



Azure Saturday 2019

An introduction to Azure Data Factory

Riccardo Perico

Nice to meet you

Riccardo Perico | rperico@solidq.com | @R1k91

[SolidQ](#)

Data Platform & BI Specialist

10 years working, training and speaking in Microsoft «Data Realm»

MCP: MTA, MCSA



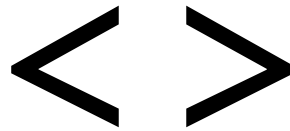
<https://www.linkedin.com/in/riccardo-perico-8b942384/>

Agenda

- Introduction to ADF v2
- Integration Runtime
- Mapping Data Flows
- Demo
- Useful information



DATA FACTORY



What ADF really is?



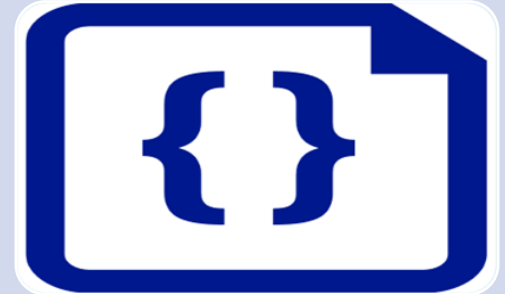
Cloud based
Data
integration
service



Orchestrates &
Automates
Data
movement and
transformation



Allows
Monitoring
and Debugging

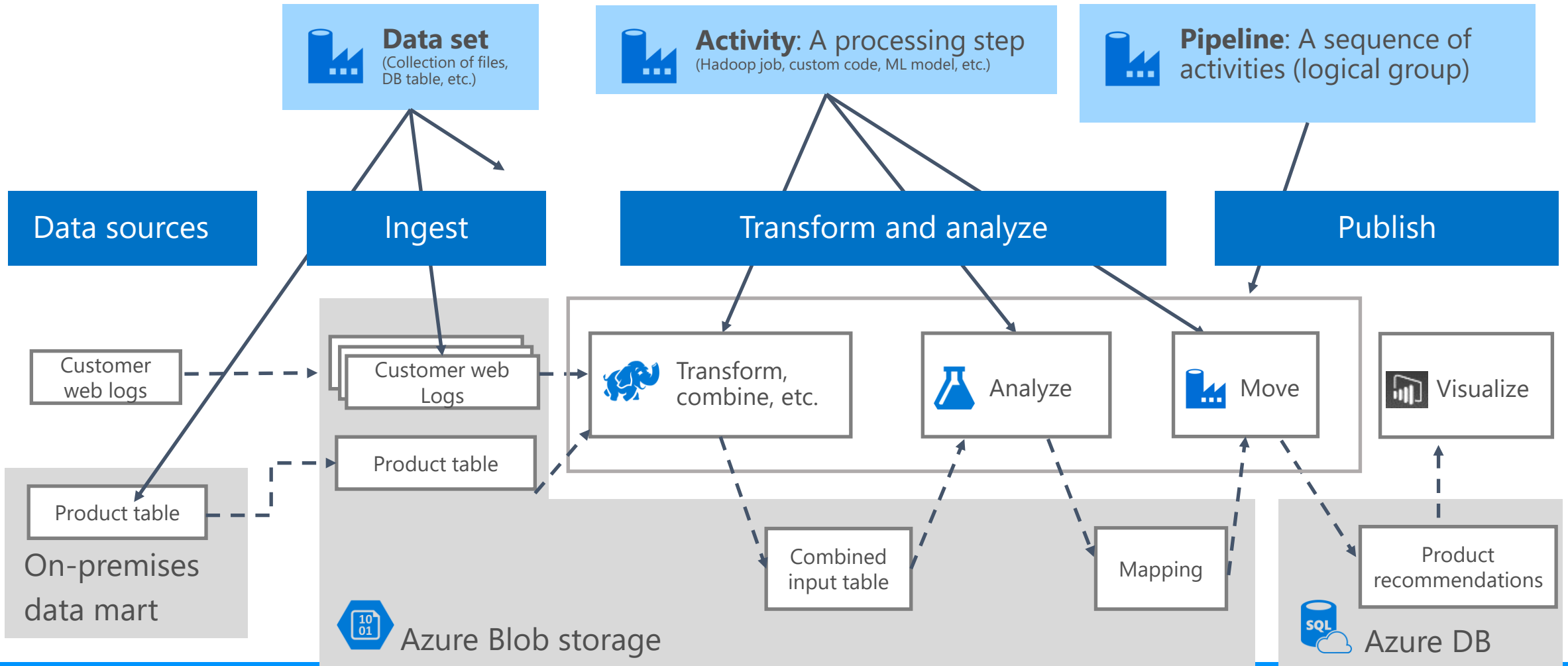























Programmable

The Big Data “problem”

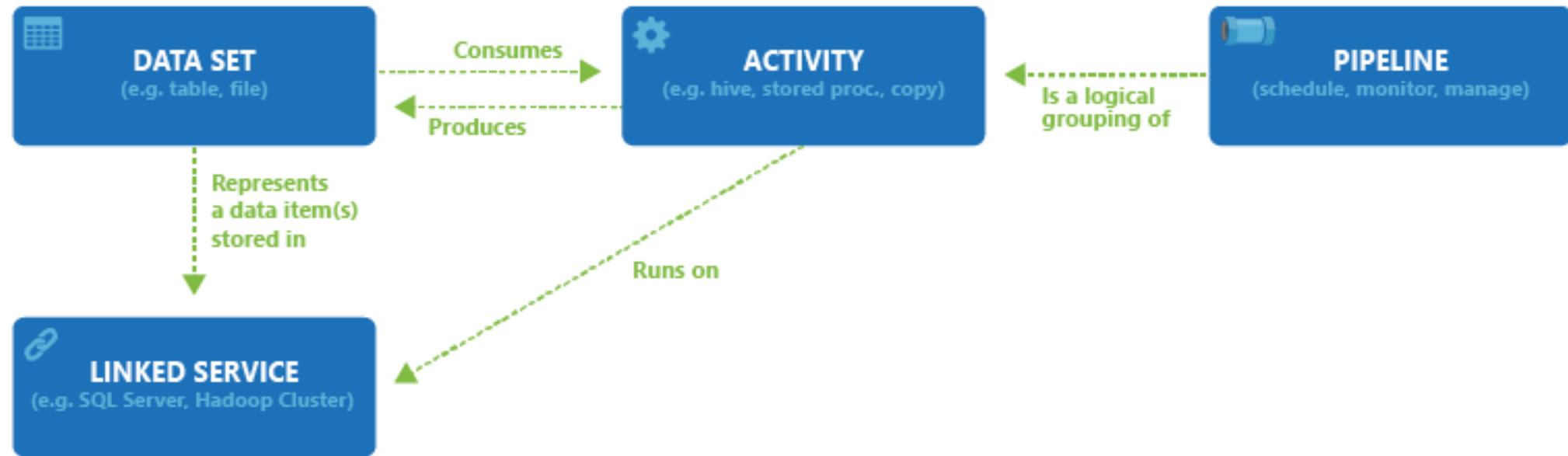


Sample Workflow



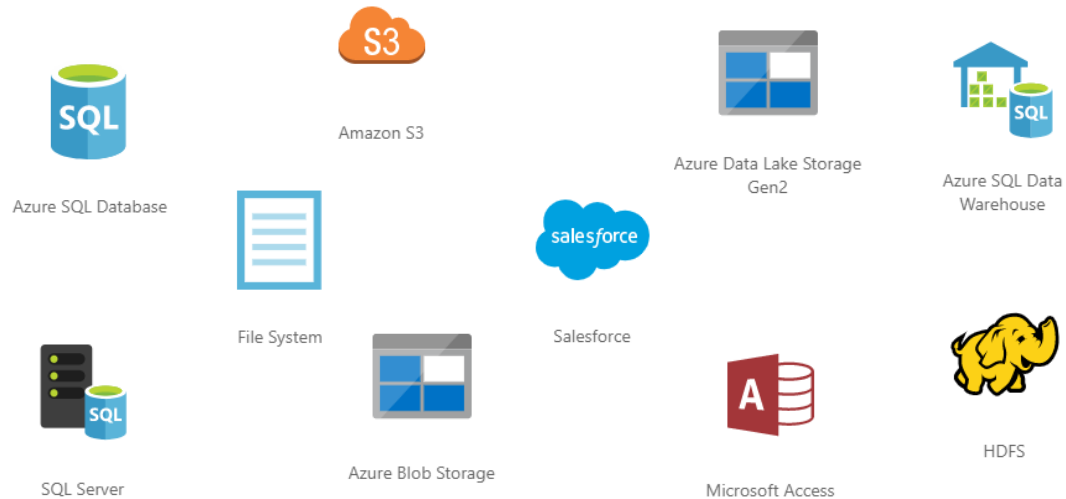
Devices	Device Connectivity	Storage	Analytics	Presentation & Action
	 Event Hubs	 SQL Database	 Machine Learning	 App Service
	 IoT Hubs	 Table/Blob Storage	 Stream Analytics	 Power BI
	 Service Bus	 Cosmos DB	 HDInsight	 Notification Hubs
	 External Data Sources	 External Data Sources	 Data Factory	 Mobile Services
			 Data Lake Analytics	 BizTalk Services

Key concepts



Linked Services

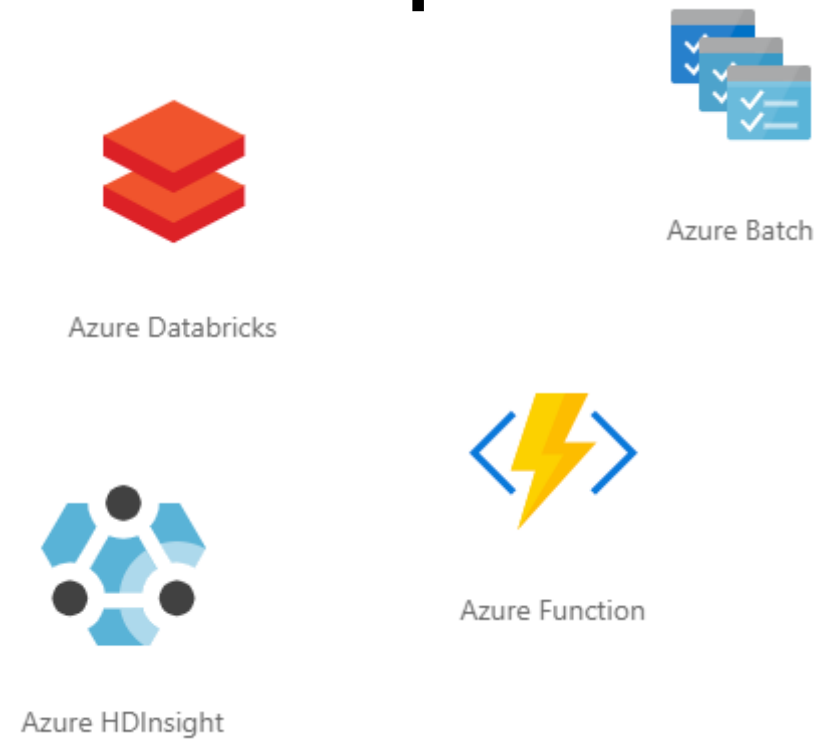
Data Stores



Input **Dataset**

Output **Dataset**

Compute

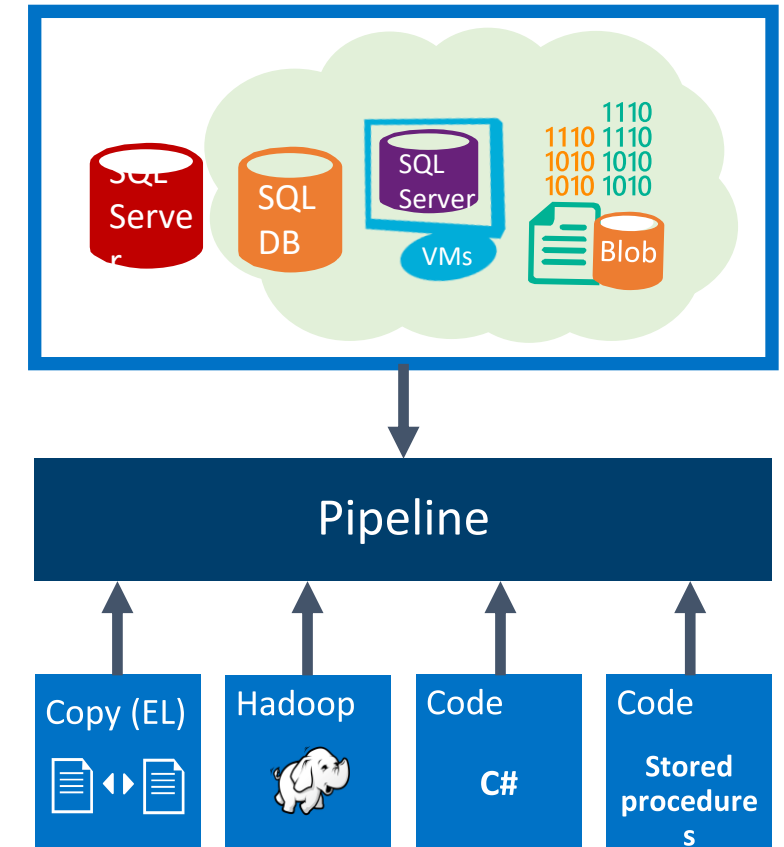


Activities & Pipelines

An **Activity** is a single task in workflow:

- *Copy* from input to output
- *Transform*
 - C#
 - Stored Procedure
 - Hadoop (Map/Reduce, Hive, Pig)
 - ML, Data Lake Analytics
 - Databricks
- *Control*
 - IF, ForEach, Until, Wait, Execute Pipeline
 - Web

Pipeline groups activities



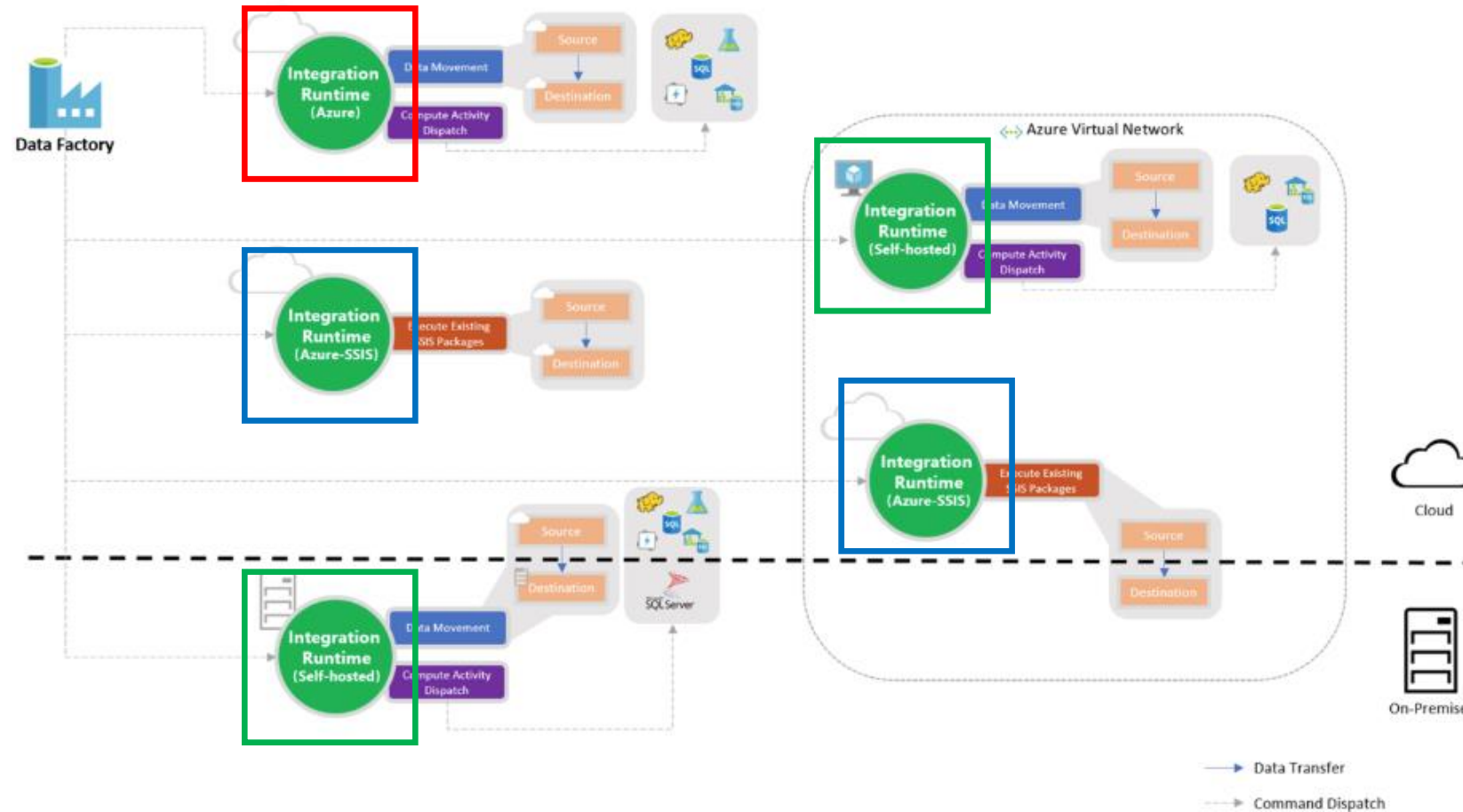
Integration Runtime

- **Bridge** between Activity and Linked Service
- **Compute environment** where activity runs or it's dispatched from

3 types of IR:

- IR Azure
- IR Self-hosted
- IR Azure-SSIS

IR Topology



ADF Location vs IR Location

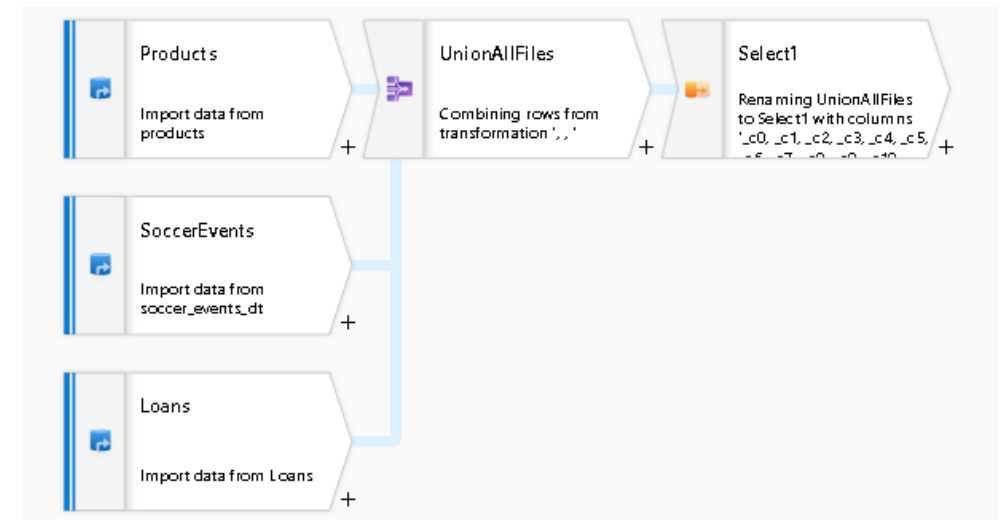
- ADF location → metadata store and triggering pipeline start
- IR location → backend compute engine location (data movement, activity dispatch and SSIS execution)

ADF Location and IR location could be different

IR can use “Auto Resolve”

Mapping Data Flows

- Based on **Spark**
- Use **Databricks** behind the scene
- A lot of **transformations** already available
- Few sources available for now
- This week **GA** announced!





Azure Saturday 2019

Demo: let's put everything together

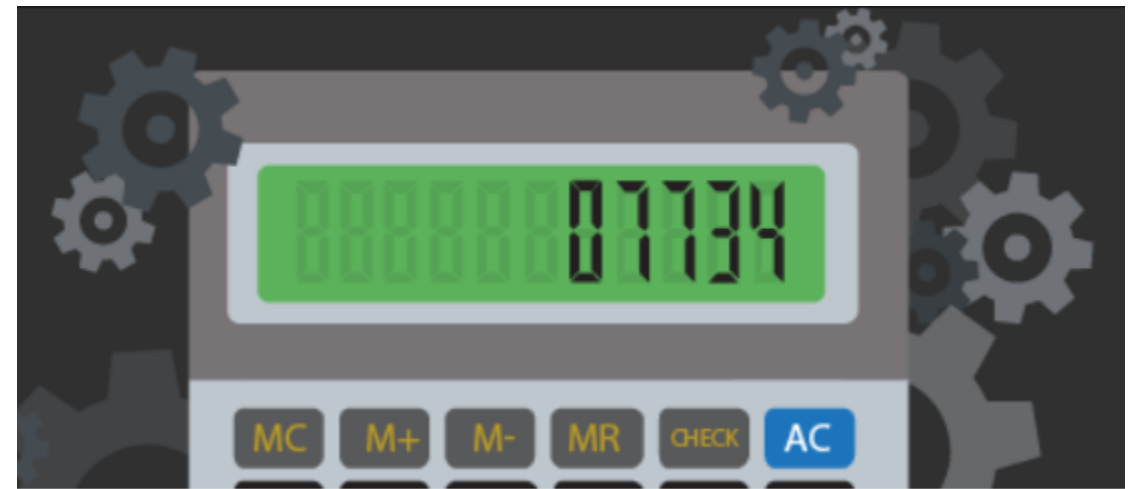
Developer Tools

- **Azure Portal**: Create, Edit. Visual and Textual
- **Visual Studio**: Integrated in VS project
- **Powershell**: cmdlets <https://docs.microsoft.com/en-us/powershell/module/azurerm.datafactories/?view=azurermps-6.13.0>
- **Azure RM Template**

Pricing

Multiple factors affect pricing

- Number of Activities run
- Volume of data moved
- SQL Server Integration Services Compute Hours
- Whether you re-running an activity



<https://azure.microsoft.com/en-us/pricing/details/data-factory/v2/>

Useful Links

- Overview: <http://tiny.cc/domwdz>
- ADF Channel 9: <http://tiny.cc/pdnwdz>
- Blog posts: <http://tiny.cc/6smwdz>
- Quick start and tutorials: <http://tiny.cc/wumwdz>
- GitHub repository – Code and examples <http://tiny.cc/1vmwdz>
- GitHub repository – Hands-on labs <http://tiny.cc/4wmwdz>
- v1 and v2 comparison <http://tiny.cc/txmwdz>



#azuresatpn



Azure Saturday 2019

Q&A



#azuresatpn



Azure Saturday 2019

Thank you!