**Big Data and Hadoop Developer Project - Social Media**

**Domain: Media**

**Create table and Load data in HIVE:**

SHOW DATABASES;

CREATE DATABASE venkat\_social\_media;

USE venkat\_social\_media;

SHOW TABLES;

CREATE EXTERNAL TABLE social\_media\_table(ques\_id BIGINT, user\_id BIGINT, score\_of\_ques BIGINT, time\_of\_ques BIGINT, tags STRING, num\_of\_views\_of\_ques BIGINT, num\_of\_ans\_ques BIGINT, unique\_ans\_id BIGINT, user\_id\_of\_ans BIGINT, score\_of\_ans BIGINT, time\_of\_ans BIGINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\_';

LOAD DATA INPATH '/user/venkatasatya9\_live/social\_media\_project/social\_media\_data.txt' OVERWRITE INTO TABLE social\_media\_table;

**Question 1) Top 10 most commonly used tags in this data set.**

**Solution:** Tool Used **PIG**.

Create tags table and load all tags data from table created above. Calculate the count of each tag by grouping the tags. Order the tags in descending order using the count and take only the top 10 rows data.

**Code:**

CREATE TABLE question\_tags(tags STRING);

INSERT INTO question\_tags\_table SELECT regexp\_replace(tags, ‘,’, “\n”) FROM social\_media\_table;

SELECT tags, count(\*) AS tag\_count FROM question\_tags\_table GROUP BY tags ORDER BY tag\_count DESC LIMIT 10;

**Question 2) Average time to answer questions.**

**Solution:** Tool Used **HIVE**.

Average time to answer the question is calculated by using adding the time to answer the question of same group by question id and dividing the total value by total number of times this question is answered.

**Code:**

SELECT ques\_id, AVG(time\_of\_ans) AS avg\_time\_of\_ans FROM social\_media\_table GROUP BY ques\_id;

**Example:**

563355\_62701\_0\_1235000081\_php,error,gd,image-processing\_220\_2\_563372\_67183\_2\_1235000501

563355\_62701\_0\_1235000081\_php,error,gd,image-processing\_220\_2\_563374\_66554\_0\_1235000551

Total of time to answer the question = 1235000501 + 1235000551 = 2470001052

Average time to answer the question = 2470001052 / 2 = 1235000526

**Question 3) Number of questions which got answered within 1 hour.**

**Solution:** Tool Used **HIVE**.

If "Number of questions which got answered within 1 hour" is about count of questions:

Get the count by of all questions answered with in 1 hour by using Time of the answer minus Time of the question is less than or equal to 3600 (1 hour in epoch time).

**Code:**

SELECT COUNT(ques\_id) FROM social\_media\_table WHERE (time\_of\_ans - time\_of\_ques) <= 3600;

If "Number of questions which got answered within 1 hour" is about displaying each question answered within 1 hour:

Show all the questions answered with in 1 hour by using Time of the answer minus Time of the question is less than or equal to 3600 (1 hour in epoch time).

**Code:**

SELECT ques\_id, (time\_of\_ans - time\_of\_ques) AS time\_diff FROM social\_media\_table WHERE (time\_of\_ans - time\_of\_ques) <= 3600;

**Question 4) Tags of questions which got answered within 1 hour.**

**Solution:** Tool Used **HIVE**.

Show all the questions and its tags answered with in 1 hour by using Time of the answer minus Time of the question is less than or equal to 3600 (1 hour in epoch time).

**Code:**

SELECT ques\_id, tags FROM social\_media\_table WHERE (time\_of\_ans - time\_of\_ques) <= 3600;