Sonal Bisla, Vaidehi Atodaria, Lakshmi Sravanthi Naupada under supervision of Prof.Roy Kucukates

TORONTO MeTROPOLITAN UNIVERSITY  MSc. in Data Science & Analytics

**Big data PROJECT ANALYSIS ON Co2 emission** Special Thanks to roy kucukates

Table of Contents

[**1. Introduction** 2](#_Toc133589998)

[**1.1 Problem Definition** 2](#_Toc133589999)

[**2. Dataset Description** 3](#_Toc133590000)

[**2.1 Attribute Description** 3](#_Toc133590001)

[**2.2 Statistics of data** 4](#_Toc133590002)

[**3. Solution Description** 5](#_Toc133590003)

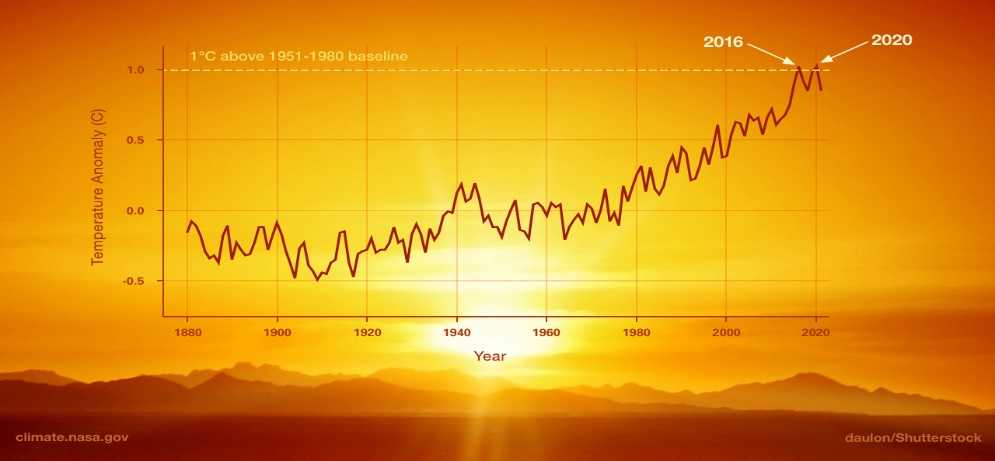
[**4. Insights** 22](#_Toc133590004)

[**5. Future Work** 27](#_Toc133590005)

[**6. References** 27](#_Toc133590006)

# 

# **1. Introduction**



Global warming is the long-term heating of earth’s surface due to raise in temperatures. In the 19th century there is a steep raise in greenhouse gases due to burning of fossil fuels and petroleum2. Each year there is raise in human activities and advancements in technology industries leading to raise of carbon dioxide especially in developed and developing countries. Among these green house gases carbon dioxide is the most dangerous gas leading to ocean acidification. Although many countries are taking steps to reduce carbon footprint there is still more to do. So, it is important to analyze the scenarios responsible for saving the world. This project is all about analyzing carbon dioxide emission factors and its impact on the world using big data tools.

## **1.1 Problem Definition**

The project problem definition is:

* To analyze different countries for the carbon dioxide emission and knowing the top countries with highest emission.
* To analyze the different sources responsible for the co2 emission.
* To analyze if there is any relation between the co2 emission and the life expectancy for the recent years.

# **2. Dataset Description**

The dataset that we selected is from world development indicators (DATA WORLD BANK WEBSITE). The variables are 52 countries and the carbon dioxide emission levels for the years between 1990 and 2019. Based on phase 1 analysis, further variables are taken up for future analysis. Detailed description is as follows.

**Dataset link:** [**https://databank.worldbank.org/source/world-development-indicators**](https://databank.worldbank.org/source/world-development-indicators)

## **2.1 Attribute Description**

**Phase 1:** Cleaning and analysing 52 countries for the levels of emission and knowing the top 5 countries with highest emission of carbon dioxide.

* + **Country name**: This column has all the 52 countries that are taken for analysis
  + **Country Code**: Code for specific country.
  + **Series code**: It is a related abbreviated column for the factors. It has short descriptions for the factors
  + **Series name**: This is our factors column that has all the factors which we have taken into consideration for our analysis
  + **Years :1990 to 2019 (CO2 emission):** It’s the overall total carbon dioxide emission in Kt (Kilotons) for the respective years.

**Phase 2***:* For the top country with highest emission looking at the sources responsible and predicting which source is highly responsible.

* + **Co2 emission from Gaseous fuel:** co2 emitted from burning of natural gas or propane.
  + **Co2 emission from Liquid fuel:** It is released by burning of liquid fuels such as gasoline.
  + **Co2 emission from Solid fuel:** The emission from solid state is from different sources like burning of fossil fuels, production and deforestation.
  + **Years selected**: 1980,2014,2018 & **Country:** USA

**Phase 3:** To detect the patterns of relation between the co2 emission and the life expectancy for the recent years.

* **Country\_code**: It contains the country code for instance here we have considered the top 2 countries on our file i.e., CHN and USA
* **Country\_name**: It contains the country names i.e., China and United States
* **Series\_code**: It is an abbreviated column for the factors. It has short descriptions for the factors
* **Series\_name**: This is our factors column that has all the factors which we have taken into consideration for our analysis
* **y\_2014, y\_2016, y\_2019**

## **2.2 Statistics of data**

Below graph shows there is a steep raise of co2 emission over the period.

Chart, line chart

Description automatically generated

# **3. Solution Description**

**PHASE 1:** The phase 1 analysis is about cleaning the dataset that involves correcting the column name, removing null values, removing any duplicate values. Once we have our clean data, we worked on the summary statistics of all the columns to determine the mean and min max values.

At this stage we will be working on analysing the trends in emission from year 1990 to 2019 and determine the top countries for emission.

**Tools used**: Tools used for this analysis are **HDFS, PySpark** and **Tableau** and parallel analysis using **Hive.**

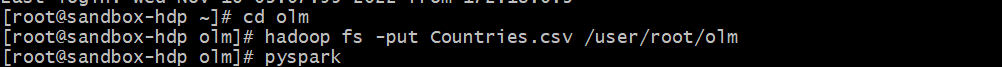
**How they are used:** HDFSsystem is used to store the data set. We used Pyspark for data preparation, cleaning and analysis as shown in the snippets below. And use Tableau for the visualisation graphs

**WHY:** Initially we started working on the data cleaning part using HIVE and Pyspark. Further analysis showed that Pyspark is giving quick results and cleaning using hive is a little challenging especially while removing the rows that are unnecessary. And Pyspark can be implemented using multiple languages Python, R, Scala and java. And the computational speed is high in Pyspark compared to hive and while working with more variables Pyspark is giving quick results.

**Code snippets:**

DATA PREPERATION:

Using HDFS as storage container



This data frame df shows the initial dataset with the variables of 52 countries and the co2 emission for all the years 1990 to 2019.

Graphical user interface

Description automatically generated with medium confidence

This schema shows the datatype of the variables

A screen shot of a computer

Description automatically generated with medium confidence

To get the count of variables we use the following code





**CLEANING:**

Correcting the column name to correct format

A screen shot of a computer

Description automatically generated with low confidence

Then we have to change the type casting for the variables changing the string type to integer for all the years.

Graphical user interface, text

Description automatically generated

**ANALYSIS:**



Text

Description automatically generated

The summary statistics of the variables and the years are shown below

Graphical user interface, text

Description automatically generated

Here we see the average value for emission is highest for the year 2019 compared to all other years.

Text

Description automatically generated

Text

Description automatically generated with medium confidence

**Phase 2:** From the phase1 analysis we have found out that China and USA are the most affected country by co2 emission. In phase2, we will analyze which factors are responsible for c02 emission. Although, there are so many factors responsible for it, we have chosen the below sources

co2 emission from gaseous fuel,

C02 emission from liquid fuel,

Co2 emission from solid fuel

For the years 1980,2014, 2016.

**Tools used:** Data Analysis tools used are **Apache hive cluster** on **Databricks cloud** platform, and visualization tool **Ambari** is used for detailed data analysis for yearly records

**Why used:** Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets.

**How:** Hive organizes tables into partitions based on partition keys for grouping similar data together.

The partitions are further categorized into buckets based on the hash function of a column in the table. These buckets are stored as a file in the partition directory.

It makes MapReduce programming easier because you don't have to know and write lengthy Java code

Each table in the hive can have one or more partition keys to identify a particular partition. Using partition, it is easy to do queries on slices of the data.

**Hive Partitioning Advantages:**

Partitioning in Hive distributes execution load horizontally.

In partition faster execution of queries with the low volume of data takes place. For example, the search population from USA returns very fast instead of searching every country.

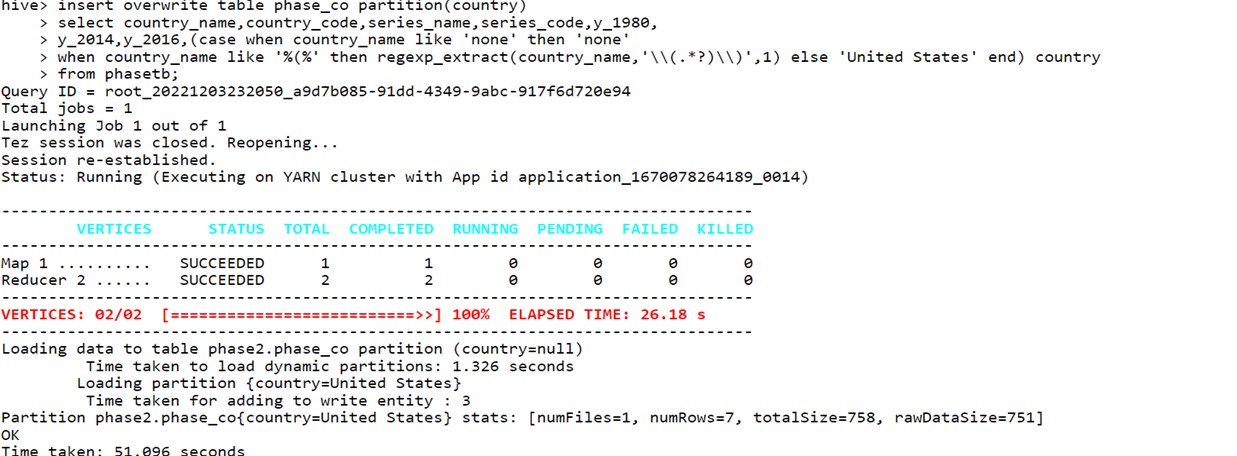
**Code snippets:**

**Non partition table and Partition(****country) table:**

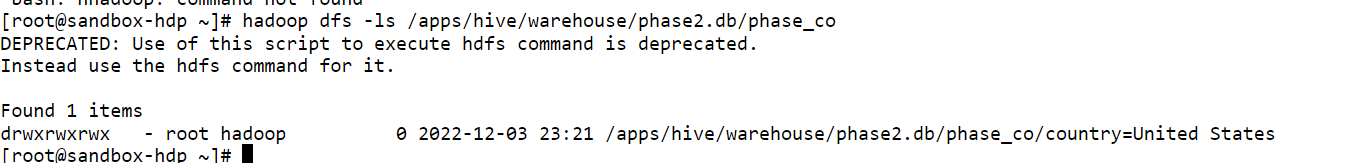
Text, letter

Description automatically generated

**Insert into table:**



**Partition on** **Hadoop cluster:**



A picture containing chart

Description automatically generated

**Database:**

Graphical user interface, text, application, email

Description automatically generated

**Phase 3**: In continuation of phase 1 and phase 2, here we are using the already cleaned data file. For phase 3, we have used elastic search tool.

**Tools used**: **Elastic Search, Kibana**

Elasticsearch is a distributed, free and open search and analytics engine for all types of data, including textual, numerical, geospatial, structured, and unstructured. Elasticsearch is built on Apache Lucene and was first released in 2010 by Elasticsearch N.V. (now known as Elastic). Known for its simple REST APIs, distributed nature, speed, and scalability, Elasticsearch is the central component of the Elastic Stack, a set of freeand open tools for data ingestion, enrichment, storage, analysis, and visualization. Commonly referred to as the ELK Stack (after Elasticsearch, Logstash, and Kibana), the Elastic Stack now includes a rich collection of lightweight shipping agents known as Beats for sending data to Elasticsearch.

**Why used:** It runs perfectly fine on any machine or in a cluster containing hundreds of nodes, and the experience is almost identical.

One can perform and combine various kind of searches irrespective of their data type which included structured, unstructured, geo and metrics data type.

**Fast performance**: By using distributed inverted indices, Elasticsearch quickly finds the best matches for your full-text searches from even very large data sets.

It also provides aggregations which can explore trends and patterns of data.

**Kibana:** Kibana is a data visualization and exploration tool. It offers powerful and easy-to-use features such as histograms, line graphs, pie charts, heat maps, and built-in geospatial support.

**How:** Create a free account and the first deployment with the deployment name ds8003 on the website <https://www.elastic.co>

Graphical user interface, application

Description automatically generated

Once we have created the deployment, we enter the elastic search page. We now will upload the csv file.

Graphical user interface, application

Description automatically generated

Once this part is done, we do the further steps as follows:

**Code snippets:**

Upload our file name life\_expectancy.csv

Graphical user interface, application, Word

Description automatically generated

Now here we can see our file was uploaded and it shows the stats which includes all the 7 column names/fields with the distinct values and distributions along with the min, max and median for the year columns respectively beside them.

A screenshot of a computer

Description automatically generated with medium confidence

Note: In Elasticsearch, an index (plural: indices) contains a schema and can have one or more shards and replicas. An Elasticsearch index is divided into shards and each shard is an instance of a Lucene index. Indices are used to store the documents in dedicated data structures corresponding to the data type of fields.

Here we have created the index with the name expectancy:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application, website

Description automatically generated

This is how our index looks below. Index health is green and is connected to the API. Ingestion status is also showing connected.

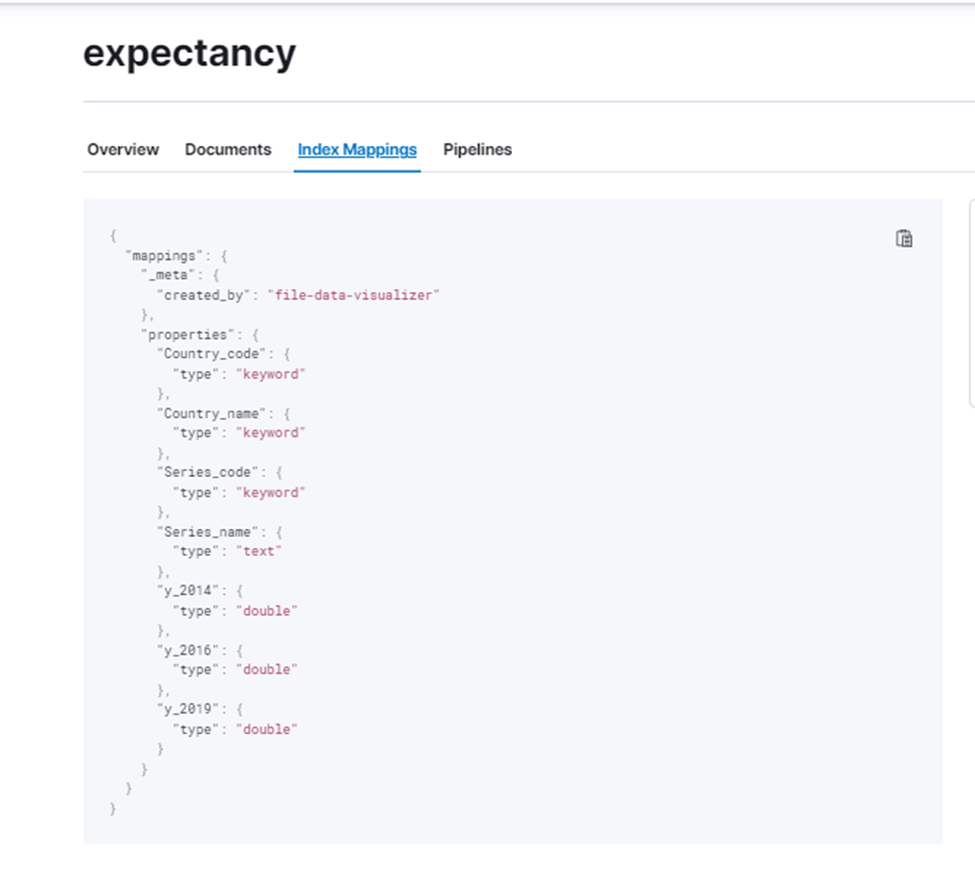
Table

Description automatically generated

This is how our expectancy API looks like. POST command is used to insert the dataGraphical user interface, text, application, email

Description automatically generated

**Index mapping:**



Transforms enable you to convert existing Elasticsearch indices into summarized indices, which provide opportunities for new insights and analytics. For example, you can use transforms to pivot your data into entity-centric indices that summarize the behavior of users or sessions or other entities in your data.

Graphical user interface, text, application, email

Description automatically generated

By clicking on a create a transform tab we will land up to a page where we can transform our data as per our requirements that we need for the better and clear visualizations further

Graphical user interface, application, email

Description automatically generated

**Select index:**

Graphical user interface, text, application, email

Description automatically generated

This gave us a pivot and group by options. Also, as we can see we are able to see our columns data without creating any tables. It generated histogram charts as well for each column.

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

Now in transforms tab it also gave us an option where we can get a better result by using group by and aggregations operations for any desired column we want.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, table

Description automatically generated

# **4. Insights**

**Phase 1:** After cleaning and analyzing each of the variables and performing exploratory analysis we have found two insights from phase 1 are as follows:

* **There is a steep raise in co2 emission from the year 1990 to 2021**
* **Among all countries the top countries with highest emission are USA and China.**

The visualisations using **Tableau** are below:

Co2 emission for the countries for the year 2019

Chart, bar chart, funnel chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Phase 2:** INSIGHTS FROM THIS PHASE 2 ARE AS FOLLOWS:

* **Among all sources, gaseous source is most responsible**

Data showing the factors with the max co2 emission out of liquid, gaseous and Solid for the years\_1980,2014 and 2016

Graphical user interface

Description automatically generated

Data showing the gaseous factor with the max Co2 emission for the year\_1980 in China and USA

Graphical user interface

Description automatically generated

Graphical user interface, website

Description automatically generated

**Phase 3:** In continuation of phase 1 and phase 2, here we are using the already cleaned data file. For phase 3, we have used elastic search tool.

Kibana is the the most effective elastic search interface for discovering data insights and performing active management of the health of the Elastic Stack. As we can see, it gives an option for Analytics. We clicked on that.

Graphical user interface, application

Description automatically generated

Inbuilt options for the visualization like dashboards, maps, graphs:

Graphical user interface, application

Description automatically generated

This is the bar graph we generated using inbuild Kibana dashboards that depicts us the below result.

Result: SH.XPD.CHEX.PP.CD (Current health expenditure per capita, PPP, international $) for US is the most dominating factor which is directly related to improving the life expectancy. While the second most responsible factor to improve the life expectancy is SP.POP.SCIE. RD. P6 (Researchers in R&D, per million people)

*y\_2019, series\_code and country\_name columns being used in this.*

Chart

Description automatically generated with medium confidence

This is another visualization below using pie chart for the same result that we saw above regarding the contribution of the factors for improving the life expectancy.

*y\_2019 and series\_code columns being used*

Chart, pie chart

Description automatically generated

# **5. Future Work**

* We will be focusing on the data for the years 2020 and 2022 and using more big data tools and cloud computing tools as well.
* We will be considering the topmost countries only which are USA and China in order to focus on the most relevant factors to reduce the greenhouse effect.
* Apart from the Co2 emission, we will be working on other factors which are directly or indirectly affecting the world and we would like to consider the pandemic phase.

# **6. References**

* *1.*[*https://climate.nasa.gov/global-warming-vs-climate-change/*](https://climate.nasa.gov/global-warming-vs-climate-change/)
* [*https://www.encyclopedia.com/environment/energy-government-and-defense-magazines/carbon-dioxide-co2-emissions*](https://www.encyclopedia.com/environment/energy-government-and-defense-magazines/carbon-dioxide-co2-emissions)
* [*https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide*](https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide)
* [*https://medium.com/bazaar-tech/apache-spark-data-cleaning-using-pyspark-for-beginners-eeeced351ebf*](https://medium.com/bazaar-tech/apache-spark-data-cleaning-using-pyspark-for-beginners-eeeced351ebf)
* [*https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262802#:~:text=Carbon%20emissions%20have%20a%20significantly,decreases%20life%20expectancy%20by%200.012%25*](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262802#:~:text=Carbon%20emissions%20have%20a%20significantly,decreases%20life%20expectancy%20by%200.012%25)
* *blogs.sap.com*
* *Lecture\_10 & Lab 8\_Elastic\_Search*
* [*https://www.elastic.co/elasticsearch*](https://www.elastic.co/elasticsearch)
* [*https://severalnines.com/blog/what-is-elasticsearch-and-why-use-it/*](https://severalnines.com/blog/what-is-elasticsearch-and-why-use-it/)
* [*https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262802*](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262802)