CBCS SCHEME

18CS72

USN | | | | | | | | 0 | 4 | 3 |

Seventh Semester B.E. Degree Examination, Dec.2023/Jan.2024
Big Data and Analytics

Time: 3 hrs. Max. Marks: 100

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

Module-1

1 a. How is Data Architecture layers used for analytics? Explain with functions of each layer. **(10 Marks)**

b. Briefly describe the three fundamental services offered by Cloud Computing. **(10 Marks)**

OR

2 a. *List the features of Grid Computing. How does it differ from clusters and cloud computing.* **(10 Marks)**

b. *Why is Data quality important in discovering new knowledge and decision making?* **(10 Marks)**

Module-2

3 a. *List Hadoop core components and explain with appropriate diagram.* **(10 Marks)**

b. *Explain the working of the Hadoop Map Reduce frame work.* **(10 Marks)**

OR

4 a. Explain the working of Hadoop – 2 Execution model (YARN Model) **(10 Marks)**

b. With a diagram, explain the concept of APACHE Sqoop to acquire relational data. ***(10 Marks)***

Module-3

5 a. Define NOSQL. Explain Big Data NOSQL or Not – only SQL with its features, transactions and solutions. **(10 Marks)**

b. Describe graph database characteristic, typical used and examples. **(10 Marks)**

OR

6 a. Explain Mongo DB with its features. **(10 Marks)**

b. Compare and contrast RDBMS and Mongo DB databases. **(05 Marks)**

c. What are the different ways of handling Big Data Problems? **(05 Marks)**

Module-4

7 a. Describe the Hive architecture components along with Hive Built – in functions. **(10 Marks)**

b. Explain with respect to Hive QL :
i) Hive QL Data Definition Language (DDL).
ii) Hive QL Data Manipulation Language (DML). **(10 Marks)**

OR

8 a. Explain the architecture, feature and applications of PIG. **(10 Marks)**

b. Illustrate by considering an example the working of the Map Reduce programming model. **(10 Marks)**

Important Note : 1. On completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages.
2. Any revealing of identification, appeal to evaluator and /or equations written eg, 42+8 = 50, will be treated as malpractice.

Module-5

9 a. How does regression analysis predict the value of the dependent variable in case of linear regression? (10 Marks)

b. Explain with example and algorithm, the working principle of Apriori process for adopting the subset of frequent item sets as a frequent itemset. (10 Marks)

OR

10 a. Define Web Mining. Discuss the broad classification of web mining and their applications. (10 Marks)

b. Define the term Social network. Explain social network as graphs with Centralities , Ranking and Anomaly Detection. (10 Marks)

* * * * *

VTU Examination – Dec- 23/Jan- 24
Solution

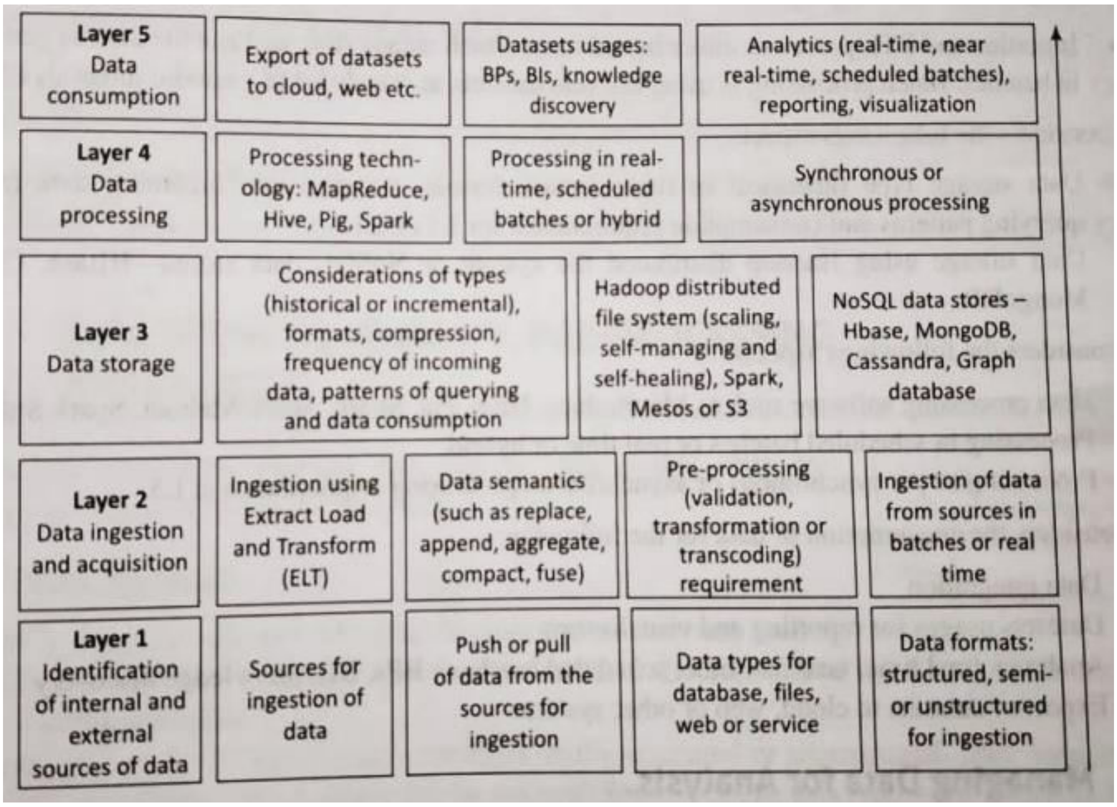
Sub:	**Big Data and Analytics**				Sub Code:	**18CS72**	Branch:	**ISE**
Exam Date:	**31/01/2024**	Duration:	**3 Hrs**	Max Marks:	**100**	Sem	**VII**	

Answer any FIVE FULL Questions — MARKS — CO

MODULE-1

1 (a) a. How is Data Architecture layers used for analytics? Explain with functions of each layer — [10] — CO1

Solution:



Layer 5 Data consumption	Export of datasets to cloud, web etc.	Datasets usages: BPs, BIs, knowledge discovery	Analytics (real-time, near real-time, scheduled batches), reporting, visualization		
Layer 4 Data processing	Processing techn-ology: MapReduce, Hive, Pig, Spark	Processing in real-time, scheduled batches or hybrid	Synchronous or asynchronous processing		
Layer 3 Data storage	Considerations of types (historical or incremental), formats, compression, frequency of incoming data, patterns of querying and data consumption	Hadoop distributed file system (scaling, self-managing and self-healing), Spark, Mesos or S3	NoSQL data stores – Hbase, MongoDB, Cassandra, Graph database		
Layer 2 Data ingestion and acquisition	Ingestion using Extract Load and Transform (ELT)	Data semantics (such as replace, append, aggregate, compact, fuse)	Pre-processing (validation, transformation or transcoding) requirement	Ingestion of data from sources in batches or real time	
Layer 1 Identification of internal and external sources of data	Sources for ingestion of data	Push or pull of data from the sources for ingestion	Data types for database, files, web or service	Data formats: structured, semi- or unstructured for ingestion	

Fig. w.r.t. logical layers in a data architecture — 5M
Functions in detail — 5M

(b) **Briefly describe the three fundamental services offered by Cloud Computing.** — [10] — CO1

Solution:

Cloud services can be classified into three fundamental types i) Infrastructure as a Service(IaaS) ii) Platform as a Service(PaaS) iii) Software as a Service(SaaS)

Cloud computing — Definition
— features — 4M

Fundamental types to be expained in detail including examples — IaaS — 2M
— PaaS — 2M
— SaaS — 2M

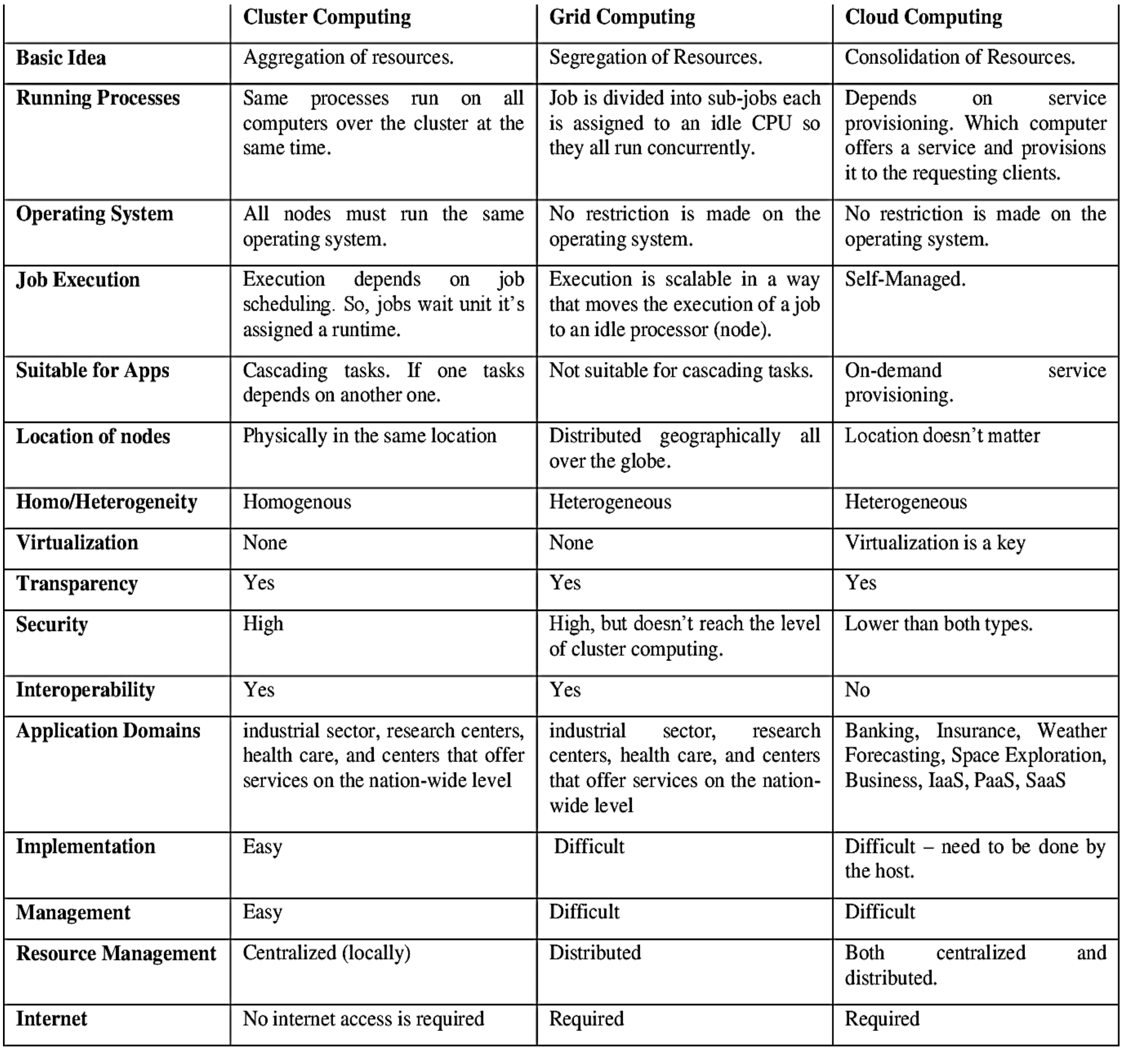
OR

2 (a) **List the features of Grid Computing. How does it differ from clusters and cloud computing.** [10] CO1

Solution:

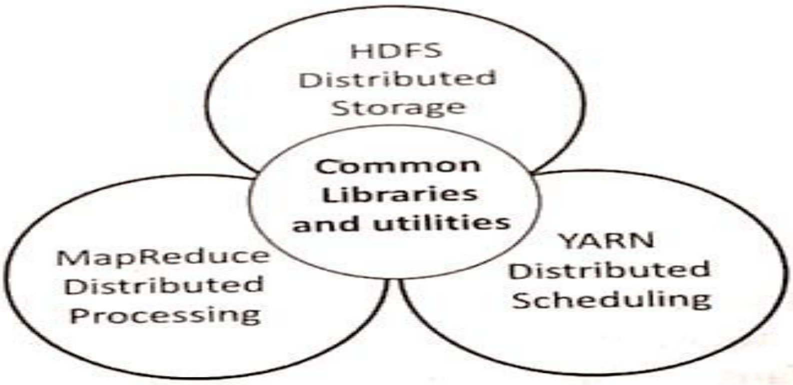
Features of Grid Computing

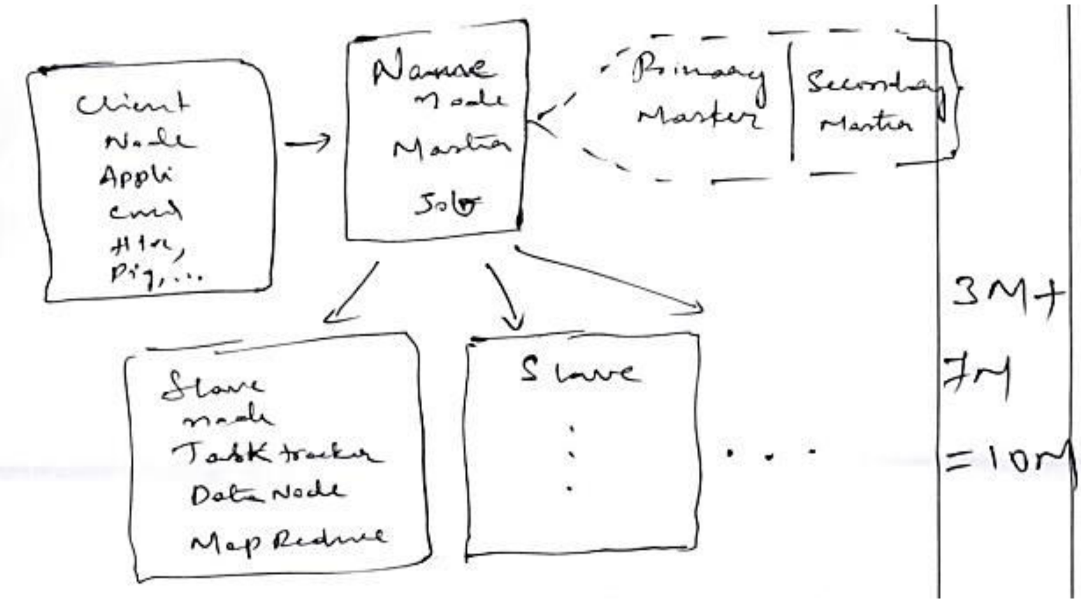
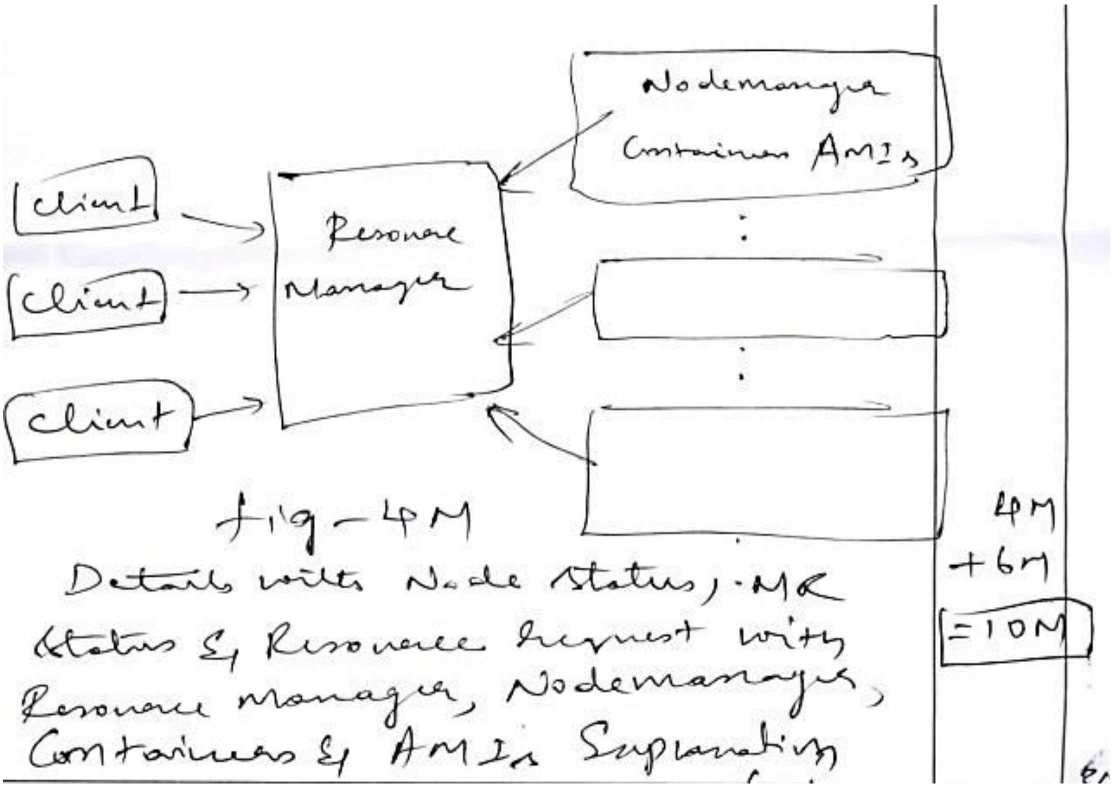
- Grid computing is similar to cloud computing ,it is scalable.
- Cloud computing depends on sharing of resources (for eg , networks , servers ,storage , applications and services)to attain coordination and coherence among resources similar to grid computing.
- Grid also forms a distributed network for resource integration. **4 Marks**

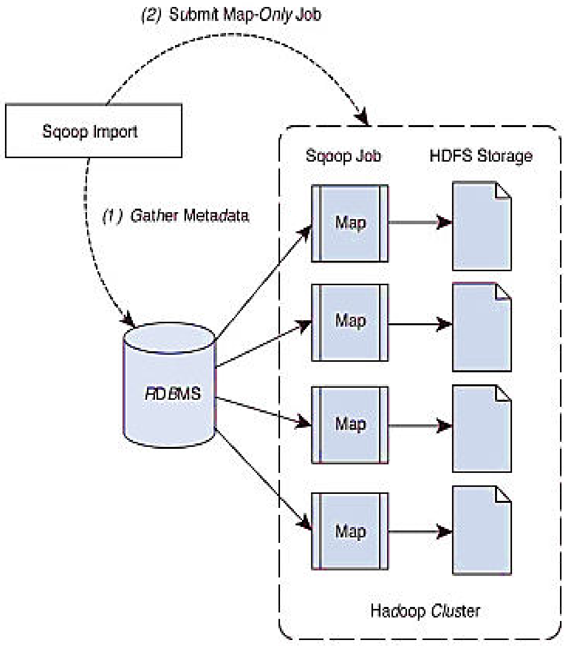
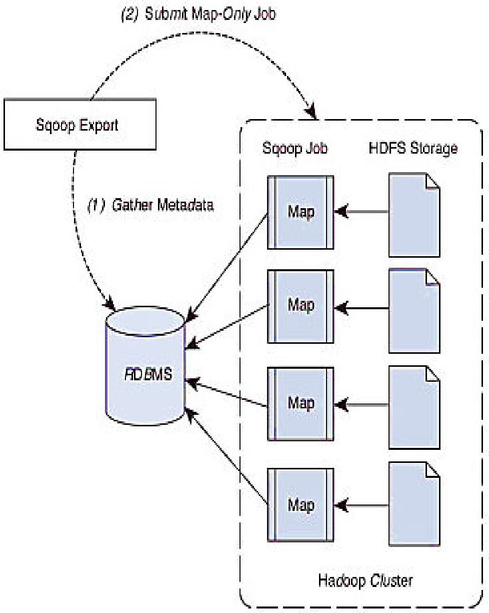


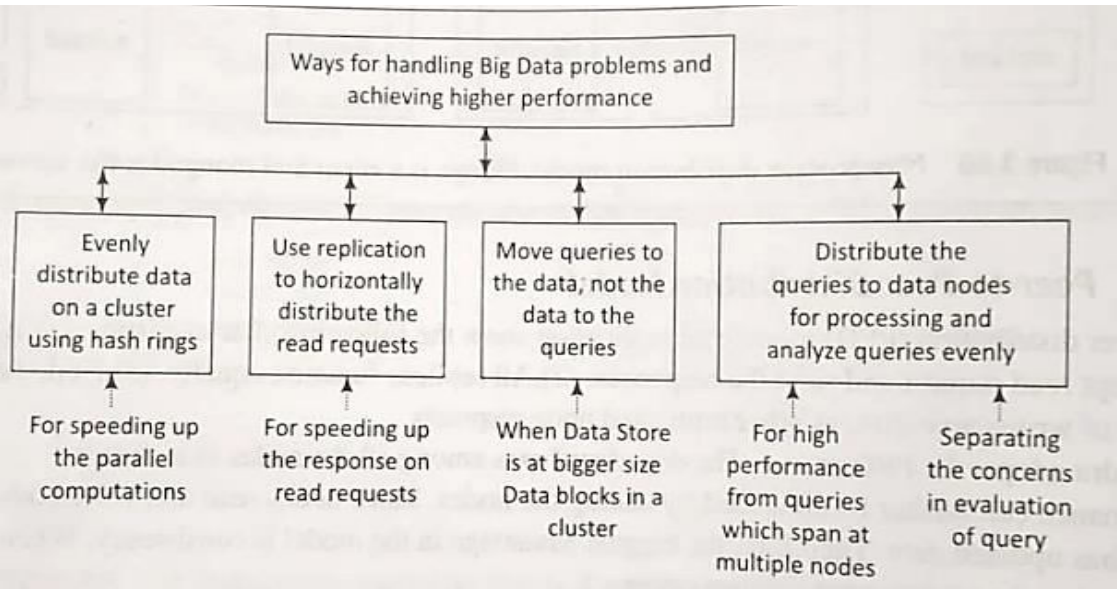
	Cluster Computing	Grid Computing	Cloud Computing
Basic Idea	Aggregation of resources.	Segregation of Resources.	Consolidation of Resources.
Running Processes	Same processes run on all computers over the cluster at the same time.	Job is divided into sub-jobs each is assigned to an idle CPU so they all run concurrently.	Depends on service provisioning. Which computer offers a service and provisions it to the requesting clients.
Operating System	All nodes must run the same operating system.	No restriction is made on the operating system.	No restriction is made on the operating system.
Job Execution	Execution depends on job scheduling. So, jobs wait unit it's assigned a runtime.	Execution is scalable in a way that moves the execution of a job to an idle processor (node).	Self-Managed.
Suitable for Apps	Cascading tasks. If one tasks depends on another one.	Not suitable for cascading tasks.	On-demand service provisioning.
Location of nodes	Physically in the same location	Distributed geographically all over the globe.	Location doesn't matter
Homo/Heterogeneity	Homogenous	Heterogeneous	Heterogeneous
Virtualization	None	None	Virtualization is a key
Transparency	Yes	Yes	Yes
Security	High	High, but doesn't reach the level of cluster computing.	Lower than both types.
Interoperability	Yes	Yes	No
Application Domains	industrial sector, research centers, health care, and centers that offer services on the nation-wide level	industrial sector, research centers, health care, and centers that offer services on the nation-wide level	Banking, Insurance, Weather Forecasting, Space Exploration, Business, IaaS, PaaS, SaaS
Implementation	Easy	Difficult	Difficult – need to be done by the host.
Management	Easy	Difficult	Difficult
Resource Management	Centralized (locally)	Distributed	Both centralized and distributed.
Internet	No internet access is required	Required	Required

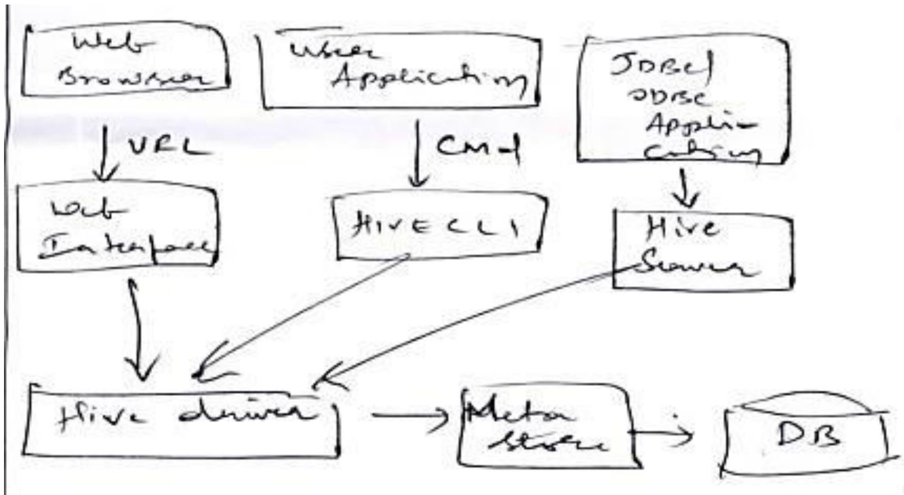
6 Marks

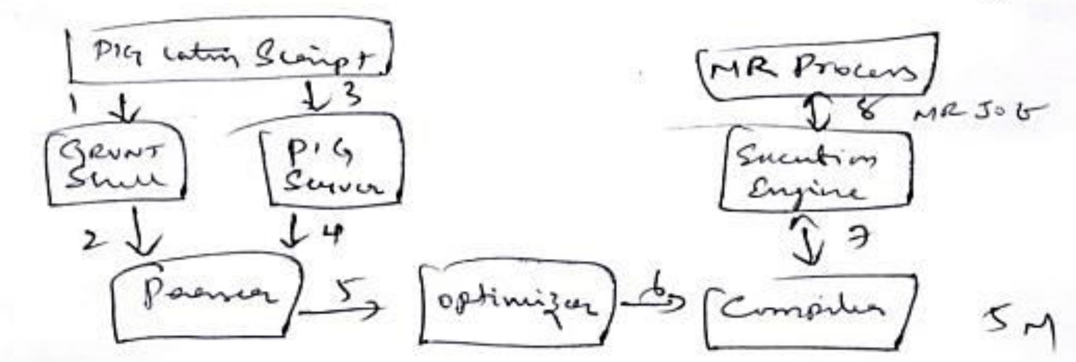
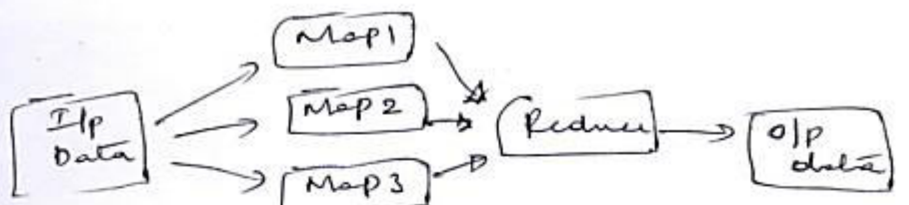
(b)	**Why is Data quality important in discovering new knowledge and decision making?** **Solution:** Defination of Data quality — 2M Five R's Relevancy Recency Range Robustness Reliability Removing Data Noise, outliers, Missing & Duplicate values. with detailed explanation & Justification for knowledge & Decision making — 8M	[10]	CO1
	MODULE-2		
3 (a)	**List Hadoop Core Components and explain with appropriate diagram.** **Solution:**  HDFS Distributed Storage Common Libraries and utilities MapReduce Distributed Processing YARN Distributed Scheduling **Diagram + Explanation – 2+8 marks**	[10]	CO2
(b)	**Explain the working of the Hadoop MapReduce Framework.** **Solution:** The client, Master, Name node & Slave nodes with detailed explanation fig — 3M Explanation - 7M	[10]	CO2

	 Client Node Appli cmd Hive, Pig,... Name Node Master Job Primary Master Secondary Master Slave node Task tracker Data Node MapReduce Slave 3M+ 7M =10M		
	OR		
4 (a)	**Explain the working of Hadoop-2 Execution model(YARN model)** **Solution:**  Nodemanager Containers AMIs client client client Resource Manager fig-4M 4M +6M =10M Details with Node Status, MR Status & Resource request with Resource manager, Nodemanager, Containers & AMIs Explanation	[10]	CO2
(b)	**With a diagram, explain the concept of APACHE sqoop to acquire relational data.** **Solution:** Diagram – 4 Marks Explanation – 6 Marks	[10]	CO2

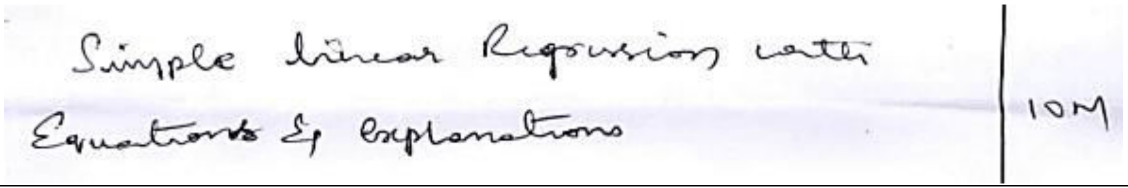
	 (2) Submit Map-Only Job Sqoop Import (1) Gather Metadata RDBMS Sqoop Job HDFS Storage Map Map Map Map Hadoop Cluster  (2) Submit Map-Only Job Sqoop Export (1) Gather Metadata RDBMS Sqoop Job HDFS Storage Map Map Map Map Hadoop Cluster Fig: Two-step Apache Sqoop data import method. Fig: Two-step Sqoop data export method.		
	MODULE-3		
5 (a)	**Define NOSQL. Explain Big Data NoSQL or Not-only SQL with its features, transactions and solutions.** **Solution:** Definition – 2M Features: Relax of ACID properties, Two properties of CAP theorem BASE flexible model — ~~1~~ 4 M Solutions: Apache's HBase, MongoDB, Cassandra, CouchDB, Oracle & Riak with uses. —— 4 M	[10]	CO3
(b)	**Describe graph database characteristic, typical uses and examples.** **Solution:** Characteristics: Specialized query lang, (RDF), different models, hyper-edges, Joins, ... — 4 M uses: link analysis, friend of friend queries, Rules & Inference, Rule Induction, pattern matching — 4 M Examples: Neo4J, Allegrograph, Hypergraph, Infinite, Titan & Flock DB. — 2 M	[10]	CO3

	OR		
6 (a)	**Explain MongoDB with the features.** **Solution:** MongoDB is Non-relational, NoSQL, Distributed, open source, document based, Cross platform, Scalable, flexible, Indexed, Multimaster, fault tolerant. Features: Physical Containers for Collections Collection stores, well defined, JSON style documents, BSON Serialization format, Efficient, Deep query ability, No Complex Joints, Distributed DB, Indexed, Atomic operations, Fast in place updates, No Configurable cache, conversion/mapping of application objects, ... 4M + 6M = 10M	[10]	CO3
(b)	**Compare and Contrast RDBMS and Mongo DB databases.** **Solution:** Comparing with Features like Data Model, Schema, Typed Data, locality, updates, Transactions, Auditing & Scaling 5M	[05]	CO3
(c)	**What are the different ways of handling Big Data Problems?** **Solution: Diagram+ Explanation – 2+3 Marks**  Ways for handling Big Data problems and achieving higher performance Evenly distribute data on a cluster using hash rings Use replication to horizontally distribute the read requests Move queries to the data, not the data to the queries Distribute the queries to data nodes for processing and analyze queries evenly For speeding up the parallel computations For speeding up the response on read requests When Data Store is at bigger size Data blocks in a cluster For high performance from queries which span at multiple nodes Separating the concerns in evaluation of query	[05]	CO3

	MODULE-4		
7 (a)	**Describe the Hive architecture components along with Hive Built-in functions.** **Solution:**  Web Browser User Application JDBC/ ODBC Application URL CMd Web Interface HIVE CLI Hive Server Hive driver Meta Store DB —6M Suplanation —4M	[10]	CO4
(b)	**Explain with respect to HiveQL.** **i) Hive QL Data Definition Language(DDL)** **ii) Hive QL Data Manipulation Language (DML).** **Solution:** i) a) CREATE DB. b) SHOW DB c) CREATE Schema d) CREATE TABLE } HIVE QL DDL Commands 5M ii) a) DROP DB b) DROP Sirema c) ALTERi TABLE d) DROP table e) LOAD Data } Hive QL DML Commands 5M	[10]	CO4
	OR		

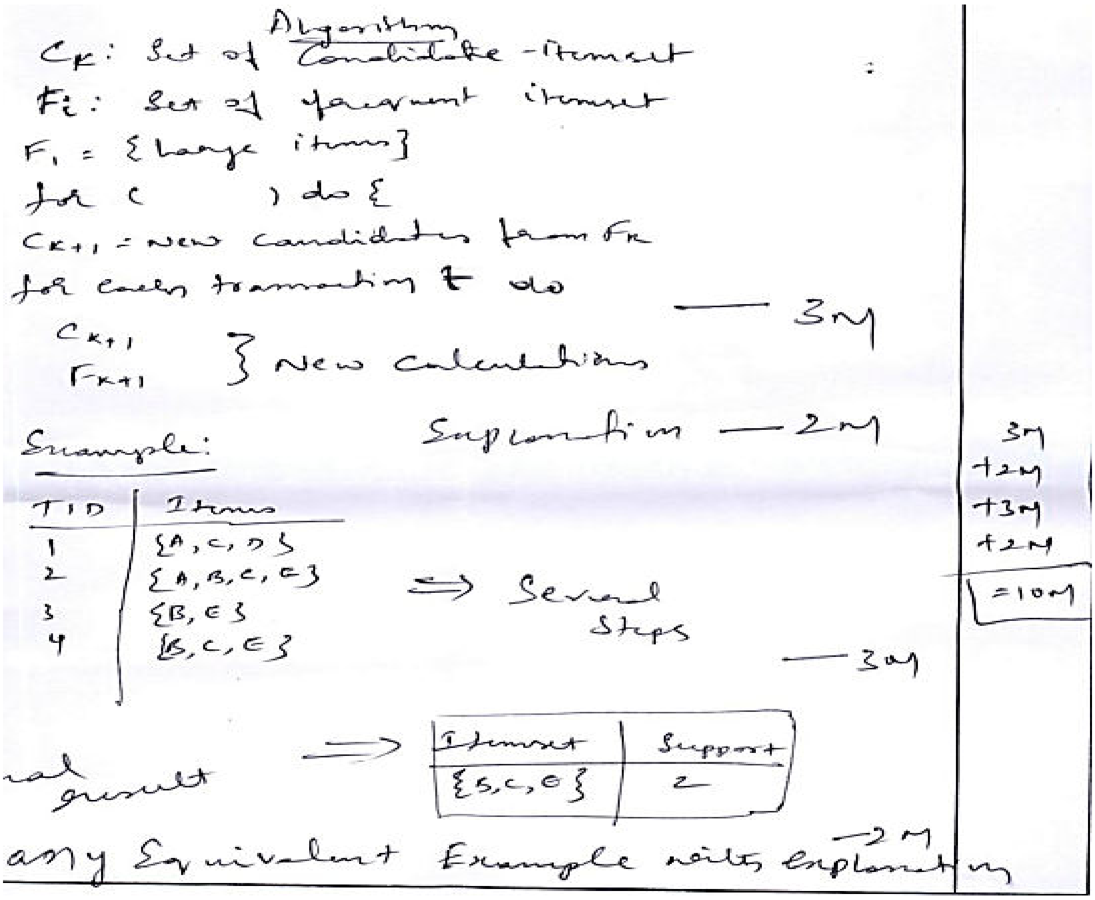
8 (a)	**Explain the architecture, feature and applications of PIG.** **Solution:**  PIG Latin Script 1 3 Grunt Shell PIG Server 2 4 Parser 5 optimizer 6 Compiler 7 Execution Engine 8 MR Job MR Process 5M Architecture Features: Scripts, UDFs, Variety of data, multiquery approach, Inconsistent schema, ETL, auto optimization, deals with whole operations, Fewer codes, live anywhere, take anything, run as if flying. −3M applications: Analysing large datasets, adhoc processing, large data sources, supports different data types, processing time sensitive data loads. − 2M	[10]	CO4
(b)	**Illustrate by considering an example the working of the Map Reduce Programming model.** **Solution:**  — 4M Diagram with detailed explanation −6M	[10]	CO4
	MODULE-5		

9 (a)	**How does regression analysis predict the value of the dependent variable in case of linear regression?** **Solution:**	[10]	CO5



Simple linear Regression with Equations & explanations | 10M

(b)	**Explain with example and algorithm, the working principle of Apriori process for adopting the subset of frequent item sets as a frequent Itemset.** **Solution:**	[10]	CO5



Algorithm

Ck: Set of Candidate-Itemset

Fi: Set of frequent itemset

F1 = {large items}

for ( ) do {

Ck+1 = new candidates from Fk

for each transaction t do — 3M

Ck+1 } New calculations
Fk+1

Example: Explanation — 2M

TID	Items
1	{A, C, D}
2	{A, B, C, E}
3	{B, E}
4	{B, C, E}

⇒ Several Steps — 3M

al result →

Itemset	Support
{B, C, E}	2

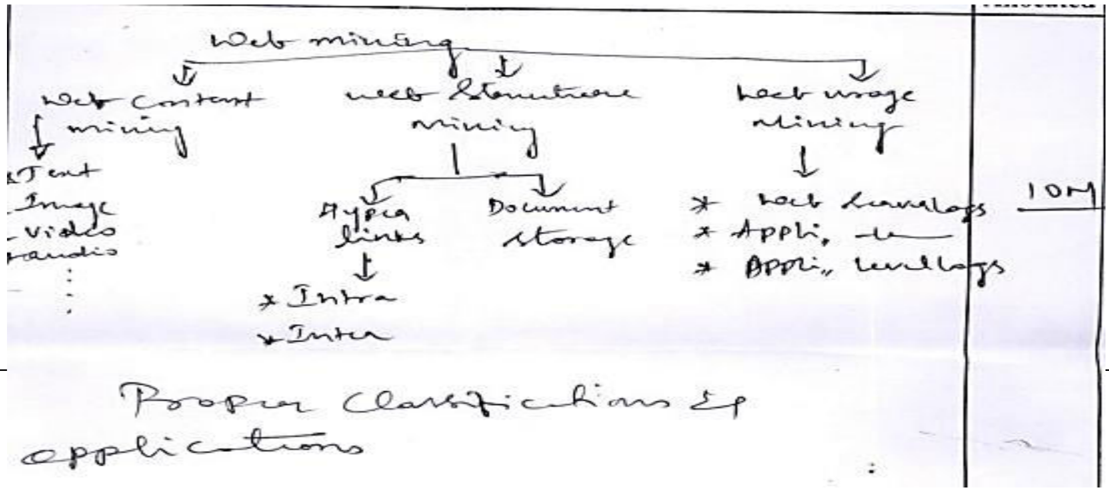
—2M

any Equivalent Example with explanation

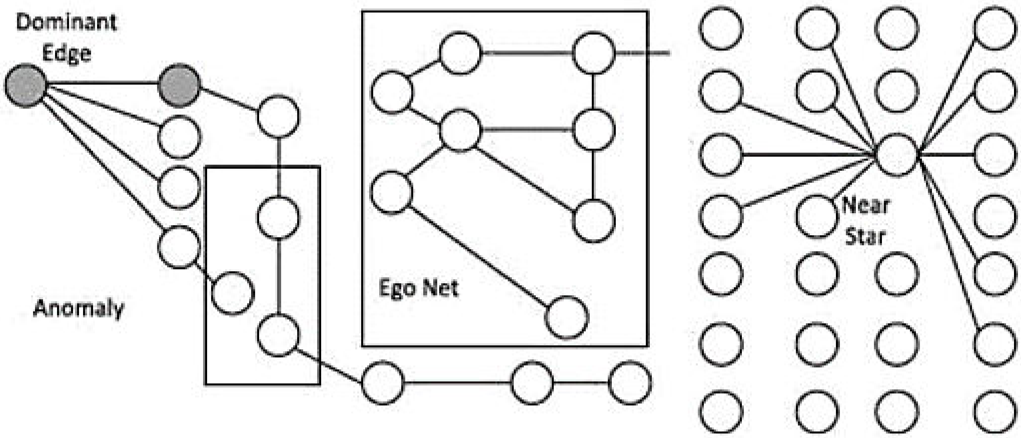
3M
+2M
+3M
+2M
=10M

OR

10 (a)	**Define web mining. Discuss the broad classification of web mining and their applications.** **Solution:**	[10]	CO5



Proper Classifications & applications

(b)	**Define the term Social network. Explain social network as graphs with centralities, Ranking and Anomaly Detection.** **Solution:** A social network is a social structure made of individuals (or organizations) called "nodes," which are tied (connected) by one or more specific types of inter-dependency, such as friendship, kinship, financial exchange, dislike or relationships of beliefs, knowledge or prestige.  Defination — 2M Centralities — 2M Ranking — 2M Anomaly Detection — 4M	[10]	CO5