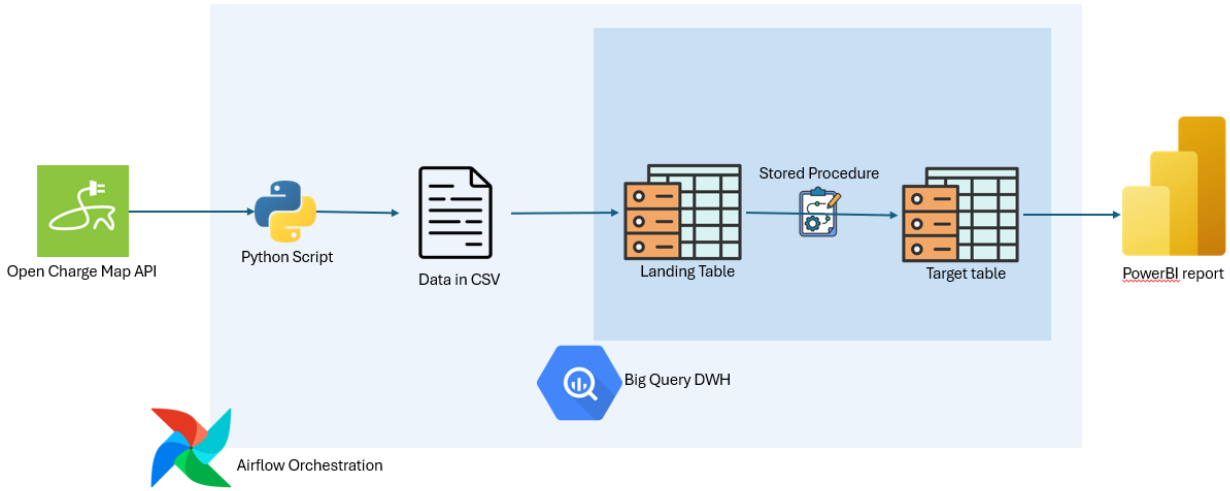
Data Sourcing

I have tried following datasets that are available.

1. National Charge point Registry (NCR) – This data set seems to be decommissioned.
2. electric Vehicle Charging Transactions from data.gov.uk – Seems to be a file submitted by the government.
3. Open Charge Map – Not the most up to date data source, but it has all the required data points along with an API that is distributed under the license Creative Commons Attribution 4.0 International (CC BY 4.0)
    a. This contains data from all over the world.
    b. Dataset get updated daily by the community.
    c. Quality of the data can vary as its community driven.

Decided to proceed with the Open Charge Map dataset.

Architecture



I built a centralized pipeline using Apache Airflow to manage the lifecycle of EV charging data. The workflow is divided into five distinct stages:

- Extract:
    - I created a Python-based task that fetches normalized EV charging data from an external API and stores it as a CSV in Google Cloud Storage (GCS).
    - I filter the Central London charging points based on the Latitude / longitude with a radius of 5KM.

- Loading:
    - I configured the pipeline to move raw data from GCS into a BigQuery Landing Table using a WRITE_TRUNCATE disposition to ensure each daily run starts with a fresh dataset.
- Data Quality Checks:
    - Before any transformation occurs, I integrated validations steps to check for duplicates and other data issues.
- Transformation:
    - I developed a BigQuery Stored Procedure that executes a MERGE statement to synchronize the landing data with the historical target table.
- Archiving:
    - I added a final step to move the source CSV to an archive folder in GCS with a datestamp (e.g., ev_charging_data_2026-01-01.csv) for future auditing.

Transformation Logic

Instead of simple overwrites, I implemented sophisticated MERGE logic within a Stored Procedure to handle Change Data Capture (CDC).

- Deduplication:
    - I used ROW_NUMBER() to select the latest status for each unique connection ID in cases where the source contains duplicates.
- Upsert Logic:
    - I programmed the procedure to insert new charging points while updating existing records with the latest power ratings, operational status, and costs.
- Derived Columns:
    - I included logic to automatically categorize chargers (e.g., "Rapid", "Ultra-Rapid") based on power_kw. (less than 3.7 = slow, less than 22 = Fast etc.)
    - is_free_to_use column flags true when the UsageCost contains the substring “free” or “0”.
- Soft Deletes:
    - I implemented a source_deleted_flag to identify charging points removed from the live API feed, allowing me to preserve historical records rather than deleting them.

Data Validation

To prevent "garbage-in, garbage-out" scenarios, I integrated two levels of validation.

- Landing Table Checks:

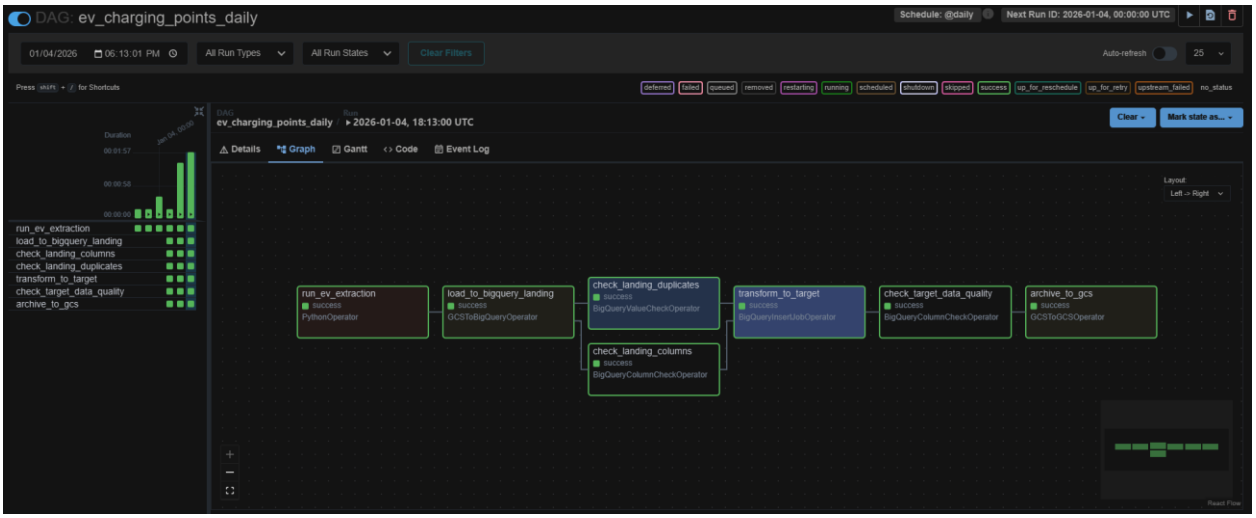
    - I set up checks to ensure critical identifiers (ID, conn_ID) are not null and that coordinates fall within a specific "Geofence" for Central London.

- Duplicate Detection:
    - I added a custom check to validate that the conn_ID is unique across the incoming batch, halting the pipeline if duplicates are found to protect the target table.
- Target Health:
    - I configured a post-transformation step to verify the target table remains healthy, checking that power values are non-negative and station IDs remain intact.

Technical Specifications

- Database Schema:
    - I designed the Target Table to be partitioned by inserted_at for cost-efficient historical queries and clustered by postcode_district and status to optimize geospatial search performance.
- Modern SQL Standards:
    - I explicitly configured all BigQuery operations to use Standard SQL (use_legacy_sql=False) to support complex table paths with hyphens and advanced analytic functions.

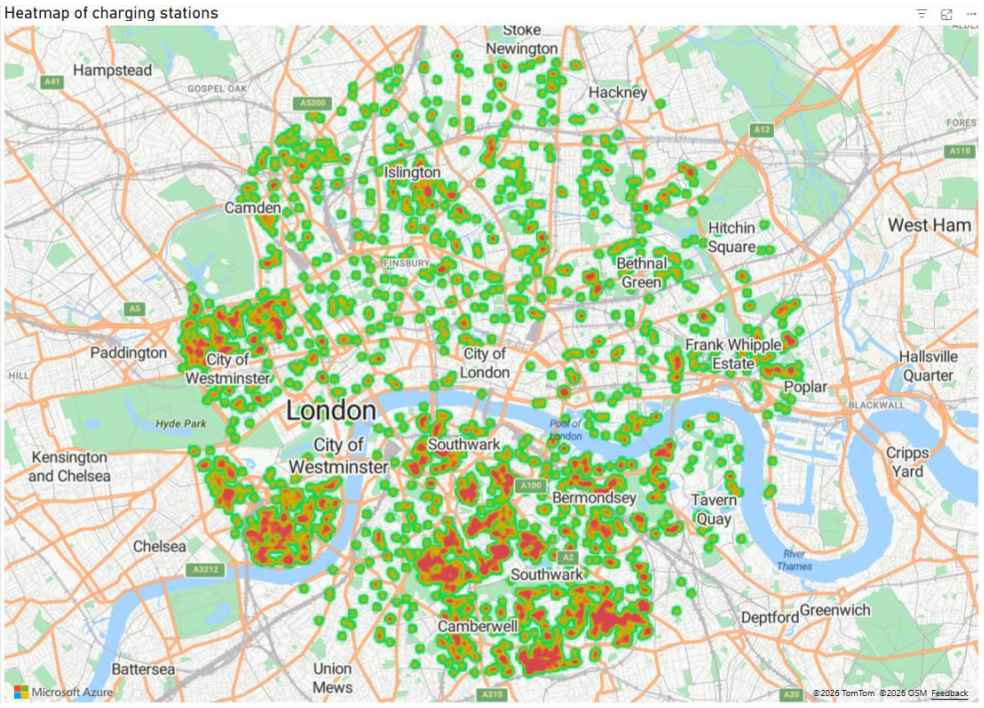
Dag Run



Key findings from the current data

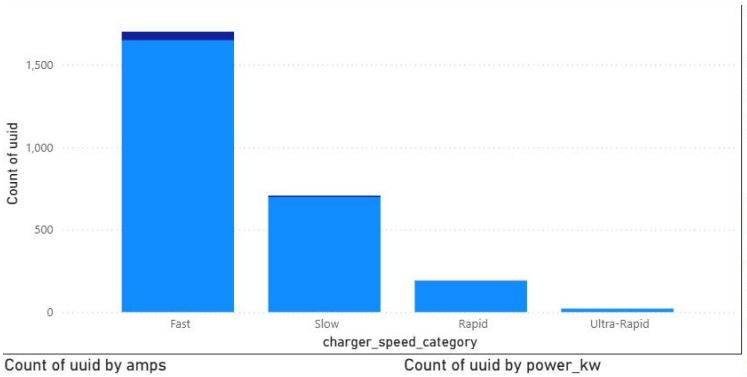
Overall scale

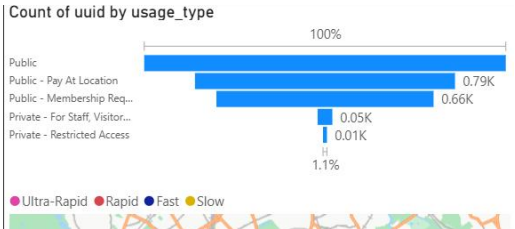
Charging Stations	2,141
Charging Connections	2618
Average connections per station	1.22
Operational	2099
Free to use	23
Operational	97%

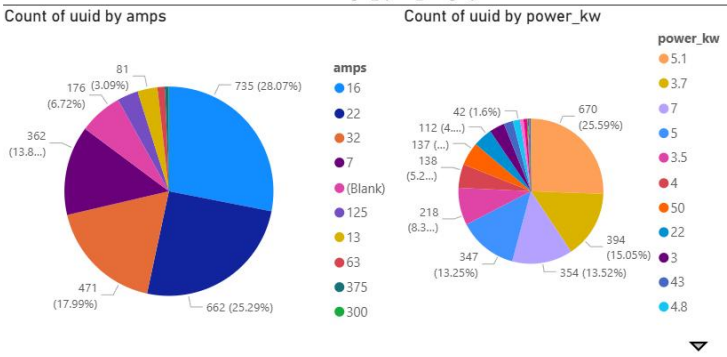


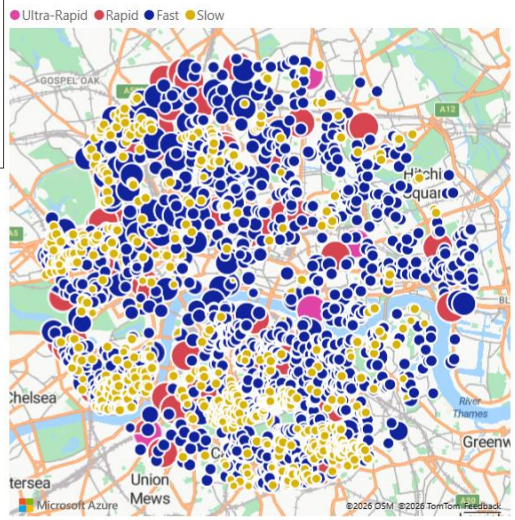
1. Average connections per station is 1.22 which means most stations have only one connector, indicating limited redundancy per location.
2. Operational chargers 97%. This suggests excellent availability and maintenance across Central London.
3. Only 23 free to use chargers. Nearly all chargers are paid, reinforcing the importance of transparent pricing and payment accessibility.

Charging drilldown

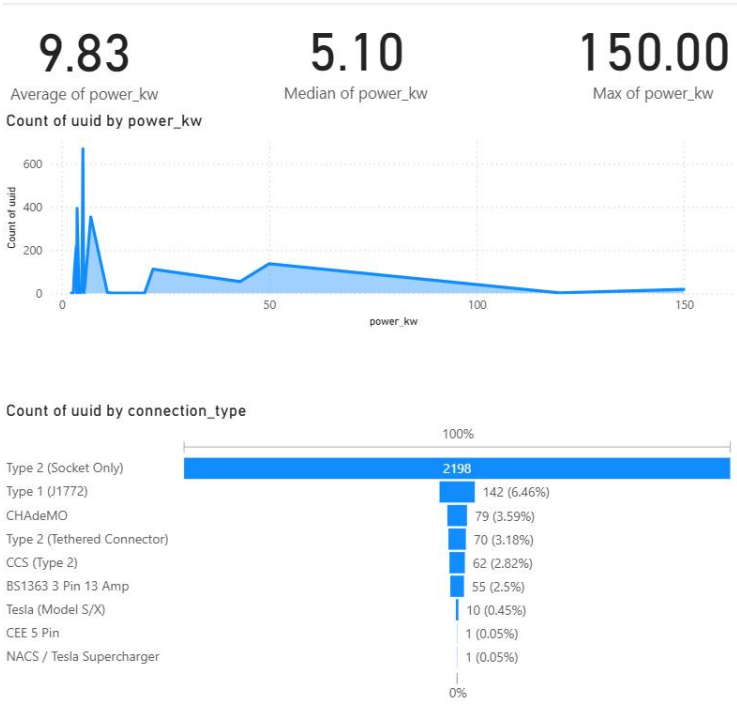


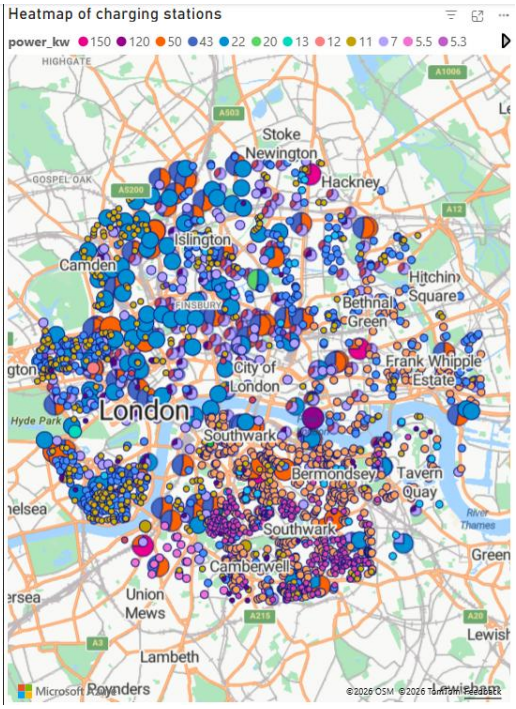






1. The network is heavily skewed towards Fast chargers.
2. Ultra-rapid chargers are extremely limited, which may constrain long-distance or high-throughput charging demand.





1. Most chargers are designed for longer dwell times (workplaces, streets, parking).
2. High-power chargers exist but are **rare and concentrated.**
3. **Type 2 (Socket Only)** – dominant, Type 1 (J1772) – still present and **3-pin domestic sockets** – still used, but low power
4. Infrastructure is largely aligned with **modern European EV standards**, but legacy connectors remain.

High Level conclusions

1. Central London has excellent charger availability and uptime.
2. Infrastructure prioritizes coverage over ultra-fast charging.
3. Ultra-rapid charging is a clear gap, especially for taxis, ride-hailing, and high-mileage users.
4. Chargers are mostly paid, indicating a mature commercial charging ecosystem.
5. Power distribution suggests overnight / workplace / destination charging is the dominant use case.