. For the dataset given below show the k-frequent itemsets for k=2, 3 and create association rules with 33% support level and 50% confidence.

Plz refer text book

Q5. Describe the various components of HDFS with a neat diagram.

Hadoop Distributed File System (HDFS) is a distributed file system which is designed to run on commodity hardware. Commodity hardware is cheaper in cost. Since Hadoop requires processing power of multiple machines and since it is expensive to deploy costly hardware, we use commodity hardware. When commodity hardware is used, failures are more common rather than an exception. HDFS is highly fault-tolerant and is designed to run on commodity hardware.

HDFS provides high throughput access to the data stored. So it is extremely useful when you want to build applications which require large data sets.

HDFS was originally built as infrastructure layer for Apache Nutch. It is now pretty much part of Apache Hadoop project.

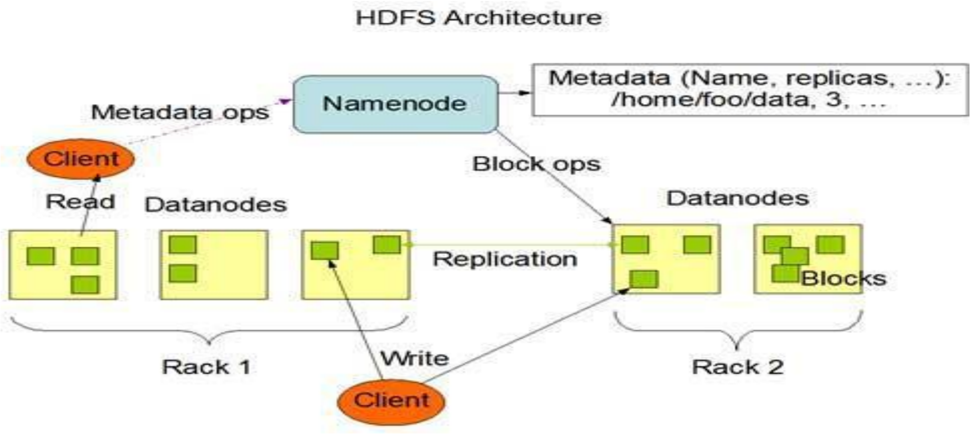


Fig 1.1 HDFS Architecture

HDFS has master/slave architecture. In this architecture one of the machines will be designated as a master node (or name node). Every other machine would be acting as slave (or data node). NameNode/DataNode are java processes that run on the machines when Hadoop software is installed.

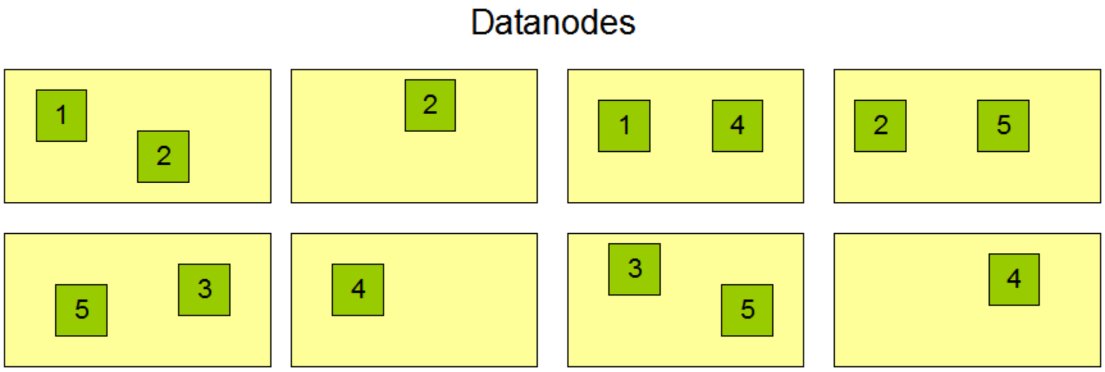
NameNode is responsible for managing the metadata about the HDFS Files. This metadata includes various information about the HDFS File such as Name of the file, File Permissions, FileSize, Blocks etc. It is also responsible for performing various namespace operations like opening, closing, renaming the files or directories.

Whenever a file is to be stored in HDFS, it is divided into blocks. By default, blocksize is 64MB (Configurable). These blocks are replicated (default is 3) and stored across various datanodes to take care of hardware failures and for faster data transfers. NameNode maintains a mapping of blocks to DataNodes.

DataNodes serves the read and write requests from HDFS file system clients. They are also responsible for creation of block replicas and for checking if blocks are corrupted or not. It sends the ping messages to the NameNode in the form of block mappings.

Q6. Write a note on the following:
a> HDFS Block replication b>HDFS safe mode c>Rack awareness

HDFS Block replication
HDFS tries to satisfy a read request from a replica that is closest to the reader. If there exists a replica on the same rack as the reader node, then that replica is preferred to satisfy the read request
If a HDFS cluster spans multiple data centers, then a replica that is resident in the local data center is preferred over remote replicas.



HDFS safe mode

On startup, the Namenode enters a special state called Safemode. Replication of data blocks does not occur when the Namenode is in Safemode state. The Namenode receives Heartbeat and Blockreport from the Datanodes. A Blockreport contains the list of data blocks that a Datanode reports to the Namenode. Each block has a specified minimum number of replicas. A block is considered safely-replicated when the minimum number of replicas of that data block has checked in with the Namenode. When a configurable percentage of safely-replicated data blocks checks in with the Namenode (plus an additional 30 seconds), the Namenode exits the Safemode state. It then determines the list of data blocks (if any) that have fewer than the specified number of replicas. The Namenode then replicates these blocks to other Datanodes.

Rack awareness

Rack– It the collection of machines around 40-50. All these machines are connected using the same network switch and if that network goes down then all machines in that rack will be out of service. Thus we say rack is down.

Rack Awareness was introduced by Apache Hadoop to overcome this issue. In Rack Awareness, NameNode chooses the DataNode which is closer to the same rack or nearby rack. NameNode maintains Rack ids of each DataNode to achieve rack information. Thus, this concept chooses Datanodes based on the rack information. NameNode in hadoop makes ensures that all the replicas should not stored on the same rack or single rack. Rack Awareness Algorithm reduces latency as well as Fault Tolerance.

Default replication factor is 3. Therefore according to Rack Awareness Algorithm:

- The first replica of the block will store on a local rack.
- The next replica will store on another datanode within the same rack.
- The third replica stored on the different rack.

In Hadoop, we need Rack Awareness for below reason: It improves

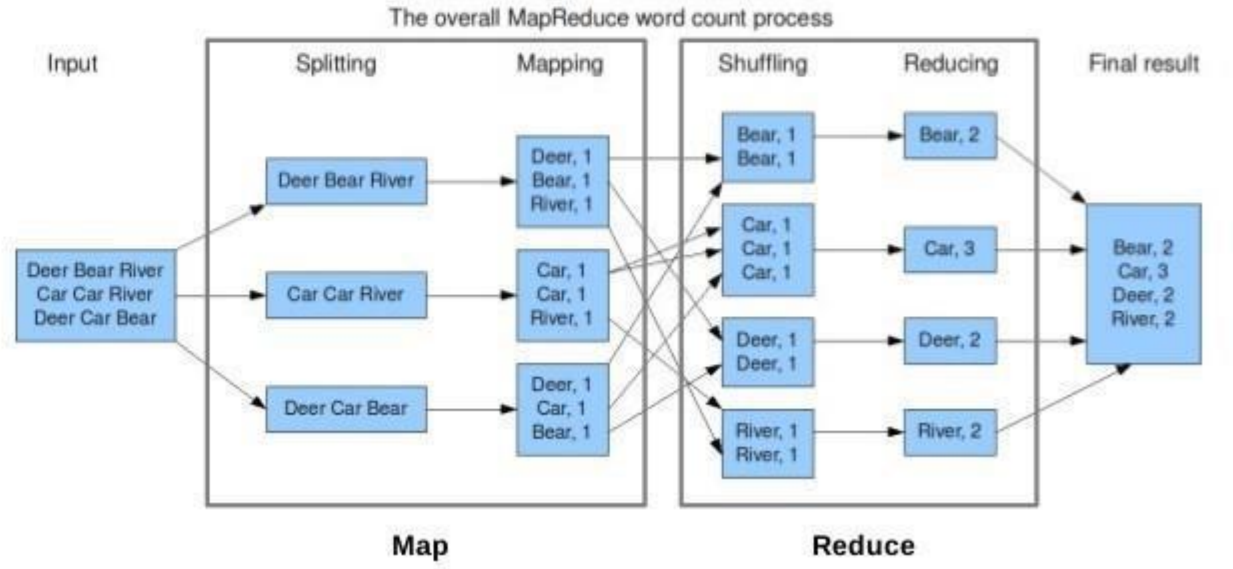
- Data high availability and reliability.
- The performance of the cluster.
- Network bandwidth.

Q7. Describe basic steps in MapReduce parallel data flow with neat diagram.

Hadoop MapReduce is a programming paradigm at the heart of Apache Hadoop for providing massive scalability across hundreds or thousands of Hadoop clusters on commodity hardware. The MapReduce model processes large unstructured data sets with a distributed algorithm on a Hadoop cluster.

The term MapReduce represents two separate and distinct tasks Hadoop programs perform-Map Job and Reduce Job. Map job scales takes data sets as input and processes them to produce key value pairs. Reduce job takes the output of the Map job i.e. the key value pairs and aggregates them to produce desired results. The input and output of the map and reduce jobs are stored in HDFS.

The following word count example explains MapReduce method. For simplicity, let's consider a few words of a text document. We want to find the number of occurrence of each word. First the input is split to distribute the work among all the map nodes as shown in the figure. Then each word is identified and mapped to the number one. Thus the pairs also called as tuples are created. In the first mapper node three words Deer, Bear and River are passed. Thus the output of the node will be three key, value pairs with three distinct keys and value set to one. The mapping process remains the same in all the nodes. These tuples are then passed to the reduce nodes. A partitioner comes into action which carries out shuffling so that all the tuples with same key are sent to same node.



The Reducer node processes all the tuples such that all the pairs with same key are counted and the count is updated as the value of that specific key. In the example there are two pairs with the key _Bear‘ which are then reduced to single tuple with the value equal to the count. All the output tuples are then collected and written in the output file.