**Exploring Environmental Footprints: Emission Analysis Across Regions with Spark and AWS**

**References:**

**[1]** Dataset: <https://www.stats.govt.nz/large-datasets/csv-files-for-download/>

**[2]** Beginning Apache Spark 2: Hienn Luu

**[3]** Code Reference: <https://spark.apache.org/docs/latest/quick-start.html>

**[4]** AWS Connect: Spark-Installation-on-AWS-UPDATED.pptx

**Introduction**

**The Significance of Apache Spark in Modern Data Analysis**

In today's data-driven world, Apache Spark is recognized for its innovation and efficiency. Its popularity stems from its fast-processing speed, easy-to-use interface, and flexibility, making it essential for companies of all sizes and industries. This report explores how Apache Spark is changing the landscape of data analysis and its impact on our daily lives. Specifically, it demonstrates how Spark can be used to analyze large datasets to understand the environmental impacts of different areas. Through this exploration, we uncover how Spark can help us make more informed decisions for a sustainable future.

Apache Spark's rapid popularity is underscored by its ability to execute data processing tasks with unparalleled speed. With claims of being up to 100 times faster than its predecessor, Hadoop MapReduce, Spark has shattered records in industry benchmarks. For instance, in the Daytona GraySort contest, Spark demonstrated its prowess by sorting 100TB of data three times faster and utilizing a fraction of the resources compared to conventional methods. This exceptional performance empowers organizations to process vast amounts of data swiftly and efficiently, driving faster insights and decision-making.

One of Spark's most compelling features is its user-friendly design, catering to a diverse audience ranging from developers to data scientists. Offering over 80 high-level data processing operators and support for multiple programming languages, including Scala, Java, Python, and R, Spark democratizes data analysis. Its intuitive interface and extensive library of tools enable users to tackle complex data challenges with ease, fostering innovation and experimentation.

Spark's versatility is another cornerstone of its appeal, providing a unified platform capable of handling diverse data processing workloads. Whether it's batch processing, interactive queries, iterative algorithms for machine learning, or real-time streaming analytics, Spark offers a comprehensive solution. By consolidating disparate technologies into a single stack, Spark streamlines operations, reducing overhead costs, and simplifying infrastructure management.

In the vast landscape of big data technologies, Spark seamlessly integrates with existing infrastructure, enhancing its adoption and utility. Compatible with distributed storage engines like HDFS and cluster management systems, Spark complements the broader big data ecosystem. This interoperability fosters synergy, enabling organizations to leverage their existing investments while benefiting from Spark's advanced capabilities.

An additional boon of Apache Spark is its open-source nature, fostering collaboration, innovation, and transparency. The accessibility of its source code empowers developers to dissect, enhance, and customize Spark to suit their specific needs. This openness not only accelerates problem-solving but also cultivates a vibrant community of contributors, driving continuous improvement and evolution.

In conclusion, Apache Spark stands as a linchpin in modern data analysis, revolutionizing the way organizations harness the power of data. Its remarkable performance, user-friendly interface, flexibility, seamless integration, and open-source ethos collectively make it an indispensable asset in the data analytics toolkit. As data continues to proliferate and complexity mounts, Spark's significance in our daily lives and its impact on data analysis will only continue to grow. Embracing Spark is not just a choice but a strategic imperative for any organization seeking to thrive in the data-driven era.

**Project Details**

This project endeavors to delve into the emissions of various regions to gain insights into environmental impacts and pollution levels. Understanding the emission patterns of different areas is crucial for several reasons. Firstly, it allows policymakers and environmental scientists to identify hotspots of pollution, enabling targeted interventions to mitigate adverse effects on ecosystems and public health. Moreover, such analysis aids in assessing the effectiveness of existing regulations and initiatives aimed at reducing emissions and combating climate change.

Attempting to conduct this analysis manually, without the aid of advanced data processing tools like Apache Spark, would be impractical and inefficient. The volume and complexity of data involved in tracking emissions across multiple regions and sources necessitate automated data processing techniques. Traditional manual methods would be labor-intensive, time-consuming, and prone to errors, making it challenging to extract meaningful insights from the vast amounts of data available.

Furthermore, manual analysis would struggle to handle the real-time nature of emission data, hindering timely decision-making and response to emerging environmental concerns. By leveraging Apache Spark's capabilities for distributed computing and parallel processing, this project can efficiently handle large datasets, perform complex calculations, and generate actionable insights in a fraction of the time it would take through manual methods.

In essence, the use of advanced data analysis tools like Apache Spark is indispensable for conducting comprehensive and timely assessments of emissions and environmental impacts. These tools not only enhance the efficiency and accuracy of analysis but also empower stakeholders to make informed decisions to safeguard the environment and promote sustainable development.

**Setting Up Spark using AWS**

1. Create an EC2 instance using Ubuntu Free Tier Micro.

* In the security settings use “All Traffic.”
* Additionally, add a key-pair value to protect the access privileges to your EC2 instance.
* Ensure that you are in the N Virginia region.

1. SSH into the EC2 instance

* Either use the internet browser to connect to the instance or ssh terminal.
* This project used MobaXTerm to connect to the EC2 instance.

1. Download the requirements for Anaconda on EC2: use the following commands

* *wget* [*http://repo.continuum.io/archive/Anaconda3-4.1.1-Linux-x86\_64.sh*](http://repo.continuum.io/archive/Anaconda3-4.1.1-Linux-x86_64.sh)
* *bash Anaconda3-4.1.1-Linux-x86\_64.sh :* Press enter when the bash shell opens and you will see the Agreement Form. Fill up at least a single detail and use . to skip. Continue to hit enter until you see the statement “Do you approve of the License Terms.” Type in ‘yes’ and follow to instructions on the screen to install Anaconda at your preferred location.
* This will install a lot of files and dependencies and might take a while.
* In case of an error related to the tar file, install the requirements based on the error messages.
* *source .bashrc*
* *jupyter notebook –generate-config*

1. Create Certifications

* *mkdir certs*
* *cd certs*
* sudo openssl req -x509 -nodes -days 365 -newkey rsa: 1024 -keyout mycert.pem -out mycert.pem

1. Configuring Notebooks

* *cd ~/.jupyter/*
* *vim jupyter\_notebook\_config.py*
* Hit I to edit and then add the following code to the document. Hit esc and then type in :wq to save and exit

A screenshot of a computer screen

Description automatically generated

1. Troubleshooting Checklist

* If you don’t get the Jupyter notebook running: Make sure you have an https: <EC2 INSTANCE>:8888
* If you are getting a “permission denied” error, it’s because the EC2 Instance is preventing access.
* In order to handle this, we need to change the root of the jupyter library. This is done using the following instructions:
* *sudo chown -R $USER /home/*
* *sudo chown ubuntu:ubuntu mycert.pem*
* *jupyter notebook* : run the server.

1. Install Java RE and SCALA that act as the Bases to run Spark

* *sudo apt-get update*
* *sudo apt-get install default-jre*
* *sudo apt-get install scala*

1. Install Spark

* Follow the next steps to install py4j
* *export PATH = $PATH:$HOME/anaconda/bin*
* *conda install pip*
* *pip install py4j*
* Follow the next steps to install Spark/Hadoop
* *wget* [*http://archive.apache.org/dist/spark/spark-2.0.0/spark-2.0.0-bin-hadoop2.7.tgz*](http://archive.apache.org/dist/spark/spark-2.0.0/spark-2.0.0-bin-hadoop2.7.tgz)
* *sudo tar -zxvf spark-2.0.0-bin-hadoop2.7.tgz*

1. Point Python to Spark

* *export SPARK\_HOME=‘/home/ubuntu/spark-2.0.0-bin-hadoop2.7’*
* *export PATH=$SPARK\_HOME:$PATH*
* *export PYTHONPATH=$SPARK\_HOME/python:$PYTHONPATH*
* Launch Jupyter Notebook: *jupyter notebook*

1. Alternate (easier way)

* Use VS Code to write on your new Jupyter Notebook.

**Example Code**

To test if our entire set up works, we can try some sample code and get the outputs. We know our setup works as we get the right required outputs as given below.

1. 3.14088
2. [(‘other’, 1), (‘the’, 2), (‘over’, 1), (‘jumped’, 1), (‘dog’, 2), (‘brown’, 2)]
3. [4, 3]
4. [1, 2, 3, 4, 5, 3, 4, 6, 7, 8]

Example 1, 2

A screenshot of a computer

Description automatically generated

Example 3, 4

A screenshot of a computer program

Description automatically generated

**Dataset**

In typical data analysis projects, acquiring and preparing the dataset involves various steps, including data collection, cleaning, and organization. However, for the sake of simplicity and to streamline the process, this project utilizes a readily available dataset provided by the government of New Zealand. The dataset, sourced from the official statistics agency, Statistics New Zealand, offers comprehensive insights into greenhouse gas emissions from industry and households for the December 2023 quarter.

The dataset, accessible via the provided link (<https://www.stats.govt.nz/assets/Uploads/Greenhouse-gas-emissions-industry-and-household/Greenhouse-gas-emissions-industry-and-household-December-2023-quarter/Download-data/greenhouse-gas-emissions-industry-and-household-december-2023-quarter.csv>), serves as the foundation for our analysis. By leveraging this official dataset, we ensure the reliability and accuracy of the information, as it is sourced from a trusted government authority.

Part of the dataset

A screenshot of a computer

Description automatically generated

**Purpose**

The code aims to analyze and visualize the total CO2 emissions over the years using Apache Spark for data processing and Pandas with Matplotlib for data visualization.

**Components**

* Spark Session Initialization: The code initializes a Spark Session named "CO2 Emissions Analysis" to work with Spark DataFrame.
* Schema Definition: A schema is defined to specify the structure of the data to be read from the CSV file. It includes fields such as "area", "sector", "year", and "emissions".
* Read Data: The code reads the CO2 emissions data from a CSV file into a Spark DataFrame. The CSV file is assumed to be tab-separated and contains columns according to the defined schema.
* Group and Aggregate: The emissions data is grouped by the "year" column, and the total emissions for each year are computed by summing up the emissions across all areas.
* Convert to Pandas: The aggregated data from Spark DataFrame is converted into a Pandas DataFrame for easier manipulation and plotting.
* Plotting: The total CO2 emissions for each year are plotted using Matplotlib. The Pandas DataFrame is used to create a bar plot showing the trend of emissions over the years.
* Visualization: The resulting visualization is a bar chart that displays the total CO2 emissions for each year. Each bar represents the cumulative emissions for a specific year, showing the trend of emissions over time.

**Dependencies**

* PySpark: Used for data processing and working with Spark DataFrame.
* Pandas: Utilized for converting the Spark DataFrame to a Pandas DataFrame.
* Matplotlib: Employed for creating the bar plot to visualize the emissions data.

**Analysis 1:**

Using Spark to analyze the dataset and group the data according to emissions by region.

CODE:  
A screen shot of a computer code

Description automatically generated

OUTPUT:

A black screen with white text

Description automatically generated

A graph with blue and black text

Description automatically generated

**Results:**

The provided data represents the total CO2 emissions over several years in the different areas of New Zealand. From the bar plot, we can observe the trend of CO2 emissions over the years. By examining the heights of the bars, we can identify the areas with the highest emissions being Waikato( 1,370,442.42), Canterbury (979,700.25), and Auckland (888,173). Identifying regions with disproportionately high emissions can help target interventions and resources more effectively.

Comparing emissions data with other relevant factors such as economic indicators, population growth, energy consumption, and environmental regulations can provide additional context and insights into the drivers of CO2 emissions.

Using historical emissions data, statistical techniques, and predictive modeling, we can forecast future emissions trends. This forecasting can help policymakers, businesses, and environmental organizations make informed decisions and develop effective strategies for mitigating climate change.

Insights from the analysis can inform the development and implementation of policies aimed at reducing CO2 emissions. Understanding the drivers of emissions and their impact on the environment and society is crucial for designing effective regulatory frameworks and incentive programs.

Overall, analyzing the data provides valuable insights into the dynamics of CO2 emissions, helping stakeholders understand the current state of emissions, identify areas for improvement, and develop strategies to address climate change effectively.

**Analysis 2:**

Using Spark to analyze the dataset and group the data according to emissions by year.

CODE:

A screen shot of a computer program

Description automatically generated

OUTPUTS:

A black and white screen with numbers

Description automatically generated

A graph of blue bars

Description automatically generated with medium confidence

**Results:**

The data represents the total CO2 emissions for each year from 2007 to 2022. From the values, we can observe fluctuations in emissions over time. They don’t seem to be extremely varied, however there is a drastic drop by about 100,000 in the year 2022. Significant events such as economic recessions, natural disasters, policy changes, and global pandemics (e.g., COVID-19) can also impact emissions levels by affecting industrial output, transportation activities, and energy demand. Analyzing what changes or factors brought about this difference can help point us in the right direction towards achieving a sustainable future.

We can identify any significant changes in emissions from year to year. Sudden spikes or drops in emissions may indicate specific events or changes in environmental policies, economic conditions, or technological advancements.

Analyzing the data over a longer time frame can help identify long-term changes in emissions patterns. This analysis can provide insights into the effectiveness of sustainability initiatives, the impact of industrial activities, and the overall trajectory of emissions reduction efforts.

The highest total CO2 emissions occurred in 2019, with a value of 450,108.21 units. This peak could be indicative of increased industrial activity, economic growth, or changes in energy consumption patterns.

There seems to be a general decreasing trend in emissions from 2019 to 2022. The total emissions decreased from 450,108.21 units in 2019 to 364,825.18 units in 2022, indicating potential efforts to reduce emissions or external factors impacting emissions levels.

Analyzing historical emissions data can provide insights into the effectiveness of past interventions and help guide future policy decisions and sustainability initiatives. Understanding the drivers of emissions trends is crucial for developing strategies to mitigate climate change and promote sustainable development.

Overall, the analysis of the provided data highlights the importance of monitoring CO2 emissions trends and identifying opportunities for reducing greenhouse gas emissions to address climate change and promote environmental sustainability.

**Other future analysis avenues:**

There are several other ways you could analyze the provided data to gain additional insights. Here are some alternative approaches:

1. Year-over-Year Growth Rate: Calculate the year-over-year growth rate of CO2 emissions to understand the rate of change over time. This analysis can help identify periods of rapid growth or decline in emissions.

2. Seasonal Analysis: Analyze seasonal patterns in CO2 emissions to identify any recurring trends or fluctuations throughout the year. This analysis can be particularly useful for understanding the impact of seasonal factors such as weather patterns, agricultural activities, and holiday seasons on emissions levels.

3. Correlation Analysis: Explore potential correlations between CO2 emissions and other relevant variables such as economic indicators (e.g., GDP growth), energy consumption, population growth, and environmental policies. This analysis can help identify factors that influence emissions levels and inform targeted interventions to reduce emissions.

4. Sectoral Analysis: Break down CO2 emissions by sector (e.g., transportation, industry, residential, commercial) to understand which sectors contribute most significantly to overall emissions. This analysis can help prioritize efforts to reduce emissions in the most impactful sectors.

5. Longitudinal Analysis: Analyze trends in CO2 emissions over longer time periods to identify long-term patterns and assess the effectiveness of past interventions. Longitudinal analysis can help evaluate progress towards emissions reduction goals and inform future policy decisions.

By applying these alternative analytical approaches, you can gain a more comprehensive understanding of the factors driving CO2 emissions and identify opportunities for mitigating climate change and promoting sustainable development.

**Conclusion:**

In conclusion, the analysis of greenhouse gas emissions using Apache Spark and the dataset provided by the government of New Zealand offers valuable insights into environmental impacts and pollution levels. By leveraging advanced data processing tools and authoritative datasets, we can uncover patterns, trends, and hotspots of emissions, facilitating informed decision-making for environmental policy and mitigation strategies.

The project demonstrates the indispensability of sophisticated data analysis techniques in tackling complex environmental challenges. Manual methods would be impractical and inefficient in handling the vast amounts of data involved, underscoring the importance of leveraging technology for meaningful insights.

Moving forward, the findings from this analysis can inform policymakers, environmental scientists, and stakeholders in implementing targeted interventions to reduce emissions and mitigate the adverse effects of climate change. By harnessing the power of data and technology, we can work towards a more sustainable future, safeguarding the health of our planet and its inhabitants.