Million Song Dataset (MSD) Analysis Tools

VE472 22SU p1-team01

UM-SJTU Joint Institute

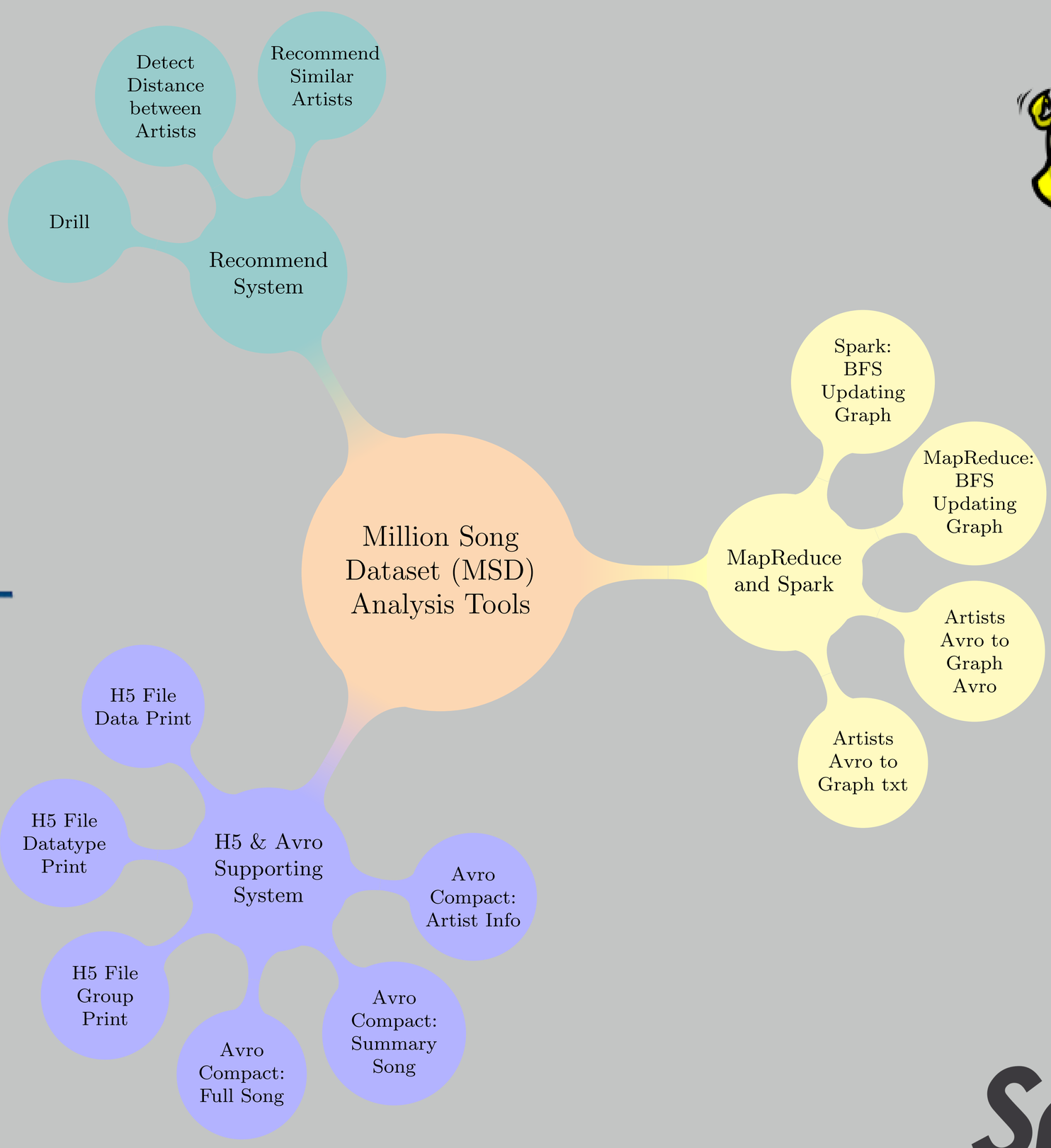
Overview

Our analysis tool provides you with recommendations based on Million Song Dataset (MSD), a dataset containing one million popular soundtracks by more than 40,000 artists. We also provides toolkits to **parse, analyze and interact** with the H5 file stored in the dataset on Hadoop.
Our toolkit is **cross-platform, flexible in terms of user-choice and easy to extend**.

Methodology

For the processing of MSD data, we construct a maven managed java project with jhdf package to read h5 and compact them into different Avro files.
For basic analysis, we use Drill with basic SQL to retrieve certain information from compacted Avro files.
For the recommendation system, we first use Java to generate graphs in both txt and Avro format and then use Python to carry out BFS in the map in both MapReduce and Spark ways. Then, we have a Python driver program to carry out the interactive job.

Feature Maps



System Requirements

Java 8 and Python3 Supported
Hadoop 3, Spark & Drill installed
Better on Linux, but Windows OK

Running Process

Transpose h5 file to Avro file
Transpose Avro file to Graph
Carry out BFS and update Graph

Developers

H5 & Avro: Kaiwen Zhang
MR& Spark: Yuxuan Zheng & Yuxiang Zhou
Recommendation System: Haoran Jin