Table of Contents

[Facebook analysis in Hadoop 1](#_Toc140496342)

[The ETL process using Nifi: 2](#_Toc140496343)

[StandardSSLContextService  configuration: 2](#_Toc140496344)

[Apache Hive: 2](#_Toc140496345)

[Hive and python: 3](#_Toc140496346)

[Python: 3](#_Toc140496347)

[Analysis in Hive: 4](#_Toc140496348)

[Questions and answers: 4](#_Toc140496349)

# Facebook analysis in Hadoop

This project aims to gain meaningful insights from Facebook data by leveraging the capabilities of the Hadoop ecosystem. This project involves collecting a representative dataset of Facebook data, including user profiles, posts, comments, likes, shares, and friend connections. The collected data is then ingested into Hadoop using Apache Nifi and stored in Hadoop Distributed File System (HDFS) for efficient and scalable storage. Hadoop ecosystem (Apache, 2022)

Next, the data is prepared through pre-processing steps such as cleaning and transforming it into a structured format using Apache Hive suitable for analysis. The transformed data is then subjected to exploration and analysis using tools like Apache Hive. User behaviour analysis is performed to uncover insights into engagement patterns and other relevant aspects. This analysis helps in understanding how users interact with Facebook and provides valuable information for improving user experience and engagement strategies.

# The ETL process using Nifi:

The data was extracted from source using two nifi processors (InvokeHTTP and PutHDFS) and ingested into HDFS (Apache, 2023)

1. InvokeHTTP
2. StandardSSLContextService  configuration: The StandardSSLContextService can be configured with a truststore, which is a Java KeyStore object containing a collection of TrustedCertEntry objects, to extract data from any secure website. Each of these objects has the public key and certificate information of a trusted entity. Apache NiFi assesses the received certificate identifying the service and makes an attempt to validate that certificate when it tries to connect to another endpoint or service over HTTPS. It checks to determine which certificate signed the leaf cert and validates that one if the endpoint certificate is not immediately contained in the truststore. This method keeps going up the certificate chain until either a trusted certificate is identified or there are none. It is located in the $JAVA HOME/jre/lib/security/cacerts directory. Depending on how java8 was installed.

You can determine the value of $JAVA\_HOME for your OS and Java version:

## StandardSSLContextService  configuration:

* **Truststore filename:** /Library/Java/JavaVirtualMachines/jdk1.8.0\_101.jdk/Contents/Home/jre/lib/security/cacerts (example)
* **Truststore password:** changeit (default value)
* **Truststore type:** JKS

# Apache Hive:

structure was then projected onto the stored data using Apache Hive (Apache Hive, 2015). Create table:

CREATE TABLE facebook\_tab (

id INT,

age int,

day int,

year int,

month int,

gender string,

tenure int,

friends int,

friends\_init int,

likes int,

likes\_recd int,

mlikes int,

mlikes\_recd int,

wlikes int,

wlikes\_recd int)

STORED AS TEXTFILE

TBLPROPERTIES("skip.header.line.count"="1")

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"separatorChar" = ",",

"quoteChar" = "\""

);

load the data into hive table:

Load data INPATH '/user/hive/project/pseudo\_facebook.csv' INTO TABLE facebook\_tab;

## Hive and python:

Connected hive to python using hiveserver2 and pyhive library in python to carry out data pre-processing.

## Python:

Carried out data pre-processing. The data contained 175 null values on the gender column and 2 null on the tenure column out of 990003 observations, other columns contain no null values. Using linear interpolation and forward fill, I cleaned the data and updated the hive table with the clean data.

# Analysis in Hive:

## Questions and answers:

1. Find out how many users there are in the dataset overall.

Command:

Select count(\*) from facebook\_tab;

A screen shot of a black screen

Description automatically generated

1. the total number of Facebook users who are over 18

command:

Select count(\*) from facebook\_tab where age>18;

A black screen with white text

Description automatically generated

1. Do male users have more friends on average than female users?

Commands:

Select avg(friends) from facebook\_tab where gender ="male";

Select avg(friends) from facebook\_tab where gender ="female";

All in one line :

Select gender, avg(friends) from facebook\_tab group by gender;

A screen shot of a computer

Description automatically generated

1. How many more likes do young individuals get on Facebook than do older users?

Command

select avg(likes\_recd) from facebook\_tab where age>=12 AND age>=30;

A black screen with white text

Description automatically generated

A black screen with white text

Description automatically generated

1. Do youthful members use their mobile phones or laptops to browse Facebook?

A black screen with white text

Description automatically generated

1. Find out the number of Facebook users for each birthday month.

A screen shot of a computer

Description automatically generated

References

Apache. (2022). Hadoop [computer software]

Apache. (2023). Nifi [computer software]

Apache Hive. (2015). Hive [computer software]

stylefix