Charotar University of Science and Technology [CHARUSAT]

CSPIT / DEPSTAR

Department of Computer Science & Engineering

Course: CS442 Data Science & Analytics (SEM-7)

QUESTION BANK

What is Big Data? Discuss Four V's of Big Data. Explain characteristics of Big Data. Explain how big data processing differs from distributed processing.

List various application of big data. How it can be used to improve business for a superstore.

Explain "Map Phase" and "Combiner Phase" in Map-Reduce. Explain working of reduce phase of Map-Reduce with an example. Explain "Shuffle & Sort" phase and "Reducer Phase" in Map-Reduce.

List various configuration files used in Hadoop Installation. What is use of mapred-site.xml?

How to create collection in MongoDB? Explain with its syntax. Explain MongoDB sharing process. Which terms are used for table, row, column and table-join in MongoDB?

What is the role of a profiler in MongoDB? Where does the writes all the data?

Explain structured, semi structured and unstructured data in terms of big data analytics.

Differentiate between: SQL and NoSQL

Write down the differences between Apache Pig and Map-Reduce. Explain each of it.

Explain the components of SPARK. Explain about the major libraries that constitute the Spark Ecosystem. How can you minimize data transfers when working with Spark?

Write down the goals of HDFS.

What is Zookeeper? What are the benefits of Zookeeper?

How Big Data Analytics can be useful in the development of smart cities

What is Compute and Storage nodes in Hadoop?

What are primary and secondary replica sets in MongoDB?

How are big data and Hadoop related to each other?

What are the applications of Big data?

Define Big Data and Explain the Five Vs of Big Data. Explain the steps to be followed to deploy a Big Data solution.

Define HDFS and YARN, and talk about their respective components.

Explain structured, semi structured and unstructured data in terms of big data analytics.

Explain working of reduce phase of MapReduce with an example.

Charotar University of Science and Technology [CHARUSAT]

CSPIT / DEPSTAR

Department of Computer Science & Engineering

Course: CS442 Data Science & Analytics (SEM-7)

Define HDFS. Describe namenode, datanode and block. Explain HDFS operations in detail. Explain characteristics of Big Data.

What is Big data? Discuss it in terms of volume and velocity.

Why do we need Hadoop for Big Data Analytics?

What is HBase? Explain storage mechanism of HBase with an example.

How to create collection in MongoDB? Explain with its syntax.

Explain CRUD operations of MongoDB with an example. Explain MongoDB sharing process.

What is NoSQL? List out the features of NoSQL. Explain types of NoSQL databases in brief

Write Map Reduce steps for counting occurrences of specific numbers in the input text file(s). Also write the commands to compile and run the code.

Explain “Shuffle & Sort” phase and “Reducer Phase” in MapReduce.

Write difference between MangoDB and Hadoop.

What is NoSQL database? List the differences between NoSQL and relational databases. Differentiate between SQL and NoSQL Explain in brief various types of NoSQL databases in practice.

Explain scaling in MangoDB

What is Resilient Distributed Dataset in Apache Spark? Explain in detail. Make a note on why RDD is better than Map Reduce data storage?

Which terms are used for table, row, column and table-join in MongoDB?

What are the common input formats in Hadoop?

Explain the core components of Hadoop. What are the different configuration files in Hadoop?

What will happen with a NameNode that doesn’t have any data? Explain NameNode recovery process.

Define respective components of HDFS and YARN

What do you understand by Rack Awareness in Hadoop?

Explain the difference between Hadoop and RDBMS.

What are the Port Numbers for NameNode, Task Tracker, and Job Tracker?

What are the different file permissions in HDFS for files or directory levels?

What are the basic parameters of a Mapper?

Charotar University of Science and Technology [CHARUSAT]

CSPIT / DEPSTAR

Department of Computer Science & Engineering

Course: CS442 Data Science & Analytics (SEM-7)

Explain the process that overwrites the replication factors in HDFS.

What are Edge Nodes in Hadoop?

Explain the core methods of a Reducer.

Explain core architecture of Hadoop with suitable block diagram. Discuss role of each component in detail.

What is Hadoop Ecosystem? Discuss various components of Hadoop Ecosystem.

List various configuration files used in Hadoop Installation. What is use of mapred-site.xml?

Define HDFS. Discuss the HDFS Architecture and HDFS Commands in brief.

What is RDD? Explain transformations and actions in RDD. Explain RDD operations in brief.

Write down the differences between Apache Pig and MapReduce.

What is Apache Pig and why do we need it?

Explain the components of SPARK.

Explain the architecture and features of HIVE. Explain Metastore in Hive.

Explain Spark components in detail. Also list the features of spark

Write a brief short note on: Spark Unified Stack

What is Zookeeper? What are the benefits of Zookeeper?

Draw architecture of APACHE PIG and explain in short.

Discuss Machine Learning with MLlib in SPARK

Discuss the applications of big data analytics in weather forecasting. List various application of big data. How it can be used to improve business for a superstore. How can Big Data Analytics be useful in the development of smart cities?

What are the benefits of Big Data? Discuss challenges under Big Data. How Big Data Analytics can be useful in the development of smart cities.(Discuss one application)

Elaborate the working of Map-Reduce Algorithm. What do you mean by heartbeat and replica in Hadoop?

Explain “Shuffle & Sort” phase and “Reducer Phase” in MapReduce.

What is Apache Pig and why do we need it? Write down the differences between Apache Pig and MapReduce.

Charotar University of Science and Technology [CHARUSAT]

CSPIT / DEPSTAR

Department of Computer Science & Engineering

Course: CS442 Data Science & Analytics (SEM-7)

Compare Raw oriented and Column Oriented database structures.

Compare Cassandra with HBase and MongoDB.

Compare RDBMS with Neo4j

Explain the working Neo4j with proper steps and diagram.

Explain the working MongoDB with proper steps and diagram.

Explain NewSQL. Explain the characteristics of NewSQL.

Explain the concept of conditional probability.

What does it mean to calculate the probability of an event given that another event has occurred? Provide a real-world example to illustrate this concept.

What is Bayes' Theorem? Explain Bayes' Theorem and describe its importance in statistical inference. Why is it particularly useful when we have limited information about probabilities?

Differentiate between conditional probability and joint probability.

Explain how these two concepts are related and give an example of each.

Real-Life Interpretation of Conditional Probability in Insurance:

In health insurance, a person's premium may be based on conditional probabilities related to age, medical history, and lifestyle. Explain how conditional probability is used in calculating insurance premiums and why it's important for insurers to consider various conditional factors.

Define Bayes' Theorem.
What does Bayes' Theorem state, and how does it relate conditional probability to prior knowledge?

Explain the components of Bayes' Theorem.
In Bayes' Theorem, what do the terms P(A|B)P(A|B)P(A|B), P(B|A)P(B|A)P(B|A), P(A)P(A)P(A), and P(B)P(B)P(B) represent?

What role does prior probability play in Bayes' Theorem?

Define the following terms related to data pre-processing:
(a) Data cleaning, (b) Data transformation, (c) Data integration, and (d) Data reduction.

What is data visualization?
Explain the importance of data visualization in data analysis and decision-making.

Charotar University of Science and Technology [CHARUSAT]

CSPIT / DEPSTAR

Department of Computer Science & Engineering

Course: CS442 Data Science & Analytics (SEM-7)

Explain how you would use bar charts and pie charts to display the frequency distribution of a categorical variable. When might you choose one type over the other?

Principal Component Analysis (PCA) is often used in high-dimensional datasets. Explain how PCA reduces the dimensionality of data and how it might affect data interpretation in visualization.

What is Hadoop?

Explain what Hadoop is and briefly describe its purpose in big data processing.

List the core components of Hadoop.

Identify the primary components of the Hadoop ecosystem and describe their functions.

What is HDFS?

Explain the Hadoop Distributed File System (HDFS), its purpose, and its primary characteristics.

Describe the roles of NameNode and DataNode in Hadoop.

Explain how these two components work together in HDFS.

What is the role of the YARN ResourceManager?

Describe the role of the ResourceManager in YARN and how it contributes to resource management in Hadoop.

What are the main functions of MapReduce in Hadoop?

Briefly explain what MapReduce is and how it processes data in Hadoop.

Describe the sequence of steps that occur when a MapReduce job is submitted in Hadoop, from the job submission to completion.

Explain how YARN schedules resources for multiple jobs in a multi-tenant environment, and describe the role of the ApplicationMaster in this process.

Why is HDFS optimized for large files? What challenges might arise if it's used for small files, and how could this issue be managed?

Explain how Hadoop handles task failures during a MapReduce job. If a node fails while processing a map task, describe the steps Hadoop takes to ensure the job can still complete successfully.

****