# Data Ingestion Classification and Naming Convention Guide

## 1. Introduction

This document outlines the best practices for classifying data ingestion volumes and defining naming conventions for configurations in Azure Synapse Analytics. The guide incorporates all critical parameters discussed, including data product, ingestion pattern, source type (File or DB), frequency, volume, SLA, sensitivity, and batch IDs, to ensure consistency and optimization in the Synapse environment.

## 2. Data Volume Classification

Data is classified based on volume, processing requirements, arrival frequency, and other factors to optimize resource allocation and pipeline efficiency.

### Benchmarks for Volume Classification

|  |  |  |  |
| --- | --- | --- | --- |
| Volume Classification | Data Size (GB/Day) | Row Count (Approx) | Processing Frequency |
| Low | Up to 1 GB | Up to 100,000 rows | Daily or weekly batches |
| Medium | 1 GB - 10 GB | 100,000 - 1 million rows | Hourly or multiple times per day |
| High | More than 10 GB | Over 1 million rows | Near real-time or continuous |

## 3. Combined Ingestion Patterns and Source Types

The following table outlines the combined ingestion patterns and source types (File or DB) to capture essential metadata for data ingestion workflows.

### Ingestion Patterns with File or DB Sources

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ingestion Pattern | Source Type | Description | Use Cases | Example Name |
| Batch | DB/File | Data is ingested in chunks at scheduled intervals (e.g., daily, weekly). | Traditional ETL pipelines, periodic data synchronization. | Config\_Analytics\_Batch\_File\_CRM\_Daily\_Low.json |
| RealTime | DB/File | Data is ingested continuously with minimal delay. | IoT sensor streams, real-time transactions. | Config\_Analytics\_RealTime\_File\_IoT\_High.json |
| MicroBatch | DB/File | Hybrid of batch and real-time, ingesting small batches at frequent intervals. | Frequent database updates or small file drops requiring quick processing. | Config\_Analytics\_MicroBatch\_DB\_Ecommerce\_Med.json |
| Delta | DB/File | Only new or changed data is ingested, typically using change tracking mechanisms. | Incremental updates from databases or appending new file versions. | Config\_Analytics\_Delta\_File\_CustomerData\_Daily\_Med.json |
| EventDriven | DB/File | Data ingestion is triggered by specific events (e.g., file uploads, database triggers). | Log updates, alert-driven pipelines. | Config\_Analytics\_EventDriven\_File\_LogData\_Low.json |
| FullLoad | DB/File | The entire dataset is ingested, usually for initial loads or periodic refreshes. | Database migrations, rebuilding data warehouses. | Config\_Compliance\_FullLoad\_DB\_AuditData\_Weekly\_High.json |

## 4. Naming Convention for Configuration Files

The following naming convention is designed to capture essential metadata, enabling organized tracking and optimized configuration management.

### Naming Format

Config\_<Data\_Product>\_<IngestionPattern>\_<SourceType>\_<Source>\_<Frequency>\_<Volume>\_<SLA>\_<Sensitivity>\_<BatchID>.json

### Naming Components Explained

1. Config - Fixed prefix indicating a configuration file.  
2. <Data\_Product> - Target system or function, e.g., Analytics, Reporting, Operations.  
3. <IngestionPattern> - Describes how data is ingested:  
 - Batch, RealTime, MicroBatch, Delta, EventDriven, FullLoad.  
4. <SourceType> - Indicates the source type:  
 - DB (Database) or File (File-based sources).  
5. <Source> - Source system or dataset, such as CRM, ERP, IoT.  
6. <Frequency> - Data arrival frequency:  
 - RealTime, Daily, Weekly, Monthly.  
7. <Volume> - Classification based on data size:  
 - Low, Med, High.  
8. <SLA> - Service Level Agreement priority:  
 - Critical, NonCritical.  
9. <Sensitivity> - Data sensitivity level:  
 - Public, Internal, Confidential.  
10. <BatchID> - Sequential ID to manage multiple batches, especially when one source exceeds 20 files (e.g., Batch01, Batch02).

### Example Config Names

1. Batch ingestion for daily CRM data from a file:  
 Config\_Analytics\_Batch\_File\_CRM\_Daily\_Low.json

2. Real-time ingestion of IoT data from a database:  
 Config\_Analytics\_RealTime\_DB\_IoT\_High.json

3. Delta ingestion of daily customer updates from a file:  
 Config\_Analytics\_Delta\_File\_CustomerData\_Daily\_Med.json

4. Event-driven ingestion of log files:  
 Config\_Analytics\_EventDriven\_File\_LogData\_Low.json

5. Full-load ingestion of weekly compliance audit data from a database:  
 Config\_Compliance\_FullLoad\_DB\_AuditData\_Weekly\_High.json

## 5. Best Practices

1. Consistency: Apply this naming convention to enable clear identification and tracking across configurations.  
2. Resource Allocation: Configure batch counts and parallel processing based on data volume to maximize performance.  
3. Monitoring: Use Synapse monitoring tools to review resource utilization and make adjustments as volume grows.  
4. Scalability: Utilize BatchID for splitting sources exceeding 20 files to maintain scalability without disruption.

## 6. Conclusion

Following this classification and naming convention ensures efficient, scalable, and organized data ingestion in Synapse. Adhering to these guidelines optimizes resource usage, improves traceability, and simplifies troubleshooting.