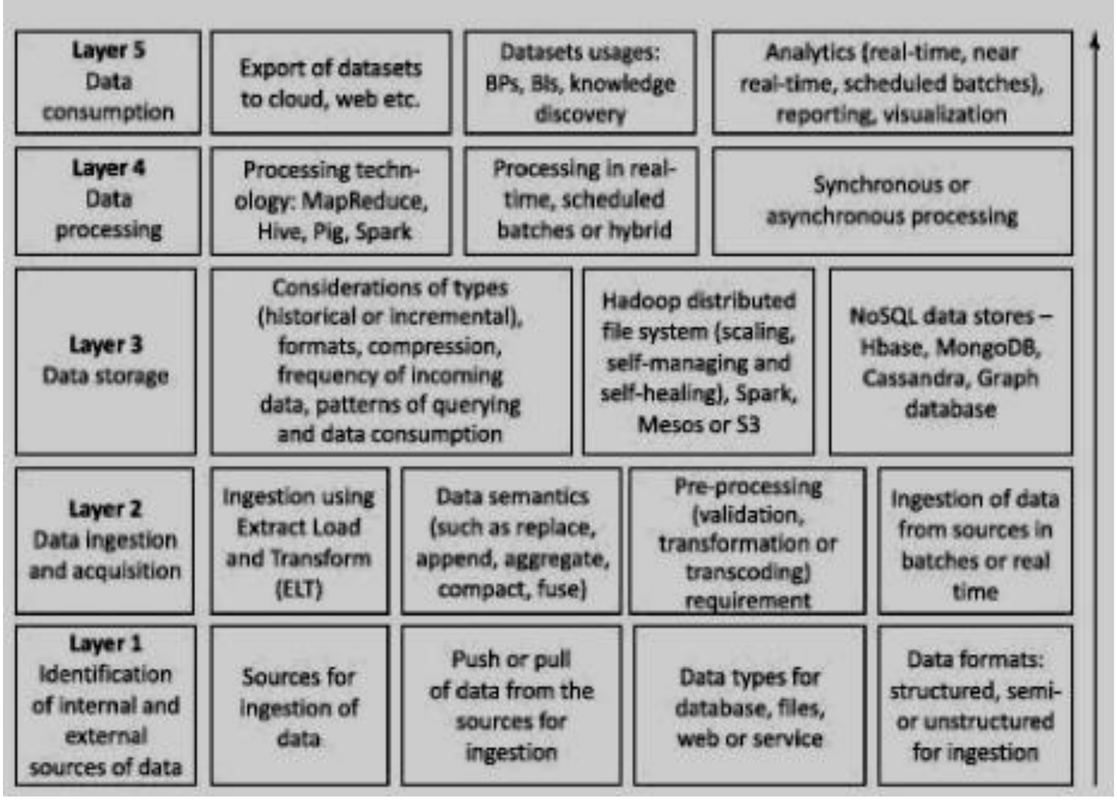
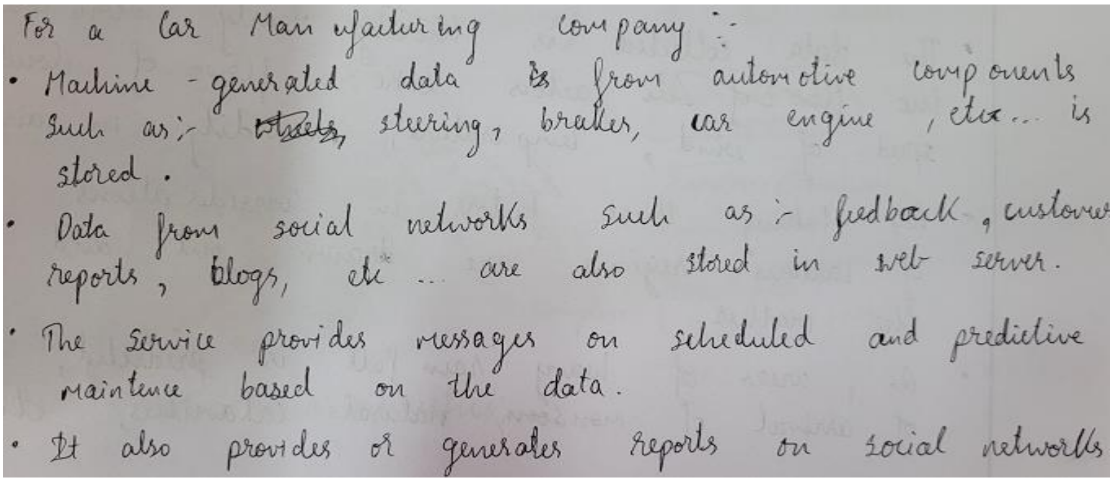
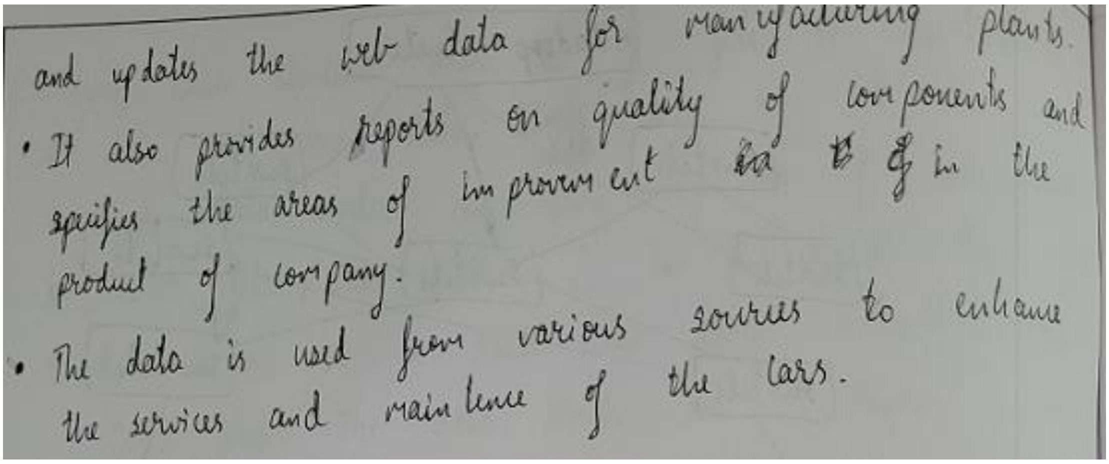
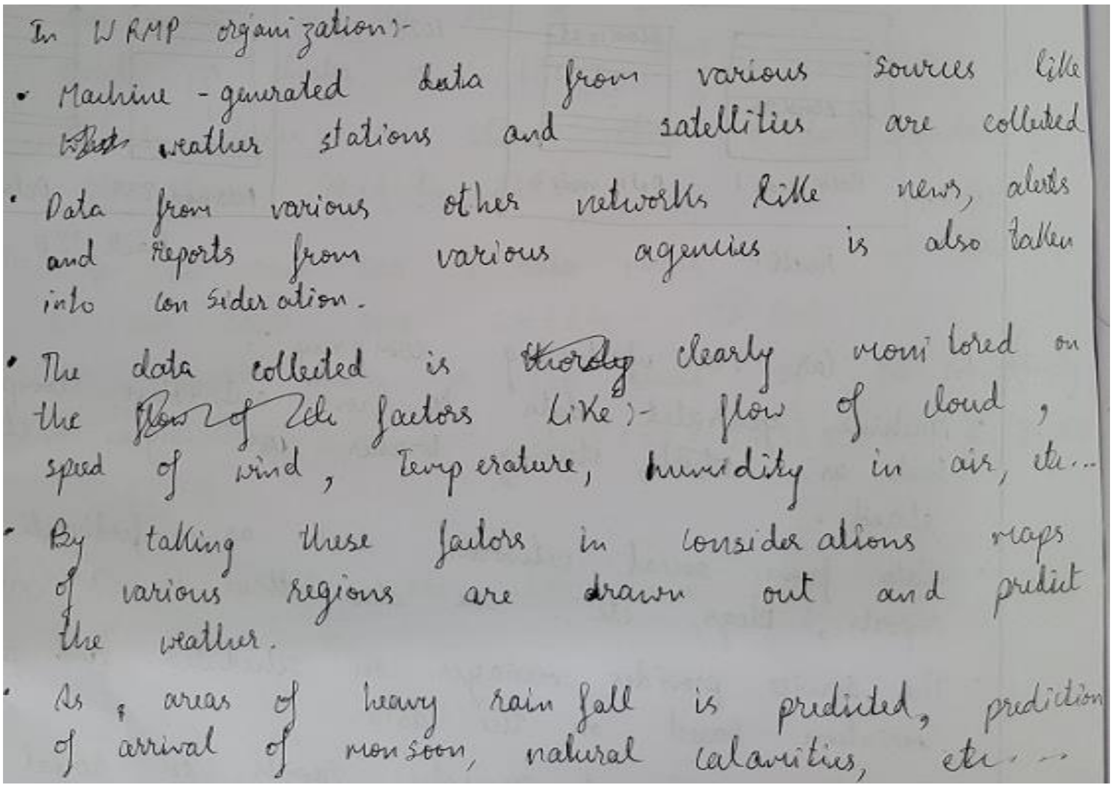
USN □□□□□□□□□□

Internal Assessment Test 1– OCT 2023
Scheme of Evaluation

Sub:	**BIG DATA and ANALYTICS**			Sub Code:	**18CS72**	Branch:	**ISE**
Date:	**30/10/2023**	Duration:	**90 min**	Max Marks: **50**	Sem/Sec:	**VII/ A, B & C**	**OBE**

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1.(a)	Discuss the functions of each of the five layers in Big Data architecture design Scheme: Definition+Diagram+ Explanation: 1+2+4 Marks 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 technology: 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	[7]	1	L2
(b)	Explain in brief about Analytics Scalability of Big Data. Scheme: Explanation of min 4 points – 3 marks Solution: Two types of scalability 1. Vertical Scalability 2. Horizontal Scalability **Vertical Scalability:** Means scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities. **Horizontal scalability:** Means scaling out [Using more multiple processors as a single entity so a business can scale beyond the computer capacity of a single server]	[3]		L2

2.	List and explain usage of Big Data Analytics in a Company for car manufacturing, marketing sales and maintenance of car service centre and WRMP Organization. Scheme:- Listing +explanation of each application : 3+3+4 Marks Solution:-  For a Car Manufacturing company :- • Machine - generated data ~~is~~ from automotive components such as:- ~~wheels~~, steering, brakes, car engine, etc... is stored. • Data from social networks such as:- feedback, customer reports, blogs, etc ... are also stored in web server. • The service provides messages on scheduled and predictive maintence based on the data. • It also provides or generates reports on social networks  and updates the web data for manufacturing plants. • It also provides reports on quality of components and specifies the areas of improvement ~~ta~~ ~~B~~ ~~of~~ in the product of company. • The data is used from various sources to enhance the services and maintence of the cars.  In WRMP organization:- • Machine - generated data from various sources like ~~that~~ weather stations and satellities are collected. • Data from various other networks like news, alerts and Reports from various agencies is also taken into consideration. • The data collected is ~~thorolly~~ clearly monitored on the ~~flow of~~ etc factors Like:- flow of cloud, speed of wind, Temperature, humidity in air, etc... • By taking these factors in considerations maps of various regions are drawn out and predict the weather. • As, areas of heavy rain fall is predicted, prediction of arrival of monsoon, natural calamities, etc...	[10]	1	L2

<table>
<tr><td>3.</td><td>

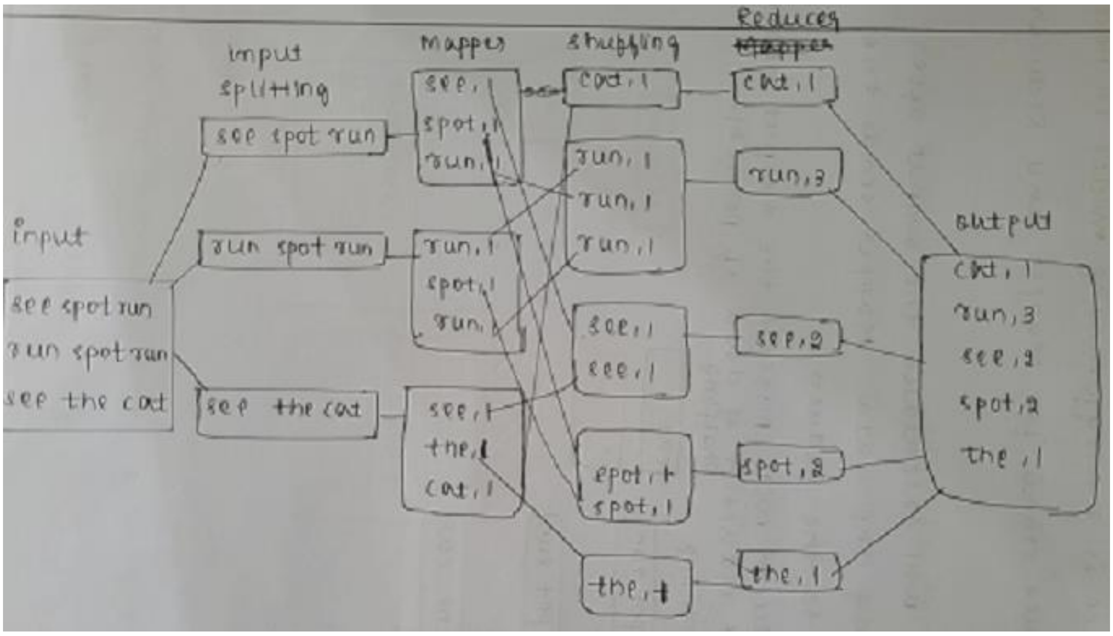
Apply **Map Reduce Programming model** for the input

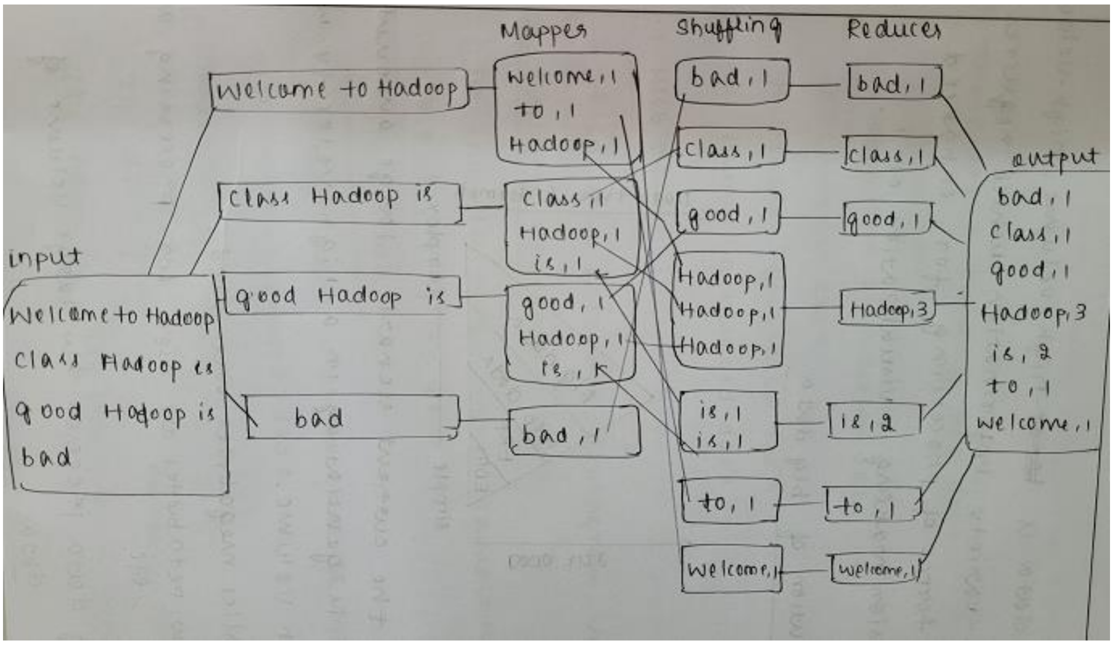
a. **“see spot run ,run spot run, see the cat** ”

b. **“ Welcome to Hadoop, class Hadoop is , good Hadoop is, bad”**

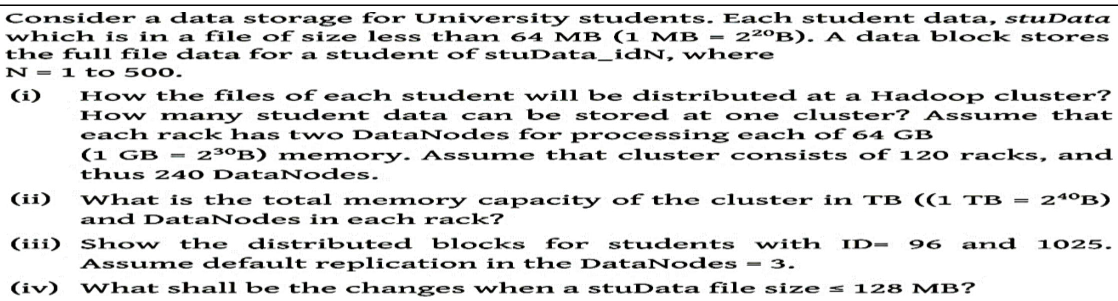
Scheme: Solving the given inputs using Map Reduce Programming Model carries 5+5 Marks.

Solution:





</td><td>[10]</td><td>2</td><td>L3</td></tr>
<tr><td>4.</td><td>



Consider a data storage for University students. Each student data, *stuData* which is in a file of size less than 64 MB (1 MB = 2^{20}B). A data block stores the full file data for a student of stuData_idN, where N = 1 to 500.

(i) How the files of each student will be distributed at a Hadoop cluster? How many student data can be stored at one cluster? Assume that each rack has two DataNodes for processing each of 64 GB (1 GB = 2^{30}B) memory. Assume that cluster consists of 120 racks, and thus 240 DataNodes.

(ii) What is the total memory capacity of the cluster in TB ((1 TB = 2^{40}B) and DataNodes in each rack?

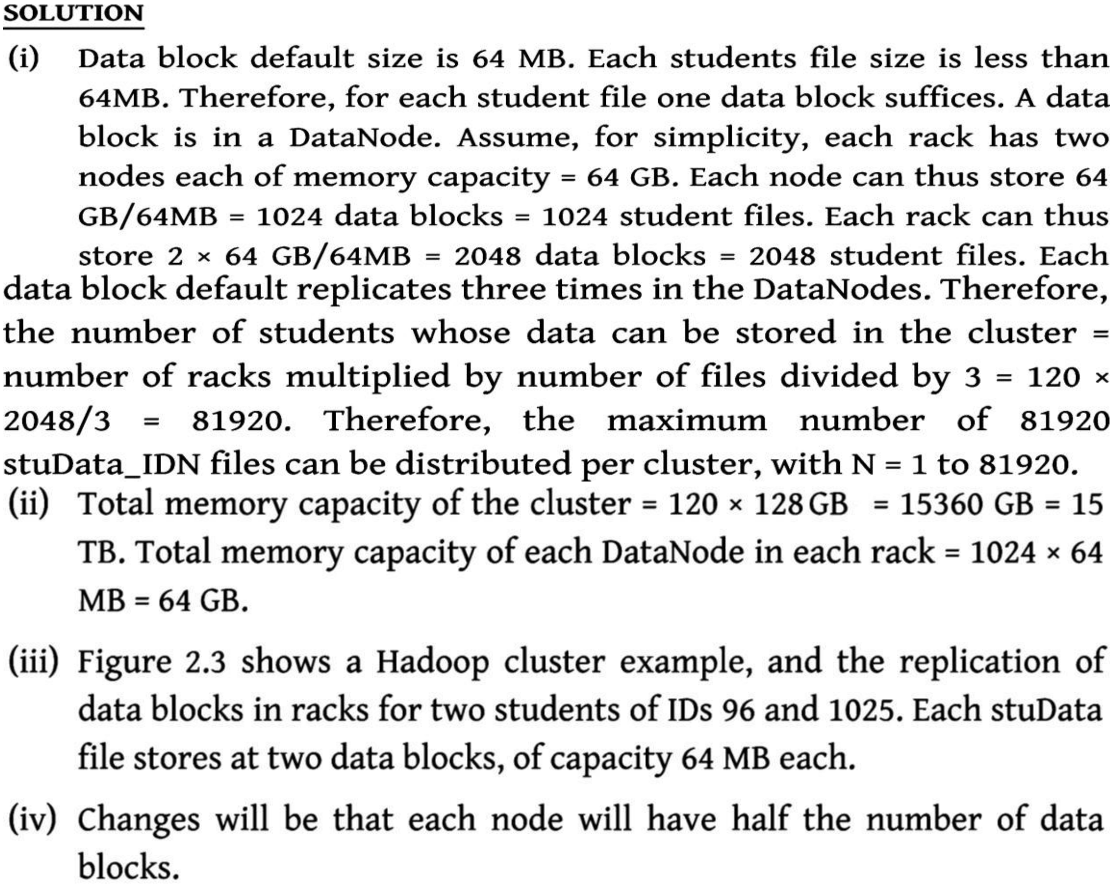
(iii) Show the distributed blocks for students with ID= 96 and 1025. Assume default replication in the DataNodes = 3.

(iv) What shall be the changes when a stuData file size ≤ 128 MB?

</td><td>[10]</td><td>2</td><td>L3</td></tr>
</table>

<table>
<tr>
<td></td>
<td>
Scheme:- Computation of Hadoop cluster distribution,data storage,total memory capacity, distributed blocks,changes in stuData file : 4+2+3+1 Marks

Solution:-



SOLUTION

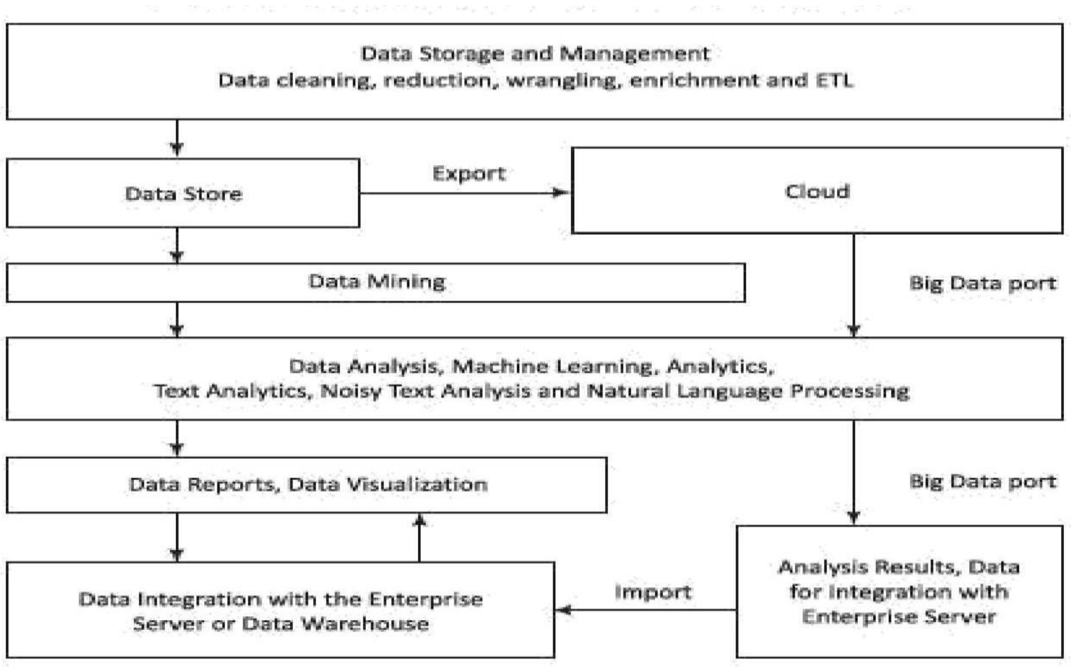
(i) Data block default size is 64 MB. Each students file size is less than 64MB. Therefore, for each student file one data block suffices. A data block is in a DataNode. Assume, for simplicity, each rack has two nodes each of memory capacity = 64 GB. Each node can thus store 64 GB/64MB = 1024 data blocks = 1024 student files. Each rack can thus store 2 × 64 GB/64MB = 2048 data blocks = 2048 student files. Each data block default replicates three times in the DataNodes. Therefore, the number of students whose data can be stored in the cluster = number of racks multiplied by number of files divided by 3 = 120 × 2048/3 = 81920. Therefore, the maximum number of 81920 stuData_IDN files can be distributed per cluster, with N = 1 to 81920.

(ii) Total memory capacity of the cluster = 120 × 128 GB = 15360 GB = 15 TB. Total memory capacity of each DataNode in each rack = 1024 × 64 MB = 64 GB.

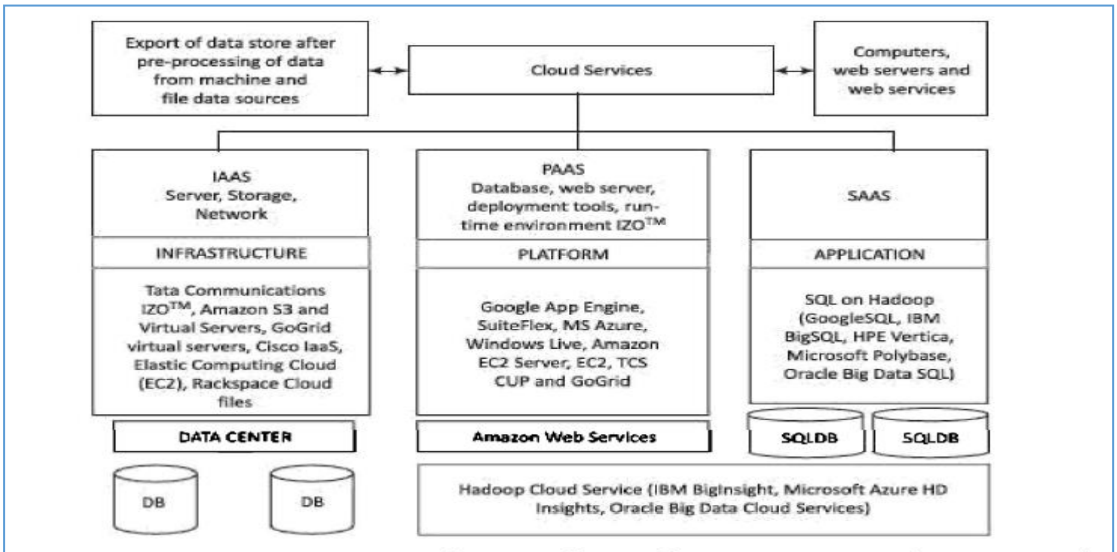
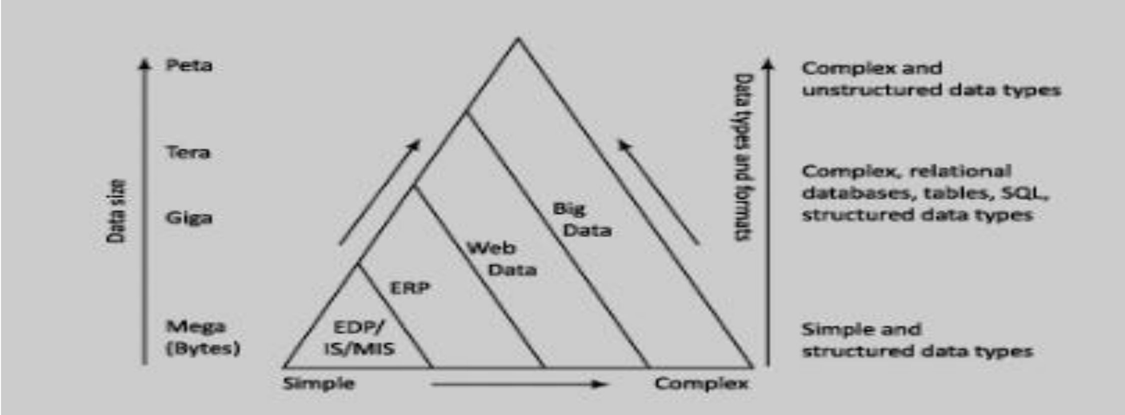
(iii) Figure 2.3 shows a Hadoop cluster example, and the replication of data blocks in racks for two students of IDs 96 and 1025. Each stuData file stores at two data blocks, of capacity 64 MB each.

(iv) Changes will be that each node will have half the number of data blocks.
</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>
With a neat diagram, illustrate the data pre-processing, analysis, visualization and data store export to cloud for big data.

Scheme: Explanation+ Diagram carries 5+5 marks



Data Storage and Management
Data cleaning, reduction, wrangling, enrichment and ETL
Data Store
Export
Cloud
Data Mining
Big Data port
Data Analysis, Machine Learning, Analytics,
Text Analytics, Noisy Text Analysis and Natural Language Processing
Data Reports, Data Visualization
Big Data port
Analysis Results, Data for Integration with Enterprise Server
Import
Data Integration with the Enterprise Server or Data Warehouse
</td>
<td>[10]</td>
<td>1</td>
<td>L2</td>
</tr>
</table>

	 Export of data store after pre-processing of data from machine and file data sources Cloud Services Computers, web servers and web services IAAS Server, Storage, Network PAAS Database, web server, deployment tools, run-time environment IZO™ SAAS INFRASTRUCTURE PLATFORM APPLICATION Tata Communications IZO™, Amazon S3 and Virtual Servers, GoGrid virtual servers, Cisco IaaS, Elastic Computing Cloud (EC2), Rackspace Cloud files Google App Engine, SuiteFlex, MS Azure, Windows Live, Amazon EC2 Server, EC2, TCS CUP and GoGrid SQL on Hadoop (GoogleSQL, IBM BigSQL, HPE Vertica, Microsoft Polybase, Oracle Big Data SQL) DATA CENTER Amazon Web Services SQLDB SQLDB DB DB Hadoop Cloud Service (IBM BigInsight, Microsoft Azure HD Insights, Oracle Big Data Cloud Services) **Figure 1.4** Data store export from machines, files, computers, web servers and web services			
6. (a)	Explain the Evolution of Big Data and their characteristics. Scheme: Explanation of Evolution +characteristics carries 3+2 marks  Peta Tera Giga Mega (Bytes) Data size EDP/ IS/MIS ERP Web Data Big Data Simple Complex Data types and formats Complex and unstructured data types Complex, relational databases, tables, SQL, structured data types Simple and structured data types	[05]	1	L2
(b)	Explain Big Data Types with examples. Scheme: Explanation of each big data type with examples carries 1+1+1+1+1 marks. Solution: 1. **Social networks and web data**, such as Facebook, Twitter, emails, blogs and YouTube 2. **Transactions data and Business Processes (BPs) data**, such as credit card transactions, flight bookings, etc. and public agencies data such as medical records, insurance business data etc. 3. **Customer master data**, such as data for facial recognition and for the name, date of birth, marriage anniversary, gender, location and income category, 4. **Machine-generated data**, such as machine-to-machine or Internet of Things data, and the data from sensors, trackers, web logs and computer systems log. Computer generated data. 5. **Human-generated data** such as biometrics data, human– machine interaction data, e-mail records, MySQL database of student grades. The following examples illustrate machine-generated.	[05]		L2

Faculty Signature CCI Signature HOD Signature