BIG DATA ANALYTICS AND VISUALIZATION LAB

Certainly! Here are some viva questions and answers for each topic:

### 1. Hadoop Setup and Configuration

\*\*Question 1:\*\* What is Hadoop and what is its primary purpose?

\*\*Answer:\*\* Hadoop is an open-source framework for distributed storage and processing of large data sets. Its primary purpose is to store and process big data in a distributed and fault-tolerant manner.

\*\*Question 2:\*\* How do you set up a Hadoop cluster using Cloudera?

\*\*Answer:\*\* You can use Cloudera Manager to set up and manage a Hadoop cluster. It provides a web-based interface to install, configure, manage, and monitor Hadoop clusters.

(Hadoop clusters are composed of a network of master and worker nodes that orchestrate and execute the various jobs across the Hadoop distributed file system)

\*\*Question 3:\*\* Explain the HDFS architecture.

\*\*Answer:\*\* HDFS (Hadoop Distributed File System) consists of a NameNode that manages metadata and DataNodes that store the actual data. The data is divided into blocks and distributed across multiple DataNodes for parallel processing.

\*\*Question 4:\*\* Provide some HDFS commands.

\*\*Answer:\*\*

- To create a directory: `hadoop fs -mkdir /user/mydirectory`

- To copy a file to HDFS: `hadoop fs -copyFromLocal localfile /user/hadoop/hdfsfile`

- To list files in a directory: `hadoop fs -ls /user/mydirectory`

### 2. MapReduce Programming

\*\*Question 1:\*\* What is MapReduce and how does it work?

\*\*Answer:\*\* MapReduce is a programming model for processing and generating large datasets. It works by breaking down a task into smaller sub-tasks, processing them in parallel, and then aggregating the results.

\*\*Question 2:\*\* Explain the Word Count example in MapReduce.

\*\*Answer:\*\* Word Count is a simple MapReduce program that counts the frequency of each word in a given set of documents. It involves a Map phase to tokenize and count words, followed by a Reduce phase to aggregate the counts.

\*\*Question 3:\*\* Provide examples of MapReduce operations for Union, Intersection, and Difference.

\*\*Answer:\*\*

- Union: Concatenate the output of multiple MapReduce jobs.

- Intersection: Identify common elements in the outputs of two MapReduce jobs.

- Difference: Identify elements present in one output but not in another.

\*\*Question 4:\*\* How does Matrix Multiplication work in MapReduce?

\*\*Answer:\*\* Matrix Multiplication in MapReduce involves breaking down the multiplication task into smaller sub-tasks, distributing them across nodes for parallel processing, and aggregating the results.

### 3. MongoDB

\*\*Question 1:\*\* What is MongoDB?

\*\*Answer:\*\* MongoDB is a NoSQL database that stores data in a flexible, JSON-like format called BSON (BSON, or Binary JSON, is the data format that MongoDB uses to organize and store data).

It is known for its scalability, flexibility, and ability to handle unstructured data.

\*\*Question 2:\*\* How do you perform CRUD operations in MongoDB?

\*\*Answer:\*\*

- Insert: `db.collection.insert()`

- Query: `db.collection.find()`

- Update: `db.collection.update()`

- Delete: `db.collection.remove()`

\*\*Question 3:\*\* What is HBASE, and how is it related to MongoDB?

\*\*Answer:\*\* HBASE is another NoSQL database that is part of the Hadoop ecosystem. It is column-oriented and designed for fast and random access to large amounts of data. While both MongoDB and HBASE are NoSQL databases, they have different architectures and use cases.

### 4. Hive

\*\*Question 1:\*\* What is Hive in the context of Hadoop?

\*\*Answer:\*\* Hive is a data warehousing and SQL-like query language system built on top of Hadoop. It provides a high-level interface to process and analyze data stored in Hadoop Distributed File System (HDFS).

\*\*Question 2:\*\* How do you create a database and table in Hive?

\*\*Answer:\*\*

- Create a database: `CREATE DATABASE mydatabase;`

- Use a database: `USE mydatabase;`

- Create a table: `CREATE TABLE mytable (column1 INT, column2 STRING);`

\*\*Question 3:\*\* Explain Hive Partitioning.

\*\*Answer:\*\* Hive Partitioning is a way to organize data in a table based on one or more columns, allowing for more efficient queries by eliminating the need to scan the entire table.

\*\*Question 4:\*\* How do you configure Hive Metastore to MySQL?

\*\*Answer:\*\* To configure Hive Metastore to MySQL, you need to modify the Hive configuration files (hive-site.xml) to point to the MySQL database for storing metadata. Update the JDBC URL, username, and password accordingly.

### 5. Pig

\*\*Question 1:\*\* What is Apache Pig?

\*\*Answer:\*\* Apache Pig is a high-level scripting language built on top of Hadoop for processing and analyzing large datasets. It simplifies the process of writing MapReduce programs.

\*\*Question 2:\*\* What is Pig Latin?

\*\*Answer:\*\* Pig Latin is the scripting language used in Apache Pig. It is a data flow language that provides operators to perform various transformations on data.

\*\*Question 3:\*\* How do you read and store data in Pig?

\*\*Answer:\*\*

- Read data: `mydata = LOAD '/path/to/data' USING PigStorage(',') AS (field1:chararray, field2:int);`

- Store data: `STORE mydata INTO '/path/to/output';`

### 6. Spark

\*\*Question 1:\*\* What is Apache Spark?

\*\*Answer:\*\* Apache Spark is an open-source, distributed computing system that provides fast and general-purpose data processing capabilities. It can be used for batch processing, real-time data streaming, machine learning, and graph processing.

\*\*Question 2:\*\* Explain RDD in Spark.

\*\*Answer:\*\* RDD (Resilient Distributed Dataset) is the fundamental data structure in Spark. It represents an immutable distributed collection of objects that can be processed in parallel.

\*\*Question 3:\*\* What is Lazy Execution in Spark?

\*\*Answer:\*\* Lazy Execution in Spark means that transformations on RDDs are not executed immediately. Instead, they are recorded, and the actual computation is performed only when an action is called.

\*\*Question 4:\*\* How do you persist an RDD in Spark?

\*\*Answer:\*\* You can persist an RDD in Spark using the `persist()` or `cache()` methods. This allows the RDD to be stored in memory or on disk and reused across multiple Spark operations.

\*\*Question 5:\*\* Provide an example of a machine learning algorithm implemented using Spark.

\*\*Answer:\*\* K-Means clustering is a common machine learning algorithm implemented in Spark. It can be used for clustering similar data points.

### 7. Visualization with Tableau

\*\*Question 1:\*\* What is Tableau?

\*\*Answer:\*\* Tableau is a data visualization and business intelligence tool that allows users to create interactive and shareable dashboards, reports, and charts.

\*\*Question 2:\*\* How do you connect to data in Tableau?

\*\*Answer:\*\* In Tableau, you can connect to data sources such as databases, Excel files, or web data connectors. You can import data or establish a live connection, depending on your requirements.

\*\*Question 3:\*\* Explain the process of building charts and analyzing data in Tableau.

\*\*Answer:\*\* In Tableau, you can drag and drop fields onto the canvas to create visualizations. You can choose from various chart types, apply filters, and analyze data dynamically.

\*\*Question 4:\*\* What is the difference between a Tableau Dashboard and a Tableau Story?

\*\*Answer:\*\* A Tableau Dashboard is an interactive collection of visualizations and sheets, while a Table

au Story is a sequence of sheets or dashboards that work together to convey a narrative.

\*\*Question 5:\*\* How can Tableau be used on the web?

\*\*Answer:\*\* Tableau offers web-based solutions, allowing users to publish and share Tableau workbooks on Tableau Server or Tableau Online. Users can access and interact with the visualizations through a web browser.

These questions cover a broad range of topics related to Hadoop, MapReduce, MongoDB, Hive, Pig, Spark, and Tableau. They should provide a good basis for assessing someone's understanding of these technologies and their applications.