Chicago Divvy Bike Sharing Data Report

Smit Shiroya, Linray Song, Kathryn Joy Mak, Pratik Parmar
Department of Information Systems,
California State University, Los Angeles

Abstract:

Americans are falling in love with bike share. Today, there are more than 100 bike-share systems across the country, operated by eight major companies. According to a report just released by the National Association of City Transportation Officials (NACTO), residents and tourists took 84 million were taken on shared micro mobility (shared bike and e-scooter) across 69 major North American cities. Approximately 36.5 million trips were taken in 2018 across 69 major North American cities, an increase of nine percent from the year previous [1]. Divvy has been a tremendous success story in Chicago, is evolving and responding to the way people want to ride. With 6,000 bikes available in 570+ Divvy stations in Chicago and Evanston. Divvy riders pedaled over 7 million miles in 2017, the equivalent of 293 trips around the globe. In just over four years, Divvy has grown to 37,000+ annual members and hundreds of thousands of riders annually. In this report, we will analyze Divvy's dataset with the aim to create insightful, rich, and trustworthy research findings.

1. Introduction

Divvy is operated by the bike share company Motivate International, Inc. under contract with the Chicago Department of Transportation (CDOT). Divvy is largely funded from membership, usage fee, and partnership between Blue Cross Blue Shield and the Federal Government. Divvy provides residents and visitors with a convenient, fun and affordable transportation option for getting around and exploring Chicago. For the low cost of $3, riders can get a thirty minutes ride and return from any of the more than 570 Divvy stations, including more than 200 that are within a quarter of a mile of a CTA or Metro Station. Divvy, like other bike share systems, consists of a fleet of specially designed, sturdy and durable bikes that are locked into a network of docking stations throughout the region [2].

1.1 How it works: Divvy Bike

Divvy is available for use year round, 24 hours a day. The bike can be unlocked through Divvy's app or station kiosk. Riders can unlock, ride, and return from any station in the system to commute to work, or school, or explore Chicago, and more. The geographical area served: Chicago's North and West Sides, much of the South and Southwest Sides. The network extends west and north into close-in suburbs such as Evanston and Oak Park. Densest coverage is in core Chicago neighborhoods such as Near North and the Loop. Riders have the option of choosing from $3 single ride, $15 day pass, or $99 annual membership as shown in Figure 1.

Figure 1. Pass Options for the Bike.

2. Research Objectives

The goal of this paper is to analyze and visualize the data of Chicago's Divvy Bike Sharing Program and present actionable insights to the actually changing industry. The objectives are:

i. Learn how to download data into the local systems in AWS.
ii. Learn how to upload it to HDFS
iii. Figure out how to manipulate and analyze the data in HDFS using HiveQL
iv. Practice how to visualize the result in Excel and Power BI.

3. About the Dataset

The data for Chicago Divvy Bikes has been retrieved from the Kaggle.com website [3]. The data contains all the bicycle data from 2013 to 2017. The file size wax 5 GB. There was one file in the CSV (Comma Separated Values) format. The total number of entries was 9.57 millions. With the data analytic tools, we can uncover new and valuable insights such as the following: where do Divvy Bike Share riders go, the distance traveled, which day of the week are most ride taken on, or which stations are most popular and more.

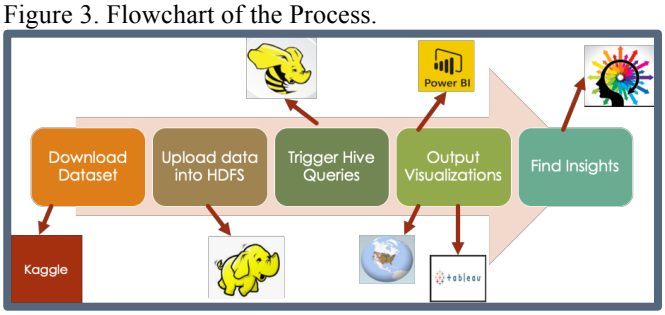
3.1 Platform Specifications

The cluster version used was Amazon Web Services. The number of nodes was 3. The memory size was 150 GB. It was a 20 vCPU with a speed of 2.5 GHz. The HDFS capacity was 147 GB and the storage space was 678 GB.

4. Analyzing the Dataset

These are the steps for the flowchart of the entire process demonstrated in Figure 3. The data was downloaded from Kaggle website and uploaded onto HDFS. Next, Hive Queries were triggered. The output was visualized, and lastly, insights were found.

Figure 3. Flowchart of the Process.



4.1 Getting Data Manually using Hadoop-Hives

Figure 4. SQL Query to get records

```
0: jdbc:hive2://localhost:10000/default> SELECT usertype, COUNT(usert
ype) FROM b_data GROUP BY usertype HAVING COUNT(usertype) > 1;
+-------------+----------+
|  usertype   |   _c1    |
+-------------+----------+
| Dependent   | 178      |
| Customer    | 1277     |
| Subscriber  | 9493780  |
+-------------+----------+
3 rows selected (40.865 seconds)
0: jdbc:hive2://localhost:10000/default>
```

SQL Query:
SELECT usertype, COUNT(usertype) FROM b_data GROUP BY usertype HAVING COUNT(usertype) > 1;

Figure 5. SQL Query to get records

```
0: jdbc:hive2://localhost:10000/default> SELECT COUNT(1),gender FROM b_data GROUP BY gender;
+---------+---------+
|   _c0   | gender  |
+---------+---------+
| 1       | NULL    |
| 2378675 | Female  |
| 7116560 | Male    |
+---------+---------+
3 rows selected (39.809 seconds)
0: jdbc:hive2://localhost:10000/default>
```

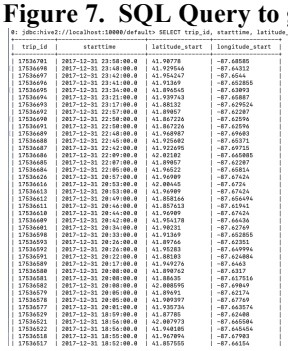
SQL Query:
SELECT COUNT(1),gender FROM b_data GROUP BY gender;

Figure 6. SQL Query to get records

```
0: jdbc:hive2://localhost:10000/default> SELECT month, COUNT(month) FROM b_data GROUP BY month HAVING COUN
T(month) > 1;
+--------+---------+
| month  |   _c1   |
+--------+---------+
| 5      | 869584  |
| 10     | 997712  |
| 11     | 643556  |
| 9      | 1189971 |
| 2      | 305975  |
| 7      | 1279505 |
| 12     | 394672  |
| 3      | 437929  |
| 1      | 271715  |
| 6      | 1178225 |
| 4      | 621688  |
| 8      | 1304703 |
+--------+---------+
12 rows selected (36.7 seconds)
0: jdbc:hive2://localhost:10000/default>
```

SQL Query:
SELECT month, COUNT(month) FROM b_data GROUP BY month HAVING COUNT(month) > 1;

Figure 7. SQL Query to get records



SQL Query:
SELECT trip_id, starttime, latitude_start, longitude_start FROM b_data ORDER BY starttime DESC LIMIT 150000;

4.2 Visualization using Power BI

In order to achieve a better quality of data analysis, the dataset was imported into Power BI for visualization. In this project, we propose to predict the potential challenges of Divvy and some suggestions to address the issues. The user composition of bikesharing systems can be quite diverse, which include short-term temporary users and long-term registered subscribers, as well as male users and female users. The predict challenges can be strongly correlated to user categories, and a clear categorization of the bike users can be a prerequisite for addressing the problem.

Figure 8. Breakdown of User Type

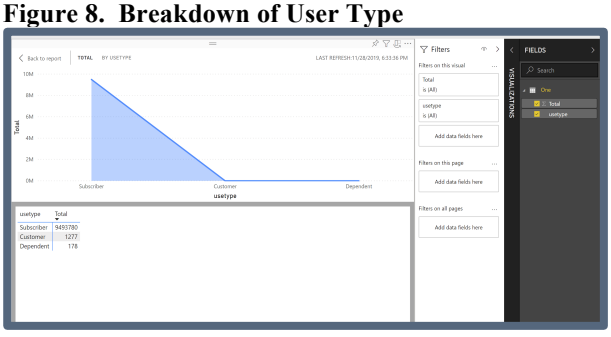


Figure 8 shows there are two types of Divvy users: subscribers and customers. Subscribers pay $99 a year and get unlimited rides under 45 minutes; these are often commuters to work. Customers pay as they go, they can purchase a daily pass for $15 and get unlimited trips on that day or for $3, one trip up to 30 minutes. These can be great options for tourists to see the city for cheap. Moreover, Divvy can consider to expand its services with universities

and associations located in Chicago to attract more potential customers such as students and employers by offering discounts. Divvy focuses on providing more affordable, environment-friendly and cheaper transportation to residents, so it started to apply Divvy for everyone policy which gives an opportunity to qualifying residents one-time annual membership with $5 of charge, which helps expanding its target market.

Figure 9. Gender Gap

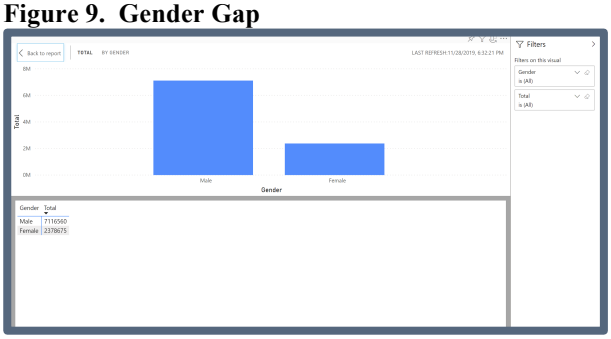
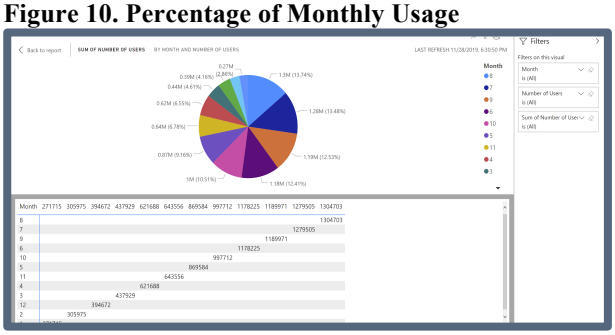


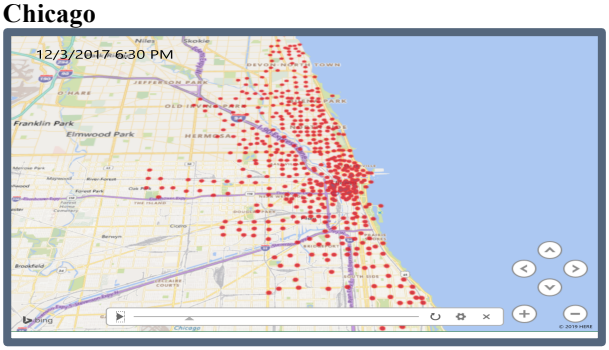
Figure 9 presents 1 out of 3 rides is taken by females meaning that male ridership occurred three more times than that of females. There is a dramatically high proportion of men using Divvy bikes over women. In this sense, a significant gender gap in Divvy bike sharing system might give us clues regarding what deficiencies transportation planners should meet so as to increase their participation in the system.

Figure 10. Percentage of Monthly Usage



From Figure 10, we observe the monthly usage of Divvy bike, but the majority of the trips concentrate within the months ranging from April to November in 2017, and the number of trips taken during the winter seasons is quite limited. Anyone that has ever been to the Midwest, Chicago in particular, knows that weather can vary greatly each month. Weather can be a great influence as to whether people want to ride a bike around Chicago or not. Figure 10 shows how those trips were distributed monthly. Given the weather patterns in Chicago, it is no surprise that the number of trips is drastically higher in the summer than in the winter months. Chicago has the best warm weather during summer time; therefore, this visualization shows from April to November, Divvy has the highest demands in bike rental.

Figure 11. Distribution of Divvy Stations across Chicago



As for Figure 11, this map shows where check-in and check-out stations are located and how vast the network of Divvy bikes reaches. As seen in the map, the stations are very clustered in the middle which is the downtown area where much of the business is done throughout the day. Figure 11 also indicates the usage of each station, it becomes very clear where the majority of the rides are taken throughout the city. Not only is downtown the business center of the city, it is also where many of the major attractions for visitors and tourists. These attractions include, Navy Pier, Millennium Park, Museum Drive, the Art Institute as well as Michigan Avenue & State St where much of the retail market is. Due to the combination of the business and attractions, both subscribers and customers are likely to frequent these stations downtown, making them more vital for Divvy to ensure that these stations are always usable to pick up and drop off bikes.

5. Conclusion:

By using historical data collected on riders, the data can help the company to have a better understanding of exactly customer travel journey and other valuable insights. Having the right data at the right times is essential to predict when people will travel, where, and how long means companies can provide exact service their customers need at the best time and at the right price. In this project, we have demonstrated how data analysis is used for making business decisions and give Divvy company an edge over competitors in a tight market.

- From 2013 to 2017, there is an increasing demand for sharing bicycles. Since the population of Chicago is relatively constant, we can assume that people tend to live healthier as time goes on.
- People use sharing bicycles more during summer as compared to winter based on the frequent usage and longer trip duration.
- Most users purchased the Annual Membership. But their trip durations are relatively shorter compared with ordinary customers and dependent. Those who purchased 24-hour pass tend to have longest trip.

- Gender imbalance is one of the important findings in Divvy. As of 2017, male riders use sharing bicycles more often than female rider according to the data.
- Most usage is in downtown Chicago. There are some stations rarely used.

References:

[1] Ink, Social. "84 Million Trips Taken on Shared Bikes and Scooters Across the U.S. in 2018." *National Association of City Transportation Officials*, 17 Apr. 2019, nacto.org/2019/04/17/84-million-trips-on-shared-bikes-and-scooters/.

[2] Divvy Bikes. (2019). *Motivate International, Inc. "About Divvy: Company & History"*. Retrieved from www.divvybikes.com/about

[3] Zhoa, J. (2017). *Chicago Divvy Bicycle Sharing Data*. Retrieved from https://www.kaggle.com/yingwurenjian/chicago-divvy-bicycle-sharing-data#data.csv