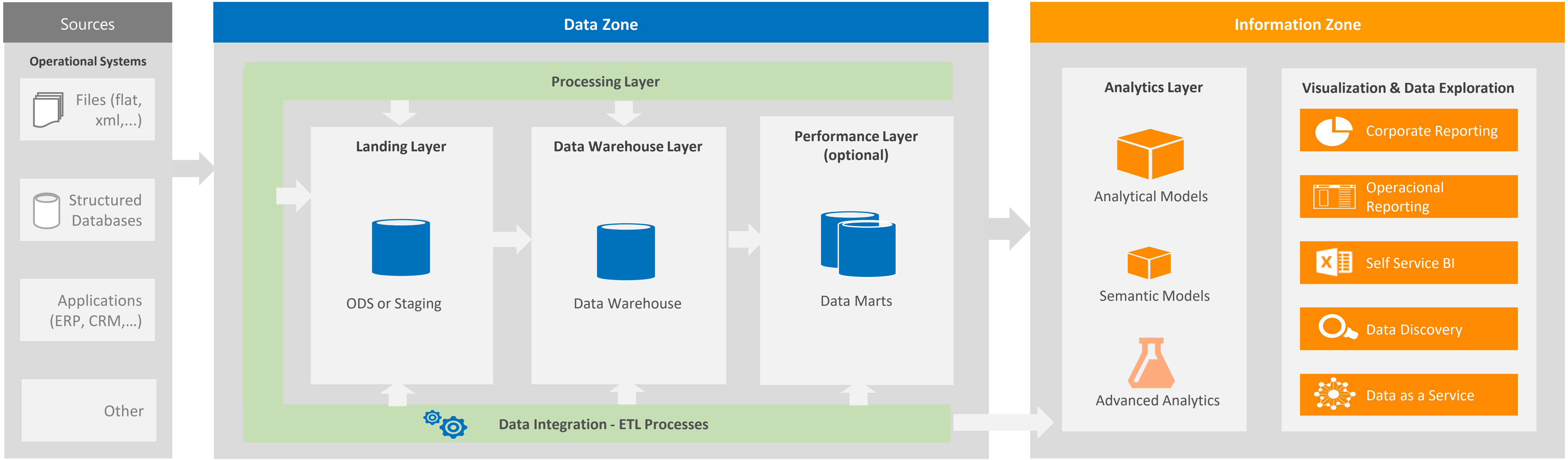
Traditional Data Warehouse Architecture



ODS – operational data store

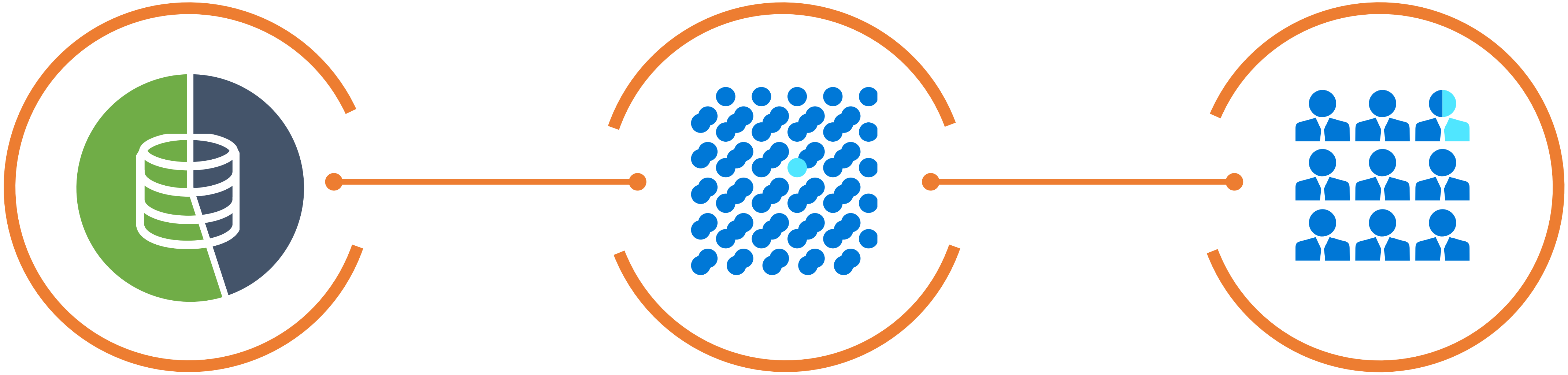
Windows iOS Android HTML5

Data Integration

Data integration is a process in which heterogeneous data is retrieved and combined as an incorporated form and structure.

- Extract, Transform and load (ETL)
- Integrate structured and unstructured data
- Multiple sources
- Multiple destinations
- Data Modeling
- Data profiling
- Data Cleansing, Data Merging / Data Enrichment

Access to data remains top issue



Less than half of structured data is actively used in decision-making

Less than 1% of the unstructured data is analyzed or used

97%* of executives find data silos harmful to their organization

*83% of executives confirm their organizations have data silos

Harvard Business Review, 2017:
https://hbr.org/2017/05/whats-your-data-strategy

American Management Association
2017 survey

Derive real value from your data

⚠ Data silos	⚠ Incongruent data types	⚠ Complexity of solutions	⚠ Multi cloud environment	⚠ Rising costs
✓ One hub for all data	✓ Support for diverse types of data	✓ Unlimited data scale	✓ Familiar tools and ecosystem	✓ Lower TCO

On-premises, hybrid, Azure

Connect with confidence

All-inclusive connectivity

More than 80 natively built and fully managed connectors, no added cost, new connectors added monthly

Efficient and resilient data transfer by leveraging the full capacity of underlying network bandwidth, up to 2 GB/sec throughput

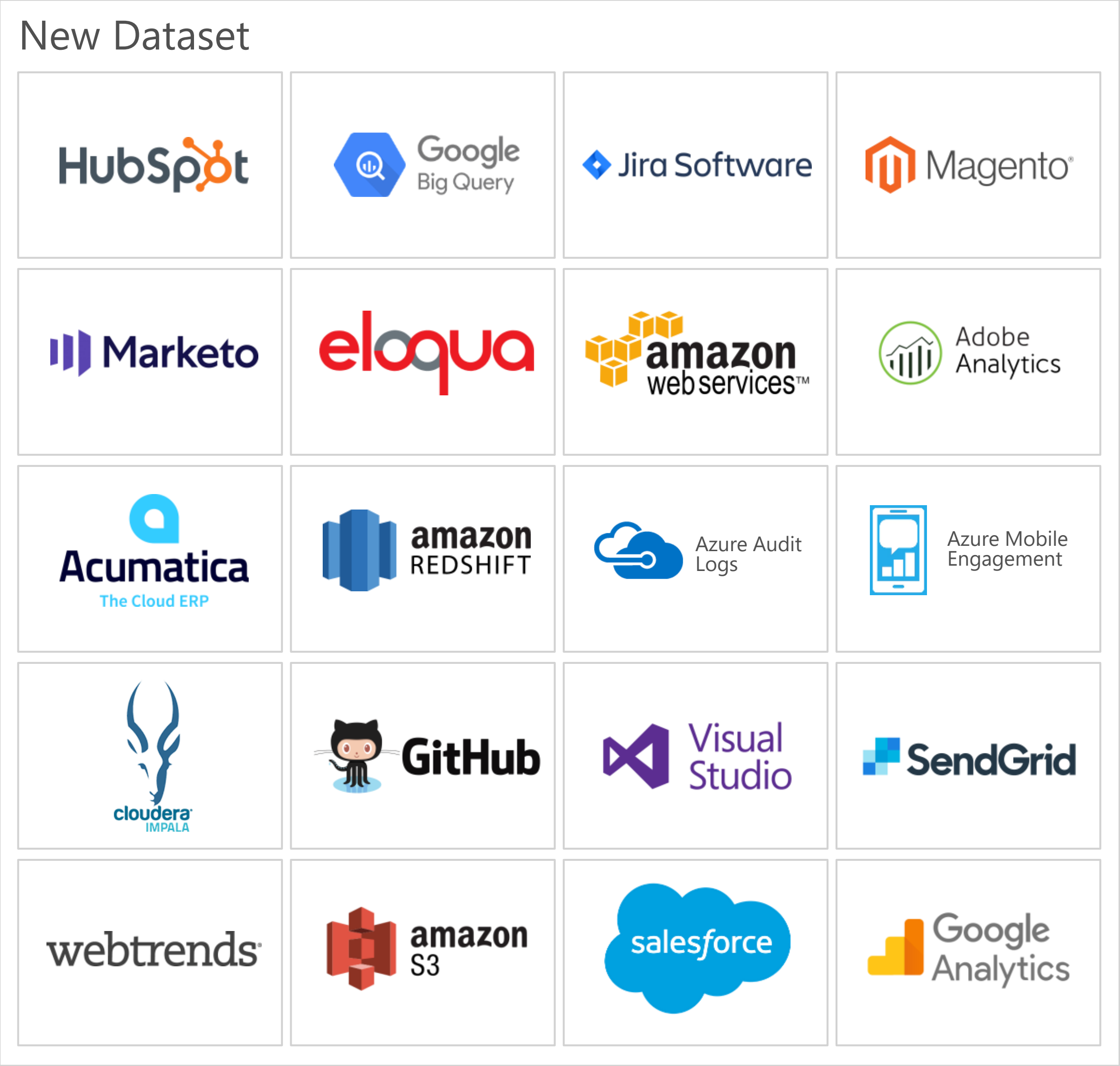
Trusted, global cloud presence

Data Factory availability in 25+ regions, with data movement available globally to help ensure compliance & reduced network egress costs.

Security & compliance peace of mind

Native integration with Azure Active Directory (AAD) and Azure Key Vault (AKV) for identity and access management to cloud solutions & applications, based on centralized policy and rules

HIPAA, HITECH, ISO/IEC 27001, ISO/IEC 27018, CSA STAR certification.



New Dataset

Reduce integration costs

Serverless, fully managed service

No infrastructure to manage, no hardware to upgrade
Scales on demand
Pay only for what you use.

One data integration service for everyone

Reduce integration tool fragmentation & costs
Flexibility to work how you please, visually or using code (Python, .NET or ARM)

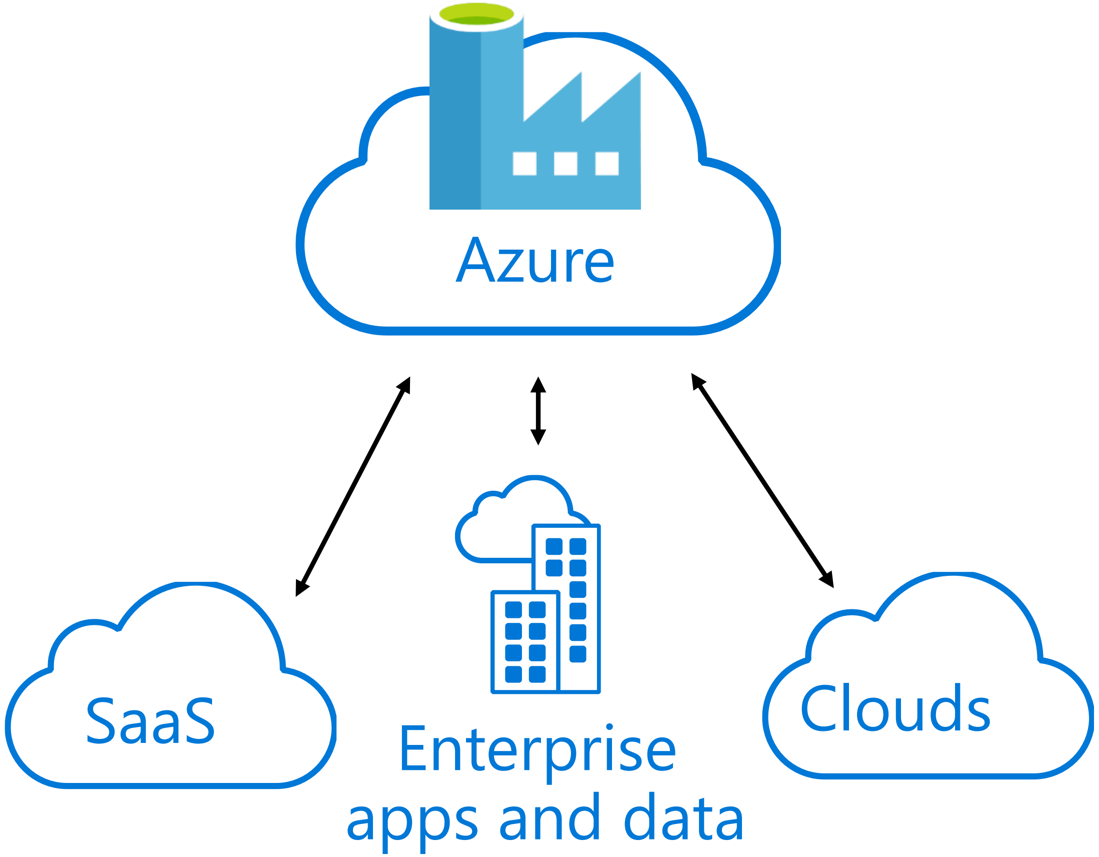
Fast and scalable transformations with Spark

Azure Databricks' Spark engine powers data transformations for fast and fully managed data transformations

.

Reduce development overhead

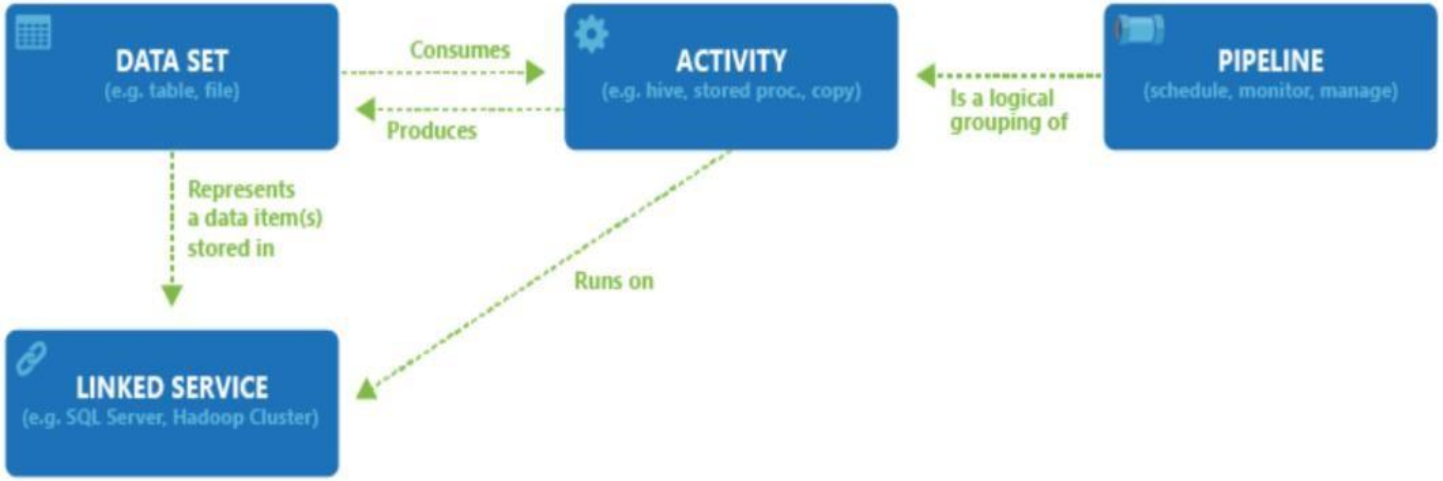
Migrate to the cloud by moving SSIS packages into Azure without redevelopment
Use existing tools for new development.
Full integration with GitHub for team collaboration.



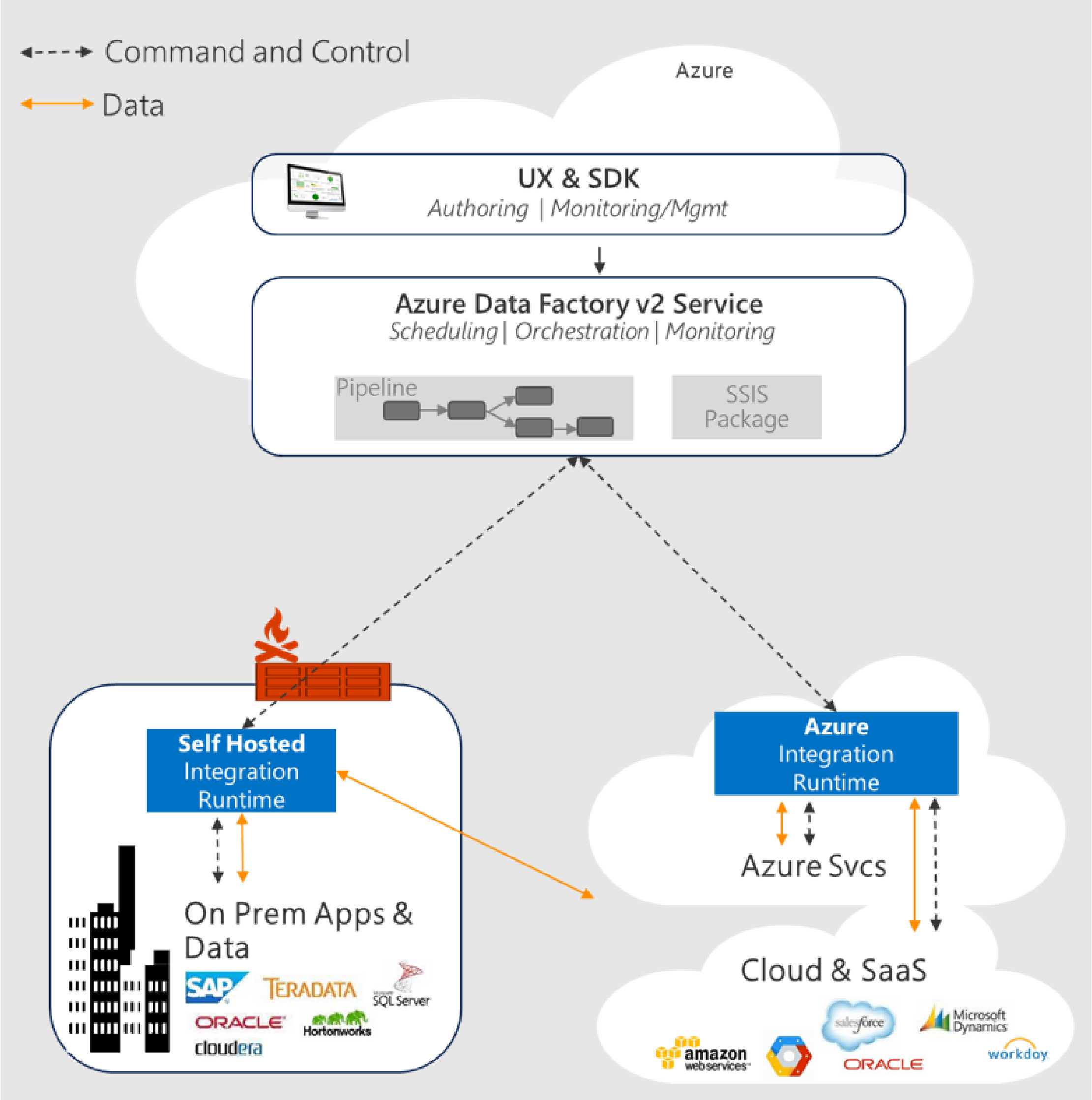
Azure Data Factory - Data Integration Service

Azure Data Factory (ADF) is a cloud-based data integration service **that orchestrates and automates the movement and transformation of data**.

It orchestrates existing services that collect raw data and transform it into ready-to-use information. ADF is used to **collect data from many different data sources, ingest and prepare it, organize and analyze it with a range of transformations, then publish ready-to-use data for consumption.**


DATA SET
(e.g. table, file)
Consumes
Produces
ACTIVITY
(e.g. hive, stored proc., copy)
Is a logical grouping of
PIPELINE
(schedule, monitor, manage)
Represents a data item(s) stored in
Runs on
LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Azure Data Factory - Data Integration Service



Data Factory

A data integration account.

Location of orchestration, service metadata

Integration Runtime (IR)

ADF's execution engine

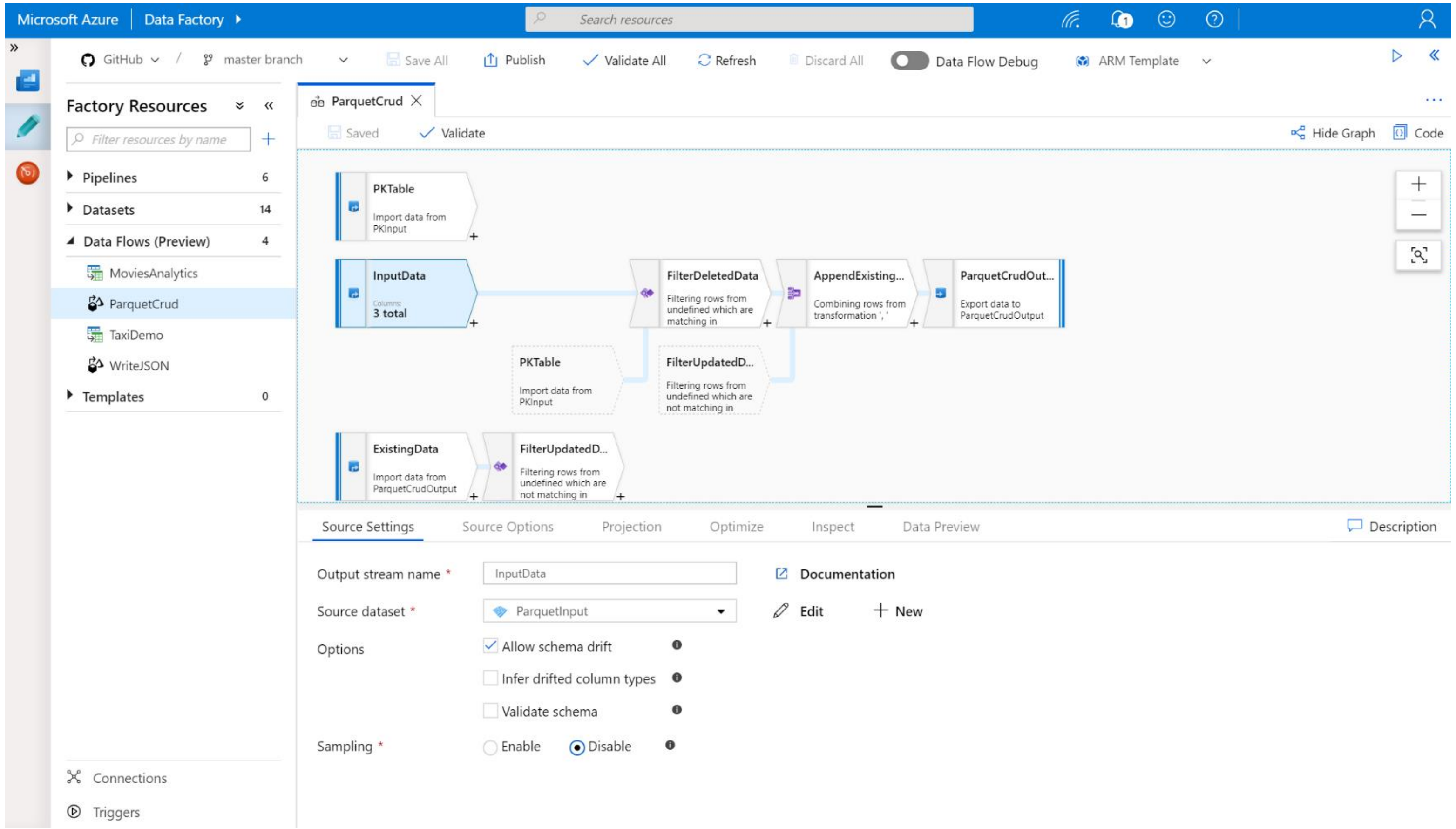
Three core capabilities:

- data movement
- pipeline activity execution
- SSIS package execution

What are Mapping Data Flows?

Data Flow is a new feature of Azure Data Factory to build data transformations in a visual user interface

- Transform at scale, in the cloud
- Code-free pipelines do NOT require understanding of Spark / Scala / Python / Java
- Serverless scale-out transformation execution engine
- Resilient data transformation Flows built for big data scenarios with unstructured data requirements
- Operationalized with Data Factory scheduling, control flow and monitoring



Output stream name * InputData

Source dataset * ParquetInput

Options ☑ Allow schema drift
□ Infer drifted column types
□ Validate schema

Sampling * ○ Enable ● Disable

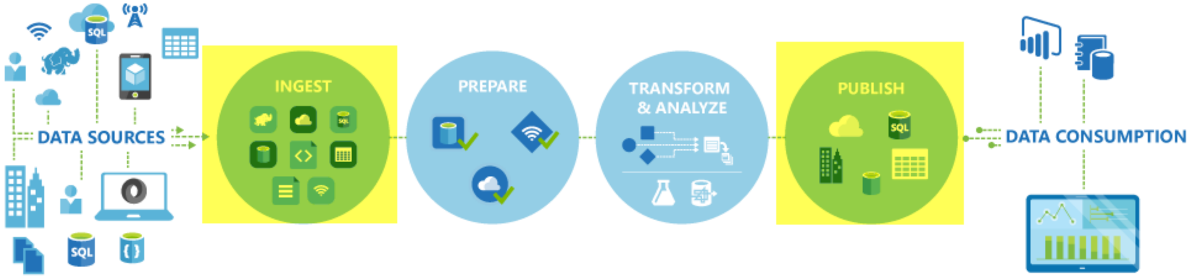
Schema Drift

- In most real-world data integration solutions, source and target data stores will change shape
    - Source data fields will change name
    - Number of columns will change over time
- Traditional ETL processes break when schemas drift
- Mapping Data Flow has built-in facilities for flexible schemas to handle schema drift
    - Patterns, rule-based mapping, byName function
    - Source: Read additional columns on top of what is defined in the dataset source
    - Sink: Write additional columns on top of what is defined in the dataset sink

Azure Data Factory Concepts

Data Factory Copy Activity

In Azure Data Factory, you can use the Copy activity to copy data among data stores located on-premises and in the cloud. After you copy the data, you can use other activities to further transform and analyze it. You can also use the Copy activity to publish transformation and analysis results for business intelligence (BI) and application consumption.



The Copy activity is executed on an integration runtime. You can use different types of integration runtimes for different data copy scenarios:

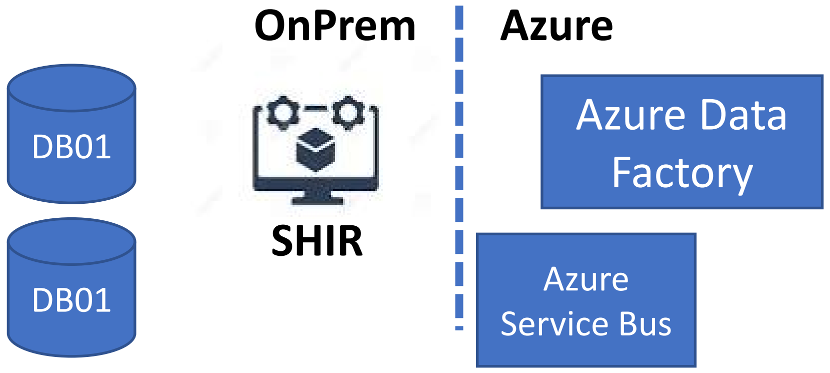
- When you're copying data between two data stores that are publicly accessible through the internet from any IP, you can use the **Azure integration runtime** for the copy activity. This integration runtime is secure, reliable, scalable, and globally available.
- When you're copying data to and from data stores that are located on-premises or in a network with access control (for example, an Azure virtual network), you need to set up a **self-hosted integration runtime**.

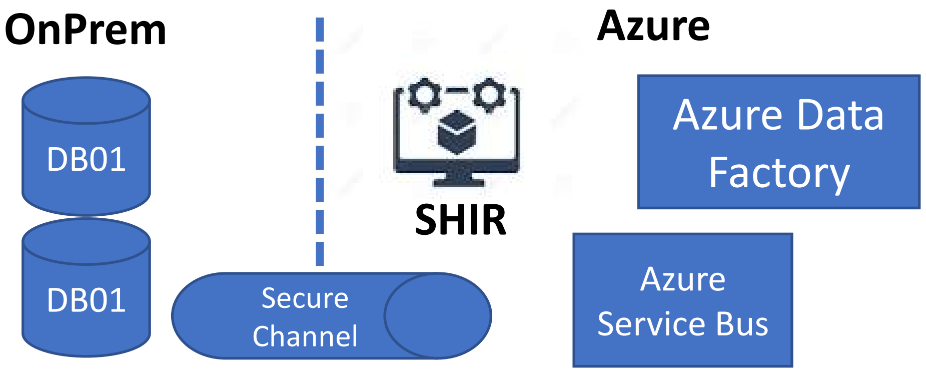
Azure Data Factory Concepts

Data Factory Self-Hosted Integration Runtime

- The Integration Runtime is a **customer managed data integration infrastructure used by Azure Data Factory to provide data integration capabilities across different network environments**. It was formerly called as Data Management Gateway.

- The integration runtime is **capable of moving data in and out of data stores within private network**, as well as dispatching activities against compute service within private network. **You can install a self-hosted integration runtime on an on-premises machine or a virtual machine inside a private network**. This was formerly called the Data Management Gateway (DMG) and is fully backward compatible. Note: An Integration Runtime instance can be registered with only one of the versions of Azure Data Factory (version 1 -GA or version 2 -GA).

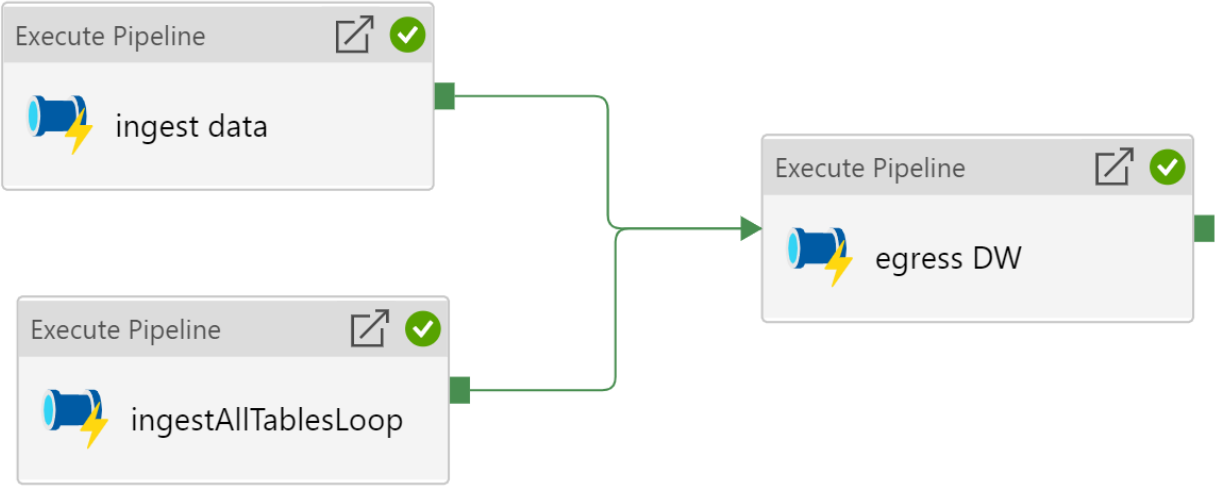




ADF Data Factory – Main pipeline

In this task, you will create the main pipeline

1. Do the previous step to add execute pipelines for the created pipelines: '… IngestAllTablesLoop and '… egress to Azure DW'
2. **Connect** the new Activity to the previous ones as described


Execute Pipeline
ingest data
Execute Pipeline
ingestAllTablesLoop
Execute Pipeline
egress DW