Class 5: Homework

Realtime and Big Data Analytics

Summer 2017

Homework

Class 5

Analytics Project

1. Start forming project teams.

   Please email me if you would like me to introduce you to other students who are looking for a partner. Teams can have up to three team members. You can also use the Forum to find teammates.

2. Start discussions with team members around potential projects - settle on up to three possible projects (if you can choose one, that's great).

   This choice is not cast in stone - you will still be able to change it if necessary. You and your team will iterate on the idea and refine it in the coming weeks.

   Enter between 1 and 3 topic ideas into the NYU Classes Assignment - just a brief, one sentence description is sufficient.

Pig Reading Assignments

3. Read Chapter 16 of TDG (Hadoop: The Definitive Guide) pp. 423-424, 426-466 (skip HCatalog).Pages 457-466 review Pig operators - helpful for completing the homework assignment.

4. Optional: Read “Pig Latin”, by Olston, Reed, Srivastava, Kumar, Tomkins, SIGMOD’04 : http://infolab.stanford.edu/~olston/publications/sigmod08.pdf (This is not on the midterm exam.)

Homework

Class 5

5. Pig Program

If you are using the Quickstart VM or Dumbo, Pig is already installed.

The homework is to write a Pig program that is equivalent to the MapReduce word search program you previously wrote.

It is ok to use Grunt, but see if you can also write a Pig script and execute the script.
Please submit your Pig program, input, output, and screenshot to NYU Classes. Your program should do the following:

a. Search the given lines of input (see b. below) for these specific search strings:

```
hackathon, Dec, Chicago, Java
```

b. Accepts a small input file to be searched containing lines of the form: ***Date,Time,Name,Tweet***
Here is the ***exact data*** to type into your input file:

```
09-Dec-16,5:00PM;#Hackatopia,Tribeca Film Hackathon: Code As A New Language For Content Creators Hackathon
28-Oct-16,7:00PM;#NYCHadoop,Hadoop-NYC Strata/Hadoop World Meetup at AppNexus NYC
31-Dec-16,3:00PM;#Hackatopia,Designers, Developers, Doers, don't miss this upcoming Chicago hackathon
```

c. Your code will search for all of the search strings in the input file and output the number of tweets that contained each search string. The matching is not case sensitive.

d. Your code should output the number of tweets that contained each search string. Using the input data above, the resulting counts will be:

```
Chicago 1
Dec 2
Java 0
hackathon 2
```

e. Upload homework to NYU Classes. To receive full credit, please hand in all of the following items by next class:
- Your program, your input file, and job output.
- Evidence that the program ran successfully (e.g. screen shots or output log)

Homework

Class 5

6. Study for Midterm exam

The midterm exam will be given at the next class meeting, it is about one hour and 15 minutes long. We will have class following the exam.

- The exam has MapReduce programming questions (pseudo-code).
- You will not have to write a Pig program, but you may be asked to interpret a simple one.
- Material from papers is not on the exam, but material from slides is, so study the slides.
- **Study your notes** - we discuss material in class in greater detail than is covered in slides.
- Re-read the book reading assignments.
- Material through Pig is included on the exam
    - Exam questions will be based on Pig **slides** covered in class only
    - Exam questions will **not** draw from the Pig book reading or the Pig paper.