Spark Analysis of NASA Data Report

Hao Wang 010702263

Spark has become the first choice for big data analyst especially from this year due to its in memory calculation. It was estimated that Spark is 100 times faster than Hadoop for large scale dataset. In this practice, I analyzed the NASA public dataset using Spark Service on the IBM bluemix cloud. I used Python and Notebook as the tool to filter the data, do the Map Reduce, and then visualized the data using pySpark API. The following are the Steps for the Spark analysis.

**Instructions**

1. Data Retrieval
   The dataset is the NASA public dataset, and can be accessed at https://data.nasa.gov/view/scmi-np9r
2. Data preprocessing
   Since each entry of the data is separated by two different lines. For easier Spark analysis, we have to process it to combine the data belong to each entry. I used python script to process the data, and the processed data is saved in the " cleaneddata.csv ". Please use this dataset for the NOTEBOOK operation. The script to process the RAW data is also in the folder for your reference.
3. Set up Spark Service on the IBM bluemix cloud.
4. NoteBook Setup
   I created the instance NoteBook (python) for data analysis.
   I uploaded the complete NoteBook in the folder, which name is NASASpark.ipynb .
5. Run the NoteBook to demo the analysis
   After loaded the cleaned dataset into the objectstore and NoteBook, you should be able to run the program in the NoteBook, and get the analysis results. The Notebook are also documented, so you can understand what each step does, and get the results.

**Question for NASA data**

How many NASA facilities are Active in each US States?

I know several very famous NASA center in the U.S. such as Kennedy center in Orlando, where I visited once before. But I really have no idea about how many NASA centers are, and which states are they located, which state has the most NASA facilities?

Those questions are the one I really want to know the answer.

**Answer from the data**

The data talks. After analyzed the data by Spark, we can see surprisingly Alabama has the most NASA center. I was thinking it's Florida, where I just came from. BUT…..Top 3 are Alabama, Virginia, and California. The number is also surprising. There are more than 100 NASA facilities running in Alabama. Most surprisingly, the number in FL is very low. I have no idea why Kennedy is so famous whereas there are only few active facilities are running there. But this is the data, where the truth is.