


Class 5: Homework

Realtime and Big Data Analytics



New York University
Courant Institute of Mathematical Sciences

Homework

Class 5

Pig Reading Assignments

1. 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.
2. *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

3. Pig Program

If you are using Dumbo, Pig is already installed and configured.

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 for all of the following strings in the input file containing tweet data:

hackathon, Dec, Chicago, Java

- b. Accept 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 (it is the same data used for the WordCount assignment):

```
09-Dec-18,6:00PM;#Hackatopia,Tribeca Film Hackathon: Code As A New Language For Content Creators Hackathon
28-Dec-18,7:00PM;#NYCHadoop,Hadoop-NYC Strata/Hadoop World Meetup at Google NYC
31-Dec-18,3:00PM;#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 lines that contained each search string. The matching is not case sensitive.

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

```
Chicago 1
Dec 3
Java 0
hackathon 2
```

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

Homework

Class 5

Analytics Project

4. Form 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.

5. Complete the Team Project Proposal (TPP) ✓

Complete the TPP template and submit a pdf of your proposal.

All team members should upload the same TPP as a pdf.

6. Draw *initial* design diagrams

Use PowerPoint, Visio, etc. to describe the design of your project. Include the software architecture (Big Data tools you'll use), the data flows, and anything else you think is important to show.

This is a first draft, you will refine it in the coming weeks.

All team members should upload the same diagrams as a pdf.

7. Create initial task list

Use the TaskList.xlsx/numbers template provided in Resources. Assign team members to tasks, and assign a due date to each task. Try to identify milestones – that will help you know if you are on or off track.

All team members should upload the same task list as a pdf.

Homework

Class 5

Analytics Project (continued)

4. Research your project. ✓

Each team member should upload a list of 5 papers relevant to your project and a short summary for one paper which includes your thoughts on how the paper is connected to your project.

It is useful to understand the state of the art before you begin a project. To do this, we read recently published papers. IEEE and ACM conferences and journals from 2017 to present are a good source.

Please do not choose papers on a Hadoop technology or other tool. Instead, choose papers related to your project thesis - the paper does not have to be in the same domain as your project. For example, do not choose a paper about Spark MLlib or Hadoop; do choose a paper about using big data in healthcare to solve some problem.

Where do I find a paper?

- Try using [GoogleScholar](#) to find papers.
- Try googling 'IEEE Big Data Analytics' for example - this brings up a bunch of conferences to pick from. Let me know if you have trouble finding a paper.
- Some other places to look for papers:
ACM KDD Conference: <http://www.kdd.org/kdd2016/>
ACM DL: <http://dl.acm.org/>
- Ask the professor if you get stuck.

List all five papers - title, authors, link where the paper can be found - at the top of your document. Below this list, add a summary (a few paragraphs) about one of the papers. This summary should include your thoughts on how the paper relates to your project. Upload to NYU Classes. (In a future homework, you will add the team summaries to the 'Related Work' section of your project paper.)

Coordinate with your teammates to ensure each member reads a different paper. Share what you learn with your teammates.

Note: The MapReduce, HDFS, and other papers already assigned cannot be used for this assignment.

Please provide the following:

- 5 papers - title, authors, link to paper
- For the summarized paper, provide:
 - Paper title
 - Paper authors
 - Link to the paper
 - Paper abstract
 - Your summary