Comprehensive Overview of the Pipeline

Purpose and Scope

The data pipeline is specifically designed to handle and process water usage data for American Samoa. The primary purpose of this pipeline is to transform raw data collected in Excel into a structured format that meets the specific analytical and reporting needs of water resource management in American Samoa. The processed data is stored securely in cloud storage and is structured to comply with the United States Geological Survey (USGS) standards for data submission. This enables seamless integration and upload of the data onto the USGS website, facilitating easy access and analysis by researchers, policymakers, and the public.

Data Source and Input

- **Data Collection**: The pipeline begins with data collection, where water usage data is gathered from various sources within American Samoa. This data typically includes measurements and estimations of water usage across different sectors, such as residential, agricultural, and industrial.
- **Input Format**: The raw data is initially received in Excel spreadsheets, which are versatile and commonly used for preliminary data recording and basic analysis in many field studies. These excel spreadsheets contain sensitive information and are accessed via a private GitHub repository.

Data Transformation Processes

- **Data Cleaning**: The first step in the transformation process involves cleaning the data. This includes correcting inconsistencies, handling missing values, and removing duplicates to ensure the accuracy and reliability of the data.
- **Data Standardization**: The pipeline standardizes data by aligning it with predefined formats and classifications, such as standardizing the names of villages and other geographical identifiers to conform with USGS requirements.
- **Data Enrichment**: The process enriches the data by deriving additional information, such as calculating aggregate water usage statistics for specific regions or periods, which are critical for comprehensive analysis.
- **Data Formatting**: Finally, the data is formatted to meet specific schema requirements of the target database (USGS website), including adjustments to field names, data types, and file formats (e.g., converting Excel files to CSV for upload).

Cloud Storage and Data Management

- **DigitalOcean Space**: Once transformed, the data is uploaded to DigitalOcean Space, a highly scalable object storage service. This storage solution is chosen for its reliability, security features, and compatibility with AWS S3 APIs, simplifying integration with other data services.
- **Security and Compliance**: A priority is ensuring data security and compliance with local and international data protection regulations. The pipeline uses secure transfer protocols and implements access controls to safeguard sensitive information.

Detailed Architecture

Data Ingestion

- **Source**: The data originates from CSV files containing various attributes related to water usage, such as SiteUUID, BeneficialUseCategory, TimeframeStart, and TimeframeEnd.
- **Tools Used**: Python libraries such as Pandas for data manipulation and Boto3 for interactions with AWS-like services (DigitalOcean Spaces in this case).

Data Processing

- **Transformations**:

  Standardizing Village Names: A dictionary mapping standardizes village names within the dataset.
  Column Formatting: Specific columns are selected and processed to align with the desired format.
  Aggregating Data: Data is aggregated based on certain key columns like SiteUUID and BeneficialUseCategory to sum up numerical values such as Amount.

Data Storage

- **Storage Solution**: Processed data is uploaded to DigitalOcean Space using Boto3 configured for S3 compatibility.
- **Security**: Credentials for accessing the space are retrieved securely, and files are managed through the API.

Data Output

- **Output Format**: The final output is a CSV file uploaded to a cloud storage space, which can then be accessed or downloaded as required.

Code Explanations

Key Functions

- **Quality Assurance/Control:** The qa_qc_checks function ensures that water usage data is within three standard deviations of the mean and that well data does not exceed the maximum monthly production capable of the well.
- **Transforming Data**: A function transform_to_site_specific_format standardizes and filters the dataset based on predefined criteria determined by USGS.
- **Well Data**: The well_data function imports necessary production well information and correctly formats it.
- **Metadata Tables:** There are a suite of functions designed for creating associated metadata tables required for the Wade 2.0 Schema needed for USGS.
- **Uploading Data**: The upload_to_digitalocean_spaces function utilizing Boto3, the pipeline uploads the processed CSV file to a specified DigitalOcean endpoint space based on required environmental variables.

End-User Documentation

1. **Introduction**

- **Document Purpose**: This document aims to guide users through accessing and utilizing the transformed data produced by the water usage data pipeline.
- **Pipeline Overview**: The data pipeline processes water usage data, transforms it according to specified rules and uploads the transformed data to a DigitalOcean Space for storage and further use.

2. **Accessing the Data**

- **How to Access**: Users can access the transformed data by downloading the CSV files from the specified DigitalOcean Space. The .csvs are publicly accessible using the object endpoint space URL.
- **File Format**: The output data is in CSV format that follows the Wade 2.0 Schema and is compatible with USGS software.