Architectures for massive data management

Apache Pig

Ioana Manolescu , Silviu Maniu
INRIA Saclay & Ecole Polytechnique
Université Paris-Sud
ioana.manolescu@inria.fr , silviu.maniu@lri.fr
http://pages.saclay.inria.fr/ioana.manolescu/
https://silviu.maniu.info/

M2 Data and Knowledge, 2017/2018
Université de Paris Saclay

Apache Pig

- A declarative framework for handling data over HDFS and by using Hadoop/MapReduce
- Uses a language called PigLatin that can specify a large set of SQL-like commands
- Pig transforms the commands into an equivalent MapReduce plan

Pig Tutorial: Running Pig

1. Download and install Pig (see instructions at https://pig.apache.org/docs/r0.7.0/setup.html); you do not need Hadoop for this lab!
2. Download the **movielens-20m** dataset https://grouplens.org/datasets/movielens/20m/
3. Unzip the data, pre-process it (remove the headers but remember what the fields stand for)
4. Download the sample PigScript https://www.lri.fr/~maniu/lab2.pig
5. Adapt it (change the location of the *ratings.csv* file in line 1 with your location)
6. Run it using `<pig_dir>/bin/pig -x local lab2.pig`

Pig Tutorial: Understanding Pig

1. See the reference document at https://pig.apache.org/docs/r0.7.0/piglatin_ref2.html for a list of PigLatin commands
2. Briefly explain what the commands in lines 1, 3, 5, 7, and 9 compute
3. Using the EXPLAIN command, show a plan for line 3; explain it

Pig Lab: Task

1. Using lab2.pig as a base compute the following:
   - Show all users which have more than 100 reviews
   - Show the total number of reviews for each movie
2. Extend lab2.pig to use more than one relation (*ratings.csv*): using *movies.csv* and *tags.csv* compute
   - The average rating for 'Documentary' movies
   - For each 'Action' movie, the total number of tags that have been added
3. Send the resulting script (lab2_extended.pig) to silviu.maniu@lri.fr , along with a readme file which contains the explanation of the 4 resulting MapReduce plans, by **October 16th, 2017 Midnight**