**Hive MapReduce - Lab Manual**

**Aim:**

To understand and implement MapReduce execution in Apache Hive, optimizing large-scale data processing through distributed computing.

**Understanding Hive with MapReduce:**

Hive converts SQL-like queries into MapReduce jobs, enabling efficient processing of large datasets in a distributed environment.

**1. Enabling MapReduce Execution in Hive**

**Definition:** Ensuring that Hive queries run as MapReduce jobs.

**Query:**

SET hive.execution.engine=mr;

**2. Creating a Table**

**Definition:** Creating a Hive table to store employee data.

**Query:**

CREATE TABLE employees (

employee\_id INT,

name STRING,

department STRING,

salary FLOAT,

year INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

**3. Loading Data into the Table**

**Definition:** Inserting data from an HDFS file into the Hive table.

**Query:**

LOAD DATA INPATH '/hdfs\_path/employees.csv' INTO TABLE employees;

**4. Running a Query that Triggers MapReduce**

**Definition:** Performing aggregations that require distributed computing using MapReduce.

**Query:**

SELECT department, AVG(salary)

FROM employees

GROUP BY department;

**Output:**

| **department** | **avg\_salary** |
| --- | --- |
| HR | 60000 |
| IT | 75000 |

.

**5. Viewing MapReduce Job Execution**

**Definition:** Checking the progress and execution details of a MapReduce job.

**Command:**

yarn application -list

**Output:**

Application ID Application Type User State FinalStatus Progress

application\_123456 MAPREDUCE hive RUNNING UNDEFINED 60%

**Result**

Hive with MapReduce enables distributed processing of large datasets, improving efficiency and scalability for complex queries. By leveraging Hadoop’s parallel computation, Hive ensures optimal data analysis for big data environments.