

Key2 Consulting

Data Ingestion Framework (DIF) Overview

**September 21, 2023**

**Table of Contents**

[Document Background 3](#_Toc146266824)

[Metadata and Configuration Information 3](#_Toc146266825)

[ADF Pipelines 4](#_Toc146266826)

# Document Background

The data ingestion process brings data from disparate systems into the data lake. The data is landed into the Bronze Zone and then transferred and transformed into the Silver Zone.

The Data Ingestion Framework (DIF) is used to consume data and is targeted at ingesting data using both full and incremental design patterns from SQL Server and delimited file data sources. This framework can be extended to include other data sources such as Excel and JSON. The goal of the framework is to create a template pipeline and inform those pipelines via metadata.

The DIF includes two specific components:

* Metadata and configuration information (as well as logging data)
* Template Azure Data Factory (ADF) pipelines to ingest data into the Azure environment.

This document provides an overview of each of the DIF components.

# Metadata and Configuration Information

The DIF database is a centralized Azure SQL repository. This database is key to the ingestion framework in that it contains all the metadata and configuration information used to drive the ingestion process. In terms of metadata information, the ingestion process relies on metadata tables that track information on four important data points:

* Systems (servers, Azure storage accounts, etc.) that serve as either a source of data or a destination for data; this table is not customer-centric, so common systems should not be entered more than once.
* Repositories that contain one or more sets of data; from a source perspective, this may be a typical database, file share, or some other type of repository of data while from a destination perspective, this will typically be a data lake folder within a given storage account; repositories reside on specific systems.
* Datasets that reside within a given repository; for a source repository, this would likely be a database table or view, a CSV or JSON file, etc. while for a destination repository (a data lake folder), this would be the subfolder name that a given data set should be written to
* Attributes of a given dataset, which would contain column information (for tables or views) or field information (for CSV or JSON files, for example) and would include details such as data type, max length, nullability, primary key designation, etc.

Along with the metadata information related to systems, repositories, datasets, and attributes, the DIF database also includes several configuration tables that are integral to the Ingestion framework. These tables support the ADF pipelines used to ingest the data by managing groups of data ingestion tasks that should be processed. The ingestion framework relies on information related to these configuration items:

* Projects, which represent a group that uses the framework (or projects within a group)
* Pipelines that define each of the processes (in this case ADF pipelines) that move data.
* Pipeline groups, which relate pipelines together in cases where a parent/child pipeline scenario exists.
* Data integration tasks define specific ingestion activities, which relate a source dataset to a destination dataset; these tasks can include details such as load types (full, incremental, etc.), source filter logic, high watermark values, query logic, etc.
* Data integration groups, which allow tasks to be grouped so that a given group of tasks can be handled together.
* Environment configuration values in cases where a given pipeline needs external configuration values such as Azure Key Vault secret names, connection information, etc.

Beyond the metadata and configuration information, the DIF database also contains a set of logging and monitoring tables for capturing and storing information related to the execution of ingestion pipelines. The logging tables allow data ingestion engineers to log results or logical details when their pipelines execute (such as start times, end times, row counts, and other results). The monitoring tables are populated automatically with data related to the execution of the pipelines within ADF. Both sets of tables can be used to review execution statuses and results regularly.

# ADF Pipelines

The ingestion framework includes a set of template ADF pipelines that align the type of data being consumed. The pipelines include the following:

* A parent pipeline (named PL\_DIF\_[DescriptiveName]\_GroupOrder) that accepts project and data integration group parameters based on the configuration tables mentioned above. This pipeline manages the ordering and precedence of the various data integration or DITasks.
* A child pipeline named PL\_DIF\_[DescriptiveName]\_TaskLoop. This pipeline determines which DITasks are associated with the DIGroupTask. The loop in this pipeline will spawn multiple threads based on the number of DITasks in this group.
* A grandchild pipeline (named PL\_DIF\_[DescriptiveName]\_Ingestion) that accepts a specific data ingestion task and looks up the information needed to copy the data from the source dataset (i.e., SQL Server table or view) to the data lake. Based on the configurations, this will handle full, full then incremental, and incremental load scenarios. Additionally, if a full load is setup to take advantage of partitioning on the source, it will multi-thread the extract for that dataset based on the partitioning strategy set up in the configuration tables.

As mentioned above, all these pipelines rely on the metadata and configuration information that is entered into the DIF database. In addition, each pipeline includes only the most basic logging, to indicate when each pipeline that is executed starts and completes.

Additional information regarding the DIF database (the objects that relate to the ingestion framework) and the template ADF pipelines can be found in these additional documents:

* Key2 Data Ingestion Framework - Configuration Database Guide
* Key2 Data Ingestion Framework – ADF Pipeline Overview