RevEnergy

**(Data Engineering Project)**

Center of Excellence

16-Aug-2023

Version 1.0

Copyright ©: 2023, Revature.

 Table of Contents

[Application Overview 3](#_Toc1555815592)

[Core Functional Requirements 3](#_Toc1194330873)

[Standard Functional Scope 3](#_Toc1763839213)

[Definition of Done 4](#_Toc2020375305)

[Competency-wise scoping 4](#_Toc1523109773)

[Non-Functional Expectations 6](#_Toc1436798500)

​​

​​

​

​

​

​

​

​​

​

​

​

​

​

​​

# Application Overview

The RevEnergy project offers insights into the trends of global energy consumption and production in various categories like coal, oil & gas, biofuel, nuclear, solar and, wind etc. This data can be utilized to generate an analysis report of energy consumption and production in the world, Identify the countries that cause the highest and least environmental impact and also to understand the changes in the carbon emissions and the deviation from fossil fuels to more environmentally friendly energy production and consumption methods implemented by the various counties in the world for over 100 years. It can also be used to understand how various causes like government policies, regulations and wars affected the trends in energy consumption and production. This will also help us to create a 10, 50 or 100 years plan for the countries to transition from fossil fuels to environmentally friendly options.

# Core Functional Requirements

As a User, I want to:

**Data Aggregation:**

1. Integrate the iso code, country, and year columns from multiple data sources into a single unified dataset to enable comprehensive cross-country energy analysis.
2. Combine energy production data with energy consumption data to assess the overall energy self-sufficiency and dependency of different countries.

**Data Analysis:**

1. Calculate the average annual growth rate and reduction rate of energy production of coal, oil & gas, biofuel, nuclear, solar, wind, etc.

**Data Mining:**

1. Identify countries with the highest percentage increase in coal production to understand emerging trends and potential environmental implications.
2. Analyze the correlation between changes in gas production and changes in oil production to uncover any potential patterns or dependencies.

**Data Visualization:**

1. Visualize the change in energy production of coal, oil & gas, biofuel, nuclear, solar, wind, etc. (both percentage and absolute) over the years for different countries, allowing stakeholders to easily compare trends.
2. Visualize the chart that illustrates the contribution of gas, oil, and other energy sources to the total energy produced for select countries, enabling a clear understanding of energy source distribution.
3. Visualize chart to depict the percentage change in energy consumption for various countries, helping to highlight regions with significant shifts in energy usage.

# Standard Functional Scope

1. Ingesting and processing various energy consumption data feeds.
2. Transforming data into analytical models.
3. Providing interactive dashboards, queries, and reports for analysis.
4. Implementing administrative controls for user management and data governance.

This scope ensures efficient data handling, advanced analysis, and user-friendly tools, facilitating the seamless transformation of raw energy consumption data into valuable insights for informed decision-making.

# Definition of Done

1. Working application demonstration.
2. Sharing the associates’ code repo for technical review with:

* Architecture
* data models
* ETL documentation

# Competency wise scoping

|  |  |  |
| --- | --- | --- |
| **Competency** | **Project Type** | **Expectations** |
| Python, SQL  Scala, SQL | REST API service | **Framework Specific**   1. Ensure the appropriate APIs are used for any of the API calls. 2. Ensure the routing is centrally configured 3. Best practices & design patterns are to be followed.     **Validation and Error Handling:**   1. Validate the inputs for their types and format. 2. Display functional-related user messages (either for input/error/output) - no system error codes or SQL error codes. 3. Handle the exceptions and errors gracefully.     **Logging:**   1. Ensure the application is using proper logging framework and methods. 2. Ensure the application’s log level is configured using configuration files so that it can be changed without changing the code. 3. Also ensure that the application logging is configured to output to the mentioned log file.     **Testing**:   1. Ensure sufficient test cases are written using appropriate testing frameworks. 2. Ensure the code coverage closed to be 80%     **Security**:   1. Ensure the SQL injection threat is taken care. 2. Ensure the CORS restriction is applied, if applicable. 3. Ensure that the secrets are stored as environment variables using secure credential storage.   **Coding Standard:**   1. Use the industry coding standards and conventions. 2. Modular based code development for better reusability. 3. Ensure proper usage of resource objects such as database connectivity objects to avoid resource leakages. 4. Ensure proper usage of design patterns and application layering (such as Business Service, DAO Layer etc.) wherever applicable. |
| Spark  Hadoop, Hive | Data Science | **Data Preprocessing:**   1. Load the world energy consumption dataset from a reliable source. 2. Handle missing data, outliers, and any data quality issues appropriately. 3. Perform data cleansing, normalization, and transformation to prepare it for analysis.   **Exploratory Data Analysis (EDA):**   1. Generate descriptive statistics to understand the dataset's basic characteristics. 2. Visualize energy consumption, production by oil & gas, coal, and other relevant attributes over time. 3. Identify patterns, correlations, and anomalies in the data through various charts and graphs.   **Data Visualization:**   1. Design interactive visualizations (line charts, bar charts, pie charts, etc.) to represent coal production trends, energy source distribution, and other relevant insights. 2. Ensure the visualizations are user-friendly, allowing users to explore and interact with the data to gain deeper understanding.   **Reporting and Presentation:**   1. Generate detailed reports summarizing the findings from the data analysis and mining processes. 2. Create presentations with clear explanations of insights, trends, and correlations for stakeholders' understanding.   **User Interaction and Exploration:**   1. Provide user interfaces or dashboards that allow stakeholders to interact with visualizations and customize views based on their interests. 2. Enable filtering, sorting, and comparison functionalities within the visualizations to facilitate data exploration. |
| Azure  AWS  GCP | ETL Pipeline | **Deployment artifacts:**   1. The deployment artifacts should be minified and obfuscated if required.   **Pipeline Orchestration**   1. Scheduled ETL jobs monitor, orchestrate and restart flows 2. Job monitoring and alerts for data quality and SLAs 3. Scalable compute resources based on workload   **Scalability and Performance:**   1. Design the ETL pipeline to handle large volumes of stock data efficiently. 2. Optimize data processing and transformation to minimize execution time.   **Monitoring and Alerts:**   1. Set up monitoring for ETL job status, performance metrics, and data quality. 2. Configure alerts to notify stakeholders about failures or anomalies in the ETL process. |

# Non-Functional Expectations

* Application development should use version control systems (e.g., Git) to manage the project codebase and facilitate collaboration.
* Application development is supposed to follow the Scrum process.

# Source Data Location

The data can be acquired from Kaggle using:

* Link to the csv file  [World Energy Consumption.csv (sharepoint.com)](https://revature0.sharepoint.com/:x:/r/sites/trainers/_layouts/15/Doc.aspx?sourcedoc=%7BA2B12687-37A8-4973-A24D-CDB64BB27D97%7D&file=World%20Energy%20Consumption.csv&action=default&mobileredirect=true)