Cropland Fire Emissions Globally Using Hadoop

Authors: Alan To, Christopher Valdepena,William Gomez, Alex Garcia, David Ybarra
Department of Information Systems, California State University Los Angeles
CIS4560-01 Introduction to Big Data
ato3@calstatela.edu, wgomez13@calstatela.edu, agarci275@calstatela.edu,
dybarra8@calstatela.edu, cvaldep3@calstatela.edu

Abstract: The paper explains the methodology used for analyzing and manipulating data related to croplands fire emissions on a global level. The major focus of the project is to understand how important tools such as Hadoop are useful to analyze data and to obtain a conclusion related to important events that are happening in the present. In addition, Excel has been used to obtain visual representation of the data. The dataset is about 2.6 GB and it has been retrieved from Kaggle.

1. Introduction

Hadoop and hive have been used as tools to keep and process the Cropland fire emissions dataset. The dataset consists of the amount of fire emissions produced in different cropland assets from companies around the world.

The dataset has been chosen because nowadays, the greenhouse effect is a real important problem to be considered. By watching the emissions produced by cropland fires we can see which country has the most emissions so we can bring awareness and encourage better measures in their cropland assets to prevent fires in their plantations.

2. Related work

We were able to locate the data and summarize it using the cleaned data that was provided in the link that was found in the free site Kaggle. The dataset provided the numbers and percentages of the amount of cropland fires around the world. All the numbers and dataset it gave us allowed us to create geography maps, bar graphs, and descriptive analytics using tools that are offered on excel. With those tools we found the countries who are emitting the most pollution on earth throughout the years, even 2022. This gives us enough information to confidently gather the data that we were observing and pinpointing the exact information that we were looking for, plus, providing more insight than we already had with the graphs. One article demonstrated the monthly emissions, estimating burn areas and contributions of different fire types. Randerson conducted the study on Global Fire Emissions [2] using data from fires to map out the area and provide estimates to the emissions usage. Comparing the monthly emissions usage to higher temporal resolutions, including monthly biosphere fluxes. Another article focuses on CO2 and Greenhouse Gas Emissions in our world [3]. Hannah conducted the study to determine how exactly CO2 and greenhouse gas emissions are negatively affecting our environment. Using the time period starting in 1850 - 2019 along with the median average temperature change, we can see that over the last few decades global temperatures have been dramatically increasing. The method of data analysis from these two articles compared to our data analysis is very similar. Using geo-mapping features to show changes within a time period, as well as gathering data to determine areas with the most impact.

3. Specifications

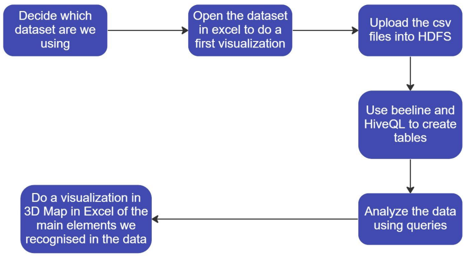
The graphs include descriptive analytics that revolve around the cropland fire and greenhouse gas emissions. Kaggle freely allows users to post data they found and makes them post sources to verify it. This allows people like our team to look through it and improve upon it by using graphs, tables, and maps to visualize it and bringing it to life instead of just seeing numbers. Bringing greater awareness. The dataset is around 2.6 GB's worth of numbers, percentages, and countries. Figure 1 shows off our specifications of the dataset using Hadoop.

Figure 1 HW Specification

Number of Nodes	5
Memory Size	60 Gb
CPU	8
Speed	1995.309

4. Implementation Flowchart

First, we had to decide which dataset we were about to use for the project. Initially, we were planning to analyze a Covid-19 dataset. However, we found out that the dataset was not really useful for the project we were planning to do. Once we realized that, we looked in Kaggle for the dataset of the project. Once we have obtained the data, we opened it in excel to do a first visualization of it. After that, we uploaded the data in csv format into HDFS. In addition to that, we use beeline and Hive to create tables and analyze the data using queries. Finally, we did a visualization using the 3D map tool from excel along with the bar graph tool.


Decide which dataset are we using
Open the dataset in excel to do a first visualization
Upload the csv files into HDFS
Use beeline and HiveQL to create tables
Analyze the data using queries
Do a visualization in 3D Map in Excel of the main elements we recognised in the data

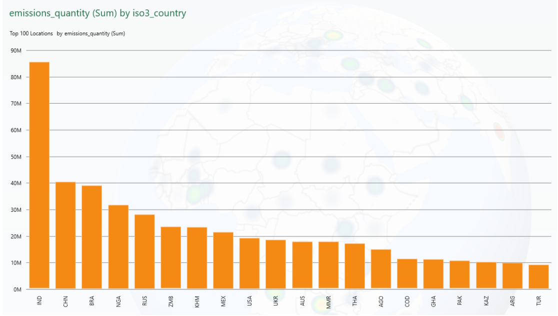
5. Data Cleaning

After we successfully downloaded our dataset we uploaded it first to the linux server then moved it to hadoop which is where the data cleaning took place. Because of how large the file is, cleaning the data in excel isn't possible as it can only load up to 10 million rows locally. We proceeded to use Hive which was built on top of Apache Hadoop. Hive allows users to read, write, and manage petabytes of data using SQL commands and was designed specifically for manipulating data. Using our own database in hadoop we ran commands that created and described tables using the emission csv file. Which was then used to visualize the data in the final steps. This was used to keep the file size fairly small to keep processes efficient. For the most part our data was fairly clean as it came from an incredibly reputable source that hires professionals to maintain their data. Data cleaning is a form of data management which helps us make sense of the data at hand, with messy data that includes anomalies, empty/null entries, errors, and inaccurate data, will prevent us from creating effective diagrams and accurate diagrams.

6. Analysis and Visualization

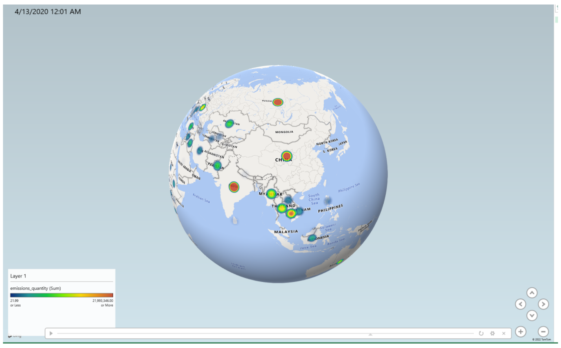
Once our team was able to verify all the data we received, we were able to get ready and present it in a way that will bring it to life. It leaves a huge impact when you see how costly these numbers are. We used maps that will primarily show who is emitting the most on earth and who is spewing the least.

Figure 1 Bar Graph



This first bar graph is indicating to us precisely the quantity of emissions throughout every country in the world. As shown, India is on top by a colossal amount and Turkey is on the far right showing off their low emissions. This is extremely vital information in order to progress in climate change and find solutions in order to intercept the damage and slow it down.

Figure 2 Geomap



For our second visualization, a map was created in order to get a clear view of all the damage caused. We are able to accurately concur just how much gas is in the air at the bottom left corner of the map, with blue being the lowest, neon green being in the middle, and highest being red. This is easy to see all in one place all at once.

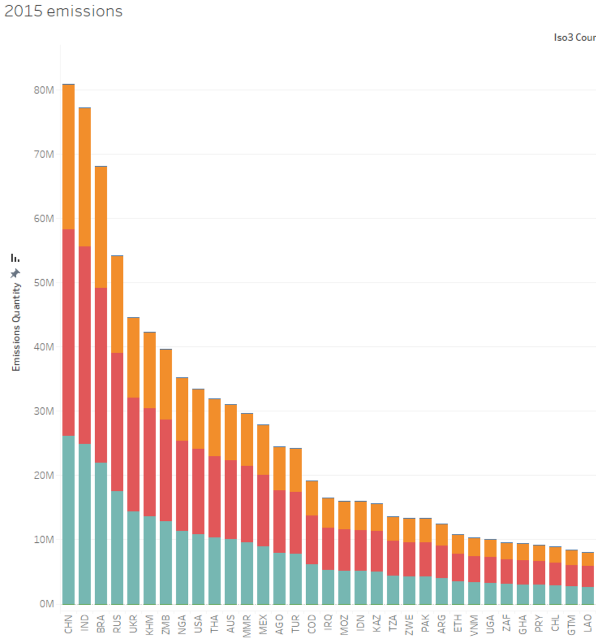


Figure 3

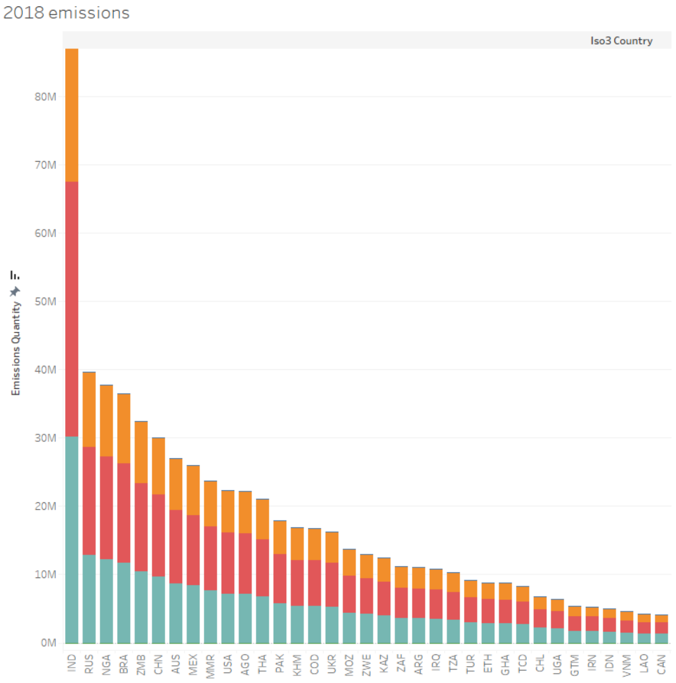


Figure 4

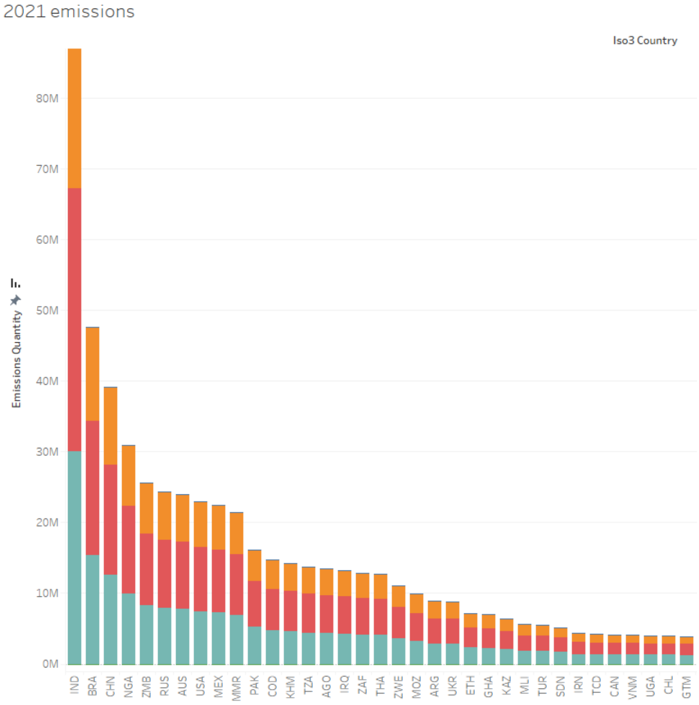


Figure 5

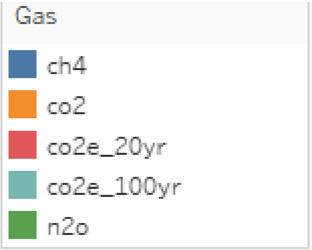


Figure 6

The snapshots above represent the total amount of emissions recorded by metric ton and country. The years include 2015, 2018, and 2021 and show the progressions with trend of countries emissions throughout time. The top three countries that consistently emit the most gasses are India, China, and Brazil. For the most part they are growing and third world countries that are responsible for a lot of the manufacturing that goes on in the world. It's incredibly hard to enforce emissions on these countries as they are ridden with problems. Countries such as India don't focus as heavily on the environment as developed countries such as the United States do due to their lack of infrastructure and enforcement. The multicolored bars in the bars represent the different types of gasses being emitted for that year. Figure 6 which is the legend which signifies the types of gasses in a particular bar.

Conclusion

To summarize the information given, we were able to cover major topics and pinpoint who exactly is emitting these toxic gasses in the air, pinpoint the exact amounts spewed in recent years, and just how much other countries have stopped emitting. From the amount of data that we were able to observe, we were able to extract all this information to show just how awful it has become over the years. More data allows us to find more important descriptive information to overcome such barriers.

References

[1] Greenhouse Gas Emissions (December 17, 2022)Kaggle. https://www.kaggle.com/datasets/michaelbryantds/greenhouse-gas-emissions-dataset

[2] Randerson, J.T., G.R. van der Werf, L. Giglio, G.J. Collatz, and P.S. Kasibhatla. 2018. Global Fire Emissions Database, Version 4.1 (GFEDv4). ORNL DAAC, Oak Ridge, Tennessee,USA. https://doi.org/10.3334/ORNLDAAC/1293

[3] Hannah Ritchie, Max Roser and Pablo Rosado (2020) - "CO_2 and Greenhouse Gas Emissions". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions'