Data Analytics on Big Data from Chicago Divvy Bike Sharing Program

Kay Mak, Pratik Parmar, Smit Shiroya, Linray Song | Dr. Ming Wang, Dr. Jongwook Woo | California State University, Los Angeles

Problem

How can Chicago Divvy Bike Sharing Program adapt to a rapidly changing business world?

Research Questions

Today, there are more than 100 bike-share systems across the country, operated by eight major companies. Divvy has been a tremendous success story, with 6,000 bikes available in 570+ 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, the company has grown to 37,000+ annual members and hundreds of thousands of riders annually. In order to stay competitive, it is important for Divvy to adapt and respond to the way people want to ride. By using historical data collected on riders, the data can help the Divvy to have a better understanding of exactly customer travel journey such as the following: where do Bike Share riders go, who are the usertypes, which months are most ride taken on, or which stations are most popular for starting rental time and more.

Research Objectives

We will analyze Chicago Divvy Bike Sharing Data with the aim to create insightful, rich, and trustworthy research findings. The learning objectives include:

- Learn how to download data to the local systems in AWS
- Upload data to HDFS.
- Analyze data in HDFS using HiveQL.
- Visualize the results in Excel, Power BI, and Tableau.

Research

Getting Data using Hadoop-Hives

```
trip_id  |      starttime      | latitude_start | longitude_start
---------+---------------------+----------------+-----------------
17536701 | 2017-12-31 23:58:00.0 | 41.90778  | -87.68585
17536698 | 2017-12-31 23:48:00.0 | 41.929546 | -87.64312
17536697 | 2017-12-31 23:42:00.0 | 41.954247 | -87.6544
17536696 | 2017-12-31 23:41:00.0 | 41.91369  | -87.652855
17536695 | 2017-12-31 23:34:00.0 | 41.896545 | -87.63093
17536694 | 2017-12-31 23:21:00.0 | 41.939743 | -87.65887
17536693 | 2017-12-31 23:17:00.0 | 41.88132  | -87.629524
17536692 | 2017-12-31 22:57:00.0 | 41.89057  | -87.62207
17536690 | 2017-12-31 22:50:00.0 | 41.867226 | -87.62596
17536691 | 2017-12-31 22:50:00.0 | 41.867226 | -87.62596
17536689 | 2017-12-31 22:48:00.0 | 41.968987 | -87.69603
17536688 | 2017-12-31 22:45:00.0 | 41.925602 | -87.65371
17536687 | 2017-12-31 22:42:00.0 | 41.922695 | -87.69715
17536686 | 2017-12-31 22:09:00.0 | 42.02102  | -87.665085
17536685 | 2017-12-31 22:07:00.0 | 41.89057  | -87.62207
17536684 | 2017-12-31 22:05:00.0 | 41.96522  | -87.65814
17536626 | 2017-12-31 20:57:00.0 | 41.96909  | -87.67424
```

SQL Query

- SELECT usertype, COUNT(usertype) FROM b_data GROUP BY usertype HAVING COUNT(usertype) > 1;
- SELECT COUNT(1),gender FROM b_data GROUP BY gender;
- SELECT month, COUNT(month) FROM b_data GROUP BY month HAVING COUNT(month) > 1;
- SELECT trip_id, starttime, latitude_start, longitude_start FROM b_data ORDER BY starttime DESC LIMIT 150000;

Materials

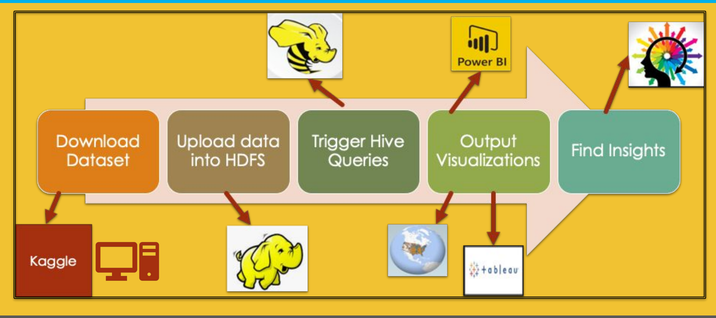
PLATFORM SPECIFICATION

- Cluster Version – Amazon Web Service
- Number of Nodes – 3
- Memory size – 150 GB
- CPU – 20 vCPU
- CPU speed – 2.5 GHz
- HDFS capacity – 147 GB
- Storage – 678 GB

ABOUT THE DATASET

Dataset	http://bit.ly/ChicagoDivvy
Data Reviewed:	2013 to 2017
File Size:	5.00 GB
Number of Files:	1
File Format:	CSV
Total no. of entries:	9.57 million

Flowchart



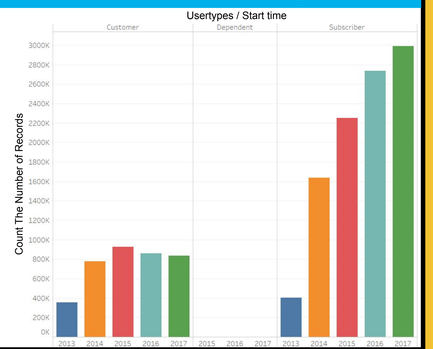
Data Visualization

How many subscribed customers, customers, & customer dependents?

```
0: jdbc:hive2://localhost:1000/default>
SELECT usertype, Count(usertype)
FROM b_data GROUP by usertype
HAVING COUNT(usertype) > 1;
```

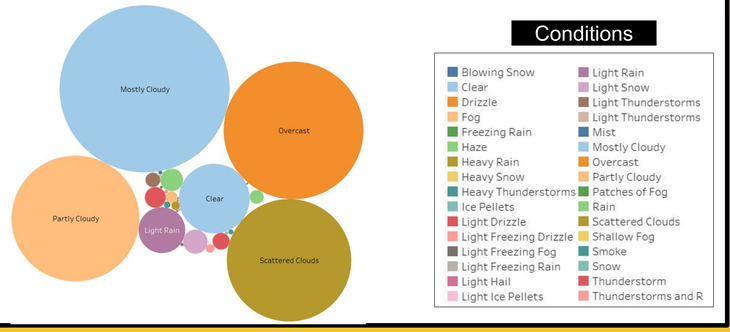
Usertype	_c1
Dependents	178
Customer	1277
Subscribed Customers	9493780

3 row select (40.865 seconds)
0: jdbc:hive2://localhost:1000/default>



What are the popular bike starting rental time blocks & locations?

Which weather conditions affect the bike rental service the most?



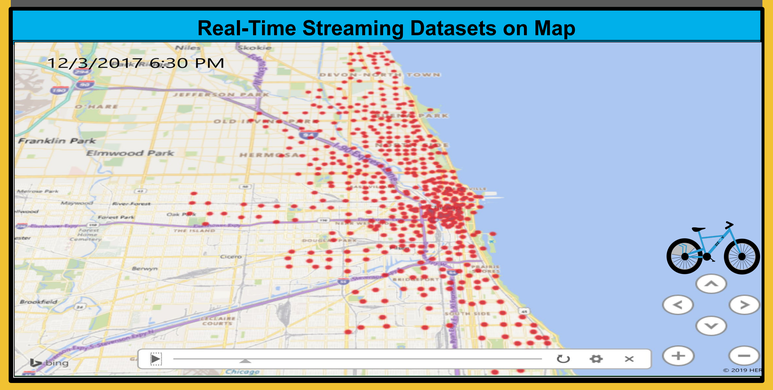
Data Analysis & Results

What is the number of rental in each of the 12 months by year?

From 2013 to 2017, there is an increasing demand for sharing bikes between April to October. Most users purchased the Annual Membership, but their trip durations are shorter compared with random customers who purchased 24-hour pass tend to have longer trip. In this research, we have demonstrated how data analysis is used for making business decisions and give Divvy company an edge over competitors in a tight market.

Conclusion

Real-Time Streaming Datasets on Map



This map shows where check-in & check-out stations are located and how vast the network of Divvy bikes reaches. It becomes clear that the majority of the rides are taken into downtown and tourist areas. Due to high demand for bike rentals near Lakeshores, Navy Pier, Millennium Park, Museum Drive, and the Art Institute. We recommend Divvy to allocate more bikes in popular rental time blocks and locations.

Works Cited

[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