The common theme of all projects is "*Bigdata Analysis using* Spark".

Each project will have 3 parts:

**Part 1: Store data and test Spark**. At the end of this part, you should have (a) your selected datasets uploaded to your EC2 instance, and (b) tested that your Spark installation works well using a small example program.

In your pdf report, include the following information:

1a. Project description. What is the goal of your analysis, and what do you expect to find? Any hypotheses?

1b. Describe your data using a table (number of records, length of each record, number of attributes, and so on), a few example records, and a description of the data.

1c. Run the following code on Spark and show screenshots of the results.

nums = sc.parallelize([1, 2, 3]);squares = nums.map(lambda x: x\*x) ;even = squares.filter(lambda x: x % 2 == 0) ;even.collect();x = sc.parallelize(["spark rdd example", "sample example"]);y = x.flatMap(lambda x: x.split(' ')) ;y.collect();

You may pick any dataset (or combinations of datasets) you like, the only constraint is that it must be at least 1 GB. Examples are:

* [https://archive.ics.uci.edu](https://archive.ics.uci.edu/ml/index.php), [tweets](http://www.cs.ucr.edu/~vagelis/classes/CS235/ProtectedSlides/Tweets2011.zip), <http://www.kaggle.com/>, [http://data.gov](http://data.gov/), <https://star.cs.ucr.edu/>, or other source
* Extra credit: Crawl web to get Web pages or images (can use Scrapy or jsoup or other software). Or collect social network data, e.g., use Twitter streaming API (may be unavailable), Reddit API, Instagram API (check if you are able to get more than 1 GB).

**Part 2: Spark**. Preprocess or analyze your data in a distributed (parallel) way using **Spark,** andstore the output in the **MySQL (**or a NoSQL database like Cassandra**)**. E.g.,

1. Locate shapes in images or any other analysis on images. You may use existing source code for image analysis and adapt it to work with Spark.
2. Compute average sentiment for the pages that contain a specific keyword (e.g. covid) for various domains
3. If you data is tabular, compute avg income by zipcode, or other group-by queries  (tabular data are unlikely to be more than 50GB so you may want to combine with other data)
4. Find most popular hashtags in Twitter for every day, or build a spatial index that for each city, has a list of tweets.
5. Compute sentiment or political party bias for each tweet

**Part 3: Build Web interface** (use your favorite web programming framework) to explore the preprocessed or analyzed data. Your Web application should read data from the MySQL database. Examples:

1. Allow searching pages by keyword. You could use something as simple as LIKE in MySQL or install Lucene for better matching.
2. Search images by shape
3. View data on map. E.g. geotagged tweets or demographic info by state or county
4. Search for a movie by actor or genre and show sentiment of reviews

**Deliverables (all in PDF):**

|  |  |  |  |
| --- | --- | --- | --- |
| Deadline | Deliverable | Description | points |
| 10/15 | Form Groups and Project Proposal.  Email to instructor (cc TA) | Each group has **four**  members. If you cannot find partners, email the TA to match you with another student ASAP.  Proposal consists of 1-2 pages describing in detail what you will do for each of the three parts of the project. |  |
| 10/29 | Part 1: Data Collection.  Submit in Canvas. Also submit zip file with your code (if any) in canvas. | Report containing Requirements, Design, Implementation, Evaluation (statistics on amount and properties of data, time it took you to collect it), Screenshots, Contribution of each team group member.  Collect your data, clean it, and store it in the lab servers; report; show your data to the TA in the lab | 10 |
| ~~11/12~~  11/17 | Part2: Spark Data Processing and store in MySQL.  Submit in Canvas. Also submit zip file with your code in canvas. | Requirements, Design, Implementation, Evaluation (execution time graphs with varying # Spark workers (1 vs. 2), and varying data sizes), Screenshots, Contribution of each team group member. Discuss your evaluation graphs.    Demo to TA in lab on 11/17. | 25 |
| tbd | Demo of Part 3 | Demo to TA in lab. |  |
| 12/08 | Final Report.  Submit in Canvas. Also submit zip file with your code in canvas. | For Part 3: Requirements, Design, Implementation, Evaluation, Screenshots, Contribution of each team group member.  For final report: Include Parts 1,2,3. Address comments you received in earlier submissions of Parts 1,2. | 65  (30 for Part 3,   10 for Part 1 revisions,   10 for Part 2 revisions; if no revisions were requested, you get the points anyways  15 for project interestingness, novelty and presentation) |

**Notes:**

You will be graded on factors including: the interestingness (10%), the technical challenges, the robustness, the cleanness of your code and documentation, and your presentation (10%).

If one of the three parts in your project is too complex, you can simplify another part.

**Additional data sources:**

1. US Census Bureau http://www.census.gov/data.html   
  
2. million songs Dataset  
http://labrosa.ee.columbia.edu/millionsong/  
  
3. usenet corpus dataset  
http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html  
  
4. Google books n-grams:  
https://aws.amazon.com/datasets/8172056142375670  
  
5. Click datasets:  
http://cnets.indiana.edu/groups/nan/webtraffic/click-dataset  
  
6. wikipedia dump: https://dumps.wikimedia.org/