Class 8: Homework

Realtime and Big Data Analytics

Summer 2017

Homework

Class 8

Analytics Project

1. Paper Drop #1

Create a first draft of your paper using the IEEE paper template (the template is in NYU Classes). Complete these sections of your paper using material that you have already produced:

a) Introduction: Use your project proposal text to create the Introduction section of your paper, and update/expand it as needed. (If you only wrote one sentence for your proposal, you should expand that now.)
b) Abstract: Write the Abstract section - it's a shortened version of the introduction.
c) Related Work: Add summaries of all papers that your team has read to the Related Work section.
d) References: Add references to the papers that your team has read to the References section of your paper.
e) Design: Add your design diagrams to the Design section of the paper. Fix diagrams as needed. Add text description.

Each team member should upload the paper to NYU Classes.

2. Code Drop #1

Develop code using only **Big Data tools** for the analytics part of your project. Use the Big Data tools specified in slides (Hive, Impala, HBase). If you want to use MlLib, Mahout, Spark, or some other tool not yet covered, or if you are not sure if the tool you want to use is permitted, please email the professor. Using R, Python, Java, Ruby, etc. is not permitted because these are not Big Data tools and they will not scale.

Please upload your in-progress code; it is ok if the code is still incomplete, it is ok if it doesn't compile.

You should be finished with cleaning up and formatting all data sources (ETL stage). Ideally, you should be running / assessing / improving your analytic starting this week.

Each team member should upload a zip of the team's analytics code to NYU Classes. (All team members upload the same thing.)

Do not include ETL code you previously submitted.

Readings

1. Please read the HBase chapter in TDG - Chapter 20 (pages 589-597 optional).

Homework - This is how the IEEE template looks

Class 10

Paper Title

Author Name
Dept. name of organization
City, Country
e-mail address if desired

Author Name
Dept. name of organization
City, Country
e-mail address if desired

Author Name
Dept. name of organization
City, Country
e-mail address if desired

Author Name
Dept. name of organization
City, Country
e-mail address if desired

Abstract—

(Use a short version of your project proposal here.)

Keywords—analytics,...

I. Introduction

(Paste your full project proposal here. When you've completed your final project, review this text and make sure it reads well and describes your final project.)

II. Motivation

(Write a couple of paragraphs describing why you think this analytic is important. Why should people care about this analytic?)

III. Related Work

(Each team member has read at least two papers related to their analytics project. Please add the paper summaries written by each team member here. Edit this section as needed to make it flow – explain the related work and how your work is similar/different/etc. Each paper reference should be added to the References section. When you refer to reference #1 in your paper, for example, use this notation: [1])

IV. Design

(Paste your design diagram here. When the design is final, you will put the final diagrams here and write some text to describe the diagrams.)

V. Results

(Future… In this section, you can describe: Your experimental setup/issues with data/performance/etc. Describe your experiments, describe what you learned. Did you prove or disprove your hypothesis? Were some results unexpected? Why?)

VI. Future Work

(Future… Given time, how would you expand your analytic? Could it be applied to other areas? Etc…)

VII. Conclusion

(Future… One or two paragraphs about the value/accuracy/goodness of your analytic.)

Acknowledgment

(This section is optional. It can be used to thank the people/companies/organizations who have made data available to you, for example. You can list any HPC people who were particularly helpful, if you used the NYU HPC.)

References

(Add references for all of the papers/texts that you refer to in your paper. You will probably want to include the papers you read that were related to your project. You may have websites to reference, the Hadoop book, the MapReduce paper, the Pig Latin paper, etc. Some references are added below as an example.)

[1] T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.

[2] A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.

[3] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implementation, 2004.

[4] S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP '03, 2003.