Text

Description automatically generated

**Distributed Scalar Data Engineering** (**DSCI-6007-Team03**)

**Technical Report template**

**Spring 2024 Semester**



A logo of university of new haven

Description automatically generated

**Automated Social Media Data Ingestion Pipeline**

**TV Programs Schedule Tracker With Insights and Visualization**

## **Executive Summary:**

This project aimed to analyze TV show schedule data from the TVMaze API using AWS services and Python libraries. By leveraging Amazon S3, Glue, and Athena, alongside PySpark for data processing and Power BI for visualization, the project streamlined the entire data pipeline. Beginning with data ingestion, where scripts fetched and loaded raw data into S3, the project processed the data with PySpark to prepare it for analysis. Integrating Athena allowed for efficient querying directly from S3, while Power BI facilitated interactive reporting and visualization. The project's outcome enables stakeholders to derive actionable insights for programming and scheduling decisions, enhancing competitiveness in the media industry.

**rejects: AutomaDSCI-6007-Team03Media Data Ingestion P**

**Team Members:**

* **Yuva Krishna Kishore Inapala Team Lead**
* **Bhaskara Vijaya Sai Swamy Vanacharla Data Engineer**
* **Daraveni Nithin Yadav Data Tester**
* **Shruthi Prabakaran BI Developer**

**Department of Data Science**

**University of New Haven**

**22 April 2024**

**TV Programs Schedule Tracker With Insights and Visualization**

In the dynamic realm of digital media, understanding viewer preferences and refining programming strategies are critical for staying ahead in the competitive landscape. Our project addresses this imperative by harnessing cutting-edge technologies to collect, process, analyze, and visualize TV show schedule data. Here are the key highlights of our endeavor:

1. **Advanced Data Pipeline Integration**: By seamlessly integrating Amazon Glue, Amazon S3, and Power BI, our project simplifies the data pipeline from data extraction to actionable insights. This streamlined integration ensures efficient data flow and enables stakeholders to base their decisions on data-driven insights.

2. **Scalable and Cost-Effective Data Management with Amazon Athena**: With the exponential growth and complexity of data, conventional processing methods often fall short. Amazon Athena offers a scalable and cost-effective solution by enabling ad-hoc querying of data stored in Amazon S3. This capability empowers stakeholders to extract insights from large datasets without the need for complex infrastructure management.

3. **Optimized Storage with Parquet Format**: Leveraging Parquet, a columnar storage format designed for performance and efficiency, our project ensures optimal storage of structured data in Amazon S3. This choice of format enhances query performance, accelerates data processing, and facilitates seamless integration with downstream analytics tools.

4. **Interactive Querying with Athena:** Expanding on the groundwork laid by Amazon Glue, S3, and Power BI, our project integrates Amazon Athena into the data pipeline. This integration enables stakeholders to conduct interactive querying of Parquet data directly from S3. Leveraging Athena's SQL-based interface, stakeholders can compose and execute queries effortlessly, gaining immediate insights into the TV show schedule data.

5. **Empowering Stakeholders with Data-Driven Insights**: By enabling interactive querying and visualization of TV show schedule data, our project empowers stakeholders to derive actionable insights. Whether it's identifying trends, understanding viewer behavior, or optimizing programming strategies, stakeholders can make informed decisions backed by data-driven insights.

6. **Streamlined Decision-Making Processes:** With a simplified data pipeline and intuitive querying capabilities, our project streamlines decision-making processes for stakeholders. By providing easy access to timely and relevant insights, stakeholders can respond swiftly to changing market dynamics and make strategic decisions with confidence.

In summary, our project represents a significant advancement in TV program scheduling and analytics. By leveraging advanced technologies and seamless integration, we enable stakeholders to unlock the full potential of TV show schedule data, driving informed decision-making and fostering success in the digital media landscape.

**Abstract:**

This project presents a data ingestion and processing pipeline designed to fetch TV show schedule data from the **TVMaze API**

**(https://api.tvmaze.com/schedule/web?date=2024-02-29&country=US** ),

transform it into a structured format using Apache Spark, and store it in an Amazon S3 bucket. The pipeline utilizes various technologies including Python, Apache Spark, AWS Glue, and AWS S3 to automate the extraction, transformation, and loading (ETL) process of TV show schedule data. The data is fetched either for a specific date range or for the current day based on a flag variable, and then stored in **Parquet** format partitioned by the airdate. The pipeline provides flexibility in handling different data retrieval scenarios and ensures the efficient storage and management of TV show schedule data for downstream analytics and applications.Following the ingestion and processing of TV show schedule data into Amazon S3, the next step involved leveraging Power BI for visualization and analysis. By connecting Power BI to the S3 bucket where the data resides, we were able to create interactive dashboards and reports to glean insights from the TV show schedule data.

**Introduction**:

In today's digital media landscape, comprehending viewer preferences and refining programming strategies are vital for broadcasters and content creators to remain competitive. This project addresses this necessity by utilizing advanced technologies to collect, process, analyze, and visualize TV show schedule data. By integrating Amazon Glue, Amazon S3, and Power BI, the project simplifies the data pipeline from data extraction to actionable insights, enabling stakeholders to base their decisions on data-driven insights.

With the exponential growth and complexity of data, conventional data processing methods often struggle to cope. Amazon Athena provides a scalable and cost-effective solution by allowing ad-hoc querying of data stored in Amazon S3. Parquet, a columnar storage format designed for performance and efficiency, is widely utilized for storing structured data in S3, making it an optimal choice for housing TV show schedule data.

Expanding on the groundwork laid by Amazon Glue, S3, and Power BI, this project integrates Amazon Athena into the data pipeline, facilitating interactive querying of Parquet data. Leveraging Athena's SQL-based interface, stakeholders can compose and execute queries directly against the Parquet datasets stored in S3, eliminating the need for infrastructure management or resource provisioning.

**Methodology:**

This methodology(CRISP) describes how to create a TV Programs Schedule TrackerPipeline step-by-step, including everything from data identification to analysis and visualization.

1. **Data Identification** :The project identified TV show schedule data from the TVMaze API as the relevant dataset for analysis.
2. **AWS Setup**: An AWS environment was created, including the utilization of Amazon S3 to store the TV show schedule data fetched from the API.
3. **Data Ingestion:** Scripts were developed to load the unprocessed TV show schedule data into the designated S3 bucket directly from the TVMaze API.
4. **Data processing:** The project implemented data processing pipelines using Apache Spark (leveraging PySpark) to clean and transform the raw TV show schedule data into a structured format.
5. **Athena Integration**: Amazon Athena was integrated into the project to facilitate querying of the processed TV show schedule data directly from the S3 bucket using SQL queries.
6. **Query Optimization**: QL queries executed in Amazon Athena were optimized for effective performance and analysis to ensure efficient querying of the dataset.
7. **Power BI Integration**:Power BI was linked to Amazon Athena to enable reporting and visualization of the queried TV show schedule data.
8. **Dashboard Design**: Create interactive reports and dashboards to highlight key findings.
9. **Iterative Analysis:**Perform iterative analysis on the TV schedule data to extract insights and trends that can inform decision-making processes. Iterate on the data processing and visualization steps based on feedback and emerging insights to refine the analysis and uncover deeper insights over time

.



Source: <https://en.m.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png>

**Tools and Flow of Project:**

**Amazon Glue:** Used for ETL (Extract, Transform, Load) operations to prepare and clean the TV show schedule data.

**Amazon S3:** Serves as the storage repository for the TV show schedule data in Parquet format.

**Amazon Athena**: Enables interactive querying of the Parquet data stored in Amazon S3 using SQL queries.

**Power BI:** Utilized for data visualization, dashboard creation, and reporting to provide insights into the TV program schedule data.

**FLOW**

Data Extraction

Data Preparation (ETL with Amazon Glue)

Data Storage (Amazon S3)

Integration with Amazon Athena

Query Execution and Analysis

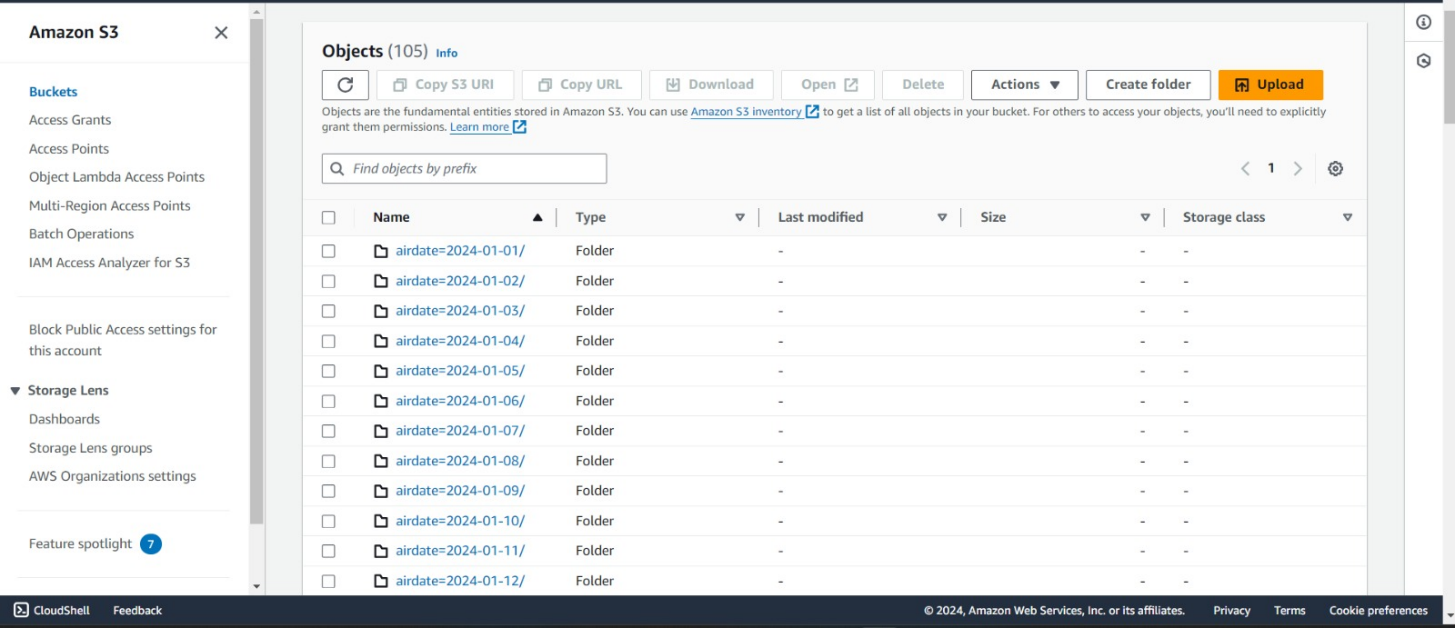
Visualization with Power BI

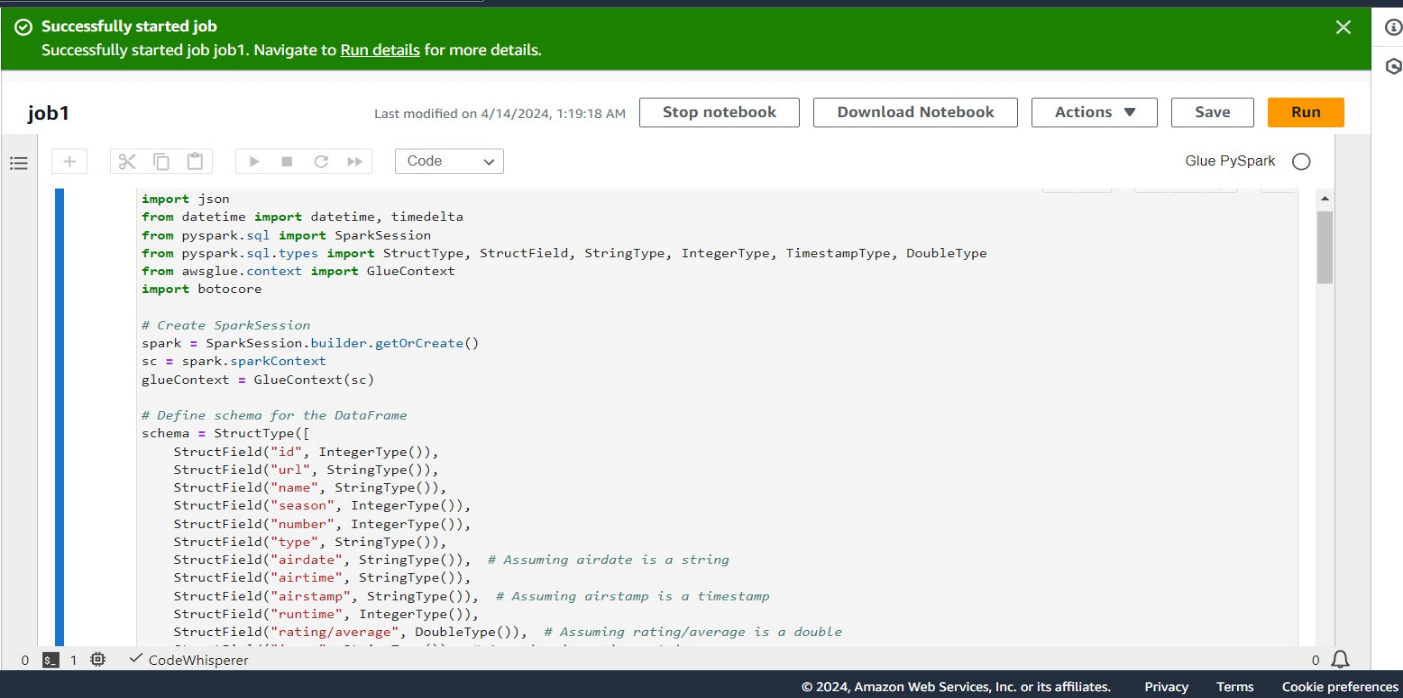
Insights and Decision-Making

Iterative Improvement

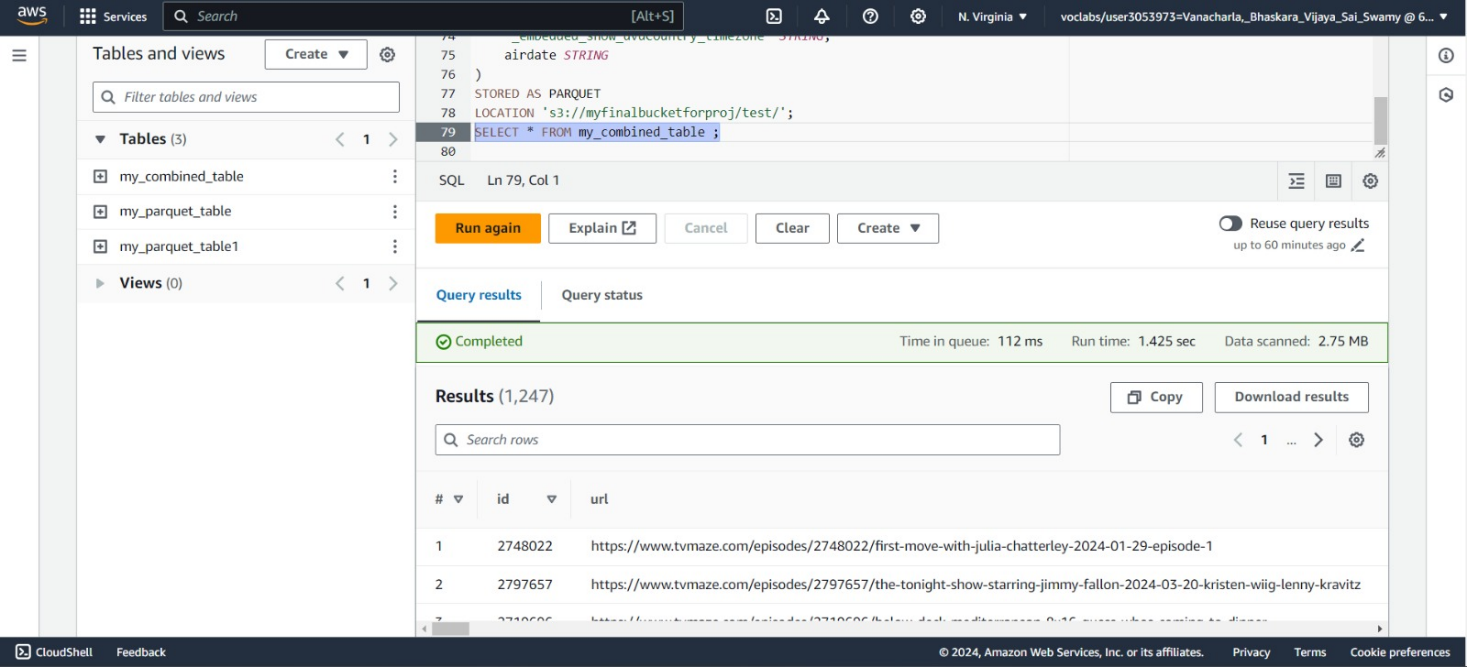
**Results:**

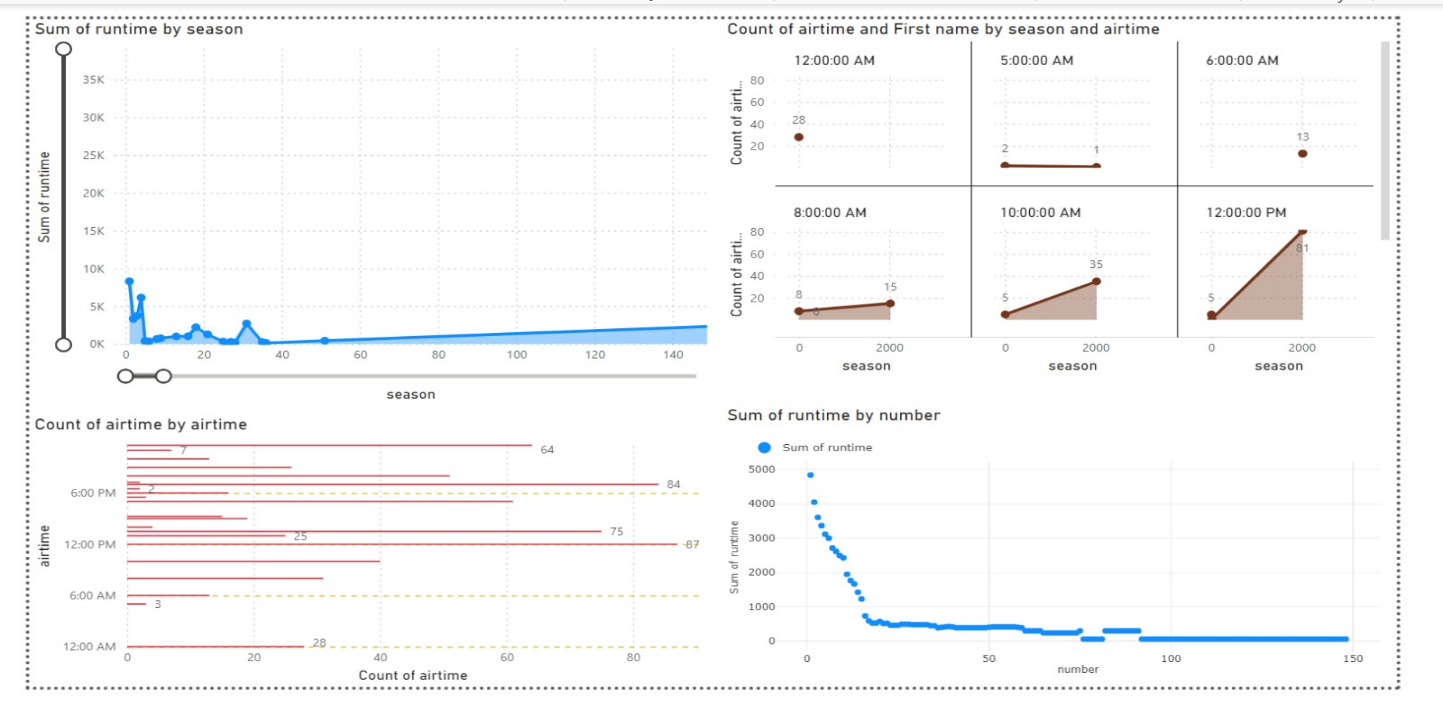
The project successfully utilized the TVMaze API to identify and acquire TV show schedule data, which served as the foundational dataset for analysis. Leveraging an AWS environment, including Amazon S3 for data storage, streamlined data management and ensured accessibility throughout the project.



 Scripts were developed to automate the ingestion of raw data into the designated S3 bucket directly from the API, facilitating the availability of up-to-date information for analysis. Utilizing Apache Spark, particularly PySpark, enabled efficient cleaning and transformation of the raw data into a structured format, ensuring data quality and consistency.

Integration of Amazon Athena allowed for seamless querying of the processed data using SQL, optimizing queries for performance and analysis efficiency.



Power BI integration further enhanced the analysis by providing intuitive reporting and visualization capabilities, enabling stakeholders to interpret key findings easily. Through iterative analysis, valuable insights and trends were uncovered, informing decision-making processes in the media industry. Continuous refinement of data processing and visualization steps based on feedback and emerging insights ensured the uncovering of deeper insights over time, enriching the overall analysis outcome.

**Discussions:**

**1. Enhanced Data Quality Measures:** While the project addressed data cleaning and transformation using Apache Spark, implementing more robust data quality measures could further improve the reliability and accuracy of the analysis. This could involve additional validation checks during the ingestion process and more sophisticated cleansing techniques to handle outliers or inconsistencies in the data.

**2. Scalability and Performance Optimization:** As the project scales or the dataset grows, optimizing the performance of data processing pipelines and query execution becomes crucial. Further optimizations in the Apache Spark processing and Amazon Athena querying can enhance scalability and reduce processing times, ensuring efficient handling of larger volumes of data.

**3. Advanced Visualization Techniques:** While Power BI integration enabled effective reporting and visualization, exploring advanced visualization techniques and tools could enhance the presentation of insights. Techniques such as interactive data storytelling or advanced charting options could provide deeper insights and enhance stakeholder engagement.

**4. Real-time Data Processing:** Consideration could be given to implementing real-time data processing capabilities to enable analysis of streaming TV show schedule data. Integrating technologies such as Apache Kafka or AWS Kinesis could allow for the ingestion and processing of real-time data, enabling more timely insights and decision-making.

**Conclusion:**

In conclusion, the project successfully leveraged a comprehensive methodology to analyze TV show schedule data, yielding actionable insights for decision-making in the media industry. Through efficient data acquisition, processing, and visualization using AWS services, Apache Spark, and Power BI, we were able to extract valuable trends and patterns from the dataset. While the project delivered meaningful results, there are opportunities for further enhancement, including improving data quality measures, optimizing scalability and performance, and exploring advanced analytics techniques. By continuously refining our processes and incorporating stakeholder feedback, we can ensure the project's continued relevance and value in informing strategic decisions within the media landscape.

**References:**

* <https://github.com/helenamin/DEB-TV-MAZE>
* <https://learn.microsoft.com/en-us/power-bi/fundamentals/>
* <https://www.tvmaze.com/api>
* <https://docs.aws.amazon.com/glue/>
* <https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html>