System Investigation Project
DATA 514
Spring 2022

The goal of this assignment is for you to gain practical experience with a data management system and communicate the results of your investigation to the class.

Summary: In short, you will pick a system of your choice, you will learn about that system and present your learnings to the class, and you will load some data into that system and write queries to retrieve data from that system. If you prefer to write a paper instead of doing an implementation, that option will also be available.

Teams: The assignment may be done individually or in teams of up to 4.

Signup - Schedule & Topic Preferences:

- Please use this google form to indicate your preferred system and preferred presentation date System Investigation Team Signup
- Please sign up by 4/5 at 11pm. After 11pm on 4/5, I reserve the right to assign you to a system and presentation date.

System:

- You will pick a data management system of your choice. Some suggested systems are listed below.
- Spark and Mongo appear to be the systems with the most interest. Our TAs have some familiarity with Mongo.
- I have significant documentation on installing and running Mongo on Google Cloud.
- I have limited documentation on running and installing Spark on Google Cloud.
- You will be able to use your first choice system.
- I will assume that you are using the system specified as your first choice in the signup sheet unless you tell me otherwise.

Cloud Platforms:

- We have $200 in Azure credits per student, you will need some of that for the homeworks, but I expect you to have left over credits.

- I have requested Google Cloud credits for use by the class and will let you know when those are available. I expect $50/person.
- You may use whatever cloud system you like if you have access to it.

Datasets:

- You may choose your own data set (e.g. from kaggle), or you may use one of these datasets.
- The datasets at that link have both a link to a data set and a suggested set of questions to be asked over the data set.
- If you pick your own data set, you will need to create your own questions.

Part 1: Project Plan (due 4/15)

- What system are you using?
- What dataset will you use?
    - Questions to be answered if you are not using one of the existing datasets
- Team roles
    - Define roles so I am aware of who is responsible for what (will be considered in grading)
- Define successful project completion.

Part 2: System Summary & Design (due 5/11)

- A short paper:
    - Summarizing the key features of your system and
    - Your initial system design
- System Installation
    - Provide verification that you have installed your system and loaded a piece of data into the system.
- System Design
    - How will you store data in your system?
    - How will you execute your queries?
- Do not need to run queries or load data at this point, but you may wish to...
- You may update definition of success at this point if you wish.
- Make a decision as to paper/implementation at this point.

Part 3: Class Presentation (weeks 5-10)

- Class presentation on the system you have selected
- Presentations will take place in Weeks 5-10 and the finals slot.
- Presentations are ~7-10 minutes per team member.

- Schedule to be announced by 4/8 and will be based on preferences in the signup sheet.
- Topics to be negotiated for the popular systems

Part 4: Implementation or Paper (due 6/7)

- Implement your queries over your data set
- Provide a demo of your implementation
- Paper alternative if you prefer. You may make a decision on paper vs. implementation when Part 2 is due.

LIst of potential systems to use:

- TensorFlow
- TACO
- System ML
- SageMaker (requires AWS)
- Mongo (MongoDB)
- Cassandra (Apache Cassandra)
- Spark (Apache Spark)
- CouchDB (Apache CouchDB)
- SimpleDB (Amazon SimpleDB)
- Accumulo (Apache Accumulo)
- Redis (Redis)
- Or … suggest a system or topic you are interested in …