COMP 6231 - Group 2
Distributed System Design

Data Analysis of Yelp Dataset Using Google Cloud Platform

Team Members:

- Anuja Somthankar (40265587)
- Darshan Kansara (40195725)
- Prachi Patel (40261038)
- Nimisha Jadav (40267767)
- Nisarg Adalja (40276285)

Contents

Problem Statement

Yelp Data Analysis: With Apache Spark, Google BigQuery, and Google Kubernetes Engine for Deployment

The 4 questions that we target are:

1. Which are the top 10 restaurants were rated 5 stars by most users?
2. Which are the top 5 usernames that have received the most compliments?
3. Which are the top 10 usernames whose reviews have been marked as most useful?
4. What is the total number of unique businesses were given more than 3 stars by at least 50 reviews in 2018 in Philadelphia?

Systems Used

- Apache Spark - Open-source distributed computing platform, offering a flexible cluster computing framework for handling big data. Because it can store intermediate data in memory rather than writing it to disk, its in memory processing capability is one of the main advantages. It can handle a range of workloads, such as streaming, interactive queries, iterative algorithms, and batch processing.
- Google BigQuery - A fully-managed serverless analytics and data warehouse platform. It can handle batch and streaming data processing. Key feature is the division of compute and storage resources in integration with GCP
- Google Kubernetes Engine - An open-source container platform that manages containerized workloads and services. It simplifies the deployment, management, and scaling of containerized applications using Kubernetes, an open-source container orchestration platform.

METHODOLOGY

- Creating a Django project on GitHub
- Docker for the project
- Creating deployment.yaml, service.yaml and autoscaler.yaml
- Creating a cluster in Google Kubernetes Engine
- Deploying to the cluster

DISTRIBUTED FEATURES

- Clustering
- Scalability
- Fault Tolerance
- Load Balancing
- Naming
- Availability

Comparison of Apache Spark and BigQuery

Google BigQuery's processing time is much faster than the Apache Spark. Following is the time comparison in seconds:

Question	PySpark	BigQuery
1	7.308	2.02
2	5.942	0.933
3	12.497	0.915
4	9.202	0.88

Demonstration

THANK YOU