Design Document
Team Mount Evans

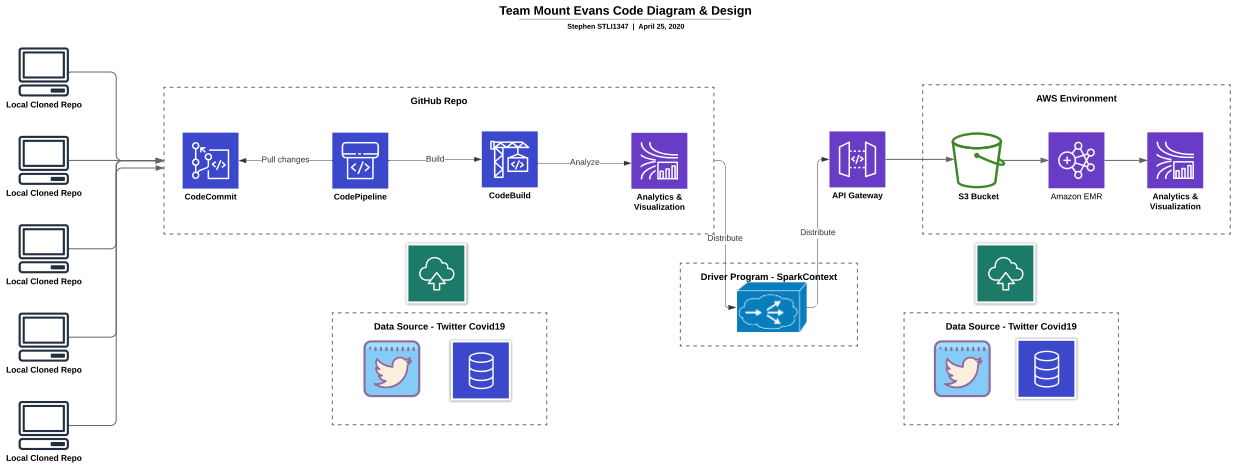
Team Mount Evan's objective for the project was to analyze large sets of Twitter data produced in the midst of the COVID-19 pandemic. The data analyzed stretched from the beginning of March 2020 to the beginning of April 2020. While the data came from all across the world, the team decided to analyze tweets that are in English only. The data was sourced from Kaggle.com. The team used GitHub as a centralized repository for shared data files and working code. The team also used AWS EMR for storage, code construction, and as a local cluster to test iterations of code. The methodology is described below:

Step 1: Distributing Raw Data
- Download datasets from Kaggle.com and upload to team Github for initial analysis
- Subset dataset for preliminary analysis. The team subset March 12, 2020 Twitter data to ~6000 rows
- Ingest subset dataset to members' individual machines for analysis. Upload working code into Github repository throughout project
- Ingest subset dataset to S3 bucket for analysis with AWS EMR for continued analysis

Step 2: AWS EMR
- Create independent cluster on AWS EMR for local analysis
- Use JupyterLab Notebooks and S3 storage to store data and develop code wihin EMR
- Update team's shared GitHub repository with developing Python and Pyspark code; finalize code

Step 3: Final Deployment
- Load data to Leeds S3 bucket
- Scale up final analysis code to handle larger Twitter datasets
- Run final analysis in Leeds EMR cluster where multiple iterations can be processed
- Introduce new data as needed

Team Mount Evans used the above procedure to explore sentiment analysis on Twitter regarding the pandemic. With the tools available in Amazon's EMR environment, the team was able to calculate and visualize the change in positive and negative sentiment over a period of time. AMS EMR proved to be an extremely useful platform that allowed for ingestion, storage, and analysis that could be scaled up or down as the project demanded.



Datasets

https://drive.google.com/open?id=1o26j9gJgmjtFRodo-4gNusJNqbMavZ56