CS5344Big Data Analytics Technology

Lab 1 (AY2019/2020 Semester 2)

I. Requirement

Find the top 10 products based on the number of user reviews getting from reviews_Musical_Instruments_5.json.gz and metadata.json.gz

II. Environment Configuration

It uses Ubuntu-16.04 pre-configuration which is downloaded from *https://nusu-my.sharepoint.com/:u:/g/personal/e0267909_u_nus_edu/EZa_73QIfdlCrivb1NyqiBQBcikJWnlz71giJeIESDZLXQ*

- Python version 2.7.12
- Spark version 2.2.1

III. Execution

- Run command: **spark-submit Lab_1.py reviews_Musical_Instruments_5.json.gz metadata.json.gz output**
- After the program running is completed, the result will be available in **output** folder

IV. Code Introduction

- Imported packages: gzip, json, ast , sys, pyspark
- Code Logic

Step 1: Create a pair RDD contains all pairs of Product ID key/Number of unique Review ID value

- Create a RDD from reading reviews_Musical_Instruments_5.json.gz
- Use map() to create a pair RDD (k,v) : k is the Product ID and v is a set which only contains one Review ID element
- Use reduceByKey() to union values for same Product ID key. Because data type of value is a set of Review IDs, so a new pair RDD which is

created, contains pairs of Product ID key and set of unique Review ID value

- Use mapValues() to iterate all value of pair RDD and get length of value . A new pair RDD is created, contains pairs of Product ID key/Number of unique Review ID value

Step 2: Create a pair RDD consist of key/value-array pairs which have key is Product ID, value is a Price array of Product ID

- Create a RDD from reading metadata.json.gz
- Use map() to create a pair RDD (k,v) : k is Product ID and v is a array which only contains one price element of the Product ID
- Use reduceByKey() to union all Price arrays of same Product ID. A new pair RDD is created, contains pairs of Product ID key/ Price array value

Step 3 : Inner join

- Use join() between a pair RDD of step 1 and a pair RDD of step 2. Whereby, it generates a new pair RDD (k,v): k is a Product ID and v is a tuple of unique Review ID count and Price array

Step 4: Show top 10 products based on the unique reviewer ID count

- From the pair RDD in step3, use sortBy() to sort the pair RDD by descending of number of Review ID count
- Use take(10) to extract the list of first 10 elements from the pair RDD
- Use parallelize() to create the new RDD from the list
- Use map() to format the new RDD and write it to output folder. Using coalesce(1) , so only one part-00000 file contains all the results