# Business Requirements Document (BRD) for Economic Data Analysis Project

## Project Overview

This project is designed to showcase a sophisticated data engineering workflow that utilizes a medallion architecture for data processing and storage. The workflow will integrate economic data into a DuckDB database, apply transformations with dbt-core, ensure data quality with Soda Core, and use Apache Airflow for workflow orchestration. The final transformed and validated data will be exported to Google Sheets, serving as a source for analytics and visualization in Tableau Public. This setup demonstrates comprehensive skills in data extraction, transformation, quality assurance, and visualization.

## Objectives

* **Medallion Architecture Implementation:** Employ a tiered data storage and processing architecture to organize data into bronze (raw), silver (cleaned/semi-processed), and gold (aggregated/transformed) layers within DuckDB.
* **Data Quality Assurance:** Integrate Soda Core to monitor and ensure the quality of data at each stage of the medallion architecture.
* **Data Visualization:** Utilize Google Sheets as an intermediary data storage for feeding into Tableau Public for visualization, showcasing the ability to derive insights from processed data.

## Requirements

### Data Source and Extraction

* **REST API for Economic Data:** Periodically pull economic data using a REST API.

### Database and Data Management

* **DuckDB:** Use DuckDB to implement the medallion architecture, segregating data into bronze, silver, and gold layers according to their processing stages.
* **dbt-core:** Apply dbt-core for data transformation, ensuring that data is processed and moved from bronze to silver and finally to gold layers as it is cleaned, aggregated, and made ready for analysis.

### Data Quality

* **Soda Core:** Incorporate Soda Core to continuously monitor and validate data quality across all layers of the medallion architecture, ensuring the integrity and reliability of the data pipeline.

### Data Workflow Orchestration

* **Apache Airflow:** Leverage Apache Airflow to orchestrate the data workflow, including extraction, loading, transformation, quality checks, and the final export to Google Sheets.

### Intermediate Storage and Visualization

* **Google Sheets:** Export the gold layer data to Google Sheets, enabling easy integration with Tableau Public for visualization purposes.
* **Tableau Public:** Use Tableau Public to create interactive dashboards and visualizations from the data stored in Google Sheets, demonstrating analytical insights.

### Containerization

* **Docker (Optional):** Consider containerizing the entire workflow, including DuckDB, dbt-core, Soda Core, and Apache Airflow components, for portability and ease of deployment.

### GitHub Repository

* **Comprehensive Documentation:** Provide thorough documentation covering the project setup, architecture, workflow, and instructions for replication.
* **All Necessary Resources:** Include source code, dbt models, Airflow DAGs, Soda Core checks, and any Docker configuration files.

## Implementation Plan

1. **Setup and Configuration:** Configure the initial setup for DuckDB, dbt-core, Soda Core, and Apache Airflow.
2. **API Data Extraction:** Implement functionality to periodically extract data from the REST API.
3. **Medallion Architecture Implementation:** Structure the DuckDB database to accommodate the bronze, silver, and gold layering of data.
4. **Data Transformation and Quality Checks:** Develop dbt models for data transformation and Soda Core checks for each stage of data processing.
5. **Orchestration and Automation:** Use Apache Airflow to automate the entire workflow, from data extraction to exporting data to Google Sheets.
6. **Visualization Setup:** Prepare and connect Google Sheets with Tableau Public for data visualization.
7. **Documentation and Sharing:** Finalize documentation and share the complete project on GitHub.

## Assumptions and Constraints

* **Technology Familiarity:** Assumes proficiency with DuckDB, dbt-core, Soda Core, Apache Airflow, Google Sheets, and Tableau Public.
* **Data Volume:** Designed for handling small to medium-sized datasets, suitable for processing within the capabilities of the selected technologies and Google Sheets.
* **Public Visualization:** Recognizes that data visualizations will be public on Tableau Public, considering any data sensitivity issues.