CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 18 August 2025. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions carrying forty (40) marks. SECTION II has three (3) practical questions carrying sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:

- **Python Software**
- **PySpark library**
- **Hadoop software**
- **Jupyter Notebook**
- **Pycharm IDE**
- **Java JDK**

SECTION I (40 MARKS)

1. Which database storage model optimises disk I/O for analytical queries by organising data in columnar format? (2 marks)

2. The Hadoop component that should be configured with a capacity scheduler to support fair, multi-tenant resource distribution across diverse data processing workloads is referred to as _____. (2 marks)

3. In a high-velocity data pipeline, which messaging platform supports real-time ingestion from multiple sources with scalability and fault tolerance? (2 marks)

4. In the context of real-time Machine Learning model predictions, which containerisation tool is typically used to package models for deployment in a Kubernetes-managed environment? (2 marks)

5. Which framework can be used to train distributed Machine Learning models and integrate naturally with HDFS for large-scale data storage? (2 marks)

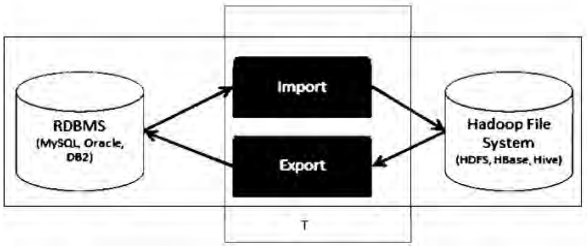
6. Which data model is used to load your data to start your analytics by transforming the data first using a set of business rules, before loading it into a sandbox? (2 marks)

7. What Hadoop component tracks MapReduce task progress and reassigns failed tasks to ensure job completion? (2 marks)

8. Java must be installed on your system before installing Hive. What command is used to verify the Java installation? (2 marks)

9. Which NoSQL database, integrated with Hadoop, uses distributed key-value storage for low-latency read-heavy ML queries? (2 marks)

10. Identify the Hadoop ecosystem tool represented by the letter **T** as used in big data management. (2 marks)



11. The type of data stored in a NoSQL database that uses collections and documents instead of tables and rows is known as _____. (2 marks)

12. The process of removing errors and combining different data sources to make them easier to analyse is known as _____. (2 marks)

13. The Hadoop MapReduce function that performs sorting and filtering of data and thereby organising it in the form of a group is _____. (2 marks)

14. The process of ensuring consistency and coherence of data across distributed big data systems is known as big data _____. (2 marks)

15. The high-level data warehousing tool in the Hadoop ecosystem that allows SQL-like querying is referred to as _____. (2 marks)

16. What is the HDFS file system command used for the recursive listing of the directories and files under a specific folder? (2 marks)

17. Apache Spark is based on a master-slave architecture, where the slaves are called _____. (2 marks)

18. State the Linux commands to start Hadoop daemons for execution. (2 marks)

19. The procedural language platform used to develop a script for MapReduce operations in Hadoop is known as _____. (2 marks)

20. The type of data collected over time, such as a year and then fed into an analytics system is known as _____. (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (f).

Use the employee attendance data in the table below to answer the questions that follow.

Employee ID	Employee name	Department	Location	Work date	Status	Hours worked
E001	Alice Otieno	HR	Nairobi	01-05-24	Present	8
E002	Brian Kimanzi	IT	Kampala	01-05-24	Absent	0
E003	Charles Moyo	Finance	Harare	01-05-24	Present	7
E001	Alice Otieno	HR	Nairobi	02-05-24	Present	9
E002	Brian Kimanzi	IT	Kampala	02-05-24	Present	8
E003	Charles Moyo	Finance	Harare	02-05-24	Present	6
E001	Alice Otieno	HR	Nairobi	03-05-24	Absent	0
E002	Brian Kimanzi	IT	Kampala	03-05-24	Present	7
E003	Charles Moyo	Finance	Harare	03-05-24	Present	8

(a) Create an external Hive table named "employee attendance" based on the dataset and insert data into the table. (5 marks)

(b) Write a HiveQL query to calculate the total hours worked by each employee across all dates. (3 marks)

(c) Write the Hive query to find the average daily hours worked per department. (3 marks)

(d) Write the Hive query to list the employees who have been absent at least once. (3 marks)

(e) Using a window function, show each employee's work records along with a cumulative total of hours worked, ordered by date. (4 marks)

(f) Write the Hive query to identify the location with the highest total hours worked. (2 marks)

Save "Question 21" document and upload. **(Total: 20 marks)**

22. Create a word processing document named "Question 22" and use it to save your answers to questions (a) and (b).

(a) Create a folder in drive C and name it "BDM". Create an Excel document shown below and save it as a comma-delimited version (CSV) file named "ElectonicInventory". (3 marks)

ProductName	Supplier	Packaging	In stock	Reorder Level
USB-C Cable 1 M	Tech Trend	Box	3200	200
Wireless Mouse	Electro Mart	Packs	1500	100
Bluetooth Earbuds	Sound Wave	Box	2800	150
HDMI Cable 2m	Connect Pro	Packs	900	50
Power Bank 10000mAh	Charge Tech	Box	1200	80
USB Flash Drive 64GB	Data Sync	Packs	600	30

(b) Use Python programming to perform the following:

(i) Import all necessary libraries for data manipulation, clustering and visualisation. (3 marks)

(ii) Load the "ElectonicInventory" dataset and create a DataFrame stock. (4 marks)

(iii) Select numerical columns from the stock DataFrame to be used for clustering and load them into an object named "X". (2 marks)

(iv) Standardise features into an object named "X_scaled". (2 marks)

(v) Apply K-means clustering using the scaled features (Assume 3 clusters). (3 marks)

(vi) Create a histogram to visualise the distribution of the in-stock values from the product data. (3 marks)

Save "Question 22" document and upload. **(Total: 20 marks)**

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) to (e) and capture screenshots.

(a) Write the Python MongoDB code to create a database called "LibraryDB" and check if the database exists. (4 marks)

(b) Write Python code that creates a collection called Books in LibraryDB using insert_one() and confirm it using collection_names(). (4 marks)

(c) Write a MongoDB Python script to drop the Loans collection if it exists in the LibraryDB database. (3 marks)

(d) In MongoDB, use find one() and projection to fetch only the title and author fields from the first document in the Books collection. Demonstrate how to count the total number of books using count_documents().

(i) Projection using find_one(). (2 marks)

(ii) Counting documents. (2 marks)

(e) Write a Python MongoDB script to calculate the number of books written by each author and display only authors with more than one book. Sort the results in descending order by book count. (5 marks)

Save "Question 23" document and upload. **(Total: 20 marks)**

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 2 December 2024. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions carrying forty (40) marks. SECTION II has three (3) practical questions carrying sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:

- **Python Software**
- **PySpark library**
- **Hadoop software**
- **Jupyter Notebook**
- **Pycharm IDE**
- **Java JDK**

SECTION I (40 MARKS)

1. The primary interface for a user to describe a MapReduce job to the Hadoop framework for execution is called _____. (2 marks)

2. The Hadoop component that is responsible for triggering the workflow actions, which in turn uses the Hadoop execution engine to execute the task is know_____. (2 marks)

3. In the Hadoop File System, which command would you use to create a directory named "cisseBDM" in the present working directory? (2 marks)

4. The data integration tools that extract data and load it into a data lake for transformation and analysis used in the data warehouse architecture are known as _____. (2 marks)

5. An architecture that creates distinct, isolated environments within a single physical infrastructure, such as a virtual machine, server or cloud platform in big data processing pipelines is known as _____. (2 marks)

6. The read-only collection of data in Spark that can be partitioned across multiple machines in a cluster, allowing for parallel computation and fault tolerance through lineage reconstruction is known as _____. ` (2 marks)

7. A type of NoSQL database that stores data in tables, rows and dynamic columns and can employ column compression techniques to reduce the storage space and enhance performance is known as_____. (2 marks)

8. The process of establishing standards, policies and procedures to manage the data effectively referred to as _____. (2 marks)

9. In the big data pipeline, the data warehouse software which is designed for reading, writing and managing tabular-type datasets and for data analysis is known as _____. (2 marks)

10. The database activity that ensures that simultaneous transactions on a DBMS do not interfere with each other and prevents common issues like dirty reads, lost updates and non-repeatable reads is called _____. (2 marks)

11. The type of computing where processing and data storage is spread across multiple devices or systems in an organisation is referred to as _____. (2 marks)

12. The big data security goal which states that information should not be altered without the authorisation of the owner is known as _____. (2 marks)

13. The structured approach that consists of the core capabilities that organisations need to take into consideration when setting up their big data organisation is called _____. (2 marks)

14. A small-sized block of data derived from another block of digital data to detect errors that may have been introduced during its transmission or storage is known as _____. (2 marks)

15. The service within Hadoop that farms out MapReduce tasks to specific nodes in the cluster and it is responsible for storing job history is known as _____. (2 marks)

16. The moral obligations of gathering, protecting and using personally identifiable information and how it affects individuals is known as _____. (2 marks)

17. The big data technology that focuses on the real-time processing of continuous streams of data in motion is referred to as _____. (2 marks)

18. The process of bringing computation to where the data resides in a Hadoop files system is known as _____. (2 marks)

19. The set of tools and processes that are used to uncover the entire journey of data from the source to the destination systems are known as _____. (2 marks)

20. In big data, the iterative process that seeks out patterns and looks at clusters, sequences of events, specific trends and time-series analysis is referred to as _____. (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (b).

(a) Use the data in the table below to answer the questions that follow:

TRANSACTION ID	ITEMS
1	Eggs, Sausage, Pork
2	Maize, Beans, Egg
3	Beef, Maize, Peas, Eggs
4	Beef, Mutton, Maize
5	Peas, Eggs, Maize, Pork
6	Maize, Peas, Beans
7	Peas, Eggs, Mutton, Maize, Pork

(i) Write the Python code to import the necessary libraries and load the transaction data. (4 marks)

(ii) Write the Python code to change the transaction data into a format suitable for the Apriori algorithm then apply the Apriori algorithm to find frequent item sets. (6 marks)

(iii) Write the Python code to generate association rules from the frequent item sets with their support, confidence, and lift values. (6 marks)

(b) Give the Hadoop Hive clauses syntax to perform the following tasks:

(i) HAVING clause to filter groups based on aggregate conditions. (2 marks)

(ii) ORDER BY clause to sort the results based on specified columns. (2 marks)

Save and upload "Question 21".

(Total: 20 marks)

22. Create a word processing document named "Question 22" and use the word processor document to save your answers to questions (a) to (d).

(a) Write the Python MongoDB code that will create a collection called "CISSE" and print a successful message. (5 marks)

(b) Type the Python MongoDB code that will insert the data in the table below into the CISSE collection. (6 marks)

Staffno	Firstname	Lastname	Salary	Gender
SX345	Victor	Njuguna	45000	Male
WD456	Anne	Moraa	23000	Female
DE467	James	Onyango	76000	Male
RE3456	Sarah	Mwende	45000	Female
TR4567	John	Kyalo	90000	Male

(c) Write the Python script to sort the data by staff number in ascending order and then by last name in descending order. (4 marks)

(d) Write the Python script to find the average salary for all staff in the CISSE collection. (5 marks)

Save and upload "Question 22".

(Total: 20 marks)

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) to (b).

(a) Create an excel document shown below and save it as "**invoicedata.csv**". (3 marks)

Invoice number	Stock Code	Description	Quantity	Unit price in dollars
5454334	5466A	Hammer	342	564
5464533	5463B	Pliers	543	456
5676563	6754D	Crowbar	345	345
4534343	7543D	Mattock	67	435
4356243	4533E	Spanner	68	657
5554334	5466A	Hammer	342	564
4356343	4533E	Spanner	68	657
4534349	7543D	Mattock	69	435
5676573	6754D	Crowbar	347	345
5464544	5463B	Pliers	550	456
5456334	5466A	Hammer	389	560
8454334	5466A	Hammer	369	561

(b) Write the Python codes to perform the following:

(i) Import the necessary Python libraries and extract data from the source. (4 marks)

(ii) Transform the data into a numeric array. (2 marks)

(iii) Separate independent variables from dependent variables. (2 marks)

(iv) Normalise the data. (3 marks)

(v) Apply K-means algorithm to create clusters. (3 marks)

(vi) Visualise the generated cluster. (3 marks)

Save and upload "Question 23".

(Total: 20 marks)

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 19 August 2024. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

SECTION I (40 MARKS)

Required resources:

- **Python**
- **Hadoop software**

1. What is the column-oriented non-relational database management system built on top of the Hadoop Distributed File System (HDFS)? (2 marks)

2. The open-source Apache project that provides a centralised service for providing configuration information, naming, synchronisation and group services over large clusters in distributed systems is referred to as _____. (2 marks)

3. The compression algorithms frequently applied in in-memory big data management to tackle the growing gap between processor speed and main memory bandwidth is called _____. (2 marks)

4. An architecture in which a single instance of a database application serves multiple customers is referred to as _____. (2 marks)

5. The data transformation activity that involves filling up the NULL values with some default values is referred to as _____. (2 marks)

6. The data warehouse architecture component, which supports the modeling of data using analytical tools, is known as _____. (2 marks)

7. The first stage in data analysis, where data analysts employ data visualisation and statistical methods to describe the characteristics of a dataset, is referred to as _____. (2 marks)

8. The big data characteristic that refers to how fast data can be generated, gathered and analysed is known as _____. (2 marks)

9. The type of database that does not enforce data types, constraints or relationships is referred to as _____. (2 marks)

10. Businesses are utilising Artificial Intelligence (AI) and Machine Learning (ML) technologies to handle more complex data management activities. List **TWO** of these activities. (2 marks)

11. Name the open-source framework that is dedicated to handling interactive queries, machine learning and real-time operations. (2 marks)

12. The big data quality consideration which state that a piece of information should not contradict another piece of information in a different source or system is referred to as _____. (2 marks)

13. The type of data that can be analysed as multiple data points per observation over time is referred to as _____. (2 marks)

14. A database activity that is used to rollback transactions and recover from crashes is referred to as _____. (2 marks)

15. In big data preprocessing, the technique that preserves mean and sample size is referred to as _____. (2 marks)

16. The tool used to easily import information from structured databases and related Hadoop systems (such as Hive and HBase) into your Hadoop cluster is referred to as _____. (2 marks)

17. A client-server model where agents collect and transport data from various sources to a central data repository is known as _____. (2 marks)

18. State a challenge encountered in ensuring big data quality. (2 marks)

19. In big data, the concurrency control problem that occurs when the second transaction selects a row, which, is updated by another transaction, is referred to as _____. (2 marks)

20. The type of big data synchronisation where information is transferred in a single direction, typically from a source to a target system is referred to as _____. (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (c).

(a) Write the Hadoop Hive statement that will create and insert data into the car table shown below: (10 marks)

Car table

CARNUMBER	CARMODEL	YEAROFMANUF	CARCOST_DOLLARS
KDP435N	TOYOTA	2014	3256
KDN546R	MAZDA	2010	4566
KDD564T	MAZDA	2011	4987
KCX453D	TOYOTA	2014	8765
KCT498K	NISSAN	2010	4567
KDC998R	TOYOTA	2015	7658
KDB321K	NISSAN	2015	4567
KDH765Y	MAZDA	2012	6989
KDC347K	BMW	2011	8765

(b) Write the Hadoop hive statement that will find the total car cost by car model and year of manufacture. (5 marks)

(c) Write the Hadoop statement that will count the number of cars that were manufactured after the year 2011 and whose cost was greater than 4500 dollars. (5 marks)

Save "Question 21" document and upload. **(Total: 20 marks)**

22. Create a word processing document named "Question 22" and use the word processor document to save your answers to questions (a) to (c).

(a) A list of numbers is given by 3,5,4,7,8,9,5,6,11,12,13,14,15,17,9. Write the python code that will pipeline the data by performing the following activities:

(i) Filtering out the even numbers. (2 marks)

(ii) Multiplying each number by 12. (2 marks)

(iii) Adding 10 to each number. (3 marks)

(iv) Calculating the average of the resulting numbers. (3 marks)

(b) Write the python MongoDB code to create a database called "Kenya" and check if the database exists. (5 marks)

(c) Write the python MongoDB code that will create a collection called "Nairobi". (5 marks)

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 22" document and upload. **(Total: 20 marks)**

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) to (d).

Use the dataset below to answer the questions that follow:

ClientID	ClientLname	Gender	ClientLocation	ClientExpenditure
201	Anne	Female	Brazil	40000
203	Alice	Female	Canada	60000
204	Andrew		Comoros	70000
205	Job	Male	Brazil	87000
206	Vincent	Male	Brazil	45000
208	Ayub	Male	Canada	56000
209	Joseph	Male	Chile	64000
210	Barrack	Male	Chile	60000
201	Anne	Female	Brazil	40000
206	Vincent	Male	Brazil	45000
	Andrew	Male	Comoros	70000
215	Victor	Male	Chile	89000

(a) Write the python code that will create a data dictionary for the data. (6 marks)

(b) Write the python code that will return a data frame with no empty cells. (4 marks)

(c) Write the python code that will remove duplicate entries from the data set. (5 marks)

(d) Write the python code that will compute the total client expenditure. (5 marks)

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 23" document and upload. **(Total: 20 marks)**

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 22 April 2024. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:
- **Python Software**
- **PySpark library**
- **Hadoop software**

SECTION I (40 MARKS)

1. _____ is the underlying general execution engine for Apache Spark platform that all other functionality is built upon. (2 marks)

2. The big data concept that addresses the needs of organisations at multiple points in their big data projects is known as _____. (2 marks)

3. When managing a variety of big data, the type of database that provides a flexible schema-less approach for storing and accessing data referred to as _____. (2 marks)

4. Give the term that **BEST** describes the data-mining task, which supports the finding of correlations between items in a database. (2 marks)

5. In data warehousing, the activity that supports incremental loading of new data on a periodic basis is referred to as _____. (2 marks)

6. The big data quality consideration which state that a piece of information should not contradict another piece of information in a different source or system is referred to as _____. (2 marks)

7. In the data warehouse architecture, the landing area for data from the source is known as _____. (2 marks)

8. The type of data warehouse architecture that contains both a traditional batch data pipeline and a fast streaming pipeline for real-time data is called _____. (2 marks)

9. _____ states that big data can handle potentially useful data regardless of where it is coming from by consolidating all information into a single system. (2 marks)

10. The algorithm which supports candidate generation in data mining is referred to as _____. (2 marks)

11. The data pipeline component that prepares big data for analysis and creates a controlled environment for downstream processes is referred to as _____. (2 marks)

12. _____ is an immutable distributed collection of datasets partitioned across a set of nodes of the cluster that can be recovered if a partition is lost, thus providing fault tolerance. (2 marks)

13. _____is a Hadoop ecosystem component which provides a distributed data warehouse for data that is stored in HDFS. (2 marks)

14. The type of data storage that efficiently writes and reads data to and from hard disk storage in order to speed up the time it takes to return a query is referred to as _____. (2 marks)

15. The type of NoSQL database that stores data in pairs and is optimised for simple and fast read/write operations is referred to as _____. (2 marks)

16. _____ is a purpose-built database that relies primarily on internal memory for data storage and enables minimal response times by eliminating the need to access standard disk drives. (2 marks)

17. The concurrency control challenge where if one user accesses data that another user has updated but not yet finalised and then the second user decides to cancel their transaction resulting in the first user having invalid data is known as _____. (2 marks)

18. _____ includes all organisational aspects that should be taken into account in a big data organisation and is vendor independent. (2 marks)

19. The type of big data deployment where the database is hosted in another organisations server and is accessed remotely using internet technologies is referred to as _____. (2 marks)

20. Which is the mutual exclusion algorithm requirement in big data that states two or more sites should not endlessly wait for any message that will never arrive? (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (c).

(a) Write the Python MongoDB code to create a database called "KASNEB" and check if the database exists. (6 marks)

(b) Write the Python MongoDB code that will create a collection called "CISSE". (4 marks)

(c) MongoDB uses mapReduce command for map-reduce operations. Explain the following mapReduce code. (10 marks)

```
>db.collection.mapReduce(
function() {emit(key,value);},  //map function
function(key,values) {return reduceFunction}, {   //reduce function
out: collection,
query: document,
sort: document,
limit: number
}
)
```

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 21" and upload. **(Total: 20 marks)**

22. Create a word processing document named "Question 22" then use the word processor document to save your answers to questions (a) and (b).

(a) Create a folder called "CISSE" in drive C. Create an Excel document shown below. Save it as a comma delimited version (CSV) file named "**Student**". (2 marks)

H12

	A	B	C	D	E
1	**StudentNumber**	**StudentFName**	**City**	**FeesPaid ($)**	
2	NE2345	James	Nairobi	400	
3	WR4657	Alice	Nairobi	564	
4	TY6453	Frida	Kisumu	354	
5	RT6576	Anne	Kisumu	456	
6	WE6534	Hellen	Nakuru	678	
7	YT6998	Joan	Nairobi	432	
8	TK7865	Derrick	Kisumu	897	
9	RY5643	Beatrice	Mombasa	876	
10	GH4563	Grace	Mombasa	457	
11	RT8976	Reagan	Nakuru	997	
12	LK4356	Benard	Mombasa	298	
13	PP5643	Maurice	Nairobi	986	
14	TB7894	Juliet	Nairobi	675	
15	YS3455	Noel	Kisumu	745	
16	QW4334	Leonard	Nairobi	908	
17	LQ3776	Vincent	Kisumu	786	

(b) Use python data analytics tool to perform the following on the data in question 22 (a):

(i) Define the necessary python libraries. (2 marks)

(ii) Extract and define the data. (2 marks)

(iii) Encode the data and transform it into a numeric array. (3 marks)

(iv) Split the data into independent and dependent attributes. (2 marks)

(v) Normalise independent variables array. (2 marks)

(vi) Decompose columns into 2 columns. (2 marks)

(vii) Use K-Means for clustering. (3 marks)

(viii) Visualise the generated clusters. (2 marks)

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 22" and upload. **(Total: 20 marks)**

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) and (b).

(a) State the Hadoop Hive commands that can be used to perform the following:

(i) Create a database named "Student". (3 marks)

(ii) Rename the database from "Student" to "Library". (3 marks)

(iii) Drop a database named "library". (3 marks)

(b) Write the Hadoop Hive command that will create and insert data into the following table. (7 marks)

Patientid	Patienftname	Patientlname	Patientwardcode	Feespaid
234	James	Opiyo	W34	20000
432	Ann	Moraa	W54	34000
654	Viola	Nekesa	W21	54000

(c) State the Hadoop Hive query that will display the patient identification, patient ward code and the fees paid by the patient for all patients who paid less than Sh.35,000. (4 marks)

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 23" and upload. **(Total: 20 marks)**

.....

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 4 December 2023. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:

- **Python Software**
- **PySpark library**
- **Hadoop**

SECTION 1 (40 MARKS)

1. The ability to split an unbounded stream of data into finite sets, based on specified criteria such as time or count so that you can perform aggregate functions such as sum or average as used in Big Data is called: (2 marks)

2. In the 5 V's of Big Data, which V explains the speed at which companies receive, store and manage data. (2 marks)

3. The programming model for processing and generating large datasets that can be distributed across a Hadoop cluster through aggregation and summarisation is called: (2 marks)

4. The process of breaking a large dataset into smaller, more manageable parts, often for distributed storage and processing is called: (2 marks)

5. The process of fixing incorrect, incomplete, duplicate or otherwise erroneous data in a data set is referred to as: (2 marks)

6. The type of Extract, Transform, and Load (ETL) tool that is free to use and whose source code is readily available thus allowing people to extend or enhance their capabilities is called: (2 marks)

7. The data warehousing activity where many steps are taken to load new or updated data into the data warehouse is known as: (2 marks)

8. The process of choosing a smaller part of a data set and using that subset for viewing or analysis is known as: (2 marks)

9. The type of data mining technique that is used to identify and analyse the relationship between variables is called: (2 marks)

10. The type of mining that concentrates on identifying rules that describe specific patterns within the data is known as: (2 marks)

11. The Hadoop mode which involves calling *"~/Hadoop-directory/bin/Hadoop"* that will execute a Hadoop operation single Java process is referred to as: (2 marks)

12. In the Hadoop Distributed File System (HDFS), what is the name given to the file segments that are stored in individual data nodes: (2 marks)

13. The big data activity of consolidating data across different sources, applications and devices while maintaining consistency is referred to as: (2 marks)

14. The type of NoSQL database where each document is a nested structure of keys and values is known as: (2 marks)

15. A central repository that stores vast amounts of raw data, both structured and unstructured, making it available for various analyses and processing is called: (2 marks)

16. A large number of features or attributes can make visualization difficult and lead to information overload. The process of reducing the complexity of Big Data while preserving the most critical information is referred to as: (2 marks)

17. A professional who uses statistical and analytical techniques to extract insights and knowledge from data place is known as: (2 marks)

18. _____database design is a database structure that does not have a predefined schema, allowing data to be stored and retrieved without the need for a rigid and fixed structure. (2 marks)

19. Name the open source framework focused on interactive query, machine learning, and real-time workloads. (2 marks)

20. The Hadoop MapReduce class that allows the user to configure the job, submit it, control its execution and query the state is referred to as: (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (g).

Write a Python script called "Question21" to perform the following tasks as used in Apache Spark.

(a) Use a spreadsheet to create the dataset called "bizsales" using the table below and save it as a CSV file. (4 marks)

Date	Product	Quantity	Price
2023-01-01	Computer	10	30000.00
2023-01-02	TV	5	50000.00
2023-01-03	Computer	8	80000.00
2023-01-04	Mobile Phone	12	18000.00
2023-01-05	TV	6	60000.00
2023-01-06	Computer	7	70000.00
2023-01-07	Mobile Phone	15	15000.00

(b) Write a Python program to load the matplotlib and the Spark session libraries and initialise a Spark session. (4 marks)

(c) Use Spark to load the dataset "bizsales" and display the result. (3 marks)

(d) Perform initial data exploration to understand your dataset such as summary statistics of the DataFrame. (2 marks)

(e) Create a well labeled and appealing histogram based on the price attribute using matplotlib to visualise the Dataset. (4 marks)

(f) Convert your Spark DataFrame to a Pandas DataFrame. (2 marks)

(g) Close the spark session. (1 mark)

Save and upload Question 21 document.

(Total: 20 marks)

22. Create a word processing document named "Question 22" and use the word processor document to save your answers to questions (a) and (b)

(a) Describe the following Hadoop MapReduce mapper code.

```
public static class Map extends Mapper<LongWritable,Text,Text,IntWritable>

{
public void map(LongWritable key, Text value, Context context) throws
IOException,InterruptedException

{
String line = value.toString ();
StringTokenizer tokenizer = new StringTokenizer(line);
while (tokenizer.hasMoreTokens())

{
value.set(tokenizer.nextToken());
context.write(value, new IntWritable(1));
}
```


(10 marks)

(b) Write the Hadoop pig codes that will perform the following:

(i) Load the data from a file "dcna.txt" into bag named "lines". The entire line should be stuck to element line of type character array. (3 marks)

(ii) Tokenize the text in the bag lines. (3 marks)

(iii) Create a bag for unique character where the grouped bag will contain the same character for each occurrence of that character. (2 marks)

(iv) Count the number of occurrences in each group. (2 marks)

Save and upload Question 22 document.

(Total: 20 marks)

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) and (b).

(a) Type the following dataset into a spreadsheet worksheet. Save the worksheet in "Analytics" folder as a comma separated version (CSV) file named ***Colleges.csv***

First	UON	Good	Yes	Y	0
First	UON	Good	Yes	Y	0
First	UON	Good	Yes	Y	0
First	UON	Good	Yes	Y	0
Second	UON	Good	Yes	N	0
Second	MOI	Bad	Yes	Y	0
Second	JKUAT	Good	Yes	N	1
Second	JKUAT	Good	Yes	N	1
Second	JKUAT	Good	No	Y	1
First	UON	Good	No	Y	2
Second	UON	Bad	No	N	2

(b) Write the python code to import the necessary python libraries and extract data from the source. (4 marks)

(c) Write the python code to transform the data into a numeric array and print the output. (3 marks)

(d) Write the python code to separate independent variables from dependent variable and normalise the data before printing the output. (5 marks)

(e) Apply K-means algorithm to create clusters and visualise the generated clusters. (6 marks)

Save and upload Question 23 document.

(Total: 20 marks)

.....

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 21 August 2023. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resource:

- Python Software

SECTION I (40 MARKS)

1. The underlying technological components and systems that are designed and employed to support the storage, processing, and analysis of large and complex data sets is known as _____. (2 marks)

2. What is the name given for software systems that distribute data and computational tasks across multiple machines (nodes) in a distributed computing environment? (2 marks)

3. Big data ethics refers to the ethical considerations and principles that guide the collection, storage, analysis, and use of large and complex datasets. Which aspect of big data ethics ensures organisations employ big data analytics for socially beneficial purposes and avoids using data in ways that could harm individuals or society? (2 marks)

4. Security is a critical aspect to consider when designing and implementing distributed applications. Which term relates to protecting sensitive data during transmission and storage by utilising secure communication protocols? (2 marks)

5. The open-source data warehouse infrastructure built on top of Apache Hadoop that provides SQL-like query for analysing and processing large datasets stored in a distributed storage system such as the Hadoop Distributed File System (HDFS) is called? (2 marks)

6. A retail company wants to leverage big data analytics to improve its business operations, customer experience, and marketing strategies. Which design solution will be used to represent the different types of data, such as transaction data, customer profiles, and product catalogs? (2 marks)

7. The component of big data architecture that includes a way to capture and store messages from real-time sources for stream processing is referred to as _____. (2 marks)

8. A data model that identifies by a unique key, and the corresponding value to be of any data type, such as strings, numbers, objects, or even complex data structures like JSON is known as _____. (2 marks)

9. The type of database design that does not rely on relational database management system to enforce any specific kind of structure is known as _____. (2 marks)

10. The process of ensuring data remains consistent and up to date across multiple systems or devices is known as _____. (2 marks)

11. What is the name given to open-source machine learning library that provides scalable algorithms and tools for implementing various machine learning tasks? (2 marks)

12. The big data pipeline feature that alerts users and provides immediate failover in the event of application failure or node failure is referred to as _____. (2 marks)

13. The type of data uncertainty that is characterised by a probabilistic distribution function is referred to as _____. (2 marks)

14. State the algorithm that Hadoop relies on to handle big data thus dividing a single task into multiple tasks. (2 marks)

15. What is the name of data mining technique that identifies relationships and patterns among variables, such as market basket analysis? (2 marks)

16. The big data concept that describes the diversity and range of different data types, including unstructured data, semi-structured data and raw data is referred to as _____. (2 marks)

17. In a data warehousing architecture, the software component that connects the databases together and makes them accessible to other applications is called _____. (2 marks)

18. Big data management is a constantly evolving field, driven by advancements in technology, data processing capabilities, and changing business requirements. What is the name given to emerging practice that applies agile and DevOps principles to data management processes? (2 marks)

19. The process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in datasets is known as _____. (2 marks)

20. What is the name given to a software architectural style that emphasises the creation of services as modular and reusable components for building applications? (2 marks)

SECTION II (60 MARKS)

21. Use the supplier table given below to answer questions (a) to (d).

(a) Type the Hadoop HIVE commands to create Supplier table whose contents are as shown below. (5 marks)

1	**SUPPLIER NAME**	**COUNTRY**	**SALES AMOUNT**
2	GREAT WALL DISTRIBUTERS	NIGERIA	$224,000
3	ORUBA OIL	GHANA	$201,000
4	BOYA METAL DEALERS	NIGERIA	$193,000
5	BIDII LAPTOPS	NIGERIA	$184,000
6	KELLER RETAILERS	GHANA	$173,000
7	BOTTOM RETAILERS	KENYA	$124,000
8	JOHNSON AND JAY	UGANDA	$124,000
9	ZEAL HONEY	UGANDA	$123,000

(b) State the Hadoop HIVE command to insert the above data into the table. (5 marks)

(c) Outline the Hadoop HIVE command to add a column named Supplier Identification. (4 marks)

(d) List the Hadoop HIVE query that will list the supplier name and country for all suppliers with a sales amount equal to or higher than $150,000. (6 marks)

(Total: 20 marks)

22. Create a word processing document named “Question 22” and use the word processor document to save your answers to questions (a) to (g).

(a) Write a Python code that stores and displays the following items from a grocery store: Bread, Meat, Seafood, Pasta and Rice stored in a python data structure called “Grocery_store”. (3 marks)

(b) Write a Python code to access Meat as one of the food stuffs from the grocery store. (2 marks)

(c) What is the output of the expression len(Grocery_store)? (1 mark)

(d) Type a Python code to access the last item from the grocery store list. (2 marks)

(e) State a Python code to add the item "Vegetables" to the end of “Grocery_store” list. (2 marks)

(f) Type a Python code to remove the item "Meat" from the list. (2 marks)

(g) Create the Excel document shown below and save it as a comma separated version (CSV) file named “**Company**”. (3 marks)

CustomerID	Gender	Age	Annual income (Ksh)	Spending score (1-100)
1	Male	33	190	39
2	Male	15	210	89
3	Female	62	200	6
4	Female	55	230	77
5	Male	45	180	40
6	Male	15	180	76
7	Female	62	190	6
8	Female	31	190	94
9	Male	23	190	3
10	Female	12	170	72
11	Male	35	170	54
12	Female	65	160	6
13	Female	20	160	63
14	Male	53	200	71
15	Male	24	210	89

(h) Write a python programming code that will retrieve data from the **company.csv** file and print the first five rows. (5 marks)

Save and upload Question 22 document. **(Total: 20 marks)**

23. Create a word processing document named “Question 23” and use the word processor document to save your answers to questions (a) to (d).

(a) State the command used to install Psycopg2 library through Python. (3 marks)

(b) Write the python code to import the necessary libraries that will support linking Python to Postgres database. (4 marks)

(c) Describe **THREE** Hadoop Map Reduce classes. (9 marks)

(d) State **FOUR** ways to ensure big data synchronisation for database systems of any size. (4 marks)

Save and upload Question 23 document.

(Total: 20 marks)

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 24 April 2023. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions carrying forty (40) marks. SECTION II has three practical questions carrying sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required resources:
- Jupyter Notebook
- Java JDK
- Hadoop software
- Pycharm IDE
- Python

SECTION I (40 MARKS)

1. To combine the strengths of both batch and real-time processing for a more complete and accurate view of data, what type of architecture is used in big data? (2 marks)

2. The technique of dealing with uncertainty in data by using statistical techniques to model it and estimating the likelihood of different outcomes is known as? (2 marks)

3. In a distributed DBMS, the technique which ensure that multiple transactions are executed simultaneously while maintaining the ACID properties of the transactions and serialisability in the schedules is referred to as? (2 marks)

4. A Hadoop ecosystem component which provides a distributed data warehouse for data that is stored in HDFS and a query language that is based on SQL is referred to as? (2 marks)

5. What is the name given to a workflow scheduler system for Hadoop that is used to manage and coordinate complex data processing workflows? (2 marks)

6. Which type of data can be analysed as multiple data points per observation over time and measured by just as many ways as the spatial data. (2 marks)

7. Give one challenge associated with ensuring the quality of big data? (2 marks)

8. What is the name given to a process used to extract data from various sources, transform it into a format that is suitable for analysis, and load it into a target data store, such as a data warehouse or a big data platform? (2 marks)

9. What is the name given to a process of data cleaning where data is augmented with additional information from external sources to improve its quality? (2 marks)

10. In big data, the best term used to describe the process of choosing a smaller part of the available data set and using that subset for viewing or analysis is? (2 marks)

11. The process of managing the speed and frequency at which data is generated, processed, and analysed in big data is known as? (2 marks)

12. In the 5 V's of big data, name the one that comes from insight discovery and pattern recognition leading to more effective operations. (2 marks)

13. A programming model used to process large datasets in a distributed environment and to write programs that can be executed on a large number of machines in parallel is known as? (2 marks)

14. When deploying a model into a big data platform, the process that involves collecting data from different sources like social media platforms, business applications and log files is referred to as? (2 marks)

15. A collection of open source tools and platforms that are used for big data storage, processing, analysis, and management is known as? (2 marks)

16. State the term used to describe correlated data that have a relationship and are predictable. (2 marks)

17. The type of database where each item is saved in its own document with a partial schema leaving the raw information untouched is referred to as? (2 marks)

18. Name the type of architecture used by Hadoop? (2 marks)

19. Big data technology that focuses on the real-time processing of continuous streams of data in motion is referred to as? (2 marks)

20. The method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise is referred to as? (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (d)

Required:
Write Hadoop HIVE command that will:

(a) Create the supplier whose contents are as shown below. (5 marks)

DEALER	COUNTRY	YEAR_ESTABLISHED	SALESAMOUNT
AUDI	GERMANY	1909	$4,000,000
ROLL ROYCE	BRITAIN	1906	$950,000
TOYOTA	JAPAN	1937	$9,000,000
VOLVO	SWEDEN	1927	$1,500,000
LAMBORGHINI	ITALY	1963	$2,500,000
FORD	AMERICA	1903	$3,500,000
INNOSON	NIGERIA	2007	$1,200,000
LANDROVER	BRITAIN	1978	$2,550,000
MERCEDEZ	GERMANY	1926	$4,500,000

(b) Insert the above data into the table. (5 marks)

(c) Add a column named dealerid. (4 marks)

(d) List the Dealer and country for all suppliers with a sales amount equal to or higher than $1,500,000. (6 marks)

Capture a screenshot to demonstrate how you have performed the above task and save it in Question 21 document.
(Total: 20 marks)

22. (a) Describe the following Hadoop MapReduce Driver code. (10 marks)

```
1. Configuration conf= new Configuration();
2. Job job = new Job(conf,"My Word Count Program");
3. job.setJarByClass(WordCount.class);
4. job.setMapperClass(Map.class);
5. job.setReducerClass(Reduce.class);
6. job.setOutputKeyClass(Text.class);
7. job.setOutputValueClass(IntWritable.class);
8. job.setInputFormatClass(TextInputFormat.class);
9. job.setOutputFormatClass(TextOutputFormat.class);
10. Path outputPath = new Path(args[1]);
11. //Configuring the input/output path from the filesystem into the job
12. FileInputFormat.addInputPath(job, new Path(args[0]));
13. FileOutputFormat.setOutputPath(job, new Path(args[1]));
```

(b) Explain the K-means code snippet for the Hadoop MapReduce mapper code below (10 marks)

```
public static class Map extends Mapper<LongWritable,Text,Text,IntWritable> {
public void map(LongWritable key, Text value, Context context) throws IOException,InterruptedException
{
String line = value.toString();
StringTokenizer tokenizer = new StringTokenizer(line);
while (tokenizer.hasMoreTokens()) {
value.set(tokenizer.nextToken());
context.write(value, new IntWritable(1));
}
```

(Total: 20 marks)

23. (a) Describe the Hadoop commands to perform the following activities:

(i) Fulfil the purpose of testing the existence of a file in the HDFS cluster. (4 marks)

(ii) Display the allocated zip file in text format. (3 marks)

(iii) Search for files in the HDFS cluster. (3 marks)

(iv) Merge one or multiple files in a designated directory on the HDFS filesystem cluster. (3 marks)

(b) Give the functions of the node manager and application master components of the Hadoop YARN architecture. (7 marks)

(Total: 20 marks)

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 5 December 2022. Morning Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

SECTION I

1. What is the name given to big data that can be stored, accessed and processed in the form of fixed format? (2 marks)

2. What is the term given to heterogeneous sources characteristic and nature of big data? (2 marks)

3. Give an example of open source software framework used to develop data processing applications which are executed in a distributed computing environment. (2 marks)

4. Suggest a suitable name for a framework that clearly defines the components, layers to be used, and the flow of information in big data. (2 marks)

5. Give the meaning of temporal data mining. (2 marks)

6. _____ is a file system that is used to scale a single Apache Hadoop cluster to hundreds of nodes. (2 marks)

7. When measuring data quality, name the factor that ensures that data is in the appropriate format. (2 marks)

8. PBC Ltd. has recently acquired a data warehouse that deals with large clusters of similar data items. Recommend the most suitable database management system for supporting this warehouse. (2 marks)

9. The examination of large amounts of data to see what patterns or other useful information can be found is known as _____. (2 marks)

10. Name **ONE** common input format used in Hadoop file system. (2 marks)

11. Name **ONE** prominent Big Data visualisation tool. (2 marks)

12. In the 5 V's of big data which one illustrates how a company can obtain data from many different sources like from in-house devices to smartphone GPS technology? (2 marks)

13. _____is a type of NoSQL database that uses tables with rows and columns to support big data design. (2 marks)

14. Write YARN in full as used in data management. (2 marks)

15. _____is a way of processing massive quantities of data that provides access to batch-processing and stream-processing methods with a hybrid approach. (2 marks)

16. State one key big data infrastructure element. (2 marks)

17. List the command used to start up all the Hadoop daemons.

18. The MapReduce framework has seven important configuration parameters. Name **ONE** of these. (2 marks)

19. In a distributed database, concurrency control must be implemented. State **ONE** concurrency control technique. (2 marks)

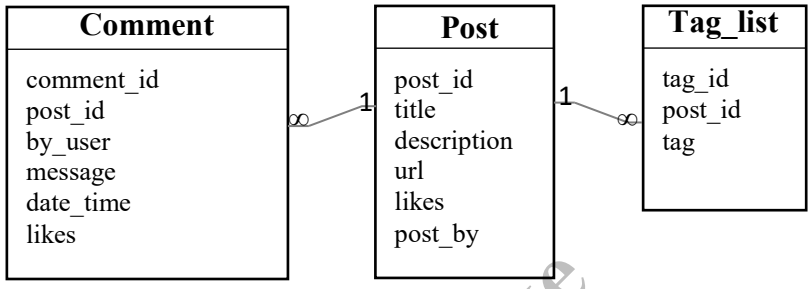
20. Name a big data technology that focuses on the real-time processing of continuous flows of data in motion. (2 marks)

SECTION II

21. Create a word processing document named "Question 21" and use it to save your answers to questions (a) to (c).

(a) Explain the **THREE** main steps for deploying a Big Data solution. (6 marks)

(b) Study the following schema of a relational database and Convert it into a MongoDB schema. (6 marks)



(c) An examiner graded students of Big Data Management as shown below. Create a stem and leaf plot and use it to explain the distribution of examination scores. (8 marks)

62, 64, 65, 65, 68, 70, 72, 72, 74, 75, 75, 75, 76,78, 78, 81, 83, 83, 84, 85, 87, 88, 92, 95, 98, 98, 100, 100.

Upload Question 21 Document.

(Total: 20 marks)

22. Create a word processing document named "Question 22" and use it to save your answers to questions (a) to (c).

(a) Explain **THREE** core Reduce functions. (6 marks)

(b) Consider a database with objects X and Y and assume that there are two transactions T1 and T2. T1 first reads X and Y and then writes X and Y. T2 reads and writes X then reads and writes Y. With adequate explanation, provide an example schedule that is **NOT** serialisable. (6 marks)

(c) Explain why the use of a Hierarchical Data File Store (HDFS) in products like Hadoop can offer significant advantages when processing Big Data. (8 marks)

Upload Question 22 document.

(Total: 20 marks)

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) to (c).

(a) Write the python MongoDB code to create a database called "BigData". (5marks)

(b) Write the python MongoDB code that will create a collection called "Lecturer" (5marks)

(c) Write the python MongoDB code that will insert the data in the table below into the "Lecturer" collection. (10 marks)

Lecturer Lastname	City	Salary	University	Specialisation
Matip	Nairobi	200000	UON	ICT
Mwaura	Mombasa	240000	JKUAT	Accounting
Odour	Nakuru	180000	KABARAK	Procurement
Ouna	Kisumu	150000	MASENO	ICT
Mwende	Kampala	200000	MAKERERE	Accounting
Njeri	Kisumu	240000	JOOUST	Mathematics
Achieng	Nakuru	305000	EGERTON	ICT
Kerubo	Nairobi	220000	KCAU	Procurement
Grant	Nakuru	140000	EGERTON	Accounting

Upload Question 23 document.

(Total: 20 marks)

……

CISSE ADVANCED LEVEL

ELECTIVE I

BIG DATA MANAGEMENT

MONDAY: 1 August 2022. Morning paper. **Time Allowed: 3 hours.**

This paper has two sections. SECTION I has twenty (20) short response questions of forty (40) marks. SECTION II has three practical questions of sixty (60) marks. All questions are compulsory. Marks allocated to each question are shown at the end of the question.

SECTION I

1. The _____precomputes results using a distributed processing system that can handle very large quantities of data. It aims at perfect accuracy by being able to process all available data when generating views. (2 marks)

2. The _____includes one fact table which is connected to several dimension tables, which can be connected to other dimension tables through a many-to-one relationship. (2 marks)

3. A set of illegal instructions that are inserted into a legitimate computer program is known as _____. (2 marks)

4. It is a column-oriented non-relational database management system that runs on top of Hadoop Distributed File System (HDFS). It provides a fault-tolerant way of storing sparse data sets, which are common in many big data use cases. It is well suited for real-time data processing or random read/write access to large volumes of data. This is the _____

5. Which is the method where analytical data is loaded into **memory** for live calculations and querying? (2 marks)

6. An _____ is an observation that lies an abnormal distance from other values in a random sample from a population. (2 marks)

7. _____ tracker plays the role of scheduling jobs and tracking all jobs assigned to the task tracker. (2 marks)

8. Identify the operator that is used to execute a shell command from the Hive shell. (2 marks)

9. Data _____ is the process of moving data from various sources into a central repository such as a data warehouse where it can be stored, accessed, analysed, and used by an organisation. (2 marks)

10. A _____is a data visualization tool that is a graphical portrayal of data that uses different colors to address different values. This difference in color representation makes it easy for the viewers to understand the trend more quickly. (2 marks)

11. MongoDB does not require a relational database management system (RDBMS), so it provides an elastic data storage model that enables users to store and query multivariate data types with ease. The key-value pair that forms the basic unit of data in MongoDB is known as a _____. (2 marks)

12. The open-source software that facilitates the collecting, aggregating and moving of huge amounts of unstructured, streaming data such as log data and events in data big data analytics is called? (2 marks)

13. In graph data structure _____ describes how a connection between a source node and a target node are related. (2 marks)

14. _____is an open-source, distributed processing system used for big data workloads and utilises in-memory caching, and optimised query execution for fast analytic queries against data of any size. (2 marks)

15. _____ is a file system that handles large data sets running on commodity hardware? It is used to scale a single Apache Hadoop cluster to hundreds of nodes. (2 marks)

16. In distributed computing the ability to support billions of job requests over massive datasets is referred to as _____. (2 marks)

17. _____ is a set of processes and activities across an organisation to support a data strategy and guide technology, people and data procedures. (2 marks)

18. This is the active and continuous management of data through its life cycle. It includes the organisation and integration of data from multiple sources and to ascertain that the data meets given quality requirements for its usage. It covers tasks related to controlled data creation, maintenance and management such as content creation, selection, validation or preservation. This activity in the big data value chain is referred to as_____. (2 marks)

19. _____ architecture is a way of processing massive quantities of data that provides access to batch-processing and stream-processing methods with a hybrid approach (2 marks)

20. There are three type of big data; structured, semi-structured or unstructured. In which category does XML data belongs to? (2 marks)

SECTION II

21. Create a word processing document named "Question 21". Use the word processor document to save your answers to questions (a) to (e) in form of screenshots.

(a) Demonstrate how you would create the table below and save it as **SALES DATA** in the **YEAR** directory of drive C then load the data of the file into an external hive table using an appropriate command. (4 marks)

Year	Month	Type	Salesperson	Region	Sales	Units
2013	January	Ice Cream	Bishop	West	$2,395.50	1597
2013	January	Ice Cream	Bishop	West	$11,761.50	7841
2013	January	Frozen Yogurt	Bishop	West	$8,943.00	5962
2013	January	Ice Cream	Bishop	West	$2,395.50	1597
2013	January	Ice Cream	Bishop	West	$11,761.50	7841
2013	January	Frozen Yogurt	Bishop	West	$8,943.00	5962
2013	January	Frozen Yogurt	Lee	Central	$14,596.50	9731
2013	January	Tasty Treats	Lee	Central	$8,793.00	5862
2013	January	Frozen Yogurt	Lee	Central	$14,596.50	9731
2013	January	Tasty Treats	Lee	Central	$8,793.00	5862

(b) Explain how you could retrieve the data from the database using an appropriate command and display the output. (3 marks)

(c) Type the Hive function that will display the Month and Region for all Sales that are greater than $11000. (5 marks)

(d) Type the Hive function that will count the Units that are realised from sales by sales person called Lee. (4 marks)

(e) Explain the Hive function that will display the year and type for all sales from the Central region then order by type in descending order (4 marks)

Upload question 21 document.

(Total: 20 marks)

22. Create a word processing document named "Question 22". Use the word processor document to save your answers to questions (a) to (c) in form of screenshots.

(a) Give the commands showing how you will use a database to create and insert data into a table called Grade whose contents are as shown below.

(8 marks)

STUDNO	STUDFNAME	STUDCSECODE	STUDCSENAME	GRADE	LECTID	LECTNAME
KB01	JAMES	DB001	DATABASE	PASS	MM003	JAMES
KB02	JAMES	DB001	DATABASE	CREDIT	MM003	JAMES
KB03	ALICE	SSOO2	PROGRAMMING	DISTINCTION	MM003	ALLAN
KB05	ALLAN	AG003	NETWORKS	CREDIT	MM006	DIANA

(b) State the Hadoop pig commands that can be used to perform the following

(i) Pasting spark code lines in the shell. (2 marks)

(i) Execute system commands and check return code. (3 marks)

(ii) Executing system commands and checking output. (3 marks)

(c) Type the command for creating a database to store tables in Hadoop HIVE. (4 marks)

Upload question 22 document.

(Total: 20 marks)

23. Create a word processing document named "Question 23". Use the word processor document to save your answers to questions (a) to (d) in form of screenshots.

(a) Type the following dataset into excel worksheet. Create an HDFS folder in drive C: and name it **BIG DATA**.

Save the excel worksheet in **BIG DATA** folder as a comma separated version (CSV) file and name it ***Candidate.csv*** (5 marks)

	A	B	C	D	E
1					
2	**CANDIDATE NAME**	**LECTURER ID**	**COURSE**	**UNIVERSITY**	**GRADE**
3	HILLARY MAINA	LEC003	ENGINEERING	JKUAT	PASS
4	GRACE AMONDI	LEC005	ACCOUNTING	UON	CREDIT
5	NIGEL MWENDIA	LEC006	FINANCE	KCAU	DISTINCTION
6	NJERI SUSAN	LEC004	ENGINEERING	JKUAT	PASS
7	JOSEPH MAKOKHA	LEC008	ARCHITECTURE	STRATHMORE	PASS
8	OYARO DISMAS	LEC007	ICT	UON	CREDIT
9	OPIYO HENRY	LEC002	HOSPITALITY	KCAU	DISTINCTION
10	ONYANGO ELVIN	LEC005	FINANCE	CATHOLIC	CREDIT
11	TABITHA MWENDE	LEC010	ACCOUNTING	UON	PASS
12	FREDRICK MWORIA	LEC009	ARCHITECTURE	STRATHMORE	DISTINCTION
13	ROY OBONYO	LEC002	HOSPITALITY	USIU	CREDIT
14	FELIX MARANGO	LEC007	ENGINEERING	KCAU	CREDIT
15	RODGERS SIMIYU	LEC008	ICT	USIU	PASS
16	FIONA NJERI	LEC004	FINANCE	JKUAT	DISTINCTION
17					

(b) Explain how you would use a command to load the data into an external hive table with column labels and show the table contents. (6 marks)

(c) Explain the procedure of moving the external table to an internal Hive table that uses the ORC format and show the table contents.

(5 marks)

(d) Stat the command for creating a partitioned table from the internal hive table on question (c) above.

(4 marks)

Upload question 23 document.

(Total: 20 marks)

……