**Weeks 2, 4, 6, 8, 10, and 12**

**✅ Week 2:**

**Hadoop Installation & Monitoring**

1. Perform setting up and installing Hadoop in its three operating modes:
   * Standalone
   * Pseudo-distributed
   * Fully-distributed
2. Use web-based tools to monitor your Hadoop setup.

**✅ Week 4:**

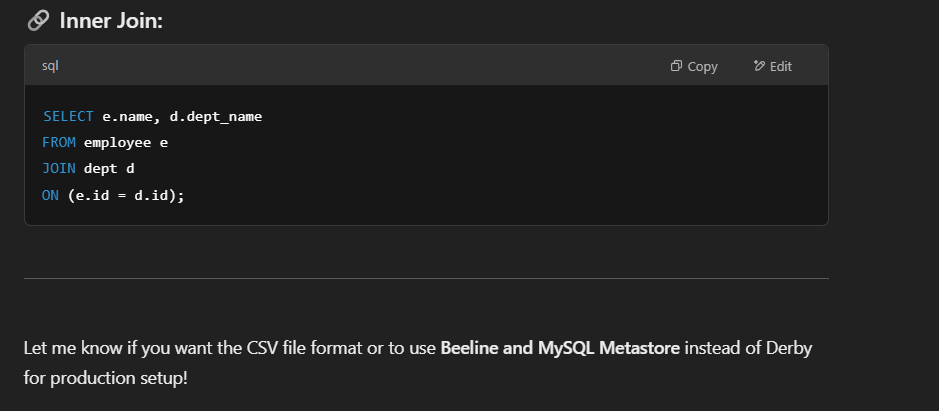
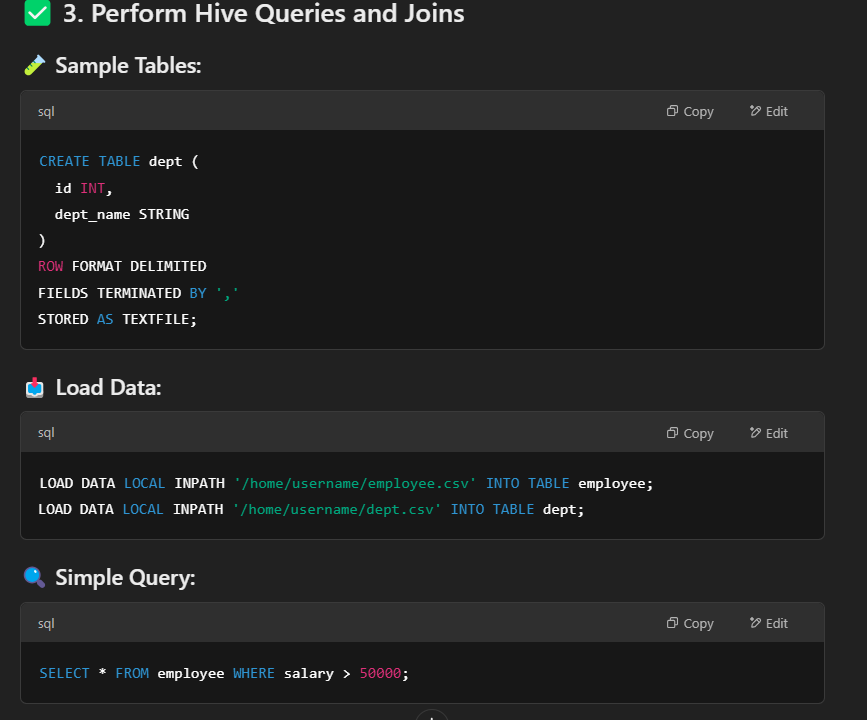
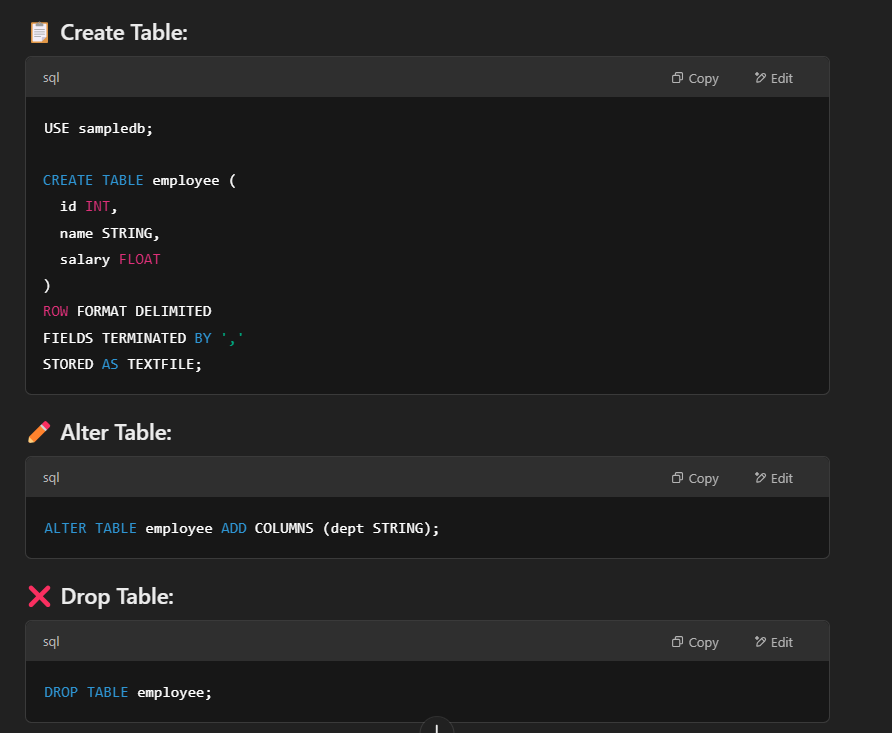
**Word Count using MapReduce**

* Run a basic **Word Count** MapReduce program to understand the **MapReduce paradigm**.

**✅ Week 6:**

**Hive Operations**

1. Install the Hive framework.
2. Create, alter, and drop **databases** and **tables** in Hive.
3. Perform **Hive queries and joins** to display and retrieve data.



**✅ Week 8:**

**Sales Data Analysis using Spark**

1. Analyze sales data stored in Hadoop Data Lake.
2. Create a DataFrame and write Spark SQL queries to compute:
   * Average sale of every customer.
3. Store the output as a **Parquet** file.

**✅ Week 10:**

**Spark RDD Operations**

1. Write a Spark program to **count the number of occurrences** of each character in a text file.

from pyspark import SparkContext

sc = SparkContext("local", "Character Count")

text\_file = sc.textFile("path\_to\_your\_file.txt")

char\_counts = (

text\_file.flatMap(lambda line: list(line)) # Break each line into individual characters

.map(lambda char: (char, 1)) # Assign each character a count of 1

.reduceByKey(lambda a, b: a + b) # Aggregate counts for each character

)

# Display results

for char, count in char\_counts.collect():

print(f"Character '{char}' occurs {count} times.")

output:

Character 'a' occurs 15 times.

Character 'b' occurs 8 times.

Character 'c' occurs 20 times.

Character ' ' occurs 50 times. # Space character count

Character '.' occurs 5 times.

**✅ Week 12:**

**Spark Joins**

* Join **Customer** and **Sales** datasets using a common key product\_id.
* Print:
  + customer\_id
  + customer\_name
  + product\_name
  + price

from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("RetailDataJoin").getOrCreate()

customer\_data = [(101, "Ravi", 1),

(102, "Keerth", 2),

(101, "Syam", 1),

(101, "Geetha", 1),

(103, "Dawn", 3)]

customer\_schema = ["customer\_id", "customer\_name", "product\_id"]

customer\_df = spark.createDataFrame(customer\_data, customer\_schema)

sales\_data = [(1, "Laptop", 50000),

(2, "Mobile", 20000),

(3, "Tablet", 15000)]

sales\_schema = ["product\_id", "product\_name", "price"]

sales\_df = spark.createDataFrame(sales\_data, sales\_schema)

joined\_df = customer\_df.join(sales\_df, "product\_id")

result\_df = joined\_df.select("customer\_id", "customer\_name", "product\_name", "price")

result\_df.show()

spark.stop()