**Data Quality Control**

**Project Title:** Data Quality Workflow Implementation for Business License Data Validation

**Project Overview:**  
This project focused on developing an automated data quality workflow to assess and validate business license datasets using AWS Glue Studio. The objective was to streamline the data validation process by separating clean records from flawed ones and storing them in designated S3 locations for further use or review.

**Project Background:**

As part of a broader data analysis initiative, it became necessary to ensure the accuracy and reliability of business license records being processed from public datasets. Issues such as missing values, inconsistent formats, and inaccurate entries had been observed, which could negatively impact downstream analytics. This project introduced a structured quality check using AWS services to clean and validate the dataset efficiently.

**Scope:**

The solution was designed to target the following key functions:

* **Data Profiling:** Analyzing the raw input dataset to detect anomalies or inconsistencies.
* **Automated Validation:** Applying transformation rules to assess field-level quality and route data accordingly.
* **Conditional Routing:** Using AWS Glue's visual editor to direct valid and invalid data through separate paths.
* **Data Storage:** Writing the results to Amazon S3 — separating “passed” (clean) and “failed” (problematic) records.
* **Traceability:** Ensuring output can be reviewed and traced for transparency and auditing.

**Methodology:**

1. **Input Data Connection**
   * Connected to an S3 bucket (license-status) as the source containing raw license data.
2. **Transformation & Quality Check**
   * Applied transformations to check quality conditions (e.g., null checks, formatting).
   * Used conditional logic to split valid and invalid records using the Glue Studio **“Conditional Router”** node.
3. **Schema Adjustment**
   * Applied schema mapping to align column formats for both output streams.
4. **Final Output**
   * Wrote “passed” data to the **S3 path: Quality Check/passed/**
   * Wrote “failed” data to the **S3 path: Quality Check/failed/**
5. **Validation & Verification**
   * Verified outputs in S3 to confirm proper routing and separation.
   * Compared file sizes and examined content to validate effectiveness of the filtering logic.

**Tools & Technologies Used:**

* **AWS Glue Studio (Visual Editor)** for ETL pipeline creation
* **Amazon S3** for raw input and validated output storage
* **Python (PySpark inside Glue)** for custom transformations (optional)
* **AWS Console** for workflow monitoring and output verification

**Deliverables:**

* A fully functional Glue Job named **busi-lic-quality check** with a visual transformation flow
* Two separate datasets stored in S3 (passed/ and failed/ folders)
* A scalable framework that can be reused for future data quality validation tasks
* Screenshots documenting each step of the workflow and output validation

**Timeline:**

This mini-project was completed in under one week, including configuration, execution, validation, and result documentation.

**Outcome:**

This project ensured that only clean and usable data flowed into subsequent analysis steps, improving data accuracy, reliability, and decision-making based on business license records. It also helped create a repeatable framework for quality checks on similar structured datasets.