UNIVERSITY OF MALAYA

Faculty of Computer Science & Information Technology

Mastre Of Data Science

Semester II Session 2021/2022

WDQ7007 Big Data Management
Group Project

Title: The Big Data ETL Process Of Social Media MOMO's Behavior Via Hadoop

Lecturer: Dr. ALI FEIZOLLAH

Group Members

No.	NAME	MATRIX NO.
1	DI BEI	S2013283
2	ZITENG ZHANG	S2149768
3	ZUOGE CHEN	S2125783
4	YINGGANG JIE	S2113868
5	JINXING LI	S2128176
6	HO WEI YAN	S2116489
7	JUNBO SHI	S2140997
8	HAN LI HUI	S2025432

Introduction

Social media applications have become essential to most individuals nowadays. Momo is a popular social media app in China. The distribution of the customer is necessary for the analysis to attract more people to use it. But for the thousands of data, how to perform it is a big problem for us. So in the research we will use hadoop technology to do the ETL and analyze the data for insight. For the tools, we will use MySQL, Sqoop, Hive and Hbase. Hadoop is an efficient platform to manage big data.

For the data, we search from a website. This dataset contains a total of 140,000 samples of user communication data with 20 columns that were collected from the social media app of Momoin in one day.

The dataset obatained from:
https://blog.csdn.net/clark123432343/article/details/124409046

Research Objectives

1. To set up the data warehouse and build a data pipeline to manage the social media data momo.
2. To do data cleaning and keep the data without noisy data and no value after EDA in hive.
3. To leverage a data access solution such as HBase so that we can better access and manage the data of social media data.
4. To do data visualization according to the behaviors of users in social media momo.

Problem statement

With the rapid growth of social media in recent years, Momo is one of the most popular apps among Chinese teenagers. Different from WeChat, Whatsapp, Facebook, qq, etc. Momo adds friends or joins interest groups through people's real time positioning. Customer losing is one of the important problems that Momo needs to consider for improving its ROI. Huge amount of user data generated every day. Integrating the user communication data and building a unified data warehouse can help it to classify, track, and monitor the business data to achieve the objectives of increasing ROI. Therefore, building the secure and reliable database or data warehouse and doing exploratory data analysis of the data records can help Momo to improve the performance of its products and avoid user loss.(Xu & Wu, 2019)(Chan, 2019)(Li, 2020)

Literature review

Stefan et al. proposed the use of social media app Momo to realize the process of user data collection. Momo will use centralized and distributed filtering methods to process and store data. Momo's data collection group mainly includes different user groups such as men and women. The system will process user data for visualization. Provide users with user behavior predictions and friend recommendations. In this paper, they also focus on the use of visual machine learning concepts on mobile and server. (Edlich & Vogler, 2013)

Xu & Wu examine a phenomenon that has become popular in mainland China in recent years, known as "stranger communication", focusing on an application known as Momo, a social discovery and dating platform widely used in China. The report begins with an overview of the post-reform and opening-up Chinese context,

which gave rise to urbanization, individualization and a sexual revolution. Thereafter, it turns to theories of strangers and cosmopolitanism, which are used to analyze stranger communication. Using cultural discourse analysis as an analytical procedure, the study presents the results of online and offline interviews conducted in both Beijing and Shanghai to examine how Momo users in urban metropolitan areas use the application in order to analyze the cultural radiation in their communication practices. Based on the data obtained, this paper analyzes MOMO Ren communication in the Chinese context. The paper analyzes MOMO Ren communication in the Chinese context and describes how Momo is designed to facilitate this communication. (Xu & Wu, 2019)

Lik et al. examined how heterosexual male users' endorsement of masculinity was related to the number of casual sex partners they encountered on Momo, based on a social constructivist view of gender. The study examined how heterosexual male users' endorsement of masculinity was related to the number of casual sex partners they encountered on Momo, based on a social constructivist perspective. The study also explored the mediating role of sexual motivation in using Momo. Analysis of survey data from 125 heterosexual male Momo users revealed an indirect positive relationship between endorsement of masculinity and the number of sexual partners, while at the same time, it was directly but negatively associated with the number of sexual partners. These contradictory associations These contradictory associations are explained by different models of the various dimensions of masculine ideology. Since unsafe sex has been found to be associated with the use of geosocial networking applications, their study also calls for the integration of the concept of implementing safer sex with the cultural ideals of masculinity. (Chan, 2019)

Li et al. identified three types of institutional privacy concerns, including information leakage, information tracking, and surveillance by apps and the Chinese government. Most participants took little action to address their privacy concerns, instead exhibiting an attitude of ignoring and accepting the privacy risks of GSNAs. In contrast, few participants developed proactive strategies In contrast, few participants developed proactive strategies to mitigate their concerns, such as abstaining or being cautious when using apps and features, and reporting privacy threats to app platforms to app platforms. Importantly, this study provides new insights into user privacy issues that are particularly relevant to the social and cultural context in China. This study provides new insights into users' privacy issues in the Chinese social and cultural context. However, since this study mainly focuses on users' privacy issues and strategies, they encourage future research to delve into solutions and strategies for mitigating users' privacy issues on social platforms in a global context.(Li, 2020)

Methodology

HDFS

I. Setting the root password in through my sql.

```
student@student-VirtualBox:~$ sudo cat /etc/mysql/debian.cnf
# Automatically generated for Debian scripts. DO NOT TOUCH!
[client]
host     = localhost
user     = debian-sys-maint
password = CIF3y6vbH6tlQXvX
socket   = /var/run/mysqld/mysqld.sock
[mysql_upgrade]
host     = localhost
user     = debian-sys-maint
password = CIF3y6vbH6tlQXvX
socket   = /var/run/mysqld/mysqld.sock
```

```
student@student-VirtualBox:~$ mysql -u debian-sys-maint -pCIF3y6vbH6tlQXvX
mysql: [Warning] Using a password on the command line interface can be insecure
.
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 3
Server version: 5.7.38-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2022, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

II. Create a database and a specific table of data with the correct column type from the dataset and check if the table framework has been created successfully in the correct database.

```
mysql> create table tb_msg (msg_time varchar(20),
    -> sender_id varchar(15),
    -> sender_sex varchar(8),
    -> sender_ip varchar(20),
    -> sender_os varchar(15),
    -> sender_network varchar(5),
    -> sender_gps varchar(30),
    -> receiver_id varchar(15),
    -> receiver_ip varchar(20),
    -> receiver_os varchar(20),
    -> receiver_network varchar(5),
    -> receiver_gps varchar(30),
    -> receiver_sex varchar(8),
    -> msg_type varchar(5),
    -> distance varchar(10));
Query OK, 0 rows affected (0.07 sec)

mysql> show tables;
+--------------------+
| Tables_in_db2_msg |
+--------------------+
| tb_msg             |
+--------------------+
1 row in set (0.00 sec)
```

III. Load the data from local to the table created.

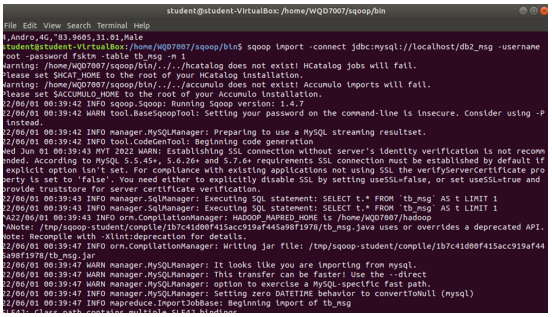
```
mysql> LOAD DATA LOCAL INFILE '~/Downloads/data2.csv' INTO TABLE tb_msg FIELDS TERMINATED BY ',';
Query OK, 68905 rows affected, 65535 warnings (1.36 sec)
Records: 68905  Deleted: 0  Skipped: 0  Warnings: 274541
```

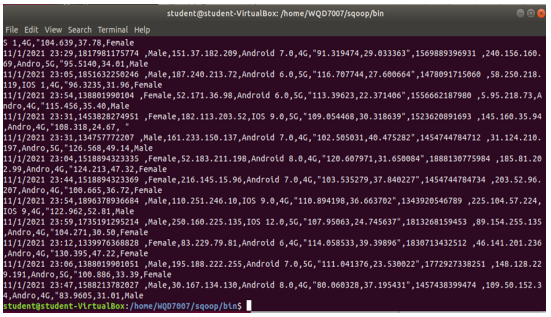
IV. Check if the data imports correctly and counts how much data has been loaded.

```
mysql> select * from tb_msg limit 5;
+-------------+-------------+------------+-------------+-------------+----------------+-------------+-------------+-------------+-------------+----------+
| msg_time    | sender_id   | sender_sex | sender_ip   | sender_os   | sender_network | sender_gps  | receiver_id | receiver_ip | receiver_os | receiver_network | receiver_gps | receiver_sex | msg_type | distance |
+-------------+-------------+------------+-------------+-------------+----------------+-------------+-------------+-------------+-------------+----------+
| 11/1/2021 17:07 | 1360740340813 | Male | 40.191.133.115 | Android 6.0 | 4G | | "82.268004 | 34.743035" | 1651865100256 | 41.252.164.18 | Andro | 5G | "92.7176 | 38.30 | Male |
| 11/1/2021 17:18 | 1370446518093 | Male | 66.93.114.125 | Android 8.0 | 5G | | "80.060328 | 37.195431" | 1346352319989 | 93.15.229.109 | Andro | 5G | "115.456 | 31.96 | Male |
| 11/1/2021 17:32 | 1473719931078 | Male | 189.227.184.152 | Android 6 | 4G | | "114.573657 | 25.549461" | 1338455304850 | 91.245.23.117 | IOS 9 | 4G | "106.699 | 33.51 | Male |
| 11/1/2021 17:25 | 1651739499485 | Male | 111.121.81.4 | Android 6.0 | 4G | | "105.595776 | 40.362789" | 1853780746279 | 89.154.255.135 | Andro | 5G | "104.650 | 40.08 | Female |
| 11/1/2021 17:37 | 1331315685893 | Female | 40.191.133.115 | Android 6.0 | 4G | | "99.340696 | 25.682909" | 1455075694893 | 8.238.89.116 | Andro | 5G | "104.197 | 33.76 | Male |
+-------------+-------------+------------+-------------+-------------+----------------+-------------+-------------+-------------+-------------+----------+
5 rows in set (0.00 sec)
```

```
mysql> Select count(*) as cnt from tb_msg;
+-------+
| cnt   |
+-------+
| 68905 |
+-------+
1 row in set (0.08 sec)
```

V. Finished by mysql, in go back to hadoop, the system will use sqoop command to transfer data from mysql to HDFS. Lastly, 'cat' will be used to check if the data sent to HDFS correctly.



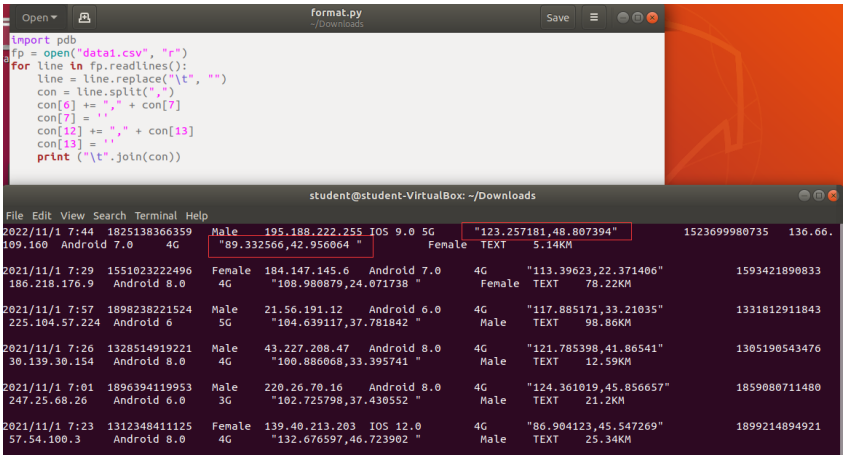


HIVE

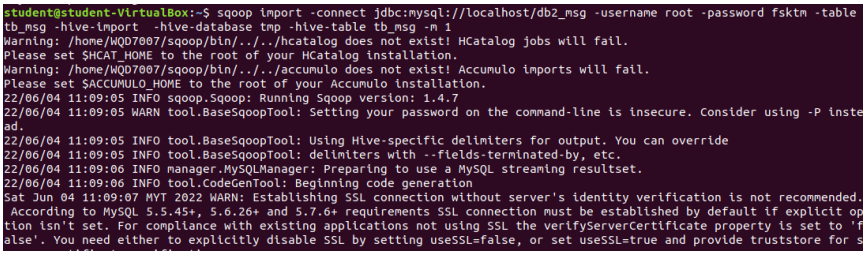
For the second part, we will introduce the hive. hive help us to word with data.

I. Data format fix

After processed,we could the data format as below :



II. Create DB and table in hive and prepare the data table in this process. Load data into hive using sqoop, we could see we successfully load data into hive



III. Data problem

A. Empty data

Problem 1: in the current data, some data fields are empty and not legal data

B. Time format

Problem 2: in the demand, the daily and hourly message volume needs to be counted, but there are no day and hour fields in the data, only the overall time field, which is difficult to handle

C. GPS location problem

Problem 3: in the demand, it is necessary to build a visual map of the region for longitude and dimension, but the GPS longitude and latitude in the data is a field, which is difficult to handle

IV. Data cleaning

Requirement 1: filter illegal data with empty fields

• where filtering

2 requirement 2: build day and hour fields through time fields

• substr function

2 demand 3: extract longitude and dimension from longitude and latitude of GPS

• split function

2 requirement 4: save the results after ETL to a new hive table

Create another table to store the cleaning data

V. Show the result after data cleaning

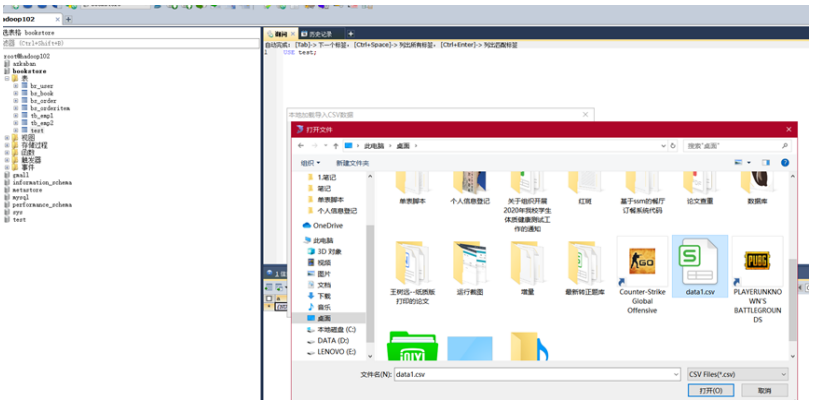
```
hive> Select msg_time, dayinfo, hourinfo, sender_gps, sender_lng, sender_lat from tb_msg_etl limit 10;
OK
2021/11/1 7:44  2021/11/1       7       "123.257181,48.807394"  123.257181      48.807394
2021/11/1 7:29  2021/11/1       7       "113.39623,22.371406"   113.39623       22.371406
2021/11/1 7:57  2021/11/1       7       "117.885171,33.21035"   117.885171      33.21035
2021/11/1 7:26  2021/11/1       7       "121.785398,41.86541"   121.785398      41.86541
2021/11/1 7:01  2021/11/1       7       "124.361019,45.856657"  124.361019      45.856657
2021/11/1 7:23  2021/11/1       7       "86.904123,45.547269"   86.904123       45.547269
2021/11/1 7:57  2021/11/1       7       "119.651311,33.519113"  119.651311      33.519113
2021/11/1 7:42  2021/11/1       7       "114.058533,39.39896"   114.058533      39.39896
2021/11/1 7:17  2021/11/1       7       "117.07569,34.010827"   117.07569       34.010827
2021/11/1 7:15  2021/11/1       7       "112.439571,33.148464"  112.439571      33.148464
Time taken: 0.191 seconds, Fetched: 10 row(s)
```

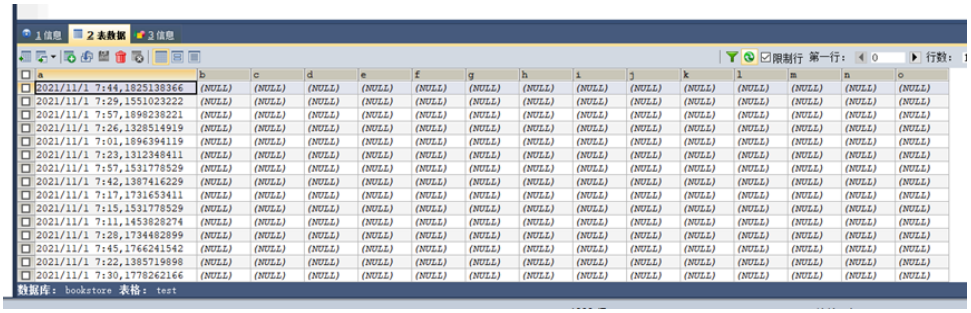
VI. Using SQL to do the EDA

1. search the sender numbers about the sex, telephone system type and network.
We found most people use android.
2. search the numbers about the sex different between female and mail
We found male would like to send messages rather than females, and the receivers are males too.

hbase

I. View the dataset.





II. Created a new table named hbase_test and a column family called info.

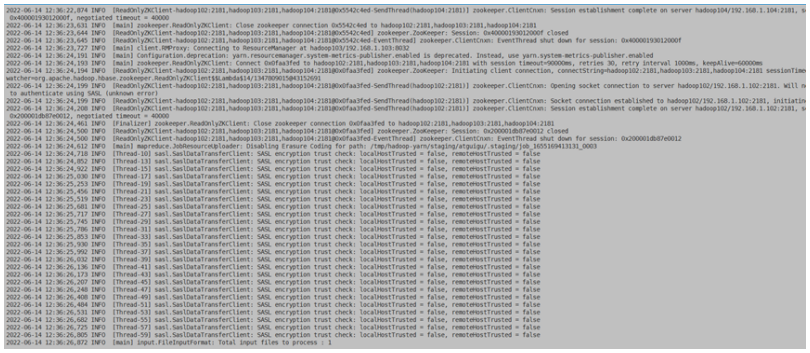
```
hbase(main):006:0> list
TABLE
users
1 row(s)
Took 0.0134 seconds
=> ["users"]
hbase(main):007:0> create 'hbase_test','info'
Created table hbase_test
Took 2.3067 seconds
```

III. Granted the permission to hbase_test.

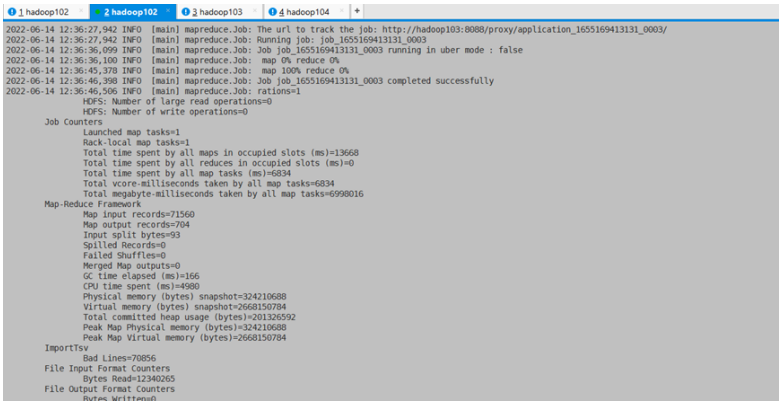
```
=> Hbase::Table - hbase_test
hbase(main):008:0> grant 'root','RWCA','hbase_test';
hbase(main):009:0* user_permission 'hbase_test'
Took 0.1255 seconds
User                          Namespace,Table,Family,Qualifier:Permission
 atguigu                      default,hbase_test,,: [Permission: actions=READ,WRITE,EXEC,CREATE,ADMIN]
 root                         default,hbase_test,,: [Permission: actions=READ,WRITE,CREATE,ADMIN]
2 row(s)
Took 0.0531 seconds
```

IV. Run the sqoop. Load the data into sqoop

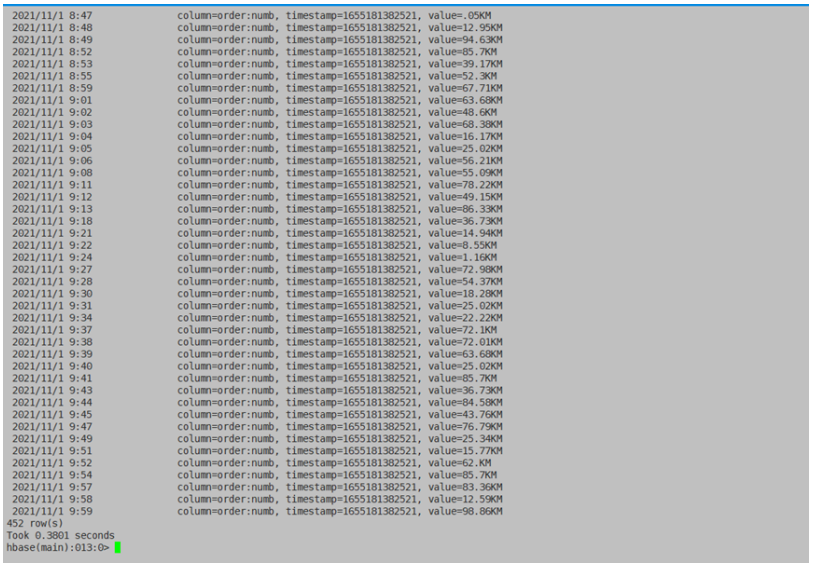
```
[atguigu@hadoop102 sqoop]$ bin/sqoop import --connect jdbc:mysql://hadoop102:3306/bookstore \
>  --username 'root' --password '123321' \
>  --split-by a \
>  --hbase-table 'hbase_test' \
>  --hbase-row-key 'a,b,c,d,e,f,g,h,i,j,k,l,m,n,o' \
>  --column-family 'info'
/opt/module/sqoop/bin/sqoop ==================
/opt/module/sqoop/bin ====================
====================/opt/module/sqoop/bin
Warning: /opt/module/sqoop/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /opt/module/sqoop/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/module/hadoop-3.1.3/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinc
SLF4J: Found binding in [jar:file:/opt/module/hbase/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2022-06-14 12:22:52,161 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
2022-06-14 12:22:52,173 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
--table or --query is required for import. (Or use sqoop import-all-tables.)
```



V. Import dataset

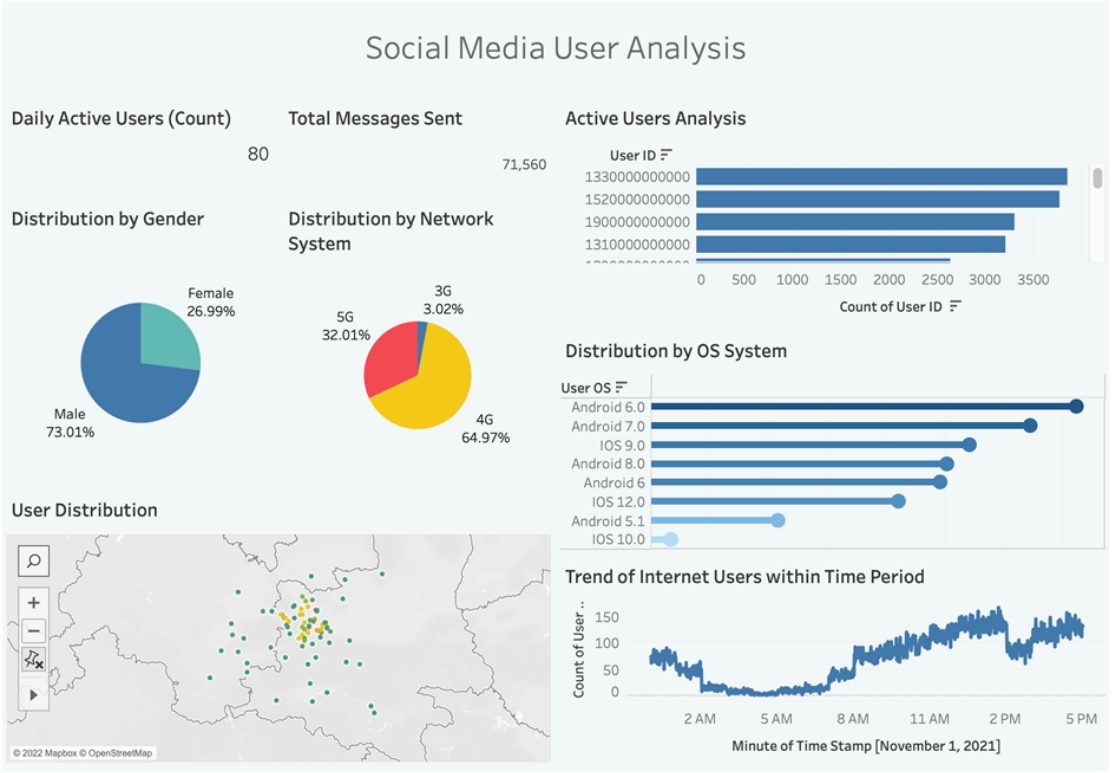


VI. Validate the data



Data Visualization

Data visualization was performed using Tableau. The tool was chosen because Tableau allows users to connect Apache Hadoop as a datasource through Cloudera's driver directly. This allowed us to complete analytics on top of massive data rapidly. We generated 6 graphs and 2 summary cards to visualize the overall performance of Momo using our dataset. We tried to conclude some key performance indicators that could assist the app management team to better keep track of their users and further instruct operation decisions.



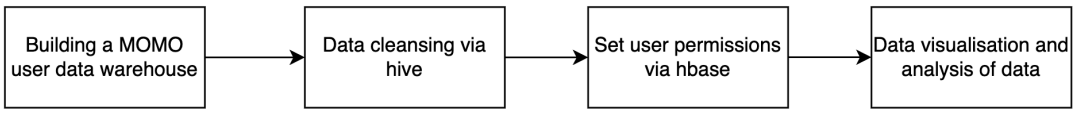
Results & discussion

No.	Data description	Description of the conclusions
1	Male: 73% Female: 27%	MOMO users are predominantly **male**
2	3G: 3.02% 4G: 64.97% 5G: 32%	MOMO users are predominantly on **4G networks**
3	Number of Android users: 73%	MOMO users are mainly on **Android OS System**
4	Most active time: 13pm	Users' internet access is concentrated in the **13pm** time slot

With the final visual analysis, it is easy to see that we have a total of 80 users who sent 71,560 messages on November 1, 2021. Over 73% of users are male and approximately 27% are female.

4G is also by far the most widely distributed, with nearly 73.01% of males online, which is a relatively large percentage. In addition, users of the Stranger platform are predominantly Android OS users.

The three most active users sent over 3,000 messages on 1 November 2021. In addition, the data shows that platform users are mainly located in the western province of Shaanxi, with internet access time slots concentrated around 1pm, with an upward trend from morning to 2pm, then gradually declining and rising.



In this assigment, we first built the momo user data warehouse through hadoop to manage momo's social media data. Then we cleaned it with hive to remove the dirty data. Then we set up data user permissions via hbase for better access and management of social media data. Finally, the momo user data was analysed through visualization of the data.

In addition, in using Hadoop, we found that the Hadoop ecosystem provides various open source tools for collecting, storing and processing data, as well as cluster deployment, monitoring and data security. Also, this research on momo user behavior data has benefited from the three components of Hadoop which are HDFS, MapReduce and HBase. and, we found that it can be applied to offline computing and mass data storage in addition to being well suited for analysing mass data. We will continue to explore the advantages and application scenarios of hadoop in future projects to really bring out its value.

Conclusions

Reference.

1.Edlich, S., & Vogler, M. (2013). Smart apps for applied machine learning on mobile devices: the MOMO project. *Https://Doi.Org/10.1117/12.2005173*, *8667*, 224–233. https://doi.org/10.1117/12.2005173

2. Xu, D., & Wu, F. (2019). Exploring the cosmopolitanism in China: examining mosheng ren ("the stranger") communication through Momo. *Https://Doi-Org.Ezproxy.Um.Edu.My/10.1080/15295036.2019.1566629*, *36*(2), 122–139. https://doi.org/10.1080/15295036.2019.1566629

3. Chan, L. S. (2019). Paradoxical Associations of Masculine Ideology and Casual Sex Among Heterosexual Male Geosocial Networking App Users in China. *Sex Roles 2019 81:7*, *81*(7), 456–466. https://doi.org/10.1007/S11199-019-1002-4

4. Li, H. (2020). Negotiating Privacy and Mobile Socializing: Chinese University Students' Concerns and Strategies for Using Geosocial Networking Applications: *Https://Doi-Org.Ezproxy.Um.Edu.My/10.1177/2056305120913887*, *6*(1). https://doi.org/10.1177/2056305120913887