# Summary

[Summary 1](#_Toc18716582)

[Executive Summary 1](#_Toc1789148610)

[Project Overview 1](#_Toc1911952379)

[Rationale 1](#_Toc975080006)

[Benefits 1](#_Toc1802167715)

[Achitecture 2](#_Toc1303843420)

[Project Implementation Stages 3](#_Toc156312044)

[1. Environment Setup 3](#_Toc163279347)

[2. Manual Data Loading to the Unity Catalog 3](#_Toc1216028323)

[3. Web Scraping to Obtain URLs 4](#_Toc481346505)

[4. Ingestion to the Bronze Layer 4](#_Toc726403586)

[5. Cleaning and Aggregation to the Silver Layer 4](#_Toc1558630502)

[6. Table Modeling for the Gold Layer 4](#_Toc904779887)

[7. Job Workflow Creation 4](#_Toc1250836897)

[8. Documentation and AI Data Accessibility 5](#_Toc315829975)

[Project Conclusion 5](#_Toc502024885)

[References 6](#_Toc148017721)

# Executive Summary

## Project Overview

The "Brazilian Tax Big Numbers End-to-End Data Application" project is a data engineering initiative that establishes an ETL (Extract, Transform, Load) pipeline to process and analyze public data on Individual Income Tax in Brazil. The implemented architecture is a Lakehouse, combining the flexibility of a data lake with the structure and governance of a data warehouse, using Databricks as the central platform. The project adopts a modular development approach and is integrated with GitHub for version control and workflow management.

## Rationale

The technical need for this project lies in the complexity and volume of the raw data, which requires structured treatment to become useful for analysis. The construction of a robust, automated data pipeline is essential to ensure data quality and reliability. The Lakehouse architecture is justified by providing a scalable and governed environment, allowing data to be progressively transformed through layers, from the raw state to a format optimized for consumption by AI Assistants (Genie AI) and BI tools.

## Benefits

Expected benefits include:

* Efficiency and Quality: The pipeline's automation, orchestrated by Databricks Workflows, ensures data is processed consistently, reducing manual errors and guaranteeing quality.
* Organization and Governance: The tiered structure (Bronze, Silver, and Gold), managed by the Databricks Unity Catalog, creates an organized and easily navigable data environment for future development and analysis.
* Optimization for AI and BI: The Gold layer provides modeled and aggregated data, ready for querying by artificial intelligence assistants like Genie, and for the creation of high-performance analytical dashboards and reports.

# Achitecture

The project adopts the Lakehouse architecture for a strategic reason: it resolves the dilemma between the flexibility of a data lake and the structure of a data warehouse. Instead of choosing between the two, the Lakehouse architecture allows us to economically store all raw data in a data lake while applying a layer of governance and ACID (Atomicity, Consistency, Isolation, and Durability) transactions to ensure data quality, reliability, and integrity, which are essential characteristics of a data warehouse.

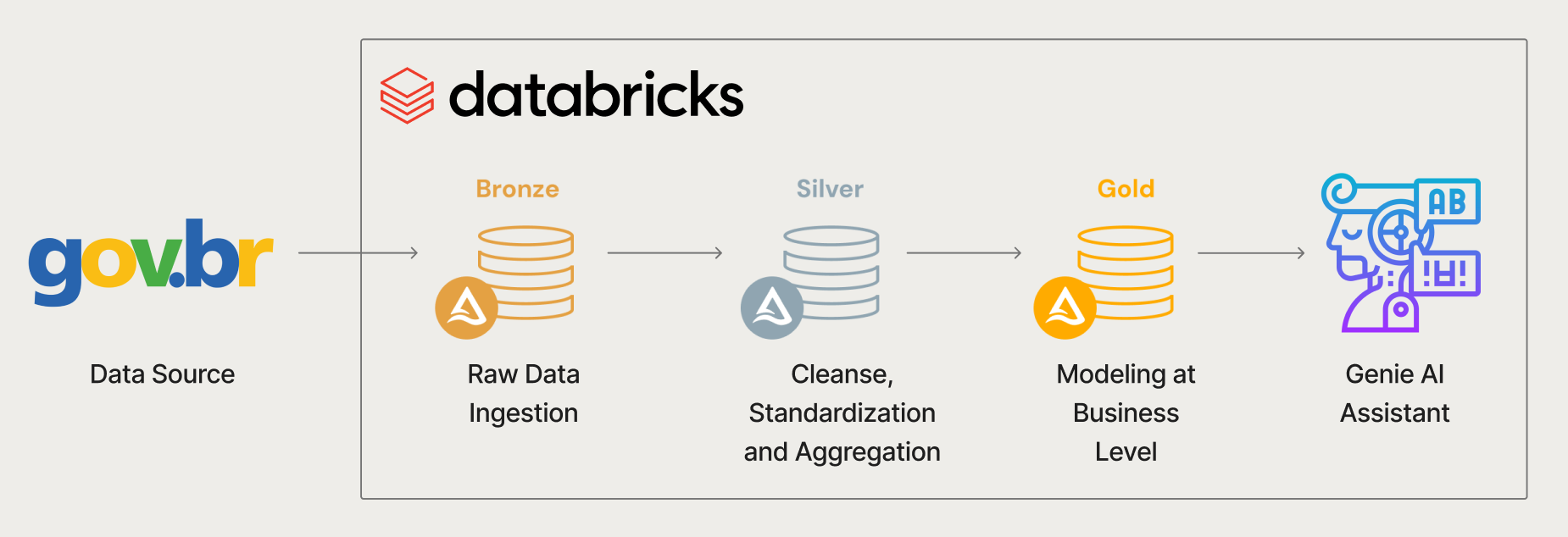
The Databricks Unity Catalog is the central piece of this architecture, acting not only as a governance system but also as a unified data catalog. The decision to implement it was motivated by the need for a single source of truth for metadata, access control, and data lineage, facilitating the secure and auditable discovery and sharing of data throughout the organization. It ensures that schema definitions, permissions, and descriptions are consistent across all pipeline layers.

The layered structure was chosen to follow data engineering best practices, creating a robust and modular pipeline:

* source\_files: This initial storage layer serves as an immutable staging area. Its existence ensures that the original raw data files are never altered, which is crucial for auditing and for the possibility of reprocessing data from the source, if necessary.
* bronze\_layer: The purpose of this layer is pure ingestion. Data is read from the staging area and converted to the Delta Lake format, maintaining the original structure and content. The use of the Delta Lake format here is fundamental as it offers ACID transactions and features like time travel, ensuring the integrity of the raw data from the moment of ingestion.
* silver\_layer: This is the data "curation" layer. The decision to create this layer was based on the need to clean, standardize, and enrich the data to remove inconsistencies and transform it into a reliable, high-quality resource for consumption.
* gold\_layer: The purpose of this layer is data modeling. The tables here are optimized for performance and designed to meet specific business use cases, such as reports and dashboards. Modeling in this layer reduces complexity and query time for end-users, such as data scientists and BI analysts.

A key technical decision was the integration of the Databricks Genie AI Assistant to democratize access to the curated data. This feature allows business users and analysts to interact with the gold\_layer tables using natural language. The AI assistant automatically translates these natural language queries into SQL commands, which are executed against the optimized tables. This approach eliminates the need for non-technical users to learn SQL, enabling them to quickly gain insights and drive data-informed decisions by simply asking questions.

The use of GitHub for code versioning, in conjunction with Kanban in GitHub Projects for workflow management, was a technical decision to promote collaboration, transparency, and project reproducibility. The direct connection of the repository to Databricks enables a Continuous Integration (CI) approach, where the code can be versioned, tested, and deployed in an automated way.



# Project Implementation Stages

## Environment Setup

Objective: To structure the development and production environment.

Description: A GitHub repository (Brazilian-Tax-Big-Numbers-End-to-End-Data-Application) was created, and a Kanban board was configured for task management. The repository was connected to a Git Folder in Databricks. The brazilian\_tax\_big\_numbers catalog was created in the Unity Catalog, along with the source\_files, bronze\_layer, silver\_layer, and gold\_layer schemas, and the source\_files volume to store the source files.

## Manual Data Loading to the Unity Catalog

Objective: To ingest raw data files into the platform.

Description: The CSV files from the dados.gov.br portal were manually loaded into the source\_files volume. An initial exploration was performed to assess the volume and superficial structure of the data, such as the number of rows and columns.

## Web Scraping to Obtain URLs

Objective: To automate the collection of data file download URLs.

Description: A web scraping script was developed to extract the table names from the source data website. These names were formatted to dynamically build the download URLs. The script downloaded and saved the data files in the source\_files volume.

## Ingestion to the Bronze Layer

Objective: To ingest the raw data from the staging area into the Delta Lake format.

Description: A dynamic ingestion notebook was created to read the CSV files from the source\_files volume, normalize the column names (by removing spaces, accents, and special characters), and save the data as Delta tables in the bronze\_layer. An orchestrator notebook was developed to manage the ingestion process, reading the list of tables and processing each one sequentially.

## Cleaning and Aggregation to the Silver Layer

Objective: To clean, standardize, and enrich the data.

Description: Several notebooks were created to process the tables from the bronze layer. Operations included cleaning null rows, defining schemas, standardizing columns, handling missing values, and aggregating by common identifiers. The resulting tables, such as delta\_bens\_dividas\_pagamentos, delta\_rendimentos, and delta\_municipio\_e\_tipo\_de\_declaracao, were saved in the silver\_layer.

## Table Modeling for the Gold Layer

Objective: To model data for analysis and consumption by Business Intelligence (BI) tools.

Description: In this step, the tables from the silver layer are consumed and joined to create wide tables and summary tables. These tables are designed to answer specific business questions and are saved in the gold\_layer, optimized for performance in analytical queries and dashboards.

## Job Workflow Creation

Objective: To orchestrate and automate the execution of the ETL pipeline.

Description: Databricks job workflows were configured to automate the execution of the notebooks. A master pipeline was created to execute the ingestion (from the volume to the bronze layer), transformation (from the bronze to the silver layer), and modeling (from the silver to the gold layer) pipelines in the correct order, ensuring a continuous and modular data flow.

Automation and Pipeline Creation Code: To facilitate automation, the code for creating the pipelines has been made available. It is crucial for the user to change variables such as notebook\_path and cluster\_id, as they are specific to each Databricks execution environment. Furthermore, the order of pipeline creation is fundamental to the success of the data flow, following the sequence: data source ingestion → ingestion to the Bronze Layer → Silver Layer transformation → Gold Layer data modeling → End-to-End Pipeline, that orchestrates all the steps in a modular way.

## Documentation and AI Data Accessibility

Objective: To make the data accessible and understandable for business users.

Description: Documentation was implemented in the Unity Catalog, using detailed descriptions and tags to enrich the table metadata. This initiative not only facilitates data governance and discovery for developers, but also serves as a knowledge base for the Databricks Genie AI Assistant. This tool, integrated into the platform, allows users without SQL knowledge to ask questions about the data using natural language. Genie automatically translates the requests into SQL queries, providing direct answers and support for frequently asked questions (FAQ), democratizing data access and accelerating the decision-making process.

# Project Conclusion

The "Brazilian Tax Big Numbers End-to-End Data Application" project represents a robust and complete implementation of a data engineering pipeline. By using a Lakehouse architecture on Databricks, the initiative demonstrates the ability to transform complex and raw government data into a valuable and accessible resource. The modular approach, with Bronze, Silver, and Gold layers, ensures data quality, governance, and optimization for analytical consumption.

The automation of the workflow through Databricks Workflows ensures efficient and consistent processing, while the integration with GitHub promotes good practices for code versioning and team collaboration.

The highlight of the project is the implementation of detailed documentation in the Unity Catalog and the integration with the Genie AI Assistant. This final step democratizes data access, allowing business users and analysts without technical SQL knowledge to extract valuable insights using only natural language. Ultimately, the project not only solves a technical problem of large-scale data processing but also empowers the organization to make faster, more informed decisions, making data a strategic asset for everyone.

# References

<https://github.com/GabrielFePL/Brazilian-Tax-Big-Numbers-End-to-End-Data-Application>

<https://github.com/users/GabrielFePL/projects/5/views/1>

<https://dados.gov.br/dados/conjuntos-dados/grandes-nmeros-do-imposto-de-renda-da-pessoa-fsica>