FINAL PROJECT – PHASE I

The final project is an essential part of this class. It will allow you to demonstrate your Big Data skills and create something that you are proud of. It can also be a valuable addition to your projects portfolio that you can demonstrate to prospective employers.

In the first phase of the project you will perform the following:

- Choose a topic that is related to Big Data and involves substantial design, analysis, programming, and validation.
- Be sure data is available. Mention source and features, such as size, of data.
- Choose your team members.
- Do preliminary research on your chosen topic and come up with an analysis document and list the future steps.

Project Ideas

Below are some of the project ideas. You can choose any one of them or choose a variation of one of the ones posted below. You are also free to propose a topic that is not listed below, provided it satisfies the requirements of Big Data and involves significant use of relevant technologies.

Below are some suggested topics.
Note: Two teams can not work on the exact same topic. Projects will be assigned on a first come first serve basis.

1. Real-time news sentiment analysis using live data from sources, such as:

- Google News, NYT API

- Social media sources → Twitter API, StockTwits

Ideas:

- Find sentiment of a stock price over time and correlate it with stock price.

- Find sentiment of a company that suffered a cyber security breach/attack. Create a vocabulary of terms associated with cyber breach/attack and correlate with stock price.

2. Using Yelp's academic dataset, predict star rating using sentiment analysis:

https://www.yelp.com/academic_dataset

3. Participate in the Yelp dataset challenge and submit a good entry:

http://www.yelp.com/dataset_challenge

4. Finding trending topics on Twitter.
- Using Twitter API, find topics that are trending in real-time. You can filter tweets by location or subject

5. Build a recommender system on DBLP's conference/publications dataset:

http://dblp.uni-trier.de/xml/

6. Find clusters in the DBLP conference/publications dataset

http://dblp.uni-trier.de/xml/

Clusters can represent authors with similar interests. You can use a text mining strategy for that.

7. [Bioinformatics] Create an approach for parallel sequence comparison using words as tokens.

** See me if you are interested in this project **

8. Take part in a Kaggle competition that involves significant amount of Big Data technologies

https://www.kaggle.com/competitions

9. Take part in the KDD cup challenge

http://www.kdd.org/kdd-cup

Note: KDD 2016 becomes live on March 4 and you can win prizes ☺
http://kddcup2016.azurewebsites.net/

10. Take part in Driven Data competitions

https://www.drivendata.org/

Phase I requirements

For this first phase, you are to do the following:

- Choose a team that consists of between 2 to 5 members.
  It's not possible to have more than 5 members in a team, as it becomes harder to assess individual contributions.

- As a team, choose a topic.
  The project should involve solving a medium or large sized problem using Big Data technologies, such as Spark, MapReduce, Pig, Hive, etc.

- Find data that you will use.
  Find the data and find its characteristics, such as number of instances, attributes, distribution of data.
  Evaluate the size of data and pre-processing requirements
  Submit a snapshot of the data.

- Propose a preliminary solution and workflow.
  This would involve sketching out some ideas on how you will solve the problem. You could also indicate the workflow for the entire project.

- Indicate your hypothesis and what you wish to accomplish.
  For example, you could say "We wish to prove that the stock price is strongly correlated with its sentiment on Twitter" or "We wish to solve this competition using Big Data technologies using xyz algorithm and come up with a more accurate solution"

Include all of the above in your report for phase I.

What to submit

Create a report containing all the parts asked for in the previous section. Also, be sure to provide names and proposed roles of all your team members.

Deadline

Midnight of March 11, 2016.