BUILDING A DATA PIPELINE

1
2
3
4
5

BUILDING BLOCKS

What are the components of Azure Data Factory?

DATA SOURCES

Overview of the data sources and data ingestion approach

BUILDING THE DATA PIPELINE

Mechanics of building a data pipeline

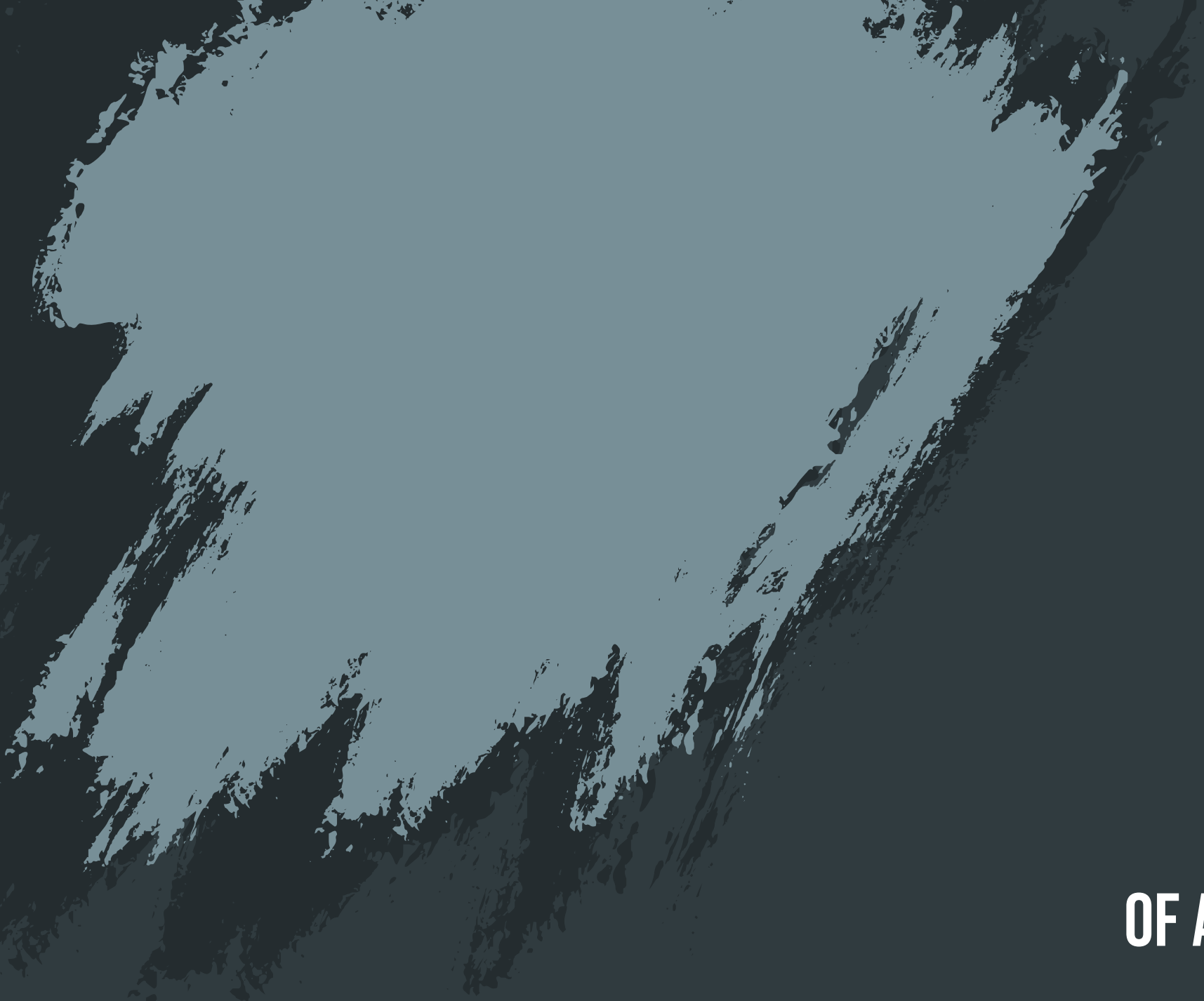
IMPORTING DATA

Importing structured or semi-structured data

NAMING CONVENTIONS

Best practices for naming conventions

MODULE 5



SECTION 1

BUILDING BLOCKS OF AZURE DATA FACTORY

BUILDING BLOCKS

What are the main components of Azure Data Factory?

TRIGGERS

- Used to kick-off a pipeline execution
- Different types of triggers

PIPELINES

- Units to perform an ETL workload
- Logical grouping of activities
- Pipelines can be chained for complex operations

INTEGRATION RUNTIME

- Compute Environment to run the activity

ACTIVITIES

- Processing step within a pipeline
- 3 types of Activities – Movement, Transformation and Control

DATASETS

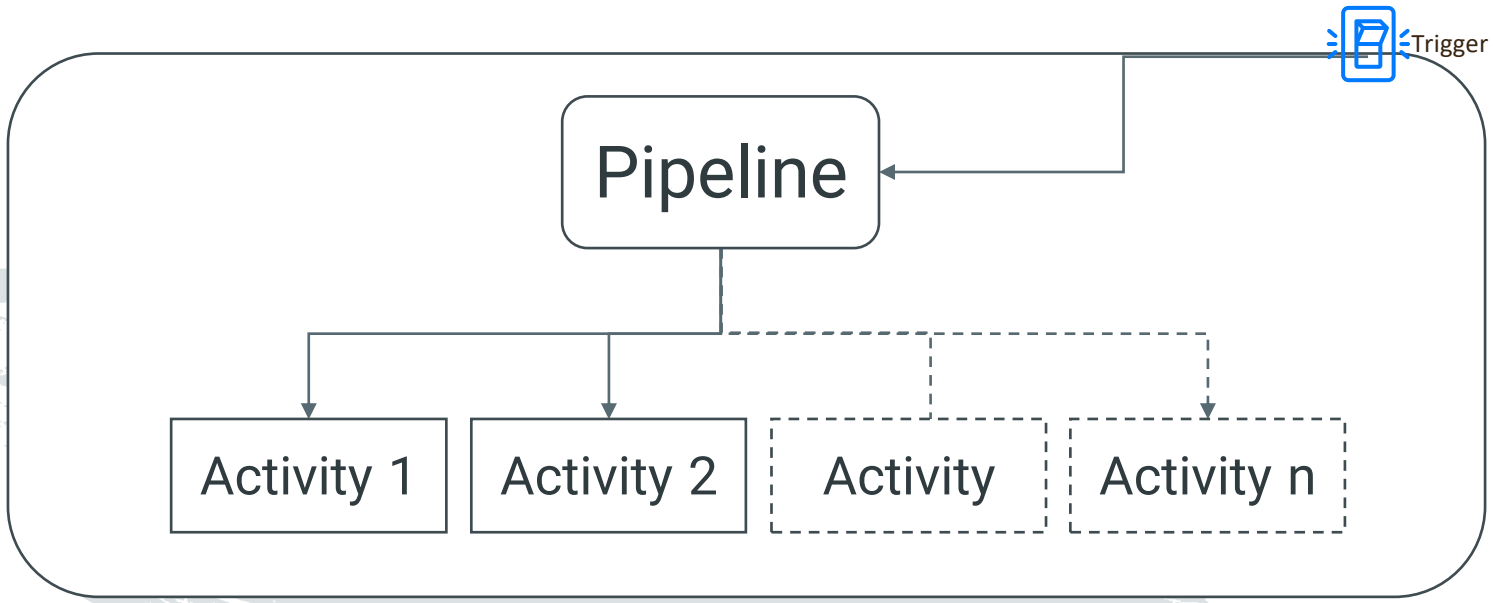
- Represent data structures within data stores

LINKED SERVICES

- Connection strings to data sources
- Can represent a data store or a compute resource

PIPELINES AND ACTIVITIES

Let's get a high-level view of Pipelines and Activities



Building blocks of an ADF Project

Pipelines and Activities

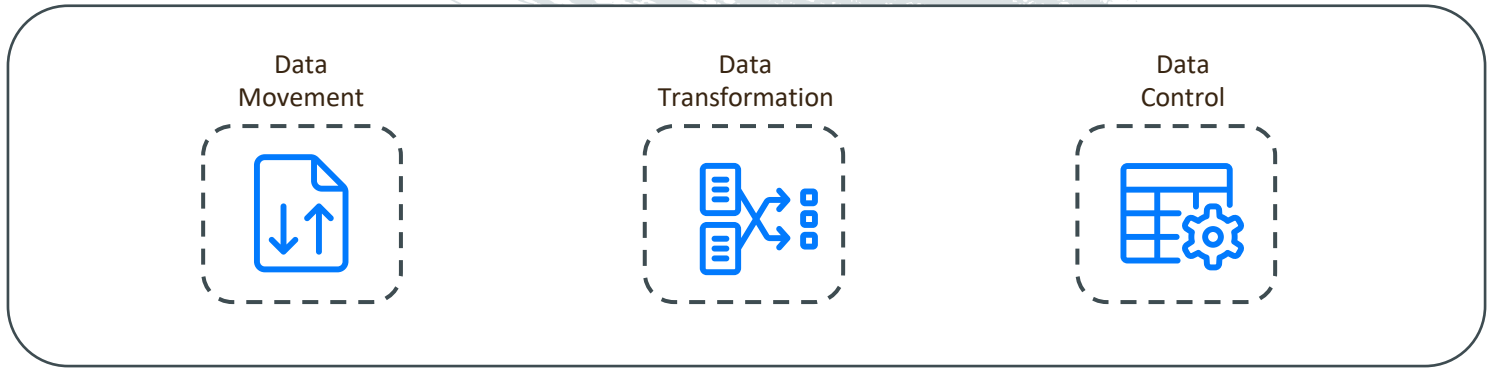
Pipelines

Logical grouping of Activities

Activities

Perform operations on the data

Activities



Data Movement Activities

Copy Activity to copy data from a source to a sink

Data Transformation Activities

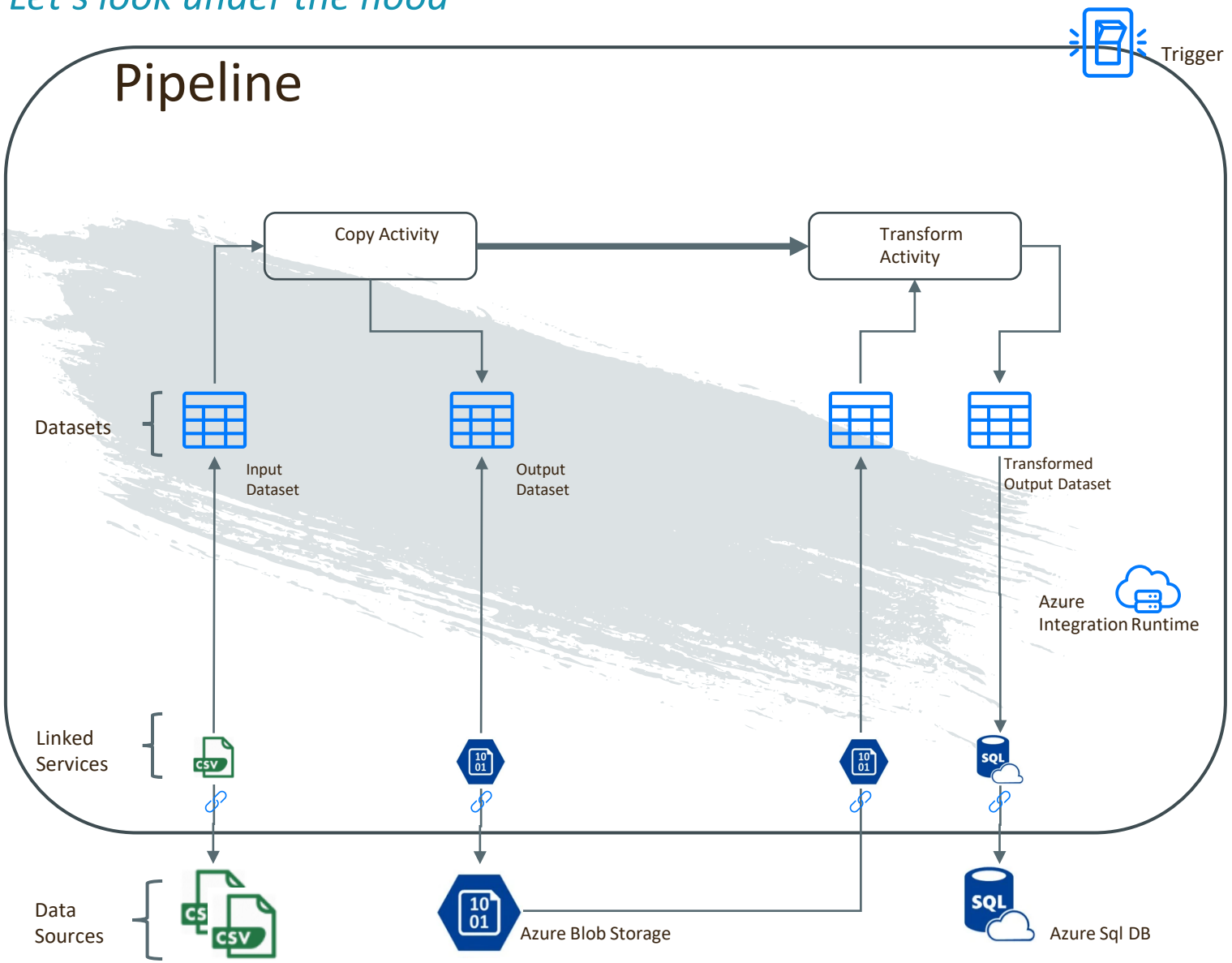
Change data – native Data Flows or External Service

Data Control Activities

Logical control activities such as for each, conditions etc.

SO HOW ARE THESE COMPONENTS RELATED IN AZURE DATA FACTORY?

Let's look under the hood



Pipelines

Group of Activities

Triggers

Switch to schedule & execute your pipeline

Activities

Operations on your dataset

Datasets

Representation of your data source

Integration Runtime

Compute Infrastructure for ADF

Linked Services

connection strings to your data sources

SECTION 2

DATA SOURCES

DATA SOURCES

Data received from the different stores

ARANCIONE

- Sales Data for Store Arancione
- CSV Files provided monthly
- Monthly sales data aggregated by product
- Sales Amounts in EUR

VERDE

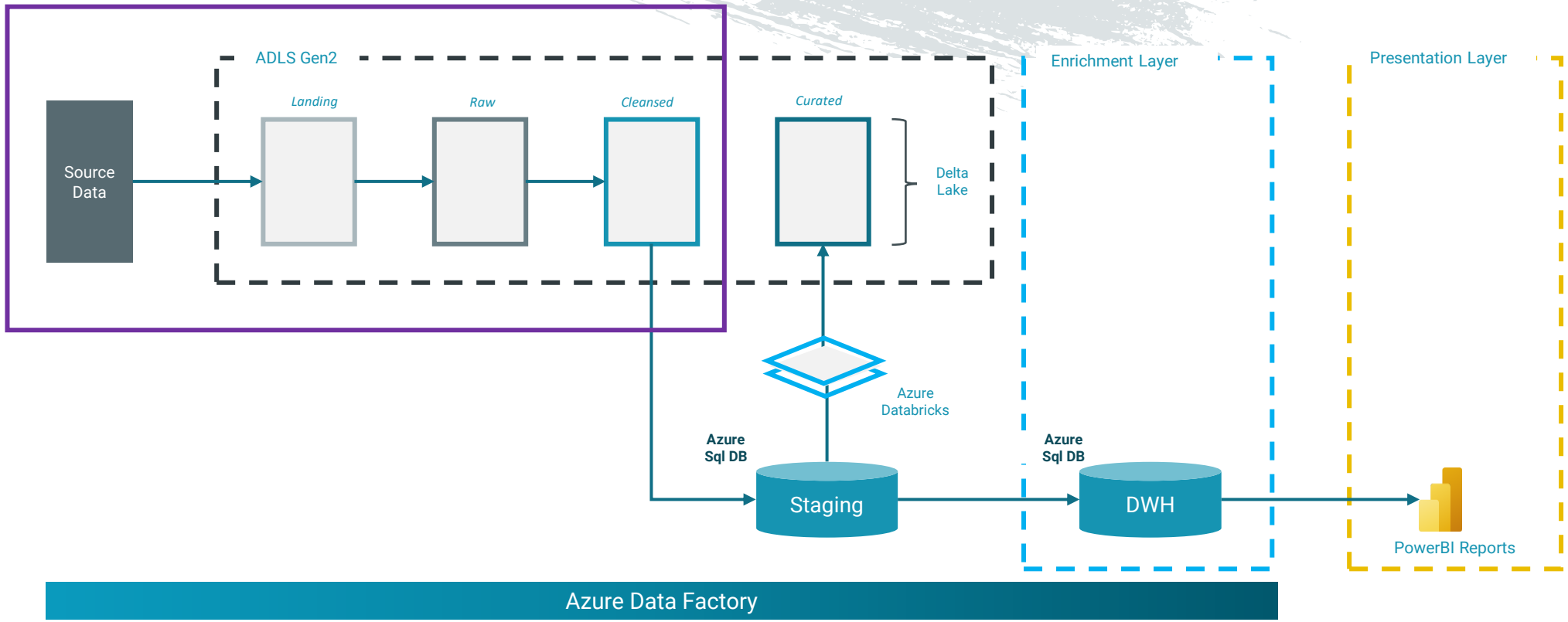
- Sales Data for Store Verde
- JSON Files provided monthly
- Monthly sales data aggregated by product
- Sales Amounts in EUR

CELESTE

- Sales Data for Store Celeste
- CSV Files provided monthly
- Daily sales data aggregated by product
- Sales Amounts in EUR and GBP

DATA INGESTION

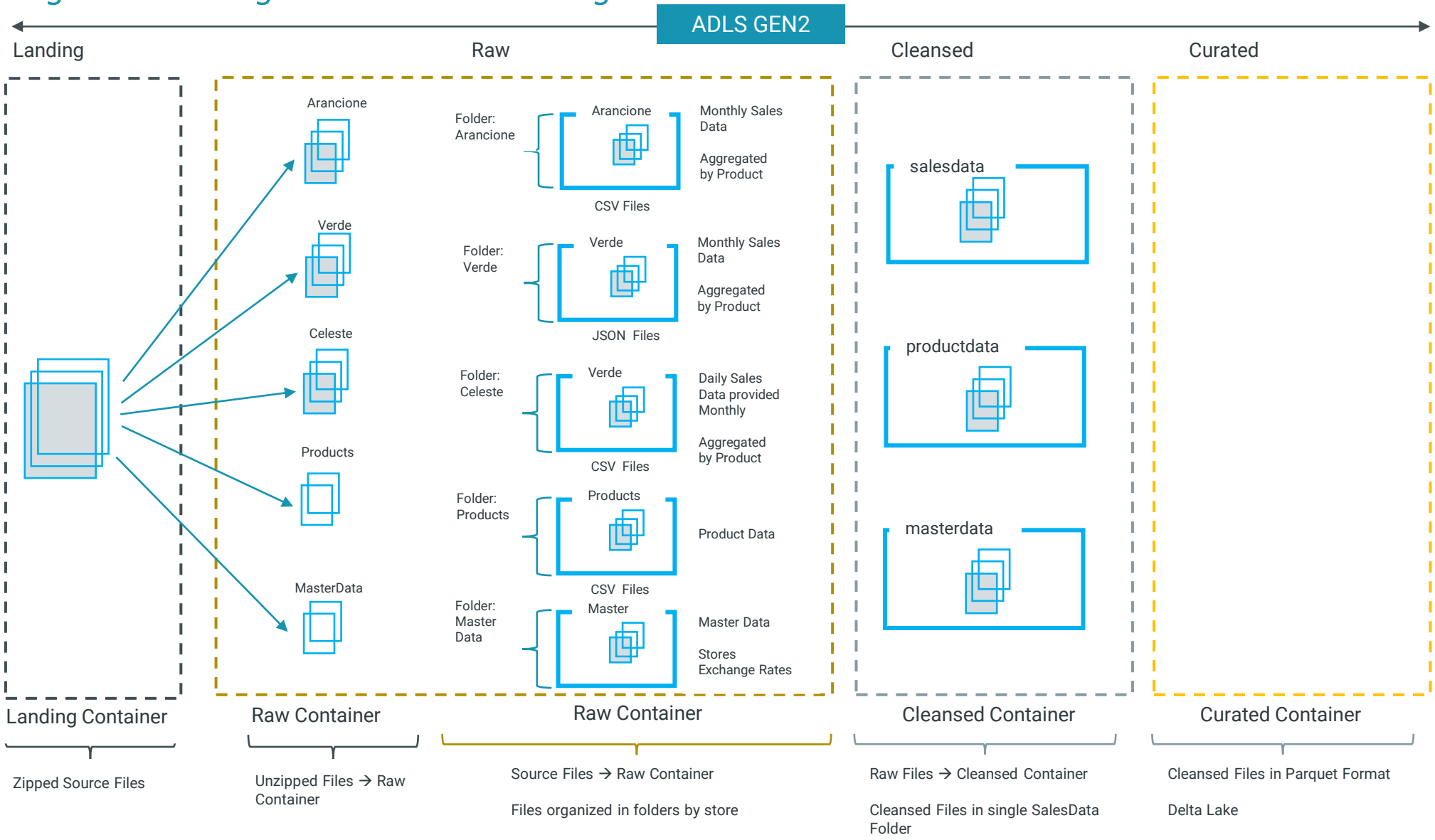
Stage of the ETL Process



Stage of the ETL Process that we will focus on

DATA SOURCES AND DATA ORGANIZATION

How the files are ingested and organized in Azure storage

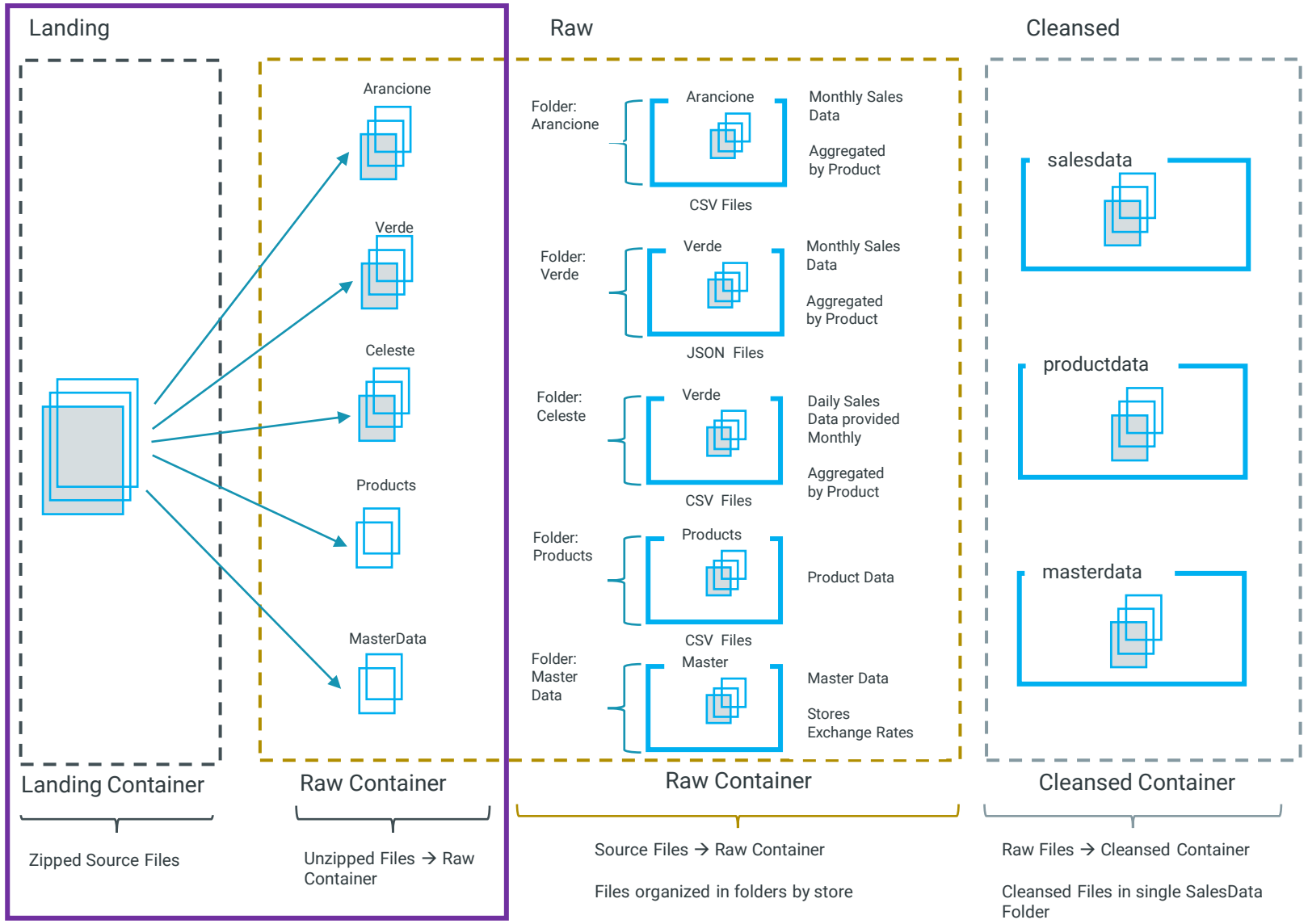


SECTION 3

BUILDING THE DATA PIPELINE

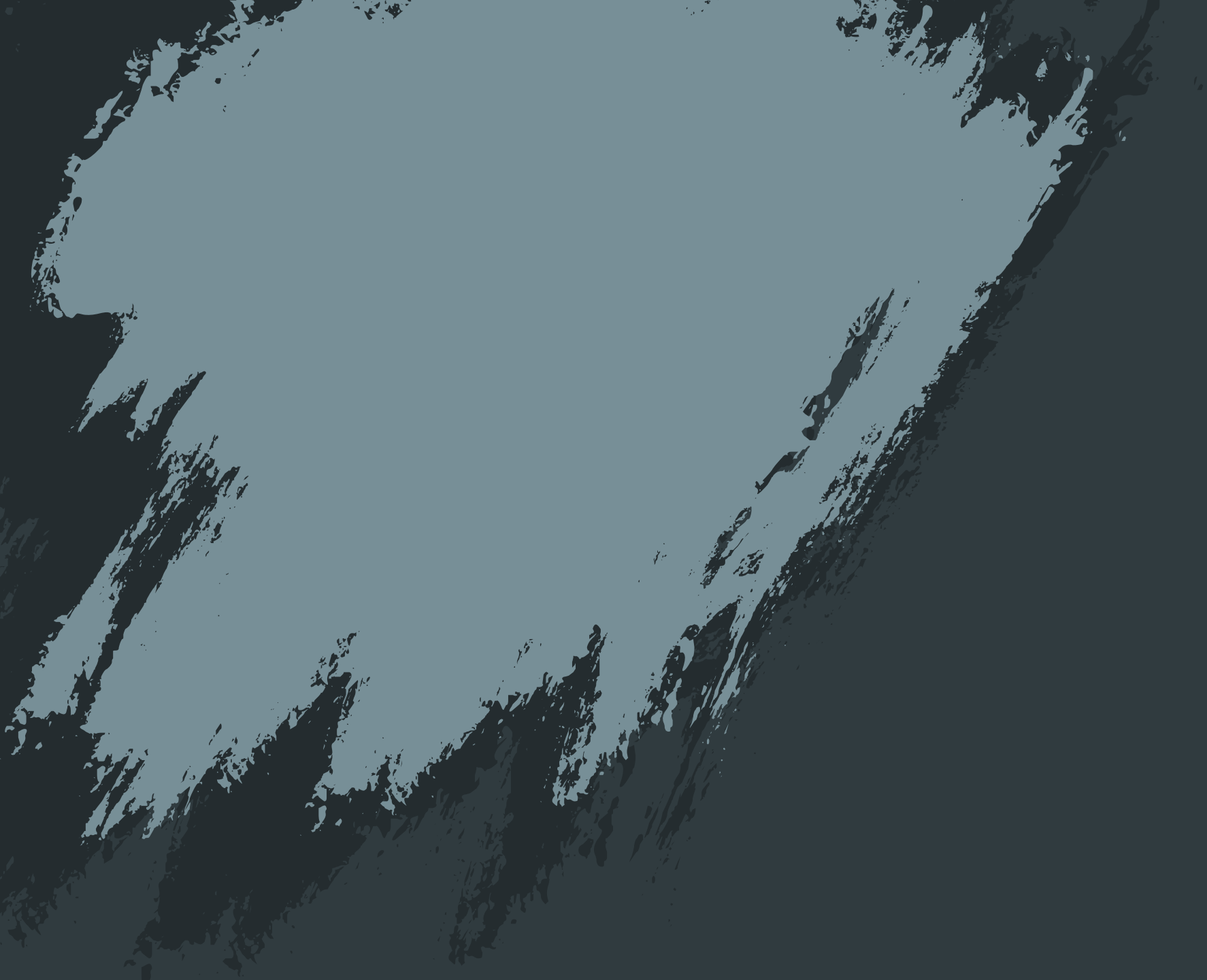
DATA SOURCES AND DATA ORGANIZATION

Ingesting source data



What we will Implement?

- *Build a data factory pipeline*
- *Copy zipped source data from the landing container*
- *And copy them as unzipped data into the raw container*
- *Review pipeline execution*
- *Organize/Review pipeline components*

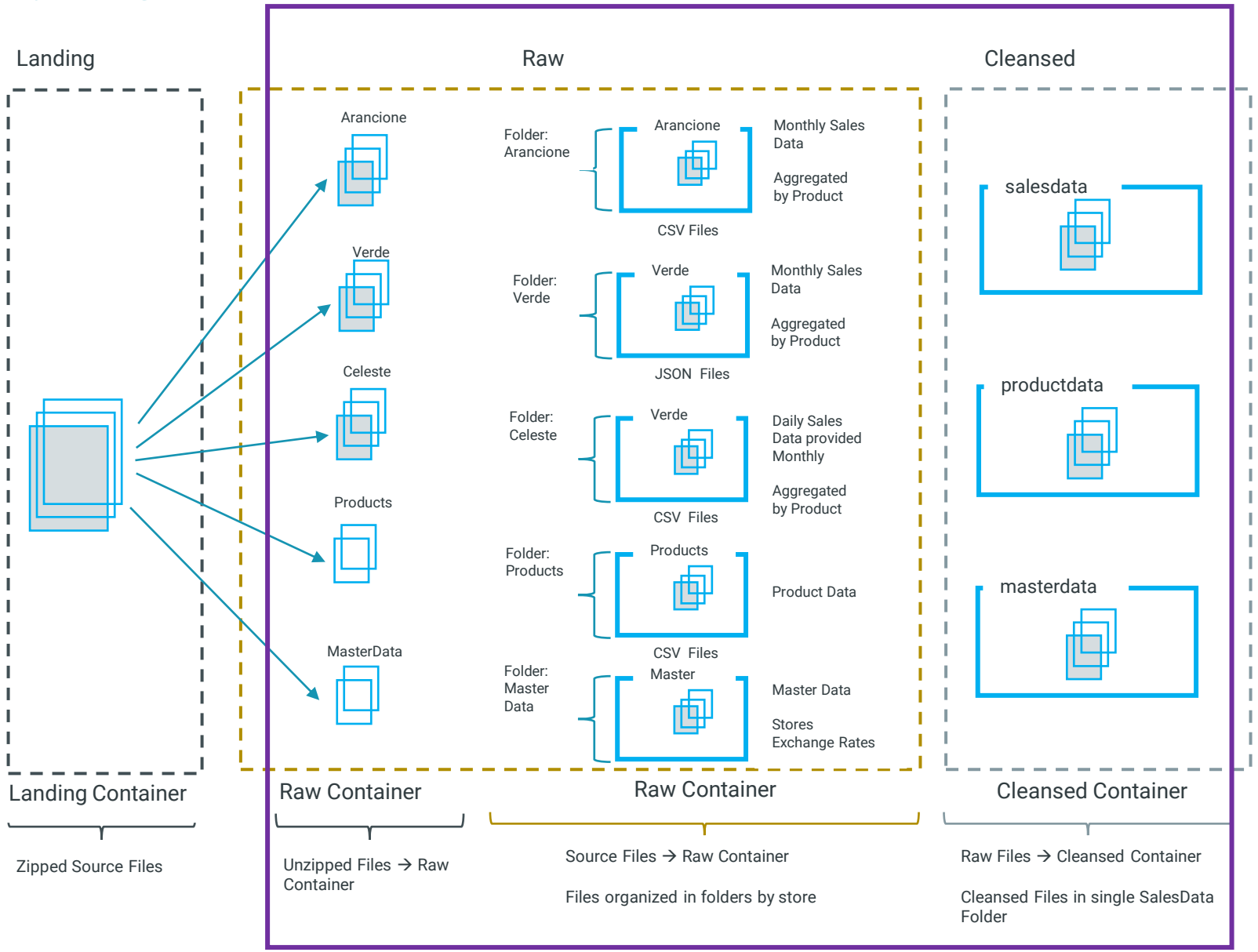


SECTION 4

IMPORTING DATA

IMPORTING DATA

Importing Semi-structured data



What we will Implement?

- *Build a data factory pipeline*
- *Copy Arancione CSV files from raw to cleansed container*
- *Review pipeline execution and results*
- *Organize/Review pipeline components*

SECTION 5

NAMING CONVENTIONS

NAMING CONVENTIONS

Naming conventions for resources and components within Azure and Azure Data Factory

RESOURCE GROUP	STORAGE ACCOUNT	ADLS GEN2	DATA FACTORY
{ABCDEF}-{ENV}-RG	ST{ABCDEF}{ENV}001	ADLS{ABCDEF}{ENV}	{ABCDEF}-{ENV}-ADF
e.g.: *vinoworld-dev-rg*	e.g.: *adlsvinoworlddev001*	e.g.: *adlsvinoworlddev*	e.g.: *vinoworld-dev-adf*
Suffix “rg” for resource group	Prefix “st” for storage	Prefix “adls” for Azure Data Lake Storage	Suffix “adf” for Azure Data Factory
Suffix “env” identifies environment	Specify “env” to identify environment	Specify “env” to identify environment	Specify “env” to identify environment

Naming the components of your project using best practices and doing it consistently makes the solution more manageable and easy to maintain

NAMING CONVENTIONS

Naming conventions for resources and components within Azure and Azure Data Factory

LINKED SERVICE FOR STORAGE	PIPELINE	DATASET	DATA FLOW
LS_ADLS{ABCDEF}_{ENV}	PL_{ABCDEF}	ABS_{TYPE}_{ABCDEF}	DF_{ABCDEF}
e.g.: *ls_adlsvinoworld_dev*	e.g.: *pl_CopySourcetoRaw*	e.g.: *abs_csv_raw_sales*	e.g.: *df_StageSales*
Prefix "ls_st" for linked service for storage	Prefix "pl" to identify pipeline	Prefix "abs" to identify Azure blob storage	Prefix "df" to identify data flows
Suffix "env" identifies environment	Follow by a descriptive name in camel case	Follow by file type whether it is CSV or JSON	Followed by a descriptive name in camel case to identify the purpose of the data flow
Note: linked service names cannot contain "-" and need only "_"	*Note: pipeline names cannot contain "-" and need only "_"*	Use a descriptive name to identify the purpose of the dataset in camel case	*Note: data flow names cannot contain "-" and need only "_"*

These are some of the suggested naming conventions for the most common components that we will use within Azure and Azure Data Factory

To see a comprehensive set of naming conventions, take a look at the naming conventions suggested by Microsoft in the resources section of this course

REFERENCES

Azure Data Factory Naming Conventions

Rules for naming Azure Data Factory entities - Azure Data Factory | Microsoft Learn

Azure Data Factory Components

Introduction to Azure Data Factory - Azure Data Factory | Microsoft Learn

Copy Data using the Copy Data Tool

Copy data by using the copy data tool - Azure Data Factory | Microsoft Learn

MODULE SUMMARY

In this module we learnt

OVERVIEW

We got an overview of the main components of ADF and how they are related to each other

We learnt about our data sources and how we will ingest the data

BUILDING BLOCKS

We then learnt how to build our first data pipeline using ADF

We learnt how to ingest data and learnt to use the copy activity within ADF

BEST PRACTICES

We learnt how to organize our data pipelines

We learnt some of the best practices in terms of naming conventions for Azure and ADF resources