



Azure Datafactory and Databricks

Sriram Gudimella

Sriram.Gudimella@valuemomentum.com

Chinmaya Kumar Bansal

Chinmaya.Bansal@valuemomentum.com



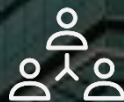
CORPORATE FACTS

Established in 2000



90+

Customers Served



2300+

Employees



23%

CAGR since inception

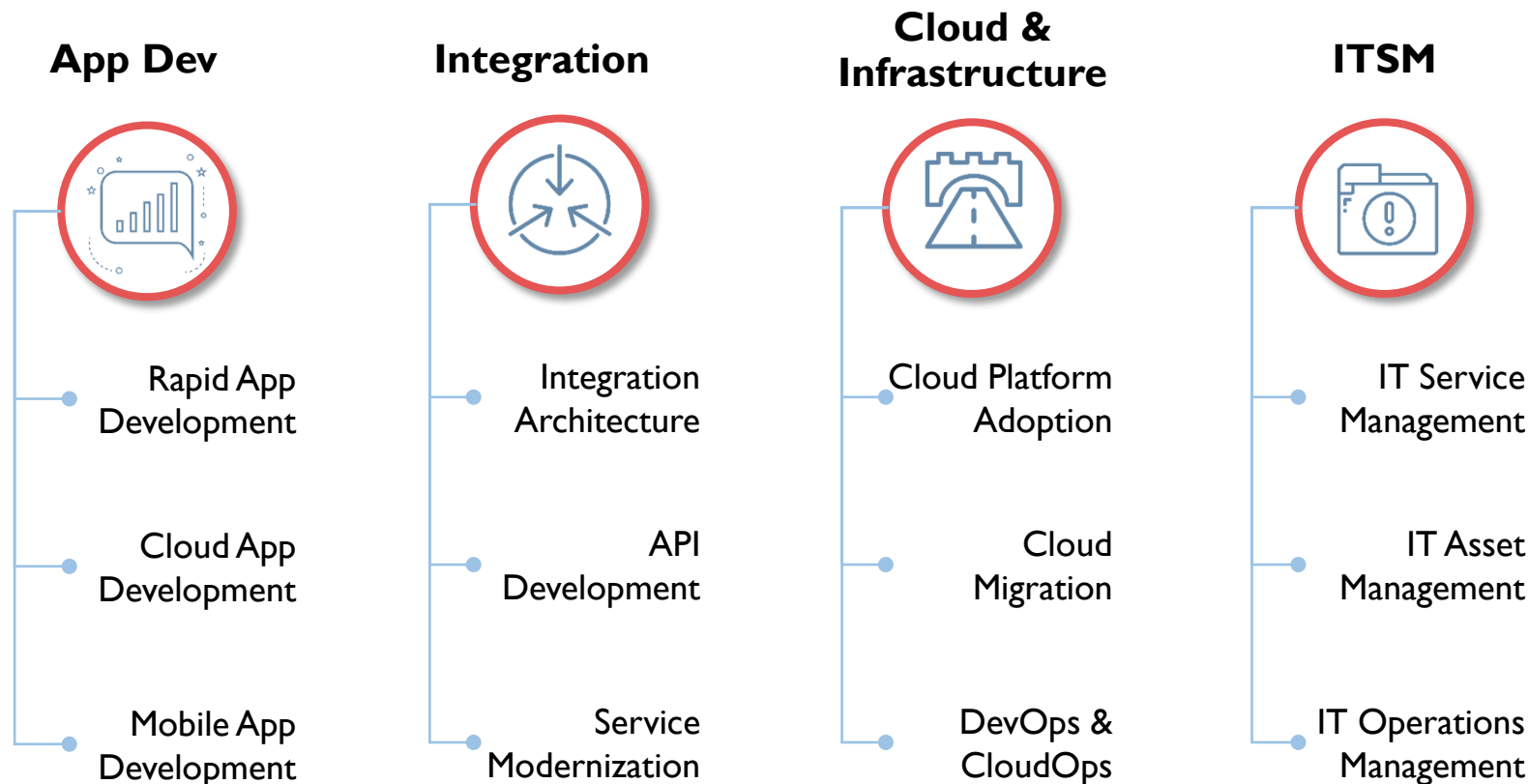


Top 10

NA P/C IT Svs Provider
by # of customers

OUR DIGITAL & CLOUD SERVICES

Customers trust ValueMomentum to rapidly deliver new experiences and stay competitive in today's digital-centric market.*

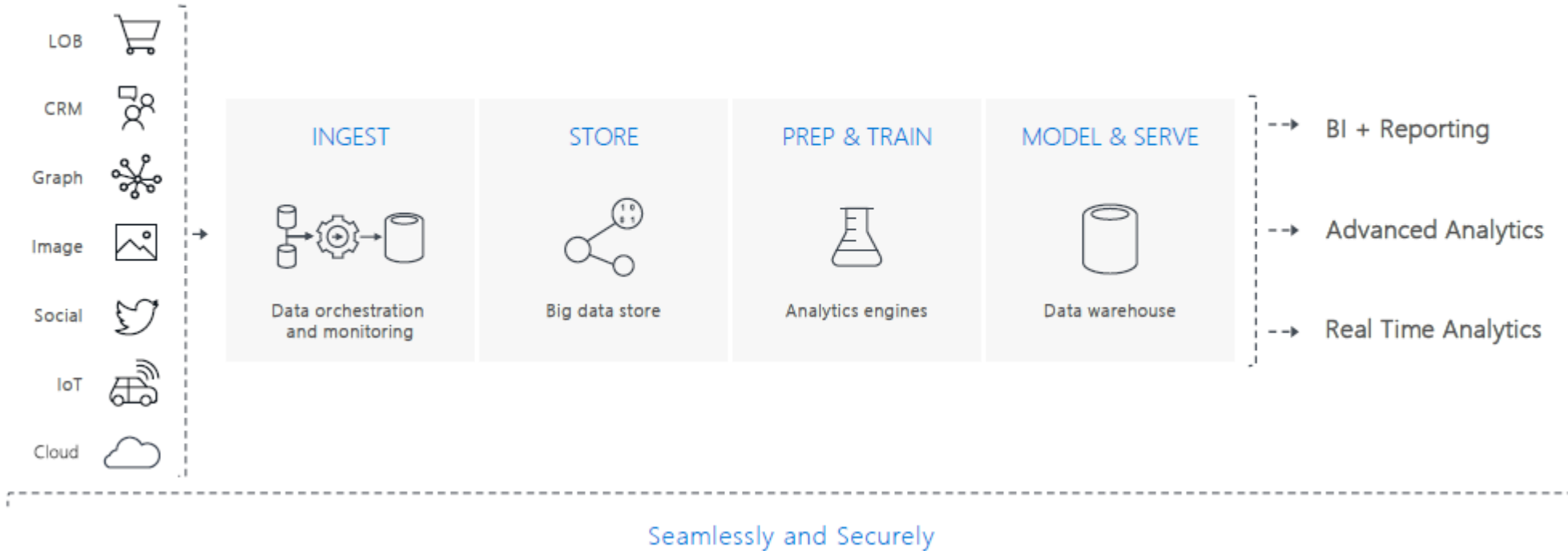


*To learn more, please log on to [ValueMomentum – Digital & Cloud Services](#)

- Typical Data Engineering Scenario
- Tool Selection Dilemma
- Selection Basis
- Demo

Typical Data Engineering Scenarios

Data Engineering Scenarios



Data Engineering Scenarios

Common scenarios



"We want to extend to untapped sources."



Modern data warehouse



"We want to use ML and AI to get deeper insights from our data."



ML and AI on big data



"We want to get insights from our devices in real-time."



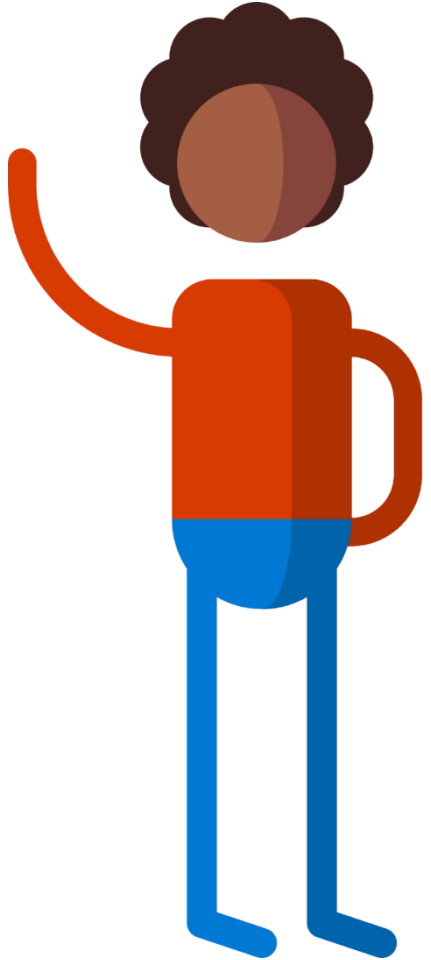
Real-time analytics



Why is Data Engineering Hard?

- Various sources/formats
- Schema mismatch
- Different representation
- Corrupted files and data
- Scalability
- Schema evolution
- Monitoring & Auditing
- Multi activity integration
- Evolve as fast as the business

Typical Project Scenario



You are a data engineer
starting a new project

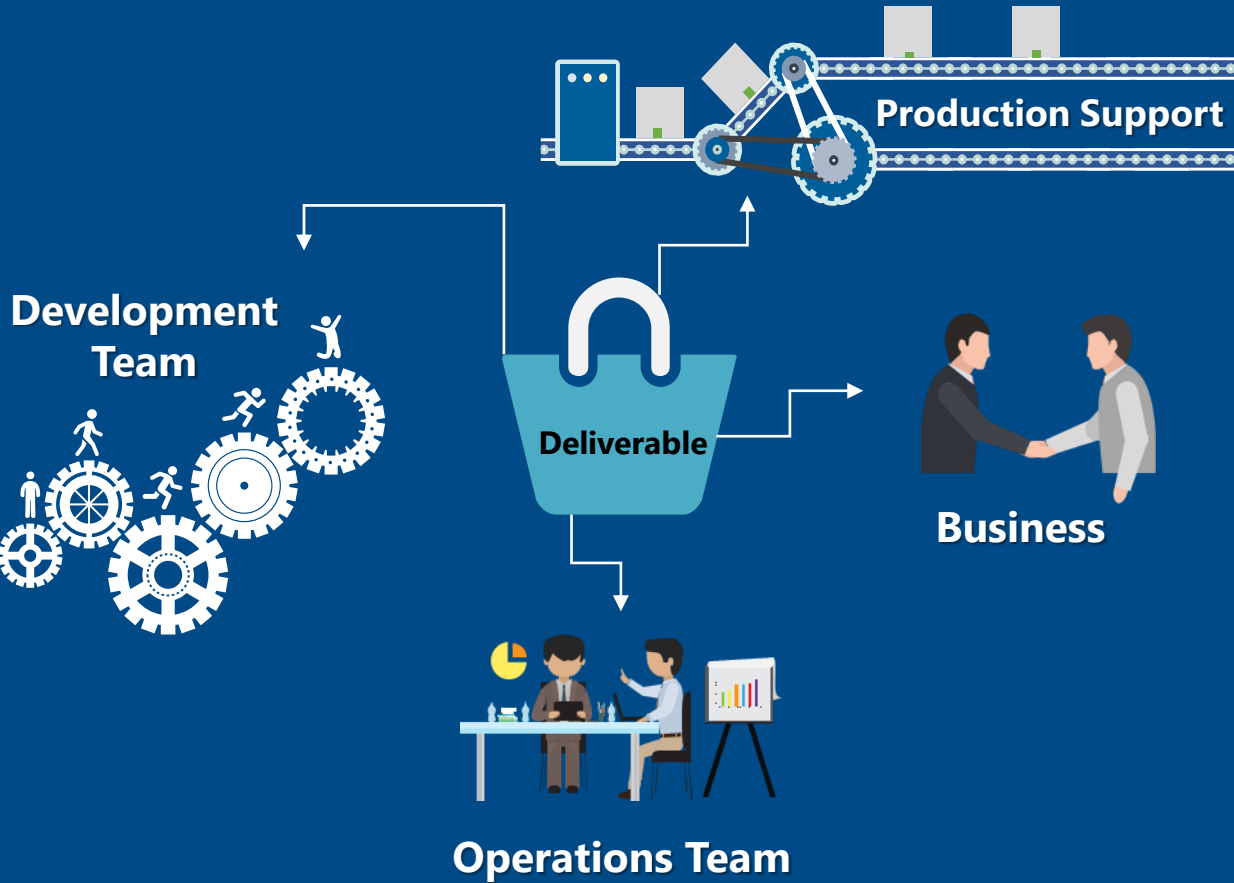
You are a Technical Manager
starting a new project



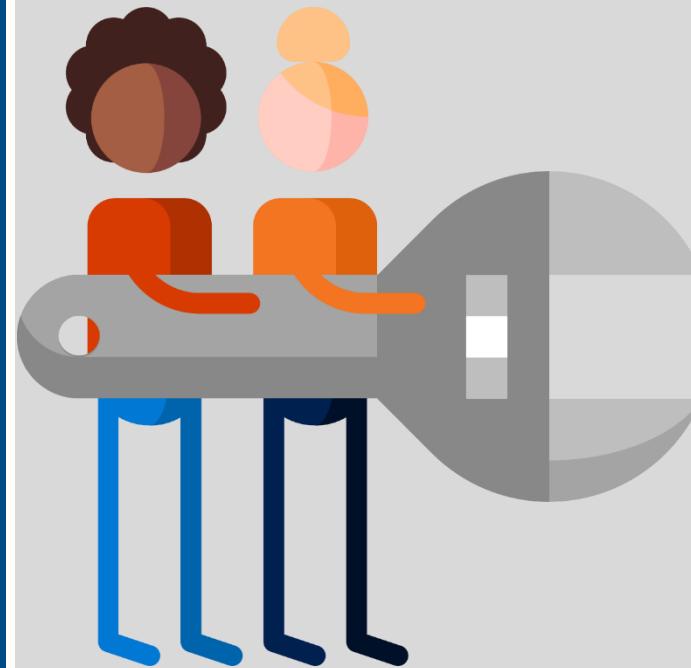
Expectations vs Reality



Expectations:

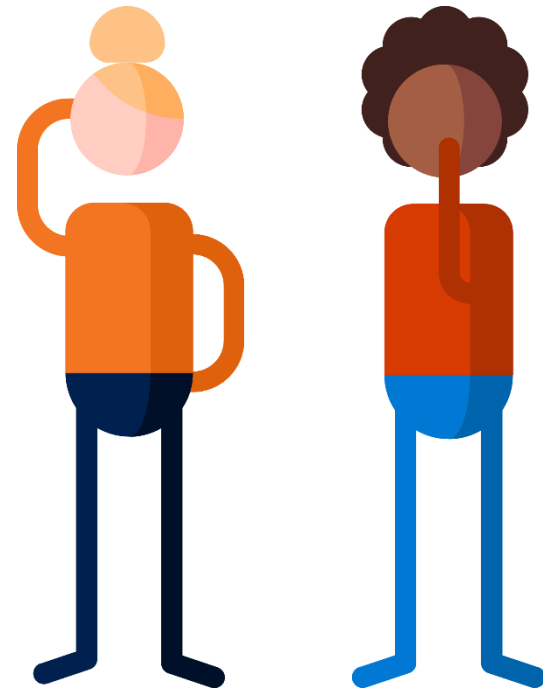


Reality:

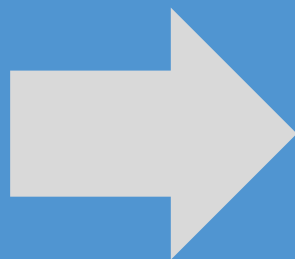


Which Tool do we pick ???

- a) SQL Server Integrations Services (SSIS)?
- b) Azure Data Factory (ADF)?
- c) Azure Databricks?
- d) All of the above?



It Depends



What does it
depend on?

Let's take
a look!

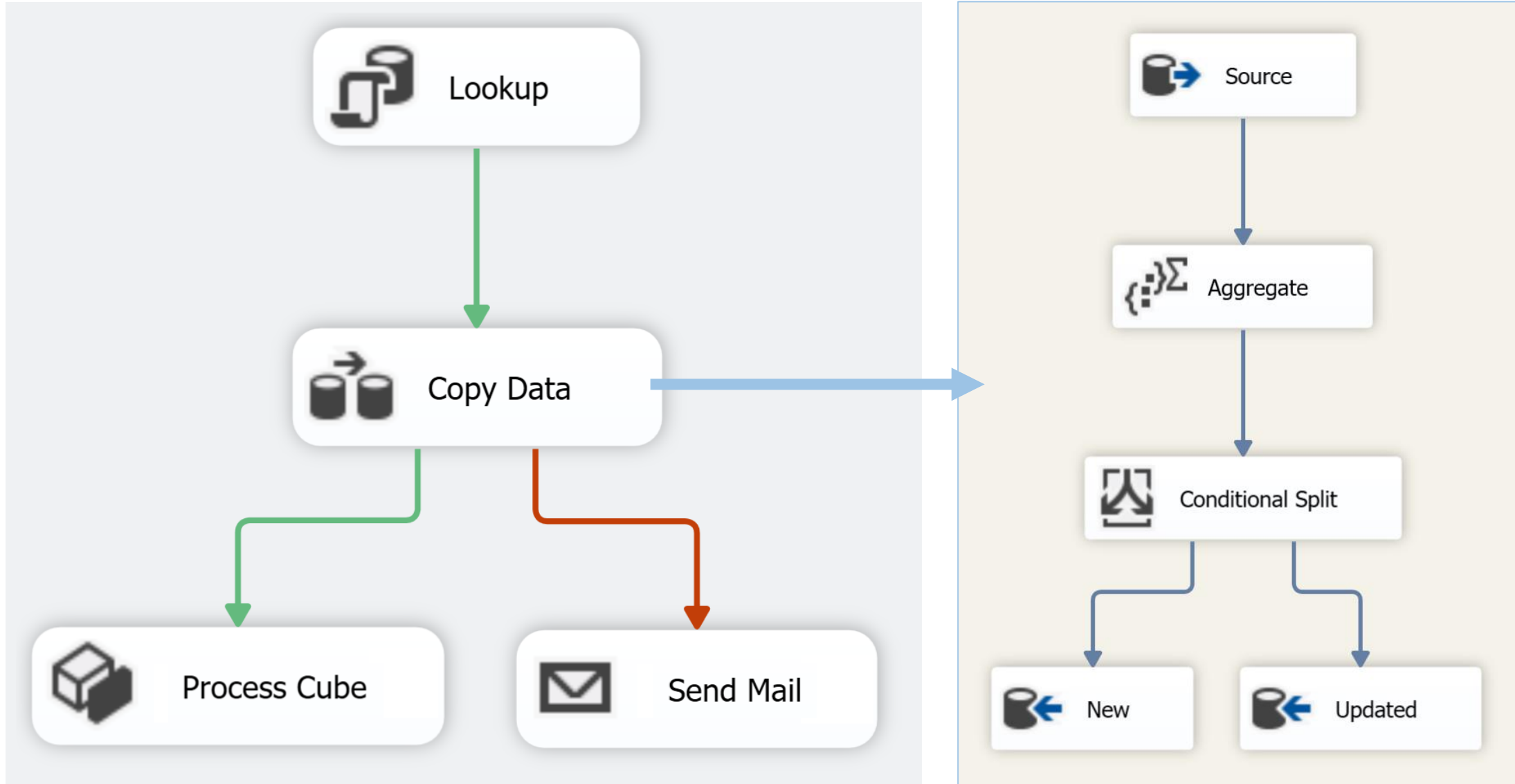


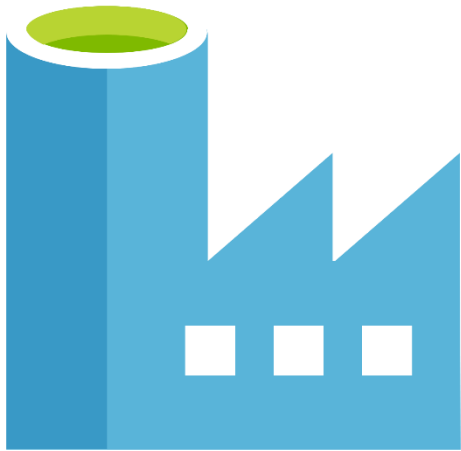
Data Integration

Extract, Transform, Load (ETL)

Hybrid

Sample Flow: SSIS



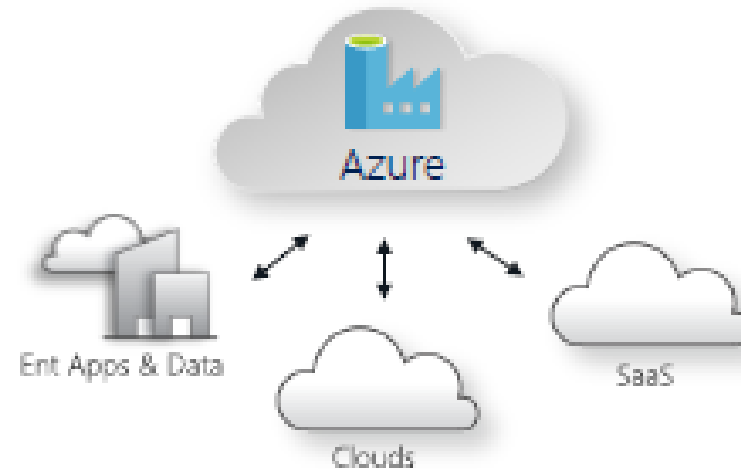


Data Movement & Orchestration

Extract, Load, Transform (ELT)

Hybrid

Data Integration Service: Serverless, Scalable, Hybrid



Hybrid Pipeline Model

Seamlessly span: on prem, Azure, other clouds & SaaS
Run on-demand, scheduled, data-availability or on event

SSIS Package Execution

Lift existing SQL Server ETL to Azure
Use existing tools (SSMS, SSDT)

Data Movement @Scale

Cloud & Hybrid w/ 80+ connectors provided
Up to 1 GB/s

Author & Monitor

Programmability w/ multi-language SDK
Visual Tools

Guided Experience to build data flows

Microsoft Azure

ansdf1 Data Factory Publish All Validate All Refresh Discard All ARM Template

Factory Resources

Filter Resources

Pipelines 13

Datasets 21

Data Flows 14

MovieDemoDataFlow

SplitTest

SelectTest

FilterTest

SortTest

UnionTest

NoOp

TaxiDemoFull

renaming

typeMatch

SourceSink

SourceSink2

formatTest

TaxiDemo

Connections

Triggers

TaxiDemoFull

Validate

TripData

Columns: 14 total

TripFare

Import data from trip_fare_full

JoinMatchedData

Inner join on TripData and TripFare

AggregateDayStats

Aggregating data by 'DayOfTheWeek' producing columns 'average_distance', 'average_passenger'

DayStatsSink

Export data to taxi_demo_day_stats

AggregateVendorStats

Aggregating data by 'VendorId' producing columns 'TotalPassengersServed', 'TotalTripTime'

VendorStatsSink

Export data to taxi_demo_vendor_stats_sink

TotalPaymentByPayment...

Export data to taxi_demo_payment_stats_sink

New Branch

Derived Column

Join

Conditional Split

Exists

Select

Aggregate

Filter

Sort

Union

Sink

Source Settings Define schema Inspect

Output stream name * TripData

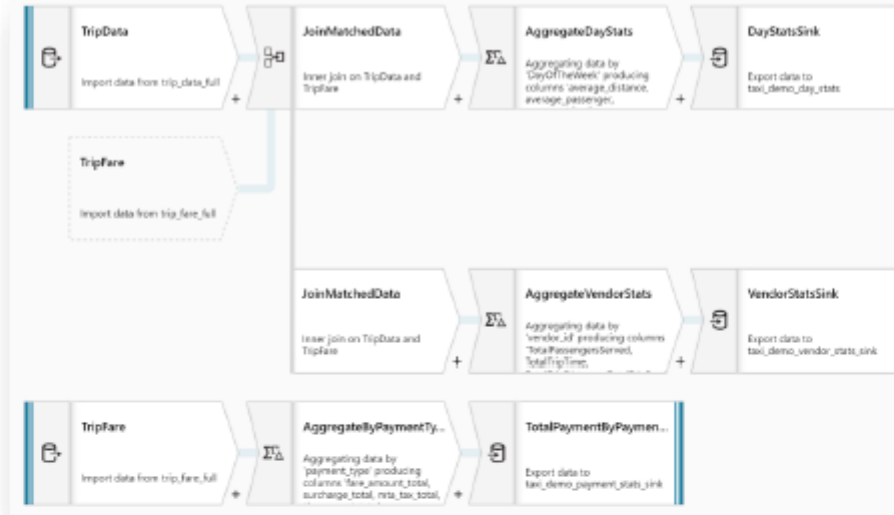
Source Dataset * trip_data_full Edit + New

Options ☐ Allow schema drift ①

Progress and Roadmap

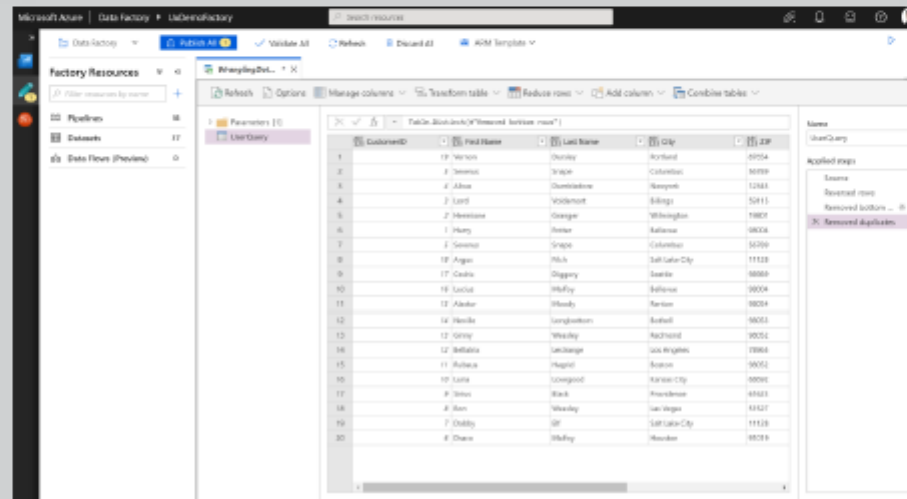


MAPPING DATAFLOW
Code-free data
transformation @scale



PUBLIC
PREVIEW

WRANGLING
DATAFLOW
Code-free data
preparation @scale



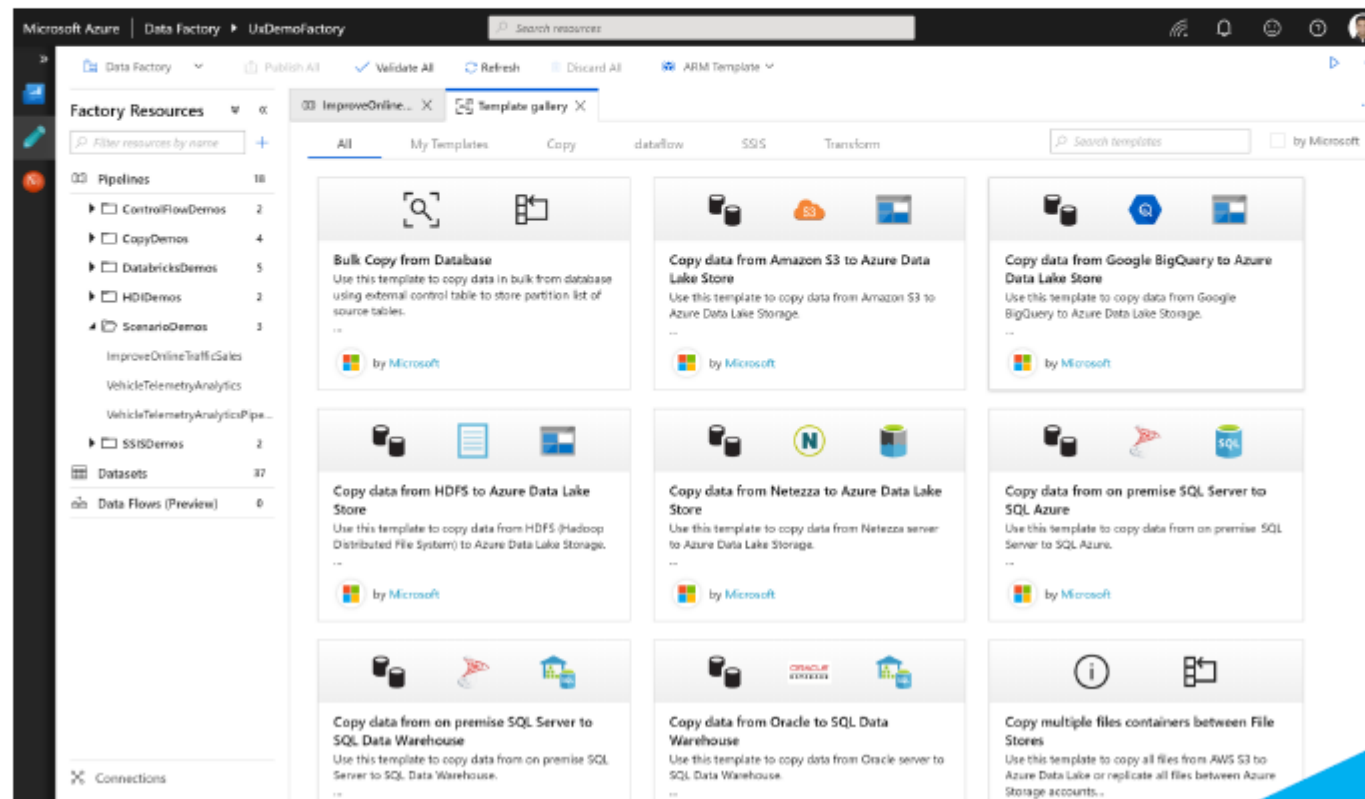
PRIVATE PREVIEW
MARCH 2019

Use Templates to quickly get started with ADF

Quickly get started with building data integration solutions

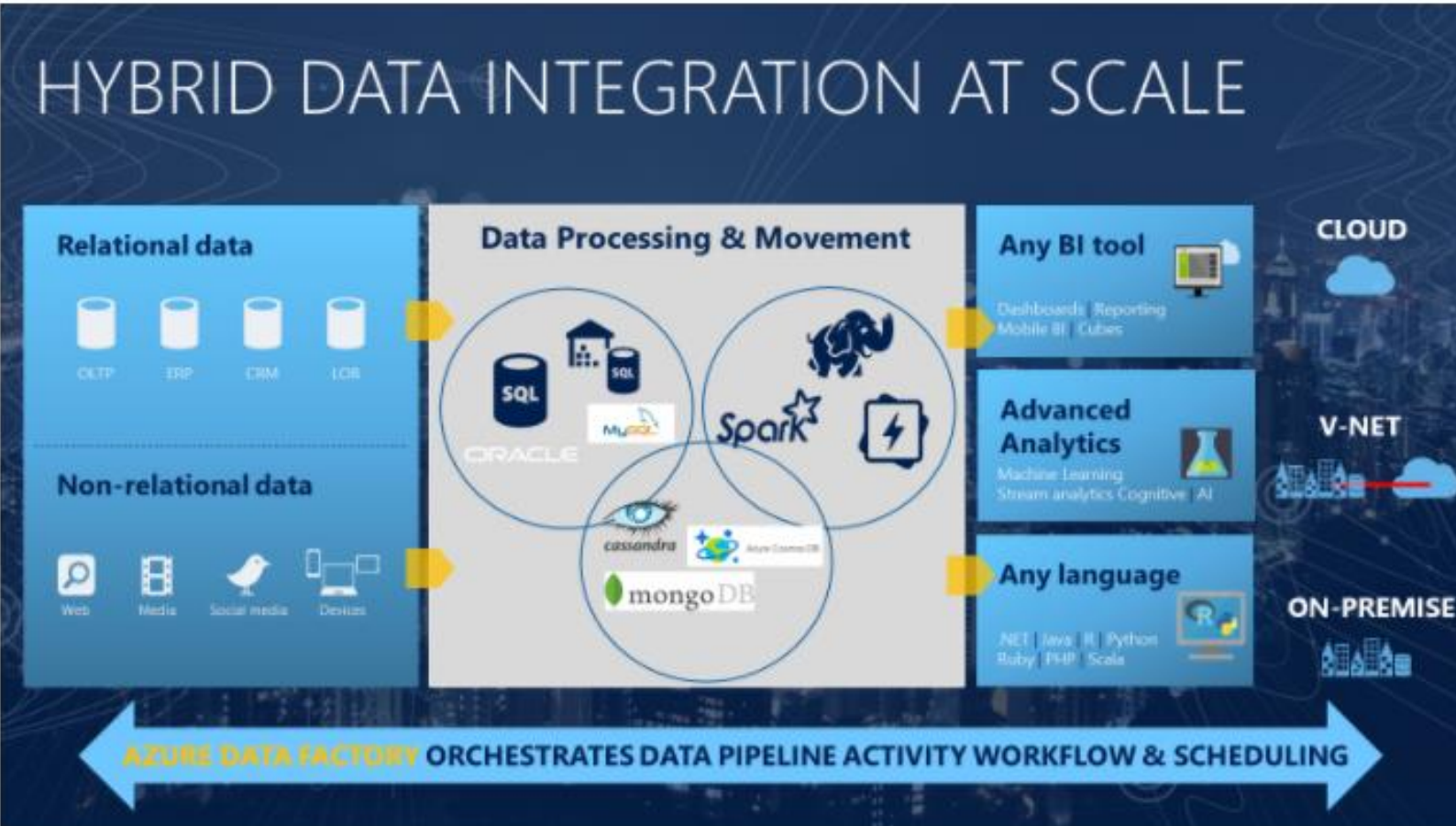
Avoid building same workflows repeatedly. Simply instantiate a template

Improve developer productivity along with reducing development time for repeat processes



JUST
RELEASED

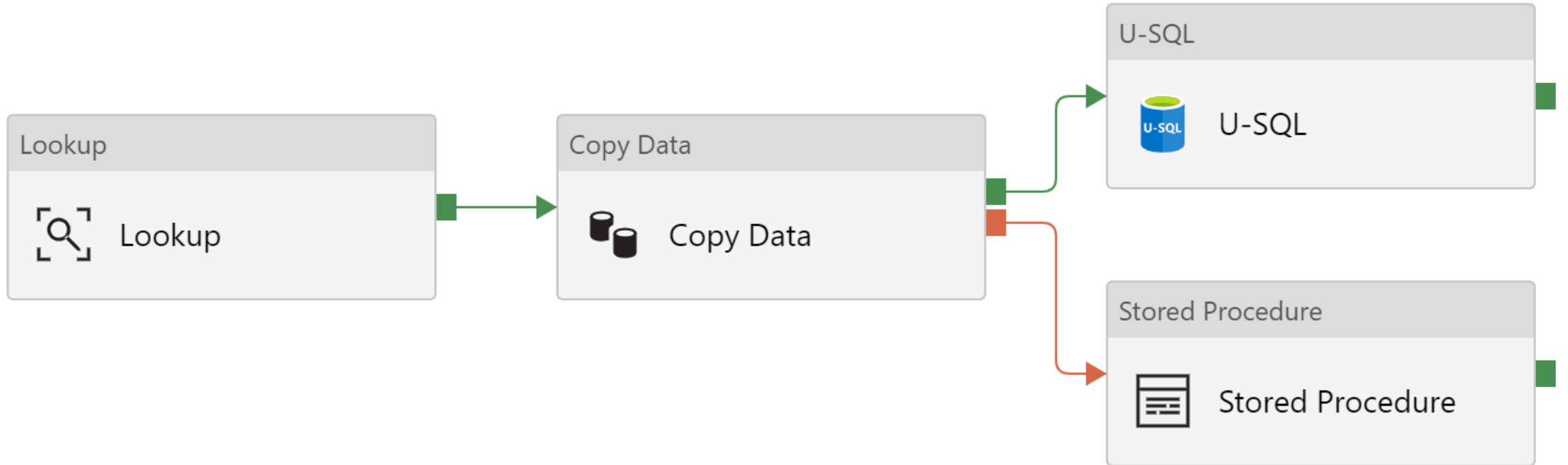
Azure DataFactory: How does it work



Top-level concepts:

1. Pipeline
2. Datasets
3. Linked services
4. Triggers
5. Control flow

Sample Flow: Datafactory





**Apache Spark + Databricks +
Enterprise Cloud = Azure Databricks**

Apache Spark-based Analytics Service

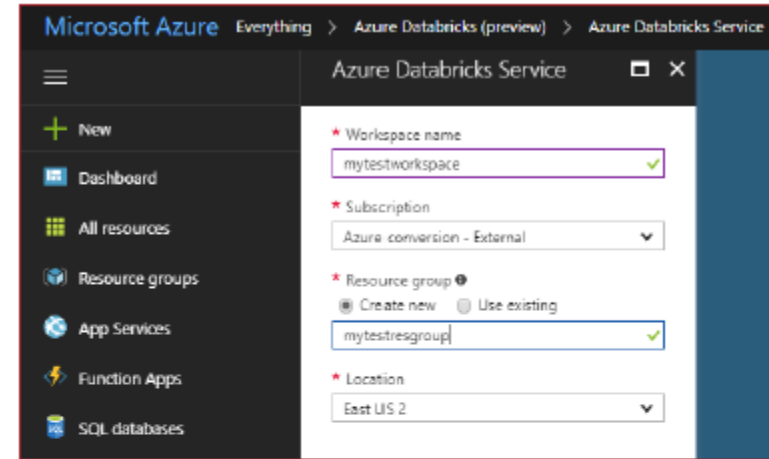
Collaborative Notebooks

Cloud

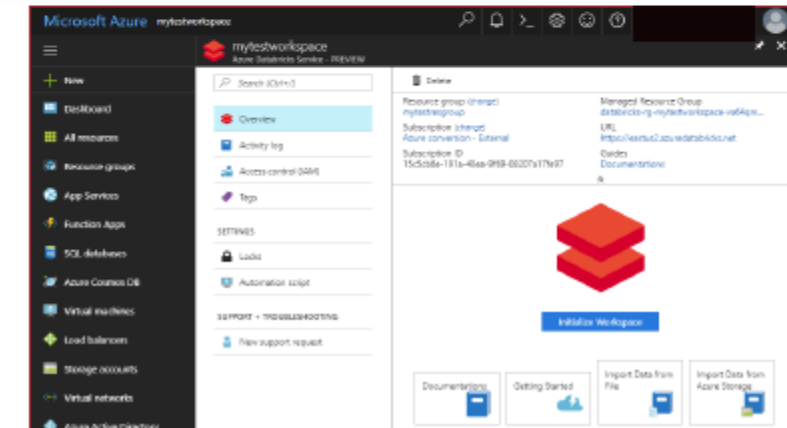


Provisioning Azure DataBricks Workspace

- Azure Databricks is provisioned directly from the Azure Portal like any other Azure service
 - In contrast, with other clouds, it has to be provisioned through the Databricks portal.
 - With Azure Databricks, the Azure Portal offers a unified portal to provision and administer Azure Databricks as well as other Azure services.
- Any Azure user with the appropriate subscription and authorization can provision Azure Databricks service*.
 - There is no need for a separate Databricks account

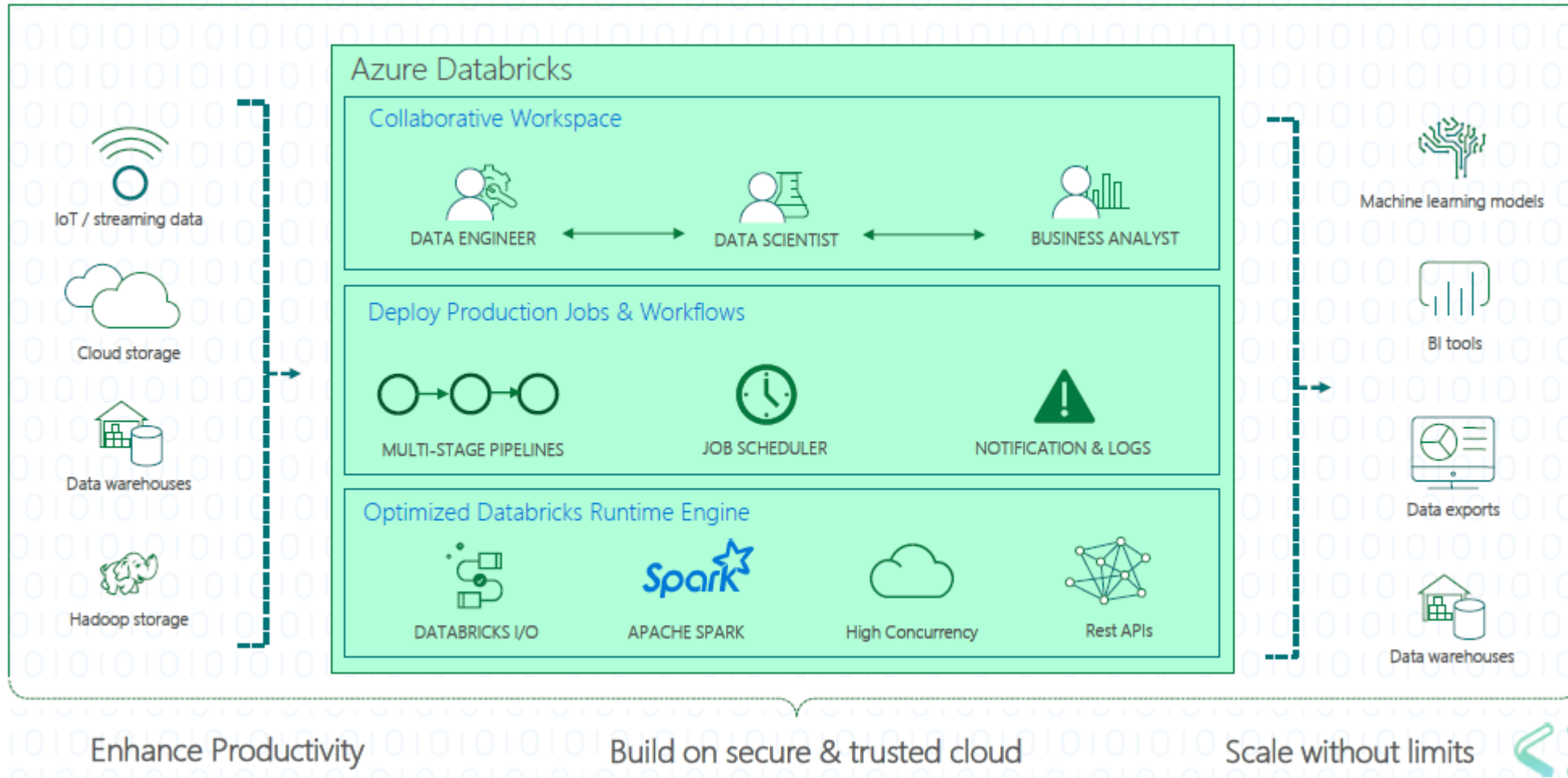


Provisioning the
Azure Databricks
Service



After provisioning
the is complete

Azure Databricks



Sample Flow: Databricks



Sample Apache Access Web Logs

```
> display(dbutils.fs.ls("/mnt/my-data/apache"))
```

path	name	size
dbfs:/mnt/my-data/apache/ex20150814.log	ex20150814.log	208693
dbfs:/mnt/my-data/apache/ex20150815.log	ex20150815.log	208693

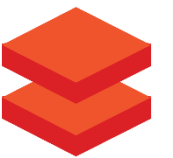
```
> myApacheLogs = sc.textFile("/mnt/my-data/apache")
myApacheLogs.take(10)
```

```
Out[11]:
[u'10.0.0.127 - 2696232 [14/Aug/2015:00:00:26 -0800] "GET /index.html HTTP/1.1" 304 428',
 u'10.0.0.104 - 2404465 [14/Aug/2015:00:01:14 -0800] "GET /Cascades/rss.xml HTTP/1.1" 304 514',
 u'10.0.0.108 - 2404465 [14/Aug/2015:00:04:21 -0800] "GET /Olympics/rss.xml HTTP/1.1" 200 499',
 u'10.0.0.213 - 2185662 [14/Aug/2015:00:05:15 -0800] "GET /Hurricane+Ridge/rss.xml HTTP/1.1" 200 288',
 u'10.0.0.203 - 2185662 [14/Aug/2015:00:05:17 -0800] "GET /index.html HTTP/1.1" 200 212',
 u'10.0.0.104 - 2696232 [14/Aug/2015:00:06:09 -0800] "GET /Cascades/rss.xml HTTP/1.1" 304 420',
 u'10.0.0.206 - 2576242 [14/Aug/2015:00:08:40 -0800] "GET /index.html HTTP/1.1" 304 343',
 u'10.0.0.213 - 2185662 [14/Aug/2015:00:09:07 -0800] "GET /Olympics/rss.xml HTTP/1.1" 304 323',
 u'10.0.0.212 - 2404465 [14/Aug/2015:00:10:29 -0800] "GET /index.html HTTP/1.1" 304 530',
 u'10.0.0.114 - 2575718 [14/Aug/2015:00:11:22 -0800] "GET /index.html HTTP/1.1" 304 341']
```

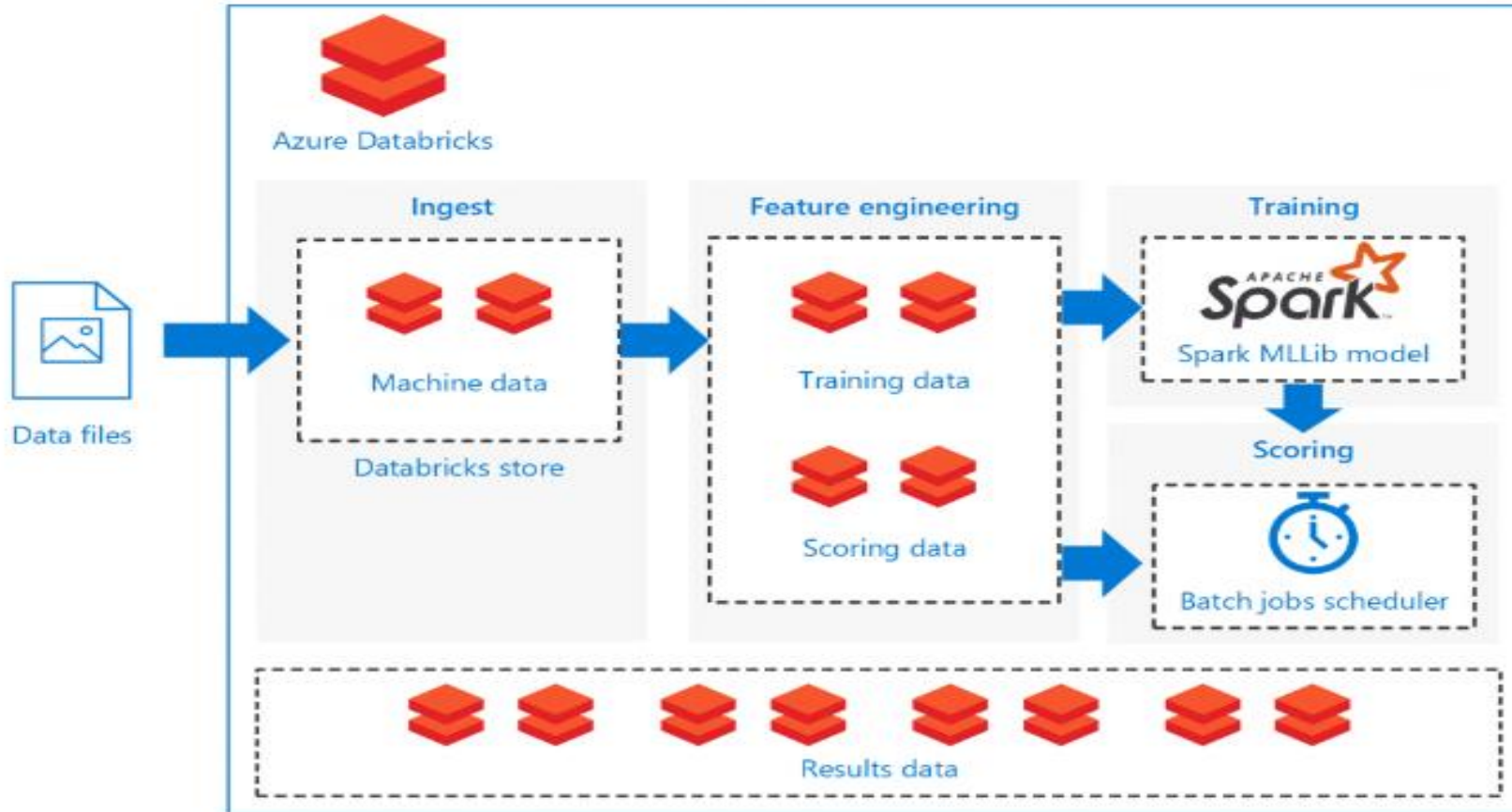
Sample Web Response Codes

```
> display(dbutils.fs.ls("/mnt/my-data/response"))
```

path	name	size
dbfs:/mnt/my-data/response/responsecodes.txt	responsecodes.txt	55



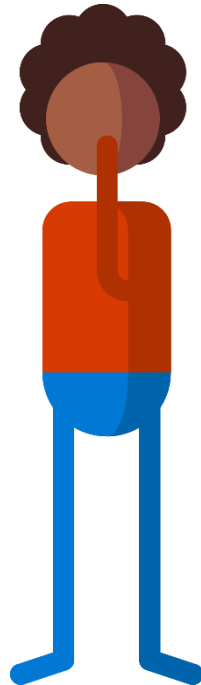
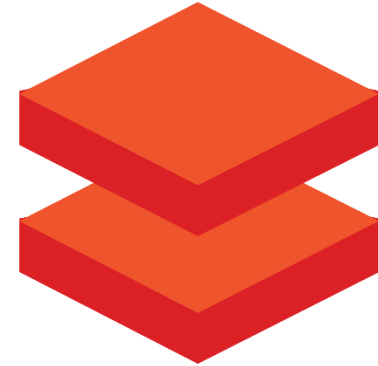
Regression model - Data Bricks



Important Components:

- ❖ Data files
- ❖ Ingestion
- ❖ Training pipeline
- ❖ Scoring pipeline
- ❖ Scheduler.

Selection Dilemma





Data

Volume
Velocity
Variety

Development

Tools
Interface
Languages

Usage

Platform
Pricing
Purpose

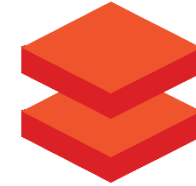
Volume



Medium



High



High

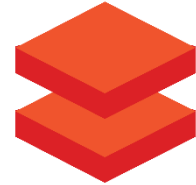
Velocity



Batch



Batch
Streaming



Batch Streaming
Real -Time

Variety



Structured



Structured
Unstructured



Structured
Unstructured

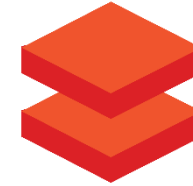
Tools



SSDT



Browser



Browser

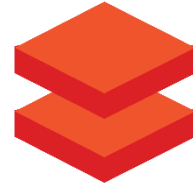
Interface



Drag and
Drop



Drag and
Drop
Code



Code

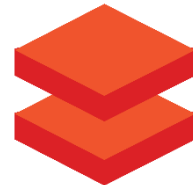
Languages



VB / C#
Biml



.NET
Python
PowerShell



SQL
Python
R
Scala

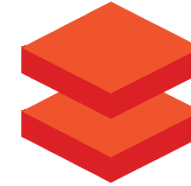
Platform



On-Premises
Own Hardware
Scale Out



Hybrid
Managed
Scale Up



Cloud
Managed
Autoscale

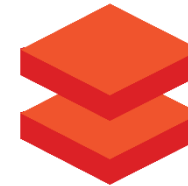
Pricing



License



Pay as you
go



Pay as you
go

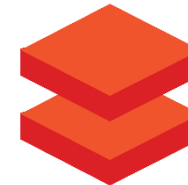
Purpose



Integration
Transformation
ETL



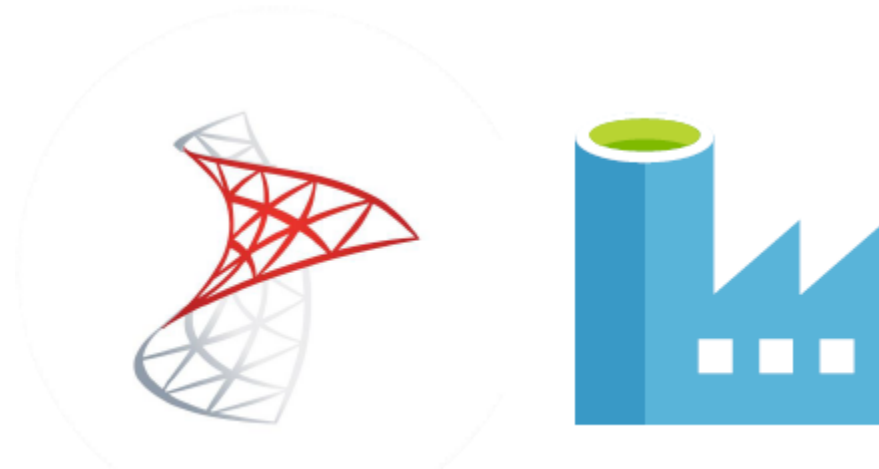
Movement
Orchestration
ETL / ELT



Preparation
Collaboration
AI / ML



Lift and Shift SSIS Execute SSIS in ADF



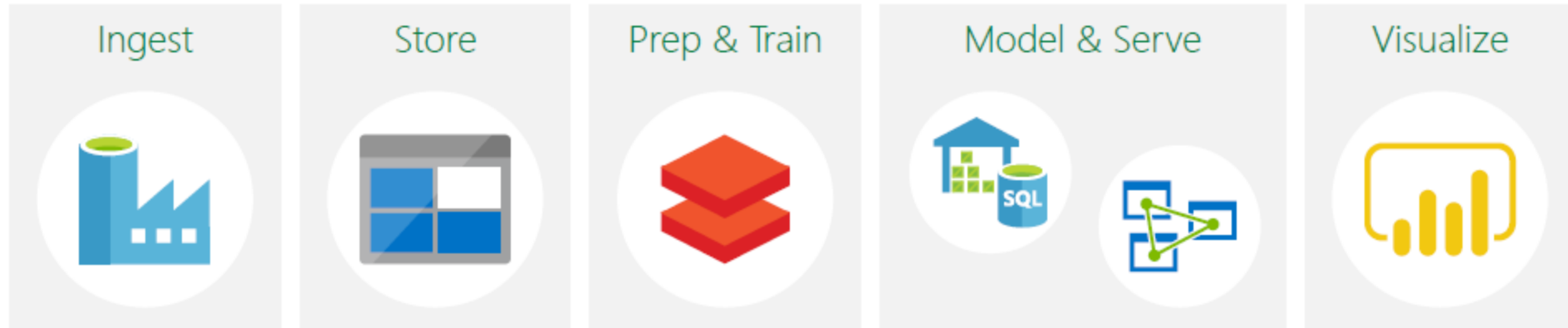


Databricks Activities
Data Flows run on Databricks



Modern Data Warehouse Implementation using Native Azure Components

Modern Data Warehousing



Demo Use Case – Azure Datafactory and Databricks

- Predicting the price of diamond.
- Linear Regression is implemented with the properties of diamond being the independent variables.
- Some of the properties used to predict the price are Carat, Color, Clarity and Depth.
- The Dataset contains 53,940 rows and 10 columns.

Demo

Use Case Snapshots



azure databricks - Google x | Access keys - Microsoft x | Clusters - Databricks x | Azure Data Factory x | Run a Databricks Notebook x | Transform data by using x | +

https://adf.azure.com/authoring/pipeline/pipeline2?factory=%2Fsubscriptions%2F1f7a5280-7f8e-454a-b5bf-9440c6c7483a%2FresourceGroups%2Fdemo1...

Microsoft Azure | Data Factory | data88

Search resources

» Data Factory | Publish All | Validate All | Refresh | Discard All | ARM Template

Factory Resources

Filter resources by name

Pipelines 2

pipeline1

pipeline2

Datasets 3

AzureBlob1

inputdataset

outputdataset

Template gallery x | Connections x | AzureBlob1 x | pipeline2 x

Activities

Search Activities

Batch Service

Custom

Databricks

Notebook

Jar

Python

Move & Transform

Lookup

Lookup1

Notebook

Notebook1

Save as template | Validate | Debug | Add trigger

General | Settings | User Properties

Name * Lookup1

Use Case Snapshots



azure databricks - Google x | Access keys - Microsoft x | Clusters - Databricks x | Azure Data Factory x | Run a Databricks Noteb x | Transform data by using x | +

https://adf.azure.com/authoring/pipeline/pipeline2?factory=%2Fsubscriptions%2F1f7a5280-7f8e-454a-b5bf-9440c6c7483a%2FresourceGroups%2Fdemo1...

Microsoft Azure | Data Factory | data88

Search resources

» Data Factory | Publish All | Validate All | Refresh | Discard All | ARM Template

Factory Resources

Filter resources by name

Pipelines 2

pipeline1

pipeline2

Datasets 3

AzureBlob1

inputdataset

outputdataset

Activities

Search Activities

Batch Service

Custom

Databricks

Notebook

Jar

Python

Move & Transform

Save as template | Validate | Debug | Add trigger

Lookup

Lookup1

Notebook

Notebook1

General | Settings | User Properties

Name * Lookup1



© 2016 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.