**Procedures Document for Federal Spending Tracker Tool**

**Procedures for Federal Spending Tracker Tool**

**Overview**

This document outlines the detailed procedures followed in the ETL pipeline project developed for the client. The project focuses on extracting data from multiple sources, transforming it, and loading it into a PostgreSQL database. The workflows are managed using Apache Airflow, leveraging AWS services for storage and processing.

**Step-by-Step Procedures**

**Pipeline 1 & 2: Webpage PDF Data Extraction (ETL\_financial\_reports, ETL\_history\_reports)**

1. **Step 1: Data Extraction**
   * Scrapes the PDF files from the designated web pages.
   * Saves the scraped PDF files in the AWS S3 raw bucket.
2. **Step 2: Data Processing**
   * Uses AWS Textract to extract relevant information from the PDF files.
   * Process the extracted data into a structured format (CSV).
   * Saves the processed data in the S3 processed bucket.
3. **Step 3: Data Loading**
   * Fetch the processed CSV files from the processed bucket.
   * Ingest the data into AWS RDS PostgreSQL tables.

**Pipeline 3: USA Spending Data Extraction (ETL\_usa\_spending)**

1. **Step 1: Data Extraction**
   * Use Python's requests library to make GET requests to the API.
   * Store the retrieved data in JSON format in the AWS S3 raw bucket.
2. **Step 2: Data Processing**
   * Transform the JSON data as needed (e.g., normalizing nested fields).
   * Save the processed data in the S3 processed bucket.
3. **Step 3: Data Loading**
   * Load the processed data into PostgreSQL tables.

**Pipeline 4: Quarterly Performance Reports (ETL\_quarterly\_performance\_reports)**

1. **Manual Task**
   * Manually download the PDF files from the provided sources and upload specific pages to the S3 raw bucket.
2. **Step 1: Data Extraction and Processing**
   * Extract and process the data from the uploaded PDF pages using AWS Textract.
   * Save the processed output in the processed bucket.
3. **Step 2: Data Loading**
   * Load the processed data into PostgreSQL tables.1. **Step 1: Data Extraction**
   * Extract USA Spending reports and quarterly performance reports using relevant extraction scripts.
   * Store the raw data in the AWS S3 raw bucket.
4. **Step 2: Data Processing**
   * Process the extracted data to convert it into a structured format.
   * Save the processed data in the S3 processed bucket.
5. **Step 3: Data Loading**
   * Ingest the processed data into AWS RDS PostgreSQL tables.

**Scheduling**

* All ETL pipelines can be scheduled to run daily or monthly to ensure up-to-date data is available.

**Error Handling**

* **Data Extraction Failures**: Retry logic is implemented for web scraping and API requests to handle intermittent network issues.
* **Data Transformation Errors**: Validation checks are applied to processed data, and any discrepancies are logged for further analysis.

**Security Considerations**

* AWS credentials are stored securely using Airflow's connections feature.
* Data in S3 is encrypted, and access is restricted through IAM policies.

**Manual Intervention Points**

* For the quarterly performance reports pipeline, manual intervention is needed to download and save specific pages from PDF files to the S3 bucket (raw).

**Lessons Learned**

* Automation of manual tasks can significantly enhance pipeline efficiency.
* Incorporating data quality checks early in the pipeline prevents downstream issues.