Python Pipeline Primer

ETL in Azure

Simon Whiteley | Adatis

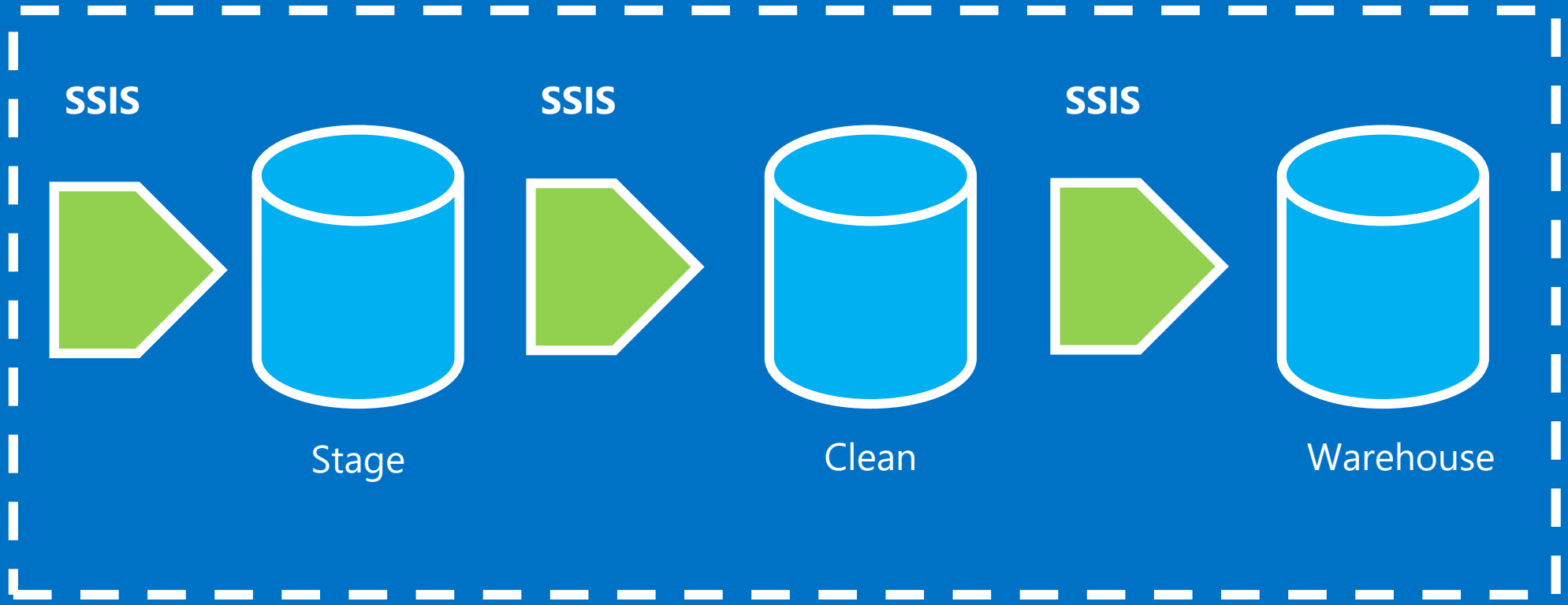
10/10/2018

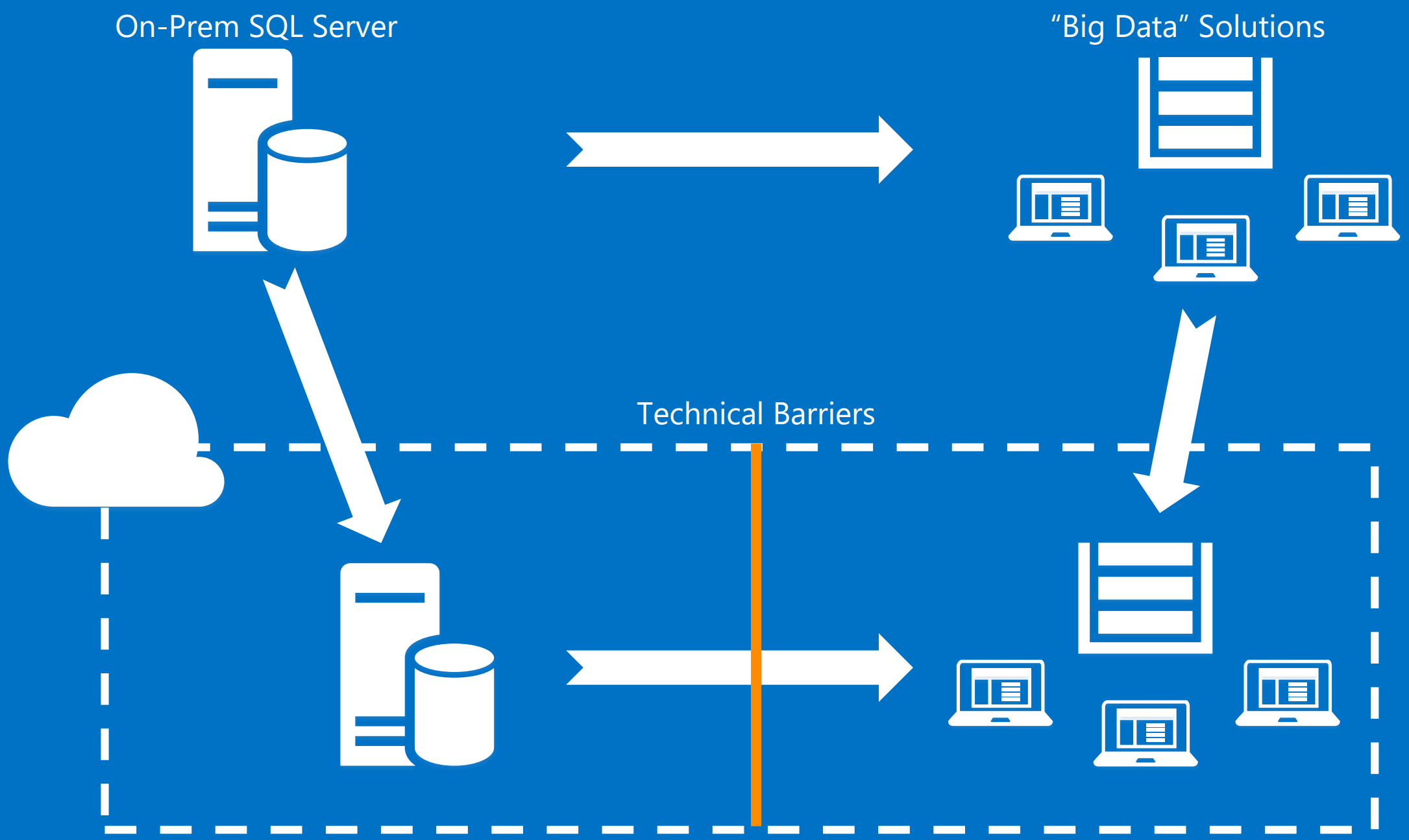
Agenda

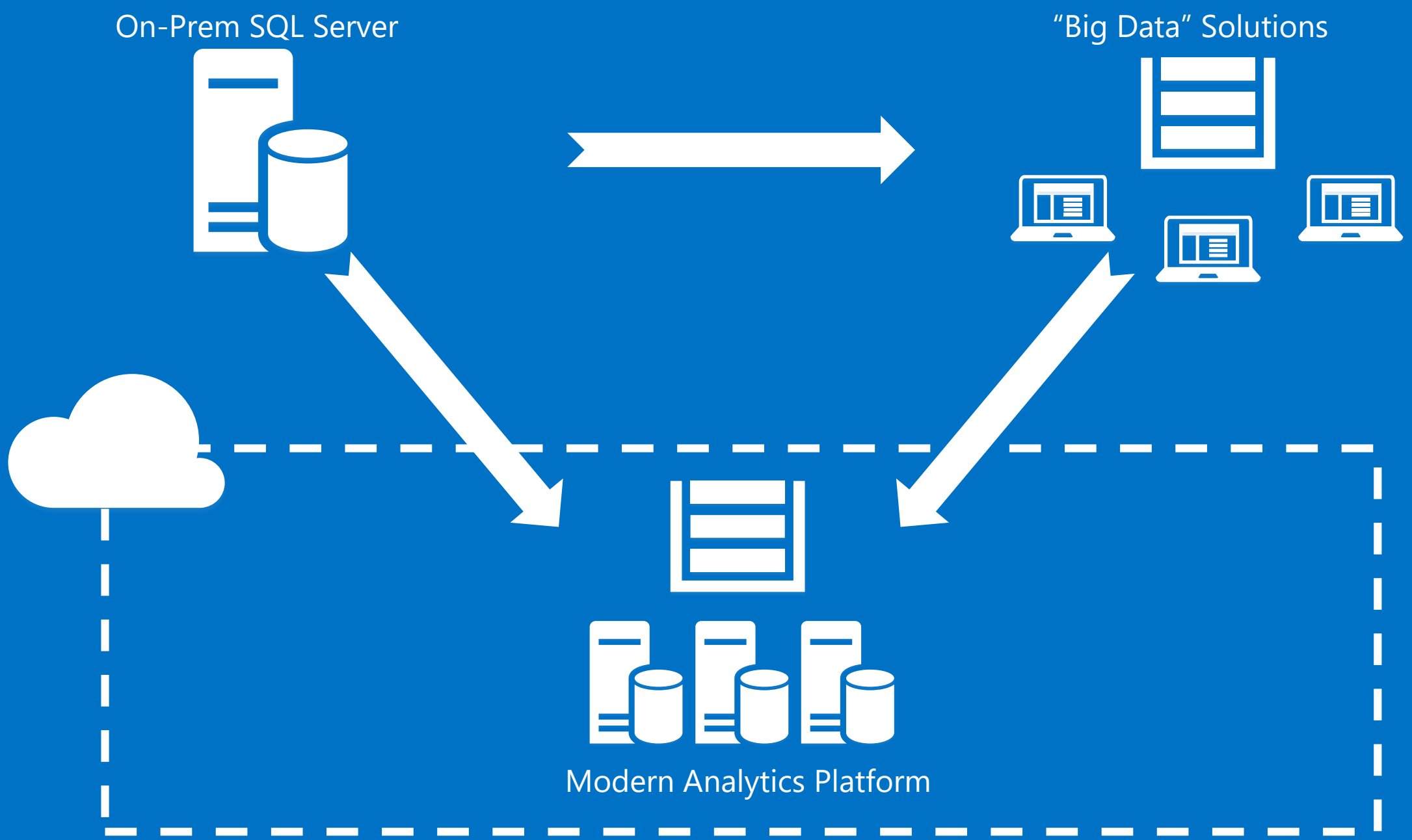
- Why Big Data processing is needed in the cloud
- What is Databricks?
- Patterns & Implementation
- Orchestration
- Data Factory Dataflows

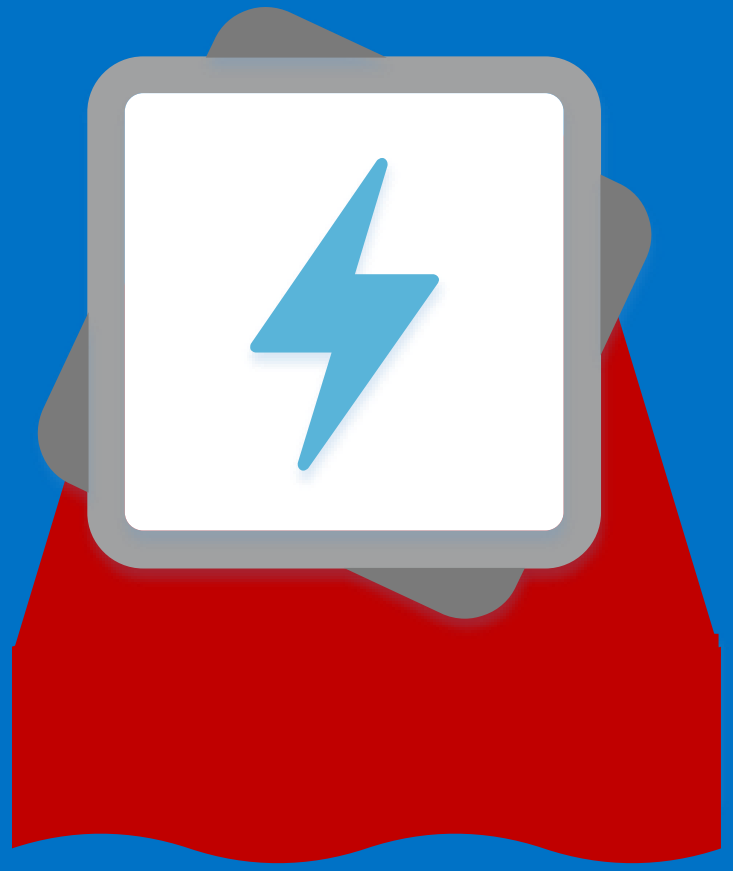
WBDPINITC

One-Box SQL BI Architecture

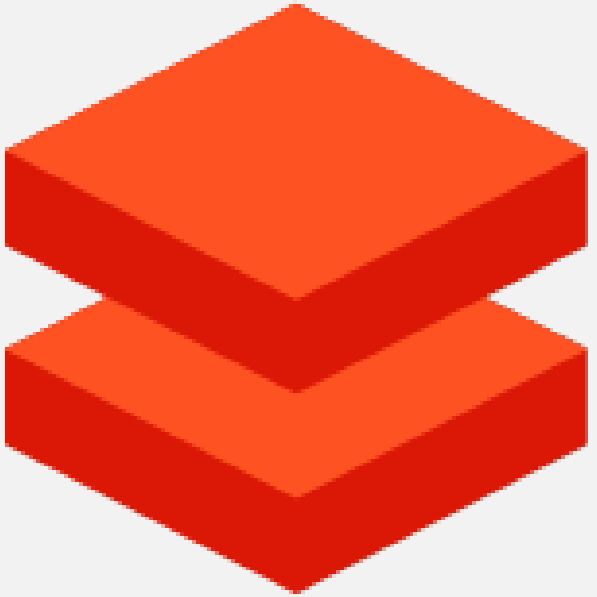



On-Prem SQL Server
"Big Data" Solutions
Technical Barriers


On-Prem SQL Server
"Big Data" Solutions
Modern Analytics Platform



DATA LAKE ANALYTICS



Azure
Databricks

adatis

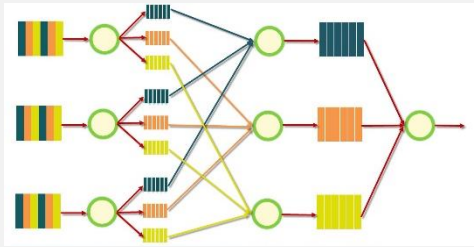
Databricks?

Google File System Papers Released

2003

Google MapReduce Papers

2004



2006

Apache Hadoop project created

Matei Zaharia starts Spark project

2012

Project donated to Apache Foundation

2013

Databricks founded by Matei

2013

2016

It’s new to Azure, not to everyone else!

Microsoft Azure

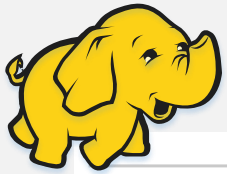
2018

adatis

So What?

- Most up to date Spark optimisations
- Doesn’t need specialist hardware
- Quicker than traditional MapReduce
- Cluster Management, Notebooks, Jobs…

adatis



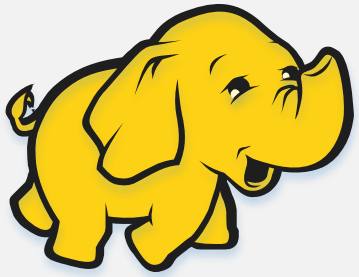
HDInsight

INSTANCE	CPU	RAM	OS	HDINSIGHT PRICE	TOTAL PRICE++
D3 v2	4	14 GB	£0.171/hour	£0.05/hour	£0.22/hour

Databricks

INSTANCE	vCPU	RAM	DBU COUNT	LINUX VM PRICE	DBU PRICE	PAY AS YOU GO TOTAL PRICE
D3 v2	4	14.00 GiB	0.75	£0.203/hour	£0.308/hour	£0.511/hour

adatis



Open Source	Open Source	Proprietary
20 min provisioning	5 min provisioning	1 min provisioning
Integrates Well	Integrates Well	Integrates Poorly
Secure	Secure	Secure
Hadoop, Spark, Kafka, Hbase, HIVE, Storm…	Spark (Python/Scala/R)	U-SQL
Slow Release Cycle	Fast Release Cycle	Slow Release Cycle

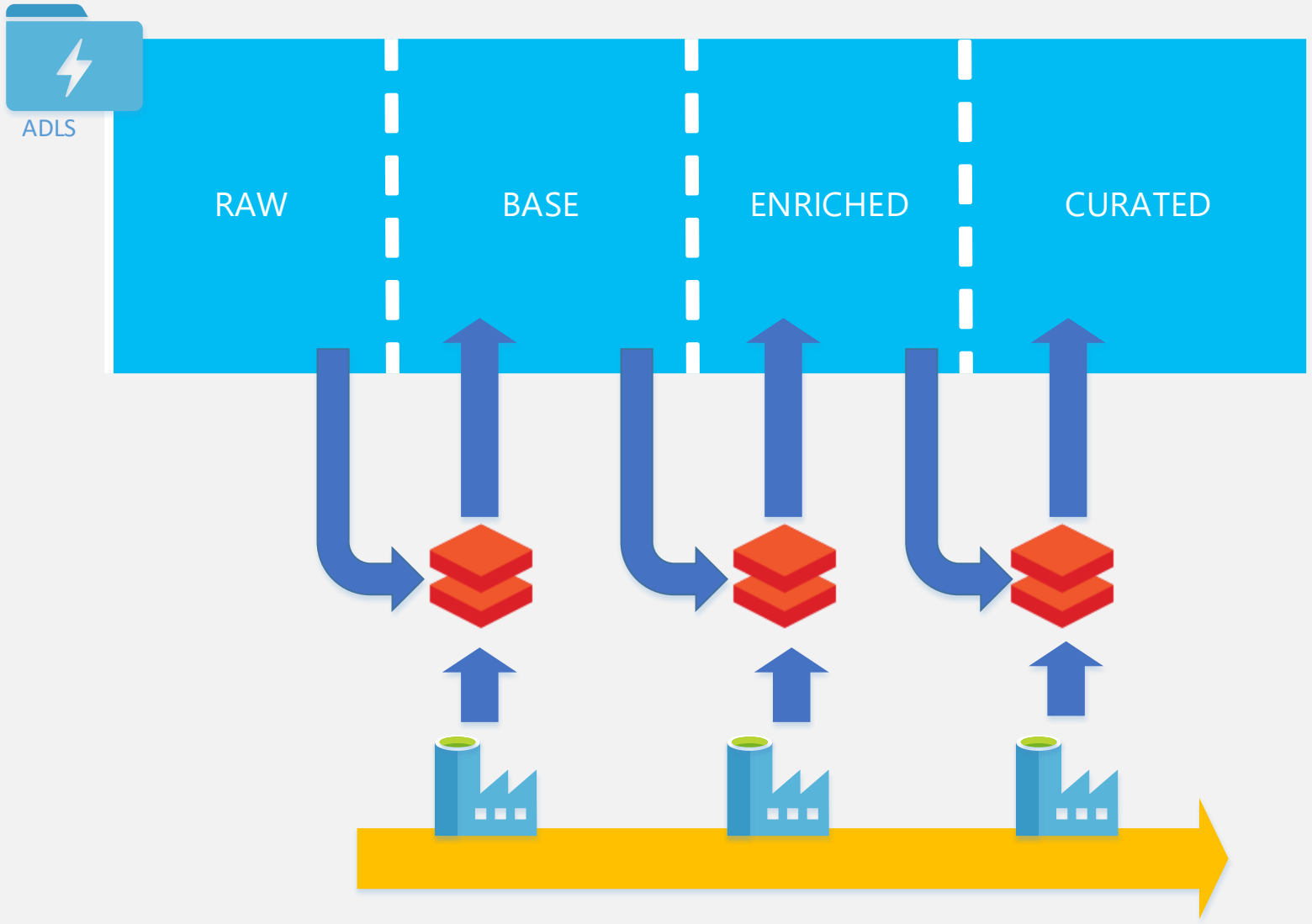
adatis

Data Scientists

Data Engineers

Data Analysts

adatis



adatis

Data Science VM Pool

Exploration & Development

Published Models

Complexity & Sophistication

Azure ML Studio

Azure ML Services

Docker

HPC

Model Hosting & APIs

Cognitive APIs

Model Services

Microsoft CNTK

Batch Scoring & Collaborative Experimentation

Data Processing & Interactive Querying

Published Data Sets

ADLS

DataBricks

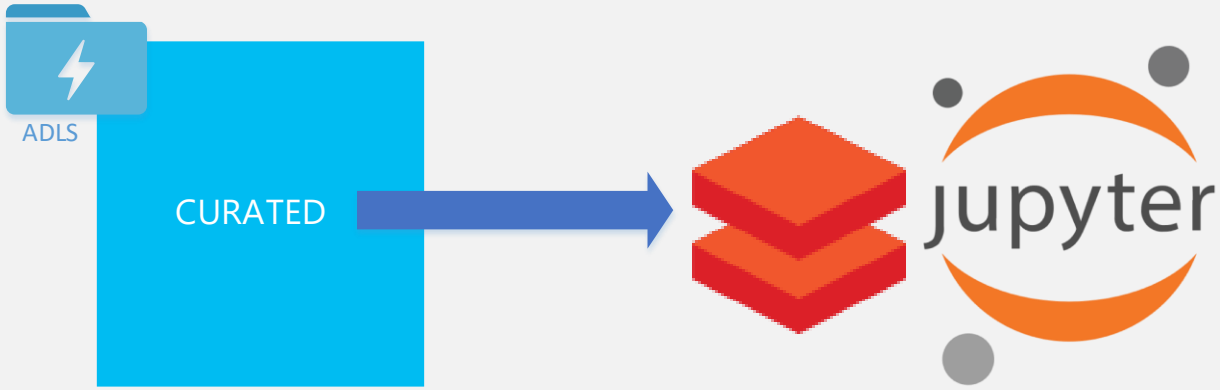
LABORATORY | EXHIBIT | MODELS

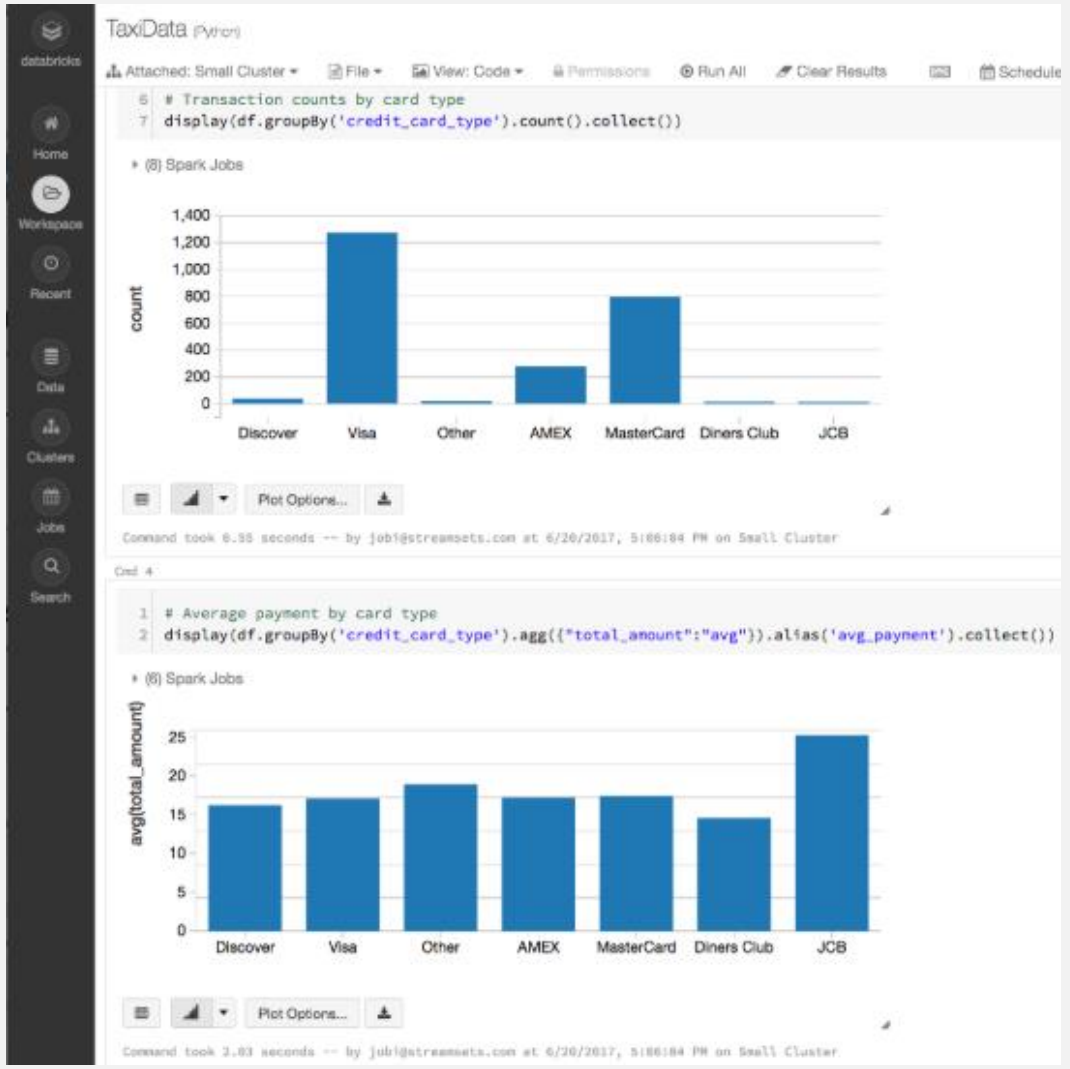
Ad-Hoc Data Collection &Transformation

RAW | BASE | ENRICHED | CURATED

DataLake Data Science EcoSystem

adatis


ADLS
CURATED
jupyter

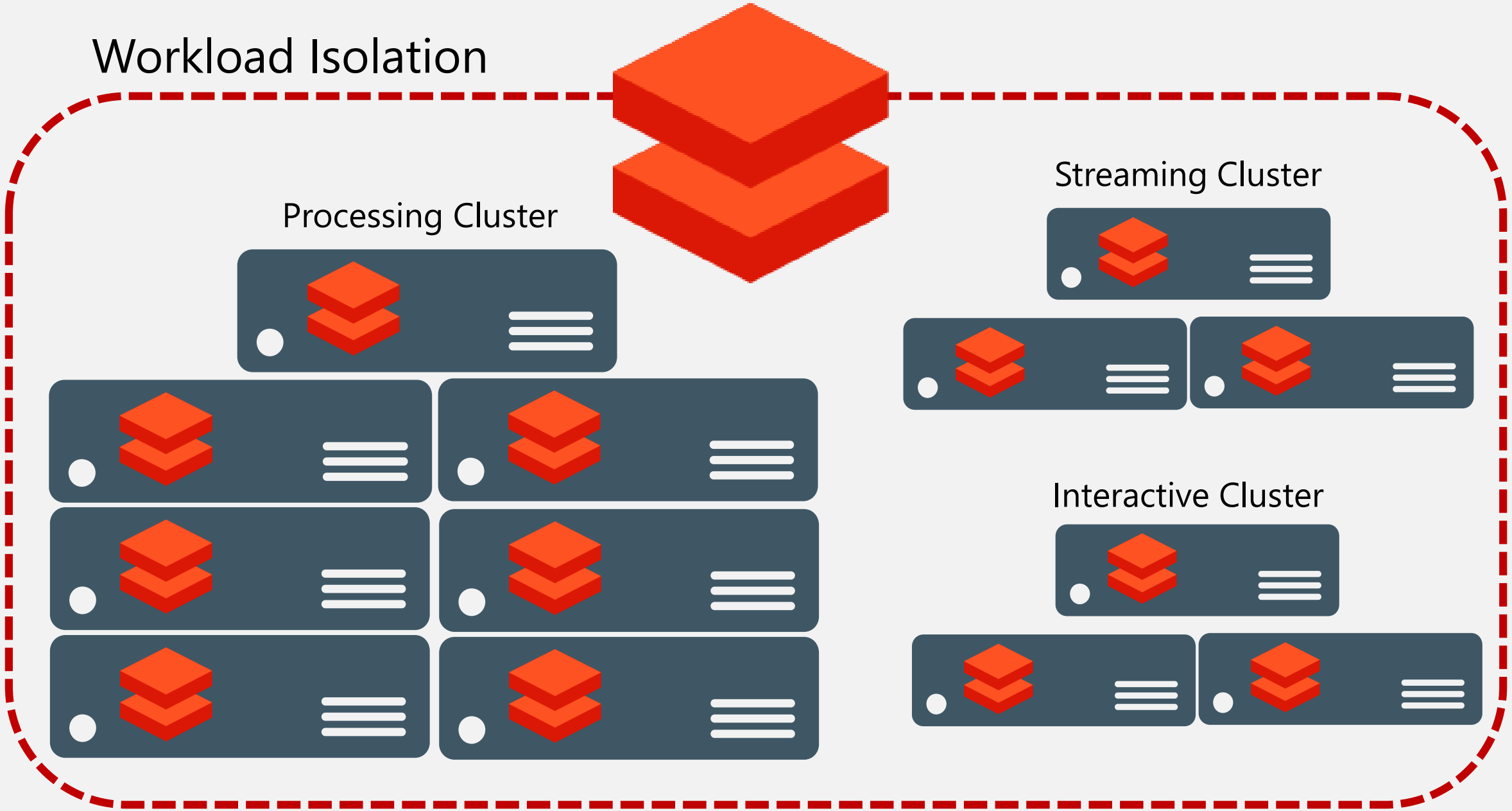

databricks
Home
Workspace
Recent
Data
Clusters
Jobs
Search
TaxiData (Python)
Attached: Small Cluster
File
View: Code
Permissions
Run All
Clear Results
Schedule
# Transaction counts by card type
display(df.groupBy('credit_card_type').count().collect())
(8) Spark Jobs
count
1,400
1,200
1,000
800
600
400
200
0
Discover
Visa
Other
AMEX
MasterCard
Diners Club
JCB
Plot Options...
# Average payment by card type
display(df.groupBy('credit_card_type').agg({"total_amount":"avg"}).alias('avg_payment').collect())
(8) Spark Jobs
avg(total_amount)
25
20
15
10
5
0
Discover
Visa
Other
AMEX
MasterCard
Diners Club
JCB
Plot Options...

adatis

Databricks Basics

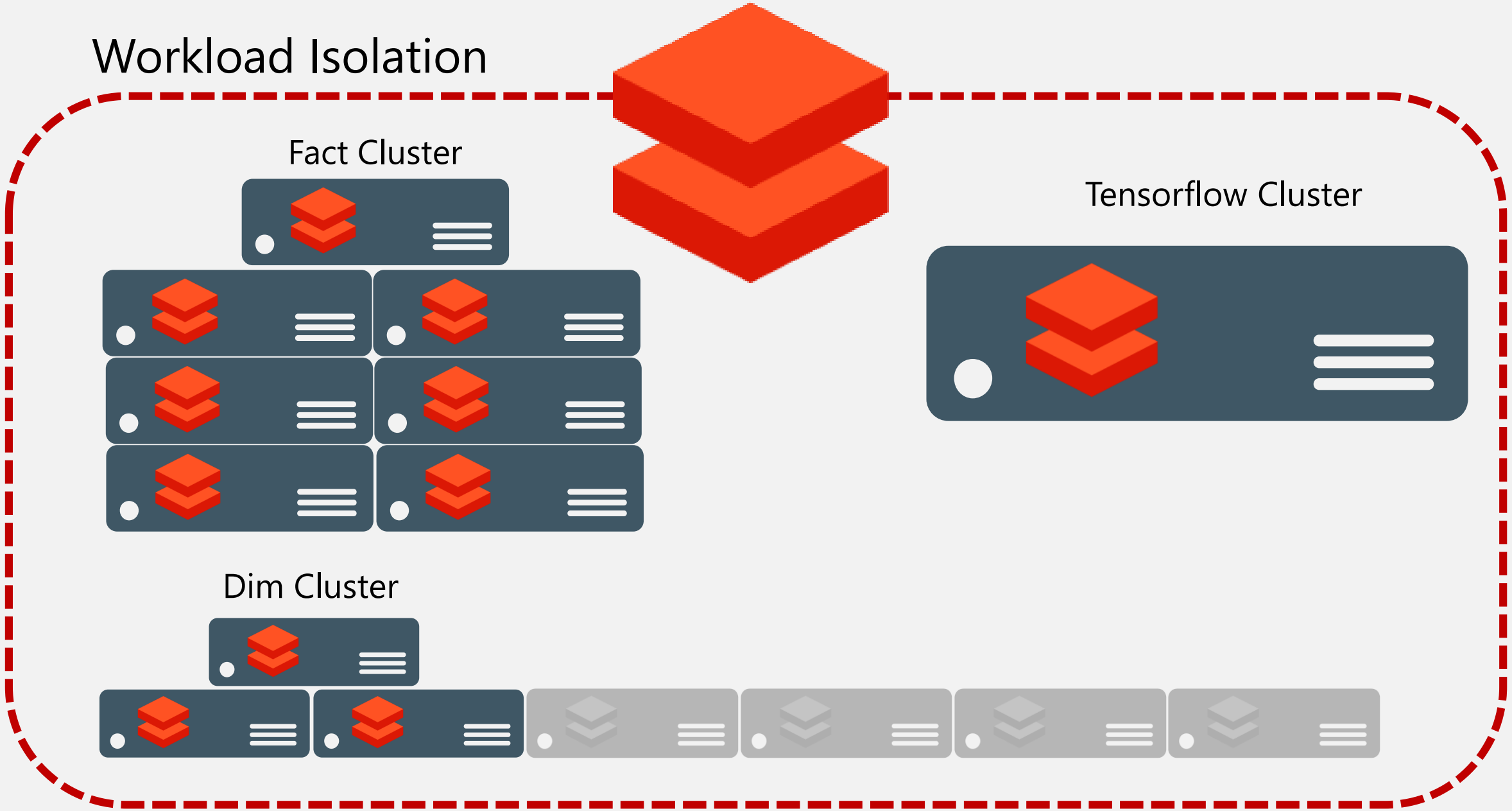
Patterns & Implementation

Workload Isolation



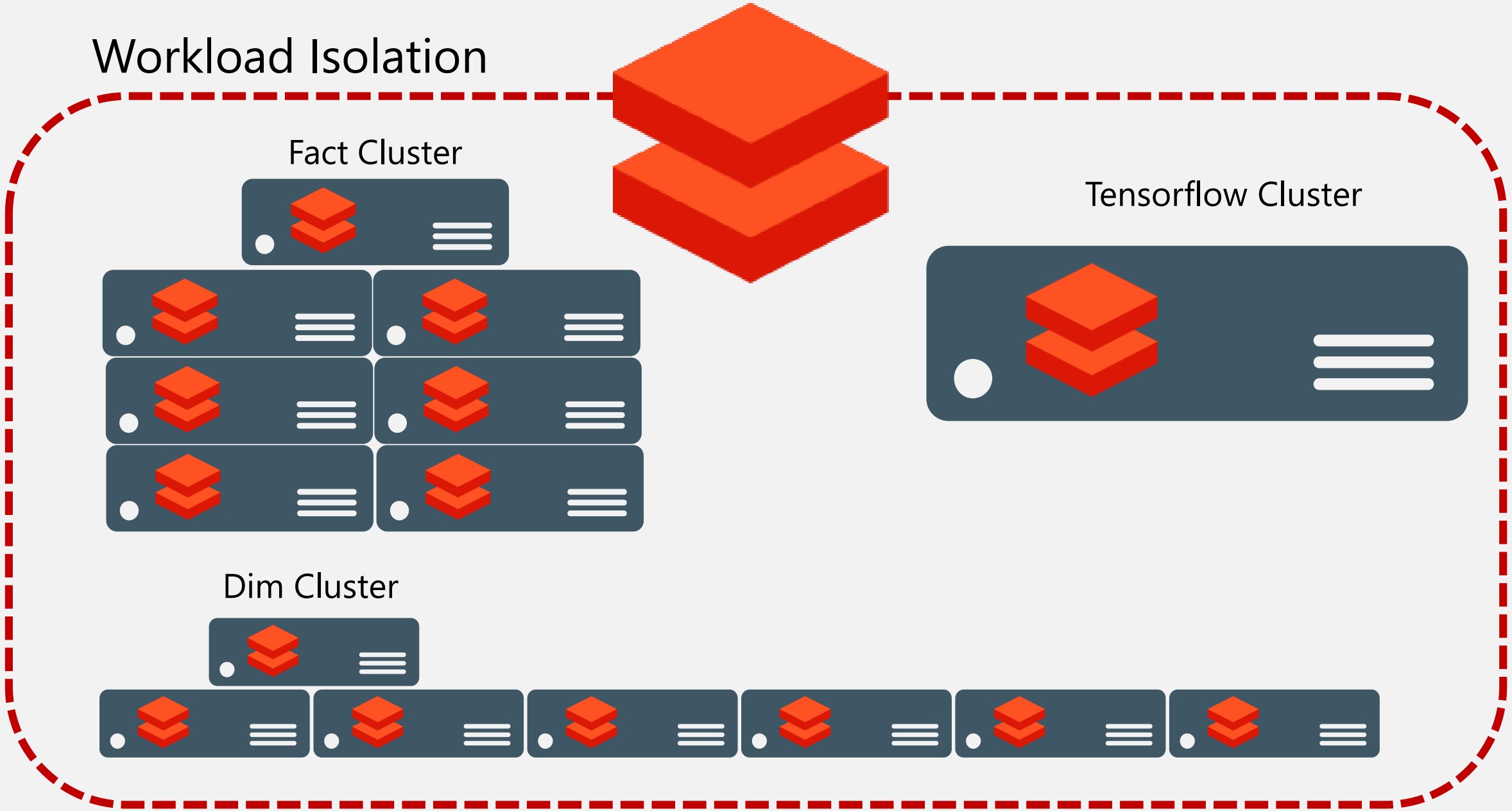
adatis

Workload Isolation



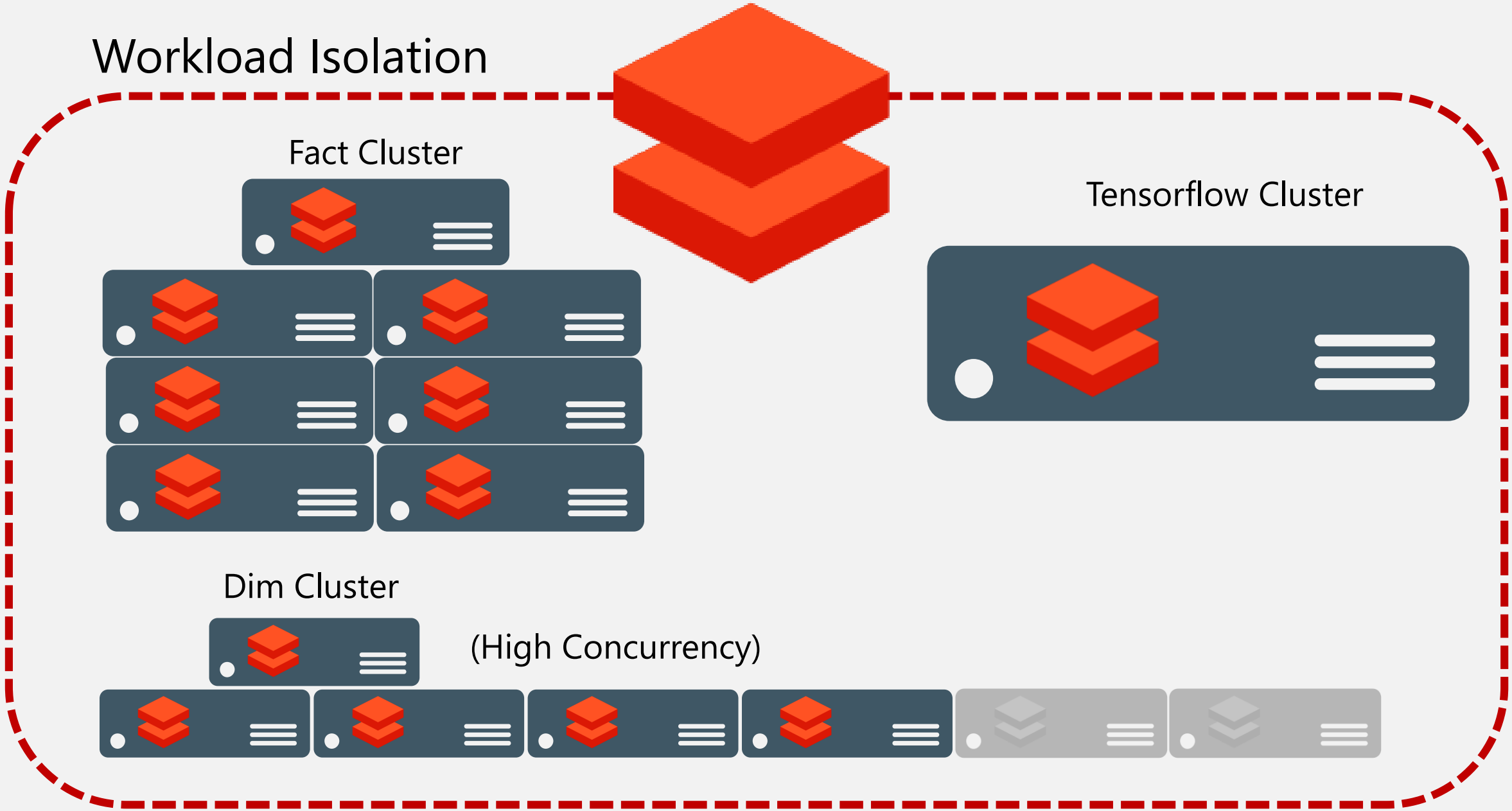
adatis

Workload Isolation

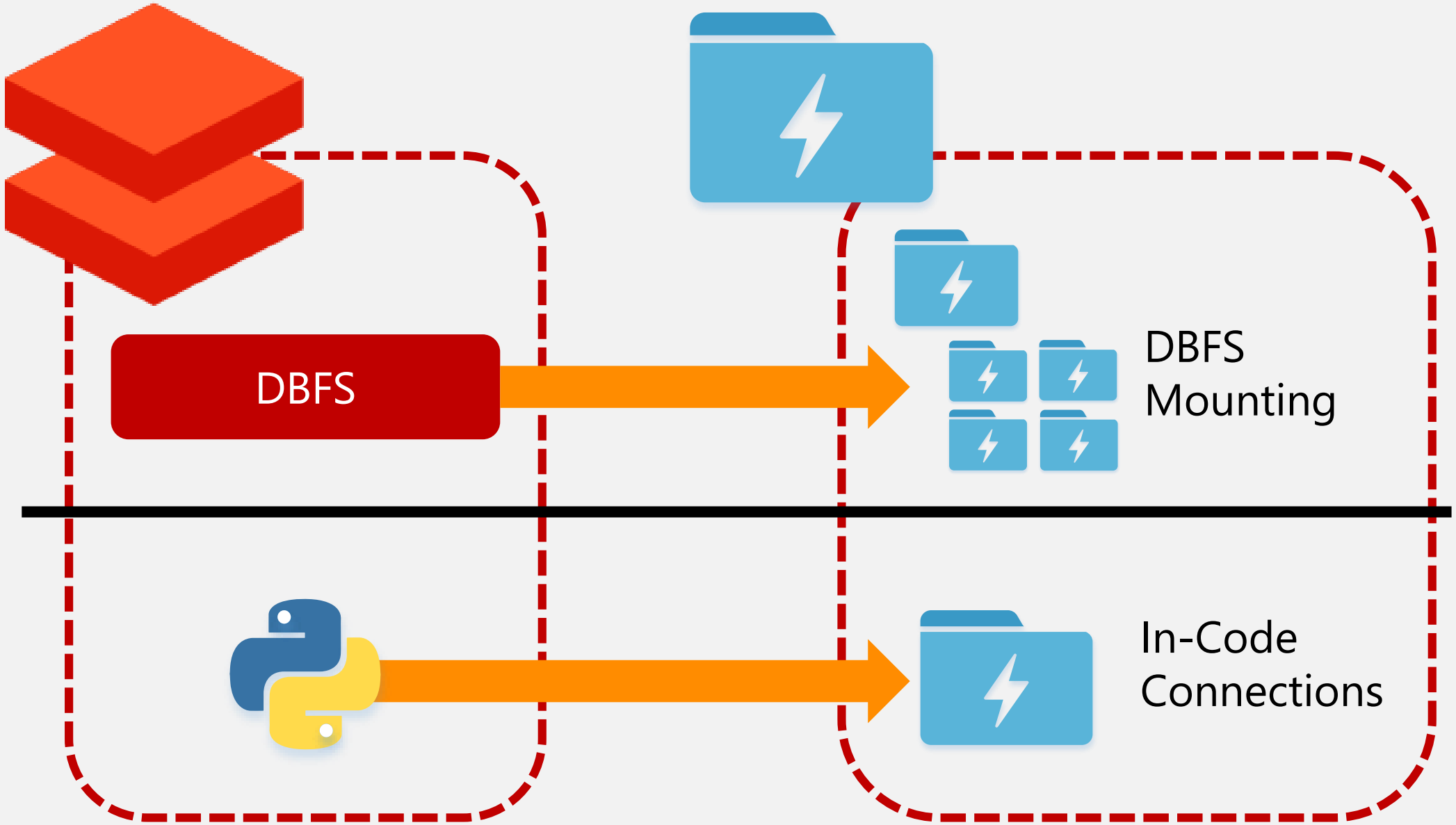


adatis

Workload Isolation

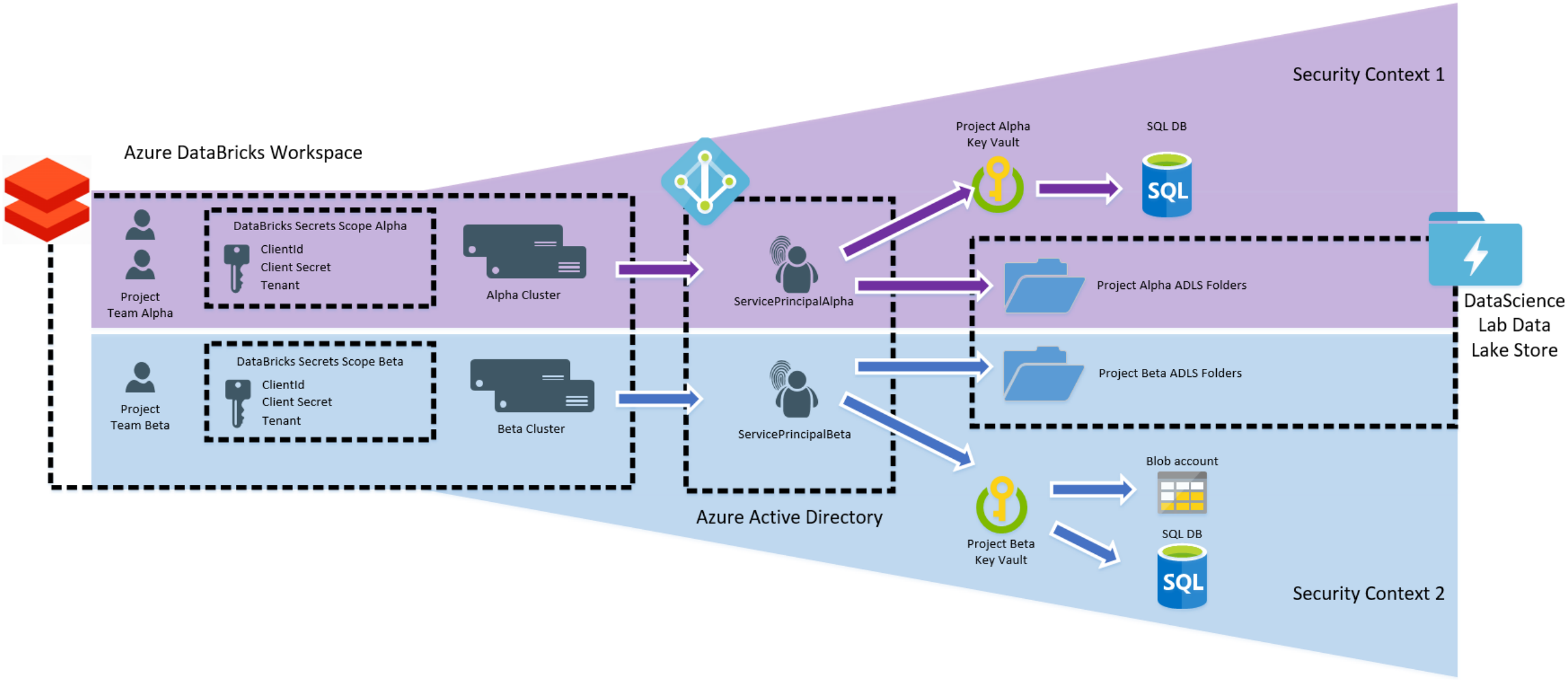


adatis



adatis

Secrets


Azure DataBricks Workspace
Project
Team Alpha
DataBricks Secrets Scope Alpha
ClientId
Client Secret
Tenant
Alpha Cluster
Project
Team Beta
DataBricks Secrets Scope Beta
ClientId
Client Secret
Tenant
Beta Cluster
Azure Active Directory
ServicePrincipalAlpha
ServicePrincipalBeta
Security Context 1
Project Alpha
Key Vault
SQL DB
SQL
Project Alpha ADLS Folders
Project Beta ADLS Folders
DataScience
Lab Data
Lake Store
Blob account
Project Beta
Key Vault
SQL DB
SQL
Security Context 2

adatis

Standard Clusters Vs Serverless pools

Serverless pools / High Concurrency

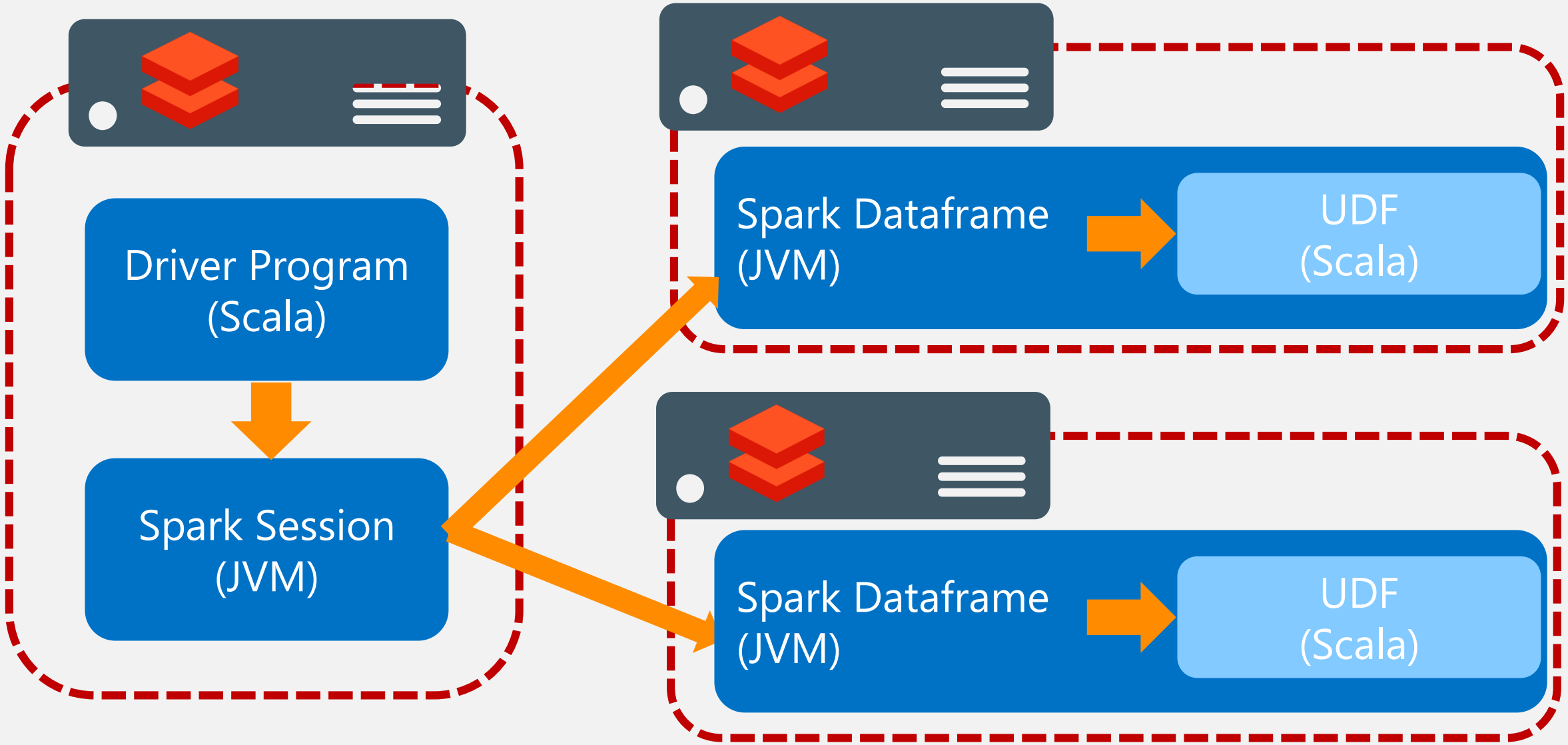
- Use SQL, Python, or R
- Want Azure Databricks to manage worker selection

Standard cluster

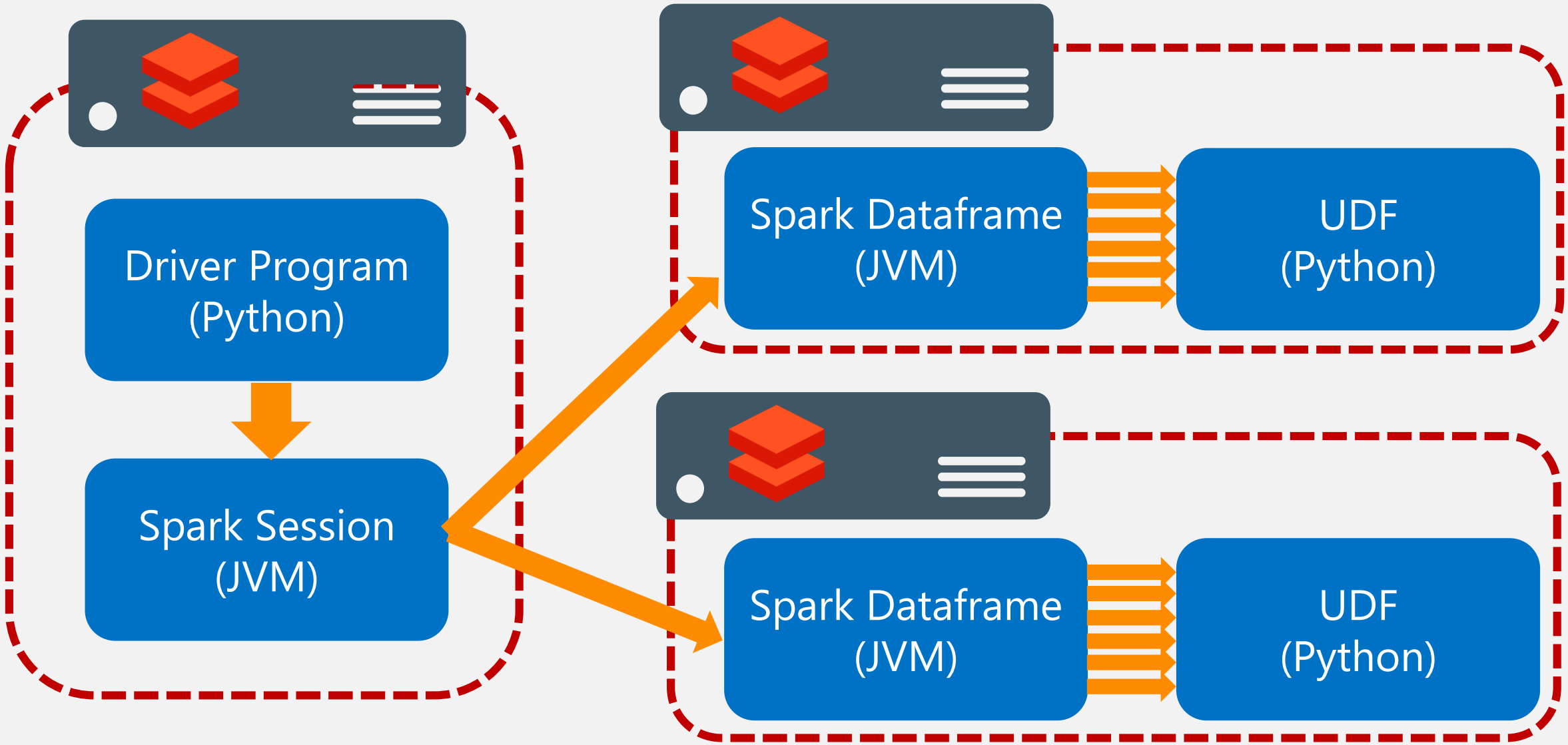
- Use Scala
- Require a specific Spark version or want to configure Spark
- Want to control some advanced parameters

adatis

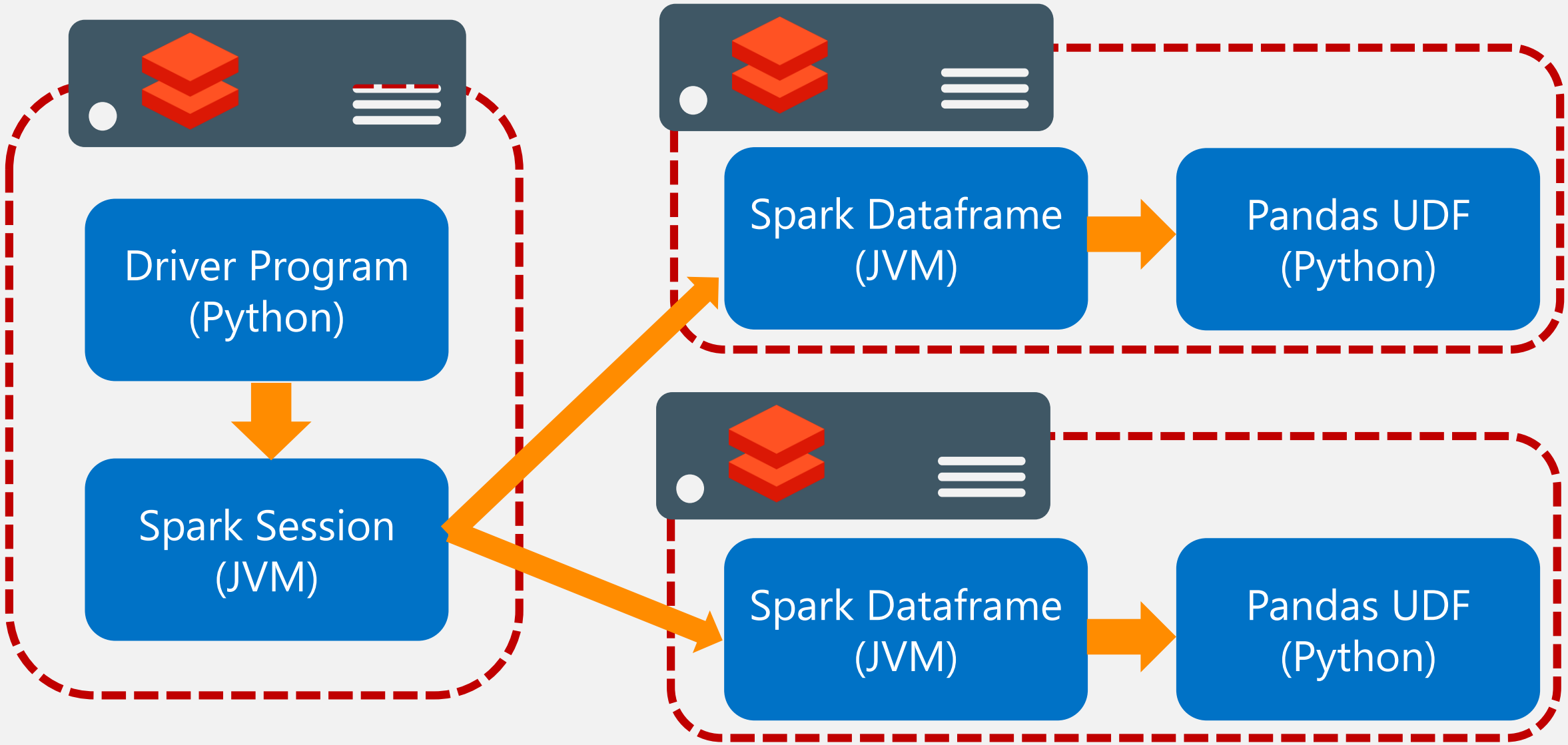
UDFs


Driver Program
(Scala)
Spark Session
(JVM)
Spark Dataframe
(JVM)
UDF
(Scala)
Spark Dataframe
(JVM)
UDF
(Scala)

adatis


Driver Program
(Python)
Spark Session
(JVM)
Spark Dataframe
(JVM)
UDF
(Python)
Spark Dataframe
(JVM)
UDF
(Python)

adatis

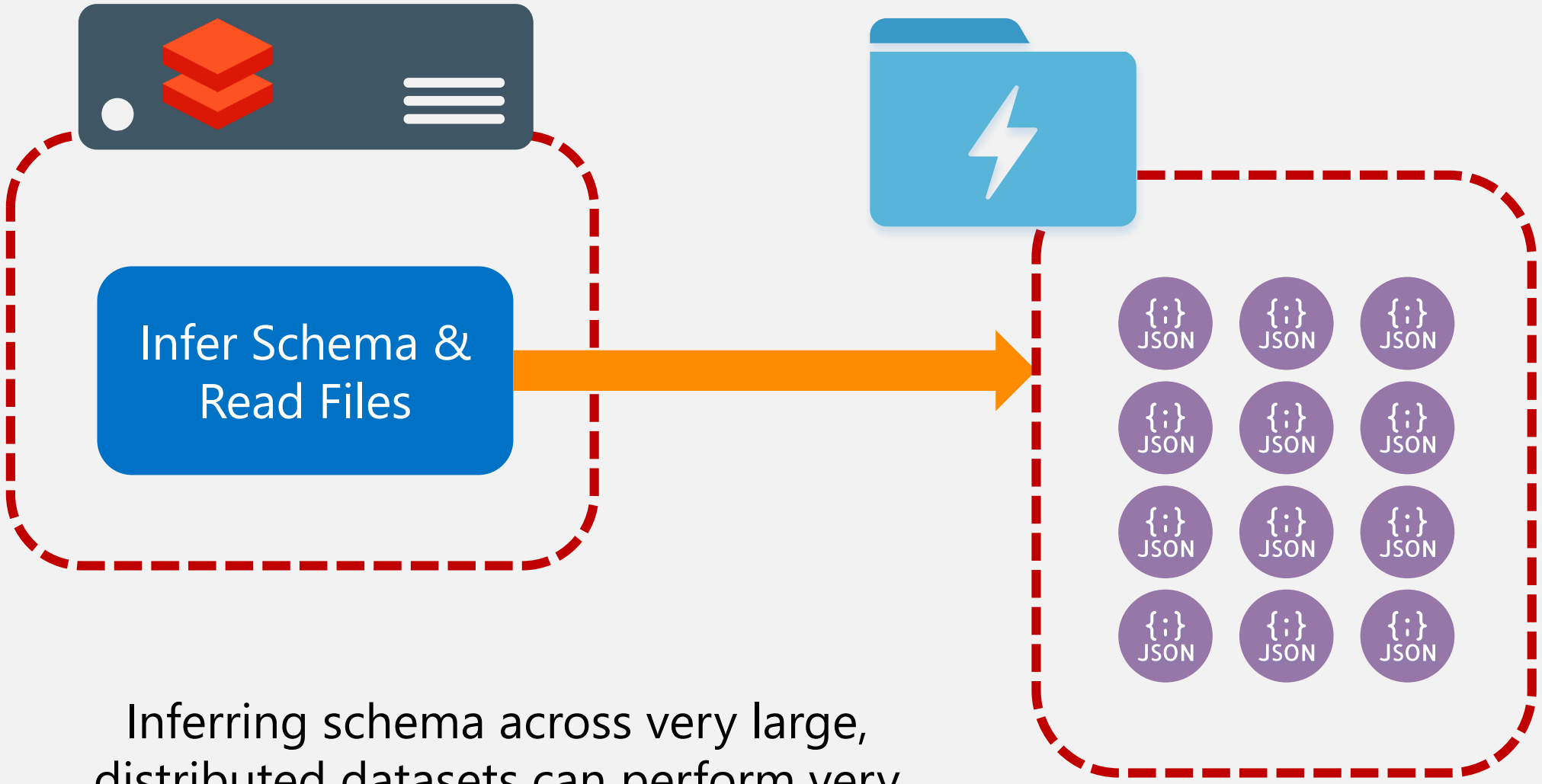

Driver Program
(Python)
Spark Session
(JVM)
Spark Dataframe
(JVM)
Pandas UDF
(Python)
Spark Dataframe
(JVM)
Pandas UDF
(Python)

adatis

Performance comparison of different UDF methods in Databricks

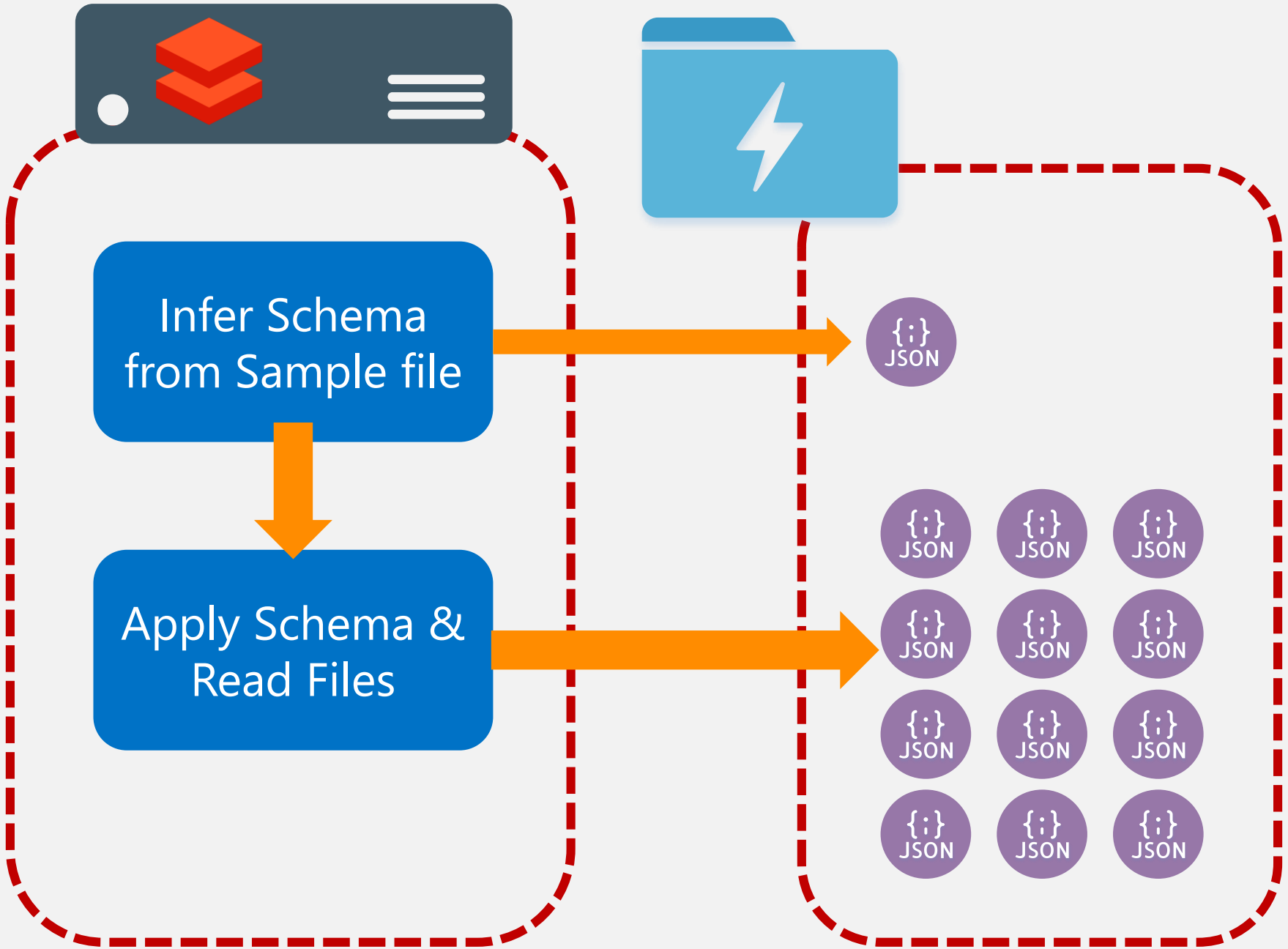
https://bit.ly/2CAXkVI

adatis



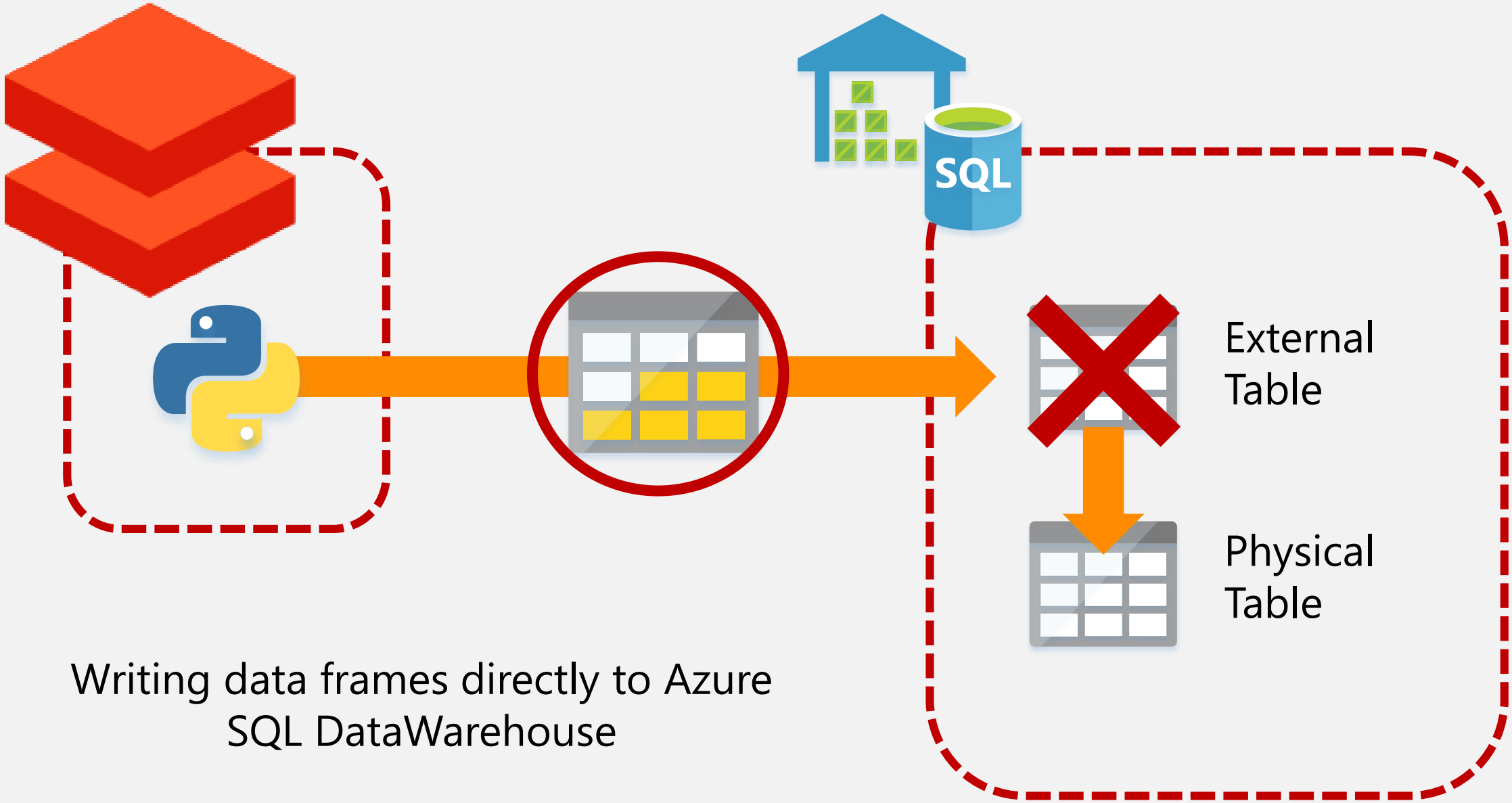
Inferring schema across very large, distributed datasets can perform very badly!

adatis



Acquire schema metadata (or infer from sample file) before reading large datasets!

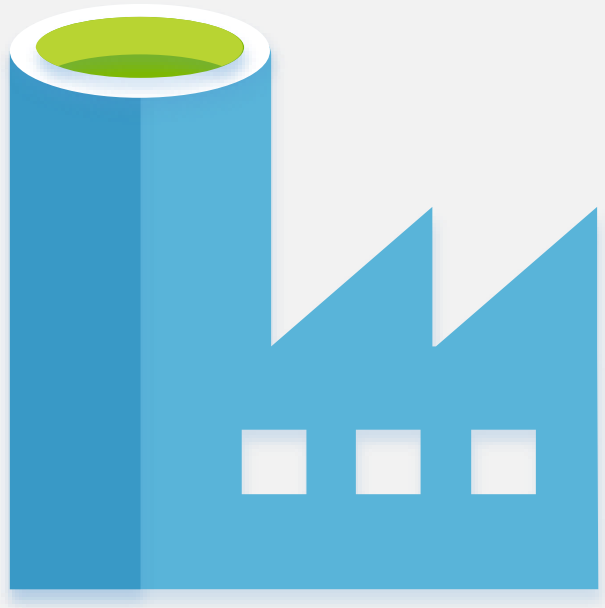
adatis

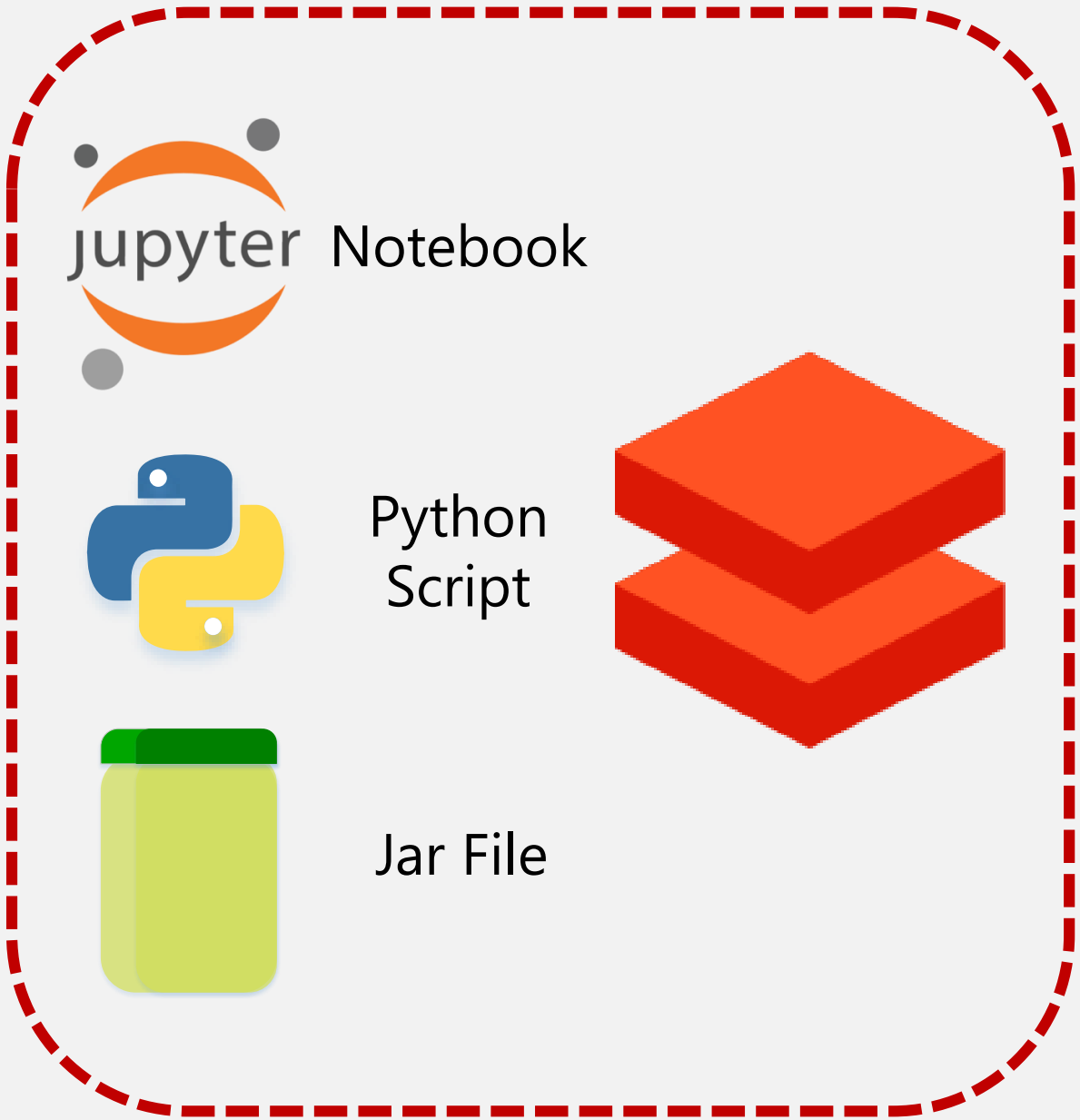


Writing data frames directly to Azure SQL DataWarehouse

adatis

Orchestration



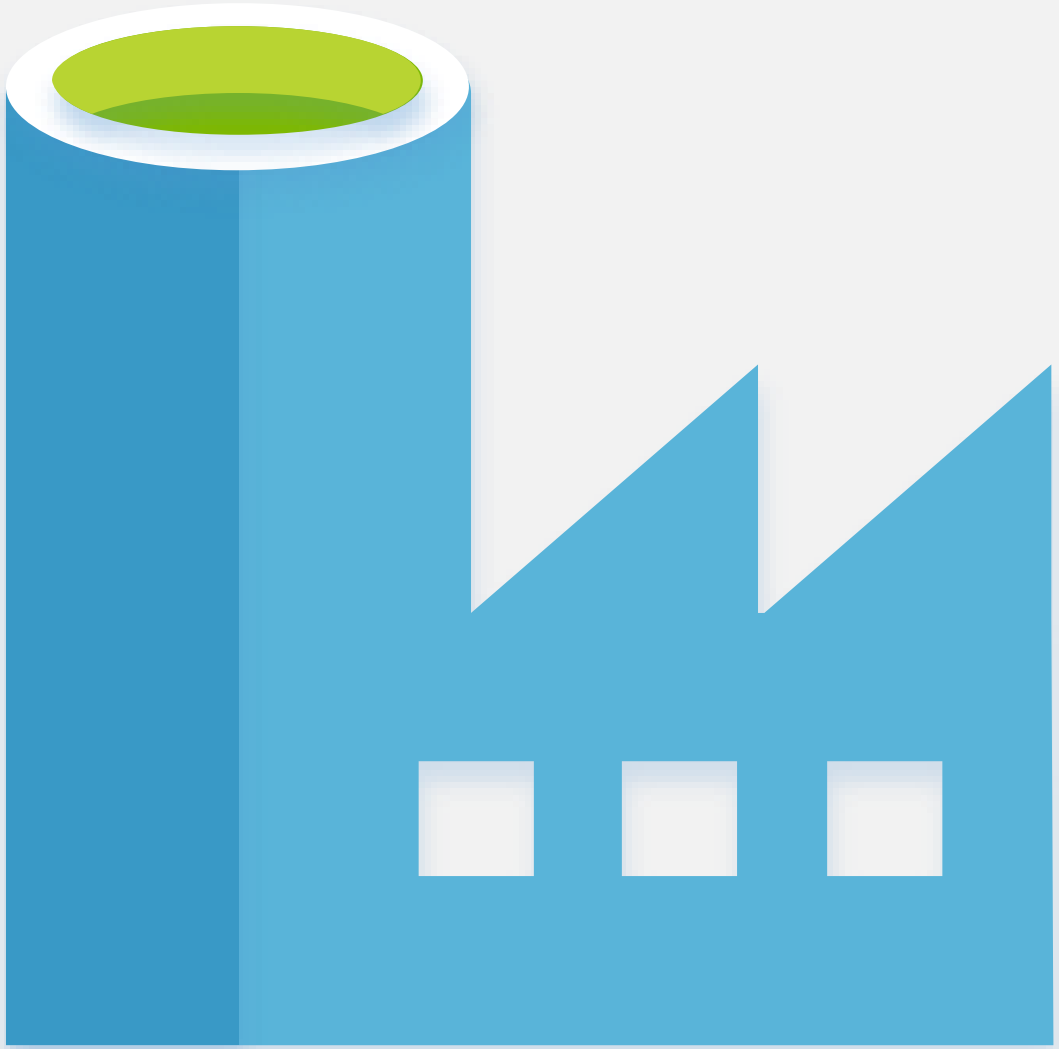


adatis

Data Factory Demo

But what if I don’t want to write any code?

adatis

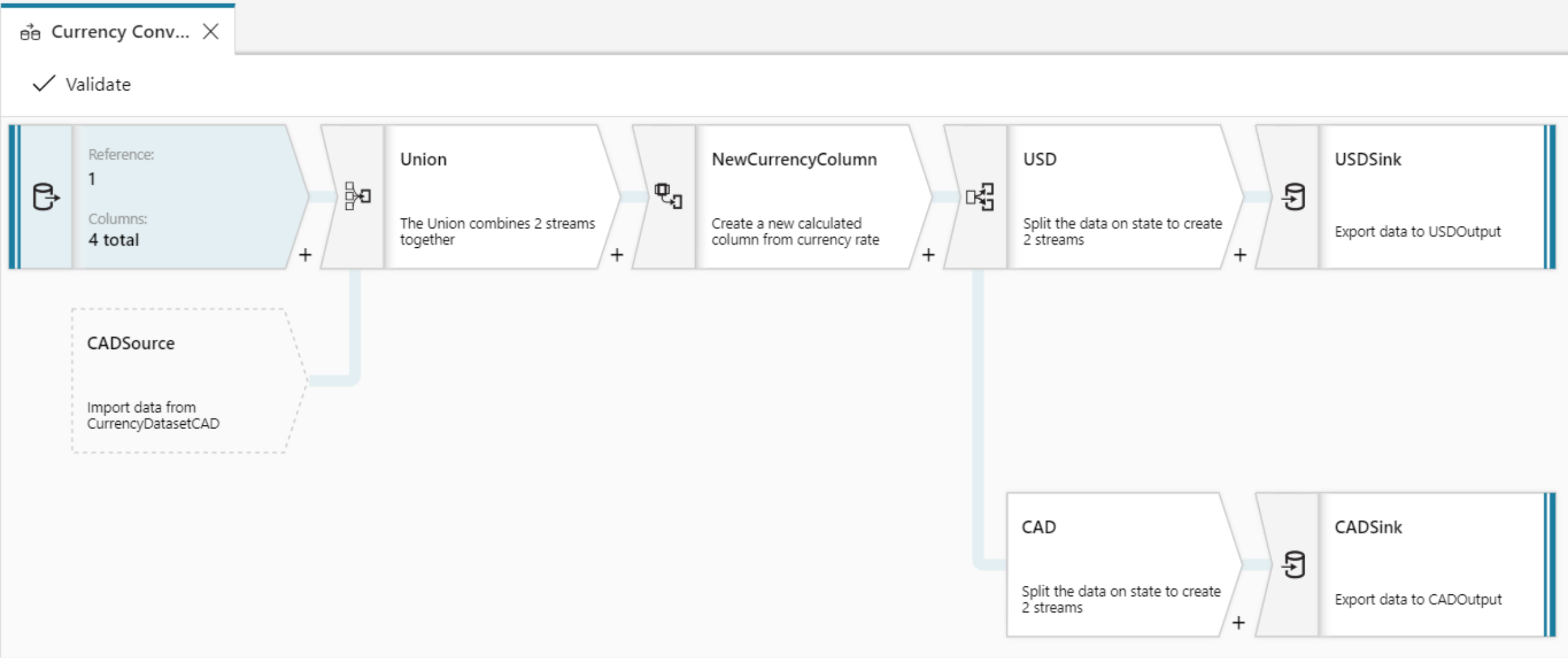


Azure Data Factory

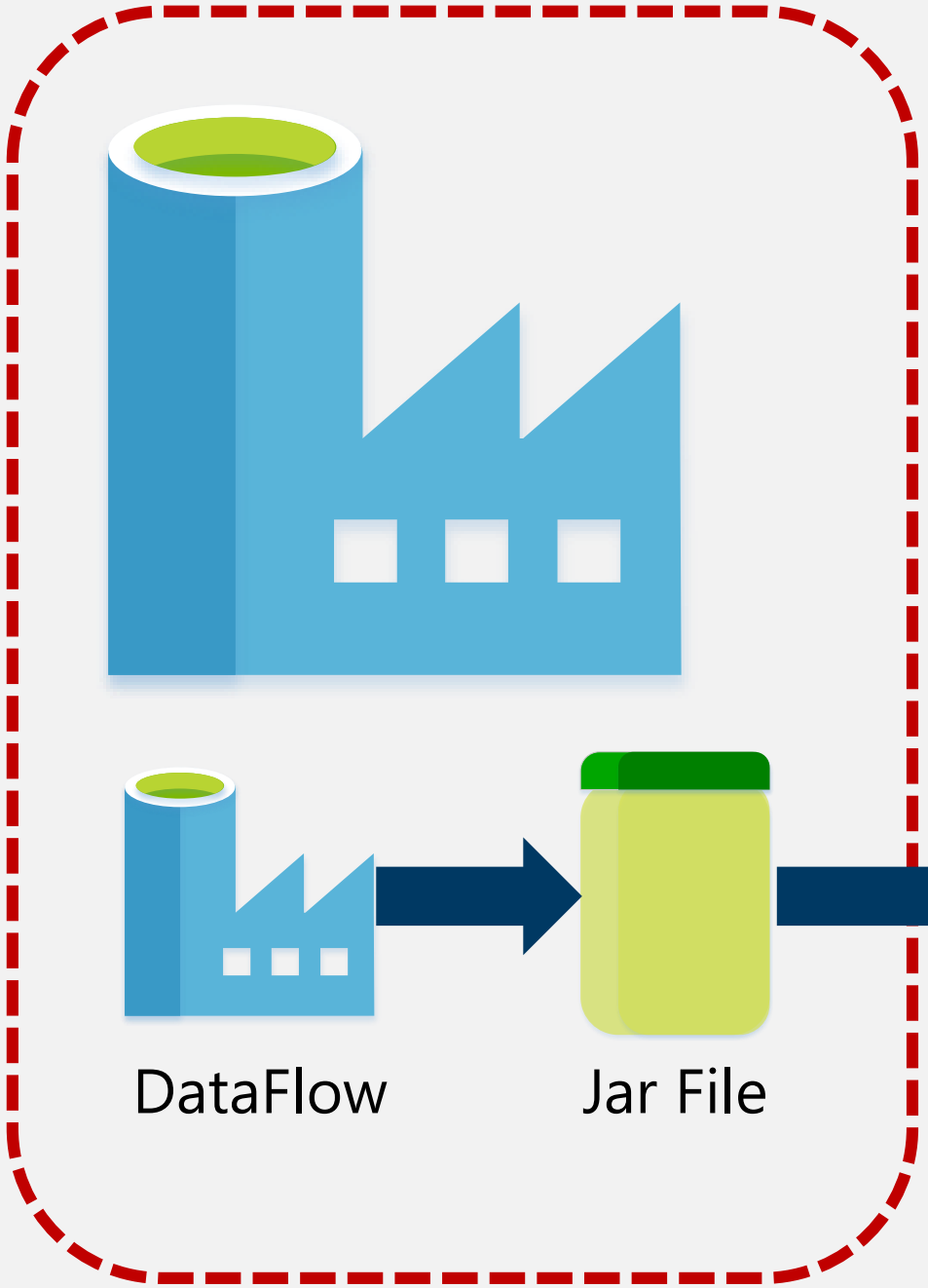
Data Flows

adatis

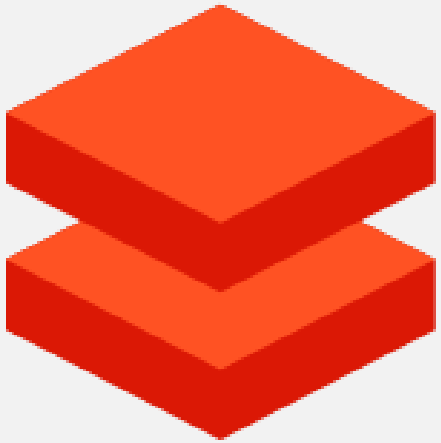
New Data Factory DataFlows can write Databricks processing packages for you!!



adatis



Dataflows will compile down to a JAR file which will be sent to the Databricks cluster for execution



This means it uses Scala!

adatis

Thanks for Listening

Simon Whiteley

@MrSiWhiteley

http://blogs.adatis.co.uk