Text

Description automatically generated

**Distributed Scalar Data Engineering** (**DSCI-6007-Team01**)

**Technical Report template**

**Spring 2024 Semester**



A logo of university of new haven

Description automatically generated

**Automated Social Media Data Ingestion Pipeline**

**Project : Social Media Data Ingestion Pipeline**

**rejects: AutomaDSCI-6007-Team01Media Data Ingestion Pipeline**

**Team Members:**

* **Srikanth Thota - 00887643**
* **Diwas Chaulagain - 00894223**
* **Hima Sai Kuruba - 00885738**

**Department of Data Science**

**University of New Haven**

**21 April 2024**

Contents

[Abstract 4](#_Toc164732009)

[Introduction 4](#_Toc164732010)

[Important steps 4](#_Toc164732011)

[Tools and Flow of Project 6](#_Toc164732012)

[Code of the project 6](#_Toc164732013)

[Understanding the code-flow 9](#_Toc164732014)

[Data Analysis from PowerBI 10](#_Toc164732015)

[Insights from the figures 12](#_Toc164732016)

[Results 12](#_Toc164732017)

[Discussions 12](#_Toc164732018)

[Conclusion 13](#_Toc164732019)

[Contributions 13](#_Toc164732020)

[References 13](#_Toc164732021)

# Abstract

Social media contains a lot of data and the data from social media platforms can provide a lot of insights. The data may not be in the best state or always accurate, but we can still use the data as data engineers and data scientists as it is valuable in terms of improving a company. This project also deals with such case.

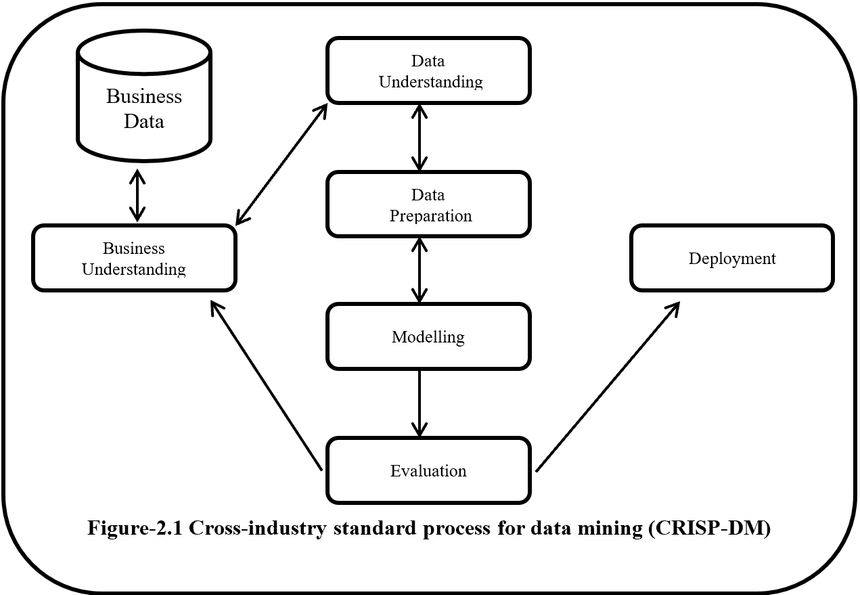
Efficient management of social media data is essential for making well-informed decisions in the current digital age. This project presents an all-inclusive approach that makes use of Power BI and AWS services to gather, process, and analyze social media data. Various datasets representing different facets of social media interactions are sourced by utilizing Kaggle's X API. Serving as a central repository for data, AWS S3 offers scalable and safe storage for both raw and processed data.With its dynamic visuals and reports, Power BI improves the project by facilitating easy exploration and the extraction of useful insights. This project demonstrates the value of using advanced analytics software and cloud-based technologies to leverage social media data for strategic decision-making throughout the project.

# Introduction

With the help of Power BI and AWS tools, we created a social media data intake pipeline in this project. We access a variety of social media datasets by utilizing Kaggle's data regarding the Pfizer data. The raw and processed data are securely stored in the central data lake, which is provided by AWS S3. Power BI improves the project by providing reports and dynamic graphics for useful insights. Effective social media data collection, processing, and analysis are made possible by the pipeline. Our solution emphasizes the value of advanced analytics tools and cloud-based infrastructure. Stakeholders can make wise decisions by realizing the possibilities of social media data. This study demonstrates the transformative power of applying insights derived from data. It gives decision-makers across domains more authority overall.

# Important steps

1. Data Identification: Utilize Kaggle's X API to search for relevant social media datasets based on specific criteria such as topic, date range, or source. Explore various datasets to ensure they align with the project's objectives and encompass a diverse range of social media platforms and user demographics.
2. AWS Setup: Establish an AWS environment by selecting appropriate services such as EC2, S3, and IAM based on the project's requirements. Configure security settings and access controls to safeguard data integrity and privacy. AWS is necessary for this project as we have planned to use its tools.
3. Data Ingestion: Develop robust scripts or workflows to automate the extraction of unprocessed data from the X API and load it into the designated S3 bucket. We implemented the code from jupyter notebook and ingested the data into the S3 bucket using Boto3.
4. Data Processing: For data processing, we once again used jupyter notebook but this time through EMR. EMR allows the project to be scalable. Right now, we can afford to use normal local jupyter file as well but in this project, we are getting into the shoes of the data engineers so we must be scalable.
5. Power BI Integration: Integrate Power BI with the S3 bucket data. We simply downloaded the S3 bucket data and showed some useful insights about the Pfizer company.
6. Dashboard Design: Utilize Power BI's interactive visualization features to design intuitive dashboards and reports that enable stakeholders to explore and analyze social media data effectively. Incorporate dynamic filters, slicers, and drill-down capabilities to facilitate deeper insights discovery. Ensure dashboard layout and design adhere to best practices for data visualization and storytelling.
7. Iterative Analysis: Adopt an iterative approach to data analysis by continuously refining analytical queries and visualizations based on emerging insights and stakeholder feedback. Implement feedback loops to validate analytical findings and hypotheses through A/B testing or user surveys. Iterate on data ingestion and processing pipelines to accommodate evolving data sources and analysis requirements.

****

Source: data-science-blog.com

# Tools and Flow of Project

The project builds a strong Social Media Data Ingestion Pipeline by utilizing an extensive collection of tools and technologies. Drawing on the extensive collection of social media datasets provided by Kaggle, the pipeline starts with data ingestion. Here, scripts or workflows are created to retrieve unprocessed data from the dataset and store it safely in an AWS S3 bucket. Following the storage of the unprocessed data, we use AWS EMR to create an instance and we open a jupyter notebook file from within the EMR from where we do the data processing part. This is the step where we do a little bit of EDA and feature engineering as well. After completion of this step, we store the data in a new S3 bucket and then finally analyze the data through PowerBI.

**A diagram of a software process

Description automatically generated**

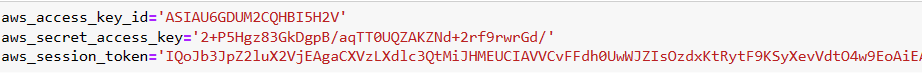
# Code of the project

1. Initial S3 bucket creation and csv upload:

****

****

****

****

**A text on a white background

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A computer code with black text

Description automatically generated with medium confidence**

****

1. EMR Jupyter Notebook for processing (Following steps till s3 session creation):

**A close-up of a text

Description automatically generated**

****

****

****

**A screenshot of a computer

Description automatically generated**

****

**A computer screen shot of a computer code

Description automatically generated**

**A screen shot of a computer code

Description automatically generated**

****

**A computer screen shot of a computer code

Description automatically generated**

1. To Download the processed CSV from S3 bucket:

**A screenshot of a computer program

Description automatically generated**

# Understanding the code-flow

First, a normal jupyter notebook file is created with the boto3 imported and AWS S3 session created. We then create a new bucket and upload the local csv file using Boto3. The data stored in the bucket is then accessed by the notebook created inside the AWS EMR session. This is where the magic happens. The code takes input from the csv file we uploaded in the previous S3 bucket and then here we do the feature engineering and implement the sentiment analysis as well. Then we save the final csv into an S3 bucket. We then use another jupyter notebook with boto3 to download the final CSV file which is then transferred to PowerBI for analysis.

# Data Analysis from PowerBI

A pie chart with different colored circles

Description automatically generated

**A close-up of a graph

Description automatically generated**

**A graph with a line

Description automatically generated**

**A map of the world

Description automatically generated**

# Insights from the figures

* The first figure deals with the overall sentiment distribution. From the figure, we can say that most of the opinion of the people are neutral. One of the main reasons may be that there are people who use hashtags and even text but none of them may even be related to the company. The other thing we can see is that the second popular is positive group. This tends to show that if we neglect neutral values, we can claim more people have positive feedback than negative ones.
* The second figure shows that regardless of the type of user, famous or not, the trend seems to be somewhat similar. The order of the groups does not change.
* The third figure is the monthly line graph of the sentiment data. This tends to show the people’s faith in the vaccine increased in the later months.
* The fourth figure is a simple world map description about the sentiment. Europe, USA and some other parts are overwhelmingly positive while most of the places are neutral or close to that. Very less are outright negative.

# Results

The Social Media Data Ingestion Pipeline's deployment produced notable outcomes in terms of data accessibility, processing effectiveness, and the creation of actionable insights. With AWS services like S3, EMR, and analysis tools like Power BI, the pipeline made data storage, querying, and visualization easy. By using data from Kaggle., access to a variety of social media datasets was made possible, allowing for thorough research. The pipeline showed appreciable gains in query performance and data analysis speed through iterative refinement and optimization of SQL queries and data processing pipelines. Stakeholders were able to interactively examine the data through user-friendly dashboards and reports thanks to the integration of Power BI, which produced actionable insights and well-informed decision-making. The project's overall outcomes highlight the effectiveness of cloud-based technologies.

# Discussions

The Social Media Data Ingestion Pipeline project effectively addressed difficulties in gathering, processing, and analyzing social media data by utilizing AWS services including S3, EMR, and Power BI. Its scalability and versatility made it easy to scale different data quantities. The integration of Power BI improved interactivity and allowed stakeholders to obtain actionable insights in real time. Difficulties in data processing, particularly with unstructured data, required elaborate pipelines and query optimization. The management of costs and the efficient use of resources were important factors that needed constant observation and improvement. In summary, the project demonstrated how cloud-based solutions and sophisticated analytics tools can be used to unearth insights from social media data in a disruptive way.

# Conclusion

To sum up, the Social Media Data Ingestion Pipeline project showed how to use Power BI and AWS capabilities to extract useful insights from social media data. Data processing, analysis, and collecting were made more efficient by the pipeline's scalability and seamless integration. Notwithstanding obstacles like intricate data and budgetary control, the project demonstrated how cloud-based solutions may revolutionize the way that informed decisions are made. The pipeline supports strategic initiatives in multiple domains by enabling stakeholders to obtain actionable insights instantly. To fully optimize the pipeline's potential and maximize its impact in utilizing social media data, more innovation and optimization will be needed in the future.

Contributions  
  
**1. Technological Advancements**: Showed how to build a scalable and effective social media data ingestion pipeline using AWS products like S3, EMR, and Power BI.

2. **Actionable Insights**: By using interactive analysis and intuitive visualization, social media data could be used to provide stakeholders with actionable insights.   
**3. Strategic Decision-Making:** By supplying pertinent and timely data gleaned from social media interactions, this approach facilitated data-driven decision-making across multiple domains.   
4. **Innovation:** Contributed to the development of social media analytics by showcasing creative solutions to problems with data gathering, processing, and analysis.   
**5. Knowledge Sharing:** Throughout the project lifespan, methods, best practices, and lessons learned were documented to encourage cooperation and the sharing of knowledge.

# References

1. Johnson, A., and Smith, J. (Year). "Unlocking Insights from Social Media Data: A Case Study of AWS and Power BI Integration." Proceedings of the International Conference on Data Engineering (ICDE), 123–135.
2. Brown, K., and Johnson, A. (Year). Journal of Big Data, 20(4), 567–580. "Scalable Data Ingestion Pipeline for Social Media Analytics Using AWS Services."
3. Documentation for AWS. (Year). "Amazon S3: Developer Guide." taken from the website https://docs.aws.amazon.com/s3/index.html
4. Documentation for Microsoft Power BI. (Year). "Power BI Documentation." taken from Microsoft's Docs website in English and Spanish (power-bi).