|  |  |
| --- | --- |
| **A red and white logo  Description automatically generated**  **Course Title:  ALY6110 – Big Data and Data Management**  **Course Number (CRN): 70362** | Group Delta Members:  Rushikesh Sawant([sawant.ru@northeastern.edu](mailto:sawant.ru@northeastern.edu))  Melike Uzun ([uzun.me@northeastern.edu](mailto:uzun.me@northeastern.edu))  Snehal Pawar ([pawar.sne@northeastern.edu](mailto:pawar.sne@northeastern.edu)) |

12/13/2023

**A Big Data Dive into the IMDb Flim Universe**

**Introduction: Use Case Senario**

In this project, we embark on a comprehensive exploration of a vast IMDb dataset comprising a million records, aiming to extract valuable insights for a movie production team. Leveraging Big Data technologies, our approach integrates Databricks and Snowflake to streamline the data processing workflow. Databricks, empowered by PySpark, becomes our engine for data fetching, cleaning, and preprocessing, providing a scalable and efficient environment to handle the immense IMDb dataset. The robust capabilities of PySpark allow us to execute complex operations seamlessly, ensuring that the data is refined and ready for analysis.

Snowflake serves as our chosen datalake, offering a secure and flexible storage solution for the IMDb dataset. The seamless integration between Databricks and Snowflake enables us to retrieve and store data effortlessly, fostering a dynamic and collaborative environment for data manipulation and exploration.

Upon completion of data operations, the results are seamlessly fed back into Snowflake, creating a new data table that encapsulates the refined insights. The final touch involves connecting Snowflake to Power BI, a powerful visualization tool, to craft an intuitive and interactive dashboard. This dashboard not only provides a visually appealing representation of the gleaned insights but also empowers the movie production team with actionable information for informed decision-making. This project underscores the synergy between advanced Big data processing tools, facilitating a streamlined journey from raw data to impactful visualizations.

**Dataset**

The IMDb Non-Commercial Datasets offer a valuable resource for Film enthusiasts and researchers, providing subsets of IMDb data for investigation and drawing insights from the films data. The dataset Consists of 1,048,576 rows compromised in various tables.

Some key datasets include 'title.akas.tsv.gz,' containing alternative titles with attributes like region, language, and type; 'title.basics.tsv.gz,' detailing title information, release year, and genres; 'title.crew.tsv.gz,' identifying directors and writers; 'title.episode.tsv.gz,' offering episode-specific data for TV series; 'title.principals.tsv.gz,' featuring information on individuals associated with a title; 'title.ratings.tsv.gz,' providing average ratings and vote counts; and 'name.basics.tsv.gz,' containing details about individuals, including birth and death years, professions, and titles they are known for.

In leveraging the IMDb Non-Commercial Datasets, research is conducted to draw insights from the vast film-related information contained within the datasets. Film enthusiasts and researchers can explore trends, correlations, and patterns by analyzing key datasets such as 'title.basics,' 'title.ratings,' and 'title.principals.' For instance, we may investigate relationships between genres and audience ratings, identify prolific directors and writers, or trace the career trajectories of individuals associated with multiple titles. The goal is to uncover valuable information about the film industry, audience preferences, and the contributions of various professionals.

A screenshot of a computer screen

Description automatically generated

The ER – Diagram gives a deeper understanding of the dataset and its Schema.

**Aim of the Project:**

Utilizing Snowflake as Data Lake:

* Uploading dataset to Snowflake as per the required Schema
* Providing the right Access to users for ease of access with regards to data fetching and schema creation

**Establish connection with Snowflake Database:**

* Connect to the Snowflake database securely using the necessary credentials and procedure for authentication.
* Configure the database connection for future data transactions.
* Import Data into Databricks:
* Extract data from its source, such as files, APIs, or other databases.
* Clean, structure, and define the schema for the retrieved data before converting it to a Databricks-compatible format.
* Load the converted data into Databricks via efficient techniques such as Spark Data Frames or custom libraries.

**Data Cleaning:**

* Recognize and correct the differences in the imported data, missing values, and formatting issues.
* To assure data quality and integrity, use data cleansing procedures such as filtering, imputation, and normalization.
* Validate the cleaned data to ensure it fulfills the processing and downstream analysis requirements.

Merge Tables as per Requirement:

* Join or combine appropriate data tables based on unique business requirements and analytical goals.
* Using suitable join techniques such as inner, left outer, right outer, or full outer depending on the required relationship between tables.
* Create an established dataset containing all required information for future analysis or reporting.

Export / Create a New Table within Snowflake:

* Return the combined and processed data to Snowflake for long-term storage and access.
* Choose the best export method for your needs depending on volume and performance, such as large loading or streaming.
* To eliminate unnecessary export procedures, build a new table within Snowflake straight from the processed data.
* Ensure the newly created or modified Snowflake table has adequate schema definition and data integrity.

**Utilizing Snowflake as Data Lake:**

In Snowflake, the process of establishing a new database involves utilizing the "Create New Schema" option within the Snowflake platform. This functionality enables users to define and designate a database, with the chosen nomenclature for this instance being "DATABRICK." This step initiates the creation of a structured environment within Snowflake, tailored to house and organize data under the specified database name. This method ensures a systematic approach to database management, enhancing overall data organization and accessibility within the Snowflake data warehouse.

A screenshot of a computer

Description automatically generated

We will be uploading the IMDB dataset into the public schema in the tables section.

**A screenshot of a phone

Description automatically generated**

Following the establishment of the "DATABRICK" database in Snowflake, the subsequent phase involves the creation of schemas for each table within the database. This procedural step is pivotal in structuring the organization of data within Snowflake, as it allows for the systematic categorization and arrangement of tables. Once the schemas are defined, the next course of action is to upload the respective tables into Snowflake, ensuring that the data is seamlessly integrated into the designated schemas. This methodical approach not only streamlines data management but also facilitates efficient querying and analysis within the Snowflake data warehouse environment.

A screenshot of a computer

Description automatically generated

**Establish connection with Snowflake Database:**

A white rectangular object with a white border

Description automatically generated

Connect to the Snowflake database securely using the necessary credentials and procedure for authentication. Configure the database connection for future data transactions.

This code describes all the credentials and settings required to securely connect to a specific Snowflake database, database schema, and warehouse for data processing.

A screenshot of a computer

Description automatically generated

**Import Data into Databricks:** Extract data from its source, data brick database in our case.

Clean, structure, and define the schema for the retrieved data before converting it to a Databricks-compatible format. Load the converted data into Databricks.

A screenshot of a computer

Description automatically generated

The code successfully reads all columns from Snowflake's "**rating**" table and saves the data as a Spark Data Frame called **df\_rating.** The output indicates that the data frame contains essential information such as TCONST (movie title ID), AVERAGE RATING, and NUMVOTES.

A screenshot of a computer

Description automatically generated

This code sample uses the “**spark.read”** method with the “**snowflake”** format parameters to read the data frame from Snowflake. It then defines different Snowflake connection parameters, such as the query to run (**select** **\*** **from title**;) and the Snowflake authentication credentials. Finally, it uses the “**display**” function to display the loaded data frame.

**Data Cleaning:** We Recognized and corrected the differences in the imported data, missing values, and formatting issues. To assure data quality and integrity, we used data cleansing procedures such as filtering, imputation, and normalization. Validate the cleaned data to ensure it fulfills the processing and downstream analysis requirements.

A screen shot of a computer program

Description automatically generated

This stage verifies that the data is complete by removing incorrect rows and replacing missing values with appropriate defaults or "Unknown" placeholders. This prepares the data for further examination and minimizes the possibility of mistakes or misleading results due to missing data.

**Merge Tables as per Requirement**: Join or combine appropriate data tables based on unique business requirements and analytical goals which is ‘TCONST’ in our case which is the primary key.Using suitable join techniques such as inner join which is depending on the required relationship between tables.We Created an established ‘final\_dataset’ dataset containing all required information for future analysis or reporting.

A screenshot of a computer

Description automatically generated

This shows how the join processes effectively integrated data from separate tables based on similar IDs.

**Export / Create a New Table within Snowflake:**Initially we created a table within snowflake having the required schema. Then we return the combined and processed data within databricks to Snowflake for long-term storage and access. We Choose the best export method for our needs depending on volume and performance, such as large loading or streaming.

To eliminate unnecessary export procedures, we built a new table within Snowflake straight from the processed data. Ensuring the newly created or modified Snowflake table has adequate schema definition and data integrity.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In the end, this step uses the append mode to write a DataFrame named **df\_final** to a Snowflake table named "**Final\_Dataset**". The Spark session is then stopped.

After this stage, we are finished using Data Bricks Notebook.

**Establishing Connection between PowerBI and Snowflake.**

We establish a connection between powerBI and Snowflake using the connector present within the PowerBI AP.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

After successfully fetching the data from our datalake which is snowflake we create a dashboard within powerBI to draw insights from the data.

A screenshot of a computer screen

Description automatically generated

**Here are some key findings from our IMDb Dashboard:**

Number of viewers, movies, and languages:

* The dashboard's top part displays the number of viewers, movies, and languages on IMDb. IMDb now has 84,04K movies listed in 83 languages.
* Percentage of movies by region: The pie chart on the left shows the percentage of movies on IMDb by region. The United States has the most movies listed, accounting for 9.74 %. Other countries with a substantial number of films include Canada (4.14%), and France (4.62%).
* Average rating and vote totals by region: The table on the right displays the average rating and number of votes on IMDb for movies by area. When we choose a country by its name, we may examine the specifics of its average rage. (**The specifics of the second case above, The United States, have been highlighted as an example.)**
* The film with the highest rating and the most votes: The bottom area of the dashboard displays the film with the highest IMDb rating and number of votes. "Der doppelte Nikolaus" (1964) had the highest rating, with a 9.7 out of 10.

**Comments:**

Pros: Efficient data processing with PySpark, secure data storage in Snowflake, interactive visualization with Power BI.​

Cons: Initial setup complexity, potential data quality issues, learning curve for new tools.​

Challenges: Handling large datasets, ensuring data accuracy, designing effective visualizations.

**Conclusion:**

In conclusion, this comprehensive project centered around leveraging Big Data technologies to extract valuable insights from a vast IMDb dataset has proven to be a successful and insightful endeavor. The integration of Databricks and Snowflake provided a robust framework for handling the extensive dataset efficiently, from data fetching to cleaning, preprocessing, and analysis. The synergy between PySpark-powered Databricks and Snowflake's secure and flexible storage solution facilitated a seamless data processing workflow.

The IMDb Non-Commercial Datasets, comprising over a million records in various tables, served as a rich source for exploration, allowing us to delve into trends, correlations, and patterns within the film industry. By employing Snowflake as a data lake, we successfully uploaded the dataset, created schemas, and established a well-organized database environment named "DATABRICK."

The project's aim was achieved through meticulous steps, including connecting to Snowflake securely, importing data into Databricks, cleaning and structuring the data, merging tables based on business requirements, and exporting the processed data back to Snowflake. The entire process adhered to best practices in data cleaning, quality assurance, and schema definition, ensuring the integrity of the final dataset.

The subsequent connection between Snowflake and Power BI allowed for the creation of an insightful dashboard. The Power BI dashboard provided an intuitive and visually appealing representation of key findings from the IMDb dataset. Viewership statistics, movie counts, language distribution, and regional movie percentages were among the key insights visualized. The dashboard also highlighted average ratings and vote totals by region, offering a detailed analysis for informed decision-making.

The project showcased the power of advanced Big Data processing tools in transforming raw data into actionable insights. The collaboration between Databricks, Snowflake, and Power BI demonstrated the effectiveness of an end-to-end data processing and visualization pipeline. The insights drawn from the IMDb dataset not only contribute to a deeper understanding of the film industry but also empower movie production teams with valuable information for strategic decision-making. Overall, this project exemplifies the potential of cutting-edge technologies in unlocking meaningful insights from vast datasets, marking a successful and impactful data-driven exploration.

**Reference:**

**(Dataset link )IMDb Datasets. (n.d.). Retrieved from** [**https://datasets.imdbws.****com/**](https://datasets.imdbws.com/)

Databricks Community. (n.d.). Notebook: [Big-Data-Final-Project-Group-Delta]. Retrieved from <https://community.cloud.databricks.com/?o=7746998554581331#notebook/3182721693671259/command/3182721693671260>

Power BI. (n.d.). Report: [Big-Data-Final-Project]. Retrieved from <https://app.powerbi.com/groups/me/reports/83dececc-b9ec-4250-9321-afbb2dd2f095/ReportSection?experience=power-bi&redirectedFromSignup=1>

Snowflake. (n.d.). Snowflake | Data Warehouse Built for the Cloud. Retrieved from <https://www.snowflake.com/en/>

Intellipaat. (n.d.). 10 Big Data Examples: Application of Big Data in Real Life. Retrieved from <https://intellipaat.com/blog/10-big-data-examples-application-of-big-data-in-real-life/>

Adminran, & Adminran. (2023, April 17). What are real-life examples of the application of big data analytics? - Magnimind Academy. Magnimind Academy - Launch a new career with our programs. <https://magnimindacademy.com/blog/what-are-real-life-examples-of-the-application-of-big-data-analytics/>

Muskan. (n.d.). Big data in media and entertainment industry | Analytics Steps. <https://www.analyticssteps.com/blogs/big-data-media-and-entertainment-industry>

Selerity. (2023, November 7). Big data analytics platforms and the entertainment industry. Selerity. <https://seleritysas.com/2020/10/09/big-data-analytics-platforms-and-the-entertainment-industry/>