**1. Overview**

This document outlines the strategic approach for ingesting data into using metadata driven Synapse pipeline, focusing on the formulation and categorization of metadata. Each data ingestion process requires a well-defined metadata structure, including a **config file** and a **meta file**, both comprising essential properties to ensure efficient and reliable data handling.

**2. Metadata Formulation (Suresh to provide the info)**

Metadata is divided into two primary files for each ingestion pipeline:

* **Config File**: Contains core configurations, including source information, target destinations, and processing logic.
* **Meta File**: Contains supplemental metadata, such as column name, data quality checks, lineage, and audit information.

**3. Criteria for Config File Categorization**

To streamline the ingestion process, config files will be grouped based on strategic criteria to optimize data flow, resource utilization, and adherence to service-level agreements (SLAs), that enhance efficiency, maintainability, and scalability. Below are the main factors and best practices for categorization:

**3.1 Data Source Categorization**

* **Objective**: Organize configuration files by their originating data sources.
* **Approach**:
  + Group configurations based on source systems (e.g., Bancs).
  + Apply source-specific validation and transformation rules.
* **Best Practice**: Maintain a registry of data sources to ensure traceability and facilitate impact analysis during source system changes.

**3.2 File Arrival Frequency**

* **Objective**: Group files based on arrival frequency (e.g., real-time, daily, weekly, monthly).
* **Approach**:
  + Real-time files can trigger continuous ingestion pipelines.
  + Daily and weekly files can follow a scheduled batch processing model.
* **Best Practice**: Create time/event-based triggers and monitor for deviations to ensure SLAs are met.

**3.3 Volume-Based Categorization**

* **Objective**: Segregate configurations based on the volume of data processed.
* **Approach**:
  + Classify as high, medium, or low volume.
  + Low-volume files can/may be batched or processed during off-peak hours to reduce resource strain.
* **Best Practice**: Define thresholds for each volume category and prioritize resource allocation for high-volume groups.

**3.3 Service-Level Agreements (SLAs)**

* **Objective**: Organize files based on the required SLA to maintain critical data availability.
* **Approach**:
  + High-priority data with strict SLAs can follow priority ingestion paths with dedicated resources.
  + Non-critical data can be processed in secondary or non-peak windows.
* **Best Practice**: Regularly review SLAs and ingestion times, optimizing pipelines to meet deadlines.

**3.4 Consumption Patterns (CDP)**

* **Objective**: Categorize data based on the consumption needs of downstream applications.
* **Approach**:
  + Frequently consumed data should be stored in readily accessible locations with low-latency processing.
  + Archive data that is accessed infrequently to optimize storage and processing.
* **Best Practice**: Implement data caching for high-frequency consumption, improving response times for analytics and reporting.

**3.7 Security and Compliance Requirements (Optional)**

* **Objective**: Categorize files based on data sensitivity and compliance needs.
* **Approach**:
  + Sensitive data can be tagged for additional security measures, like encryption and access restrictions.
  + Non-sensitive data can follow standard processing without additional security overhead.
* **Best Practice**: Implement access controls and audit trails for sensitive groups to ensure compliance.