Overview and Introduction to Hive

William Evans, Medhini Bhat, Harshini S, Janaaki, Bhava Nidhi, Lokhesh Kumar

Introduction

Apache Hive

Apache Hive is a data warehouse and ETL tool built on Hadoop, offering an SQL-like interface for querying and analyzing large datasets in HDFS. It translates HiveQL queries into MapReduce jobs, making big data processing easier. Designed for data warehousing and analytics, Hive is optimized for batch processing but not for OLTP.

Components of Hive:

HCatalog

It is a Hive component and is a table as well as a store management layer for Hadoop. It enables user along with various data processing tools like Pig and MapReduce which enables to read and write on the grid easily.

WebHCat

It provides a service which can be utilized by the user to run Hadoop MapReduce (or YARN), Pig, Hive tasks or function Hive metadata operations with an HTTP interface.

Modes of Hive:

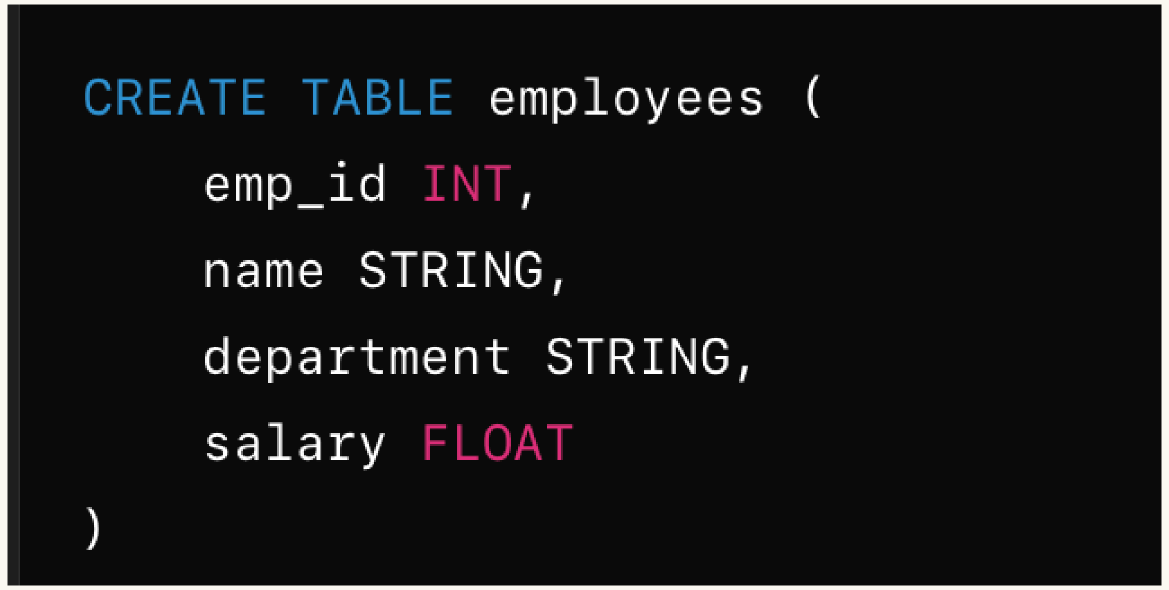
Local Mode

It is used, when the Hadoop is built under pseudo mode which has only one data node, when the data size is smaller in term of restricted to single local machine, and when processing will be faster on smaller datasets existing in the local machine.

Map Reduce Mode

It is used, when Hadoop is built with multiple data nodes and data is divided across various nodes, it will function on huge datasets and query is executed parallelly, and to achieve enhanced performance in processing large datasets.

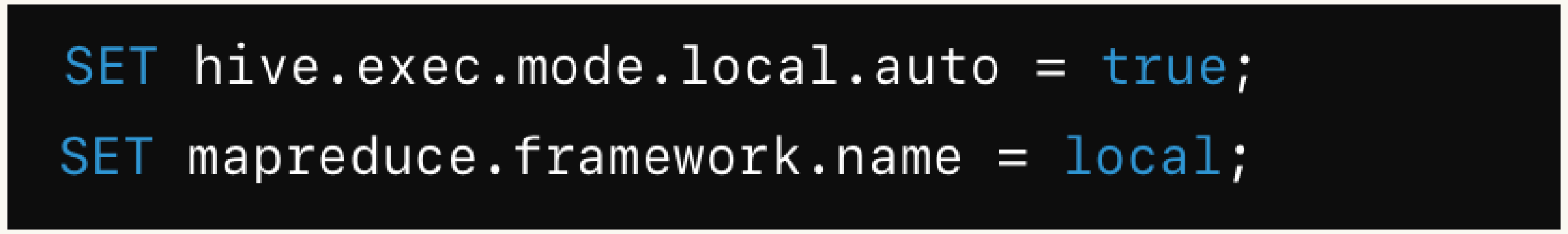
Example



```
CREATE TABLE employees (
    emp_id INT,
    name STRING,
    department STRING,
    salary FLOAT
)
```

Structure of the DB

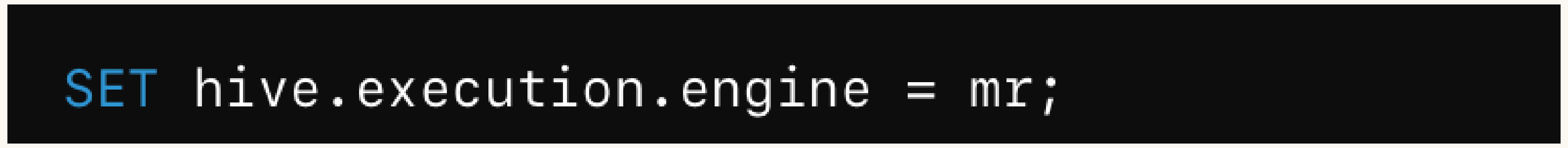
Local Mode



```
SET hive.exec.mode.local.auto = true;
SET mapreduce.framework.name = local;
```

Only for small datasets (a few hundred records)

Map Reduce Mode



```
SET hive.execution.engine = mr;
```

For large datasets (millions of records)

```
SELECT department, AVG(salary) AS avg_salary
FROM employees GROUP BY department;
```

Query

Characteristics of Hive:

1. Requires databases and tables before loading data.
2. Manages and queries only structured data in tables.
3. Optimized for structured data with UDFs, unlike raw MapReduce.
4. Uses partitioning for better query performance.
5. Supports multiple file formats (TEXTFILE, SEQUENCEFILE, ORC, RCFILE).
6. Uses Derby for single-user metadata and MySQL for multi-user metadata.

Features of Hive:

1. Supports indexing (bitmap, compaction) for faster queries.
2. Stores metadata in RDBMS for efficient query execution.
3. Provides built-in UDFs for data manipulation and extensibility.
4. Supports compression algorithms like DEFLATE, BWT, Snappy.
5. Stores schemas in a database and processes data in HDFS.
6. Designed for OLAP workloads.
7. Uses Hive Query Language (HiveQL) for querying.

Advantages

Scalability

Apache Hive is designed to handle large volumes of data, making it a scalable solution for big data processing.

Familiar SQL-like interface

Hive uses a SQL-like language called HiveQL, which makes it easy for SQL users to learn and use.

Integration with Hadoop ecosystem

Hive integrates well with the Hadoop ecosystem, enabling users to process data using other Hadoop tools like Pig, MapReduce, and Spark.

Supports partitioning and bucketing

Hive supports partitioning and bucketing, which can improve query performance by limiting the amount of data scanned.

User-defined functions

Hive allows users to define their own functions, which can be used in HiveQL queries.

Disadvantages

Limited real-time processing

Hive is designed for batch processing, which means it may not be the best tool for real-time data processing.

Slow performance

Hive can be slower than traditional relational databases because it is built on top of Hadoop, which is optimized for batch processing rather than interactive querying.

Steep learning curve

While Hive uses a SQL-like language, it still requires users to have knowledge of Hadoop and distributed computing, which can make it difficult for beginners to use.

Limited flexibility

Hive is not as flexible as other data warehousing tools because it is designed to work specifically with Hadoop, which can limit its usability in other environments.

Integration with Hadoop ecosystem

Thank You