12/16/2020

MISSION SPACE-X

ETL PROJECT

Members:

Tejal Kotkar

Hibo Dahir

Savita Hirilall

Phanit Tameerug

Data analysis and management requires data integration. These data may come in different formats and sources such as csv files, json, files, or html tables. Integration of various datasets is a key for data analysis. In other words, identifying the dataset, extracting and reading the data, cleaning and structuring it in the desired format, and loading the clean data into a database for storage is imperative for analysis and management. Integrating this which is an important part of working with data.

For this ETL project we have used Space-X data (2006-2020) from Kaggle & space-x API.

**Introduction:**

SpaceX designs manufacture and launch advanced rockets and spacecraft and is a privately owned company founded by Elon Musk in 2002. SpaceX has gained worldwide attention for a series of historic milestones by revolutionizing space transportation, with the goal of making life multiplanetary.

**Data Source:**

**Kaggle**: <https://www.kaggle.com/rohanrao/rspacex-data?select=launches.csv>

**API**: [https://documenter.getpostman.com/view/2025350/RWaEzAiG#e232e64a-58a2-4bc0-af42-eb20499425cc](https://documenter.getpostman.com/view/2025350/RWaEzAiG#00ac651a-8ba2-4b4c-858a-4034dd1254fa)

**Process:**

* Extract :

Extracted data as CSV files from Kaggle. Below are the file names and their respective API’s are used to get the data not available in CSV files.

* Capsules.csv:
  + API - <https://api.spacexdata.com/v3/capsules>

A capsule is a pressurized section, allows for the transport of people as well as environmentally sensitive cargo.

* Payloads.csv:
  + API - https://api.spacexdata.com/v3/payloads

The payload is the carrying capacity of an aircraft or launch vehicle, usually measured in

terms of weight. The payload of a vehicle may include cargo, passengers, flight crew.

* Launches.csv & launchpads.csv:
  + API : <https://api.spacexdata.com/v3/launches> & https://api.spacexdata.com/v3/launchpads

A platform from which rockets or spacecraft are launched.

* Rocket.csv :
  + API : https://api.spacexdata.com/v3/rockets

A rocket is a missile, spacecraft, aircraft or other vehicle that obtains thrust from a rocket engine.

* Transform:

Below process was followed to transform the data for all csv’s.

* Loaded the csv files (downloaded from Kaggle) into dataframes.
* Inspected dataframes and noted the data (column structure) that was missing and available in API.
* Retrieved additional data from API for each respective dataframe.
* Added new columns to Data Frame and loaded a data from API to those columns.
* Kept only required columns.
* Renamed & Rearranged columns.
* Cleaned any missing values.
* Wrote DF to CSV files.
* Load:

Used relational database PostgreSQL to load the final production database.

Below process was followed.

* After transformation step, 5 DF’s were generated each corresponding to respective csv file.
* Created another 2 DF’s to act as connection between these DF’s by reading data from csv & API both
  + Bridge – Holds the ids from Launches, Rocket, Launchpads & Payloads.
  + Launch\_cap – Holds ids from Launches & Capsules.
* Wrote additional DF’s to csv files.
* Read all cleaned csv files to DF’s.
* For Bridge DF, checked if all ids are present in their respective primary DF’s and created another DF with that.
* Created database Mission\_SpaceX with the help of interface & schema with the SQL queries in pgAdmin4.
* Using SQL Alchemy loaded all the DF’s to tables below:
  + Capsule
  + Rocket
  + Payload
  + Launches
  + Launchpad
  + Launch\_cap
  + Bridge
* Retrieved the loaded data from above created SQL tables for verification.

**Potential analysis**

On these datasets we created could be inquiries such as:

* Number of successful & failed launches.
* Are there any launches failed because of same reason & how many?
* Relationship between Cost of launch and success rate of mission.