Project Draft

Title: Visualizing Divvy Chicago Bike Data: Redshift and Power BI Analytics

Team: Topic 4

Jayanth Chidananda	jchidananda@hawk.iit.edu	A20517012
Devin Wu	dwu38@hawk.iit.edu	A20484154
Harshali Vasant Gaikwad	hgaikwad@hawk.iit.edu	A20538035
Kalebu Patan	kpatan@hawk.iit.edu	A20446212

Introduction:

Since its inception in 2013, the Divvy Chicago bike system has flourished as a public bike-sharing initiative. Boasting over 800 stations and 6,000 bikes, it caters to a daily ridership exceeding 85,000 individuals. Its impact is evident in the reduction of traffic congestion and the enhancement of air quality, both attributed to the increased use of bicycles in Chicago. The extensive data generated by the Divvy system offers valuable insights into user behaviors and highlights opportunities for system enhancement.

Project Description:

Utilizing Amazon Redshift, this project aims to analyze and visualize data sourced from the Divvy Chicago bike system. With a plethora of information on bike usage, including trip details such as start and end times, locations, and distances traveled, the Divvy system offers rich insights. The primary objectives are:

1. Cleanse and format Divvy data to align with Redshift requirements.
2. Seamlessly load data into Redshift from S3, accommodating various file formats.
3. Perform thorough data analysis, spotlighting popular stations, frequently traveled routes, and peak usage periods.
4. Communicate findings effectively using Power BI's flexible charting functionalities.

Literature Review:

Studies leveraging Divvy data have unearthed valuable insights into travel behaviors and user demographics. One analysis pinpointed popular Divvy stations in downtown Chicago and surrounding neighborhoods, while another delved into user trends based on demographic and bike type distinctions. In terms of data warehousing solutions, AWS Redshift, Azure SQL Data Warehouse, Snowflake, and MemSQL offer distinct advantages. Redshift excels with its optimized SQL querying and seamless integration with AWS services, while Azure SQL Data Warehouse provides scalability and hybrid deployment options. Snowflake stands out for its

cloud-native architecture and support for structured and semi-structured data, while MemSQL offers real-time analytics capabilities and in-memory performance for transaction processing. Each platform caters to different use cases, requiring careful consideration of performance needs and integration requirements.

Data Collection and Preparation:

The dataset utilized for this project comprises trip data sourced from Divvy, encompassing trip start and end times, locations, and rider type, spanning from January 2023 to December 2023. Each dataset, obtained in CSV format, underwent meticulous organization to ensure optimal compatibility. With its substantial sample size, originality, completeness, relevance, and attribution to a cited source, the dataset maintains integrity for robust analysis.

Following the loading of Divvy bike trip data, the team undertook data preparation measures. This involved validating the unique ride_id primary key and consolidating bike types. Additionally, new columns were derived from start and end times to denote ride length, day of the week, and date. Cleanup procedures included the removal of rows with negative ride lengths and outliers, as well as addressing null values in the ending station names of classic bike trips.

Data Ingestion:

The complete dataset was initially stored in an Amazon S3 bucket to facilitate ingestion into AWS Redshift. To accomplish this, a Redshift cluster was configured with a dc2.large node and was granted the necessary permissions to access the S3 buckets. Subsequently, a database named 'Divvy_database' was established using the query editor interface on our cluster. By utilizing the 'load data' option, a direct connection was made between S3 and the Redshift cluster. This connection enabled the automatic creation of a table that included all data points for the year 2023, spanning from January to December.

Data Analysis and Visualization:

The data analysis tasks involve pinpointing the most frequented Divvy stations, identifying popular travel routes, and determining peak usage times of the Divvy system. Subsequently, the analysis results are visualized using Power BI, leveraging its versatile charting capabilities encompassing bar charts, line charts, scatterplots, heat grids, and scorecards.

Conclusion:

This project utilizes Amazon Redshift and Power BI to analyze Divvy Chicago Bike-sharing data, revealing usage patterns and user behavior. Stakeholders can use these insights to strategically place new bike stations, allocate resources effectively and enhance the bike-sharing service.

References:

1. https://divvybikes.com/system-data
2. https://www.kaggle.com/code/devisangeetha/divvy-bike-share-eda-network-analysis
4.https://www.linkedin.com/pulse/data-analysis-visualizations-chicago-divvy-bikes-sharing-mazarei/
5.https://medium.com/codex/exploratory-data-analysis-cyclistic-bike-share-analysis-case-study-1b1a00475a4f
6.https://towardsdatascience.com/predicting-hourly-divvy-bike-sharing-checkouts-per-station-65b1d217d8a4
7. https://www.slideshare.net/HanbitChoi1/divvy-bike-use-data-analysis-and-recommendations

Citations:

J. Zhang, X. Pan, M. Li and P. S. Yu, "Bicycle-Sharing System Analysis and Trip Prediction," 2016 17th IEEE
International Conference on Mobile Data Management (MDM), Porto, Portugal, 2016, pp. 174-179, doi:
10.1109/MDM.2016.35.

Milestones:

Team Member	Task	Responsibilities	Challenges	Due Date
Kalebu Patan	Data Collection	Collect Divvy bike system data for analysis.	Handling diverse data sources and formats.	April 1st, 2024
Kalebu Patan	Data Preparation	Prepare the data through cleaning and structuring for analysis.	Addressing missing data and inconsistencies	April 5th, 2024
Jayanth Chidananda	Data Cleaning and Ingestion	Reviewing data format and transferring data to the redshift database.	Integrating redshift and S3.	April 12th, 2024
Jayanth Chidananda	Data Analysis	Analysis of Divvy data to identify popular stations and routes.	Writing effective queries to combine data points.	April 15th, 2024

Harshali Vasant Gaikwad	Data Integration	Integrated Redshift with powerbi using ODBC driver	Made changes in the inbound rules and added TCP/IP connection to allow traffic.	April 22nd, 2024
Harshali Vasant Gaikwad	Data Visualization	Creating a dashboard for rider type, member type , routes and total rides.	Designing interactive and informative visuals.	April 25th, 2024
Devin Wu	Data Visualization	Developing a comprehensive dashboard featuring advanced visualizations.	Designing a unified view with data from analysis.	April 28th, 2024
Devin Wu	Performance Comparison	Differentiate performance of Redshift, Snowflake and Azure SQL Data Warehouse	Performance analysis of various data storage platforms	May 1st, 2024