A Wildfire Can Make Things Dire

Wildfire Data Analysis: Group 20

By: George Orduno Galicia, Diane Shan, Dong Tran

Student Information:

Diane Shan (dshan017; 862148900) - Task 1
Dong Tran (dtran038; 860975199) - Task 2
George Orduno Galicia (gordu004; 862148001) - Task 3

Overview:

We analyzed a dataset that consists of wildfire occurrences in the US, where each point in the dataset is an occurrence of a wildfire along with other relevant information such as, fire intensity, location, and date. First we prepared the data to be processed by adding a new "County" attribute then converting the file into Parquet format. After that, we did both spatial and temporal analyses on the data. Spatial analysis was conducted by computing the total fire intensity for each county over a given start and end time. Temporal analysis was conducted by computing the total fire intensity each month over all time when given a specific county by name and plotting the results on a line chart.

We used a combination of Beast and SparkSQL to perform our analysis. We chose these systems because they integrate well together to complete everything needed to do a proper analysis on the data. Beast naturally integrates with Spark and can efficiently spatially partition big data. Beast can also load, decompress, convert, compress, and write files in parallel. It can be used to help generate visualizations as well which is an integral part of our analysis. SparkSQL allows us to use both DataFrames and SQL queries in our code which helps us to process and get the data we need.

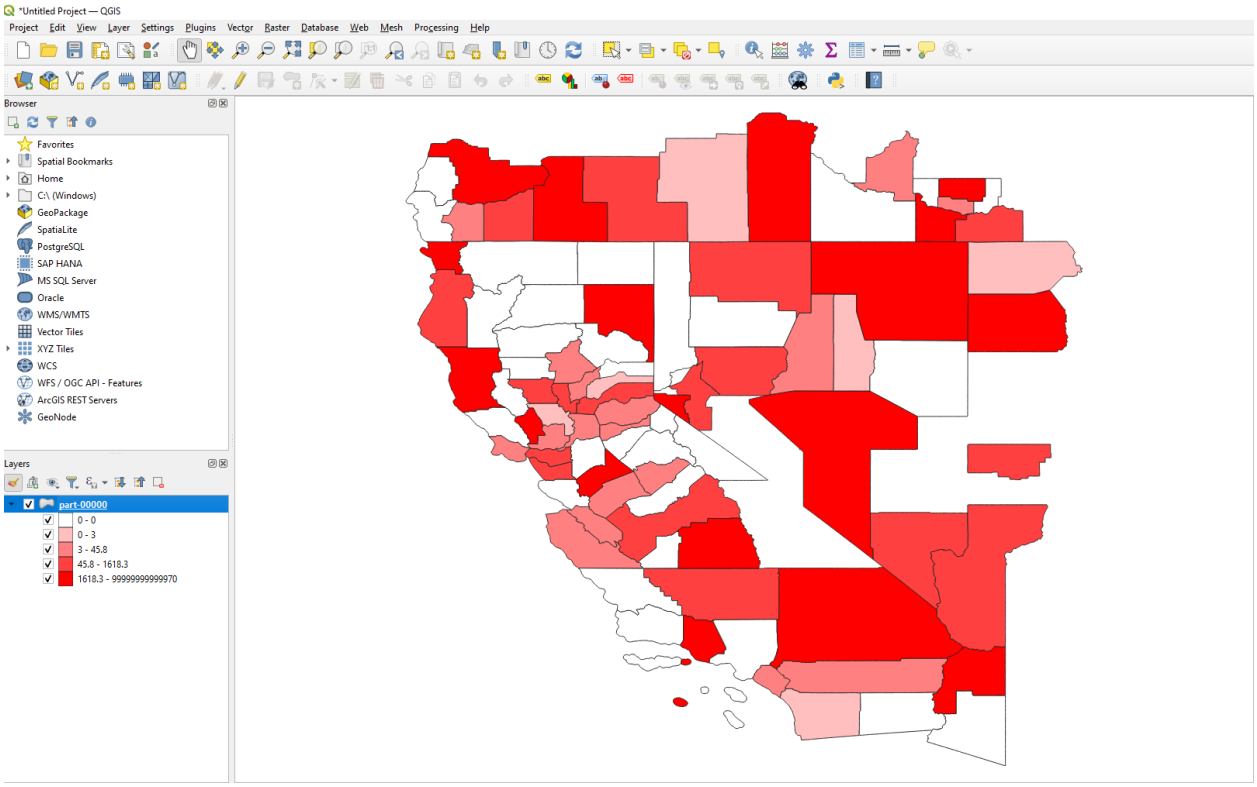
Task 1:

The parquet format is helpful for this project because we are working with large files and the parquet format is able to store these files using a lot less space than CSV format. As can be seen from the table below, the difference between the size of the CSV file and the parquet file for the same data is pretty big. This smaller size makes the file easier to work with because it's faster to scan through the file and find the data that we need. When doing the next two tasks with the visualizations, it's better to get the input file in parquet format because it takes less time to scan through the file and perform analytics.

Dataset	CSV Size	Parquet Size
1,000	1.3 MB	35 KB
10,000	13.2 MB	241 KB
100,000	132.3 MB	2.322 MB

Task 2:

Results for the 10K file with date range 01/01/2016 - 12/31/2017:



Task 3:

Results for the fire intensity by year-month for the wildfires 100k file in Riverside county:
Loaded in the parquet wildfires dataset. Found the county's GEOID by filtering counties using STATEFP and the user's input. Then selected all the wildfires related to that county and computed the total fire intensity and grouped the results by year and month.