Big Data Analysis on Million Song Dataset (MSD)

ECE4710: Methods and Tools for Big Data

Yiding Chang   Yifan Shen   Kexuan Huang   Qinhang Wu

July 5, 2022

Overview

- Milestone 0: HDF5 Data Pre-process
- Milestone 1: Drill Database Query
- Milestone 2: Advanced Data Analysis

Section 1

Milestone 0: HDF5 Data Pre-process

Goals

1. Compact small hdf5 files into larger one
2. Read hdf5 file and extract the information
3. Convert hdf5 to Avro with Apache Avro

0. Environment

- Python: 3.9.2
- Install PyGreSQL env

```
$ sudo apt-get install libpq-dev
```

- Install python packages

```
$ python3 -m pip install -r requirements.txt
```

- Remark: some syntax are modified to accommodate Python 3, for example,

```
print hdf5_path
```

1. Compact small hdf5 files into larger one

Creates an aggregate file from all song hdf5 files in a given directory

- Usage:

  ```
  $ python3 create_aggregate_file.py ↩
    ↪ <H5 DIR> <OUTPUT.h5>
  ```

- Input: a directory contains hdf5 song files
- Output: an aggregate hdf5 song file
- Remark: Remove the existing file having the same name as the output file before running the Python script

  ```
  $ rm -f <OUTPUT.h5>
  ```

2. Read `hdf5` files and extract the information

Quickly display all we know about a single/aggregate/summary `hdf5` song file

- Usage:

```
$ python3 display_song.py [FLAGS] ←
  → <HDF5 file> <OPT: song idx> <OPT: getter>
```

- Input: a `hdf5` song file
- Output: specified field content
- Remark: `getter` arguments must correspond to getters in `hdf5_getters.py`. Please refer to this file:

  `m0/schema/valid_field.log`

3. Convert hdf5 to Avro with Apache Avro

Convert a single/aggregate song file from hdf5 to Avro

- Usage:

```
$ hdf5_to_avro.py [-h] -s <SCHEMA> -i <HDF5> -o <AVRO>
```

- Input: an Avro schema file, a hdf5 song file to be converted
- Output: an Avro song file
- Remark: field names in schema file must correspond to getters in hdf5_getters.py. A sanity check will be performed before conversion. Please refer to this file:

  m0/schema/valid_field.log

4. Run hdf5 file pre-process pipeline

- Usage:

```
$ chmod +x m0.sh
$ ./m0.sh
```

Reference

1. MSongsDB https://github.com/tbertinmahieux/MSongsDB
2. MSongsDB Field List http://millionsongdataset.com/pages/field-list/
3. Apache Avro Documentation https://avro.apache.org/docs/current/index.html

Section 2

Milestone 1: Drill Database Query

Section 3

Milestone 2: Advanced Data Analysis