redapt

Synapse Accelerator

Metadata-driven ELT Framework

Framework Overview

Batch Processing of on-prem or external data sources into the cloud

Currently supports RDBMS or flat file sources

Metadata stored in a config database in Azure

Dynamic ELT pipelines in Synapse

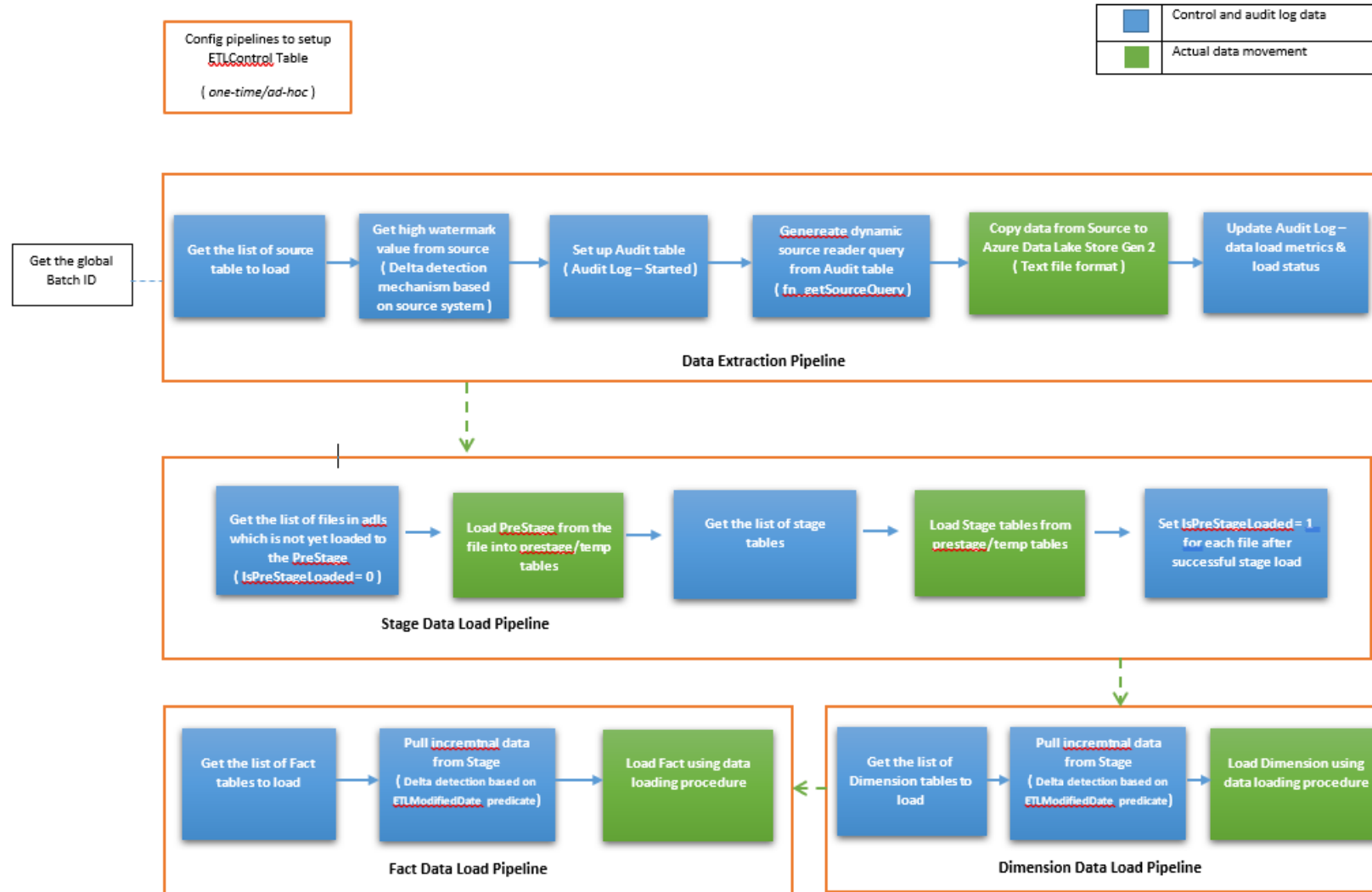Audit logging of pipeline executions

High Level Process Flow

Config pipelines to setup
ETLControl Table

(one-time/ad-hoc)

■ (blue)	Control and audit log data
■ (green)	Actual data movement

Get the global
Batch ID

Data Extraction Pipeline

Get the list of source table to load → Get high watermark value from source (Delta detection mechanism based on source system) → Set up Audit table (Audit Log – Started) → Genereate dynamic source reader query from Audit table (fn_getSourceQuery) → Copy data from Source to Azure Data Lake Store Gen 2 (Text file format) → Update Audit Log – data load metrics & load status

Stage Data Load Pipeline

Get the list of files in adls which is not yet loaded to the PreStage (IsPreStageLoaded = 0) → Load PreStage from the file into prestage/temp tables → Get the list of stage tables → Load Stage tables from prestage/temp tables → Set IsPreStageLoaded = 1 for each file after successful stage load

Fact Data Load Pipeline

Get the list of Fact tables to load → Pull incremtnal data from Stage (Delta detection based on ETLModifiedDate predicate) → Load Fact using data loading procedure

Dimension Data Load Pipeline

Get the list of Dimension tables to load → Pull incremtnal data from Stage (Delta detection based on ETLModifiedDate predicate) → Load Dimension using data loading procedure

Figure: ELT Framework Architecture
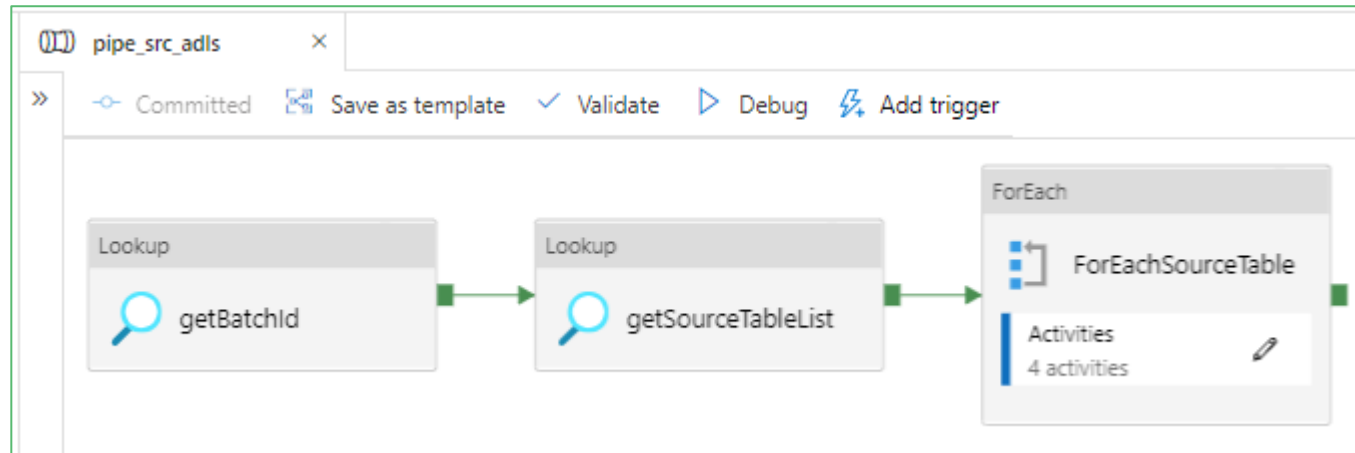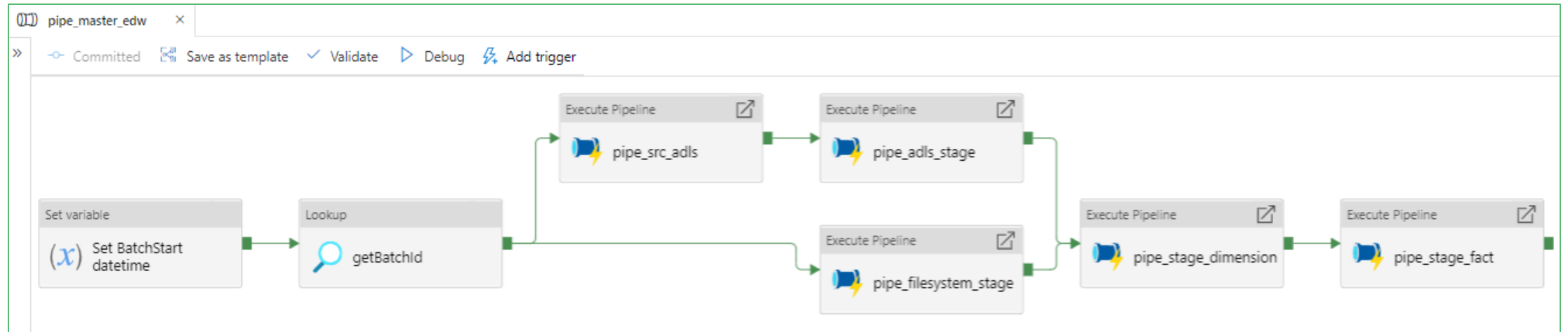
redapl

Config Database

Metadata in CTL schema

This database contains the control objects that dictate the sequence and flow of data from the source, through the bronze, silver, and gold data layers

Synapse Pipelines

Metadata Result Set From Lookup

Lookup activity pulls record set from CTL tables, defining source to sink attributes

For-Each loop executes Copy Data activity for each record, running in parallel

ETLControlID	SourceSystem	SourceType	SinkType	Source	Sink	Transform	ColumnMetadata	DeltaDetect
13	AW	SqlSource	Parquet	AdventureWorksLT.SalesLT.ProductModel	dls\raw\AdventureWorksLT\SalesLT\ProductModel\{YYYY}\{MM}\{DD}\ProductModel_*.pq	SqlSourceParquet	*	Predicate
20	AW	SqlSource	Parquet	AdventureWorksLT.SalesLT.Customer	dls\raw\AdventureWorksLT\SalesLT\Customer\{YYYY}\{MM}\{DD}\Customer_*.pq	SqlSourceParquet	*	Predicate
28	AW	SqlSource	Parquet	AdventureWorksLT.SalesLT.Address	dls\raw\AdventureWorksLT\SalesLT\Address\{YYYY}\{MM}\{DD}\Address_*.pq	SqlSourceParquet	*	Predicate
32	AW	SqlSource	Parquet	AdventureWorksLT.SalesLT.SalesOrderHeader	dls\raw\AdventureWorksLT\SalesLT\SalesOrderHeader\{YYYY}\{MM}\{DD}\SalesOrderHeade...	SqlSourceParquet	*	Predicate
50	AW	SqlSource	Parquet	AdventureWorksLT.SalesLT.Product	dls\raw\AdventureWorksLT\SalesLT\Product\{YYYY}\{MM}\{DD}\Product_*.pq	SqlSourceParquet	*	Predicate
57	AW	SqlSource	Parquet	AdventureWorksLT.SalesLT.ProductModelProductDes...	dls\raw\AdventureWorksLT\SalesLT\ProductModelProductDescription\{YYYY}\{MM}\{DD}\Pro...	SqlSourceParquet	*	Predicate

Audit Logging

Each data movement is considered a "load"

"Loads" are tied together into a "Batch", which includes all data movements for the end-to-end processing of the master pipeline

Audit logging tracks the statuses and volume of each load, and allows for some limited automated recovery in the event of a failure

ETLBatchId	ETLMaxValue	SourceQuery	RowsInserted	RowsUpdated	RowsDeleted	ETLStartTime	RunTimeInSecond	LoadType	LoadStatus	ErrorMsg	IsLoadSuccess
1	2021-05-05 16:57:19.043	SELECT * FROM AdventureWorksLT.SalesLT.Customer ...	847	0	0	2021-05-05 16:57:50.160	13	Incremental	Completed	NULL	1
1	2021-05-05 16:59:53.400	NULL	847	0	0	2021-05-05 16:59:53.600	11	Incremental	Completed	NULL	1
1	2021-05-05 16:58:57.150	NULL	456	0	0	2021-05-05 16:58:57.200	12	Incremental	Completed	NULL	1
1	2021-05-05 16:57:19.043	SELECT * FROM AdventureWorksLT.SalesLT.Address s ...	456	0	0	2021-05-05 16:58:00.620	14	Incremental	Completed	NULL	1
1	2021-05-05 16:58:57.150	NULL	456	0	0	2021-05-05 17:01:18.860	14	Incremental	Completed	NULL	1
1	2021-05-05 16:59:53.400	NULL	847	0	0	2021-05-05 17:01:23.100	22	Incremental	Completed	NULL	1

redapt