`Drexel University
College of Computing and Informatics

INFO 323 – Cloud Computing and Big Data

Project Proposal

Benji Bui

Recommendation system using business reviews using Apache Spark on GCP

---

1. PROBLEM DESCRIPTION: (Clearly describe the data analytics problem...)

My problem description is that I have some free money and I am looking to invest in a restaurant. I want to understand what factors contribute to a restaurant's success.

2. DATA SETS: (Describe the data sets…)

I am using Yelp Open Dataset. It includes 3 JSON files:

- business.json (118.9 MB) : business attributes, categories, locations, etc
- review.json (5.3 GB): user reviews with star rating and timestamps
- user.json (3.4 GB): user information

For the dataset, I will convert them into RDDs using Spark. They will then be transformed into structured DataFrames, analyzed for valuable insights, and stored in Google Cloud Storage (GCS) in Parquet format for storage efficiency. Spark will also be used to perform large-scale analytics, including filtering, aggregation, joins, and trend analysis.

3. PROJECT GOALS: (Outline the specific goal…)

- Identify top-performing restaurant types and attributes based on average star ratings and review count.
- Analyze geographic trends: Which cities/states have higher restaurant success rates?
- Detect temporal patterns: What months/seasons have higher review volumes or better ratings?
- Generate a feature-based success model to recommend.

4. CLOUD COMPUTING COMPONENTS: (Detail how you will leverage cloud platforms and tools…)

- Google Cloud Storage (GCS): Used to store both raw JSON files and cleaned datasets in Parquet format for efficient storage and access during processing and analysis.

- Apache Spark: Spark is used to clean the data, perform joins across datasets, compute aggregations, and generate analytical features from the Yelp data.

- Vertex AI Workbench: A managed Spark environment will be used to run PySpark scripts for scalable data processing.