

**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. According to the report from Kepios, there are around 4.65 billion social media users worldwide until April 2022, accounting for 58.7% of the world population. (Kemp, 2022) These users are generating trillion of data on these platforms. Data generated on social media platform are massive in volume, high velocity and vary type of format.

Momo is a social media and instant messaging applications that allows users to communicate with friends and explore new connections in China. (Wikipedia contributors, 2022). The app allows users to form and build social interactions based on shared geographical regions and interests. The software reported that there were 110.9 million monthly active users in March 2022. There was a decline of 5.3 million as compared to March 2021. (*Hello Group Inc. Announces Unaudited Financial Results for the First Quarter of 2022*, 2022)

Identifying potential users is crucial for business development, by understanding user’s behavior could effectively assist business to potential users or detect the reasons of user churn. However, considering the massive amount of high velocity data generated on social media platform daily, it is challenging for businesses to extract valuable information to instruct business decision. Therefore, 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 considered being an efficient platform to manage big data.

# 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 loss is one of the important problems that Momo needs to consider for business development. 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 a 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)

# Research Objectives

1. To set up the data warehouse and build a data pipeline to manage the social media data generated by Momo users.
2. To remove noisy data and meaningless data after the exploratory data analysis on 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 perform data visualization based on Momo users’ data.

# Research Questions

1. How to set up a data warehouse and build data pipeline to manage social media data generated by Momo users?
2. How to remove noisy data and meaningless data using Hive?
3. How to build data access solution using HBase?
4. How to perform data visualization based on Momo users’ data?

# 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 focuses mainly on users' privacy issues and strategies. 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)

In the research of Bello-Orgaz et al. (2016), the author suggested that the convergence of big data technology and classical machine learning algorithms has created new and exciting issues in domains such as social media and social networks. These new difficulties are primarily concerned with issues like as data processing, storage, and representation, as well as how data may be utilized for knowledge discovery, analyzing user behavior, visualizing, and monitoring data.

# Methodology

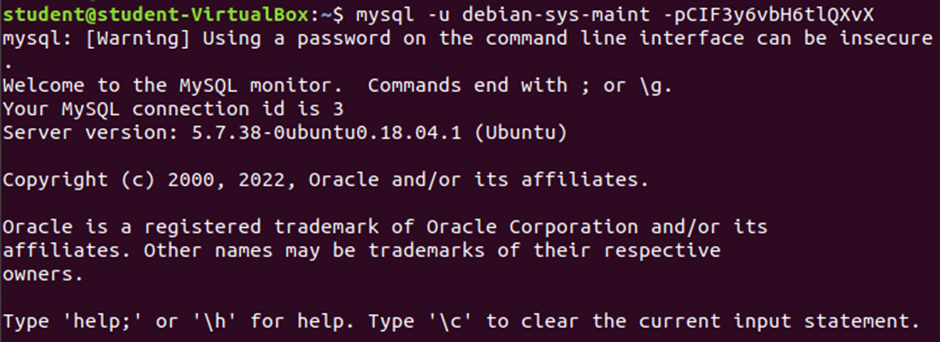
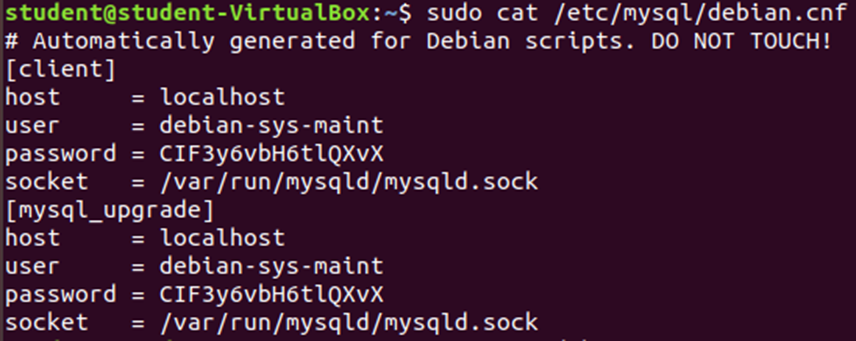
## Data Collection

We obtained the dataset from internet. 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 Momo on 1st November 2021.

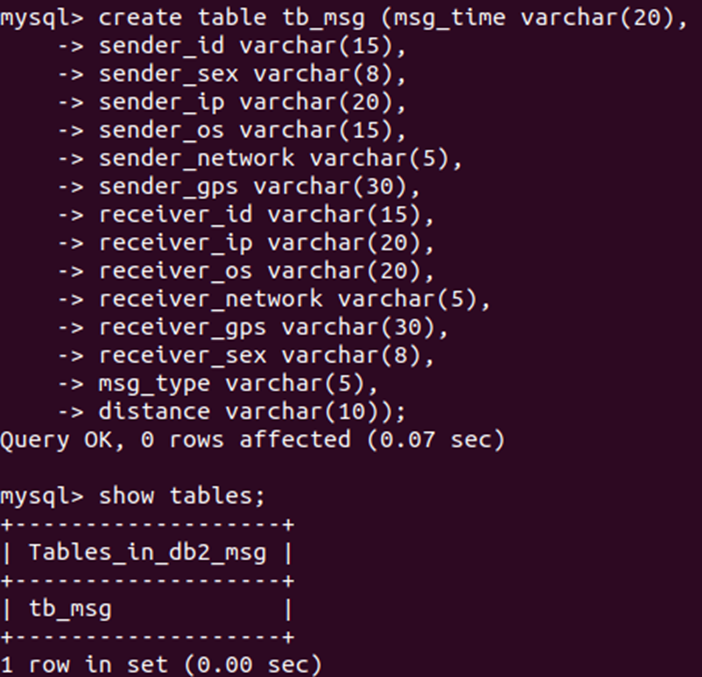
The dataset obtained from:  https://blog.csdn.net/clark123432343/article/details/124409046

## HDFS

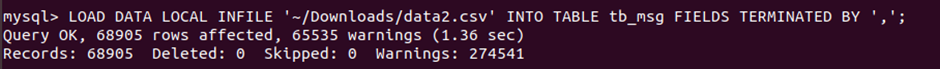
1. Setting the root password in through MySQL



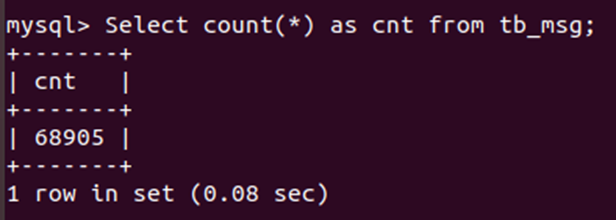
1. Create a database and a specific table of data with the correct column type from the dataset and check if the table framework has been created successfully in the correct database.



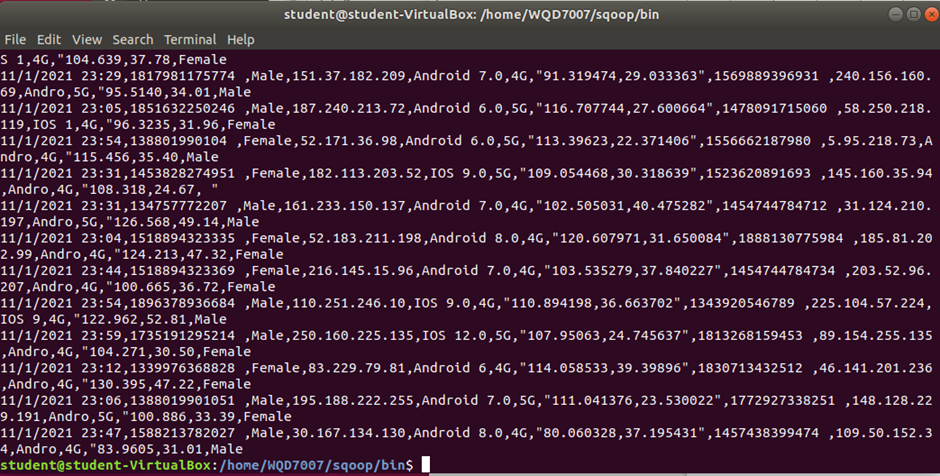
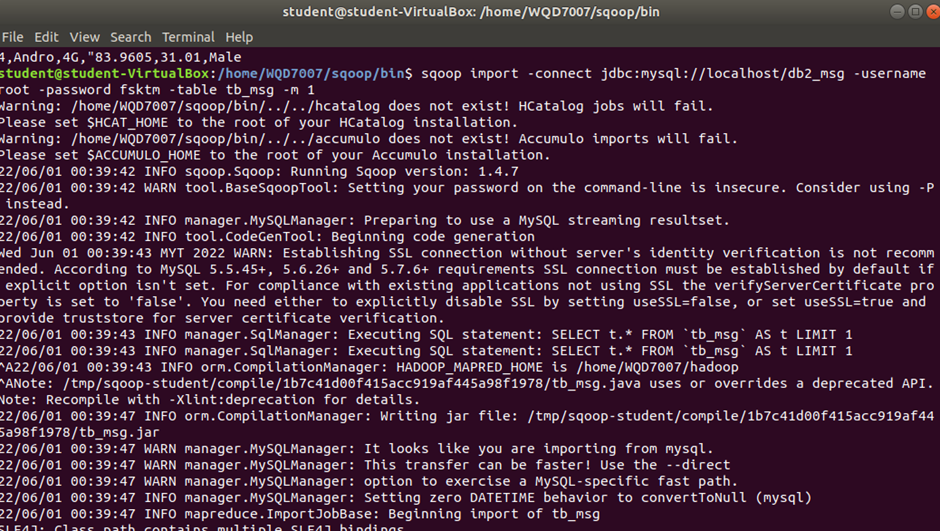
1. Load the data from local to the table created.



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



1. 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.

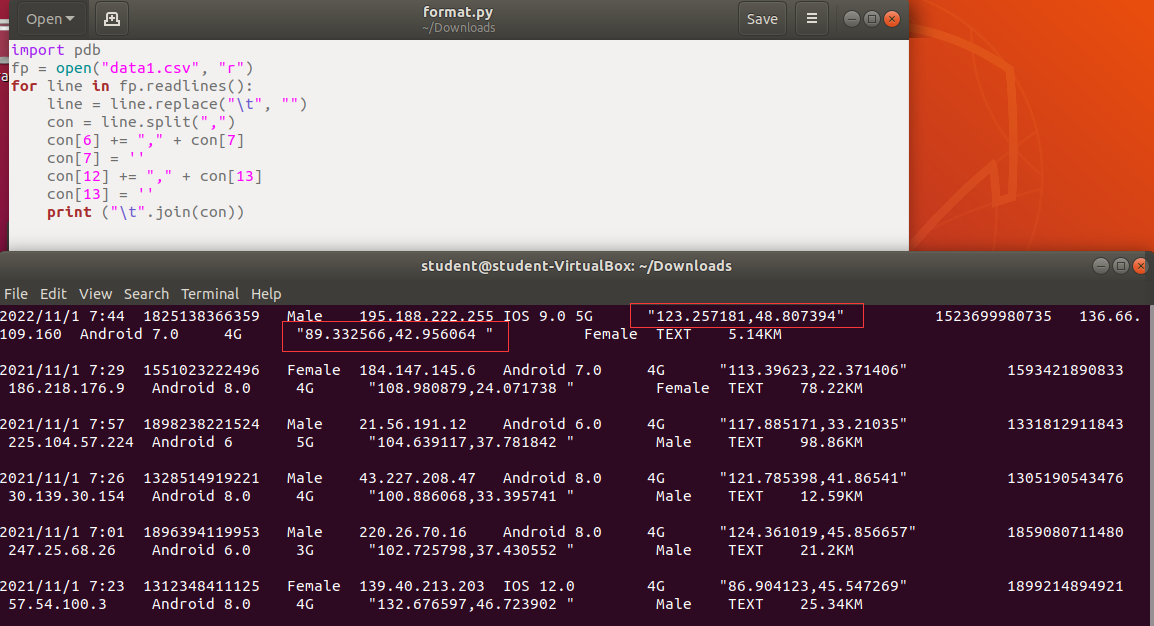


## 

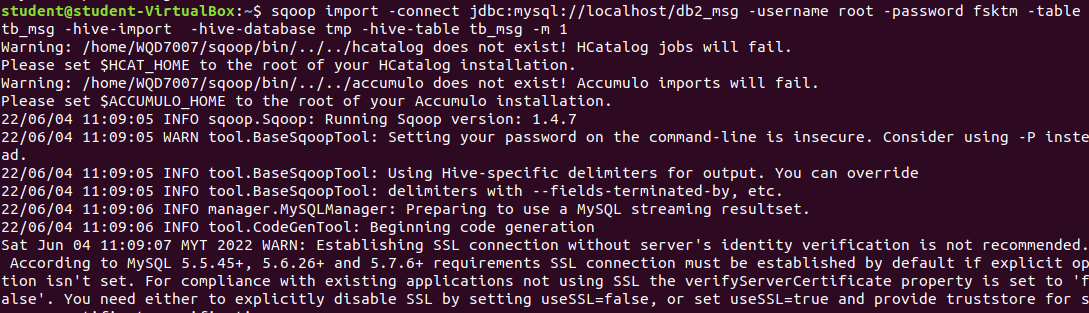
## Hive

For the second part, we will introduce Hive to assist us on data processing.

1. Amend data format: after processed, we could observe the data format as below:



1. Create database 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.



1. Data problem
2. Empty data

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

1. Time format

Problem 2: According to our objectives, 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

1. GPS location problem

Problem 3: According to our objectives, 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

1. Data cleaning

Requirements: filter illegal data with empty fields

* + where filtering

Requirement 2: build day and hour fields through time fields

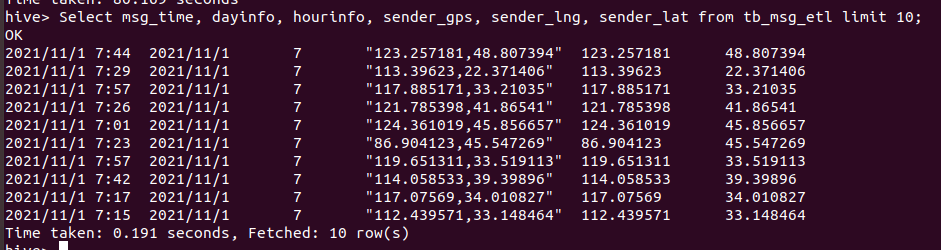
* + substr function

Requirement 3: extract longitude and dimension from longitude and latitude of GPS

* + split function

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

1. Create another table to store the cleaning data
2. Show the result after data cleaning

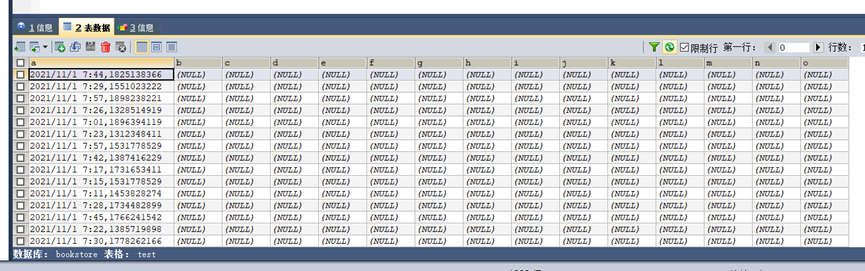
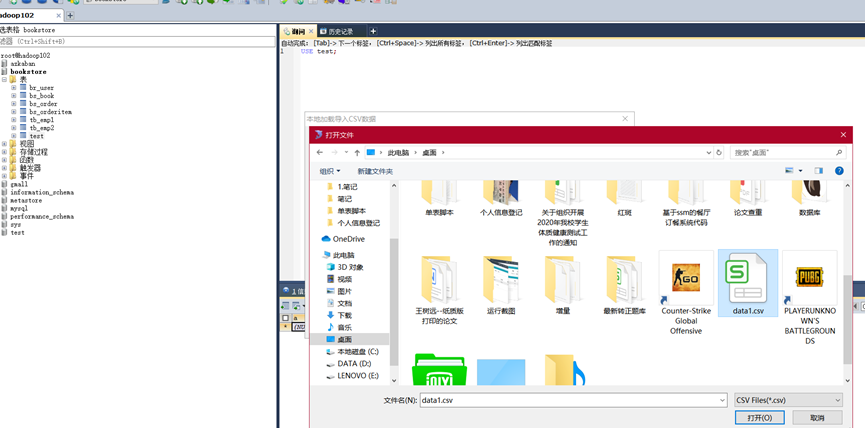


1. Using SQL to do the Exploratory Data Analysis (EDA)
2. Explore the distributions of sender through gender, telephone system type and network dimension.
3. We found most people use android.
4. Discover the distributions of users from gender perspective.
5. We found that the number of male users is higher than female users, and the majority of receivers are male users as well.

## 

## HBase

1. View the dataset.



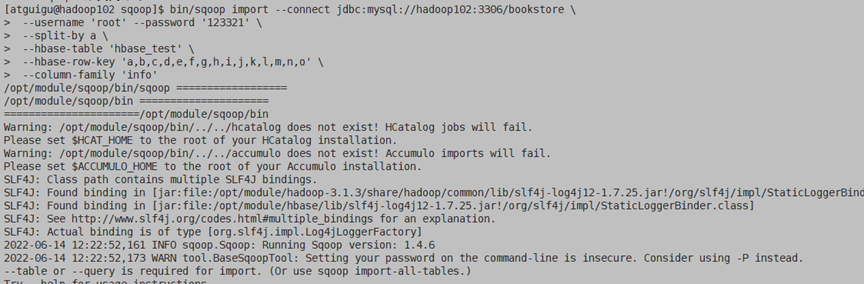
1. Created a new table named hbase\_test and a column family called info.

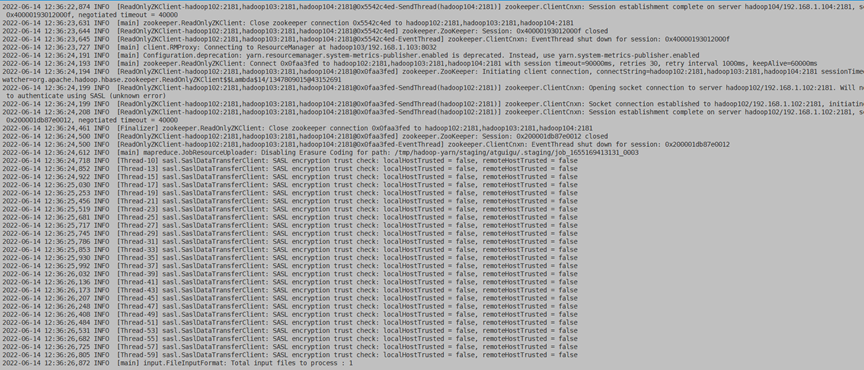


1. Granted the permission to hbase\_test.



1. Run the Sqoop. Load the data into Sqoop





1. Import dataset



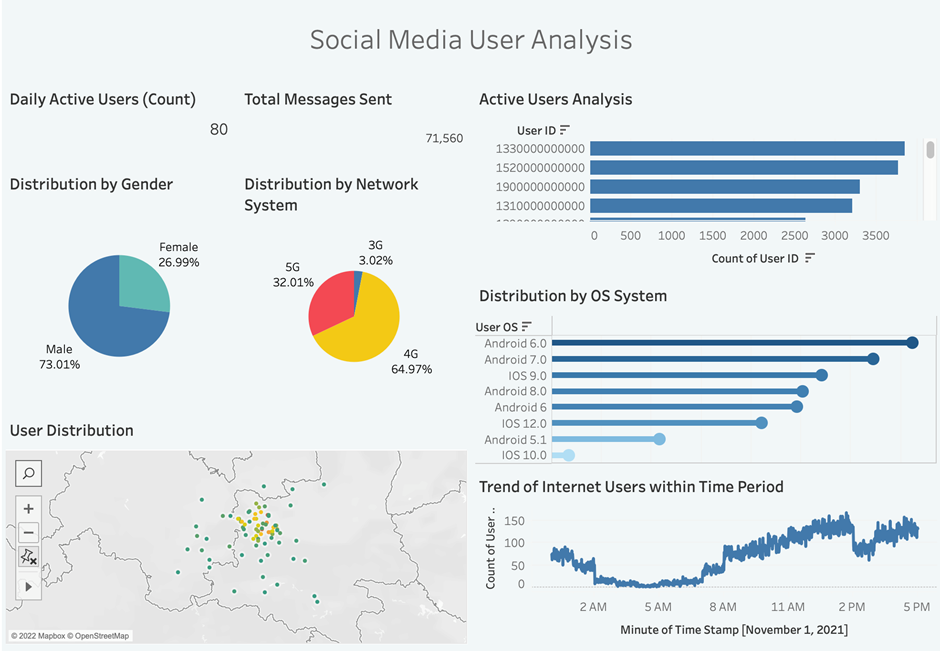
1. Validate the data

# 

## 

## Data Visualization

Data visualization was performed using Tableau. The tool was chosen because Tableau allows users to connect Apache Hadoop as a data source 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.



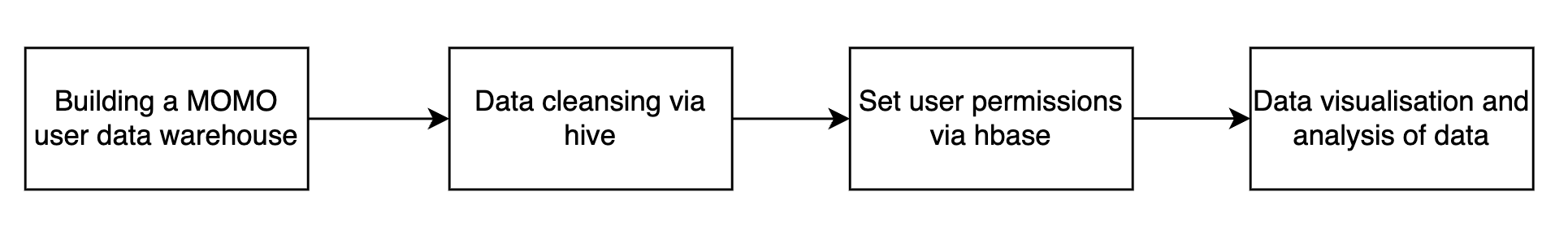
# Results and Discussion

|  |  |  |
| --- | --- | --- |
| **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 **1pm** 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 research project，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 analyzed 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. We found that it can be applied to offline computing and mass data storage in addition to being well suited for analyzing mass data. We will continue to explore the advantages and application scenarios of Hadoop in future projects to really bring out its value.

# Conclusions

Social big data has become an important issue in a wide range of study domains due to the vast number and quick expansion of social networking applications and systems. In this paper, we examine the process of building a secure and reliable data warehouse using HDFS, MapReduce and Hbase to manage and store massive amount of high velocity data and used Tableau to perform user analysis dashboard.

It is proposed to expand the research to text analytics and sentimental analysis using users text message data in the future. Besides, current study has been limited to one day data, future study can consider data with longer time span.

# Reference

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, 45–59. https://doi.org/10.1016/j.inffus.2015.08.005

Chan, L. S. (2019). Paradoxical Associations of Masculine Ideology and Casual Sex Among Heterosexual Male Geosocial Networking App Users in China. *Sex Roles*, *81*(7–8), 456–466. <https://doi.org/10.1007/s11199-019-1002-4>

Edlich, S., & Vogler, M. (2013). Smart apps for applied machine learning on mobile devices: the MOMO project. *SPIE Proceedings*. <https://doi.org/10.1117/12.2005173>

*Hello Group Inc. Announces Unaudited Financial Results for the First Quarter of 2022*. (2022). <https://momoinc.gcs-web.com/static-files/d25103a9-c151-4ab3-80c6-d0c0f0831396>

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management*, *44*, 1–12. <https://doi.org/10.1016/j.ijinfomgt.2018.09.003>

Kemp, S. (2022, May 4). *Digital 2022: Global Overview Report*. DataReportal – Global Digital Insights. <https://datareportal.com/reports/digital-2022-global-overview-report>

Li, H. (2020). Negotiating Privacy and Mobile Socializing: Chinese University Students’ Concerns and Strategies for Using Geosocial Networking Applications. *Social Media + Society*, *6*(1), 205630512091388. https://doi.org/10.1177/2056305120913887

Wikipedia contributors. (2022, June 8). *Momo (software)*. Wikipedia. <https://en.wikipedia.org/wiki/Momo_(software)>

Xu, D., & Wu, F. (2019). Exploring the cosmopolitanism in China: examining *mosheng ren* (“the stranger”) communication through Momo. *Critical Studies in Media Communication*, *36*(2), 122–139. https://doi.org/10.1080/15295036.2019.1566629