LivePerson API Development Document

12/20/2022

Created by the Brian Bolden, CAP Development

Revision History

Version	Date	Description	Author
1.0	12/20/2022	Initial Draft	Brian Bolden

Contents

Document Details

Stakeholders

Stakeholder	Job Tile	Organization
	Consultant	IT Data Management
	Software Developer	

Review and Approval

Role	Name	Organization	E-Signature / Date
Sr Manager		Data Engineering	

Document Control

This document is controlled by the Data Engineering team and is located here:

https://<location>

Any copy of this document stored or retrieved from a site other than the official site is considered uncontrolled. This procedure should be reviewed and approved annually for necessary changes.

NOTE: The audience for this document is internal. However, it can be provided to external users if necessary.

References

This policy references the following information sources:

Reference/Standard/Compliance Name

1 Introduction

1.1 Overview

This document contains the complete pipeline process, configuration, and parameters for the LivePerson At Scale Analytics Platform (ASAP). LivePerson is a Software as a Service (SaaS) provider that develops customer engagement and conversational commerce applications and platforms. Their primary products are:

LiveEngage

A messaging channel integration that allows companies to create AI-powered chatbots that can answer messages directly from customers. Developers can coordinate how conversations are handled and use a mixture of human and bot responses.

LivePerson Insights

A tool that inputs structured and unstructured customer chat transcripts and outputs business insights. It can provide customer service agents with customer sentiments, conversational health and performance KPIs.

1.2 Objective

To give developers detailed information on the LivePerson pipeline process.

1.3 Scope

This document is designed to support the engineering team. It is assumed that readers already have a grasp on Azure Data Factory pipeline terminology.

2 Azure Resources

2.1 LivePerson Pipeline Workstream

This section lists the workstream used to maintain the LivePerson pipeline. The following table highlights the workstream, related technical activities, and the skills required to maintain the workstream.

Table 1. LivePerson Pipeline Activities and Skills

Workstream	Technical Activities	Skills Required
Data Flow/Data Engineering	Data Flow Pipelines	Azure Resources, including: • Azure Data Factory • Azure Data Lake Gen2 (ADLSG2) • Azure Synapse Analytics

2.2 LivePerson Pipeline Capabilities

LivePerson must have the pipeline capabilities listed in the following table:

Table 2. LivePerson Pipeline Capabilities

Capability	Details	Priority
Write instructions to existing LivePerson Messenger APIs for ingesting data via LivePerson Messenger.	Some data can be accessed at a single conversation level, but not exported. Historical data beyond three months can be accessed via API only. Person number, account type, and other data fields are only accessible via API.	High
Identify and delete member data.	This is a requirement for unauthenticated messaging.	High
Store and access messaging event data for extended periods of time.	Requires functionality to generate year-over-year data reports and year-end reports. Capable of storing conversations historically without a Member Interaction Tracking (MIT) connection, which is not supported by Business Online Banking (BOB). Current state is that LivePerson only stores 13 months of data within LiveEngage, which makes required reporting inaccessible—Contact Center (CC) Leadership must do manual calculations which are less accurate. BOB cannot use LivePerson as a messaging solution, due to the conversation storage issue. This may mean no messaging in BOB, or an entirely separate, disconnected platform.	High
Store and access all messaging event conversation data fields.	Need access to the complete collection of data on conversations and users, including person numbers, account types, channel, device, etc. Current state is that data exports cannot be tied to members or any data points outside of internal LivePerson conversation IDs.	High
Access LivePerson data from Tableau, Power BI, or other internal reporting tools.	Provide the ability to report out on member journeys, engagement level by user, account type and/or channel, and member effort (based on number of contacts per intent). This expanded data will be utilized for improved Product Backlog Item (PBI) dashboards, CC Tableau dashboards, and potential new reporting platforms (such as Spiral).	High

Capability	Details	Priority
Store LivePerson chat transcripts and associated intent and sentiment data.	Need the ability to store LivePerson chat transcripts and associated intent and sentiment data.	Medium
Filter and sort conversation data.	Refined filtering allows users to view conversations both at a granular and holistic level, improving research and reporting on channel performance and the member experience.	Low

3 Infrastructure

This section provides background on the development tools, current solution architecture, solution data flow, etc. that the LivePerson team used for this solution.

3.1 Workflow Diagram

This section provides background on the architecture components used to develop the pipeline model. A high-level workflow diagram is listed below.

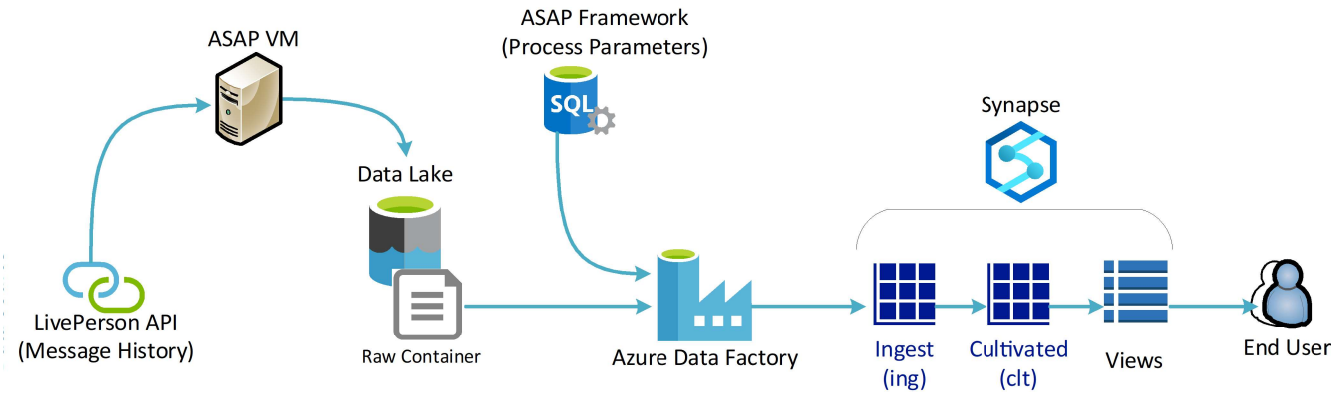

Figure 1. LivePerson Workflow

ASAP VM Locations

<drive location here>

Data Lake Location

The **RAW Layer Directory Structure** looks like this (HH – hour is optional):

{*layer*}/{*subject area*}/{*YYYY*}/{*MM*}/{*DD*}/{*HH*}

The **RAW Layer File Names** look like this:

{*schema/subject area descriptor*}_{*data_object_name*}_{*yyyyMMddHHmmssff*}_{*file_num*}.{*file_type*}

The **LivePerson Data Lake location** is:

raw/liveperson/messaginghistory/ingest

Azure Services

Azure Services used by the LivePerson team to build the solution are listed in the following table.

Table 3. LivePerson Azure Services

Service Name	Resource Type	Purpose
ASAP VM	Server	Serves as an intermediary between the API and ADLS. Takes information from a batch job that runs Python code to execute an API call. The API then returns the dataset, which gets stored in the raw container on ADLS.
asapadf{environment}{svc_num}	Data Factory (V2)	Pulls the data from external sources into any storage containers. Automation/Orchestration mechanism for daily triggering of pipelines.

Service Name	Resource Type	Purpose
asapadls{environment}{svc_num}	ADLS	Azure data lake storage account used for data storage, security, and Industry stand reliability.
asapsyn{environment}	Synapse SQL Pool	SQL Pool is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP) to quickly run complex queries across petabytes of data.

Azure Data Factory(V2)

Azure Data Factory (ADF) is a cloud-based data integration service that allows users to:

- Create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.
- Create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.
- Process and transform data by using compute services such as Azure Data Lake Analytics and Azure Machine Learning.

The figure below lists ADF capabilities at a high level (Ingest, Orchestrate, Transform Data, Configure SSIS):

Figure 2. ADF Service Capabilities

Azure Data Lake Storage Gen2

Azure Data Lake Storage Gen2 (ADLSG2) is a set of capabilities dedicated to big data analytics. It includes the following characteristics and features:

- Built on top of Azure blob storage.
- Converges the capabilities of ADLSG1.
- Makes Azure storage the foundation for building Enterprise Data Lake on Azure. A fundamental part of ADLSG2 is the hierarchical namespace to the blob storage.
- Uses Hierarchical Namespaces to organizes objects and files into a hierarchy of directories for efficient data access.

Synapse SQL Pool

Synapse SQL Pool (formerly SQL Data Warehouse) is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP) to quickly run complex queries across petabytes of data. As such, it is a key component for end-to-end big data solutions in the Cloud.

The figure below lists Synapse SQL Pool capabilities at a high level (Ingest, Explore and Analyze, Visualize):

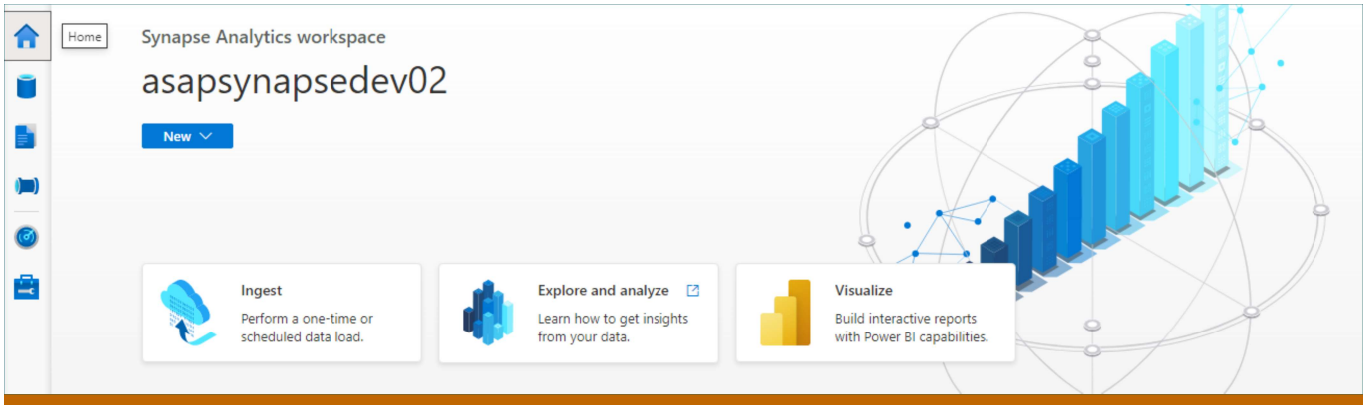

Figure 3. Synapse SQL Pool Capabilities

3.2 Solution Data Flow & Orchestration

Azure Data Factory

ADF's function is to fetch data from available sources. LivePerson uses two ADF pipelines:

- LP_MessagingHistory
- LP_SkillSegment

The main role of these pipelines is to incrementally load data from ADLS into Synapse. The pipelines are scheduled to run based on a trigger. This trigger can be changed as per business needs. Detailed information about each activity in the LivePerson pipelines is described in 3.3 LivePerson Pipeline Orchestration Details.

Azure Data Lake Storage

The Azure VM retrieves data from the LivePerson API and stores in ADLS raw layer.

raw/liveperson/messaginghistory/ingest: A directory in the ADLS raw container that contains files pulled from the LivePerson API. Data is stored in the hierarchical format in ADLS as shown in the below screenshot.

Azure Synapse Analytics

The Synapse SQL pool (asapsynapsedev02 for the dev environment) contains the views for user consumption of data from the cultivated layer. It manages the processing and user consumption of data. A schema exists that contains all the tables details in the Synapse SQL pool.

Table 4. Synapse SQL Pool Table Details

Server Name	SQL Database	Schema.Key Table
asapsynapsedev02 (dev)	asapsyndev (dev)	clt_lp.mh_agentparticipants.sql
asapsynapsetest (test)	asapsyntst (test)	clt_lp.mh_campaign.sql
asapsynapsepp (preprod)	asapsynpp (pre-prod)	clt_lp.mh_consumerparticipants.sql
		clt_lp.mh_conversationsurveys.sql
		clt_lp.mh_info.sql
		clt_lp.mh_interactions.sql
		clt_lp.mh_messagerecords.sql
		clt_lp.mh_messagescores.sql
		clt_lp.mh_messagestatuses.sql
		clt_lp.mh_sdes_events.sql
		clt_lp.mh_SkillSegments.sql
		clt_lp.mh_transfers.sql
		clt_lp.ss_SkillSegments.sql
		ing_lp.mh_agentparticipants.sql
		ing_lp.mh_campaign.sql
		ing_lp.mh_consumerparticipants.sql
		ing_lp.mh_conversationsurveys.sql
		ing_lp.mh_info.sql
		ing_lp.mh_interactions.sql
		ing_lp.mh_messagerecords.sql
		ing_lp.mh_messagescores.sql
		ing_lp.mh_messagestatuses.sql
		ing_lp.mh_sdes_events.sql
		ing_lp.mh_SkillSegments.sql
		ing_lp.mh_transfers.sql
		ing_lp.ss_SkillSegments.sql

Ingest(ing) Layer

Ingest is a non-persistent intermediary layer used to land untransformed data from raw prior to transformation to a consumable data set in the cultivated layer. Data in the ingest layer is deleted/truncated prior to being loaded from the raw layer files and only contains data for the current processing window.

Ingest Table Naming Standard

ing_{src acronym}.{src_schema}_{table_name}

Example: ing_lp.mh_campaign.sql

Cultivated(clt) Layer

Cultivated tables are persistent with data in a final transformed state intended for consumption by users. Only a subset of power users gets read-only access for the clt tables. All other users consume data via consumption views built on the clt tables. The clt tables are typically loaded via a stored procedure for each table or in some a case a ADF Dataflow.

Cultivated Table Naming Standard

clt_{src acronym}.{src_schema}_{table_name}

Example: clt_lp.mh_interactions.sql

Load/Transformation Stored Procedure

The load store procedures are triggered in the ADF extract to cultivated pipeline. The role of these store procedures is to load the data from ingestion layer to the cultivated layer tables. Stored procedures are listed in the following table.

Table 5. LivePerson Stored Procedures

Stored Procedure	Description
usp_mh_agentparticipants_load.sql usp_mh_campaign_load.sql usp_mh_consumerparticipants_load.sql usp_mh_conversationsurveys_load.sql usp_mh_info_load.sql usp_mh_interactions_load.sql usp_mh_messagerecords_load.sql usp_mh_messagescores_load.sql usp_mh_messagestatuses_load.sql usp_mh_sdes_events_load.sql usp_mh_skillsegments_load.sql usp_mh_transfers_load.sql usp_ss_skillsegments_load.sql	All stored procedures load data from the ingest schema table to the cultivated table. SPs are not reusable across multiple tables—they are not dynamic, but specific to the table that each one loads.

Load Stored Procedure Naming Standard

ing_{src system name/acronym}.usp_{src_schema}_{table_name}_load

Example: ing_dna.usp_osibank_ACCT_load

Stored Procedure Logic

Most pipelines follow the incremental data load patterns for the clt load stored procedures. The incremental stored procedure logic supports full table and delta source extracts for Type-I tables.

1. DROP and CREATE work table to temporarily hold the transformation data set:
    a. wrk_{src acronym}.{src_schema}_{table_name}
2. Load the work table from the ingest schema:
    a. Dedupe based on the natural/business key and the last updated date of the table.
    b. Set the DML_FLAG for each unique record by joining to the clt table and determining if the record exists. Values: I, U, D
3. DELETE any existing rows from the clt where the wrk table DML_FLAG = ‘U’ for the row.
4. INSERT all rows from the work table into the clt table.
5. DROP the work table.

Consumption Layer

The consumption layer consists of views built on the cultivated layer tables with permission managed by security group(s) created for each source system. To get access to the cultivated layer tables the users in place a ServiceNow Additional Access request for the security group(s). The views typically don’t have any transformation logic on the columns except for obfuscating PII.

Consumption Layer (View) Naming Standard

{src system name/acronym}.{src_schema}_{table_name}

Example: dna.osibank_ACCT

3.3 LivePerson Pipeline Orchestration Details

ADF plays the role of extracting data from the available sources and loading the data into the target ADLS or database targets. ADF also manages the orchestration of data flows. The following pipelines are standard patterns.

Below are the LivePerson pipelines:

- LP_MessagingHistory
- LP_SkillSegment

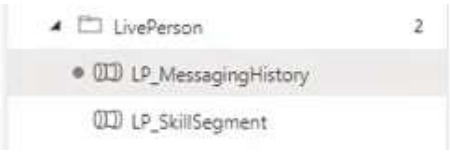

Figure 4. LivePerson Pipelines

Pipeline architecture is basically the same for each. Detailed information for all the Azure resources used to build the LivePerson pipeline are listed below.

Get Primary Ingest Folders

This component lists the files present in the ADLS ingest folder.

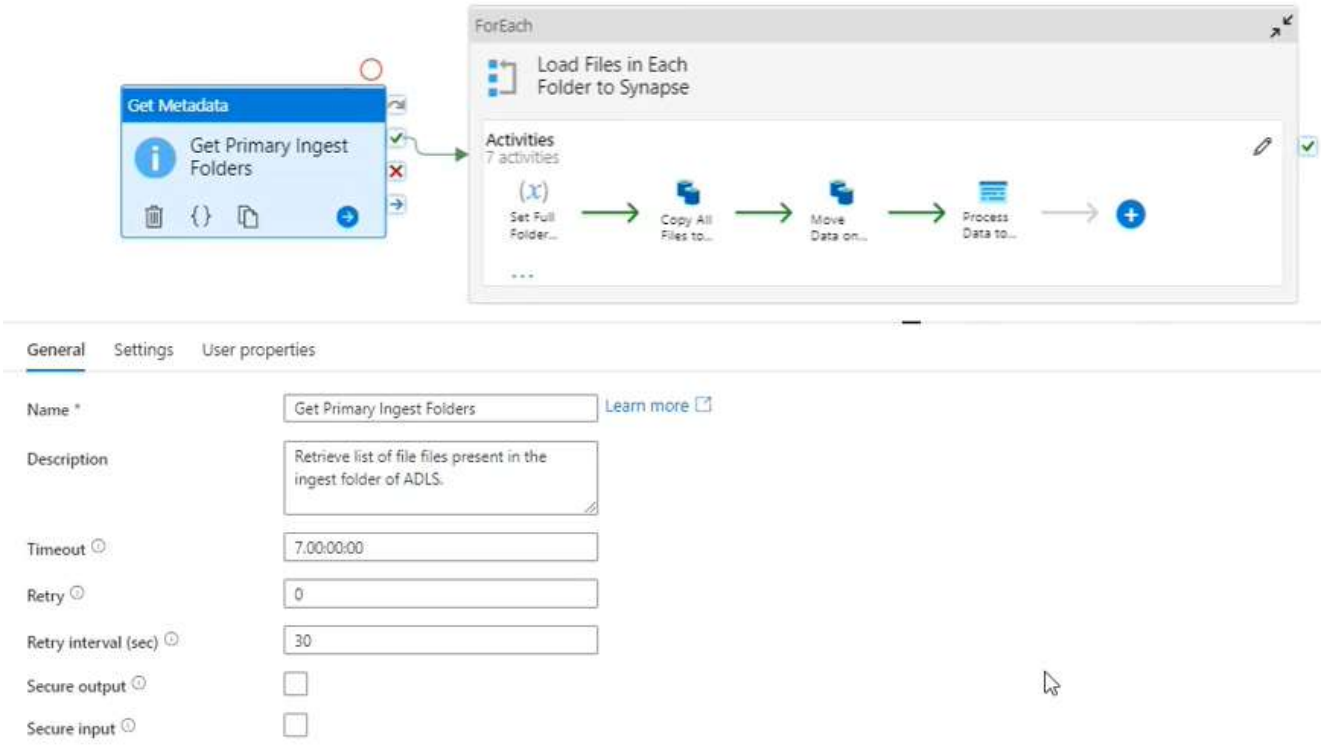

Figure 5. Get Primary Ingest Folders

Load Files in Each Folder to Synapse

This component ingests the file list used to load data into Synapse.

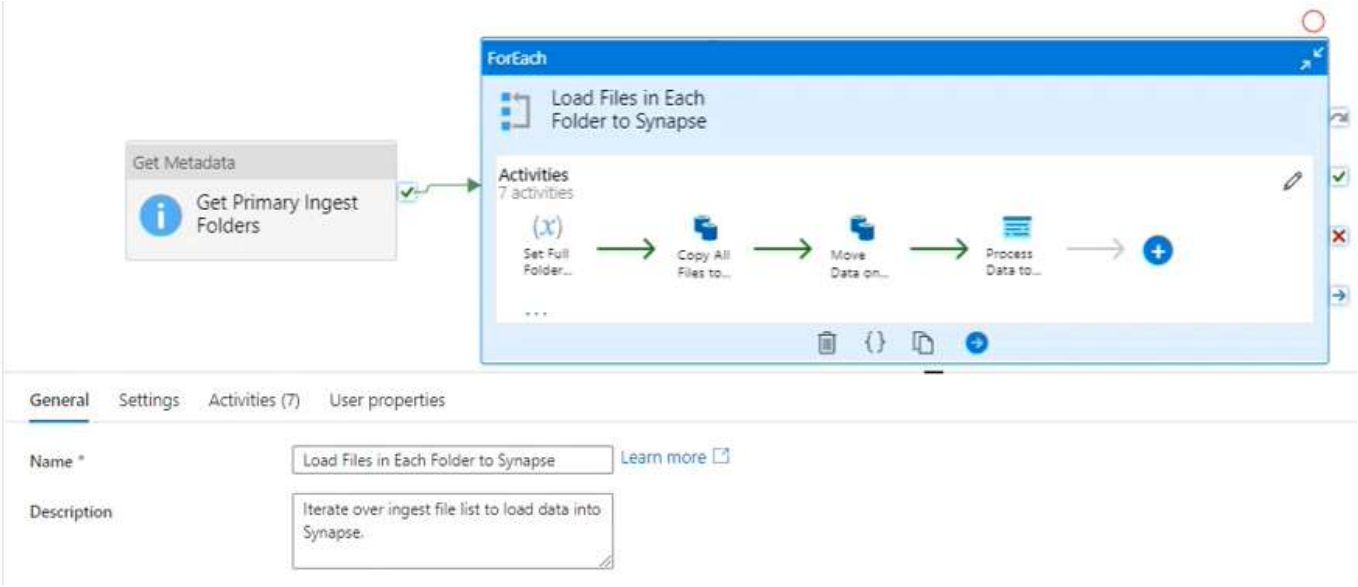

Figure 2. Load Files in Each Folder to Synapse

Set Full Folder Name

This component generates the file path for ingesting files.

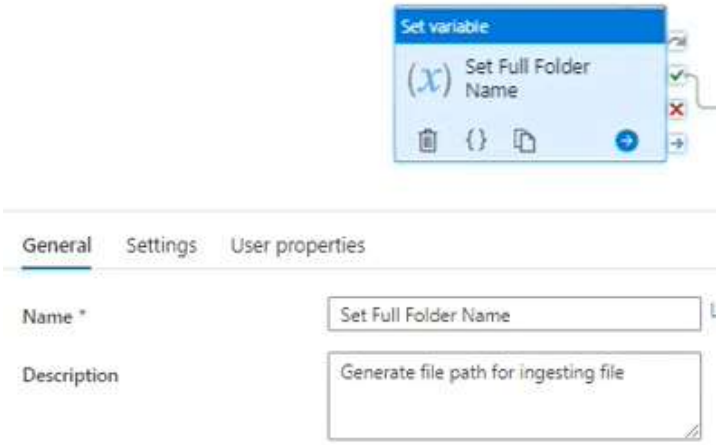

Figure 3. Set Full Folder Name

Copy All Files to Synapse Ingest

This component loads files from the ingest file to the ingest layer in Synapse.

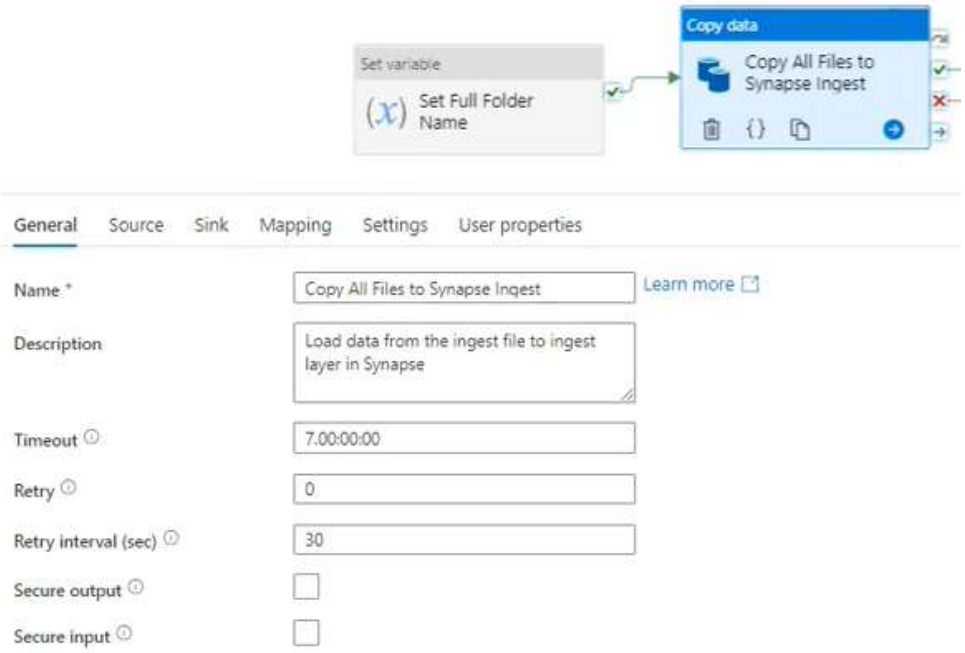

Figure 4. Copy All Files to Synapse Ingest

Move Data on Successful Load

This component moves the file from the ingest folder to the “complete” folder.

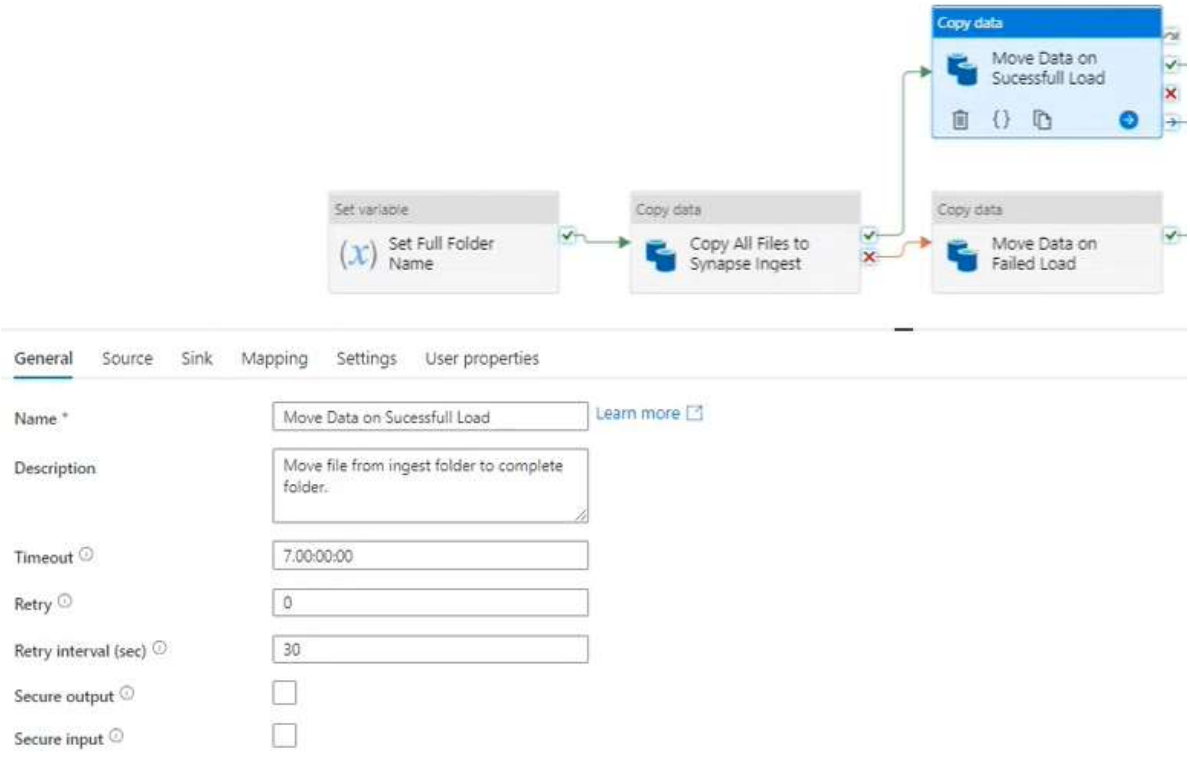

Figure 6. Move Data on Successful Load

Move Data on Failed Load

This component moves the file from the ingest folder to the “failed” folder.

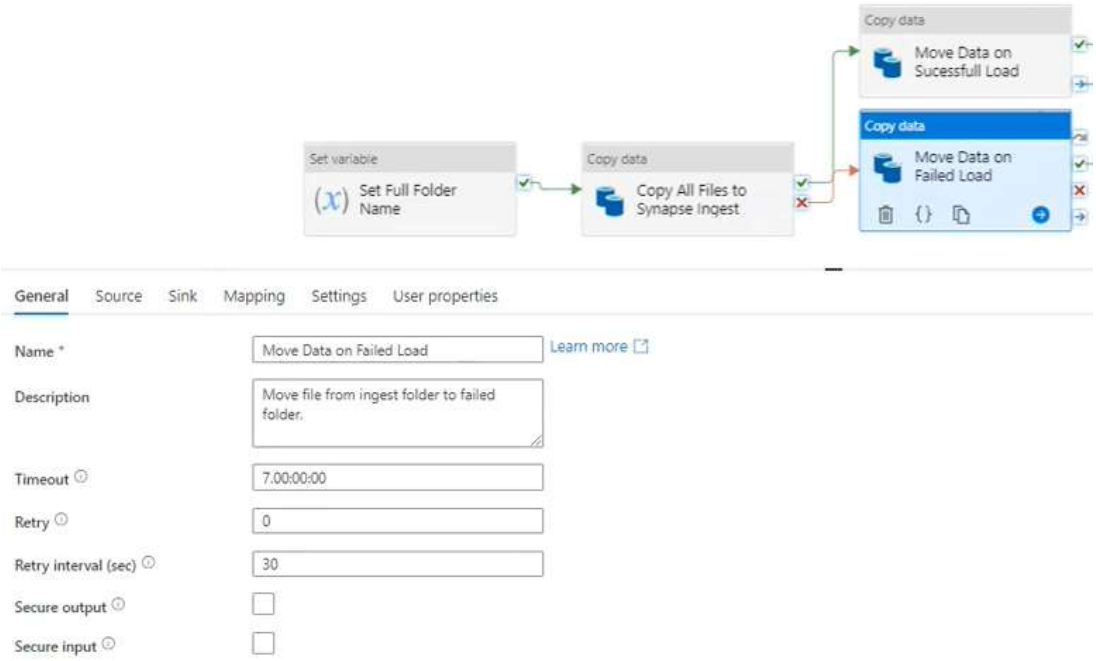

Figure 5. Move Data on Failed Load

Process Data to Cultivate in Synapse

This component runs the stored procedure to load data from the “ingest” layer to the “cultivated” layer:

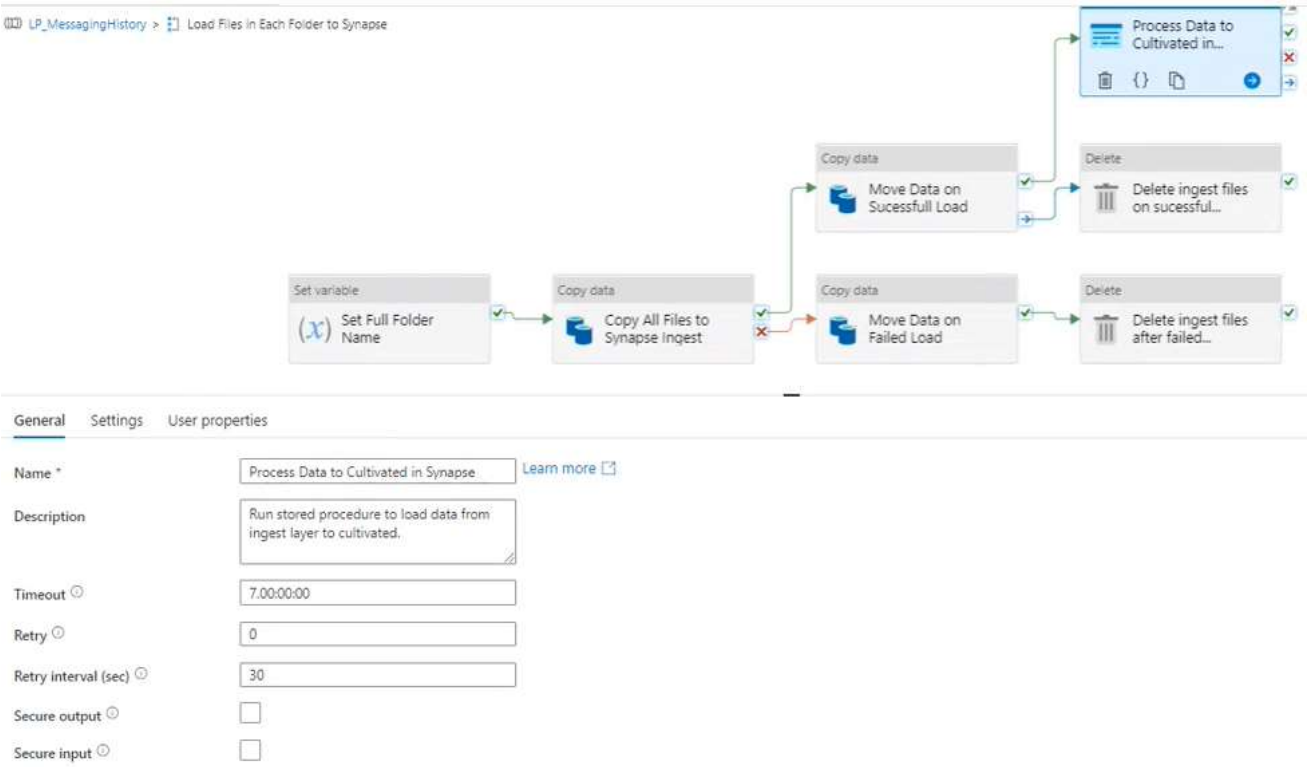

Figure 7. Process Data to Cultivate in Synapse

Delete Ingest Files on Successful Processing

This component deletes files from the ingest folder once processing has completed successfully:

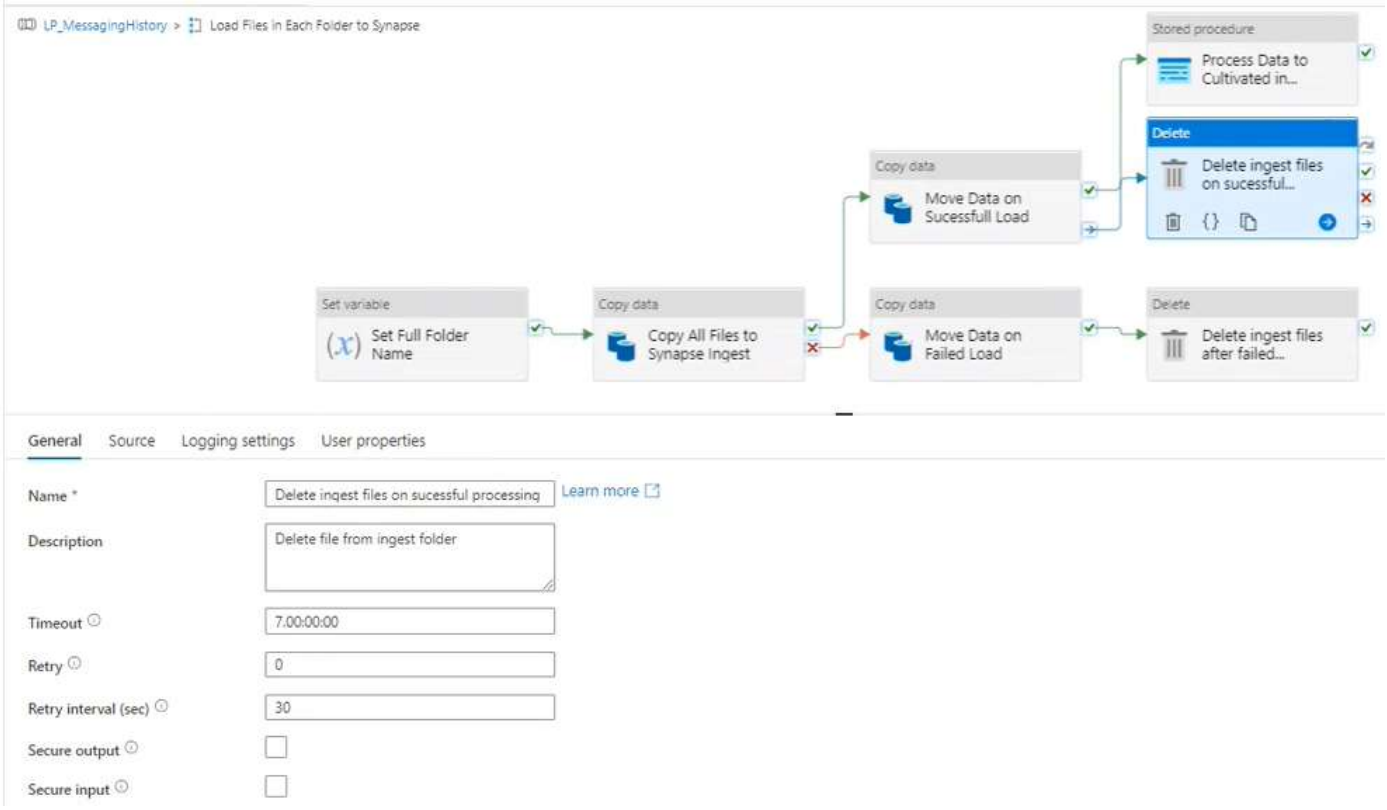

Figure 8. Delete Ingest Files on Successful Processing

Delete Ingest Files After Failed Processing

This component deletes files from the ingest folder when processing fails:

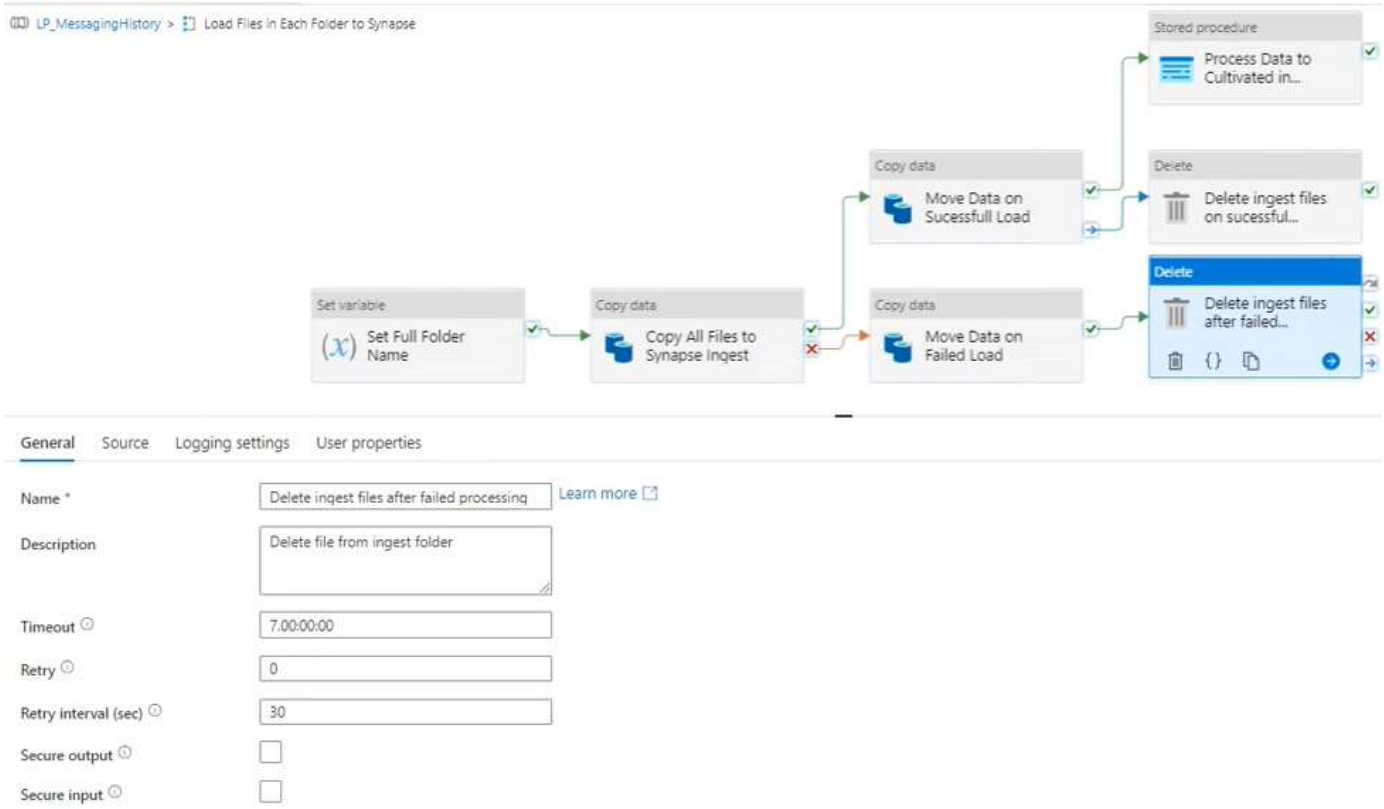

Figure 9. Delete Ingest Files After Failed Processing

Data Lake Pipeline Connection Details

ADF pipeline connection details to Data Lake are listed in the following table:

Table 6. Data Lake Connection Details

Linked Service	Dataset	Key Vaults
ASAP_ADL	ds_copy_raw_amperity	asap-dev-kv1 (dev) Test: asapkvtst (test) asapkvpp02 (preprod) asapkv (prod) Secret: kv-asapframe-connectionstring

Synapse Pipeline Connection Details

ADF pipeline connection details to Synapse are listed in the following table:

Table 7. Synapse Connection Details

Server Names	Linked Service	Dataset	Key Vaults	SQL Database
asapsynapsedev02.sql.azuresynapse.net (dev) asapsynapsetest.sql.azuresynapse.net (test) asapsynapsepp.sql.azuresynapse.net (pre-prod) asapsynapse.sql.azuresynapse.net (prod)	ls_asapsyn	ds_asapsyn	asap-dev-kv1 (dev) asapkvtst (test) asapkvpp02 (preprod) asapkv (prod) Secret: kv-asapsyn-password	asapsyndev (dev) asapsyntst (test) asapsynpp (preprod) asapsyn (prod)

- **NOTE**: The information about tables, stored procedures and views of synapse SQL pool is documented in section 3.2 Solution Data Flow & Orchestration.

4 Troubleshooting Issues

4.1 Synapse Pre-Prod Maximum Concurrent Session Limit Exceeded

Issue

When the user attempts to run pipelines in Pre-Prod with the number of parallels (concurrent sessions) as 45, there are times the pipelines fail with the following error message:

> Execution fail against sql server. Please contact SQL Server team if you need further support. Sql error number: 111219. Error Message: 111219;The maximum concurrent session limit of 512 exceeded. To learn more, please visit: http://aka.ms/dwsoftlimits.

Cause

This appears to be an issue with the Synapse configuration. The capacity limits for dedicated SQL pool in Azure Synapse Analytics varies based on the configuration.

The below screenshot shows the DWU for the dedicated SQL pool in PREPROD Synapse Analytics.

There are limits on the number of queries that can execute concurrently. This limitation varies based on the selected DWU:

- DWU1000c and above support a maximum of 1024 open sessions.
- DWU500c and below supports a maximum concurrent open session limit of 512. ***This is the current synapse configuration.***

When the concurrency limit is exceeded, the request goes into an internal queue where it waits to be processed and the pipeline fails and displays the above error message.

Resolution

1. Capacity limits for dedicated SQL pool in PREPROD Azure Synapse Analytics needs to be increased from DWU500c to DWU1000c.
2. DWU1000c and above support a maximum of 1024 open concurrent sessions.
3. If not possible, go for Work around options below.

Workarounds

Rerun the appropriate pipeline in PREPROD ADF.

OR

Schedule the appropriate pipelines to run at a different time of the day in PREPROD.

5 Appendices

Appendix A Naming Convention Details

The naming conventions used in the ADF pipeline adhere to the following standards:

Table 8. Pipeline Naming Conventions

Pipeline	Naming Convention
Pipeline Name	pl_<*source_name*>_main
Data Lake (ADLSG2)	raw/<*source_name*>/*year*/*month*/*day*/*filename_timestamp*.csv
Synapse Ingest	**Table Name** clt_<*source_name*>.*sourceschema_tablename* **Stored Procedure** Ing_<*source_name*>.usp_ing_*tablename*_load
Synapse (Consumption Layer)	**Table Name** clt_<*source_name*>.*sourceschema_tablename*

Appendix B Source Control and Azure DevOps

The following lists include Azure resources used to manage and deploy data migration workflows.

Azure Repos

- ADF: Tools.ASAP.ADF
- ASAPFrame: Tools.ASAP.ASAPFrame
- Synapse: Tools.ASAP.Synapse

Build and Release

Build and release is automated via YAML pipelines that are triggered with a pull request. The following documents detail the pipelines and the process:

- ADF Build and Release
- Database Build and Release

Appendix C Useful Queries

To be completed when available.

Appendix D Source to Target Mapping

To be completed when available.

Appendix E Business Terms and Acronyms

Table 9. Business Terms and Acronyms

Term (Acronym)	Definition
At Scale Analytics Platform (ASAP)	An in-house platform that includes Azure data-related tools, including Azure SQL, Synapse, Data Factory, ADSL 2.0, function apps, logic apps, and Integration Runtime.
Azure Data Factory (ADF)	Azure's cloud ETL service for scale-out serverless data integration and data transformation. It offers a code-free user interface for authoring, monitoring, and management. You can also lift and shift existing SSIS packages to Azure and run them with full compatibility in ADF. SSIS Integration Runtime offers a fully managed service, so you don't have to worry about infrastructure management. Refer to Microsoft's Azure Data Factory documentation for more information.
Azure Data Lake	A centralized repository designed to store, process, and secure large amounts of structured, semi structured, and unstructured data. It can store data in its native format and process any variety of it, ignoring size limits.
Cloud-Based Automation Framework (CBAF)	Streamlines tasks or processes to improve efficiency and reduce manual workload. Cloud-based automation defines the deployment and management of tasks to be automated, and cloud orchestration arranges and coordinates those defined tasks into a unified approach to accomplish intended goals. enables IT admins and Cloud admins to automate manual processes and speed up the delivery of infrastructure resources on a self-service basis, according to user or business demand.
Contact Center (CC)	
Data Pipeline	A data pipeline is a series of data processing steps. If the data is not currently loaded into the data platform, then it is ingested at the beginning of the pipeline. Then there are a series of steps in which each step delivers an output that is the input to the next step. This continues until the pipeline is complete. In some cases, independent steps may be run in parallel. Data pipelines consist of three key elements: a source, a processing step or steps, and a destination. In some data pipelines, the destination may be called a sink. Data pipelines enable the flow of data from an application to a data warehouse, from a data lake to an analytics database, or into a payment processing system, for example. Data pipelines also may have the same source and sink, such that the pipeline is purely about modifying the data set. Any time data is processed between point A and point B (or points B, C, and D), there is a data pipeline between those points.

Term (Acronym)	Definition
Enterprise Data Warehouse (EDW)	Stores an organization's data from sources across the entire business. An EDW enables data analytics, which can inform actionable insights. A smaller data warehouse may be specific to a business department or line of business (like a data mart). A central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence. A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. Active Data Warehousing is the technical ability to capture transactions when they change, and integrate them into the warehouse, along with maintaining batch or scheduled cycle refreshes. An active data warehouse offers the possibility of automating routine tasks and decisions. The active data warehouse exports decisions automatically to the On-Line Transaction Processing (OLTP) systems. Real-time Data Warehousing describes a system that reflects the state of the warehouse in real time. If a query is run against the real-time data warehouse to understand a particular facet about the business or entity described by the warehouse, the answer reflects the state of that entity at the time the query was run. Most data warehouses have data that are highly latent – or reflects the business at a point in the past. A real-time data warehouse has low latency data and provides current (or real-time) data.
Enterprise Strategic Services (ESS)	Provides many strategies for businesses trying to improve their service delivery and internal processes.
Extract, Transform, Load (ETL)	Brings data from multiple sources into a centralized database or unified data repository by: • Extracting data from its original source • Transforming data by deduplicating it, combining it, and ensuring quality • Loading data into the target database/repository
Fact tables	Fact tables vary in structure. Some are daily snapshots that contain one record per item (e.g., account) per day. These tables can grow quite large, but they provide a picture of the data elements on any given day. The daily snapshot tables normally have a corresponding month-end snapshot table. Other updateable fact tables only show the current view of a set of data like the address or party associated with a given account.
Informatica	Informatica is an American software development company founded in 1993. It is headquartered in Redwood City, California. Its core products include Enterprise Cloud Data Management and Data Integration. Informatica uses BI techniques & tools to perform ETL.
LiveEngage	A messaging channel integration that allows companies to create AI-powered chatbots that can answer messages directly from customers.
Member Interaction Tracking (MIT)	Information recorded on a member's profile to indicate how they were assisted by an associate. MIT information should be specific and factual.
Massively Parallel Processing (MPP)	The coordinated processing of a program by multiple processors that work on different parts of the program, with each processor using its own operating system and memory.

Term (Acronym)	Definition
Orchestrator	Orchestrator is an automation software tool that allows a user to automate the monitoring and deployment of data center resources. It ties disparate tasks and procedures together using a graphical user interface Runbook Designer to create reliable, flexible, and efficient end-to-end solutions in your IT environment.
Online Banking (OLB)	Online banking allows customers of a financial institution to conduct financial transactions on a secure website operated by the institution, which can be a retail or virtual bank, credit union or building society. For access, a customer having personal Internet access must register with the institution for the service and set up a password or a pin for customer verification.
Product Backlog Item (PBI)	A single element of work that exists in the product backlog. PBIs can include user stories, epics, specifications, bugs, or change requirements.
Synapse SQL Pool	Refers to the enterprise data warehousing features that are available in Azure Synapse Analytics.