22MCA2052

NATION FIRST

USN	1	B	Y								

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(An Autonomous Institute affiliated to Visvesvaraya Technological University, Belagavi)

SEMESTER END EXAMINATION QUESTION PAPER

Second Semester MCA Degree Examination

Regular / Make-up / Arrears / Supplementary

BIG DATA ANALYTICS

Time: 3 hrs. **Max. Marks: 100**

Note: Answer FIVE full questions, choosing ONE full question from each module.

Q. No	Module - 1	Marks	CO, RBT
1a.	What is Big Data, and where does it come from? How does it work? Describe any five characteristics of it.	10	CO1, K2
b.	Construct Box plot for given data: 51,17,25,39,7,49,67,41,20,2,43,13.	5	CO1, K2
c.	Explain Visual Data Exploration and Missing Values.	5	CO1, K2
	OR		
2a.	List the satisfactory requirements of a good analytical model and explain.	10	CO1, K2
b.	Describe outlier. How do you detect and treat outliers? Give Example.	05	CO1, K2
c.	Describe the various methods to categorise the data.	05	CO1, K2
	Module – 2		
3a.	What is predictive analytics? Define neural networks. Explain application of neural networks in predictive analytics.	10	CO2, K2
b.	Describe confusion matrix, accuracy, precision, recall, specificity with respect to classification model evaluation.	10	CO2, K2
	OR		
4a.	Differentiate between predictive analytics and descriptive analytics.	05	CO2, K2
b.	Explain the concept of random forest with an example.	05	CO2, K2
c.	Discuss any 2 techniques used for classification in predictive analytics.	10	CO2, K2
	Module – 3		
5a.	Elaborate on the need for tools like Hadoop.	05	CO3, K2
b.	Assume a Hadoop cluster is configured to use block size of **64 MB** and replication factor of **3**. A user has a file “BMSIT_M.dat” of **525 MB** and user stores it in HDFS. What is the size of the last block of the file? How many blocks are required on HDFS for storing the file?	05	CO3, K2
c.	Discuss the critical or core components of Hadoop and their working along with a neat diagram.	10	CO3, K2
	OR		

6a.	Discuss how high availability is achieved in Hadoop.	05	CO3, K2
b.	What is MapReduce. Discuss its application with a neat diagram.	05	CO3, K2
c.	Describe the sequence of events flow when client writing data in HDFS with a neat diagram.	10	CO3, K2
	Module – 4		
7a.	How is Apache Spark different from MapReduce? Explain	05	CO4, K2
b.	Describe the important components of the Spark ecosystem with a neat diagram.	10	CO4, K2
c.	What is the significance of Resilient Distributed Datasets in Spark?	05	CO4, K2
	OR		
8a.	With a neat diagram, Explain how Spark runs applications with the help of its architecture.	10	CO4, K2
b.	What are the different cluster managers available in Apache Spark? List them.	05	CO4, K2
c.	Explain the lazy evaluation in Spark.	05	CO4, K2
	OR		
	Module – 5		
9a.	Illustrate how tables are partitioned in Hive.	05	CO5, K2
b.	List the components used in Hive Query Processor. Explain with diagram.	10	CO5, K2
c.	Discuss various ways through which data is added to Hive tables.	05	CO5, K2
	OR		
10a.	Discuss how internal and external tables are supported in Hive.	08	CO5, K2
b.	Mention various data types supported by Hive. Explain with an example each.	06	CO5, K2
c.	Explain the Trim and Reverse functions in Hive with examples.	06	CO5, K2

Course Outcomes (COs): At the end of the course, the student will be able to

COs	Statements
CO-1	Identify the business problem for a given context and frame the objectives to solve it using data analytics tools.
CO-2	Differentiate various types of analytics algorithms and context of their application.
CO-3	Illustrate the architecture of HDFS and MapReduce.
CO-4	Explore Spark architecture and its language APIs.
CO-5	Write Hive queries against large datasets on clusters.

******

" Success is not final; failure is not fatal… it is the courage to continue that counts ".

USN | 1 | B | Y | | | | | | | |

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(An Autonomous Institute affiliated to Visvesvaraya Technological University, Belagavi)

SEMESTER END EXAMINATION QUESTION PAPER

Second Semester MCA Degree Examination, September / October – 2023

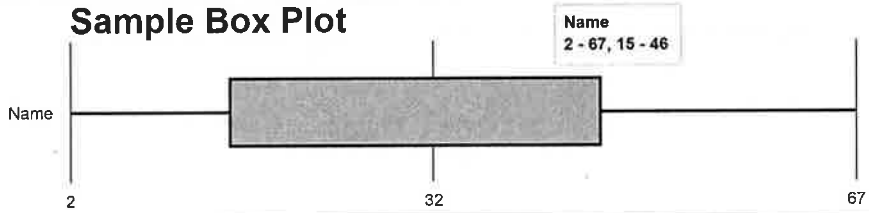
BIG DATA ANALYTICS

Time: 3 hrs. **Max. Marks: 100**

Note: Answer FIVE full questions, choosing ONE full question from each module.

<table>
<tr><th>Q.#</th><th>Module - 1</th><th>Marks</th></tr>
<tr><td>1a.</td><td>What is Big Data, and where does it come from? How does it work? Describe any five characteristics of it.</td><td>10</td></tr>
<tr><td colspan="2">Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them. But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.

Big data analytics is used in nearly every industry to identify patterns and trends, answer questions, gain insights into customers and tackle complex problems. Companies and organizations use the information for a multitude of reasons like growing their businesses, understanding customer decisions, enhancing research, making forecasts and targeting key audiences for advertising.

The 5 V's of big data (velocity, volume, value, variety and veracity) are the five main and innate characteristics of big data. + Detailed explanation.</td><td>2 m
3 m
5m</td></tr>
<tr><td>b.</td><td>Construct Box plot for given data: 51,17,25,39,7,49,67,41,20,2,43,13.</td><td>5</td></tr>
<tr><td colspan="2">1. Calculate quartile values from the source data set.
2. Calculate quartile differences.
3. Create a stacked column chart type from the quartile ranges.
4. Convert the stacked column chart to the box plot style.

 </td><td></td></tr>
<tr><td>c.</td><td>Explain Visual Data Exploration and Missing Values.</td><td>5</td></tr>
<tr><td colspan="2">Visual data exploration is a mandatory intial step whether or not more formal analysis follows. When combined with descriptive statistics, visualization provides an effective way to identify summaries, structure, relationships, differences, and abnormalities in the data.
Often times no elaborate analysis is necessary as all the important conclusions required for a decision are evident from simple visual examination of the data. Other times, data exploration will be used to help guide the data cleaning, feature selection, and sampling process.
Regardless, visual data exploration is about investigating the characteristics of your data set. To do this, we typically create numerous plots in an interactive fashion.</td><td></td></tr>
</table>

Scheme and Solution

2a.	List the satisfactory requirements of a good analytical model and explain.	10
	A good analytical model should satisfy following requirements: – Achieve business relevance – Statistical significance and predictive power – Interpretability – Justifiability – Operationally Efficient – Economic Cost – Regulation and Legislation + Explanation	
b.	Describe outlier. How do you detect and treat outliers? Give Example.	05
	Outliers are extreme observations that are very dissimilar to the rest of the population. Actually, two types of outliers are: • Valid observations (e.g salary of boss is $1 million) • Invalid observations (e.g age is 300 years) We can detect and treat outliers in the following methods • Detecting outliers using the Z-scores • Detecting outliers using the Inter Quantile Range(IQR) Below are some of the methods of treating the outliers • Trimming/Remove the outliers • Quantile based flooring and capping • Mean/Median imputation	
c.	Describe the various methods to categorise the data.	05
	Data classification methods • Manual interval. • Defined interval. • Equal interval. • Quantile. • Natural breaks (Jenks) • Geometrical interval. • Standard deviation.	
3a.	What is predictive analytics? Define neural networks. Explain application of neural networks in predictive analytics.	10
	Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior. A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. The applications of neural networks are: 1. Facial Recognition 2. Stock Market Prediction 3. Social Media 4. Aerospace 5. Defence	

6. Healthcare
7. Signature Verification and Handwriting Analysis
8. Weather Forecasting

b.	Describe confusion matrix, accuracy, precision, recall, specificity w.r.t. classification model evaluation.	10

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy

$$\textit{Recall} = \frac{TP}{TP + FN} \qquad \textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{F - measure} = \frac{2*Recall*Precision}{Recall + Precision}$$

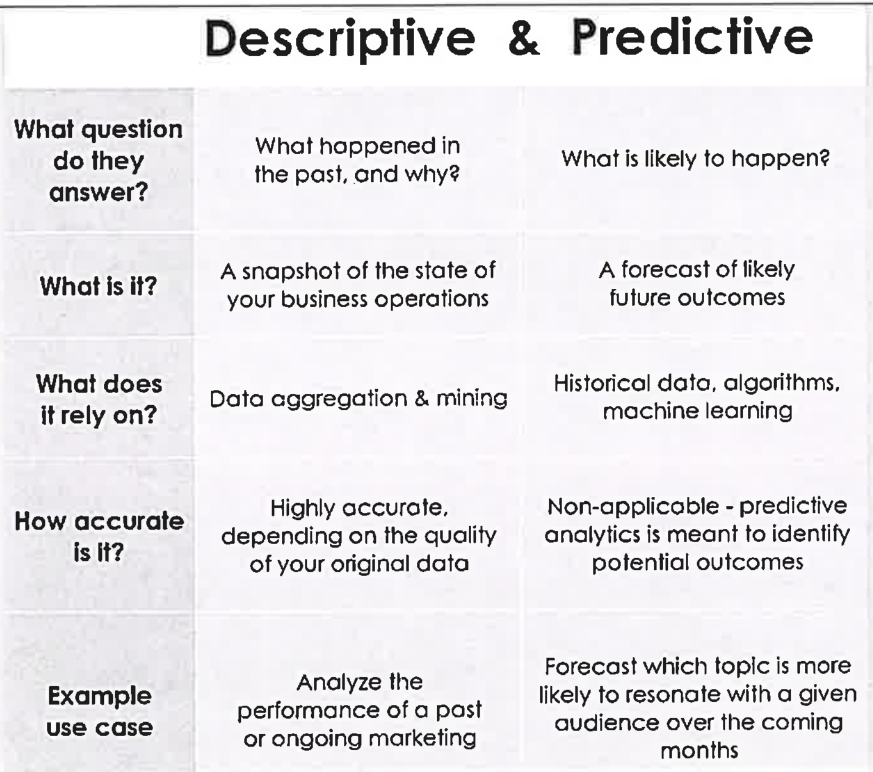
Accuracy (ACC) $\mathrm{ACC} = \dfrac{TP + TN}{TP + TN + FN + FP} = \dfrac{TP + TN}{P + N}$

Sensitivity (SN) $\mathrm{SN} = \dfrac{TP}{TP + FN} = \dfrac{TP}{P}$

Specificity (SP) $\mathrm{SP} = \dfrac{TN}{TN + FP} = \dfrac{TN}{N}$

Scheme and Solution

4a. Differentiate between predictive analytics and descriptive analytics. 05

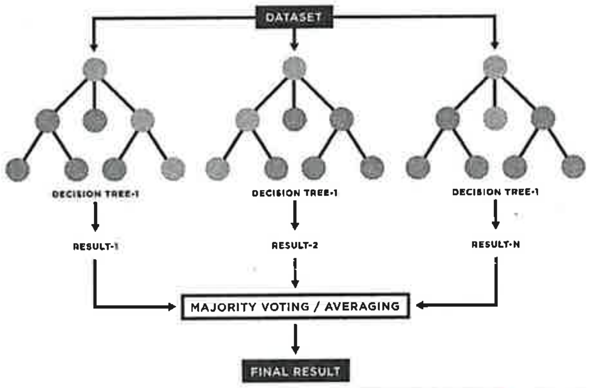


Descriptive & Predictive

	Descriptive	Predictive
What question do they answer?	What happened in the past, and why?	What is likely to happen?
What is it?	A snapshot of the state of your business operations	A forecast of likely future outcomes
What does it rely on?	Data aggregation & mining	Historical data, algorithms, machine learning
How accurate is it?	Highly accurate, depending on the quality of your original data	Non-applicable - predictive analytics is meant to identify potential outcomes
Example use case	Analyze the performance of a past or ongoing marketing	Forecast which topic is more likely to resonate with a given audience over the coming months

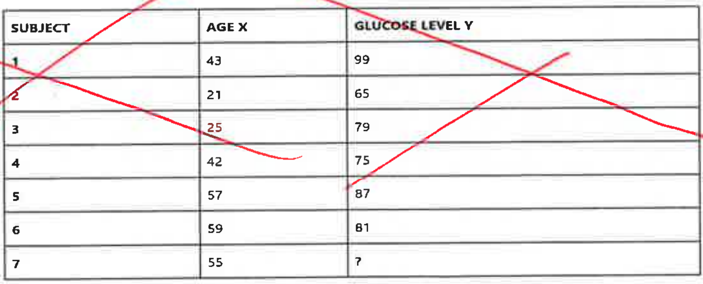
b. Explain the concept of random forest with an example. 05

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.



4

c. Using the Linear Regression on following dataset, prediction – the value of y for the given value of x = 55. 10



SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81
7	55	?

Any 2 classification Methods - 5M each.

Solution:

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable.
- Explain the impact of changes in an independent variable on the dependent variable.

The dependent variable is the variable we wish to explain and

Independent variable is the variable used to explain the dependent variable

The key steps for regression are simple:

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

The regression model is described as a linear equation that follows.

y is the dependent variable, that is, the variable being predicted.

x is the independent variable or the predictor variable.

There could be many predictor variables (such as *x*1, *x*2, . . .) in a regression equation. However, there can be only one dependent variable (*y*) in the regression equation.

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

$$\boxed{\hat{Y}_i = b_0 + b_1 X_i}$$

Value of X for observation i

Here we need to find the value of b₀, b₁ using the following equation.

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{n\,(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Step 1: *Make a chart of your data, filling in the columns in the same way as you would fill in the chart if you were finding the Pearson's Correlation Coefficient.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X²	Y²
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Step 2: Use the following equations to find b_0 and b_1.

Find b_0:

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b_0 = \frac{(486)(11409) - (247)(20485)}{6\,(11409) - (247)^2}$$

$$b_0 = \frac{4848979}{7445} = 65.14$$

Find b_1:

$$b_1 = \frac{n\,(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{6\,(20485) - (247)(486)}{6(11409) - (247)^2}$$

$$b_1 = \frac{2868}{7445} = 0.385335$$

Step 3: *Insert the values into the equation.*

y' = b_o + b_1 * x

y' = 65.14 + (0.385225 * x)

Step 4: *Prediction – the value of y for the given value of x = 55*

y' = 65.14 +(0.385225 *55)

y' =86.327

5a.	List out the problems encountered during parallel computing. What problems Hadoop could solve?	05
Problem: Efficient Storage and Reliability *Solution :HDFS Block Storage with Replication factor Problem: Processing bigdata * Solution: Map Reduce Model.		

tools like

Discuss the need of Hadoop.

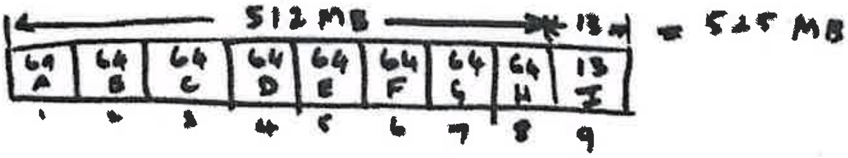
5a.	List out the problems encountered during parallel computing. What problems Hadoop could solve?	05
Problem: Efficient Storage and Reliability □ Solution :HDFS Block Storage with Replication factor Problem: Processing bigdata □ Solution: Map Reduce Model.		
b.	A Hadoop cluster is configured to use block size of 64 MB and replication factor of 3. A user has a file "BMSIT_M.dat" of 525 MB and user stores it in HDFS. What is the size of the last block of the file? How many blocks are required on HDFS for storing the file?	05

Soln.

The default block size is 64 MB

Replication factor = 3

The file Size = 525 MB



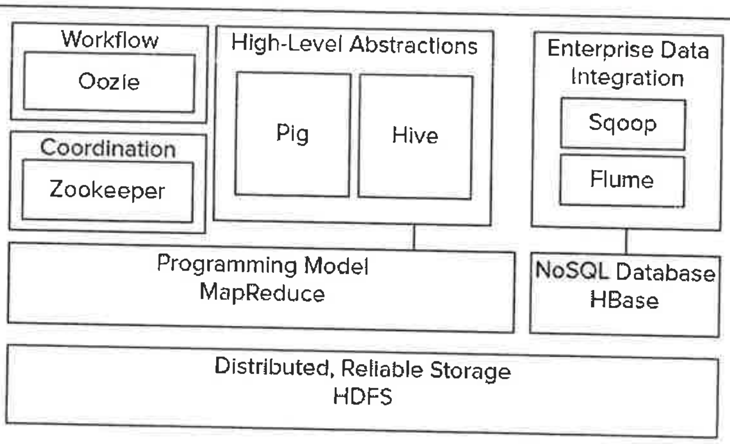
∴ 9 Blocks are used to allocate file Size of 525 MB

Since Replication factor is 3

Total Blocks required = 3 × 9

= 27 Blocks

c.	Discuss the critical or core components of Hadoop and their working along with a neat diagram.	10

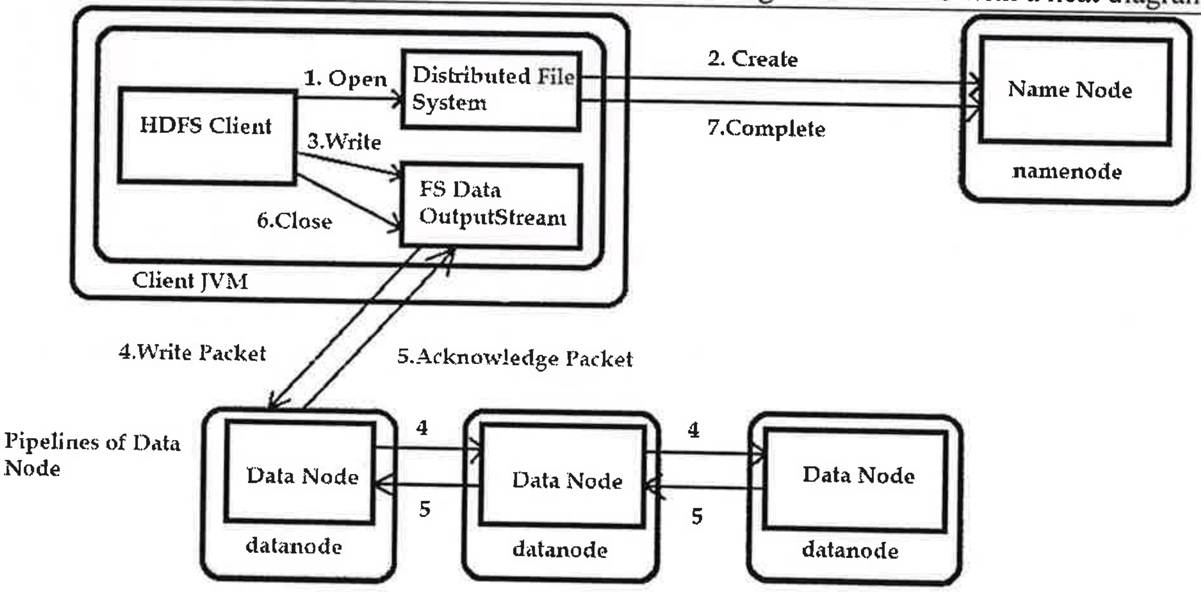


Core components of the Hadoop ecosystem

+ Explanation of each component

Scheme and Solution

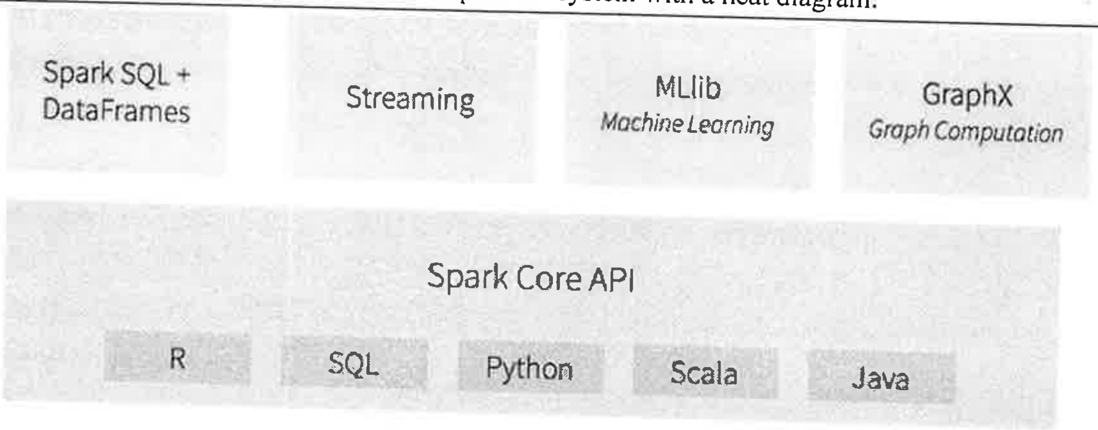
High availability in Hadoop.

6a.	Horizontal scaling refers to adding additional nodes to the infrastructure to cope with new demands. Explain the importance of Hadoop Federation.	05
HDFS federation provides namespace scalability, performance advantage because of scaling read/write throughput, and isolation of applications and users in a multitenanted environment. Namespace Scalability: - Namespace Scalability: Federation adds namespace horizontal scaling. Large deployments or deployments using lot of small files benefit from namespace scaling by allowing more NameNodes to be added to the cluster. - Performance: File system throughput is not limited by a single NameNode. Adding more NameNodes to the cluster scales the file system read/write throughput. - Isolation: A single NameNode does not offer namespace isolation in a multi-user environment. By using multiple NameNodes, different categories of applications and users can be isolated to different namespaces.		
b.	What is the procedure to compare two HDFS files? Discuss with an example. MapReduce.	05
Compare two files in HDFS : Binary (non-text and non-flat. For example- images) files : It compares the checksum of the files. The baseline expected is a set of checksums. The comparison is based on matching the checksum of the baseline to the target file. The diff command gives you the differences in an orderly way so content-wise you should be careful to use it in file comparison. For example let file X have three lines each having A , B , C respectively. The second file Y has C B A.		
c.	Describe the sequence of events flow when client writing data in HDFS with a neat diagram.	10
 HDFS Client 1. Open Distributed File System 2. Create 7.Complete Name Node namenode 3.Write 6.Close FS Data OutputStream Client JVM 4.Write Packet 5.Acknowledge Packet Pipelines of Data Node Data Node datanode 4 5 Data Node datanode 4 5 Data Node datanode + Explanation		
7a.	How is Apache Spark different from MapReduce? Explain	05

Difference Between MapReduce and Spark

S.No.	MapReduce	Spark
1.	It is a framework that is open-source which is used for writing data into the Hadoop Distributed File System.	It is an open-source framework used for faster data processing.
2.	It is having a very slow speed as compared to Apache Spark.	It is much faster than MapReduce.
3.	It is unable to handle real-time processing.	It can deal with real-time processing.
4.	It is difficult to program as you required code for every process.	It is easy to program.
5.	It supports more security projects.	Its security is not as good as MapReduce and continuously working on its security issues.
6.	For performing the task, It is unable to cache in memory.	It can cache the memory data for processing its task.
7.	Its scalability is good as you can add up to n different nodes.	It is having low scalability as compared to MapReduce.
8.	It actually needs other queries to perform the task.	It has Spark SQL as its very own query language.

b.	Describe the important components of the Spark ecosystem with a neat diagram.	10

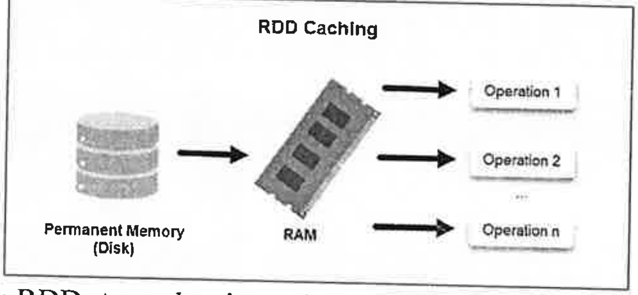


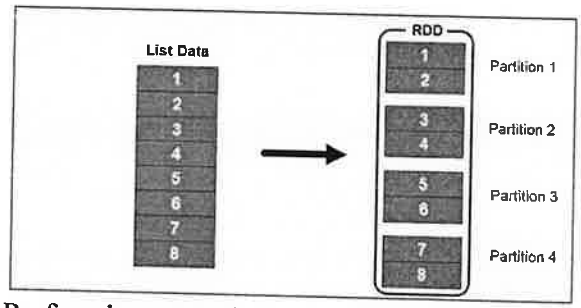
+ Explanation

c. What is the significance of Resilient Distributed Datasets in Spark? **05**

Resilient Distributed Datasets (RDDs) are the primary data structure in Spark. RDDs are reliable and memory-efficient when it comes to parallel processing. By storing and processing data in RDDs, Spark speeds up MapReduce processes. RDDs address MapReduce's shortcomings in data sharing. When reusing data for computations, MapReduce requires writing to external storage (HDFS, Cassandra, HBase, etc.). The read and write processes between jobs consume a significant amount of memory.

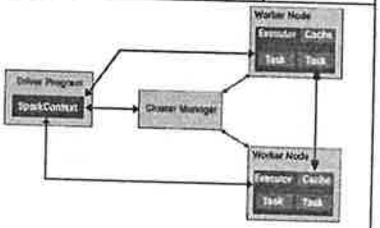
Furthermore, data sharing between tasks is slow due to replication, serialization, and increased disk usage. RDDs aim to reduce the usage of external storage systems by leveraging in-memory compute operation storage. This approach improves data exchange speeds between tasks by 10 to 100 times. Speed is critical when working with large data volumes. Spark RDDs make it easier to train machine learning algorithms and handle large amounts of data for analytics.





An RDD stores data in read-only mode, making it immutable. Performing operations on existing RDDs creates new objects without manipulating existing data. RDDs reside in RAM through a caching process. Data that does not fit is either recalculated to reduce the size or stored on a permanent storage. Caching allows retrieving data without reading from disk, reducing disk overhead. RDDs further distribute the data storage across multiple partitions. Partitioning allows data recovery in case a node fails and ensures the data is available at all times.

8a. With a neat diagram, Explain how Spark runs applications with the help of its architecture. **10**



The high-level components that are a part of the architecture of the Apache Spark application:

- The Spark Driver
- The Spark Executors
- The Cluster Manager

Modes of Execution

An execution model helps determine where the resources mentioned previously are physically located when the application is run. There are three modes of execution to choose from:

- Cluster Mode
- Client Mode
- Local Mode

b. What are the different cluster managers available in Apache Spark? List them. **05**

Cluster manager is a platform (cluster mode) where we can run Spark. Simply put, cluster manager provides resources to all worker nodes as per need, it operates all nodes accordingly.

We can say there are a master node and worker nodes available in a cluster. That master nodes provide an efficient working environment to worker nodes. There are three types of Spark cluster manager. Spark supports these cluster manager:

- Standalone cluster manager
- Hadoop Yarn
- Apache Mesos

c. Explain the lazy evaluation in Spark. **05**

Lazy Evaluation in Sparks means Spark will not start the execution of the process until an Action is called. Once an Action is called, Spark starts looking at all the transformations and creates a DAG. DAG is sequence of operations that need to be performed in a process to get the resultant output. If Spark could wait until an Action is called, it may merge some transformation or skip some unnecessary transformation and prepare a perfect optimized execution plan.

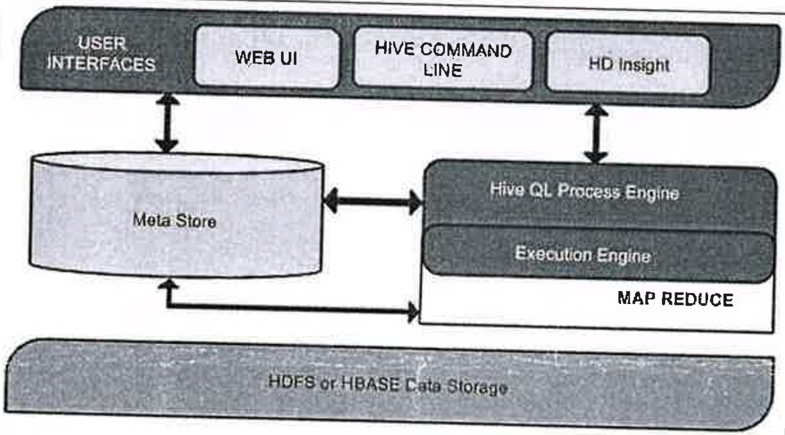
9a. ~~Explain the process to access subdirectories recursively in Hive queries.~~ Elaborate a table Partitioning in hive of Hive **05**

To process directories recursively in Hive, we need to set below two commands in hive session. These two parameters work in conjunction.

```
hive> Set mapred.input.dir.recursive=true;
hive> Set hive.mapred.supports.subdirectories=true;
```

Now hive tables can be pointed to the higher level directory. This is suitable for a scenario where the directory structure is as following: /data/country/state/city

b. List the components used in Hive Query Processor. Explain with diagram. **10**



+ Explanation

c. ~~What is a Hive variable? What do we use it for? Explain with an example.~~ Discuss various ways through which data is inserted into Hive Table. **05**

Hive variables are basically created in the Hive environment that is referenced by Hive scripting languages. They allow to pass some values to a Hive query when the query starts executing. They use the source command. + Example

10a. ~~What is the maximum size of a string data type supported by Hive? Explain how Hive supports binary formats.~~ Discuss how internal and external tables are processed in Hive. **08**

~~The maximum size of a string data type supported by Hive is 2 GB. Hive supports the text file format by default, and it also supports the binary format sequence files, ORC files, Avro data files, and Parquet files.~~

- Sequence file: It is a splittable, compressible, and row-oriented file with a general binary format.
- ORC file: Optimized row columnar (ORC) format file is a record-columnar and column-oriented storage file. It divides the table in row split. Each split stores the value of the first row in the first column and follows subsequently.
- Avro data file: It is the same as a sequence file that is splittable, compressible, and row-oriented but without the support of schema evolution and multilingual binding.
- Parquet file: In Parquet format, along with storing rows of data adjacent to one another, we can also store column values adjacent to each other such that both horizontally and vertically datasets are partitioned.

b.	Mention various data types supported by Hive. Explain with an example each.	06

Hive data types are categorized in numeric types, string types, misc types, and complex types. Explanation with example.

c.	Explain the Trim and Reverse functions in Hive with examples.	06

The trim function will delete the spaces associated with a string.

Example:

```
TRIM(' INTELLIPAAT ');
```

Output:

```
INTELLIPAAT
```

To remove the leading space:

```
LTRIM('INTELLIPAAT');
```

To remove the trailing space:

```
RTRIM('INTELLIPAAT');
```

In the reverse function, characters are reversed in the string.

Example:

```
REVERSE('INTELLIPAAT');
```

Output:

```
TAAPILLETNI
```

******