DS4300: Large Scale Storage and Retrieval

Prof. Rachlin, Northeastern University

The Challenger Accident




Problem Description:

On January 28th, 1986, the space shuttle Challenger exploded approximately seventy seconds after lift-off, killing all seven astronauts on board. According to the Rogers Commission that investigated the accident, this tragedy was caused by the unusually cold temperatures reducing the resiliency of the rubber O-ring seals connecting each section of the solid rocket boosters. The loss of resiliency led to a gas leak at one of the O-ring joints and the eventual explosion.

How cold was it that day at Cape Canaveral? In this assignment you will analyze temperature data for weather stations within 100 km of the NASA facility. Do your analysis in Scala and PySpark. I recommend using Jupyter Notebooks, but you can alternatively write stand-alone a spark application or make use of the Databricks ecosystem to do your analysis in the cloud. Submit your Jupyter notebook or PySpark / Python application code for grading.

The basic approach is

1. Read the station and temperature data

2. Filter and clean up the data. There may be some duplicates, and many stations are missing GPS coordinates

3. Identify all weather stations within 100 km of Cape Canaveral (Note: not all of the stations necessarily recorded a temperature on any given day.) When measuring distances between two GPS locations, make use of a haversine function to correctly account for the curvature of the Earth. Here is an implementation in python. In your solution, you'll want to implement a user-defined function (UDF) so that you can apply this function across the Spark DataFrames as with any other column-oriented ("vectorized") function. https://stackoverflow.com/questions/4913349/haversine-formula-in-python-bearing-and-distance-between-two-gps-points

4. Use inverse distance weighting (https://gisgeography.com/inverse-distance-weighting-idw-interpolation/) to estimate the temperature at Cape Canaveral on January 28, 1986. (Weather stations closer to the Cape Canaveral will be given more weight than stations that are far away.). Only consider stations within 100 km that recorded a temperature on January 28th.

5. Plot the temperature at Cape Canaveral for every day in January 1986. Comment on your visualization. What might have happened if NASA had decided to delay the launch for a couple of days?

What to submit: Submit your code (.ipynb or .py) and visualizations (if not already embedded in your notebook file.)