Summary of DBT Pokemon project

### Project Overview

For this project, I will integrate Pokémon data from Generations 1 through 7.

### Tools

* Project Management: GitHub
* Data Pipeline: DBT Cloud
* Data Storage: BigQuery
* Source: CSV file (Kaggle)

### Setup

GitHub

* Create a new GitHub project named dbtproject\_pokemon.
* Add this repository to the list of repositories available for DBT Cloud.

BigQuery

* Create a new BigQuery project called dbtpokemon.
* Set up a BigQuery service account named dbtproject, with roles: BigQuery Job User and Data Editor.
* Generate a JSON key for the service account.
* Enable the BigQuery API.

DBT Cloud

* Upload the JSON key to DBT Cloud for connection.
* Designate the development dataset as dbtprojectpokemon\_sl.
* Link the DBT project to the GitHub repository: dbtproject\_pokemon.
* Initialize the project using the "initialize" button and commit/sync changes.

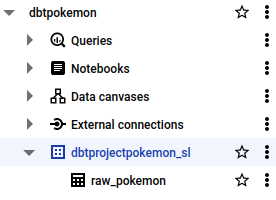
### Data Import

Source

* The source data is a CSV file from Kaggle: [kaggle.com/datasets/rounakbanik/pokemon](https://www.kaggle.com/datasets/rounakbanik/pokemon).
* Add the raw CSV file to the seeds folder in the GitHub repo, rename it to raw\_pokemon.csv, and commit the changes.
* In DBT Cloud, pull from the remote repository and create a new branch named sl\_dev.
* Verify that the source file appears in the seeds folder.
* Run dbt seed to import the data into the dataset.

Data Exploration

* Access the data in the table dbtprojectpokemon\_sl.raw\_pokemon.
* Perform a SELECT query to identify useful data and determine relevant dimensions



Data Exploration

* I execute the following select statement: SELECT \* FROM (SELECT COUNT(\*) as C FROM `dbtpokemon.dbtprojectpokemon\_sl.raw\_pokemon` GROUP BY pokedex\_number) WHERE C > 1
* The result of 0 columns indicates that pokedex\_number is the primary key of the raw dataset.
* Document the columns available in the raw dataset.
* Select and retain only the columns needed for analysis.
* Normalize the data to an extent suitable for analytical purposes.
* Refer to the tabledesign.txt for a draft of the database schema.

### DBT Project

Project

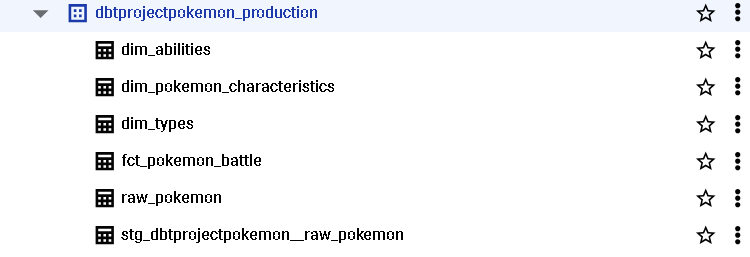
* The project is available in the GitHub repository.

Deployment

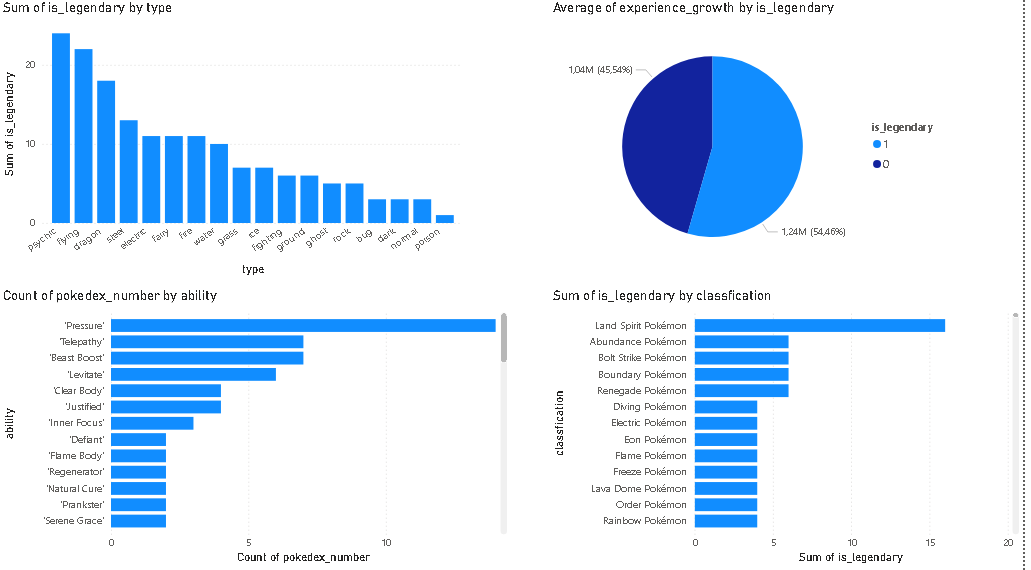
* In DBT Cloud, navigate to Deploy → Environments.
* Create a production environment named dbtprojectpokemon\_production with the dataset dbtprojectpokemon\_production.
* In DBT Cloud, go to Deploy → Jobs.
* Set up a job called Deploy dbtprojectpokemon\_production to run dbt build --full-refresh using the newly created environment.
* For testing, run the job manually using the “Run Now” option instead of scheduling it.

**Finalisation**

Schema in BigQuery

****

**Small analysis of Legendary Pokemon using MS Power Bi**

****