Extract Transform Load Analyze

Raj Shah | Joe Kunz | Jorge Melchor | Luis Gonzalez

ETL PROCESS

SOURCE SYSTEMS

- Kaggle Datasets
- GitHub

EXTRACT, TRANSFORM, LOAD

- Extract: CSV
- Transform: Select & Rename columns, Joins
- Load to RDBMS

DATA WAREHOUSE

- ERD Model
- SQL Schema
- PostgreSQL Database

Extract

- Import CSV
- Create Dataframe from imported CSV: types_df, moves_df, pokemon_df

	id	identifier	generation_id	damage_class_id
0	1	normal	1	2.0
1	2	fighting	1	2.0
2	3	flying	1	2.0
3	4	poison	1	2.0
4	5	ground	1	2.0

Types

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False

Pokémon

	id	identifier	generation_id	type_id	power	pp	accuracy	priority	target_id	damage_class_id	effect_id	effect_chance	contest_type_id	contest_eff
0	1	pound	1	1	40.0	35.0	100.0	0	10	2	1	NaN	5.0	
1	2	karate-chop	1	2	50.0	25.0	100.0	0	10	2	44	NaN	5.0	
2	3	double-slap	1	1	15.0	10.0	85.0	0	10	2	30	NaN	5.0	
3	4	comet-punch	1	1	18.0	15.0	85.0	0	10	2	30	NaN	5.0	
4	5	mega-punch	1	1	80.0	20.0	85.0	0	10	2	1	NaN	5.0	

Moves

Transform

- Create new transformed Dataframe: types_transformed, moves_transformed, pokemon_transformed
- Filter dataframe with select columns
- Rename columns
- Create new index column
- Capitalize first string letter in Types column

id	TID	Type
1	1	Normal
2	2	Fighting
3	3	Flying
4	4	Poison
5	5	Ground

MID	Moves	TID
1	Pound	1
2	Karate-Chop	2
3	Double-Slap	1
4	Comet-Punch	1
5	Mega-Punch	1

id	Pokemon_Name	Type	Attack	Defense	Speed
1	Bulbasaur	Grass	49	49	45
2	Ivysaur	Grass	62	63	60
3	Venusaur	Grass	82	83	80
4	VenusaurMega Venusaur	Grass	100	123	80
5	Charmander	Fire	52	43	65

Transform

Merge Pokémon & Types Transformed Table

- Outer Merge on Type column from types_transformed_df & pokemon_transformed_df
- Set Pokemon_Name as index

id	TID	Type
1	1	Normal
2	2	Fighting
3	3	Flying
4	4	Poison
5	5	Ground

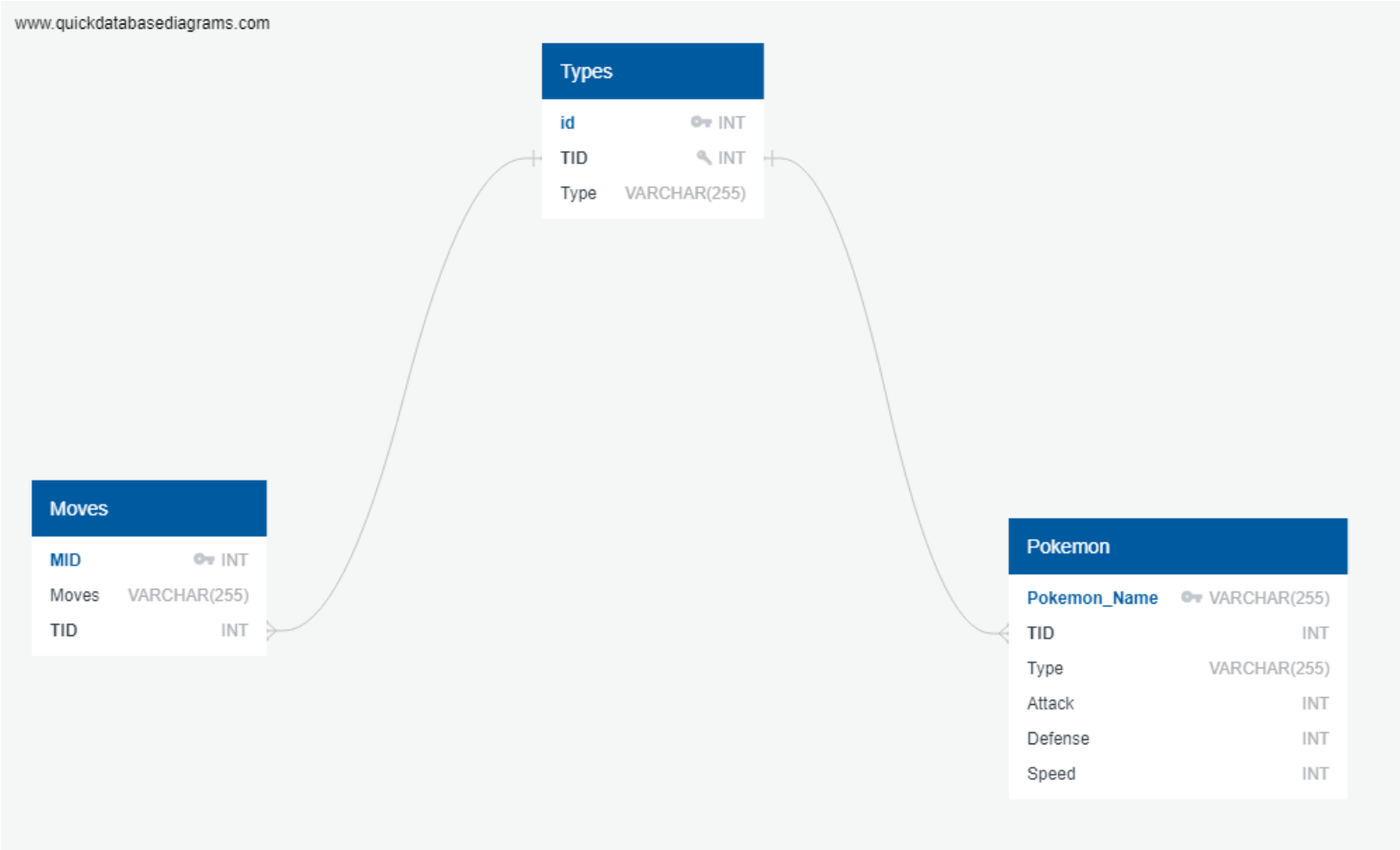
id	Pokemon_Name	Type	Attack	Defense	Speed
1	Bulbasaur	Grass	49	49	45
2	Ivysaur	Grass	62	63	60
3	Venusaur	Grass	82	83	80
4	VenusaurMega Venusaur	Grass	100	123	80
5	Charmander	Fire	52	43	65

Pokemon_Name	TID	Type	Attack	Defense	Speed
Bulbasaur	12	Grass	49.0	49.0	45.0
Ivysaur	12	Grass	62.0	63.0	60.0
Venusaur	12	Grass	82.0	83.0	80.0
VenusaurMega Venusaur	12	Grass	100.0	123.0	80.0
Oddish	12	Grass	50.0	55.0	30.0

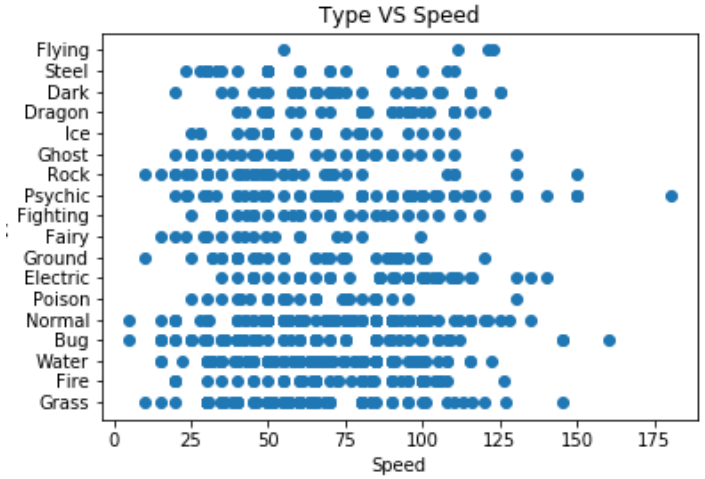
Load

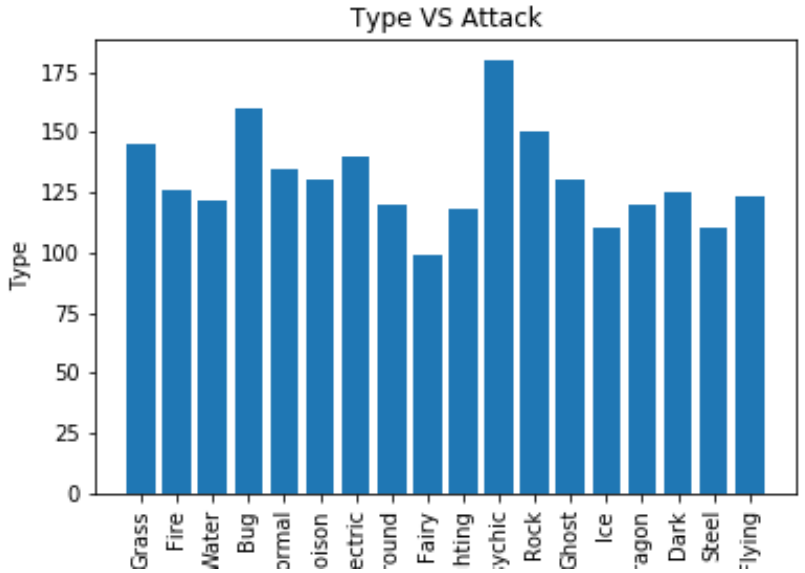
- Create Pokemon_DB in PostgreSQL
- Import schema to database
- Create connection string to PostgreSQL DB : `<username>:<password>@localhost:5432/pokemon_db`
- Load DataFrame to pokemon_db

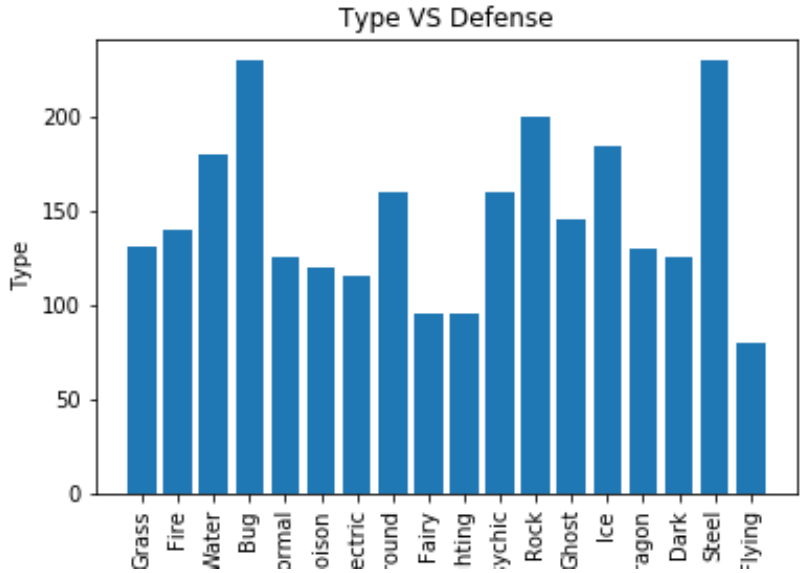
```
ETL-Project

Types
-
id PK INT
TID INT UNIQUE
Type  VARCHAR(255)

Moves
-
MID INT PK
Moves VARCHAR(255)
TID INT FK >- Types.TID

Pokemon
-
Pokemon_Name VARCHAR(255) PK
TID INT FK >- Types.TID
Type VARCHAR(255)
Attack INT
Defense INT
Speed INT
```



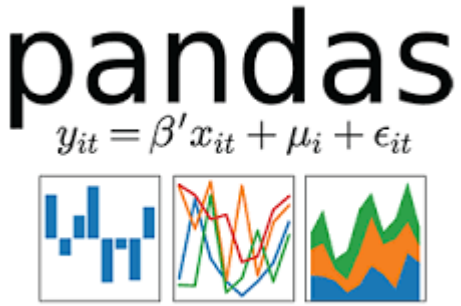
Project Analysis

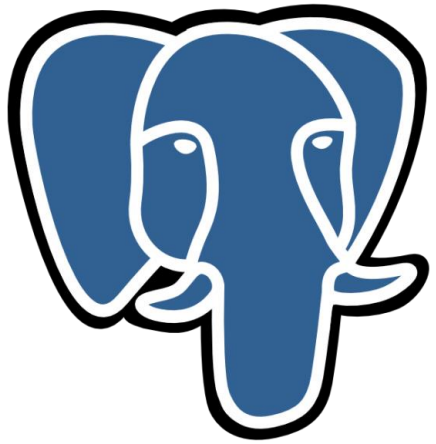


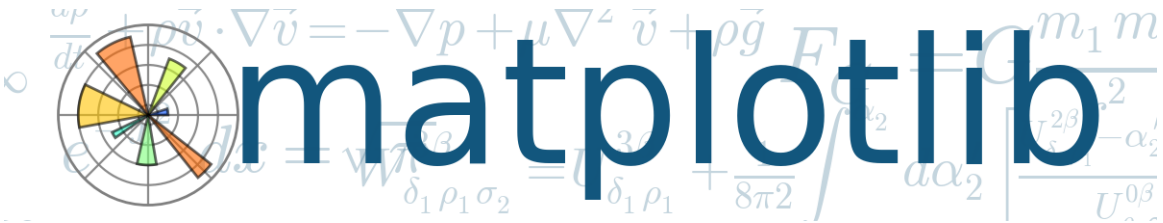




Packages & RDBMS







PostgreSQL

Data Sources

Pokémon List:
https://www.kaggle.com/abcsds/pokemon

Types Table:
https://github.com/veekun/pokedex/blob/master/pokedex/data/csv/types.csv

Moves List:
https://github.com/veekun/pokedex/blob/master/pokedex/data/csv/moves.csv

Thank You