# Reddit Data Pipeline Documentation

## Overview

This document outlines the process of creating a data pipeline for extracting, transforming, and loading Reddit data into AWS services, including S3, Glue, and Redshift. The steps involve setting up AWS resources, running ETL jobs, and integrating with Redshift for data warehousing and analytics.

## Step 1: Setting Up Docker and Airflow

The project began by setting up Docker and Apache Airflow (version 2.10) to orchestrate the data pipeline. Used Python 3.9 for scripting and created Airflow DAGs to handle job automation.  
A screenshot of a computer

Description automatically generated

## Step 2: Extracting Data from Reddit

Leveraged the Reddit API to extract posts from specific subreddits. Created Python scripts within the Airflow DAGs to extract this data dynamically based on subreddit filters and timeframes.

A screenshot of a computer

Description automatically generated

## Step 3: Storing Data in S3

The extracted data was saved as CSV files and loaded into an AWS S3 bucket using Airflow tasks. The raw data was stored in a dedicated `raw` folder within the S3 bucket.  
A screenshot of a computer

Description automatically generated

## Step 4: Transforming Data Using AWS Glue

Used AWS Glue to process and transform the data:  
- Concatenated three columns (`edited`, `spoiler`, and `stickied`) into a new column named `ESS\_updated`.  
- Dropped the original columns after concatenation.  
- Saved the transformed data in the `transformed` folder in the same S3 bucket.  
A screenshot of a computer

Description automatically generated

## Step 5: Validating Data with Athena

Queried the transformed data using Amazon Athena to ensure the transformations were correctly applied. Validated the data schema and structure before loading it into the data warehouse.  
A screenshot of a computer

Description automatically generated

## Step 6: Loading Data into Redshift

Set up an Amazon Redshift cluster and defined the namespace for the data warehouse. Created a schema in Redshift to store the transformed data. Loaded the data from S3 to Redshift using the `COPY` command.

## Step 7: Visualizing Data

Connected to Redshift to query and analyze the data directly. Created visualizations and dashboards using Amazon QuickSight to gain insights from the Reddit data.  
A screenshot of a computer

Description automatically generated

## Notes

- Ensure that all IAM roles and policies are correctly configured.  
- Use appropriate partition keys in Glue and Redshift for large datasets.  
- Monitor and optimize Glue and Redshift for performance.