IMAAD IMRAN HAJWANE

202101132 / 21

TOPIC: INSTALLATION OF HIVE IN UBUNTU

BIG DATA LAB 4

What is Hive?

Apache Hive is a data warehousing and SQL-like query system built on top of Hadoop. It allows users to write queries in a language called HiveQL (similar to SQL) to interact with data stored in Hadoop Distributed File System (HDFS). Hive abstracts the complexity of Hadoop's underlying MapReduce framework, allowing users to query and manage large datasets without needing to write low-level MapReduce code.

Basic Components of Hive:

1. **HiveQL (Hive Query Language)**:
    - A SQL-like query language used to perform operations like data retrieval, insertion, and analysis.
2. **Metastore**:
    - Stores metadata about the databases, tables, columns, data types, and table locations in HDFS. The metastore allows Hive to retrieve information quickly when executing queries.
3. **Driver**:
    - Acts as a controller for managing the execution of the queries, compiling the queries into MapReduce jobs, and executing them.
4. **Compiler**:
    - Converts HiveQL queries into MapReduce tasks. The query is parsed and optimized before sending it to the execution engine.
5. **Execution Engine**:
    - Executes the MapReduce jobs generated by the compiler and returns the results back to the user.
6. **Hive Server**:
    - It allows clients to connect to Hive and submit queries. This server can handle multiple requests from various clients simultaneously.

7. **CLI (Command Line Interface)**:
    - Provides an interface where users can execute HiveQL queries and interact with the Hive system.

Architecture of Hive:

The architecture of Hive is composed of several key components that interact with Hadoop and provide a user-friendly interface for managing large datasets.

1. **User Interface**:
    - Users interact with Hive via a web UI, command line interface (CLI), or JDBC/ODBC interface. This is where queries are written and submitted to Hive.
2. **HiveQL Process Engine**:
    - Once a query is submitted, the HiveQL process engine parses the HiveQL, checks for errors, and compiles it into a directed acyclic graph (DAG) of MapReduce tasks.
3. **Metastore**:
    - The metastore holds all metadata about the tables in Hive, such as table schemas, column types, and table locations. This metadata is crucial for query execution, as it helps in locating the data and understanding the structure of the datasets.
4. **Execution Engine**:
    - The execution engine is responsible for the actual execution of the queries. It takes the compiled MapReduce jobs and runs them in the Hadoop cluster.
5. **HDFS** (Hadoop Distributed File System):
    - Hive stores its data on HDFS, Hadoop's underlying file system. All data queried or inserted via Hive is stored in HDFS.
6. **Driver**:
    - The driver is responsible for managing the query lifecycle. It coordinates between different stages of the query, including the compilation, optimization, and execution phases. It also handles the communication between Hive and the metastore.

Usage of Hive:

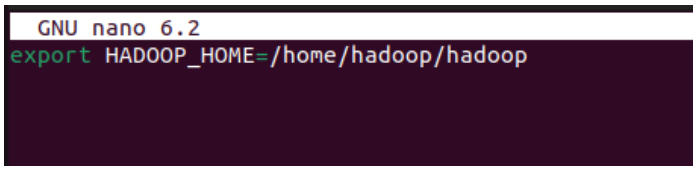
1. **Data Warehousing**:
    - Hive is primarily used for performing data warehousing tasks on large datasets. It provides the ability to run complex analytical queries over large amounts of structured and semi-structured data.
2. **Handling Structured Data**:
    - Hive is ideal for handling structured data, with support for multiple formats such as text files, ORC, Parquet, etc. It allows users to organize data into tables, partitions, and buckets.
3. **Batch Processing**:
    - Hive is designed for batch processing of data. It is not meant for real-time querying but excels in analyzing historical data stored in HDFS.
4. **SQL-Like Interface for Hadoop**:
    - Hive makes Hadoop accessible to analysts and data scientists who are familiar with SQL, offering a more user-friendly way to interact with large datasets without needing to learn MapReduce programming.
5. **Integrating with BI Tools**:
    - Hive can be integrated with business intelligence (BI) tools that use JDBC or ODBC for reporting and data analysis, making it a suitable choice for organizations looking to use Hadoop for business analytics.

Installation of HIVE:

- Downloading hive files

- Locating hive folder after extraction

- Setting path for hive in Hadoop



- Creating hive and warehouse folder in hdfs

- Updating warehouse part with desired values