Anushna Prakash & Preston Stringham
DATA 516 - Scalable Data & Algorithms
October 29, 2021

Project Proposal: Benchmarking TPC-H And IMDb Data On Popular DBMSs

1. What question are you planning to address?
    a. How do the different proprietary database systems perform in terms of query runtime on large data sets?
2. What system are you using? Redshift/snowflake/spark/sql-server/postgres/etc/etc? Do you have access to it?
    a. We will use Redshift, Snowflake, and Spark.
3. What data are you planning to use? Do you have access to it? If you know details, describe briefly how large it is, in how many files does it come, in what format does it come (csv/json/xml/whatever), and whether you thought how you will import it into the system.
    a. We will use TPC-H and the publicly-available IMDb dataset from https://www.imdb.com/interfaces/
    b. TPC-H
        i. This data is 10GB
        ii. It comes in 8 tables: customer, supplier, orders, region, nation, part, partsupp & lineitem
        iii. It comes in a .tbl file
        iv. It can be imported from an s3 bucket using a URL
    c. IMDb
        i. This data is compressed, with each file being a minimum of 32KB compressed. With all files uncompressed we expect the data to be roughly 4GB
        ii. It comes in 7 tables: name.basics, title.akas, title.basics, title.crew, title.episode, title.principals, title.ratings
        iii. It comes in a .tsv.gz format and can be downloaded directly onto the machine from the https://datasets.imdbws.com/ website.
4. What do you hope to report in your project? For example, a graph showing the runtime as a function of the data size; or a bar chart showing the runtimes for 10 queries with features X enabled and with feature X disabled.
    a. We will report bar charts showing the runtimes for several queries with increasing levels of complexity (3 joins, group-bys/aggregates, etc).
    b. We will develop a suite of queries to run on each database management system
5. If you work in a team of two, tell us about how you are planning to split the work.
    a. Each of us will design the queries that meet the complexity needs of the benchmarking task, for example if we determine that 2 queries with 4 or more joins is necessary, then each of us will design one query each.
    b. Each of us will take our own time to test the queries and produce results and record warm cache timing from one of our respective machines if the queries need to be run locally as opposed to on a cloud server.

c. We will both contribute results to the final presentation/paper based on the work each of us performed in part 5b.