DATA516 Project Proposal

Aviva Munshi & Nizan Howard

1. **What question are you planning to address?**

Our topic is on Sampling, specifically on strategies for accelerating joins and group by aggregating queries. We’re going to try and see how to analyze the different sampling techniques that affect the efficiency of complex joins in group by aggregate queries for big data on execution time and accuracy.

1. **What system are you using? Redshift/snowflake/spark/sql-server/postgres/etc/etc? Do you have access to it?**

For our project, we have access to two powerful database systems: Snowflake and PostgreSQL. Our decision on which to use will depend on the specific needs and goals of our project, as well as the features and capabilities of each system that can best help us achieve our objectives.

1. **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.**

We’ll use the [DataBank international debt statistics](https://databank.worldbank.org/source/international-debt-statistics). It’s an open-source dataset that contains information about the amount of debt in USD owed by developing countries across several categories.

1. **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.**

* **Execution Time**

We'll create some line graphs to help us understand how the time it takes for our queries to run relates to the different sample sizes we use. These graphs will show us how the time changes as we work with larger or smaller samples.

* **Visualizations on Accuracy vs Efficiency Trade-off**

We're going to use scatter plots and line graphs to give us a clear picture of how we can balance the accuracy of our query results against the time it takes for those queries to complete. This will involve looking at how accurate our results are versus how long the queries take when we use different sampling rates or methods.

* **Recommendations**

After we've gathered all the data and analyzed the findings, we'll provide suggestions on which sampling methods are the best fit for various types of queries or different situations. This will help guide us on when to use which method based on our specific needs.

1. **If you work in a team of two, tell us about how you are planning to split the work.**

In our two-member team, we’ve structured our workflow to optimize our efficiency and output while also taking into account each other's strengths. After our discussion, we’ve divided our responsibilities in a way that enhances our collaboration and the quality of our work.

Aviva is responsible for the initial phase, which involves the initial queries and the preliminary tests - basically responsible for creating the foundation of the project.

Nizan will then follow up, building upon the initial work that Aviva drafted by delivering into more in-depth analysis, fine-tuning the queries, and conducting additional experiments.

Both Aviva and Nizan will be contributing equally to the documentation and reporting of the project by working on it together.

Project Milestones:

* Submit via [Gitlab](https://gitlab.cs.washington.edu/jackkhuu/csed516-2023au/), in the porject directory. E.g. project-milestone.pdf
* This is a preliminary draft of your final report.
* Summarize what you did already and what you are still planning to do.
* If you have any preliminary graphs, show them! You can change them later. Mention how you split the work, in case you are working in a team.

I am planning to meet with each of you individually shortly after you submit the milestone.

Meet:

* Download and play around with postgre sql before than ( meet at 12 inperson/ your house )

(Discuss thursday: )

* Take two of the preliminary sql joins and run them and track the time on data
* Run them again and track the time
* Take the data (graph it )

Table schema:

CREATE TABLE WorldDebt

(

Country VARCHAR(50),

CountryCode VARCHAR(50),

Indicator TEXT,

IndicatorCode TEXT,

DebtAmount FLOAT

);

s3://data516-aviva-nizan/scalable\_int\_debt.csv

copy table name from 's3://uwdb/tpch/uniform/1GB/customer.tbl' REGION 'us-west-2' CREDENTIALS 'aws\_iam\_role=<>' delimiter '|';

copy WorldDebt from ‘s3://data516-aviva-nizan/scalable\_int\_debt.csv’ REGION ‘us-west-2’

CREDENTIALS ‘aws\_iam\_role=<>’ delimiter ‘|’;