

Disclaimer: The questions given here are indicative of the format of questions in the MCQ section of the end semester exam, and not exhaustive. The final question paper for your exam will be set by the University and will have a wide range of questions. Please use this model paper in conjunction with your other course materials to prepare well for your exams.

Jain deemed to be University		Course Title: Big Data Management & Analytics				
S no	Question	A	B	C	D	Correct Answer
1	What happens when the same data is stored multiple times in a database?	Data redundancy	Data inconsistency	Data is overwritten	Data is lost	Data redundancy
2	What is the aim of NoSQL?	NoSQL is not suitable for storing structured data.	NoSQL databases allow storing non-structured data.	NoSQL is a new data format to store large datasets.	NoSQL provides an alternative to SQL databases to store textual data.	NoSQL is a new data format to store large datasets.
3	NoSQL databases are used mainly for handling large volumes of what type of data?	unstructured	structured	semi-structured	half-structured	unstructured
4	Which of the following is a characteristic of a NoSQL database?	Uses JSON	Needs a schema	Required JOINS	Uses tables for storage	Uses JSON
5	Cassandra was developed at?	IBM	Facebook	Google	Oracle	Facebook

Disclaimer: The questions given here are indicative of the format of questions in the MCQ section of the end semester exam, and not exhaustive. The final question paper for your exam will be set by the University and will have a wide range of questions. Please use this model paper in conjunction with your other course materials to prepare well for your exams.

Jain deemed to be University		Course Title: Big Data Management & Analytics				
S no	Question	A	B	C	D	Correct Answer
6	Cassandra uses which of the following in the background to allow the nodes to communicate with each other and detect any faulty nodes in the cluster.	Gossip Protocol	Commit Protocol	Mem Protocol	Talk Protocol	Gossip Protocol
7	Apache Cassandra is a massively scalable open source database of?	SQL	NoSQL	NewSQL	OldSQL	NoSQL
8	Point out the wrong statement.	Cassandra delivers continuous availability, linear scalability, and operational simplicity across many commodity servers	Cassandra has a “masterless” architecture, meaning all nodes are the same	Cassandra also provides customizable replication, storing redundant copies of data across nodes that participate in a Cassandra ring	Apache Cassandra is a relational database management system	Apache Cassandra is a relational database management system
9	Which of the following database uses key-based indexing?	RDBMS	AWS Lambda	Apache Cassandra	Apache Hive	Apache Cassandra
10	What is the maximum size the Memtable can take with 64GB of total storage?	32GB	64GB	16GB	4 GB	16GB
11	What type of database is Apache Cassandra?	Relational database	NoSQL database	Graph database	Columnar database	NoSQL database

Disclaimer: The questions given here are indicative of the format of questions in the MCQ section of the end semester exam, and not exhaustive. The final question paper for your exam will be set by the University and will have a wide range of questions. Please use this model paper in conjunction with your other course materials to prepare well for your exams.

Jain deemed to be University		Course Title: Big Data Management & Analytics				
S no	Question	A	B	C	D	Correct Answer
12	You need a schemaless database. Which Amazon database service provides that solution?	ORDS	Aurora	Redshift	DynamoDB	DynamoDB
13	Which of the following API is used to modify one or more attributes in an item in AWS DynamoDB?	UpdateItems	Update_Items	UpdateItem	Update-item	UpdateItem
14	Which statement regarding DynamoDB Streams is False?	There are four Stream views.	When a Stream is enabled, it holds changes to the entire database.	DynamoDB and its Stream are highly resilient.	Streams can be used to invoke Lambda functions.	When a Stream is enabled, it holds changes to the entire database.
15	An online social media platform receives data from multiple users. Apart from the user's profile ID number, all other details are received without any prior knowledge of their format. As a Cloud Specialist, what would you recommend the platform?	RDBMS table with each row indicating each profile ID	HTML page with profile ID	Kinesis Stream with DynamoDB	Kinesis Stream with Amazon EMR	Kinesis Stream with DynamoDB

Disclaimer: The questions given here are indicative of the format of questions in the MCQ section of the end semester exam, and not exhaustive. The final question paper for your exam will be set by the University and will have a wide range of questions. Please use this model paper in conjunction with your other course materials to prepare well for your exams.

Jain deemed to be University		Course Title: Big Data Management & Analytics				
S no	Question	A	B	C	D	Correct Answer
16	What is Lambda Architecture?	An architectural pattern for building web applications	An architectural pattern for building data processing systems	An architectural pattern for building distributed computing systems	An architectural pattern for building mobile applications	An architectural pattern for building data processing systems
17	What is the purpose of the batch layer in Lambda Architecture?	To perform real-time processing of incoming data	To store and process large volumes of historical data	To provide low-latency access to query results	To serve as a buffer for incoming data before processing	To store and process large volumes of historical data
18	What is the purpose of the speed layer in Lambda Architecture?	To perform real-time processing of incoming data	To store and process large volumes of historical data	To provide low-latency access to query results	To serve as a buffer for incoming data before processing	To perform real-time processing of incoming data
19	What are the three types of data sources in Lambda Architecture?	Historical data, real-time data, and metadata	Batch data, streaming data, and unstructured data	Structured data, semi-structured data, and unstructured data	Internal data, external data, and user-generated data	Batch data, streaming data, and unstructured data

Disclaimer: The questions given here are indicative of the format of questions in the MCQ section of the end semester exam, and not exhaustive. The final question paper for your exam will be set by the University and will have a wide range of questions. Please use this model paper in conjunction with your other course materials to prepare well for your exams.

Jain deemed to be University		Course Title: Big Data Management & Analytics				
S no	Question	A	B	C	D	Correct Answer
20	Which service in YARN allocates resources to an application?	Node Manager	Resource Manager	YARN Manager	Application Manager	Resource Manager
21	Changes to data in the Hadoop framework are logged into which log files?	Log File, Edit Log	Log File, FS Search	Edit Log File, FS Image	Edit Log Directory, FS Search	Edit Log File, FS Image
22	Point out the correct statement.	Hive Commands are non-SQL statement such as setting a property or adding a resource	Set -v prints a list of configuration variables that are overridden by the user or Hive	Set sets a list of variables that are overridden by the user or Hive	Hive is a relational database management system (RDBMS) that uses the Hadoop Distributed File System (HDFS) for storing and processing data	Set -v prints a list of configuration variables that are overridden by the user or Hive

Disclaimer: The questions given here are indicative of the format of questions in the MCQ section of the end semester exam, and not exhaustive. The final question paper for your exam will be set by the University and will have a wide range of questions. Please use this model paper in conjunction with your other course materials to prepare well for your exams.

Jain deemed to be University		Course Title: Big Data Management & Analytics				
S no	Question	A	B	C	D	Correct Answer
23	What is Apache Hive?	A data warehouse system built on top of Hadoop that provides SQL-like query capabilities to data stored in HDFS	A relational database management system that provides SQL-like query capabilities to data stored in HDFS	A distributed file system that allows for parallel processing of large datasets	A data processing framework that provides stream processing capabilities to data stored in HDFS	A data warehouse system built on top of Hadoop that provides SQL-like query capabilities to data stored in HDFS
24	What is Databricks?	A cloud-based data processing and analytics platform	An on-premises data processing and analytics platform	A machine learning library	An open-source data processing engine	A cloud-based data processing and analytics platform
25	Which of the following operations can be used to convert a DataFrame column from one type to another type?	col().cast()	convert()	castAs()	col().coerce()	col().cast()
26	Which of the following data structures are Spark DataFrames built on top of?	Arrays	Strings	RDDs	Vectors	RDDs

Disclaimer: The questions given here are indicative of the format of questions in the MCQ section of the end semester exam, and not exhaustive. The final question paper for your exam will be set by the University and will have a wide range of questions. Please use this model paper in conjunction with your other course materials to prepare well for your exams.

Jain deemed to be University		Course Title: Big Data Management & Analytics				
S no	Question	A	B	C	D	Correct Answer
27	What is the primary purpose of using Spark in a data pipeline using Flume, Spark, and S3?	To process and analyze data	To store data	To ingest data	To query data	To process and analyze data
28	What is the primary purpose of using S3 in a data pipeline using Flume, Spark, and S3?	To process and analyze data	To store data	To ingest data	To query data	To store data
29	What is streaming data?	Data that is generated and processed in batches	Data that is generated and processed in real-time	Data that is generated and processed offline	Data that is generated and processed on a schedule	Data that is generated and processed in real-time
30	What is the primary challenge of analyzing streaming data?	Processing large amounts of data quickly	Storing large amounts of data reliably	Ensuring data accuracy and completeness	Performing complex queries and analysis	Processing large amounts of data quickly