22MCA2052

USN	1	B	Y								

NATION FIRST

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

Q. No	Module - 1	Marks	CO & RBT
1a.	What are Outliers and discuss causes of outliers	6	CO1, K2
b.	What is Big Data? Demonstrate the working of analytical processing Model.	10	CO1, K2
c.	Explain the different sources of Big data and Explain?	4	CO1, K2
	OR		
2a.	Define Big Data and Characterise 5V's of Big Data.	10	CO1, K2
b.	Construct Box plot for given data: 51,17,25,39,7,49,67,41,20,2,43,13.	4	CO1, K2
c.	Explain the schemes to deal with missing values in a dataset.	6	CO1, K2
	Module – 2		
3a.	Discuss the logistic regression role in predictive analytics.	7	CO2, K2
b.	Justify the need of having support vector machines for predictive analytics.	7	CO2, K2
c.	Differentiate the following i) predictive analytics and descriptive analytics. ii) hierarchical and non-hierarchical clustering.	6	CO2, K2
	OR		
4a.	Discuss various performance methods to evaluate classification models.	8	CO2, K2
b.	Describe the role of decision trees in predictive analytics.	7	CO2, K2
c.	Describe bagging and boosting concepts in predictive analytics.	5	CO2, K2
	Module – 3		
5a.	Demonstrate core architecture of Hadoop with suitable block diagram. Discuss role of each component in detail.	10	CO3, K2
b.	Demonstrate different stages of MapReduce with an example.	10	CO3, K2
	OR		

6a.	Analyse the anatomy of writing data into a file in HDFS with a neat diagram.	**10**	**CO3, K2**
b.	Describe the role of combiner functions in MapReduce processing.	**10**	**CO3, K2**
	Module – 4		
7a.	Discuss the architecture of Spark and Spark's language APIs.	**10**	**CO4, K2**
b.	Discuss the dataframes and datasets in Spark.	**10**	**CO4, K2**
	OR		
8a.	Elaborate on Spark's toolset with a neat sketch.	**10**	**CO4, K2**
b.	Bring out the importance of lazy evaluations in Spark.	**5**	**CO4, K2**
c.	Discuss the overview of structured API.	**5**	**CO4, K2**
	Module – 5		
9a.	Analyze architecture of APACHE HIVE. Explain different ways of inserting/loading data to HIVE tables with example.	**10**	**CO5, K2**
b.	Illustrate HiveQL Data definition and Data manipulation commands with example.	**10**	**CO5, K2**
	OR		
10a.	Summarize various data types supported by HiveQL with an example.	**8**	**CO5, K2**
b.	Design the hive solution for the following queries in HiveQL. Creating student table with USN, name of the student, year of join(yoj) ,course, email address, phone number and CGPA i) Find the total no of students in each course. ii) List the student who is having maximum and minimum CGPA iii) Create a view to store the details of the students whose CGPA is greater than the 9.00. iv) List the students in the ascending order of their course	**8**	**CO5, K2**
c.	Illustrate Metastore in HIVE.	**4**	**CO5, K2**

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

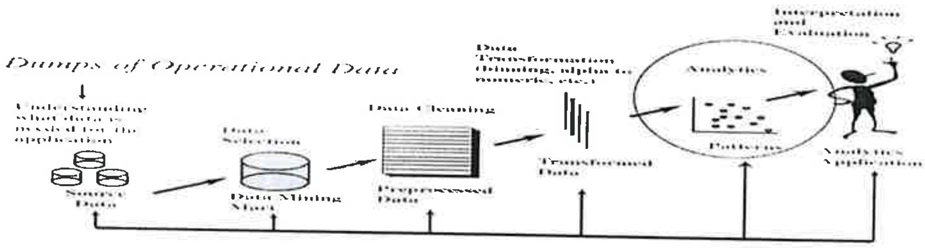
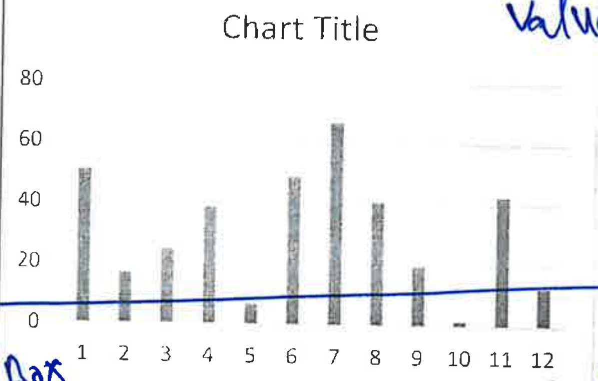
COs	**Statements**
CO-1	Identify the business problem for a given context and frame the objectives to solve it using data analytics tools.
CO-2	Differentiate various types of analytics algorithms and context of their application.
CO-3	Illustrate the architecture of HDFS and MapReduce.
CO-4	Explore Spark architecture and its language APIs.
CO-5	Write Hive queries against large datasets on clusters.
K1- Remembering K2 - Understanding K3 – Applying K4- Analyzing K5 - Evaluating K6 -Creating	

"Success is the progressive realization of a worthy goal."

****

DESCRIPTIVE QUESTIONS PATTERN

COURSE CODE: **2052** COURSE NAME **BIG DATA ANALYTICS**

Q. No	Scheme and Solutions	Marks
1 a)	Brief description of outliers Types of outliers Causes of outliers • Data entry errors (human errors) • Measurement errors (instrument errors) • Experimental errors (data extraction or experiment planning/executing errors) • Intentional (dummy outliers made to test detection methods) • Data processing errors (data manipulation or data set unintended mutations) • Sampling errors (extracting or mixing data from wrong or various sources) • Natural (not an error, novelties in data)	3 1 2
b)	Big Data Definition with examples  Dumps of Operational Data Understanding what data is needed for the application Source Data Data Selection Data Mining Mart Data Cleaning Preprocessed Data Data Transformation (binning, alpha to numeric, etc.) Transformed Data Analytics Patterns Interpretation and Evaluation Analytics Application Explanation of each component Business understandings, data exploration, data preparation, data processing, data analytics, data evaluation	2 3 5
1)c) ~~1 c)~~	~~Social Networks~~ ~~Social Networks provide human-sourced information from:~~ ~~Traditional Business Systems~~ ~~These organizations offer customers services or products~~ ~~Internet of Things~~ Machine Data Social Data Transactional Data	4
~~2 a)~~ 2a) ~~b)~~	**Big Data is characterized in terms of following Vs:** Volume, Velocity, Variety, Veracity, ~~Validity, Vulnerability; Vulnerability~~ Value, Variability, Explanation of any 5  	2*5
2b)	Box ~~Bar~~ plot explanation with fig	4M
2c)	Methods to treat missing values — Replace, Delete, Keep – 2M each Deletion, Mean/ ~~Mode~~/ Median ~~Imputation~~ , Prediction Model ,~~KNN~~ Imputation ~~with~~ examples	6M

DESCRIPTIVE QUESTIONS PATTERN

COURSE CODE: **2052** COURSE NAME **BIG DATA ANALYTICS**

Q. No	Scheme and Solutions	Marks
3 a)	**Explanation & examples** Expl – 5M, fig – 2M Logistic regression aims to measure the relationship between a categorical dependent variable and one or more independent variables (usually continuous) by plotting the dependent variables' probability scores. A categorical variable is a variable that can take values falling in limited categories instead of being continuous.Logistic regression techniques have recently experienced a surge in demand due to the increasing use of Machine Learning, as this is one of the most commonly used algorithms. Its applications aren't limited to specific industries or use cases, making it a commonly used and flexible analytics technique compared to other analytics methods.	5+2
3b)	**Explanation & examples** Expl – 4M, fig – 3M Non-linear SVM means that the boundary that the algorithm calculates doesn't have to be a straight line. The benefit is that you can capture much more complex relationships between your datapoints without having to perform difficult transformations on your own. The downside is that the training time is much longer as it's much more computationally intensive.	5+2
3) c)	See tables below	3M 3 M

3) c)

Descriptive Analytics	**Predictive Analytics**
tells you what happened in the past These are generally pre-canned reports, dashboards and MIS, operational reports etc. E.g. Profit per store, per region. Sales through various channels.	It determines what might happen in 'future'. This needs larger data set expertise and tool set. e.g. : Which channels are likely to perform better in next quarter based on past data.
is the branch of the advanced analytics which is used to make predictions about unknown future events.	is a preliminary stage of data processing that creates a summary of historical data to yield useful information.
is generally used to produce correlation, cross tabulation, frequency etcetera. These techniques are determined to find the regularities in the data and to reveal patterns. The other application of descriptive analysis is to discover the captivating subgroups in the major part of the data.	The primary objective of **predictive analysis** is to predict future results instead of current behaviour. It involves the supervised learning functions used for the

3M

Hierarchical	**Non-hierarchical**
Hierarchical Clustering involves creating clusters in a predefined order from top to bottom .	Non Hierarchical Clustering involves formation of new clusters by merging or splitting the clusters instead of following a hierarchical order.
It is considered less reliable than Non Hierarchical	It is comparatively more reliable than Hierarchical

3 M

Clustering.	Clustering.
It is considered slower than Non Hierarchical Clustering.	It is comparatively more faster than Hierarchical Clustering.
It is very problematic to apply this technique when we have data with high level of error.	It can work better then Hierarchical clustering even when error is there.

4a) Performance Evaluation Measures for Classification Models?

- Confusion Matrix
- Precision
- Recall/ Sensitivity
- Specificity
- F1-Score
- AUC & ROC Curve with examples

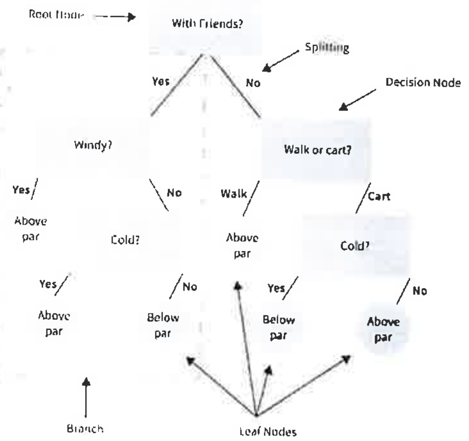
Expl. of any 4 — 2M each

~~3.5~~ 4

4b) Decision trees tend to be the method of choice for predictive modeling because they are relatively easy to understand and are also very effective. The basic goal of a decision tree is to split a population of data into smaller segments. There are two stages to prediction. The first stage is training the model—this is where the tree is built, tested, and optimized by using an existing collection of data.

Expl — 4M

Fig — 3M



~~2.5*2~~

4c) Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation. Boosting in general builds strong predictive models.

2½M each

Advantages:

- Reduces over-fitting of the model.
- Handles higher dimensionality data very well.
- Maintains accuracy for missing data.

Disadvantages:

Since final prediction is based on the mean predictions from subset trees, it won't give precise values for the classification and regression model.

Boosting is used to create a collection of predictors. **In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree.** When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly. This process converts weak learners into better performing model.

Advantages:

- Supports different loss function (we have used 'binary:logistic' for this example).
- Works well with interactions.

Disadvantages:

- Prone to over-fitting.
- Requires careful tuning of different hyper-parameters.

DESCRIPTIVE QUESTIONS PATTERN

COURSE CODE: **2052** COURSE NAME **BIG DATA ANALYTICS**

Q. No	Scheme and Solutions	Marks
5 a)		3+7

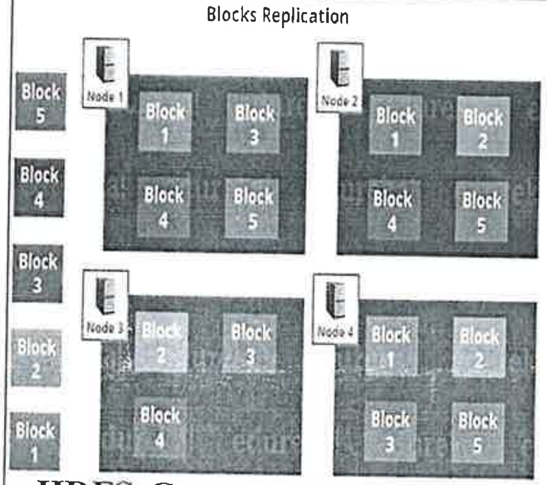
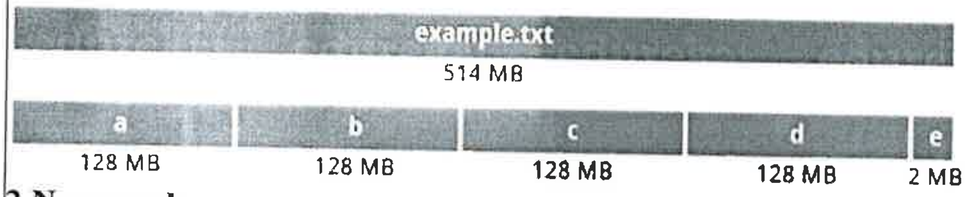


fig – 3M
Expl – 7M

HDFS Concepts

a. **Blocks**



2.Namenode

Name Node is the single point of contact for accessing files in HDFS and it determines the block ids and locations for data access. So, Name Node plays a Master role in Master/Slaves Architecture where as Data Nodes acts as slaves. File System metadata is stored on Name Node.

File System Metadata contains majorly, File names, File Permissions and locations of each block of files. Thus, Metadata is relatively small in size and fits into Main Memory of a computer machine. So, it is stored in Main Memory of Name Node to allow fast access. Tasks of HDFS NameNode (Job Tracker)

i. Manage file system namespace.
ii. Regulates client's access to files.
iii. Executes file system execution such as naming, closing, openingfiles and directories.

ii. Datanode

Tasks of HDFS DataNode(Task Tracker)

iii. DataNode performs operations like block replica creation, deletion,and replication according to the instruction of NameNode.
iv. DataNode manages data storage of the system.

Replication Management:

5b)

Pg 5

6a)

3+7

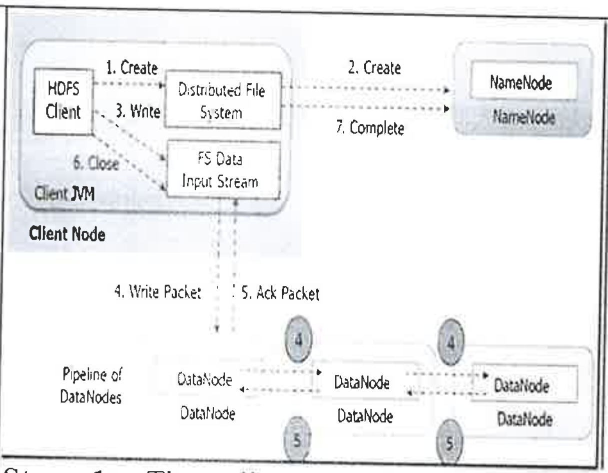


Fig – 3M

Expl. of Steps – 7M

Step 1: The client creates the file by calling create() method on DistributedFileSystem. **Step 2:** DistributedFileSystem makes an RPC call to the namenode to create a new file in the filesystem's namespace, with no blocks associated with it. The namenode performs various checks to make sure the file doesn't already exist and that the client has the right permissions to create the file. If these checks pass, the namenode makes a record of the new file; otherwise, file creation fails and the client is thrown an IOException. TheDistributedFileSystem returns an FSDataOutputStream for the client to start writing data to. **Step 3:** As the client writes data, DFSOutputStream splits it into packets, which it writes to an internal queue, called the data queue. The data queue is consumed by the DataStreamer, which is responsible for asking the namenode to allocate new blocks by picking a list of suitable datanodes to store the replicas. The list of datanodes forms a pipeline, and here we'll assume the replication level is three, so there are three nodes in the pipeline. TheDataStreamer streams the packets to the first datanode in the pipeline, which stores the packet and forwards it to the second datanode in the pipeline. **Step 4:** Similarly, the second datanode stores the packet and forwards it to the third (and last) datanode in the pipeline. **Step 5:** DFSOutputStream also maintains an internal queue of packets that are waiting to be acknowledged by datanodes, called the ack queue. A packet is removed from the ack queue only when it has been acknowledged by all the datanodes in the pipeline. **Step 6:** When the client has finished writing data, it calls close() on the stream.

5a) 5b)

~~**Sample code**~~ Map Reduce

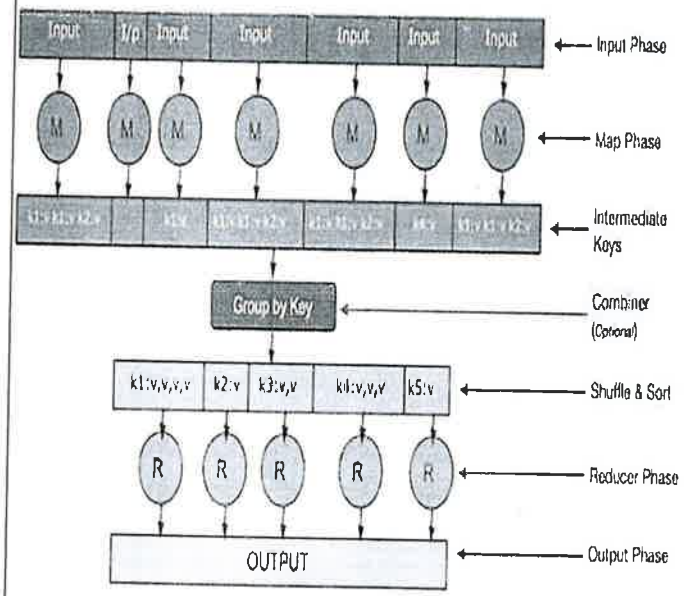


Fig – 4M

Expl – 2M

Example – 4M

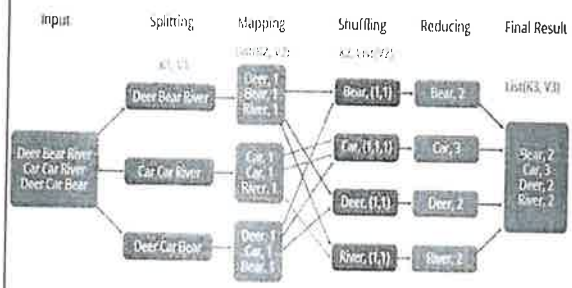
(i)Map Phase

Pg 6

(ii) Combiner Phase
(iii) Shuffle and Sort Phase
(iv) Reducer Phase

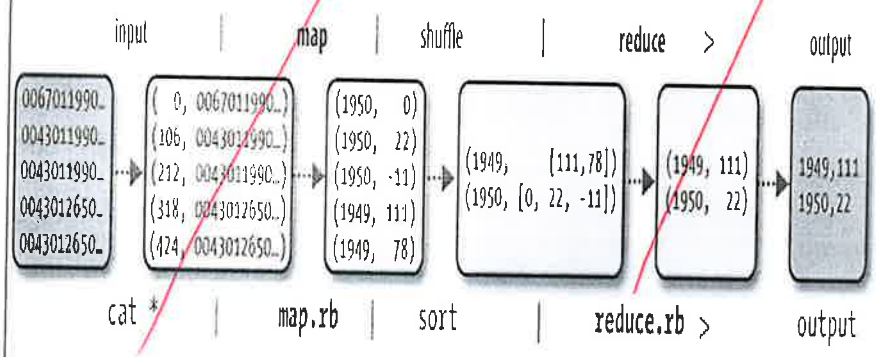
Word count examples

The Overall MapReduce Word Count Process



6b)

Logical Flow of Map Reduce data:
The input to our map phase is the raw NCDC data. We choose a text input formatthat gives us each line in the dataset as a text value. The key is the offset of the beginning of the line from the beginning of the file, but as we have no need for this, we ignore it. Our map function is simple. We pull out the year and the air temperature, since these are the only fields we are interested in. In this case, the map function is just a data preparation phase, setting up the data

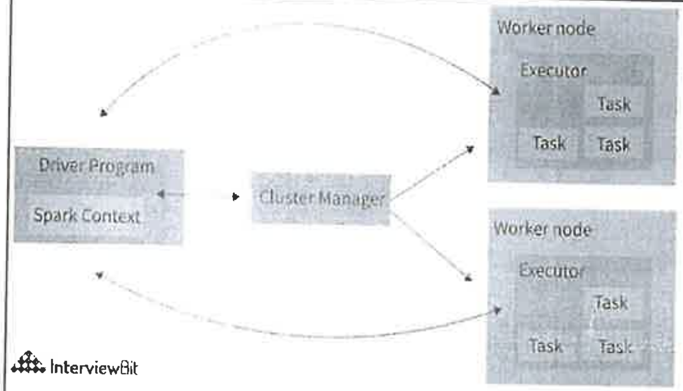
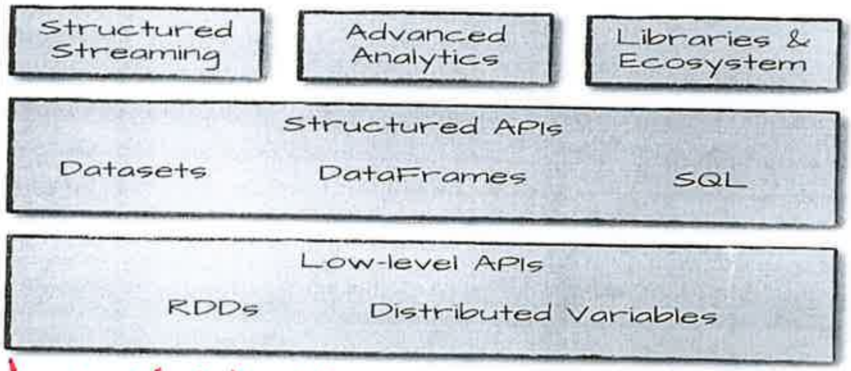


Discuss the role of Combiner function in MapReduce Processing — 10M

Combiner function - MapReduce Setup - Need - 3M
Figure - 3M
Explanation with Example - 4M

DESCRIPTIVE QUESTIONS PATTERN

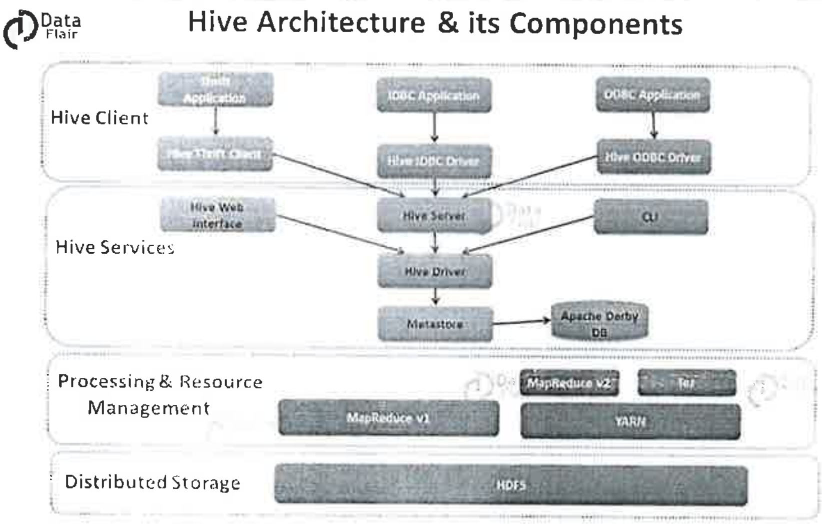
COURSE CODE: **2052** COURSE NAME **BIG DATA ANALYTICS**

Q. No	Scheme and Solutions	Marks
7 a)	 Worker node Executor Task Task Task Driver Program Spark Context Cluster Manager Worker node Executor Task Task Task InterviewBit figure – 2M Explan – 4M Lang. APIs – 4M Spark has integration with a variety of programming languages such as Scala, Java, Python, and R. Developers can write their Spark program in either of these languages. This freedom of language is also one of the reasons why Spark is popular among developers. If we compare this to Hadoop MapReduce, in MapReduce, the developers had only one choice: Java, which made it difficult for developers from another programming languages to work on MapReduce.	
7b)	Significance of dataframe, Features of dataframe Creating dataframe Dataframe operations Applications of dataframe Data frame – 5M Data Set – 5M	10M ~~2*5~~
8 a)	 Structured Streaming Advanced Analytics Libraries & Ecosystem Structured APIs Datasets DataFrames SQL Low-level APIs RDDs Distributed Variables fig – 3M Expl – 7M	
b)	Lazy Evaluation **~~Spark Transformation~~** is a function that produces new RDD from the existing RDDs. It takes RDD as input and produces one or more RDD as output. Each time it creates new RDD when we apply any transformation. Thus, the so input RDDs, cannot be changed since RDD are immutable in nature. With features	~~2+2~~ 5M
c)	Three types of structured API Dataframe, Dataset . ~~SQL tables with exa~~mples 2½M each	~~2*3~~ 5M

Pg 8

DESCRIPTIVE QUESTIONS PATTERN

COURSE CODE: **2052** COURSE NAME **BIG DATA ANALYTICS**

Q. No	Scheme and Solutions	Marks
9 a)	 Data Flair Hive Architecture & its Components Hive Client Thrift Application JDBC Application ODBC Application Hive Thrift Client Hive JDBC Driver Hive ODBC Driver Hive Services Hive Web Interface Hive Server CLI Hive Driver Metastore Apache Derby DB Processing & Resource Management MapReduce v2 Tez MapReduce v1 YARN Distributed Storage HDFS fig – ~~4~~ 3 M Exp – ~~5~~ 3 M Insert/Load – 4M Using insert statement Load data statements	~~6~~ ~~4~~
b)	DDL commands — 4M CREATE, ALTER,DROP, DML Commands HIVEQL DML — 6 M SELECT,LOAD, GROUP BY , ORDER BY, JOIN	~~[illegible]~~
10 a)	Primitive data types, date time, string types, Misc types, complex data types with examples	2*4
10 b)	i) CREATE TABLE IF NOT EXISTS Movie (Movieid int, MovieName String, MovieYear String, Language string, Actor string) ii) SELECT MovieName FROM movie WHERE MovieYear=2020 iii) SELECT language ,count(*) from Movie group by languages iv) create view stdmv as select * from Movie where language = kannada v) Select MovieName from movie where MovieYear=" "	2*4
c)	The component that stores all the structure information of the various tables and partitions in the warehouse including column and column type information, the serializers and deserializers necessary to read and write data and the corresponding HDFS files where the data is stored.	~~2+2~~ – 4M