Analyzing climate changes impacts using big data Hadoop

Silvana Greca, Ingrid Shehi and Jonuz Nuhi

University of Tirana, Faculty of Natural Sciences, Department of Informatics, Bulevardi Zogu I, Tirana, Albania

Abstract

This paper aims to analyze climate change data using Hive and Hadoop, two big data processing frameworks. We collected data from various sources and used Hive to store and manage the data, and Hadoop to process it. By using these tools, we were able to perform complex queries and analysis on large datasets with ease. This paper also used Super Set, a data visualization tool, to create interactive dashboards that display the results of our analysis. The dashboards help users to explore the data and gain insights into climate change trends. Our findings show that the temperature in the city of Durrës has increased by 1.1°C since the pre-industrial era. This paper demonstrates the usefulness of big data processing tools for analyzing climate change data and provides valuable insights into the impact of global warming on our planet.

Keywords

Climate change, big data processing, Hadoop, Hive, Data analysis, Data visualization, Superset

1. Introduction

In recent years, the rapid growth of data has become a ubiquitous challenge in various domains, including climate science. The scale and complexity of climate data pose significant challenges to traditional data processing and analysis techniques. However, big data processing tools, such as Hadoop and Hive, have emerged as promising solutions to handle large-scale data processing and analysis.

Hadoop is a distributed computing framework that can store and process large datasets in parallel, while Hive is a data warehousing tool that provides a SQL-like interface for managing and querying large datasets stored in Hadoop.

These tools have already been used in various domains, including climate science, to process and analyze large datasets efficiently. In particular, Hadoop's distributed processing capabilities make it well-suited for analyzing large climate datasets.

In this paper, we present a study that leverages Hadoop and Hive to analyze climate change data. Our study collects data from various sources, including temperature measurement data, and uses Hadoop's distributed processing capabilities to store and process the data in parallel. We then use Hive to manage and analyze the data, including running complex queries and creating visualizations. Finally, we use Superset, a data visualization tool, to create interactive

Proceedings of RTA-CSIT 2023, April 26–27, 2023
Tirana, Albania

EMAIL: silvana.greca@fshn.edu.al(S.Greca);
ingrid.shehi@fshnstudent.info(I.Shehi);
jonuz.nuhi@fshnstudent.info(J.Nuhi)
ORCID: 0000-0002-2435-233X (S.Greca)

dashboards that provide insights into climate change trends.

The objective of this paper is to demonstrate the usefulness of big data processing tools for analyzing climate change data and to provide valuable insights into the impact of global warming on our planet. The paper is structured as follows: in the next section, we review related work on using big data processing tools for climate science. We then describe the methodology used in our study, including data collection, processing, and analysis. Next, we present our results and discuss the implications of our findings. Finally, we conclude the paper with a summary of our contributions and future research directions.

By using big data processing tools, we can gain deeper insights into climate change patterns and trends, which can help inform policy decisions and mitigation strategies. We hope that this study will inspire further research on the use of big data processing tools for climate science and contribute to the ongoing efforts to address the urgent issue of climate change.

2. Literature overview

The use of big data processing tools such as Hadoop and Hive for analyzing climate change impacts has been gaining increasing attention in recent years. In this section, we will discuss some of the notable studies that have leveraged these tools to analyze climate data and investigate the impacts of global warming on our planet.

One example of such studies is the work of Hossein Hassani, Xu Huang, Emmanuel Silva. [10], who used Big Data to analyze climate data collected from multiple sources. Their study focused on identifying and characterizing heatwaves and their impacts on human health. By leveraging the processing power of Hadoop, they were able to efficiently process large volumes of data and identify significant trends and patterns. Their results demonstrated the potential of big data processing tools for analyzing the complex and dynamic impacts of climate change.

Another notable study is the work of Hua-Dong Guo, Li Zhang, Lan-Wei Zhu. [11], who used Big Data. Their study reviews the advances of climate change studies based on Earth observation big data and provides examples of case studies that utilize Earth observation big data in climate change research, such as synchronous satellite–aerial–ground observation experiments, which provide extremely large and abundant datasets. With the era of global environment change dawning, Earth observation big data will underpin the Future.

In a similar vein, the work of Thanos Papadopoulos, M.E. Balta [12] leveraged big data analytics to analyze climate data for various challenges. Their research addresses the impact of climate change on businesses, operations, and supply chains by identifying and discussing how these challenges and opportunities can be better pursued. Their opinion paper proposes ideas about future research on BDA and climate change could follow to facilitate the transition to a sustainable future.

Overall, these studies highlight the potential of big data processing tools for analyzing climate change impacts and providing valuable insights into the complex and dynamic relationships between climate and the environment. By leveraging the processing power of tools for Big Data, researchers can efficiently process and analyze large volumes of data, which can inform policy decisions and mitigation strategies for addressing the urgent issue of climate change.

3. Hadoop and it’s functionality

Hadoop is an open-source framework that is used to efficiently store large amount of data from gigabytes to petabytes. Instead of using one large computer to store and process the data, Hadoop uses a set of clusters to analyze massive datasets [1]. The processing of datasets is done through HDFS. HDFS enables the rapid transfer of data between computer nodes. When the data is ready to be processed, it is sent to Map Reduce and split into smaller datasets. After the data is divided each of those datasets get a mapping function. Then the shuffling start which just send the data from the mapper to the reducers. The Reducing stage is about summarizing the effects of the previous stages and reducing them to a small set of values [2]. The use of Hadoop is very cost-effective because the commodity hardware is very cheap. HDFS besides storing large amount of data it can store them in different formats. Since Hadoop process the data in parallel the speed of our output is very fast and also Hadoop creates a duplicate of the data in all the nodes so if a node crash or burn, we have a copy of it so data is not lost as it is shown in the figure 1.

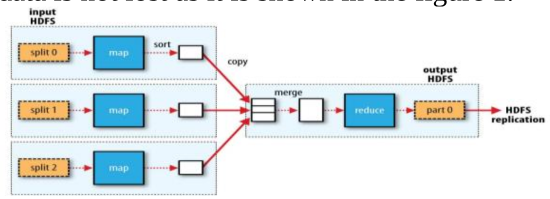


Figure 1: MapReduce data flow with a single reduce task

4. The Hive

Hive, an open-source data warehousing solution tool built on top of Hadoop Distributed File System (HDFS) that allows users to query and analyze large datasets using a SQL-like language called HiveQL [3].

The connection between Hive and Hadoop is established through the use of a driver program that acts as a mediator between Hive and the Hadoop cluster as is shown in the figure 2.

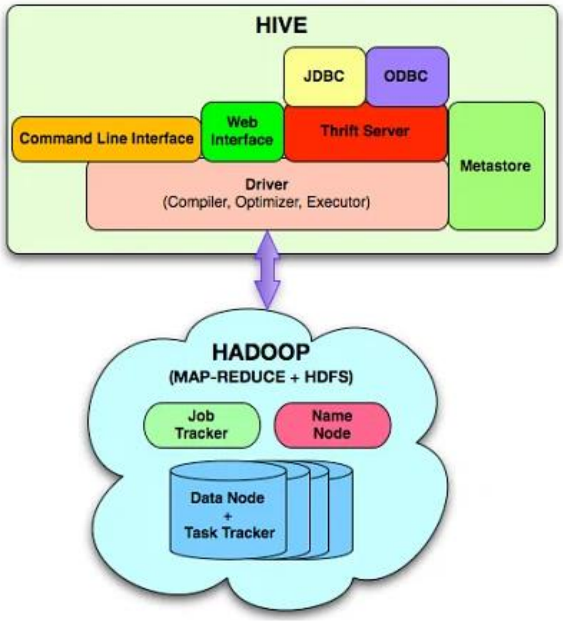


Figure 2: Connection with Hadoop

The driver program is responsible for translating HiveQL queries into MapReduce jobs, submitting them to the Hadoop cluster for processing, and returning the results back to Hive. Results are aggregated and returned to the user as the output of the Hive query [4].

4.1 Hive components

The components of hive are shown in the figure 3.

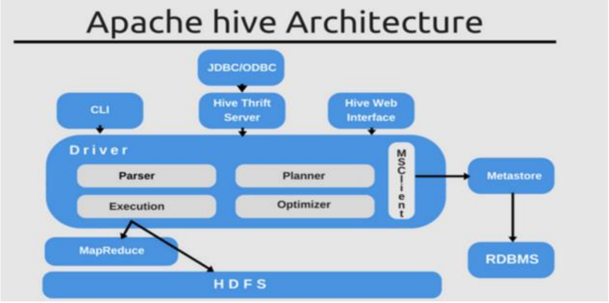


Figure 3: Hive Components

- Hive Client - Hive allows writing applications in various languages. It supports different types of clients such as: Thrift Server, JDBC Driver, ODBC Driver.
- Metastore - The Metastore is a database that stores metadata about the tables and partitions in Hive. It keeps track of the schema, location, and other properties of the data stored in Hive. The Metastore supports multiple databases, including MySQL, PostgreSQL, and Derby.
- HiveQL - HiveQL is a SQL-like language that is used to query data stored in Hive. It provides a familiar syntax for SQL users and supports many SQL operations, including joins, aggregations, and subqueries.
- Driver - The Driver is responsible for receiving queries from the user, generating an execution plan, and coordinating the execution of the plan with the other components of Hive.
- Compiler - The Compiler takes the query plan generated by the Driver and generates a series of MapReduce or Tez jobs to be executed on the Hadoop cluster. The Compiler is responsible for optimizing the query plan to improve performance.
- Execution Engine - The Execution Engine is responsible for executing the MapReduce or Tez jobs generated by the Compiler. The Execution Engine runs on the

Hadoop cluster and processes the data stored in HDFS.

- Hive Server - The Hive Server is a server process that exposes a Thrift interface, allowing clients to connect to Hive and execute queries using various programming languages, such as Java, Python, and R. Hive Server 2 (HS2) is the preferred version of the server [5].

5. Using Superset to visualize the data

Data is stored in tabular mode. To help viewers to understand the data, a visualization tool is needed. Apache Superset is a tool from ASF (Apache Software Foundation) which provides a way to data visualization and exploration from simple line to highly detailed geospatial charts. It allows integration with most relational databases. Hive is used as a data source from superset.

5.1 Superset and Hive connection to visualize data

The chart in the figure 4 shows the flow of our work. First, we configured Hadoop in our environment and after that we configured Hive, a tool to work with our data. We extracted the data in Hive and after that we configured Superset. After Superset was successfully configured, we connected it with Hive so it could visualize our data and Superset also uses queries so we can filter the data we want to visualize.

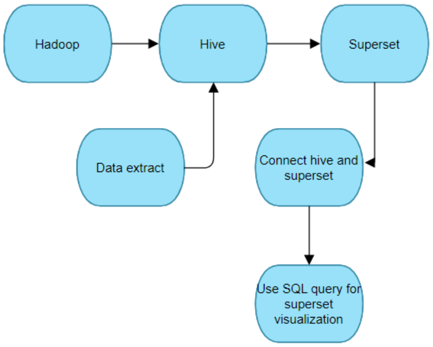

Hadoop
Hive
Superset
Data extract
Connect hive and superset
Use SQL query for superset visualization

Figure 4: Activity and flow chart

The recommended way to connect Superset with hive is by using the pyhive library. The pyhive library is a python interface for interacting with hive databases. For connection to a Hive database, you will need to configure a new data source in the Superset interface and specify the connection details for the Hive database. Once the connection is established, Superset can execute HiveQL queries on the data stored in Hive. Superset can't store the data in itself it gets the data from the existing database. When you create a new chart or dashboard in Superset, you can write a SQL query to pull the data from Hive, or use the Superset Query Builder interface to generate the query for you [6].

5.2 Advantages of Superset

As a visualization tool Superset stands out because using it the user have total access over the data. Using Superset you can add users to your database, provide access to them and track their behavior. To create a visualization of the data we can create query in Superset and the queries are created the same way as in any SQL based database. Since Superset is a visualization tool it is created for non-programmers so it can be used by anyone with a basic understanding of SQL. Superset is also accessible as web application or app [7]. The figure 5 shows the connection between Superset and Hive.

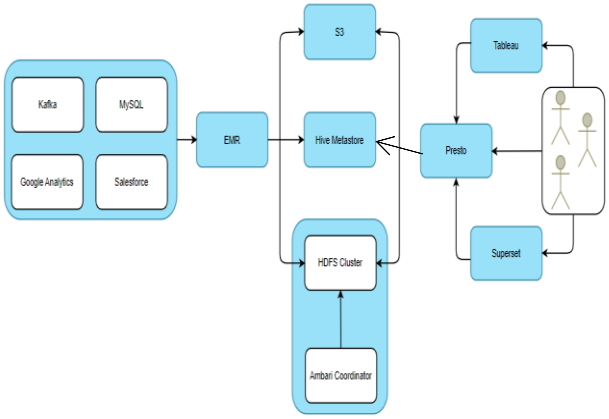


Figure 5: Superset connection with Hive

6. Big Data Climate Change and visualizing

6.1 About climate change data

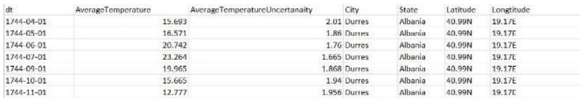
To use Hadoop, we have created an ubuntu environment and everything we have used is in a live server. To secure our work environment we have used a firewall. A live server is a term used to describe a server that is currently in operation and accessible on the internet or a network. It is a server that is running and serving web pages, applications, or other content to clients who are requesting it. A firewall is a security system that is designed to prevent unauthorized access to or from a private network. It acts as a barrier between a trusted network (such as a corporate or home network) and an untrusted network (such as the internet) [8].

Some say climate change is the biggest threat of our age. There are a lot of organizations that collect and use climate trends data. We have taken our data from NOAA’s MLOST, NASA’S GISTEMP and UK’s HADCRUT [9].

The newer data is collected by the Berkley Earth, which is affiliated with Lawrence National Library. The Berkeley Earth Surface Temperature Study combines 1.6 billion temperature reports from 16 pre-existing archives. It is nicely packaged and allows for slicing into interesting subsets (for example by country). They publish the source data and the code for the transformations they applied [9].

6.2 Dataset Overview

This Dataset focuses on data related to the city of Durrës, Albania, covering a time period between 1744 and 2013. There are six columns of data: date, country, city, average temperature, longitude and latitude. Each column has a specific type of data associated with it. The date column contains information about the time period during which the climate data was recorded. This column is typically represented as a timestamp or date format, such as yyyy-mm-dd. The country and city column contains information about the geographic location where the climate data was collected. This column is typically represented as a string data type, with the name of the country and city. These data column is essential for analyzing climate data for specific regions and identifying trends and patterns in climate data for different countries. The average temperature column contains information about the temperature at the location where the data was collected. This column is typically represented as a numeric data type, with the temperature value recorded in degrees Celsius. The average temperature column is essential for analyzing climate data for temperature trends over time and identifying changes in temperature patterns due to climate change. The longitude column contains information about the east-west position of a point on the earth's surface. The latitude provides information about the north-south position of a particular location where the temperature data was recorded. It is typically represented as a numeric data type, with the latitude value recorded in degrees. In summary, each column in a climate data file contains specific types of data that are critical for analyzing climate data and identifying trends and patterns in temperature and climate variables over time. The date, country, average temperature, and longitude columns provide essential information that enables researchers to gain insights into the impact of climate change on the environment and identify strategies for mitigating its effects. The figure 6 shows the format of the data file.

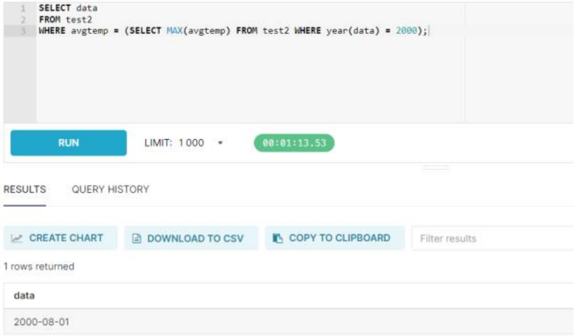


dt	AverageTemperature	AverageTemperatureUncertanaity	City	State	Latitude	Longtitude
1744-04-01	15.693	2.01	Durres	Albania	40.99N	19.17E
1744-05-01	16.571	1.86	Durres	Albania	40.99N	19.17E
1744-06-01	20.742	1.76	Durres	Albania	40.99N	19.17E
1744-07-01	23.264	1.665	Durres	Albania	40.99N	19.17E
1744-09-01	19.965	1.868	Durres	Albania	40.99N	19.17E
1744-10-01	15.665	1.94	Durres	Albania	40.99N	19.17E
1744-11-01	12.777	1.956	Durres	Albania	40.99N	19.17E

Figure 6: CSV file data

6.3 Visualizing data

Superset is a popular open-source Business Intelligence tool that allows users to easily create visualizations and dashboards based on their data. After data is stored successfully, we started using superset so we could make data visualization. Superset is deployed in docker. Docker is a popular platform for deploying and managing containerized applications. Running Superset in Docker allows you to easily deploy and manage Superset as a containerized application. While working with this data we have made an analyzation of the data for Albania like finding the maximum, minimum and average of temperatures in different times. Figures 7 through 10 display a selection of these examples.

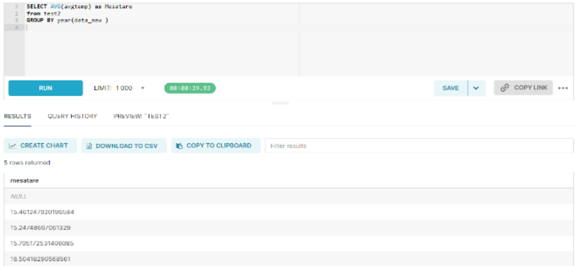


```
SELECT data
FROM test2
WHERE avgtemp = (SELECT MAX(avgtemp) FROM test2 WHERE year(data) = 2000);
```

1 rows returned

data
2000-08-01

Figure 7: The hottest day from 2000 to 2013

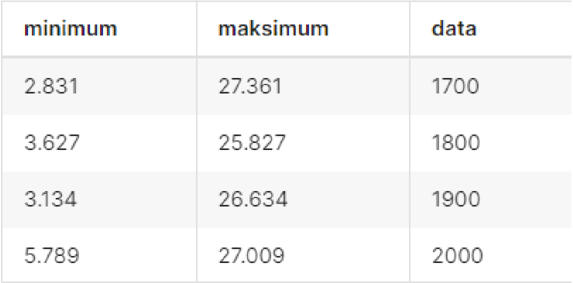


```
SELECT AVG(avgtemp) as Mesatare
from test2
GROUP BY year(data_new )
```

5 rows returned

mesatare
NULL
15.461247920196584
15.247486670613 29
15.705172531406085
18.50418290568561

Figure 8: Average of each century



minimum	maksimum	data
2.831	27.361	1700
3.627	25.827	1800
3.134	26.634	1900
5.789	27.009	2000

Figure 9: Result query for average of each century

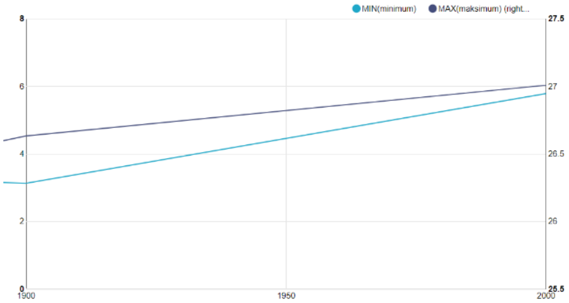


Figure 10: Graphic for average of each century

The area chart graphic from Superset visualization is a powerful tool for visualizing data related to climate change as it is shown in figure 11. By plotting the average temperature from 1800, 1900, and 2000 for the city of Durres in Albania, we can see the stark changes in temperature over the past century. The graphic shows a clear upward trend, with temperatures steadily increasing from 1800 to 2013. This visual representation of the data makes it easy to see the significant impact that climate change has had on the temperature in Durres. It highlights the urgency of taking action to address this issue, as well as the importance of using data visualization tools like Superset to help us better understand and communicate complex environmental issues.

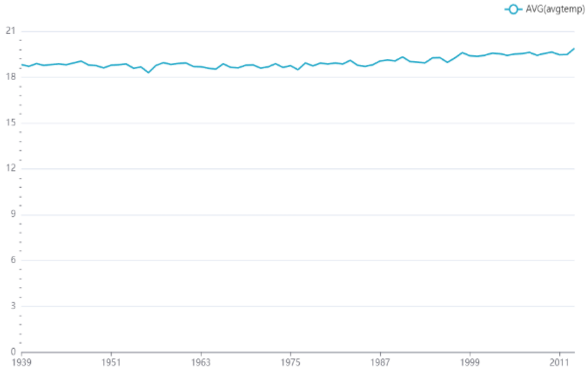


Figure 11: Area Chart Superset

7. Conclusions

In this work, we have demonstrated the effectiveness of Hadoop, Hive, and Superset in managing and analyzing data related to climate change. The use of Hadoop allowed us to process and store large amounts of data, while Hive provided a powerful SQL-like interface for querying and analyzing the data. Finally, Superset enabled us to visualize the data in interactive dashboards and charts, providing a comprehensive view of our findings. Through our analysis, we have identified significant patterns and trends related to climate change. Our findings indicate that global temperatures are rising, and extreme weather events are becoming more frequent. These changes have far-reaching consequences for our planet, including the loss of biodiversity, the displacement of communities, and the exacerbation of social and economic inequalities. Found out that the data have been increased from 1700 to 2013 with an average of about 1.1°C. In conclusion, our work demonstrates the power of advanced technologies like Hadoop, Hive, and Superset in analyzing complex issues like climate change. By leveraging these tools, we were able to gain critical insights into the impact of climate change and develop evidence-based strategies to address it. Moving forward, we believe that the use of advanced technologies and sound research methods will be essential in addressing the pressing issues facing our planet, including climate change.

8. References

[1] White, T. (2015). Hadoop: The definitive guide (4th ed.). O'Reilly Media.

[2] Smith, J. (2017). Hadoop: Processing large datasets with HDFS and MapReduce. Journal of Big Data, 4(1), 1-10

[3] Hive: A Data Warehousing Tool on Hadoop. Journal of Computer Science and Technology, 34(1), 1-21. doi: 10.1007/s11390-019-1909-3 (2019)

[4] Building a high-level dataflow system on top of Map-Reduce: The Pig experience. Proceedings of the VLDB Endowment, 2(2), 1414-1425

[5] Programming Hive: Data Warehouse and Query Language for Hadoop" by Edward Capriolo, Dean Wampler, and Jason Rutherglen

[6] Ganta, S. R., & Ranganatham, M. (2020). Analyzing and Visualizing Big Data Using Hive and Superset. International Journal of Computer Sciences and Engineering, 8(7), 263-268.

[7] Wang, Y., Han, R., & Wang, S. (2020). Application of Superset in Big Data Analysis Platform. 2nd International Conference on Computer Science and Software Engineering

[8] Neupane, Kishan & Haddad, Rami & Chen, Lei. (2018). Next Generation Firewall for Network Security: A Survey. 1-6. 10.1109/SECON.2018.8478973.

[9] Climate Change: Earth Surface Temperature Data https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

[10] Hassani, Hossein & Huang, Xu & Silva, Emmanuel. (2019). Big Data and Climate Change. Big Data and Cognitive Computing. 3(1). 10.3390/bdcc3010012.

[11] Hua-Dong, Guo & Zhang, Li & Zhu, Lanwei. (2015). Earth observation big data for climate change research. Advances in Climate Change Research. 6. 10.1016/j.accre.2015.09.007.

[12] Papadopoulos, Thanos & Balta, Maria. (2021). Climate Change and big data analytics: Challenges and opportunities. International Journal of Information Management. 63. 102448. 10.1016/j.ijinfomgt.2021.102448.