Spotify Data Pipeline

Jason Kim

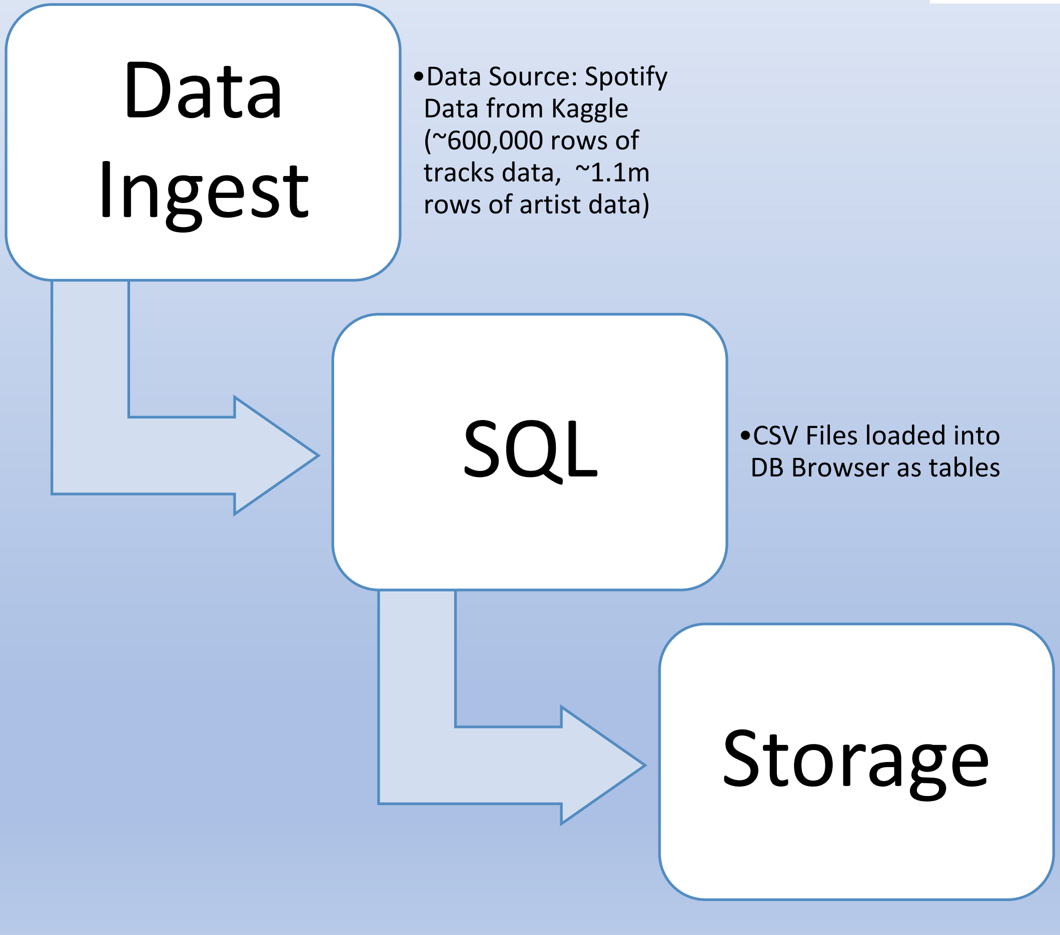
Metis

30 April 2021

Popularity

Song popularity is a metric that artists may seek to maximize

Data Ingest and Storage



Data Ingest

- Data Source: Spotify Data from Kaggle (~600,000 rows of tracks data, ~1.1m rows of artist data)

SQL

- CSV Files loaded into DB Browser as tables

Storage

- Tables of data stored into SQL database

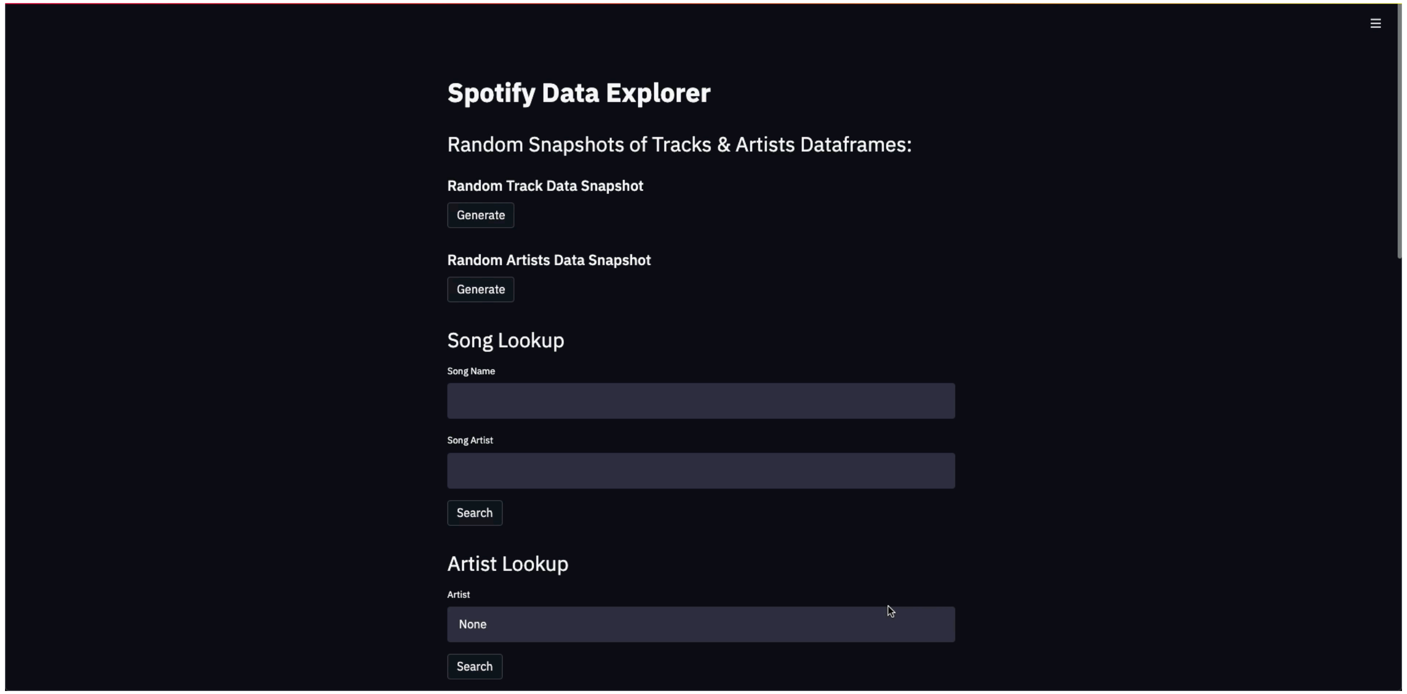
Processing

Processed data using pandas

Deployment

- Deployed Streamlit app:
  - Provides access to specific track and artist data
- Linear Regression model with SKLearn to predict Popularity of songs based on several features
- Visualizations with Seaborn

Spotify Data Explorer Demo



Spotify Data Explorer

Random Snapshots of Tracks & Artists Dataframes:

Random Track Data Snapshot

Random Artists Data Snapshot

Song Lookup

Song Name

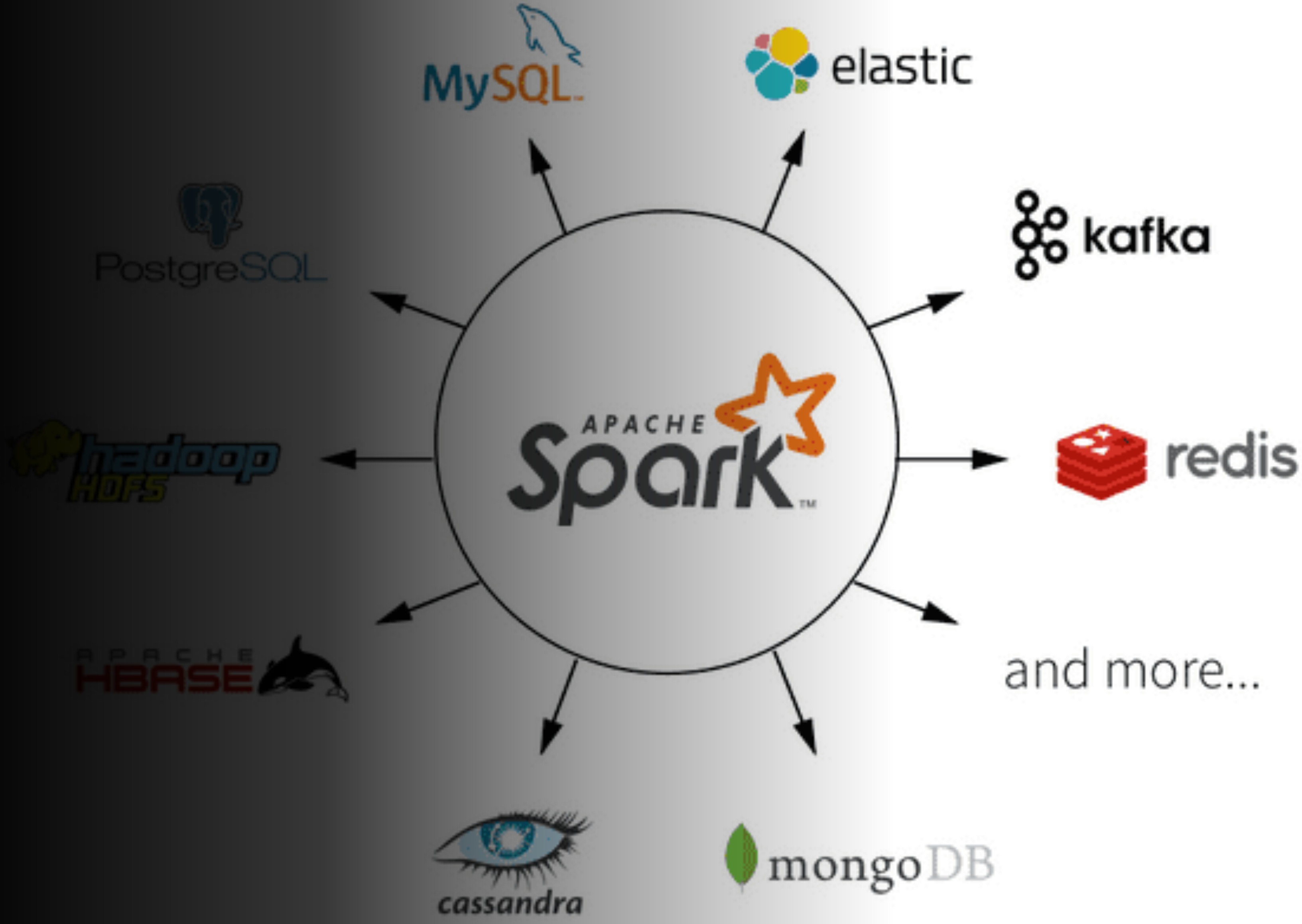
Song Artist

Artist Lookup

Artist: None

Future Work

- More accurate prediction model with further feature engineering with Spark ML
- Create access to further visualizations on Streamlit/Flask App
- Integration with Spotify API and further utilization of Spark


MySQL
elastic
PostgreSQL
kafka
APACHE Spark
hadoop HDFS
redis
APACHE HBASE
and more...
cassandra
mongoDB