**PROJECT DOCUMENTATION**  
*metro\_bike\_2024\_datawarehouse  
by: Christian Cartagena*

**Purpose:** The goal of this project is to analyze metro bike trip data to identify patterns, trends, and insights using a data warehouse.  
  
**Dataset Scope:** The dataset covers bike-sharing data for Q3 2024. This time period was chosen to focus on identifying seasonal trends and peak usage patterns during the summer months. While the data is limited to one quarter, it provides sufficient granularity for detailed analysis.  
  
**Tools used**: BigQuery, DbSchema, Power BI, Lucid charts  
  
**GitHub Repository:** All project files, including ETL scripts, schema diagrams, and visualizations, are available at [GitHub Repository](https://github.com/chriscarts/metro_bike_datawarehouse_2024_q3).

**Overview of the data:**   
**Rows:** 134,919   
**Columns:** 15   
**Attributes:** *trip\_id, duration, start\_time, end\_time, start\_station, start\_lat, start\_lon, end\_station, end\_lat, end\_lon, bike\_id, plan\_duration, trip\_route\_category, passholder\_type, bike\_type***Dictionary***:* [*Metro Bike 2024 - Dictionary*](https://1drv.ms/x/c/7a6541984c4cdcb1/EVRvs2m1XupInyB732g__fMBave6ycTLQSP5-zbSb1PkdQ?e=bAvuFj&nav=MTVfezAwMDAwMDAwLTAwMDEtMDAwMC0wMDAwLTAwMDAwMDAwMDAwMH0)**Source:** [Metro Bike Share Open Data (Los Angeles).](https://bikeshare.metro.net/about/data/)  
The dataset was downloaded in raw format (CSV) and imported into BigQuery for ETL processing.

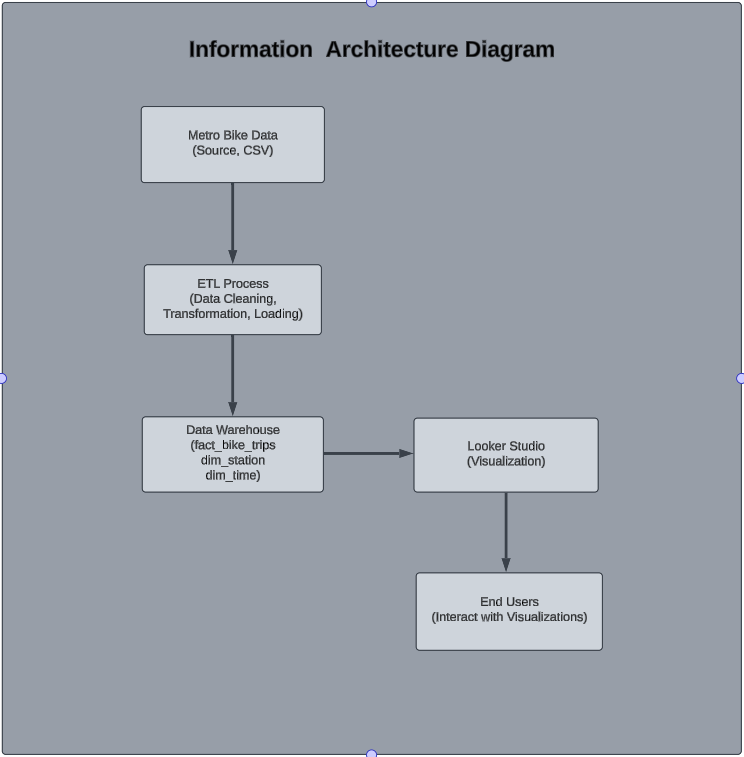
**Requirements**

**Business Requirements:**  
1. **Optimize Bike Station Placement**: Identify high-demand stations based on trip frequency to optimize resource allocation and ensure adequate bike availability.  
2. **Understand Peak Usage Times**: Analyze daily and seasonal trends to determine peak hours/days for bike-sharing activity, helping improve scheduling and resource planning.  
3. **Improve Pricing Models**: Study trip duration patterns to develop tailored pricing strategies that align with user behavior.  
  
**Functional Requirements:**  
1. The system must aggregate trip data to identify the busiest stations by trip count.  
2. The system must calculate daily and seasonal trip trends to highlight peak usage times.  
3. The system must compute average trip duration by station to support pricing strategies.  
4. The system must visualize data with heatmaps, bar charts, and line charts for user-friendly analysis.  
5. The system must allow querying for specific dimensions (e.g., station name, trip date) to enable detailed analysis.

**Compliance with Requirements:**  
1. The dataset was sourced from <https://bikeshare.metro.net/about/data/>.  
2. The data contains granular trip-level information and is not aggregated.  
3. The dataset meets the minimum requirement of 10 columns and 7,500 rows.

**Information Architecture**

**Centralized Data Warehouse:**  
A system designed to store, process, and analyze bike-sharing trip data efficiently. It provides a foundation for visualizations, reporting, and querying through user-friendly tools.

**Data Components:**  
1. **Fact Table (fact\_bike\_trips):** Contains core trip metrics like trip duration and station details.  
2. **Dimension Tables:**  
**dim\_station:** Provides metadata for bike stations (e.g., name, latitude, longitude).  
**dim\_time:** Breaks trip dates into year, month, and day for time-based analyses.  
  
**Structure:**  
1. **Key Information Components:**  
**ETL Process:** Transforms raw trip data into structured tables optimized for analysis. Ensures data consistency and scalability.  
**Storage:** All data is stored in Google BigQuery, enabling seamless querying and integration with visualization tools like Looker Studio.  
  
2. **Roles & Permissions:**  
**Admin Role:** Manages ETL pipelines, data security, and schema updates.  
**Analyst Role:** Accesses cleaned data for querying and visualization, focusing on generating insights.  
  
3. **User Interaction Flow:**  
- Data flows from raw sources (e.g., Metro Bike Data API) into the warehouse via ETL pipelines.  
- Users interact with the warehouse using SQL (via BigQuery) or through dashboards (via Looker Studio).  
- Insights generated support decision-making for business requirements like optimizing station placement and understanding usage patterns.  
  
**Diagram:**Refer to the **Information Architecture Diagram** to visualize the flow of data from source to end-user interactions.  
  


**Benefits**

* **Centralized Data Storage**: All trip-related data is stored in a structured format for easy access.
* **Efficient Analysis**: Pre-processed data enables faster querying and reporting.
* **Actionable Insights**: End users gain valuable insights through interactive visualizations.

**Data Architecture**

The data architecture for this project is designed to ensure seamless integration, processing, storage, and retrieval of Metro Bike Share data. The system includes three key components: data sourcing, processing (ETL), and storage (data warehouse).

**Data Flow and Processing Steps:  
1. Data Source:**- The raw data is obtained from the **Metro Bike Share Open Data API** or downloaded as a CSV file. **-** This dataset contains granular trip-level information, including attributes such as trip duration, start and end station, and geographic coordinates.

**2. ETL Process:**

**1. Extract:** The raw data is ingested into BigQuery, ensuring accurate and complete data capture. **2. Transform:  
-** Data is cleaned to handle missing values, inconsistent column names, and redundant entries. **-** Fact and dimension tables are created: **-** fact\_bike\_trips stores core metrics such as duration and plan\_duration for analysis. **-** dim\_station provides metadata about stations, including latitude and longitude. **-** dim\_time breaks down trip\_date into granular components (year, month, and day). **3. Load:** The processed data is loaded into the data warehouse (BigQuery), ensuring optimal structure for querying and analysis.

3. **Data Storage:** The data warehouse is organized into a **star schema**, with fact\_bike\_trips at the center, linked to dim\_station and dim\_time through foreign keys.

**4. Data Retrieval and Visualization:  
- BigQuery:** Serves as the data warehouse, allowing SQL queries to retrieve specific insights. **- Looker Studio (Visualization):** Connects to the data warehouse to generate heatmaps, bar charts, and line graphs, enabling end users to interact with visualized data.

**How the Schema Supports Data Architecture:**

**Fact Table (fact\_bike\_trips):**  
- Centralizes metrics for analysis.  
- Links to dimensions via foreign keys for filtering and aggregation.

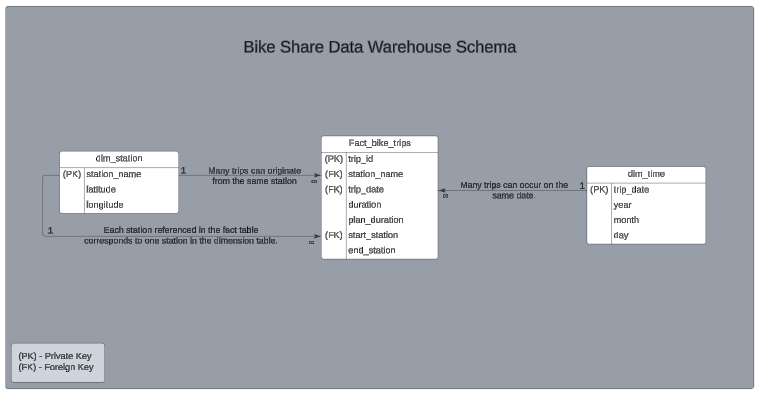
**Dimension Tables (dim\_station and dim\_time):**  
- Provide metadata to enrich analytical capabilities.  
- Facilitate efficient querying by reducing redundancy and enabling drilldowns.

**Conclusion:**This data architecture ensures: **- Scalability:** The data warehouse can handle additional data as new trips are recorded. **- Consistency:** The star schema design maintains data integrity. **- Efficiency:** Queries and visualizations are optimized for performance.

**Schema Design**

**Fact table:** *fact\_bike\_trips stores trip-level metrics such as duration, plan duration, trip date, starting station, and ending station.*  
**Dimension tables:**  
**dim\_station:** Provides station metadata, including station name, latitude, and longitude.  
**dim\_time:** Breaks down trip\_date into year, month, and day.  
**Relationships:** **Fact table links:**  
 - fact\_bike\_trips links to **dim\_station** on station\_name, representing the starting station for each trip. - fact\_bike\_trips links to **dim\_time** on trip\_date, representing the date of the trip.  
  
**Diagram Details:** The schema represents the relationships between the fact and dimension tables:  
**- dim\_station** provides metadata for stations, supporting queries related to station-specific insights.  
**- dim\_time** helps analyze trip trends over time.

**The cardinality indicates:**  
- Many trips can originate from a single station (1 to ∞ relationship with dim\_station).  
- Many trips can occur on the same date (1 to ∞ relationship with dim\_time).



**ETL Process**

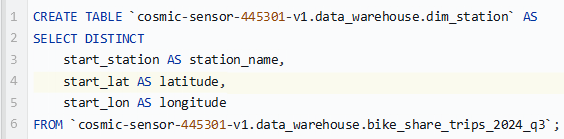
The ETL (Extract, Transform, Load) process for this project involved creating three key tables: *dim\_station, dim\_time, and fact\_bike\_trips*. These tables were built from the raw dataset *(bike\_share\_trips\_2024\_q3)* stored in BigQuery. Each step ensured clean, structured, and meaningful data for analysis.

**1. Creating dim\_station**

The dim\_station table serves as a dimension table containing metadata for bike stations, including the station name, latitude, and longitude. Extracts and stores metadata about starting stations.

**Steps Taken:**  
1. Extract unique start\_station names and their geographic coordinates (start\_lat, start\_lon) from the raw dataset.  
2. Deduplicate records using DISTINCT to ensure unique entries.  
3. Map these fields to station\_name, latitude, and longitude in the dim\_station table.

**SQL Script:**



**Challenges:** A leading space in the *start\_lon* column caused DbSchema to misinterpret the column name.  
**Solution:** Recreating the table on BigQuery with the correct name for *start\_lon* fixed the issue.

**2. Creating dim\_time**The dim\_time table is a dimension table that breaks down the trip\_date into smaller components, such as year, month, day, and day of the week.

**Steps Taken:**  
1. Extracted the date component from start\_time.  
2. Used SQL functions like *EXTRACT()* to break down start\_time into meaningful parts.  
3. Ensured unique rows with *DISTINCT*.

**SQL Script:**

A computer screen shot of text

Description automatically generated

**Challenges:** Initially included *EXTRACT(HOUR FROM start\_time)* but removed it after realizing the data didn’t include time components at the hour level.  
**Solution:** Simplified the table to focus on date-related components.

**3. Creating fact\_bike\_trips**The *fact\_bike\_trips* table is the primary fact table containing trip metrics such as duration and plan duration. It also includes *keys (station\_name, trip\_date)* to link to the dimension tables.

**Steps Taken:**

1. Selected key columns from the raw dataset.  
2. Renamed columns for consistency *(e.g., start\_station to station\_name).*  
3. Included foreign keys to link with *dim\_station* and *dim\_time*.

**SQL Script:**

A screen shot of a computer code

Description automatically generated

**Challenges:** None encountered during this step, as issues with columns (e.g., leading spaces) were already resolved while creating the dimension tables.

**ETL Issues Resolved**

**Leading Space in Column Names:**

Issue: The *start\_lon* column had a leading space, which caused errors in DbSchema.

Solution: Used backticks in SQL queries to handle the column correctly.

**Unnecessary Hour Extraction in dim\_time:**

Issue: Attempted to extract hour from start\_time, but the data didn’t include time components.

Solution: Removed *EXTRACT(HOUR)* from the script and focused on date components.

**Summary:** This structured ETL process ensured the data warehouse schema was optimized for querying and analysis. The challenges encountered were minimal and successfully resolved, paving the way for robust data exploration and visualization.

**Data Validation**

**Join Test Results:**

**1. fact\_bike\_trips and dim\_station  
Purpose:** To validate that the *station\_name* column in *fact\_bike\_trips* correctly links to the *station\_name* column in *dim\_station*. **Query Used:**

A screenshot of a computer code

Description automatically generated

**Results:** The query returned accurate mappings of trip data (*trip\_id*, *duration*) to station details *(latitude, longitude)*.  
**Validation:** Successful join, no null or mismatched values detected.

**2. fact\_bike\_trips and dim\_time  
Purpose:** To validate that the *trip\_date* column in *fact\_bike\_trips* correctly links to the *trip\_date* column in *dim\_time*. **Query Used:**

A computer code with text

Description automatically generated

**Results:** The query successfully linked trip data *(trip\_id, duration)* with time details *(year, month, day).* **Validation:** Successful join, no null or mismatched values detected.

**Insights and Analysis**

**Metrics and Queries:**

**1. Top 5 Busiest Stations**

**Query Used:**

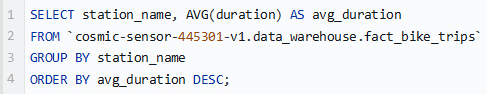
A close-up of a computer code

Description automatically generated

**Insight:** Station 4214, with 7,439 trips in Q3 2024, is the busiest starting point in the network. Its popularity likely stems from strategic location, accessibility, or proximity to high-demand areas. This makes it a key hub for optimizing bike availability, improving user experience, and guiding future station placement decisions.

**2. Average Trip Duration**

**Query Used:**



**Insight:** Trip durations per station varied between 9 and 110 minutes, with Station 4600 recording the longest average duration of 110 minutes. This suggests Station 4600 may serve longer-distance commuters or leisure riders, providing an opportunity to tailor services and pricing for these users.**3. Trips Over Time**

**Query Used:**

A computer code with black text

Description automatically generated

**Insight:** The highest daily trip count occurred on August 18th, 2024, with a total of 2,385 trips. This peak might suggest a significant event or high-demand day. Consistent activity levels were observed throughout Q3 2024, with slight fluctuations on weekends.  
  
**4. Total Trips Per Station**

**Query Used:**

A close-up of a computer code

Description automatically generated

**Insight:** The busiest station is Station 4214 with 7,439 trips. Top-performing stations highlight high-demand areas likely due to their proximity to busy hubs such as downtown areas or transit intersections. On the other hand, some stations exhibit much lower usage, suggesting possible over-allocation of resources or insufficient user engagement.

**Visualizations**

**1. Bar Chart: Top 5 Busiest Stations**

A graph of blue bars

Description automatically generated

**Description:** This bar chart displays the five stations with the highest trip counts during Q3 2024.

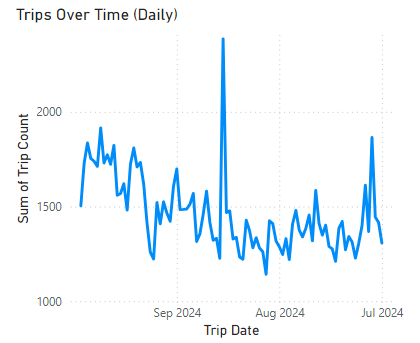
**Insights:** The chart highlights the top 5 busiest bike stations based on trip counts during Q3 2024. Station 4214 had the highest usage with 7,439 trips, significantly exceeding other stations, suggesting its strategic importance in the bike-sharing network. **Technical Notes:** Data aggregated using the station\_name dimension and Record Count metric, sorted in descending order.

**2. Average Trip Duration by Station**

A graph of blue bars

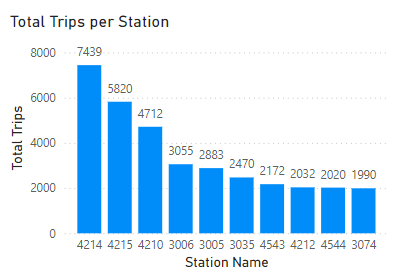
Description automatically generated

**Description:** This bar chart visualizes the average trip duration (in minutes) for each bike station during Q3 2024. The chart provides insights into station-specific trip patterns and highlights stations with unusually long or short average durations.  
  
**Insights:** The chart reveals that Station 4600 has the longest average trip duration of 110 minutes, followed by Stations 4529 and 4545 with 107 and 95 minutes, respectively. This suggests these stations might serve areas with longer commute requirements or higher recreational use. Stations with lower durations may indicate shorter, utilitarian trips. **Technical Notes:**   
**1. Data Source**: fact\_bike\_trips table from the data warehouse. **2. Metric:** Average duration (in minutes) grouped by station\_name. **3. Sorting:** Data is sorted in descending order of average trip duration to highlight stations with the highest values. **4. Filtering**: Full dataset of stations is included, with a horizontal scroll bar to accommodate all entries in the visual.  
  
**3. Trips Over Time (Daily)**



**Description:** This line chart displays the daily number of bike trips throughout Q3 2024, highlighting fluctuations in usage over time.  
  
**Insights:** The chart reveals a consistent pattern of daily bike usage with noticeable peaks and troughs. A significant spike in activity occurred on a specific day, potentially indicating an event or promotional activity. Lower activity on certain days could correspond to weekdays, holidays, or weather-related conditions. **Technical Notes:**   
**1. X-axis:** trip\_date (Daily intervals). **2. Y-axis:** Sum of trip\_count (Total trips per day). **3. Chart Type:** Line Chart, sorted chronologically by date. **4.** Data sourced from the fact\_bike\_trips table and aggregated using the query for daily trip counts.

**4. Trips Over Time (Daily)**



**Description:** This bar chart showcases the total number of trips initiated at each station during Q3 2024, emphasizing usage distribution across all stations.  
  
**Insights:** Station **4214** is the most utilized station with **7,439 trips**, followed by Station **4215** and Station **4210** with **5,820** and **4,712 trips**, respectively. A steep drop-off is observed after the top 3 stations, suggesting a significant disparity in station usage. Lower-performing stations, such as **4544** and **3074**, indicate areas with potential for infrastructure improvement or targeted marketing efforts. **Technical Notes:**   
**1. Data Source:** Google BigQuery; fact\_bike\_trips table.  
**2. Visualization Type:** Clustered Bar Chart. **3. Axis Configuration:**- X-axis: Station Name.  
- Y-axis: Total Trips (Count of trip\_id).

**Conclusion**

The **metro\_bike\_2024\_datawarehouse** project successfully demonstrated the power of data warehousing and visualization tools in analyzing large-scale datasets. Through a carefully designed schema and the use of tools like BigQuery, DbSchema, Lucid Charts, and Power BI, the project provided actionable insights into metro bike trip activity during Q3 2024, addressing the updated business and functional requirements effectively.  
  
**Key Achievements:**

**Data Organization:  
1.** The development of a robust star schema with a central fact table and two-dimensional tables ensured efficient storage, retrieval, and analysis of trip data, station metadata, and time details. **2.** The schema's relationships facilitated seamless querying, enabling detailed analysis of key metrics such as trip duration, trip frequency, and station utilization.

**Insightful Visualizations:  
1.** Visual tools such as bar charts and line charts provided clear and actionable insights into busiest stations, average trip durations, total trips per station, and trips over time. **2.** The visualizations aligned with the business requirements, highlighting key trends and patterns to support decision-making for optimizing station placement, understanding peak usage times, and improving pricing strategies.

**Resolved Challenges:  
1.** Addressed data inconsistencies, such as leading spaces in column names and unnecessary transformations in the ETL process, to ensure accurate analysis. **2.** Improved schema design by clearly defining relationships between tables and aligning the data model with business objectives.

**Key Findings:**

**Busiest Stations**: Station 4214 recorded 7,439 trips, establishing it as a significant hub. This station's high usage underscores its strategic location and importance in the bike-sharing network.

**Daily Usage Trends**: A peak daily trip volume of 2,385 occurred on August 18th, 2024, possibly correlating with an event or high-demand day. Seasonal fluctuations revealed consistent activity patterns, with weekends and specific dates exhibiting spikes.

**Trip Duration Variations**: Station 4600 had the longest average trip duration of 110 minutes, indicating potential use for longer commutes or recreational purposes. Conversely, other stations exhibited shorter durations, reflecting different usage patterns.

**Station Utilization**: The distribution of total trips per station identified underutilized stations, offering opportunities for targeted marketing or infrastructure improvements.

**Reflections and Future Work:**

This project emphasized the importance of meticulous data preparation, schema design, and validation in deriving meaningful insights. Each step in the ETL process and visualization design played a critical role in addressing the project requirements.

In the future, the following enhancements could be explored:  
  
**1. Incorporating Additional Data Sources**:  
Include weather data, local events, or demographic information to provide deeper contextual analysis and improve demand forecasting.  
  
**2. Automation and Real-Time Analysis**:  
Automate the ETL pipeline to allow real-time data processing and visualization.  
  
**3. Advanced Analytics**:  
Implement predictive modeling to forecast demand patterns and optimize resource allocation dynamically.  
  
The **metro\_bike\_2024\_datawarehouse** project effectively met its objectives, delivering valuable insights to optimize bike-sharing operations. The comprehensive data architecture and visualizations not only addressed the specified requirements but also laid a foundation for future data-driven enhancements. By leveraging advanced tools and methodologies, the project highlighted the critical role of data in driving informed decision-making for bike-sharing systems.