Big Data - Case Study

Subject - Big Data Analytics and Architecture

PROJECT

Mental Health and Social Media Balance Analysis

Mental Health and Social Media Balance Analysis Using Apache Hive

Project Overview

This project focuses on performing data analysis and extracting meaningful insights from a *Mental Health and Social Media Balance* dataset using **Apache Hive**.
The primary objective is to utilize Hive' s SQL-like capabilities to explore the relationship between **screen time**, **social media usage**, and **mental well-being**.

The analysis demonstrates how structured mental health and digital behavior data can be managed efficiently on a **Big Data platform (Hadoop/Cloudera)** and queried using **HiveQL** for analytical reporting and decision-making support.

Dataset Description

The dataset, `Mental_Health_and_Social_Media_Balance_Dataset.csv`, contains detailed information about users' social media habits and mental health indicators.

Attributes include:

User_ID — unique identifier for each participant.
Age — age of the participant.
Gender — demographic category (Male/Female/Other).
Platform — primary social media platform used.
Screen_Time — average time (in hours) spent on social media per day.
Mental_Health_Score — self-reported mental health rating on a standardized scale.

Objectives

The key objectives of this project are:

To **import and store CSV data** into Hive tables efficiently.

To perform **analytical queries** on mental health and social media usage patterns.
To extract **data-driven insights** such as:

- Average screen time by gender or platform.
- Correlation between screen time and mental health.
- Platform-wise mental health score comparisons.
- Age group– based differences in social media habits.
- Identifying trends in digital wellness behavior.

Technologies Used

Apache Hive
Hadoop Distributed File System (HDFS)
Cloudera QuickStart Environment
HiveQL (SQL-like query language)
CSV file ingestion via HDFS

Steps Performed

1. **Database and Table Creation**

   Created a new database `social_media_analysis` in Hive.

   Defined an **external table** schema corresponding to the CSV structure:

```
CREATE EXTERNAL TABLE IF NOT EXISTS mental_health_balance (
  User_ID STRING,
  Age INT,
  Gender STRING,
  Platform STRING,
  Screen_Time DOUBLE,
  Mental_Health_Score DOUBLE
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar" = "\""
)
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/mental_health_data/';
```

2. **Data Loading**

   Uploaded the CSV file from local storage (`/home/cloudera/Desktop/mental.csv`) to HDFS.

   ```
   hdfs dfs -put /home/cloudera/Desktop/mental.csv
   /user/hive/warehouse/mental_health_data/
   ```

3. **Data Analysis Using HiveQL**

   Executed multiple analytical queries to explore mental health and social media balance:

   `SELECT COUNT(*)` → total number of records.

   `GROUP BY` → gender and platform analysis.

   `AVG()` → average screen time and mental health scores.

   `ORDER BY` + `LIMIT` → identify top platforms or high-risk groups.

4. **Insight Extraction and Reporting**
   - Generated summarized analytical results for mental wellness trends.

Key Insights

1. **Average Screen Time** — The mean daily screen time among users indicates moderate-to-high engagement across social media platforms.
2. **Gender-wise Screen Time** — Males and females show comparable screen usage, with slight variations in platform preference.
3. **Platform Popularity** — Instagram and YouTube emerged as the most-used platforms.
4. **Mental Health Patterns** — Users with higher screen time often recorded lower mental health scores.
5. **Age Group Analysis** — Teenagers and young adults (13– 30 years) had the highest screen times.
6. **Platform Impact** — Professional platforms (e.g., LinkedIn) users reported better mental health averages.

7. **Correlation Trend** — A negative correlation was observed between `Screen_Time` and `Mental_Health_Score`.
8. **High-risk Segment** — Users exceeding 5+ hours of daily screen time tend to have below-average well-being.
9. **Gender-based Mental Health** — Average mental health scores were slightly higher for females in the dataset.
10. **Overall Digital Well-being** — Data indicates that balanced screen habits are associated with improved mental health ratings.

Conclusion

This project demonstrates how **Apache Hive** can be leveraged for large-scale analysis of behavioral and mental health data.
By combining **Big Data storage (HDFS)** and **HiveQL querying**, the study reveals actionable insights into how digital habits affect psychological well-being.

The approach showcases Hive' s effectiveness in:

Managing structured datasets.
Performing SQL-like analytics on large-scale data.
Supporting **data-driven decisions** in mental health awareness and digital wellness research.

Use Database

```
hive> USE social_media_analysis;
OK
Time taken: 0.066 seconds
hive>
```

Create table

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS mental_health_balance (
    >   User_ID STRING,
    >   Age INT,
    >   Gender STRING,
    >   Platform STRING,
    >   Screen_Time DOUBLE,
    >   Mental_Health_Score DOUBLE
    > )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    >   "separatorChar" = ",",
    >   "quoteChar" = "\""
    > )
    > STORED AS TEXTFILE
    > LOCATION '/user/hive/warehouse/mental_health_data/';
OK
Time taken: 1.186 seconds
hive>
```

1. Average Screen Time of Users

```
Time taken: 1.201 seconds, Fetched: 10 row(s)
hive> SELECT AVG(Screen_Time) AS avg_screen_time
    > FROM mental_health_balance;
```

Output

```
Total MapReduce CPU Time Spent: 8 seconds 660 msec
OK
6.304
Time taken: 85.479 seconds, Fetched: 1 row(s)
hive>
```

2. Average mental health score

```
hive> SELECT AVG(Mental_Health_Score) AS avg_mental_health_score
    > FROM mental_health_balance;
Query ID = cloudera_20251029083636_a0742dd7-d860-48e7-957c-de16dfb3265c
Total jobs = 1
Launching Job 1 out of 1
```

Output

```
Total MapReduce CPU Time Spent: 7 seconds 410 msec
OK
6.618
Time taken: 56.396 seconds, Fetched: 1 row(s)
hive>
```

3. Gender-wise average screen time

```
hive> SELECT Gender, AVG(Screen_Time) AS avg_screen_time
    > FROM mental_health_balance
    > GROUP BY Gender;
Query ID = cloudera_20251029083939_1ae8080c-b120-49d6-ba7f-9f5cf677da6f
Total jobs = 1
```

Output

```
Total MapReduce CPU Time Spent: 6 seconds 220 msec
OK
Female  6.279475982532751
Gender  NULL
Male    6.294354838709677
Other   6.6521739130434785
Time taken: 44.006 seconds, Fetched: 4 row(s)
hive>
```

4. Gender-wise average mental health score

```
hive> SELECT Gender, AVG(Mental_Health_Score) AS avg_mental_health_score
    > FROM mental_health_balance
    > GROUP BY Gender;
Query ID = cloudera_20251029084343_73623982-e469-4000-abf6-118237317add
Total jobs = 1
```

Output

```
Total MapReduce CPU Time Spent: 5 seconds 610 msec
OK
Female  6.602620087336245
Gender  NULL
Male    6.633064516129032
Other   6.608695652173913
Time taken: 47.52 seconds, Fetched: 4 row(s)
hive>
```

5. Most used social media platform

```
hive> SELECT Platform, COUNT(*) AS user_count
    > FROM mental_health_balance
    > GROUP BY Platform
    > ORDER BY user_count DESC
    > LIMIT 1;
Query ID = cloudera_20251029084747_a191f592-a571-45b4-84b9-d5988936801a
```

Output

```
Total MapReduce CPU Time Spent: 17 seconds 430 msec
OK
6.2     19
Time taken: 176.617 seconds, Fetched: 1 row(s)
hive>
```

6. Platform-wise average screen time

```
hive> SELECT Platform, AVG(Screen_Time) AS avg_screen_time
    > FROM mental_health_balance
    > GROUP BY Platform
    > ORDER BY avg_screen_time DESC;
Query ID = cloudera_20251029085454_8c689be0-e5d2-4984-bcbc-7dae2adbc601
Total jobs = 2
```

Output

```
Total MapReduce CPU Time Spent: 16 seconds 340 msec
OK
2.2     9.5
1.5     9.0
1.7     9.0
1.8     9.0
2.1     9.0
2.3     9.0
2.6     9.0
2.5     8.833333333333334
2.0     8.5
3.3     8.285714285714286
3.2     8.2
3.5     8.2
1.0     8.0
2.9     8.0
3.4     8.0
3.8     8.0
4.2     7.7272727272727275
2.7     7.666666666666667
3.0     7.666666666666667
3.7     7.666666666666667
3.6     7.4
```

7. Correlation between screen time and mental health

```
hive> SELECT
    >   CASE
    >     WHEN Screen_Time > 4 AND Mental_Health_Score < 5 THEN 'High Screen Tim
e, Low Score'
    >     WHEN Screen_Time < 2 AND Mental_Health_Score > 7 THEN 'Low Screen Time
, High Score'
    >     ELSE 'Other'
    >   END AS Category,
    >   COUNT(*) AS count
    > FROM mental_health_balance
    > GROUP BY
    >   CASE
    >     WHEN Screen_Time > 4 AND Mental_Health_Score < 5 THEN 'High Screen Tim
e, Low Score'
    >     WHEN Screen_Time < 2 AND Mental_Health_Score > 7 THEN 'Low Screen Time
, High Score'
    >     ELSE 'Other'
    >   END;
Query ID = cloudera_20251029085252_1747d10f-7115-479a-a2ab-449fce577564
Total jobs = 1
```

Output

```
OK
High Screen Time, Low Score     43
Other   458
Time taken: 79.608 seconds, Fetched: 2 row(s)
hive>
```

8. Average screen time by age group

```
hive> SELECT
    >   CASE
    >     WHEN Age BETWEEN 13 AND 18 THEN 'Teen'
    >     WHEN Age BETWEEN 19 AND 30 THEN 'Young Adult'
    >     WHEN Age BETWEEN 31 AND 45 THEN 'Adult'
    >     ELSE 'Older Adult'
    >   END AS Age_Group,
    >   AVG(Screen_Time) AS avg_screen_time
    > FROM mental_health_balance
    > GROUP BY
    >   CASE
    >     WHEN Age BETWEEN 13 AND 18 THEN 'Teen'
    >     WHEN Age BETWEEN 19 AND 30 THEN 'Young Adult'
    >     WHEN Age BETWEEN 31 AND 45 THEN 'Adult'
    >     ELSE 'Older Adult'
    >   END;
Query ID = cloudera_20251029090000_34cf09d4-2f92-45d5-bc6a-3504a47f1e1a
```

Output

```
Total MapReduce CPU Time Spent: 10 seconds 700 msec
OK
Adult   6.339055793991417
Older Adult     5.967741935483871
Teen    6.68
Young Adult     6.264516129032258
Time taken: 90.383 seconds, Fetched: 4 row(s)
```

9. Average mental health score by platform

```
hive> SELECT Platform, AVG(Mental_Health_Score) AS avg_mental_health
    > FROM mental_health_balance
    > GROUP BY Platform
    > ORDER BY avg_mental_health ASC;
Query ID = cloudera_20251029090303_4fd4dbbc-8ba8-4c8e-967d-a7b4ec628c18
Total jobs = 2
```

Output

```
Total MapReduce CPU Time Spent: 15 seconds 570 msec
OK
Daily_Screen_Time(hrs)  NULL
2.3     3.0
1.5     3.5
1.7     3.5
2.5     3.8333333333333335
2.0     4.0
2.6     4.0
2.9     4.0
3.0     4.0
2.1     4.333333333333333
2.2     4.5
4.0     4.714285714285714
3.5     4.8
2.7     4.833333333333333
3.3     4.857142857142857
1.0     5.0
3.6     5.2
3.9     5.25
3.8     5.333333333333333
```

10. Identify users with extreme screen time (top 5)

```
hive> SELECT User_ID, Screen_Time, Mental_Health_Score
    > FROM mental_health_balance
    > ORDER BY Screen_Time DESC
    > LIMIT 5;
Query ID = cloudera_20251029090808_423b9db1-5a7a-490e-9c46-cc7640439080
```

Output

```
Total MapReduce CPU Time Spent: 7 seconds 210 msec
OK
User_ID Sleep_Quality(1-10)     Stress_Level(1-10)
U116    9.0     5.0
U165    9.0     5.0
U061    9.0     5.0
U217    9.0     5.0
Time taken: 79.122 seconds, Fetched: 5 row(s)
hive>
```