Advanced Data Management (CMM524)

Laboratory #10: Macros

1. Aims

- Use macros to reuse Pig Latin code.
- Create samples of a dataset.

2. Outcomes

In completing this exercise, you should be able to:

- Write a simple Pig Latin macro to create sample of a dataset.

3. The Movie & Movie Rating Domain

In this lab, we have 3 datasets on movies and their ratings given by users. These datasets are exported from a MySQL database using Sqoop.

3.1. Downloading the Datasets

Download the following datasets from the resources of Lab 10:

- `movie`
- `movierating`
- `user`

- Unzip these files and put them into a folder in the Unix file system of your virtual machine.
  - Note: If you download them in the host OS (i.e. Windows), you may need to copy-and-paste the files from the host OS into the virtual machine.
- Depending on whether you want to work in Pig MapReduce or local mode, you can optionally upload these files to HDFS.

3.2. Creating Sample Datasets

Instead of using the original datasets, we would like to use smaller samples to test our scripts.

- Create a text file `makeSample.pig` with the following macro:

```
define makeSample(ORIGINAL_DATASET,SAMPLE_RATE,SAMPLE_DATASET)
returns void
{
/*
load the original dataset
create a sample with the give sample rate
store the sample into a sample dataset

Note: You should not fix the names of the files/directories or
sampling rate but use the parameters to the macro.
*/
};

/*
Invoke the macro here to create samples for the movie, movierating
and user datasets.
*/
```

- Complete the macro.
- Add Pig Latin to invoke the macro and create sample datasets:
  - Use a 5% sample rate for the `movie` dataset.
  - Use a 1% sample rate for the `movierating` dataset.
  - Use a 3% sample rate for the `user` dataset.
- Run your script to create the 3 smaller sample datasets.
- Check that the 3 sample datasets are created properly.

3.3. Percentage of Valid Years

You may notice that some movies' `year` values are missing. These movies have a value 0 in their `year` field.

- Write a Pig Latin script to calculate the percentage of movie with a valid year.
  - Hints:
    - One approach is to first find out the total number of movies, ignoring their `year` field validity.
    - Then find out the number of movies with a valid `year` value.
    - Finally find a way to combine these 2 results and calculate a percentage.
- Run your script on the sample dataset. Note the answer.
- Run your script on the original big dataset. Is your sample a good representation in term of the `year` field validity?

3.4. The Highest Rated Movies

One way to evaluate the quality of a movie is its average rating from all users.

- Write a Pig Latin script to calculate the average rating of all movies.
- List the top 50 highest average rated movie in descending order.

3.5. Distribution of Ratings

Some people argue that viewers tend to write reviews after they watched a really good or bad movie, but not those mediocre ones. To verify this claim, we will calculate the distribution of ratings in the 1-5 range.

- Write a Pig Latin script to calculate the percentage distribution of the ratings in the 1,2,3,4,5 categories.
  - Does the result suggest that people are more likely to say something when a movie is either very good or very bad?

3.6. Other Possible Analysis

Here are some questions that you can try to answer:

- Is there a co-relation between a person's profession and his/her favourite movie genre?
- Do matured audience tend to enjoy old movies more?
- Do males like action movie more than females?
  - Note: This one requires the extra datasets `genre` and `moviegenre`. Download them from Moodle.