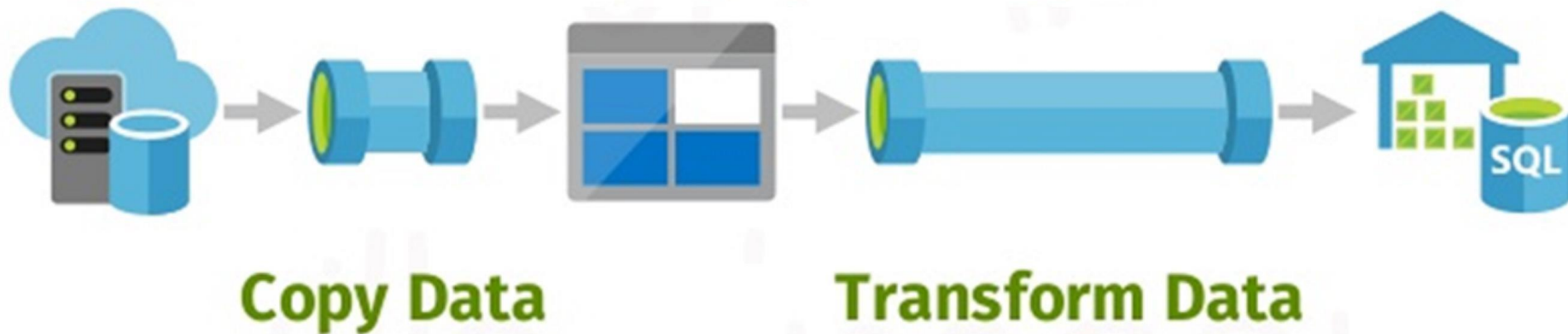




# Azure Data Factory

Cloud version of SSIS

# What can you do in Azure Data Factory?



## Copy Data

More than 80 connectors to different services are available



## Transform Data

Using newly added Data Flow, now Data Factory is complete cloud based ETL tool.



**Azure Data Factory**

## Definition:

Azure Data Factory (ADF) is a hybrid data integration service that enables you to quickly and efficiently create automated data pipelines – without having to write any code!



**Azure Data Factory**

- Hybrid Data Integration Service
- Simplifies ETL at scale
- Enables modern data integration
- Drag and drop interface
- Over 80 connectors available
- Move, transform and save data
- Managed Service
- Create Data Driver workflows
- Orchestrate and automate data movement
- Transform and store data
- Operationalize the process
- ETL or ELT scenarios

# Data Factory on Azure Ecosystem

01

Migration?

Data Factory excels in periodic data loads and transformation instead.



02

Streaming?

ADF can orchestrate, but there are other dedicated services for streaming



03

Transformations?

Data flows for simple ones, but you can use Databricks or HDInsight for more complex transforms



# SSIS vs Data Factory Cluster Types

## SSIS

More code-free transformations  
On Premises connectors (e.g excel)

## Data Factory

Much higher scalability  
Cloud and SaaS Connectors  
Event based Triggers  
Can use SSIS Packages



## Data Factory considerations

### Two versions

ADF V2 is the current and improved version

### Build options

PowerShell,  
.Net, Python,  
REST, ARM

### Highly integrated

DevOps, Key  
Vault, Monitor,  
Automation

### No data storage

Need to persist  
data by the end.

### Security standards

HTTP/TLS  
whenever  
possible





# **Azure Data Factory Components**

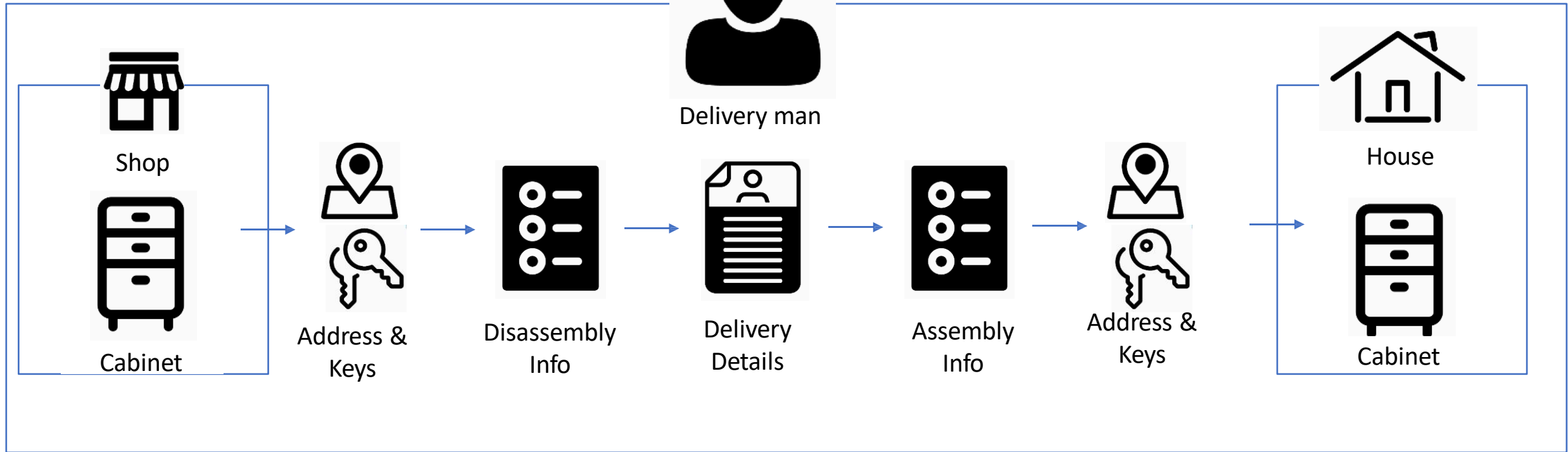




Delivery Manager



Delivery man

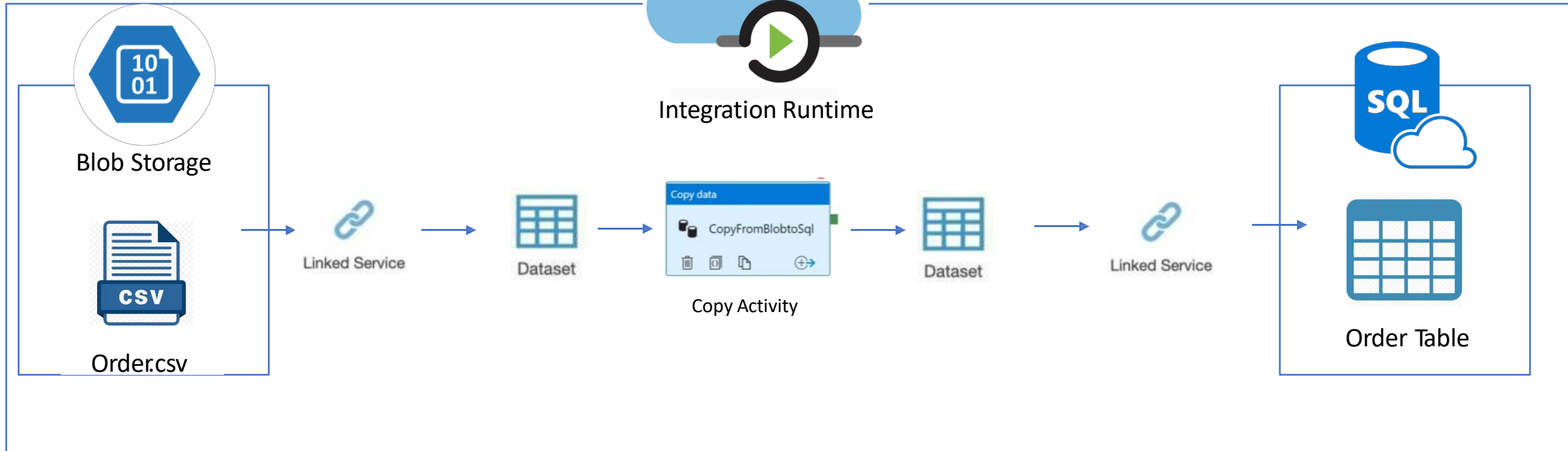




Data Factory Pipeline



Integration Runtime



# Data Factory vs SSIS

## Cluster Types

### Azure Data Factory

Pipeline

Linked Service

Source

Sink

Activity

Data Flow

### SSIS

Package

Connection manager

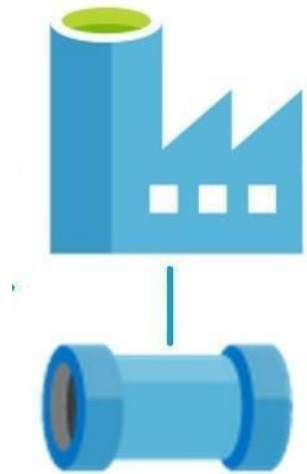
Source

Destination

Control flow task

Data flow





Data Factory  
Pipeline

- Data Factories can contain one or more pipelines
- Logical group of Activities
- Manage Activities as a set
- One Pipeline can have one or more activities

## Azure Data Factory Activities

- Represents a processing step in the pipelines
- Actions to perform on data
  - Ingest data
  - Transform data
  - Store data
- Can be linked
  - Execute sequentially or
  - Run in parallel



# Activity types

01

## Data movement activities

Copy data amongst data stores located on-premises and in the cloud

Data stores – Blob storage, Cosmos DB, Amazon Redshift, Google BigQuery Hive, Maria DB...etc.



02

## Data transformation activities

Transform and enrich data

e.g. Hive, Pig, MapReduce, Spark or Databricks



03

## Control activities

Control pipeline flow

e.g. ForEach, Web

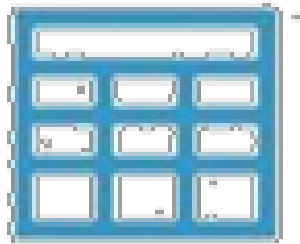




## Data Flows

- Data Flow is a new feature of Azure Data Factory (ADF) that allows you to develop graphical data transformation logic that can be executed as activities within ADF pipelines.
- Two types:
  - Mapping
  - Wrangling





## Dataset

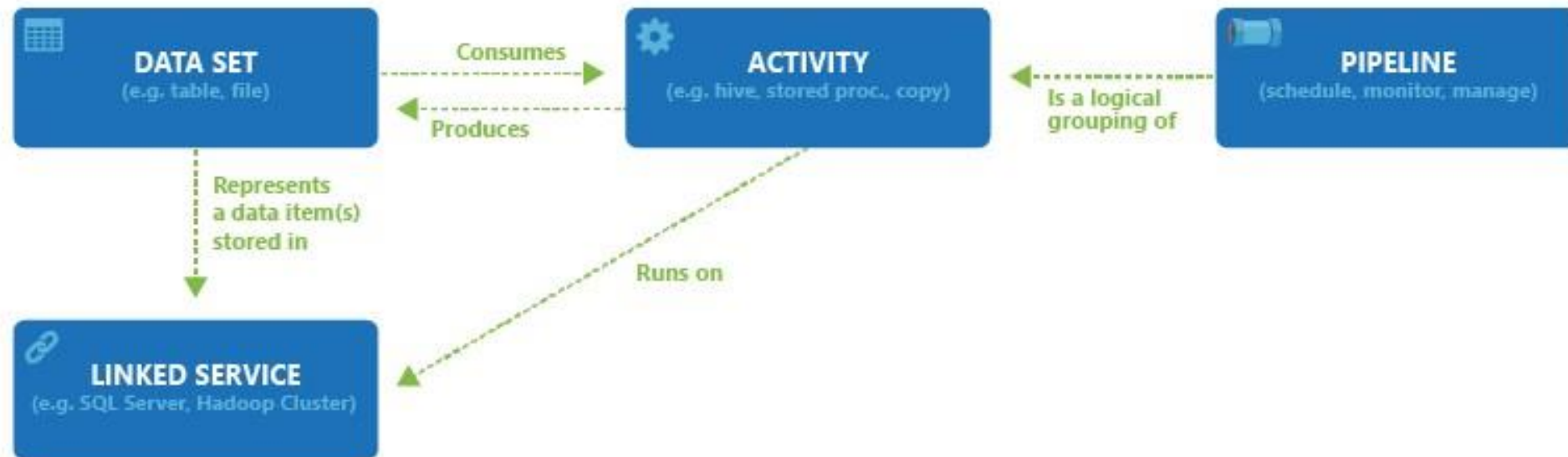
- Simply point or reference the data
- Reference data used in an Activity
  - Files
  - Folders
  - Documents
  - Tables




## Linked service

- Similar to connection string
- Represent the connection information to connect to external resources
  - Datastores like Azure SQL Server
  - Compute resource e.g. Spark Cluster

# ADF Components



A background image showing a desk with two laptops, a tablet, and some papers. The text 'Integration Runtimes' is overlaid on the left side of the image.

## Integration Runtimes

- Provides fully managed, serverless compute infrastructure
  - You don't have to worry about infrastructure provision, software installation, patching, or capacity scaling.
  - Pay only for duration of actual use
- Bridges between the activity and linked service
  - Activity defines the action
  - Linked service define the location



## Integration Runtimes

- **Data Integration Capabilities**
  - **Data Flow**
  - **Data Movement**
    - Format conversion, column mapping, serialization/deserialization etc.
    - Provides the native compute to move data between cloud data stores in a secure, reliable, and high-performance manner.
  - **Activity dispatch** (e.g. Databricks Notebook, HDInsight Hive, pig, spark activity, SP, ADL Analytics U-SQL activity)
  - **SSIS Package execution**

# Q&A

