RecomMuse

YOUR SOUNDTRACK TO THE PAST, PRESENTS THE FUTURE

Introduction

RecomMuse is a data-driven song recommendation system that utilize Hadoop, Spark, Drill to analyze million song dataset. Specifically recommend songs that make you feel nostalgic!

Data Processing

We use a pure Java HDF5 Library (**JHDF**) to extract H5 data, **Snappy compression** to squeeze them smaller, and lean Avro schema to store relevant features. This leads to a **99.4%** storage reduction!

MapReduce vs Spark

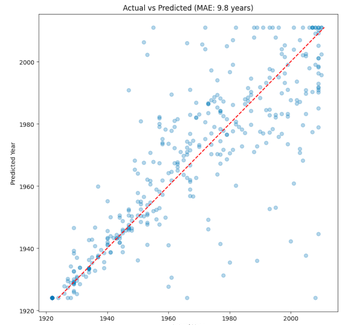
Iteratively Refining Graph State

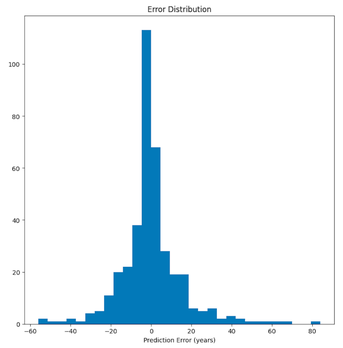
- Each job provides updated artist states for the next BFS level.
- The final output provides all data necessary to reconstruct the shortest path.

Getting **+20% boost** when using spark

Year Prediction

Using **only timbre** features, our model predicts a song's release year with an average error of just **9.82 years (MAE)** and an **RMSE of 15.38** —proving that musical timbre alone carries a surprising temporal signature!





Results of Linear Regression with different Principal Components (k)

$\mathbf{R^2}$ dropped: $0.26(k = 52) \rightarrow 0.21(k = 32) \rightarrow 0.14(k = 10)$

RMSE increased: $9.37 \rightarrow 9.73 \rightarrow 10.14$

SGD Result $k = 32$: R^2 0.21 $\rightarrow$ indicates reached limitations of LR

Drill Queries

You can easily interact with the data through Drill **SQL database query**! Let's find out some facts together!

- The age of the youngest and oldest song is **14** and **103** years respectively.
- The album with the maximum number of songs in it is **Greatest Hits** with **2014** tracks!
- The artist name of the longest song is **Mystic Revelation of Rastafari** with a song duration of **almost an hour**!

Kantaphat Leelakunwet, Choo Lee Wen, Arsen Aghayan, Ibrahim Daurenov