Problem Statement-1

a. Write HDFS shell commands for the following

2. To Print Version of installed Hadoop
3. For listing the files and directories present in HDFS ‘path’ directory present under root (/path).
4. To Delete an empty directory named as **XYZ.**
5. To fetch the usage instructions of **mkdir** command
6. To Copy ‘file1.txt’ from ‘InputDir’ to ‘OutputDir’ as file2.txt
7. Write command for copy the ‘testfile’ of the hadoop filesystem (present under root) to the local file system (pwd)
8. Write command for display the content of the ‘sample’ file present in newDataFlair directory of HDFS (under root).
9. Write command for copy local file named **file1** under the present working directory(pwd) of local file system to the Hadoop filesystem under root.

Problem Statement-2

Part 1

Write hive query for below:

1. Write hive query to create databases name: **anotherDB**
2. Write hive query to CREATE EXTERNAL TABLE in anotherDB name orders1 with order_id, order_date, order_customer_id , order_status.
3. Write a hive query to load data in order1 table using a file which is available in the local file system.

Part 2

Using the CUSTOMERS and ORDERS provided dataset as a hive table write code for performing below joins

1. **Inner join**
2. **Left outer join**
3. **Right Outer Join**
4. **Full Outer Join**

Part 3 :

A web log dataset extract is provided.
Dataset has 4 columns, they are –

1. IP
2. Time,
3. URL,
4. Response Status.

Write Hive-QL to perform & answer below tasks.

1. Create a Hive table to represent the web server log data by defining the table schema with appropriate column names and data types for IP, Time, URL, and Status.

2. Load the web server log data into the Hive table using the LOAD DATA statement or by creating an external table pointing to the log file location.

3. Retrieve the count of log entries in the table.

4. Fetch the top N URLs based on the number of hits.

5. Count the number of log entries per IP address.

6. Fetch top 5 IP addresses with the highest average number of requests per hour.

7. Fetch top 10 most visited URLs along with their visit counts.

8. Fetch the average response time for each hour of the day.

9. Which IP addresses have made more than 100 requests in total?

10. How many unique URLs were visited by each IP address?

Instructions : Please upload your solutions along with the output screenshots in one pdf.