Big Data and Intelligent Analytics
Spring Semester 2021

INSTRUCTOR: Sri Krishnamurthy
analyticsneu@gmail.com

Assignment 1: Three experiments with Big data

In this assignment, you will develop a data pipeline to ingest, process, store it so you can access it through different means.

Preparation:

- Review the tutorials we discussed in class on Snowflake, Apache Beam, Dataflow, Accessing Amazon S3 buckets, Sevir etc.
- Review: https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SEVIR_Tutorial.ipynb
- https://www.ncdc.noaa.gov/stormevents/ftp.jsp

Case:

You are a freshly minted data scientist engineer at WeSpace Inc. which leverages space and weather data to build forecasting systems! It is exciting and you just bought your first Tesla to celebrate. Post Covid, you drive to your office and your manager is kind enough to offer the corner office with the view of the ocean on one side and mountains on the other. She checks in to make sure you don't miss the freshly baked cookies that are at the kitchen. During group lunch, the team is engrossed in a fresh challenge. They plan to build a nowcasting system leveraging satellite data and weather datasets but everyone in the team wants to build models!

Where is the data ? You ask. Everybody stares at you and pretend it isn't a problem and continues the conversation about Deep Learning and other problems. Your manager

interjects and asks your question to the group again. "That is an important question! Where is the data?".

Some team members say it is all on AWS and in csv files and HDF5 format.

"How do we plan to access it?" You ask!

"Well, we will get there when we get there!", a team member says!

Your manager in your 1:1, commends you for asking the hard questions and gives you your first assignment! First, we need to try out how to access the data. She suggests 3 architectures and asks you to experiment and provide your recommendations.

Here is the dataset catalog:
https://github.com/MIT-AI-Accelerator/eie-sevir/blob/master/CATALOG.csv

- https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SEVIR_Tutorial.ipynb
- https://www.ncdc.noaa.gov/stormevents/ftp.jsp
- https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Export-Format.pdf

Gives you information on the datasets we need for the project.

Experiments needed:

1. AWS S3

- **Leave the data here where it is: https://registry.opendata.aws/sevir/**
- **Upload a sample of data to Amazon S3**
- **Experiment with Amazon Glue and build a pipeline.**
- **Show how to use Amazon Quicksight to query and sample datasets**

2. Google

- **Download sample datasets and move it to Google storage buckets**

- **Experiment with Google Dataflow and build a pipeline**
- **Show how Google Bigquery and Data studio to query and sample datasets. (https://cloud.google.com/bigquery/docs/visualize-data-studio)**
-

3. Snowflake

- **Download sample datasets and move it to Snowflake**
- **Use Sql-alchemy and Apache Superset to query and sample datasets.https://superset.apache.org/docs/databases/snowflake**

Questions to answer:

1. **How can you combine satellite data + storm data?**
2. **Which architecture would you recommend for machine learning purposes?**
3. **What are the challenges working with different architectures?**
4. **Do a cost/benefit analysis for each architecture**

Deliverables (Due March 4th 11.59am):

1. A 2-5 page report in https://github.com/googlecodelabs/tools format to illustrate your understanding of how various steps are integrated in the pipeline.
2. A fully functional pipelines that when run generates the outputs when initiated to publish data into buckets.
3. You will be given 10 minutes to present your company analysis in class on March 4th.