CS 7280 Special Topics in Database Management Spring 2024

Project 3: Big Data Analytics

Objectives:

1. Understanding Hadoop Ecosystem and Data Analytics
2. Become familiar with MapReduce programming and Spark
3. Gain experience with research on big data and data analytics

This will be a group project (by 2 students) for one semester. The main purpose of this project is to become familiar with Big Data platform, including Hadoop system, MapReduce programming, and cloud based big data solutions (e.g., Google Big Query). You need to follow the instruction to conduct the project.

Phase 1 (15%): Selecting Data Set - Due: March 20, 2024 (Wed)

- Each student researches on any data that you are interested in, and collect the information about the data.
- Find any characteristics of the data you select, and describe why you are interested in
- If possible, prepare 3~4 sample data, which can be either real data or manipulated one.
- Make 2~ 3 pages of Powerpoint file as a report
- Submit the PPT file to Canvas
    - **PPT, PPTX or PDF file format ONLY**

at least 1GB
json/csv format

Phase 2 (15%): Defining Problems – Due: March 27, 2024 (Wed)

- In this 2nd phase, you are going to research on the following topics based on the data you selected in Phase 1:
    - What you can analyze using the selected data in terms of Hadoop HDFS with Spark and MapReduce programming, and Google Big Query using GCP. That means your project MUST be implemented by MapReduce including Java and other streaming programming in Phase 4-5 and Spark.
        - 1 MapReduce Programming
        - 1 Spark
        - 1 Hive or Pig
        - 1 Google Big Query using GCP
    - How you can collect the data. That means your data MUST be uploaded to HDFS using VM in Phase 4-5.
- Make 2~ 3 pages of Powerpoint file as a report
- Submit the PPT file to Canvas
    - **PPT, PPTX or PDF file format ONLY**

Phase 3 (20%): Preparing Proposal – Due: April 3, 2024 (Wed)

- Prepare a proposal using a MS word template: A proposal template can be found at Canvas
  - **DOC, DOCX or PDF file format ONLY**
- Prepare and submit 5~10 pages of Powerpoint file for presentation
  - **PPT, PPTX or PDF file format ONLY**
- Then, submit 10 minutes presentation video to Canvas
  - Submit a link such as YouTube, or record your presentation using Canvas
- In your proposal, you need to consider how to prepare the final deliverable of following outputs
  1. Write-up
  2. Source code
  3. Data set
  4. Poster

  ** Note that this is a plan to prepare 1 ~ 4 above. NOT implementation right now.
- Then, submit your proposal to Canvas
- Prepare for 5 mins presentation for your proposal (submit PPT file also)

Phase 4 (25%): Implementation – Due: April 10, 2024 (Wed)

1. Preparing Data and Upload to HDFS. You can use variety of ways to prepare your data set including:
   - Use API provided by each website, such as Facebook API, Twitter API and Flickr API
   - Use benchmarking data sets, such as
     - UCI data set: http://archive.ics.uci.edu/ml/datasets.html
     - Wikipedia database: https://en.wikipedia.org/wiki/Database_testing
   - Government database
     - US Census data: http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
     - NOAA weather data: https://www.ncdc.noaa.gov/cdo-web/
   - Implement Data collection program using Web query
   - Synthesized data set
   - Use googling
2. You data set MUST have at least 100,000 instances (or rows)
3. Upload your data set into HDFS (VM)
4. Implement MapReduce program on your Developer VM and Spark
   - You can use Java or any Steaming with other program language such as Python.
     - 1 MapReduce Programming
     - 1 Spark
     - 1 Hive or Pig
5. Submit your source code to Canvas and download link for your data set

- All source files should be compressed with TAR (e.g., tar cvf XXX.tar) on VM (**JAR, TAR or ZIP file format ONLY**)
- For the dataset, you can upload it to Google Drive (or any Web hard) and then send a link when you submit your source

6. Then, submit 10 minutes demo video to Canvas
   - Submit a link such as YouTube, or record your presentation using Canvas

Phase 5 (25%): Presentation of Project – Due: April 17, 2024 (Wed) before class.

1. Writing-up (at least 4 pages with IEEE format). You must use IEEE format.
   - o DOC, DOCX or PDF file format ONLY
2. Poster (36 x 24 inches Powerpoint file). You can use one of templates provided on Canvas.
   - o PPT, PPTX or PDF file format ONLY
3. Submit your paper and poster to Canvas
4. Make 8 ~ 10 pages of Powerpoint file and submit to Canvas
   - o PPT, PPTX or PDF file format ONLY
5. Then, prepare 8 minutes final presentation on April 27, 2022 (Wednesday)

Submission

You will submit your program using Canvas. If you have any trouble to use blackboard, you can contact TA or instructor.

Grading

15	Phase 1
15	Phase 2
20	Phase 3
25	Phase 4
25	Phase 5

Bonus +20 for high quality writing-up that can be submitted to either conference or journal paper.

.