# Project Overview

In order to track growth in the burgeoning New York City energy storage system (ESS) market, we will be utilizing project and interconnection queue data from public utilities and state agencies, from which key market insights can be gained.

For the sake of this particular project, we are asked to extract, transform, and load (ETL) several datasets and perform any necessary transformations (cleaning, joining, filtering, aggregating, etc.) to produce a single “clean” dataset. This dataset will then be loaded into a final production database, which can be queried for any subsequent analyses desired.

The ETL process makes up a crucial part of the *Analytics Paradigm* employed by data analysts, shown below:

A screenshot of a cell phone

Description automatically generated

Figure 1 - Analytics Paradigm

# Data Sources

We will be utilizing data sourced from the following publicly available resources:

1. **Con Edison Interconnection Queue Data**

* **Description:** list of all Con Edison interconnection applications, including energy storage systems located in NYISO Zone J (NYC).
* **Source:** <http://www3.dps.ny.gov/W/PSCWeb.nsf/All/286D2C179E9A5A8385257FBF003F1F7E>
* **Shape:** (34759, 31)
* **Filename:** *‘data/csv/con\_ed\_public\_january\_2020.csv’*

1. **NYISO Interconnection Queue Data**

* **Description:** list of all NYISO interconnection applications, including energy storage systems located in NYISO Zone J (NYC).
* **Source:** <http://www3.dps.ny.gov/W/PSCWeb.nsf/All/286D2C179E9A5A8385257FBF003F1F7E>
* **Shape:** (891, 19)
* **Filename:** *‘data/csv/nyiso\_interconnection\_queue.csv’*

1. **NYSERDA DER Integrated Data System – Facilities List**

* **Description:** list of all NYSERDA-funded energy storage system facilities, including projects located in NYISO Zone J (NYC).
* **Source:** <https://der.nyserda.ny.gov/>
* **Shape:** (1213, 38)
* **Filename:** *‘data/csv/nyserda\_der\_metric\_data\_facilities.csv’*

1. **NYSERDA DER Integrated Data System – Projects List**

* **Description:** list of all NYSERDA-funded energy storage system projects, including projects located in NYISO Zone J (NYC).
* **Source:** <https://der.nyserda.ny.gov/>
* **Shape:** (1174, 35)
* **Filename:** *‘data/csv/nyserda\_der\_metric\_data\_projects.csv’*

# ETL Process

The flow of processes beginning at extraction through loading into the final production database is shown below:

A screenshot of a cell phone

Description automatically generated

Figure 2 - ETL Process Diagram

# Step 1. Extraction

Because the data from each organization is only provided in the form of Excel files – generally updated quarterly and non-automated, it was easiest to simply download the files, convert to CSV format, and load into Pandas dataframes to clean.

# Step 2. Transformation

To get into useable format to load into the final database, the following operations were performed for each dataframe:

**coned\_df:**

1. *Keep only non-null ESS projects (does not include thermal/ice storage)*
2. *Remove unwanted columns*
3. *Rename columns*
4. *Drop projects in Westchester*
5. *Rename boroughs*
6. *Convert zipcode and power to integers*
7. *Convert dates to datetime*
8. *Create new dataframe for completed projects*
9. *Export clean CSVs*

**nyiso\_df:**

1. *Keep only wanted columns*
2. *Rename columns*
3. *Keep only ESS projects in Zone J*
4. *Rename boroughs*
5. *Drop if Borough = Union*
6. *Convert dates to datetime*
7. *Export clean CSV*

**nyserda\_projects\_df:**

1. *Keep only wanted columns*
2. *Rename columns*
3. *Change values to integers*
4. *Replace 0s with NaN*
5. *Keep only projects in Zone J and remove thermal storage projects*
6. *Convert dates to datetime*
7. *Export to CSV*

**nyserda\_facilities\_df:**

1. *Keep only wanted columns*
2. *Rename columns*
3. *Replace 0s with NaN*
4. *Keep only projects in Zone J and remove thermal storage projects*
5. *Concatenate coordinates or mapping*
6. *Export to CSV*

NYSERDA Projects and Facilities datasets can be merged on ‘facility\_id’ in Pandas or later using SQL commands.

# Step 3. Load

All datasets are now loaded into a relational database ‘nyc\_ess\_projects.db’ with a Python script ‘create\_database.ipynb’, which uses using Pandas, SQLite3 and SQLAlchemy modules.

The entity relationship diagram is seen below:

A close up of a piece of paper

Description automatically generated

Figure 3 - ERD Diagram

Finally, viewing the database in DB Browser for SQLite:

A screenshot of a social media post

Description automatically generated

Figure 4 - Database in DB Browser

A screenshot of a social media post

Description automatically generated

Figure 5 - Database in DB Browser