12 APR 2024

22MCA2052

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

Q. No	Module - 1	Marks	CO & RBT
1a.	Present an overview of Big Data? Explain Analytics Process model.	10	CO1, K2
b.	Big Data is characterised by various Vs. Describe *4 Vs* of Big Data.	5	CO1, K2
c.	Calculate Z scores and detect the outlier for the following data. Mu=40, SD=10. Data = 30, 50, 10, 40, 60, 80.	5	CO1, K2
	OR		
2a.	Discuss the *applications* of Big Data Analytics in marketing, risk management, government, web and logistics.	10	CO1, K2
b.	Explain the *requirements* to satisfy a good analytical model.	5	CO1, K2
c.	Construct *Box plot* for given data: 51,17,25,39,7,49,67,41,20,2,43,13.	5	CO1, K2
	Module – 2		
3a.	Describe the scenarios in which predictive and descriptive analytics can be used.	10	CO2, K2
b.	Explain *association rules* as applied in descriptive analytics.	5	CO2, K2
c.	Bring out the relevance of *decision trees* in data analytics with an example.	5	CO2, K2
	OR		
4a.	Elaborate on any 2 techniques used for *classification* under predictive analytics.	10	CO2, K2
b.	Describe *k-means* clustering with example.	5	CO2, K2
c.	Compare and contrast *bagging and boosting*.	5	CO2, K2
	Module – 3		
5a.	Explain HDFS architecture with a neat diagram.	10	CO3, K2
b.	Explain with a diagram, MapReduce data flow with a single reduce task and multiple reduce task.	10	CO3, K2
	OR		

6a.	With a neat diagram, explain the anatomy of reading data from a file in HDFS.	**10**	**CO3, K2**
b.	What is Map Reduce? Sketch a neat diagram and explain the logical data flow in Map Reduce?	**10**	**CO3, K2**
	Module – 4		
7a.	What is the role of Apache Spark? With suitable diagram illustrate the architecture of a Spark Application.	**10**	**CO4, K2**
b.	Describe Spark's MLLib.	**5**	**CO4, K2**
c.	Discuss the difference between a single machine and dataframe data distributed over a cluster.	**5**	**CO4, K2**
	OR		
8a.	Describe Structured API logical planning process and physical planning process.	**10**	**CO4, K2**
b.	Discuss datasets of Spark.	**5**	**CO4, K2**
c.	What is Lazy Evaluation? Compare and contrast between Transformations and Actions.	**5**	**CO4, K2**
	Module – 5		
9a.	Illustrate the Hive modules in the Hadoop Ecosystem.	**10**	**CO5, K2**
b.	Discuss how data is inserted/loaded into Hive tables.	**5**	**CO5, K2**
c.	Describe any 3 key data types supported by Hive.	**5**	**CO5, K2**
	OR		
10a.	Illustrate application of Aggregate functions, Mathematical functions and Table generating functions.	**10**	**CO5, K2**
b.	Discuss how dynamic partition inserts are accomplished in hive.	**5**	**CO5, K2**
c.	Discuss how internal and External Tables are created in Hive.	**5**	**CO5, K2**

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

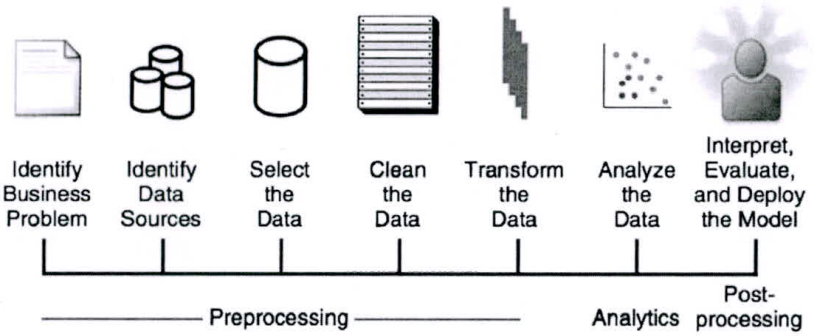
COs	Statements
CO-1	Identify the business problem for a given context and frame the objectives to solve it using data analytics tools.
CO-2	Differentiate various types of analytics algorithms and context of their application.
CO-3	Illustrate the architecture of HDFS and MapReduce.
CO-4	Explore Spark 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	

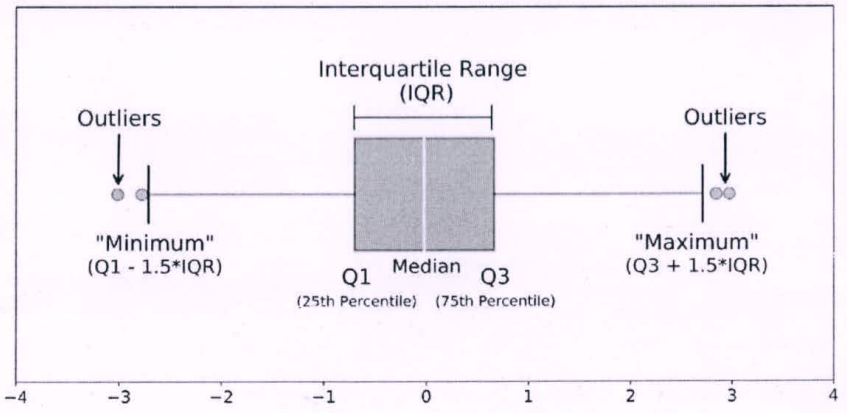
"Success is the progressive realization of a worthy goal."

****

DESCRIPTIVE QUESTIONS ONLY

Course Code: 22MCA2052 **Course Name: Big Data Analytics**

Module - 1		
Q.No.	**Questions**	**Marks**
1a.	**Overview of Big Data – 2 M** **Analytical Process Model Diagram – 2M** **Explanation of Analytical Process Model – 6 M** **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big Data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered. Overview of the Analytics Process Model  **Steps involved in Analytics Process Model:** 1. Define the business problems to be solved 2. All source-data need to be identified 3. All data to be gathered as a pool 4. Data cleaning 5. Analytical model estimation 6. Model interpretation and evaluation	10
1b.	**Mention of 4 to 6 Vs – 1 M - Volume, Velocity, Value, Variability, Veracity, Variety** **Explanation of any 4 Vs – 4 M**	5
1c.	Calculation of Z score and detection of the outlier for the following data. **Mu=40, SD=10. Data = 30, 50, 10, 40, 60, 80.** Z Score with +4 is an outlier in the given data. That means the value corresponding to +4 is 80 is an outlier. **Calculation – 3 M** **Final Discussion – 2M** | ID	Age	Z-Score | |---|---|---| | 1 | 30 | (30 – 40)/10 = –1 | | 2 | 50 | (50 – 40)/10 = +1 | | 3 | 10 | (10 – 40)/10 = –3 | | 4 | 40 | (40 – 40)/10 = 0 | | 5 | 60 | (60 – 40)/10 = +2 | | 6 | 80 | (80 – 40)/10 = +4 | | … | … | … | | | $\mu = 40$ $\sigma = 10$ | $\mu = 0$ $\sigma = 1$ |	5

2a.	**Big Data Applications – Any Five domains - 2 M Each** **Marketing:** Response Modelling, Net Lift Modelling, Retention Modelling, Market Basket Analysis, Recommender Systems, Customer Segmentation **Risk management:** Credit Risk Modelling, Market Risk Modelling, Operational Risk Modelling, Fraud Detection. **Government:** Tax Avoidance, Social Security Fraud, Money Laundering, Terrorism Detection **Web:** Web Analytics, Social Media Analytics, Multivariate Testing. **Logistics:** Demand Forecasting, Supply Chain Analytics. **Others:** Text Analytics, Business Process Analytics, Sentiment Analytics	10
2b.	**Requirements to satisfy a good analytical model – Any 5 – 1 M Each** Business relevance Statistical performance Interpretability Justifiability Operationally efficient Economical Local and international regulation and legislation	5
2c.	***Box plot*** for given data: **51,17,25,39,7,49,67,41,20,2,43,13** **Box Plot Construction – 3 M** **Q1, Q2, Q3, IQR Values computation – 2 M**  	5

DESCRIPTIVE QUESTIONS ONLY

Course Code: 22MCA2052 **Course Name: Big Data Analytics**

3a.	**Scenarios for Predictive Analytics Application. – 2 M Each – Any Five** **Any 5 use cases / scenarios like healthcare, marketing, supply chain, logistics, web analytics etc.**	10
3b.	**Association Rules: 5 M** • Association rules are used to find correlations and co-occurrences between data sets. • They are ideally used to explain patterns in data from seemingly independent information repositories, such as relational databases and transactional databases. • It is employed in Market Basket analysis, Web usage mining etc. • Association rules typically start from a database of transactions. • Each transaction consists of a transaction identifier and a set of items. • Association rules are usually represented in the form $X \rightarrow Y$, where X (also called rule Antecedent) and Y (also called Consequent) are disjoint item sets (i.e., disjoint conjunctions of features). • Ex: If a customer buys bread, he's 70% likely of buying milk." In the above association rule, bread is the antecedent and milk is the consequent. • Association rule shows how frequently an item set occurs in a transaction. • Association rules are created by analysing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships.	5
3c.	**Decision Trees – Discussion – 3 M** **Example Diagram – 2 M**	5
4a.	**Any 2 techniques from Logistic Regression, Decision Trees, Neural Networks, SVM, Ensemble Methods (Bagging, Boosting, Random Forest).** **Each 5 M with Explanation and Example Diagram**	10
4b.	**k-means Clustering** **Discussion – 3 M** **Example – 2 M**	5

4c. Bagging vs Boosting – Any 5 Parameters – 1 M Each — 5

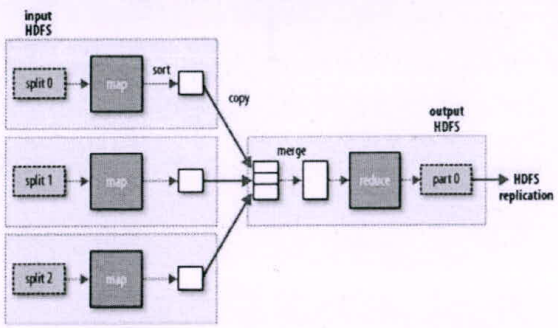
Bagging	Boosting
The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
Each model receives equal weight.	Models are weighted according to their performance.
Each model is built independently.	New models are influenced by the performance of previously built models.
Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.
If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
In bagging base classifiers are trained parallelly.	In boosting base classifiers are trained sequentially.
Aims at decreasing variance (Overfitting leads to high variance – Overfitting means that the model performs well on the training data but does not perform accurately in the evaluation set.)	Aims at decreasing bias (Bias is a phenomenon that skews the result of an algorithm either in favour or against an idea).
Example: The **Random forest** model uses Bagging.	**Example:** The **AdaBoost** algorithm uses Boosting.

5a. **Hadoop Architecture – Diagram – 4 M** — 10

Discussion on HDFS, YARN, MapReduce – 2 M each

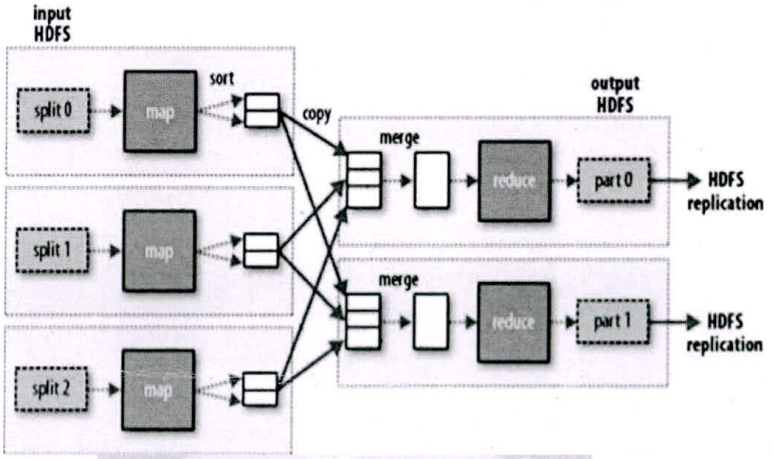
5b. MapReduce processing involves multiple mappers in general. However, when it comes to reducer, it may have one in general or increase it carefully as the need be but not in more numbers. Elaborate the MapReduce data flow with a no reduce task, single reduce task and multiple reduce tasks. — 10

Single Reduce Task – Diagram, Discussion – 5M
Multiple Reduce Tasks – Diagram, Discussion – 5 M



DESCRIPTIVE QUESTIONS ONLY

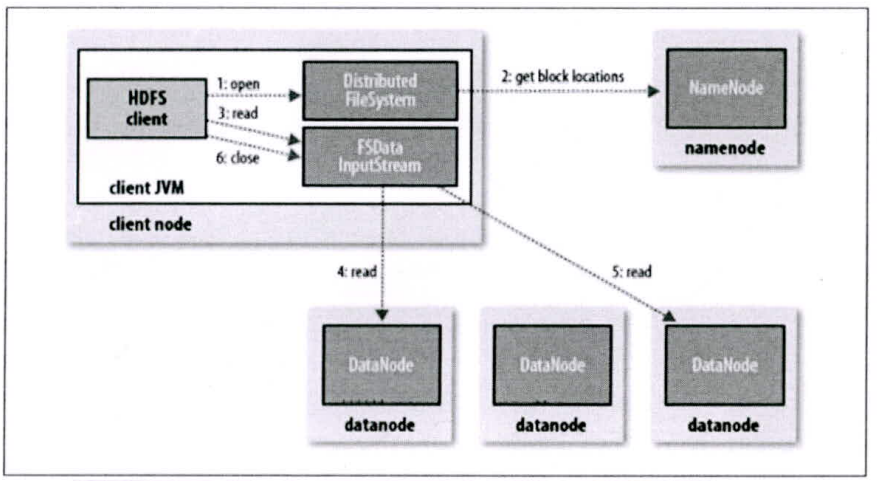
Course Code: 22MCA2052 **Course Name: Big Data Analytics**



6a. **Anatomy of reading Data from a file in HDFS.** 10

Diagram – 4 M

Explanation of Steps involved – 6 M



6b. **MapReduce Logical DataFlow** 10

Diagram – 4 M

Explanation with Example – 6 M

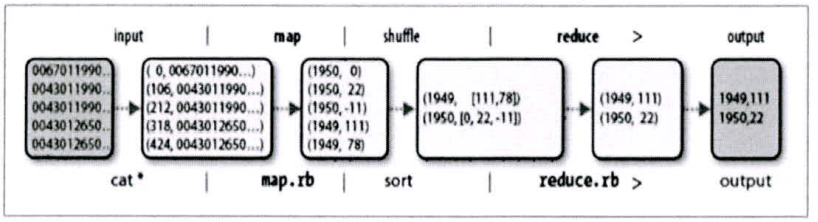
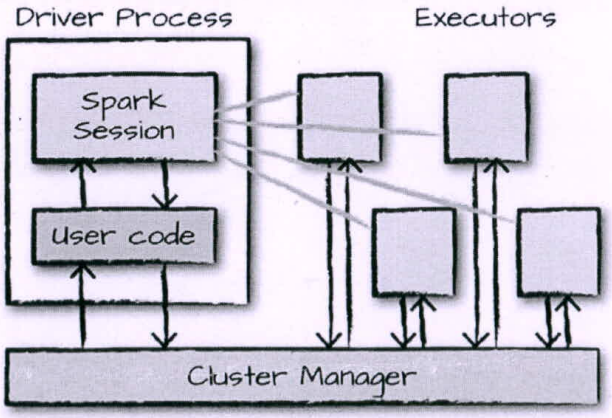
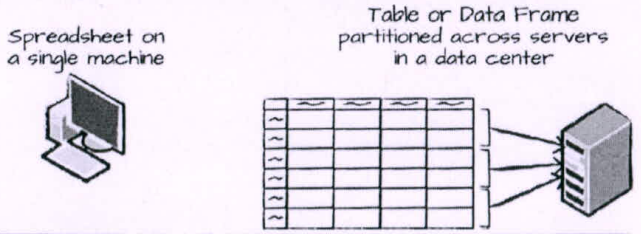
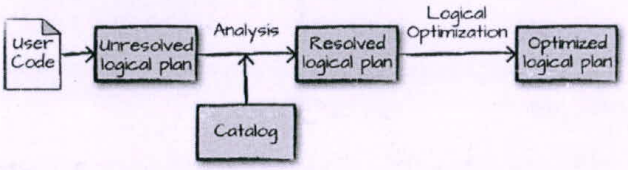
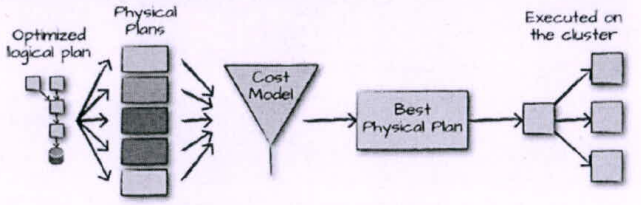
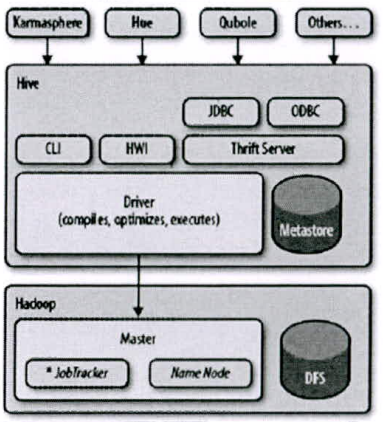


Figure 2-1. MapReduce logical data flow

7a.	**About Apache Spark – 2 M** **Architecture of Spark – Diagram – 3 M** **Explanation of Spark's Architecture – 5 M** 	10
7b.	**Note on Spark MLib – 5 M**	5
7c.	Difference between a single machine and dataframe data distributed over a cluster. **Diagram – 2 M** **Discussion – 3 M** 	5
8a.	**Structured API Logical Planning Process – Diagram, Explanation - 5 M**  **Structured API Physical Planning Process – Diagram, Explanation - 5 M** 	10

8b.	**Note on Datasets – Type-safe Structured APIs**	5

8c.	**Lazy Evaluation – 1 M** Lazy evaluation means that Spark will wait until the very last moment to execute the graph of computation instructions. In Spark, instead of modifying the data immediately when you express some operation, you build up a *plan* of transformations that you would like to apply to your source data. By waiting until the last minute to execute the code, Spark compiles this plan from your raw DataFrame transformations to a streamlined physical plan that will run as efficiently as possible across the cluster. **Difference between TRANSFORMATIONS and ACTIONS – 2 M Each** Transformations allow us to build up our logical transformation plan. To trigger the computation, we run an action. An action instructs Spark to compute a result from a series of transformations.	5

9a.	**Hive Modules in Hadoop Ecosystem** **Diagram – 4 M** **Explanation – 6M** 	10

9b.	**Usage of CASE…WHEN…THEN Statements.** The CASE … WHEN … THEN clauses are like if statements for individual columns in query results. **Explanation – 2 M** **Example – 3 M** Data insertion/loading in (2) ways 2½ Mark ``` hive> SELECT name, salary, > CASE > WHEN salary < 50000.0 THEN 'low' > WHEN salary >= 50000.0 AND salary < 70000.0 THEN 'middle' > WHEN salary >= 70000.0 AND salary < 100000.0 THEN 'high' > ELSE 'very high' > END AS bracket FROM employees; John Doe 100000.0 very high Mary Smith 80000.0 high Todd Jones 70000.0 high Bill King 60000.0 middle Boss Man 200000.0 very high Fred Finance 150000.0 very high Stacy Accountant 60000.0 middle ... ```	5

Course Code: 22MCA2052 **Course Name: Big Data Analytics**

9c.	**Partitioned-Managed Tables of Hive.** **Explanation – 3 M** **Example – 2 M** 3 Key data types Mention – 1M Explanations 2–4M.	5
10a.	**Aggregate functions, Mathematical functions and Table generating functions.** **Write up on these functions - 1 M** **Discussion of any three from each group – 3 M Each**	10
10b.	**Dynamic Partition in Hive - Hive supports a dynamic partition feature, where it can infer the partitions to create based on query parameters.** **Explanation – 3 M** **Example – 2 M**	5
10c.	**Internal and External Tables are created in Hive.** **Internal Tables/Managed Tables – Explanation, Example – 2 ½ M** **External Tables – Explanation, Example – 2 ½ M** **********	5