Project 5 (Spark)

Matthew Lueder

CIS 677

Overview

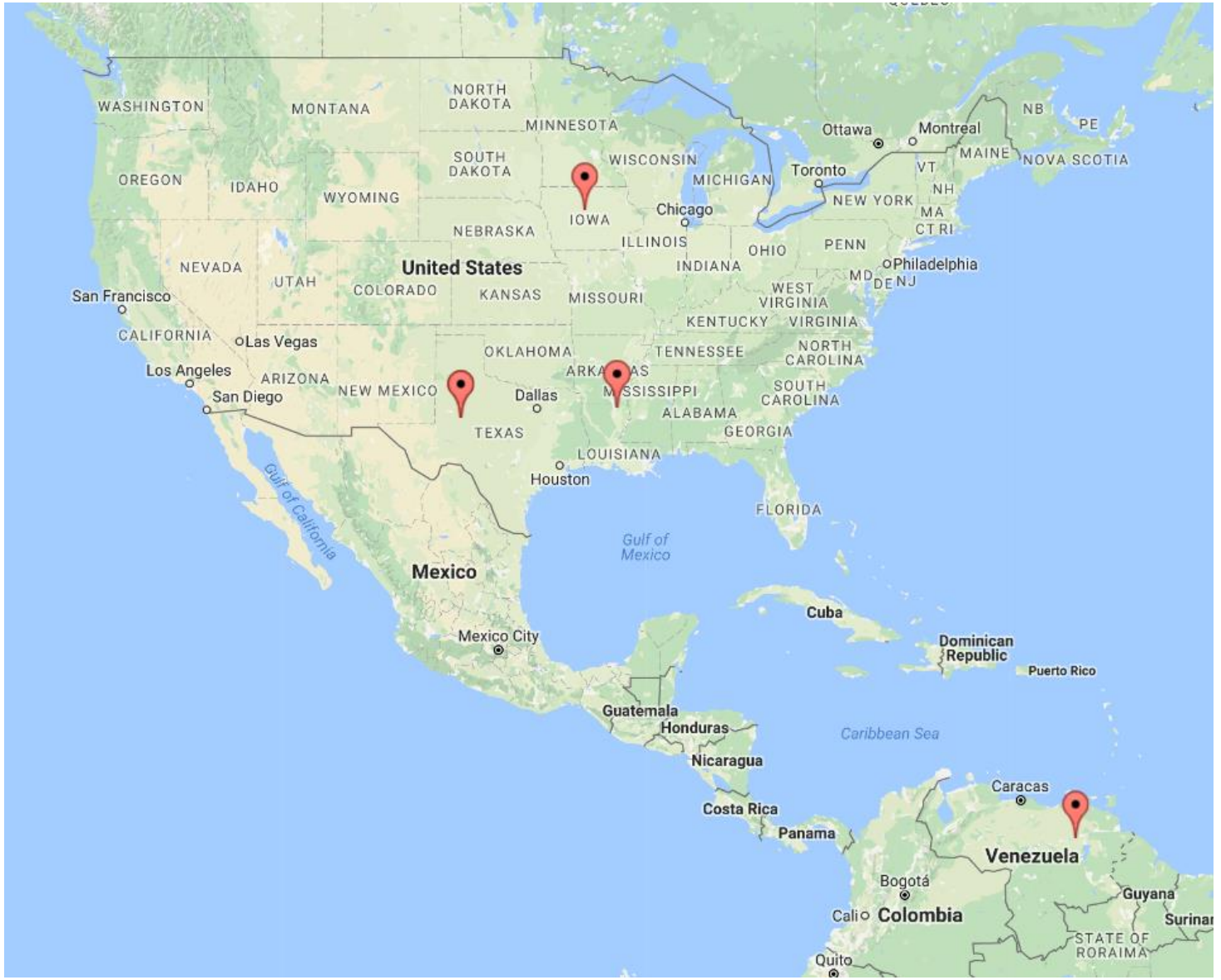
- Goal: Create a Spark-based MapReduce application that finds the highest yearly temperatures from 2008-2012 from a large collection of weather data
- A performance analysis was performed to test how the time it takes to execute the program relates to the number of machines in the cluster
- Coordinate data was taken from the weather stations where the highest temperatures occurred to allow me to create a visualization to show location.
- The analysis was performed on machines in the EOS lab.
  - The number of machines in the cluster was varied from a maximum of 31 to a minimum of 5.
- The data used is a set of compressed files obtained from the National Climatic Data Center
  - 20.64 GB
- My program works by:
  1. Loading all weather data from relevant years into an RDD
  2. Reducing the information in each dataset to contain only data on year, temperature, temperature data quality, and coordinates
  3. Restructuring this data so that year for each entry is treated as a key and the rest of the data is stored in a tuple and treated as the associated value
  4. Filtering out erroneous/suspect data
  5. Performing a reduction which finds the maximum value of temperature based on key

Results - output

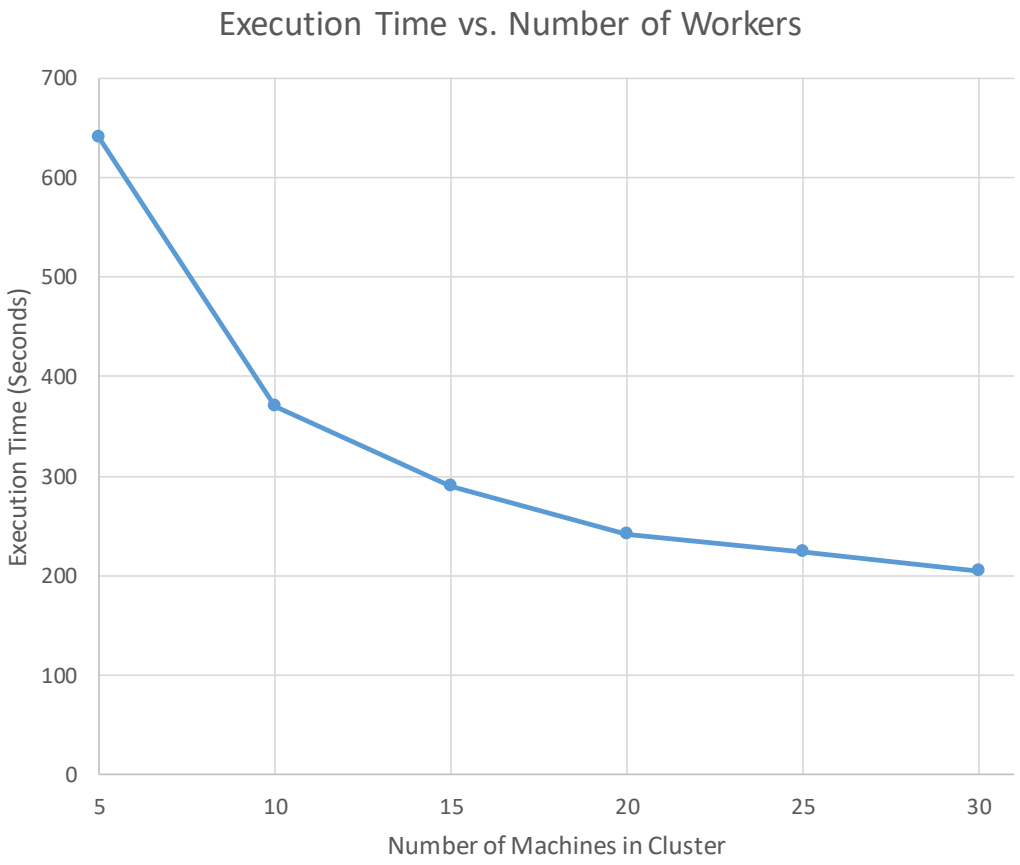
- 2008
  - Highest Temperature = 61.0 °C
  - Latitude = +32.213
  - Longitude = -101.520
- 2009
  - Highest Temperature = 61.0 °C
  - Latitude = +36.200
  - Longitude = -81.650
- 2010
  - Highest Temperature = 61.7 °C
  - Latitude = +32.213
  - Longitude = -101.520
- 2011
  - Highest Temperature = 61.8 °C
  - Latitude = +42.436
  - Longitude = -93.867
- 2012
  - Highest Temperature = 60.0 °C
  - Latitude = +8.150
  - Longitude = -63.550

Results - Map

- The temperature record for 2008 and 2010 occurred at the same weather station in Texas



Results – Execution Time



Number of Machines in Cluster	Elapsed Time (seconds)
5	639.978431
10	370.13217
15	290.360211
20	242.194077
25	223.006742
30	205.1571629

Discussion

- The results are suspicious.
  - 60 °C = 140 °F
  - The highest ever recorded temperature is 56.7 °C according to the World Meteorological Organization
  - The record for each year according to my data was above 60 °C
  - Likely that data which was marked as good is actually bad
- Locations for the highest temperatures make sense, besides one year which had the highest temperature in Iowa
- It is important when trying to adjust the number of workers, to not edit the slaves file before calling the stop-all script. This will leave the workers you removed from the file running. I learned this the hard way, completing my analysis then coming to the realization that all my trials were done on the same number of machines.
- Execution time decreases with the number of workers steeply at first but then starts to level off. This trend is similar to what we have observed with OpenMP and number of cores, and MPI and number of machines.