Python Pipeline Primer

ETL in Azure

Simon Whiteley | Adatis

https://github.com/SiWhiteley/DatabricksETL

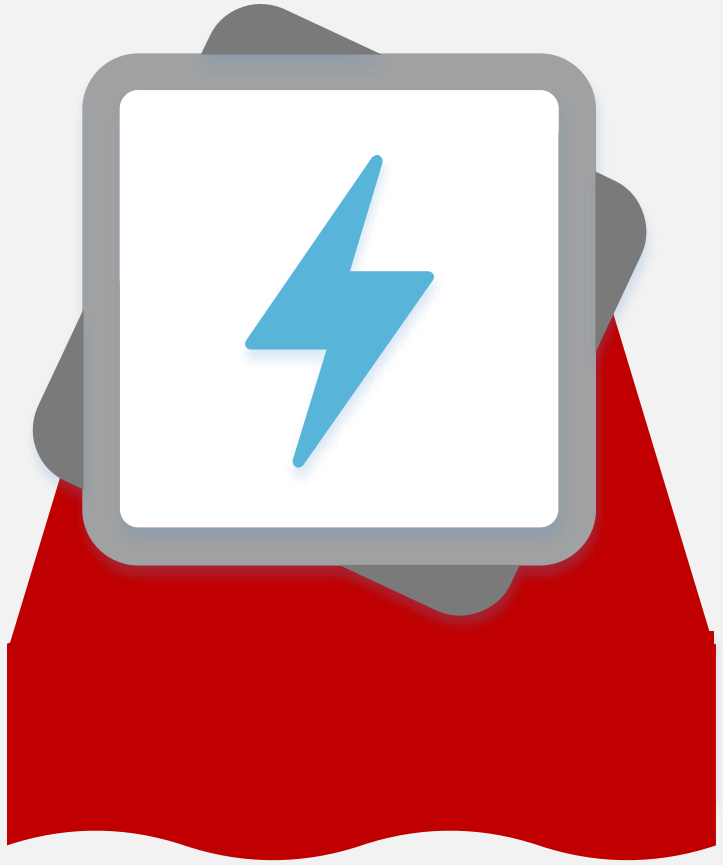
Agenda

What is Databricks?

Patterns & Implementation

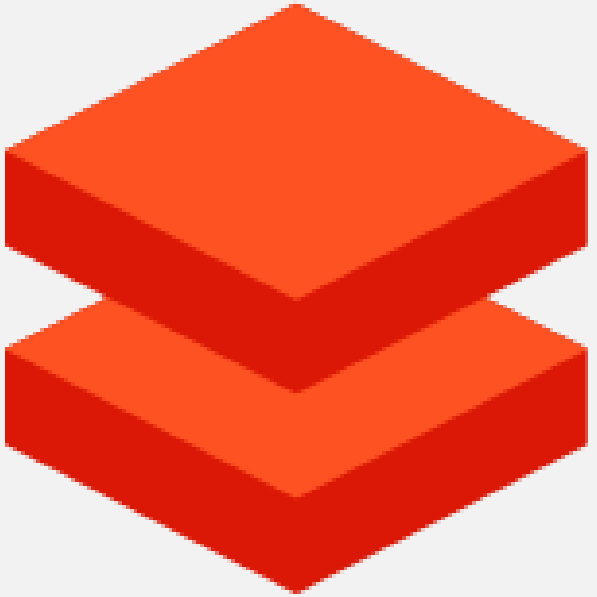
Orchestration

Data Factory Dataflows



DATA LAKE ANALYTICS

adatis



Azure
Databricks

adatis

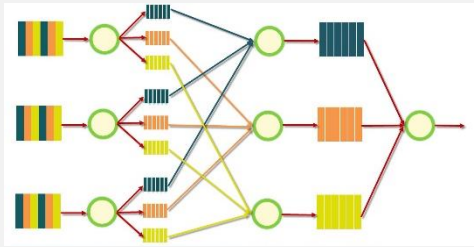
Databricks?

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

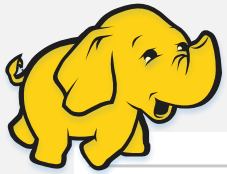
2018

adatis

So What?

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

adatis



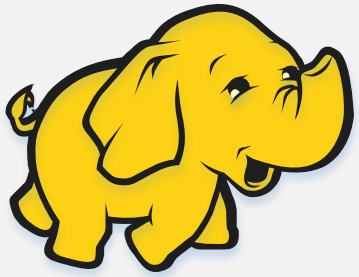
HDInsight

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	1 YEAR RESERVED (% SAVINGS) TOTAL PRICE	3 YEAR RESERVED (% SAVINGS) TOTAL PRICE
D3 v2	4	14.00 GiB	0.75	£0.208/hour	£0.168/hour	£0.376/hour	£0.296/hour (~21%)	£0.25/hour (~34%)

adatis



Open Source

20 min provisioning

Integrates Well

Secure

Hadoop, Spark, Kafka, Hbase, HIVE, Storm…

Slow Release Cycle

Open Source

5 min provisioning

Integrates Well

Secure

Spark (Python/Scala/R)

Fast Release Cycle

Proprietary

1 m rovi ning

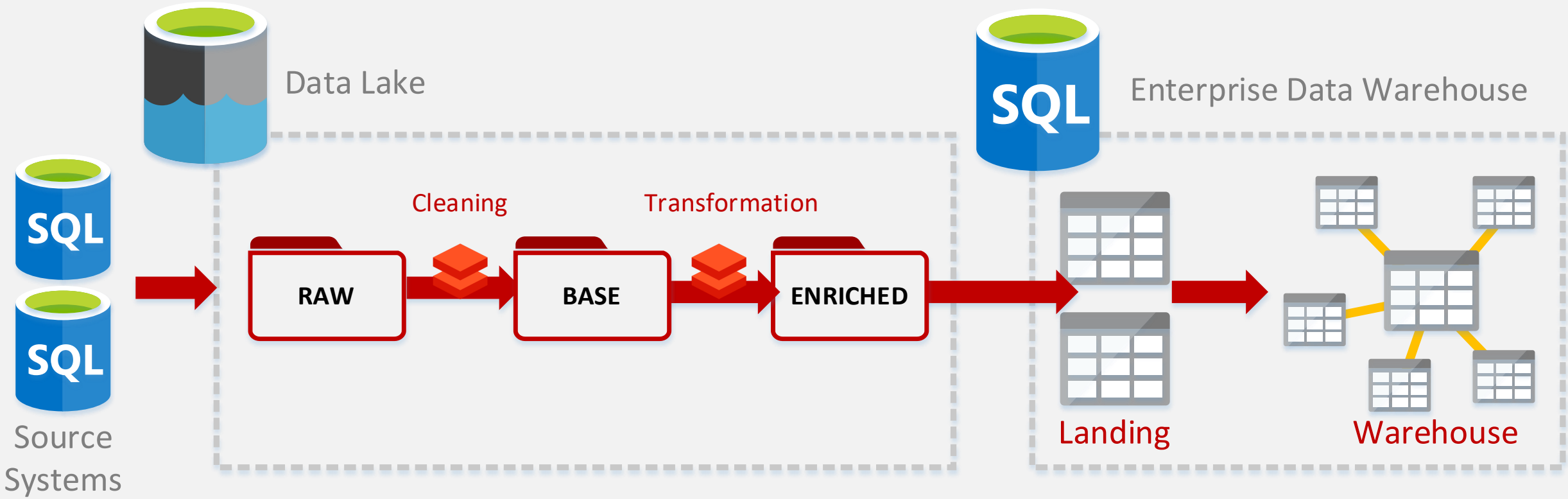
Integr Poorly

cu

U-SQL

ow Release Cyc

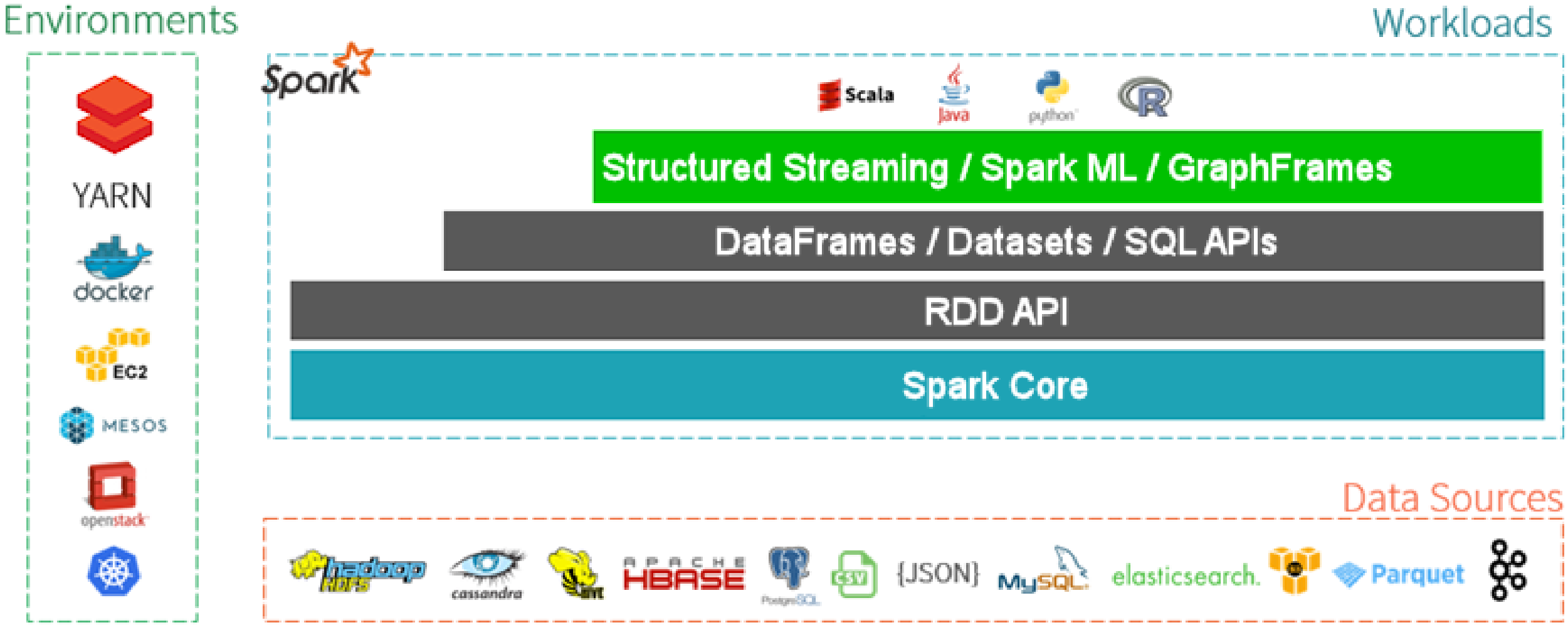
adatis

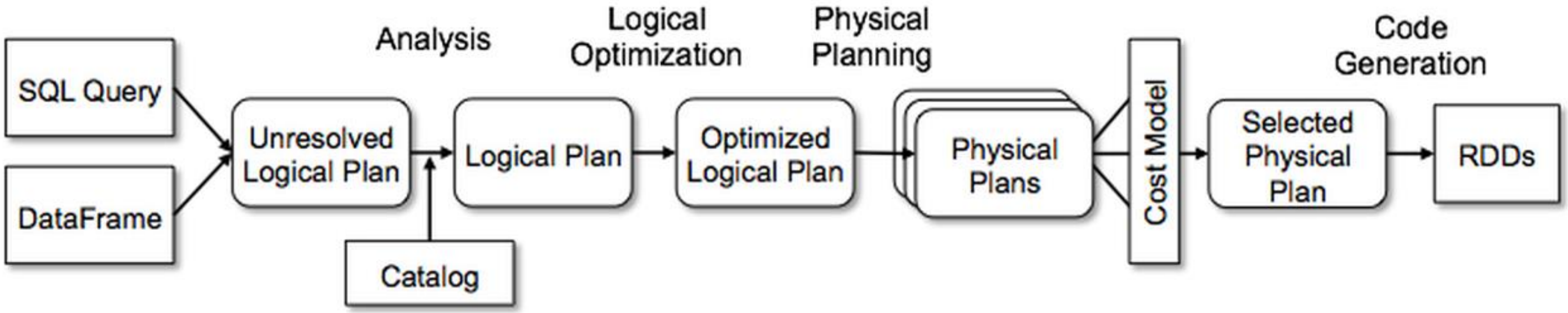

Data Lake
SQL
Enterprise Data Warehouse
SQL
SQL
Source Systems
Cleaning
Transformation
RAW
BASE
ENRICHED
Landing
Warehouse

adatis

Databricks Basics

Under the Hood

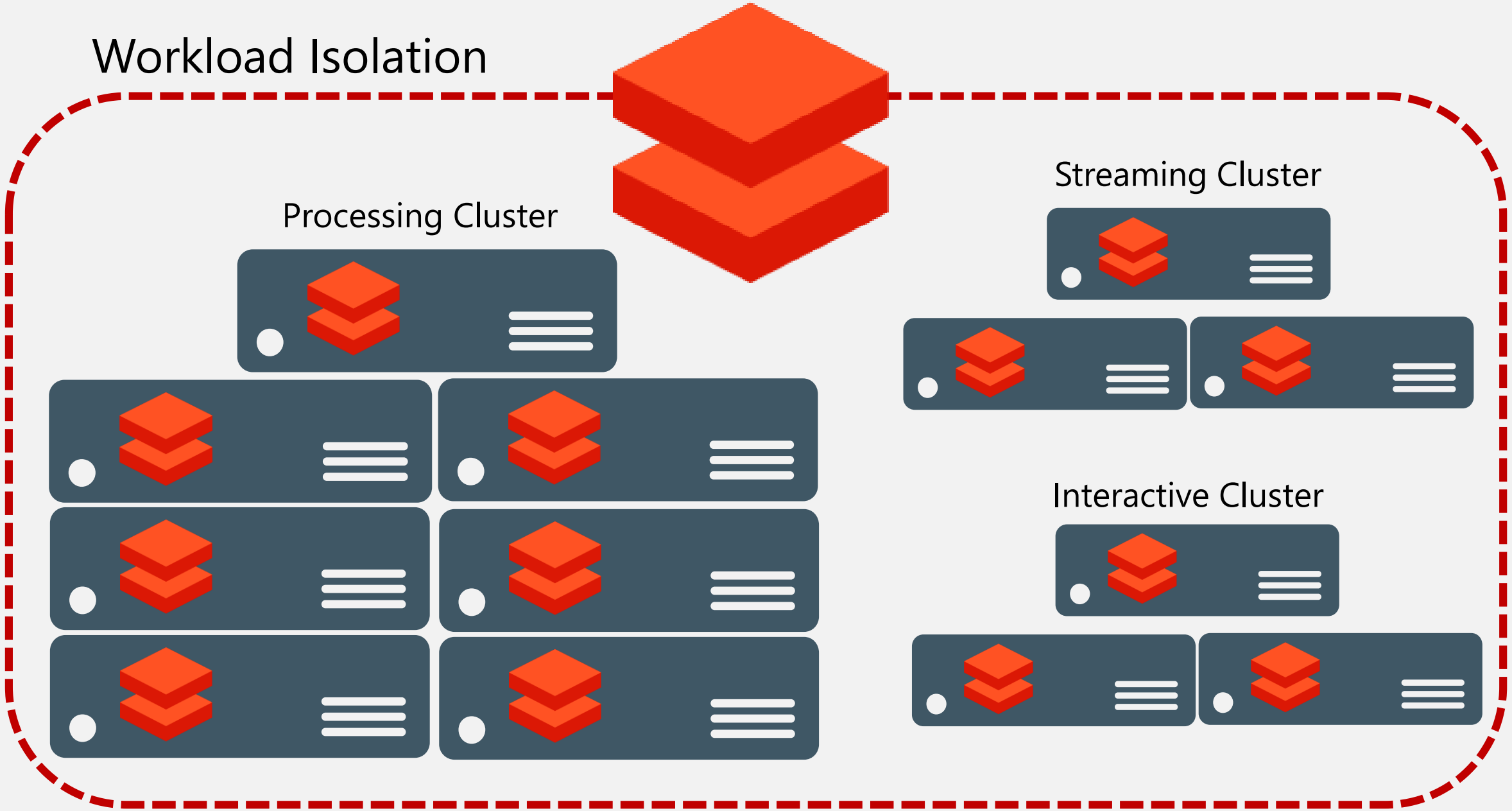



SQL Query
DataFrame
Analysis
Unresolved Logical Plan
Catalog
Logical Plan
Logical Optimization
Optimized Logical Plan
Physical Planning
Physical Plans
Cost Model
Code Generation
Selected Physical Plan
RDDs

adatis

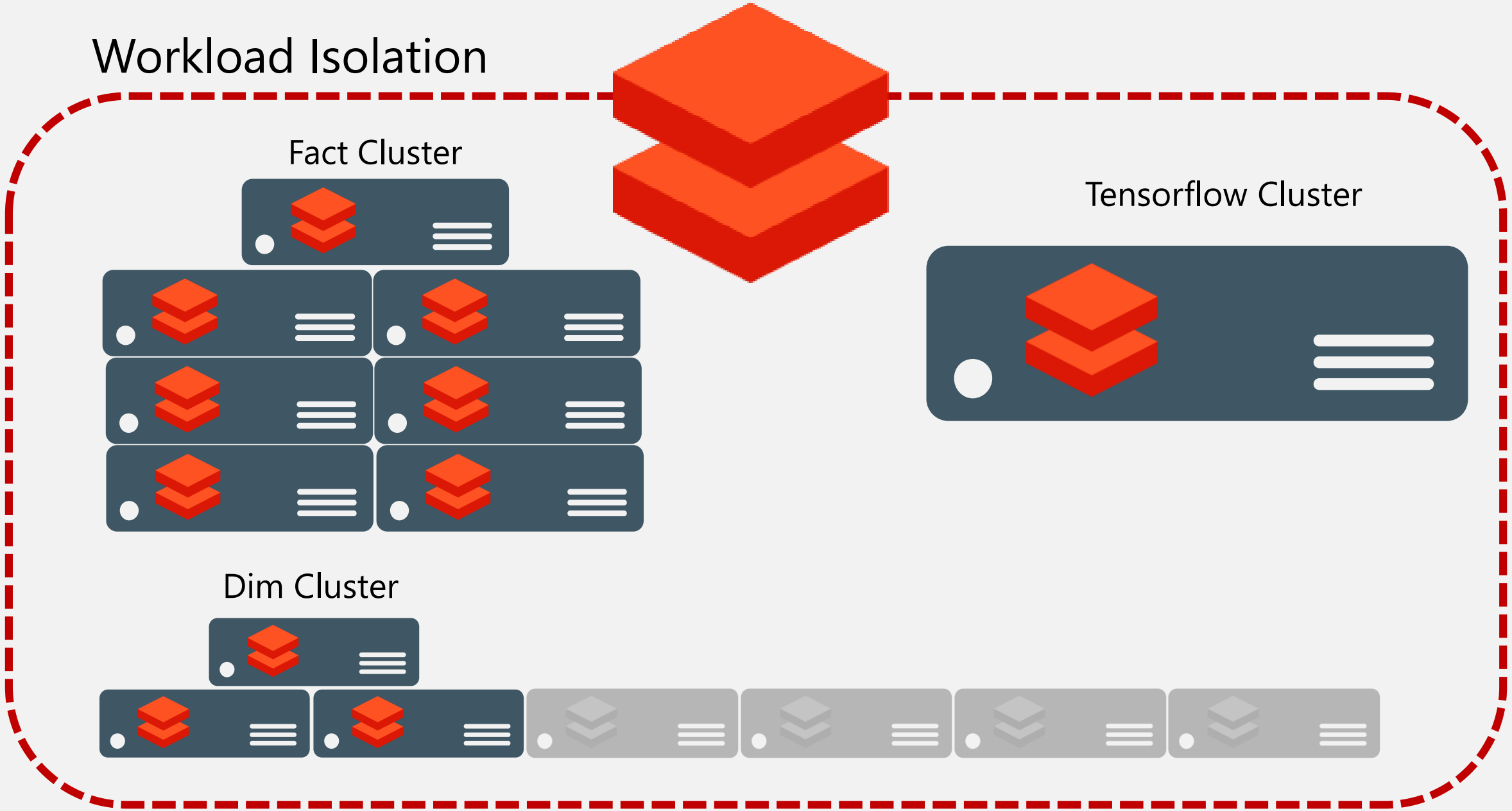
Patterns & Implementation

Workload Isolation



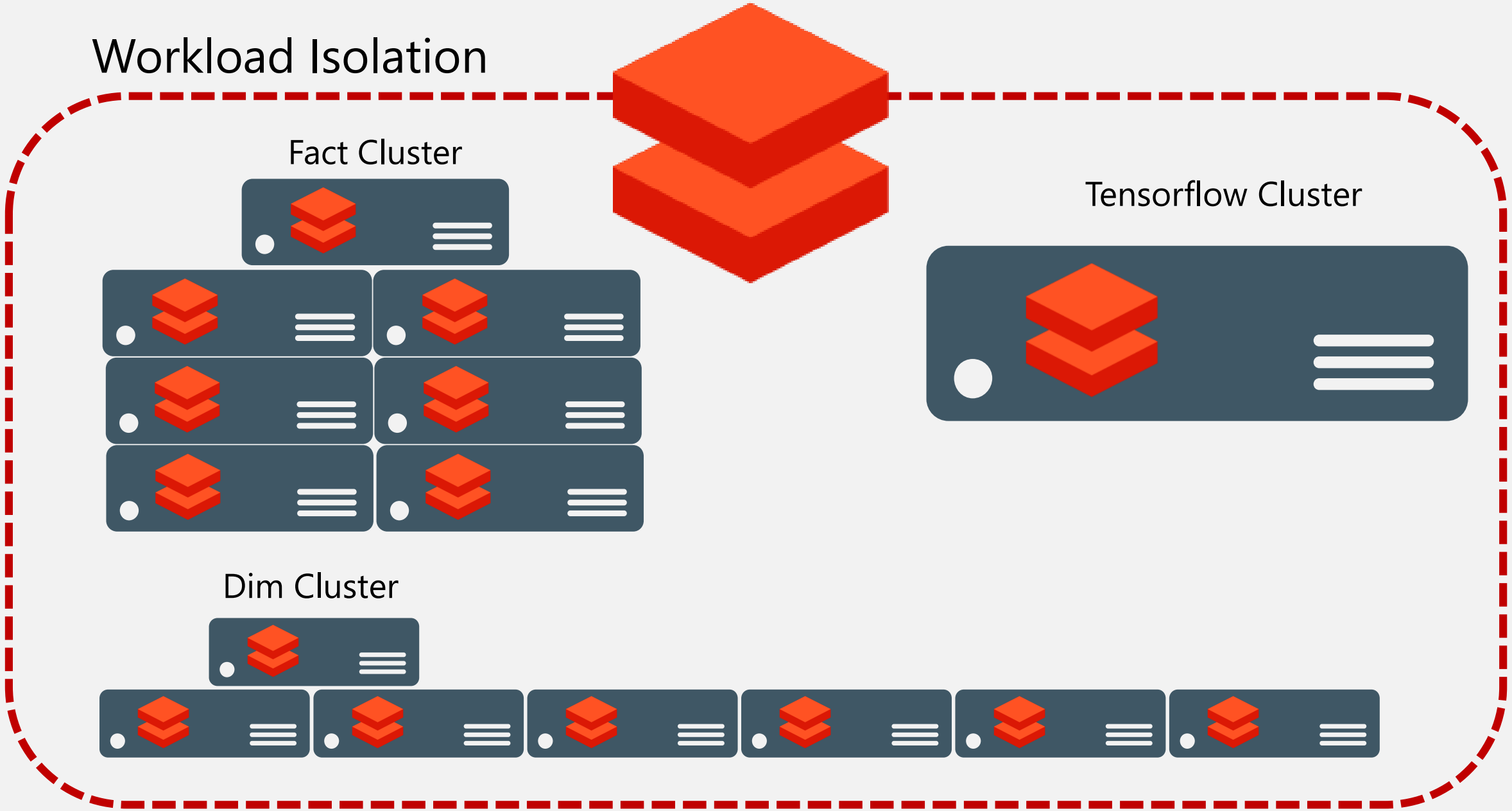
adatis

Workload Isolation



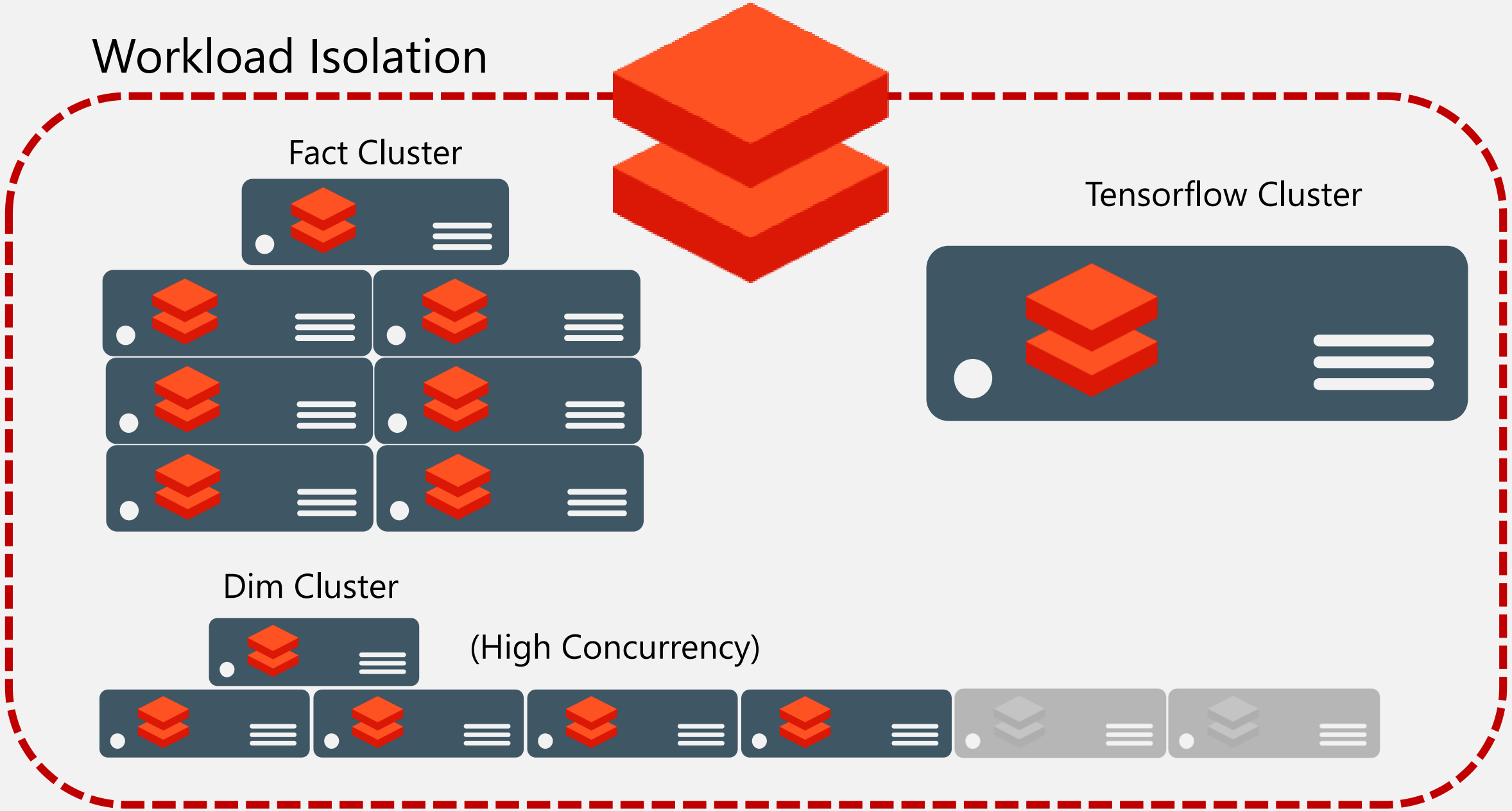
adatis

Workload Isolation



adatis

Workload Isolation



adatis

So…. What size?

Size of Driver

What is the largest dataset that we will perform need to return to the user? What actions do we need to perform outside of the spark engine? How performance / memory intensive is it? How many concurrent workers does my driver need to handle?

Size of Worker

What is the largest data set/single partition that needs to fit on a single executor?
How much memory should be left over for performing calculations?
How fast should each executor finish their job?

Number of Workers

What is the total amount of data that needs to be held in memory (both for in transformation queries, and cached tables)
How much concurrency do I need?

Example workloads

Standard ETL Load – Small Data

Small cluster, shared across multiple low priority workloads. Autoscaling for better concurrency

Standard ETL Load – Large Data

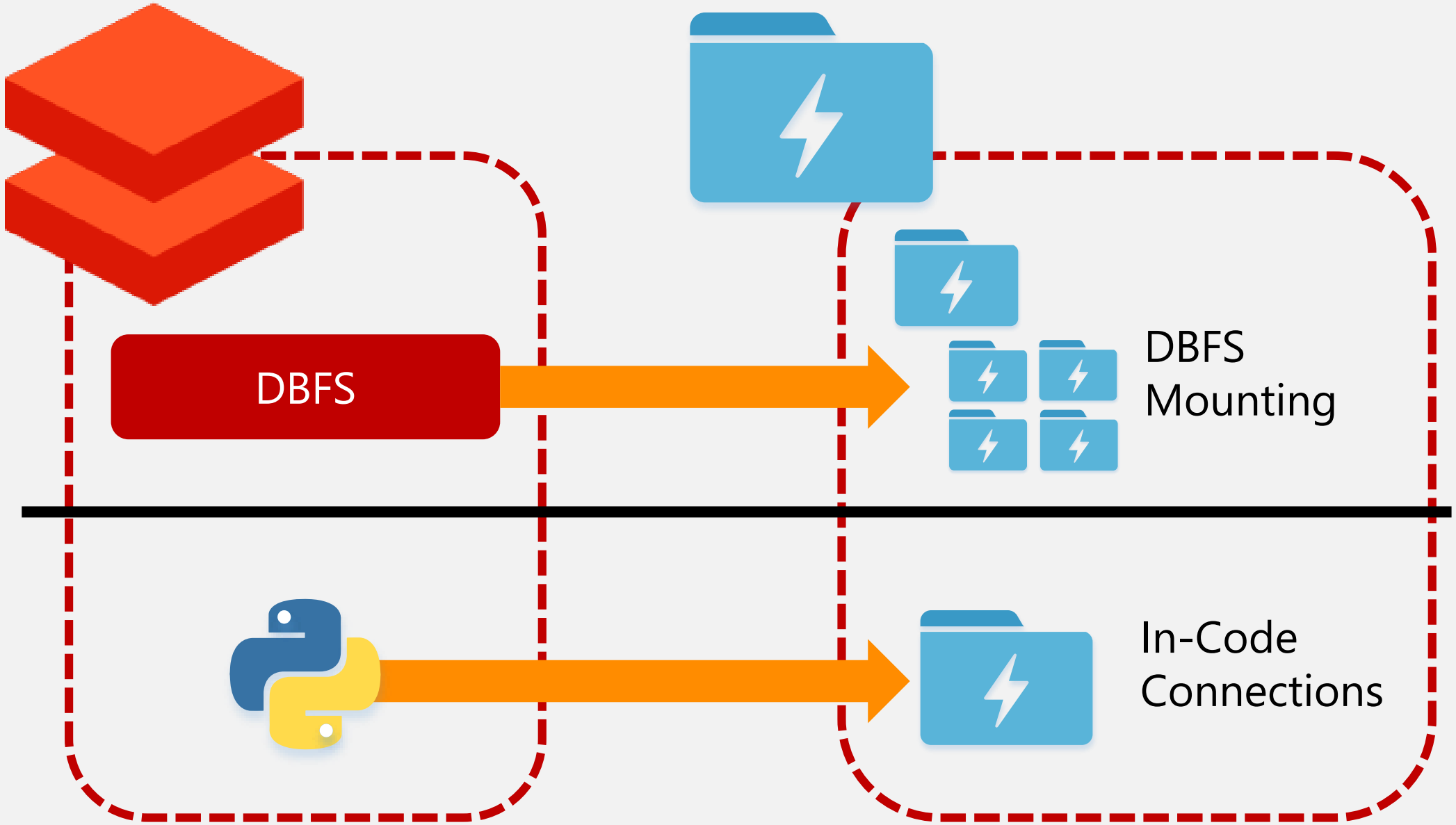
Many small worker nodes, assuming transformations can be distributed

Large Data Science Process

Fewer high-power worker nodes, each executor needs more compute power to train models

Analytics Load – Small Data

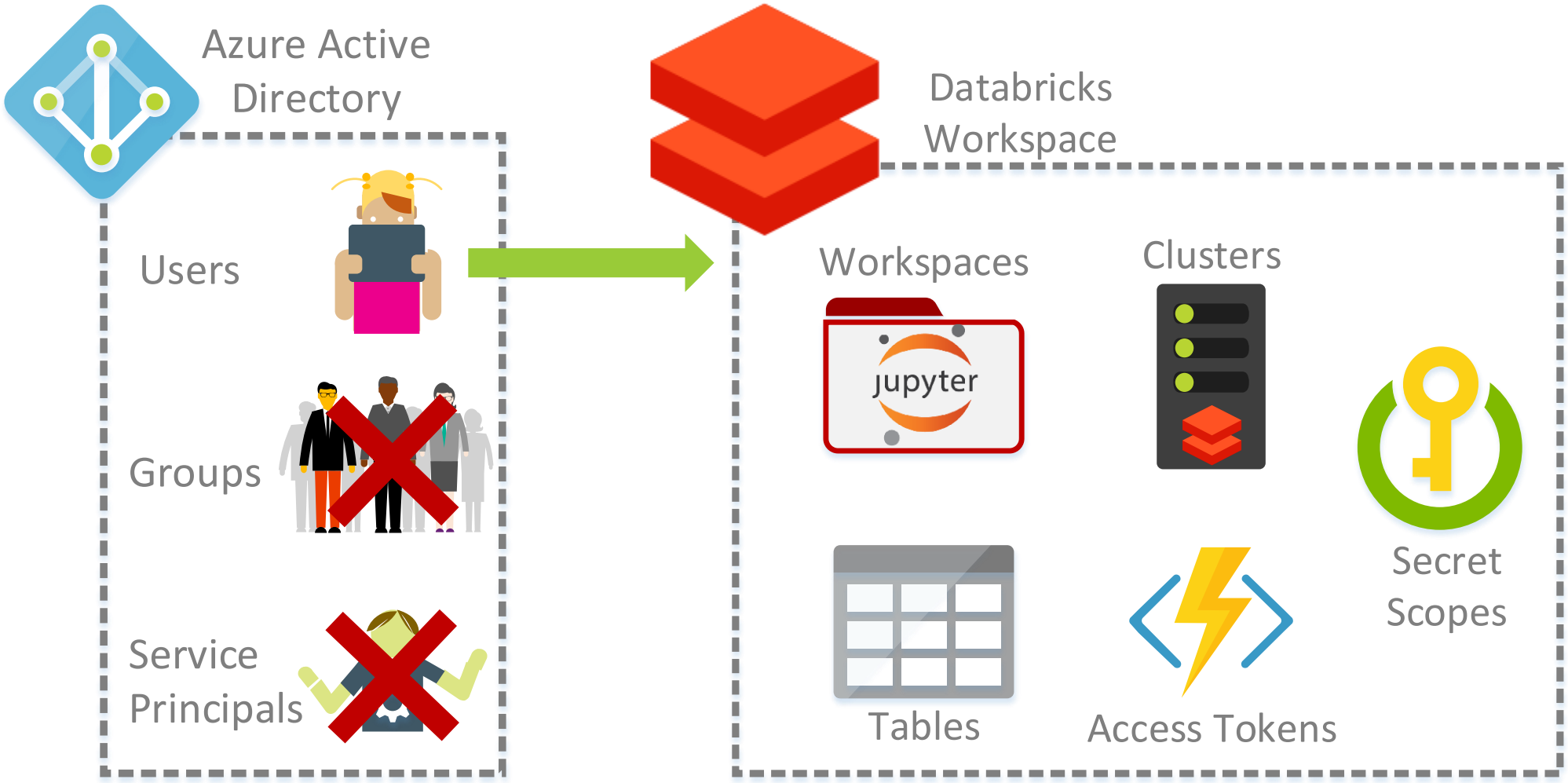
Few, low-powered worker nodes with autoscaling. If using cached tables, needs memory for full data set, plus additional transformation capacity



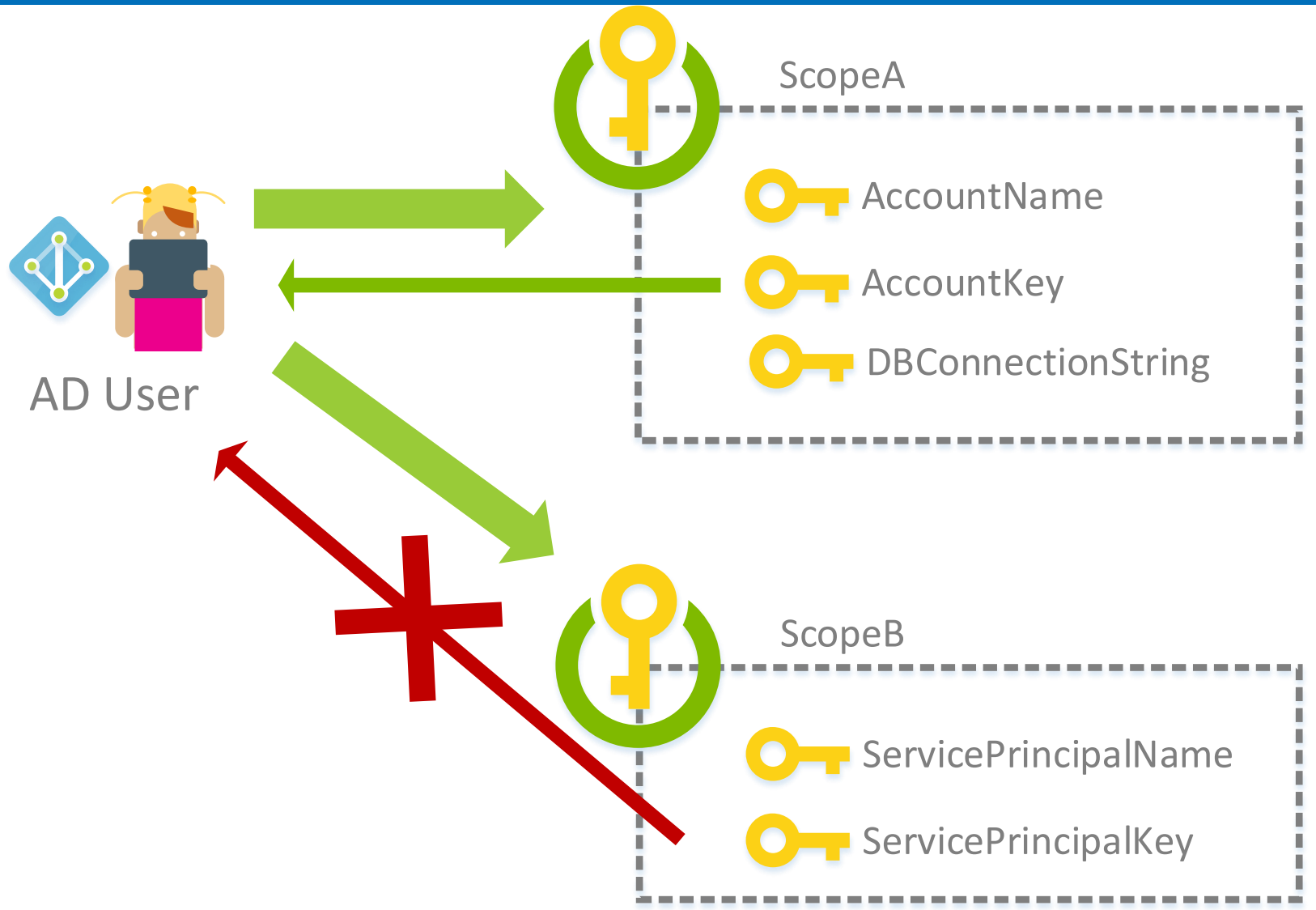
adatis

Secrets

User Management



Databricks Secrets

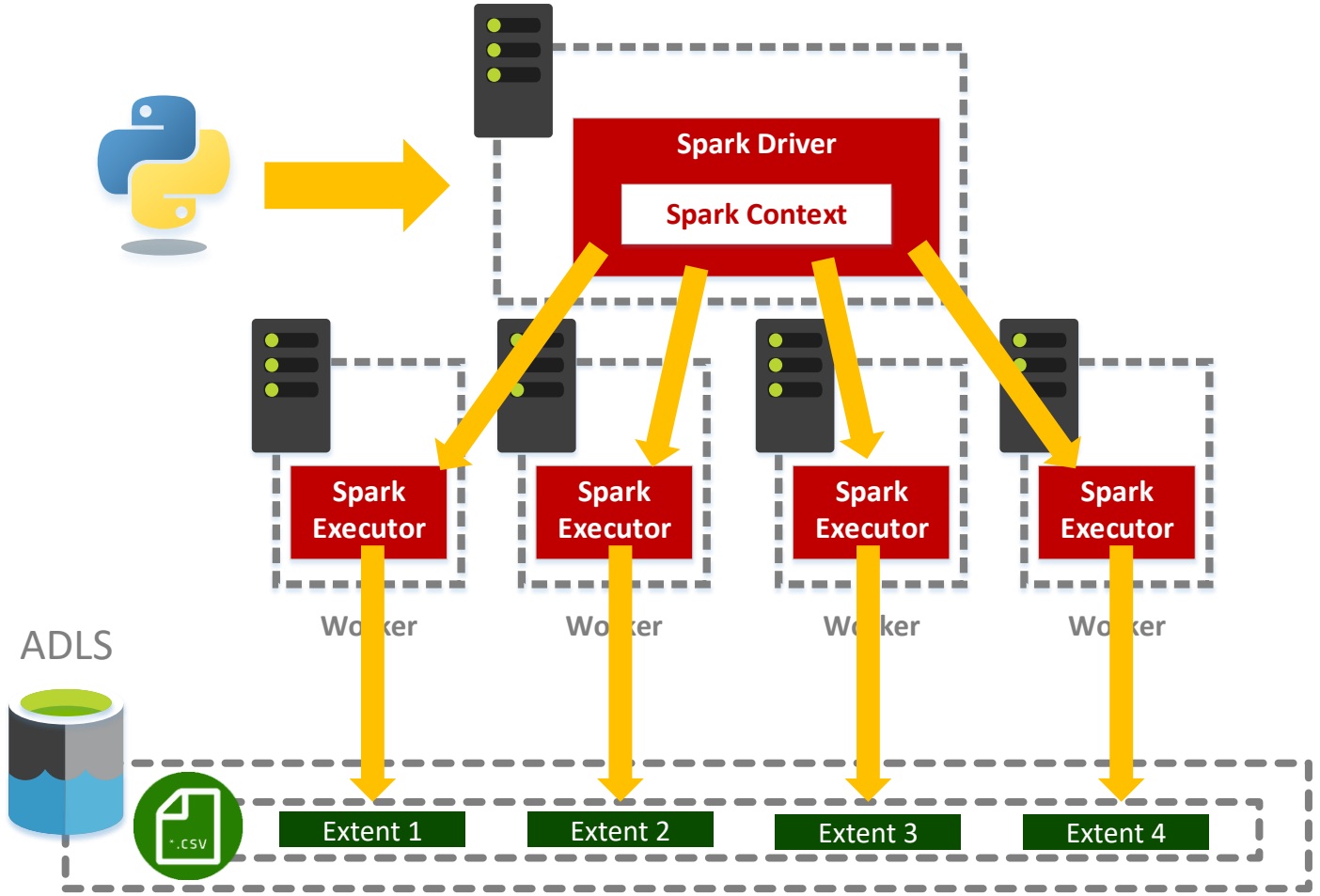


Secrets are never displayed in databricks notebooks, even if you have access!

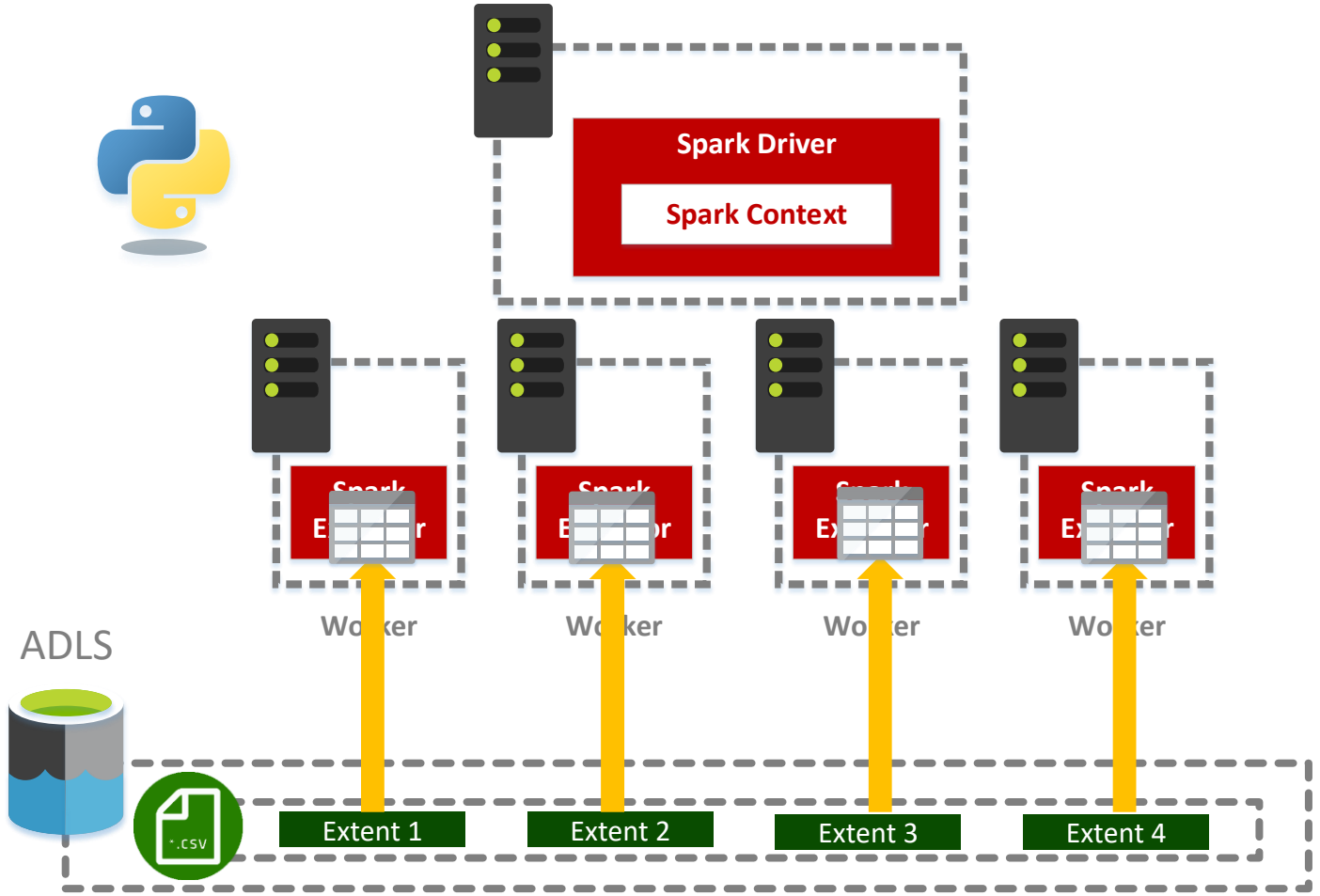
Any attempt to display the value will return **[REDACTED]**!

Executions

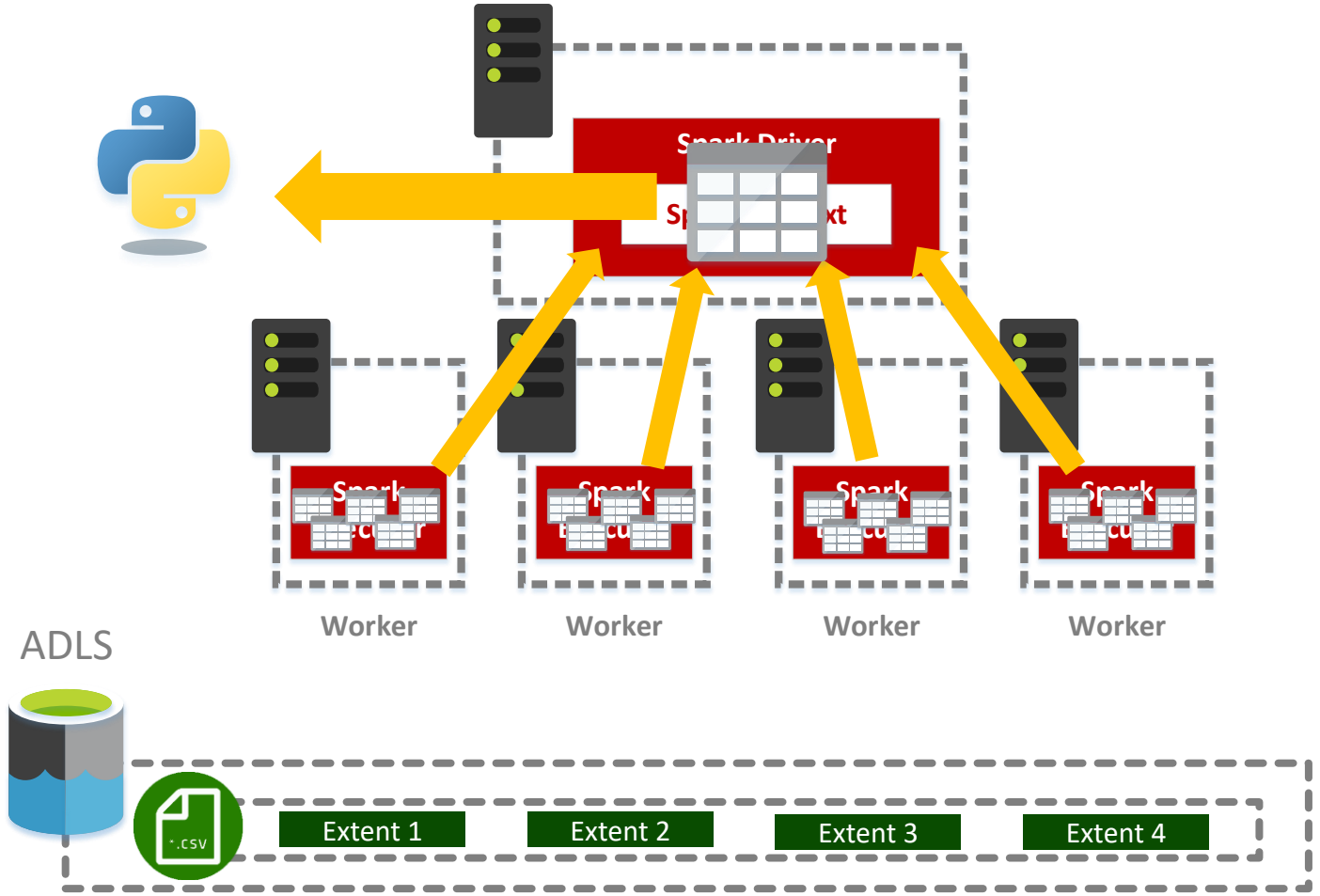
Distributed Compute



Distributed Compute



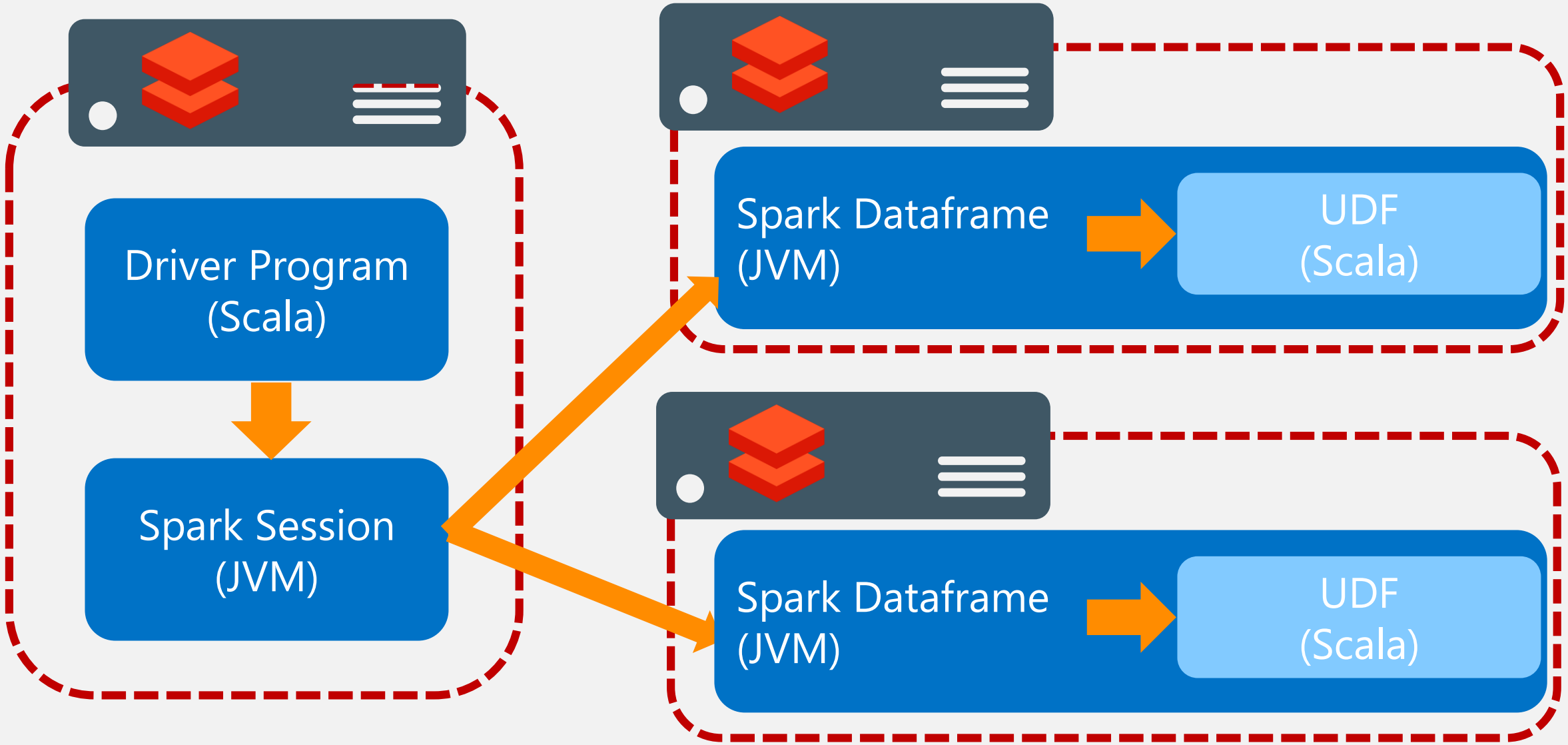
Distributed Compute



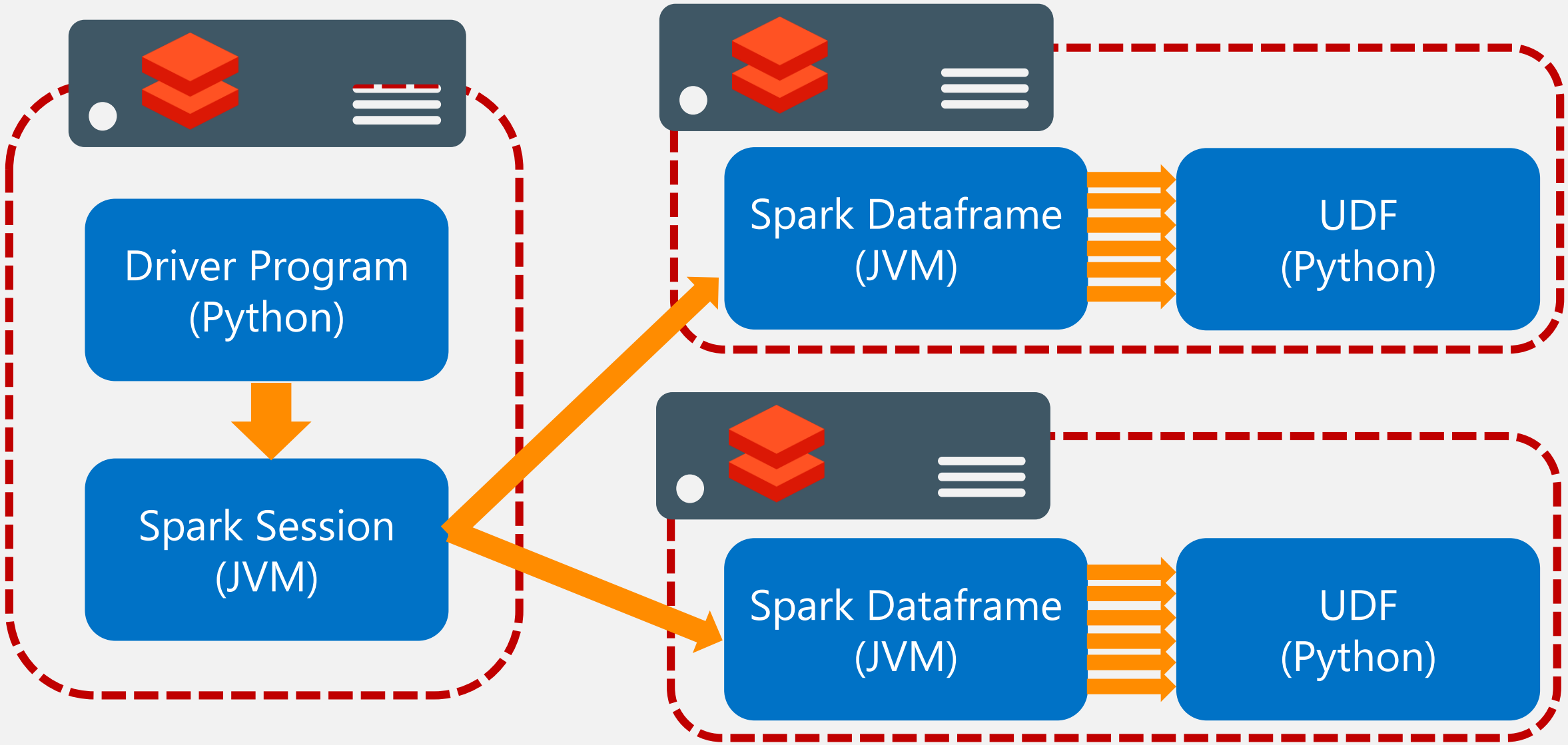
All languages perform the same

…except…

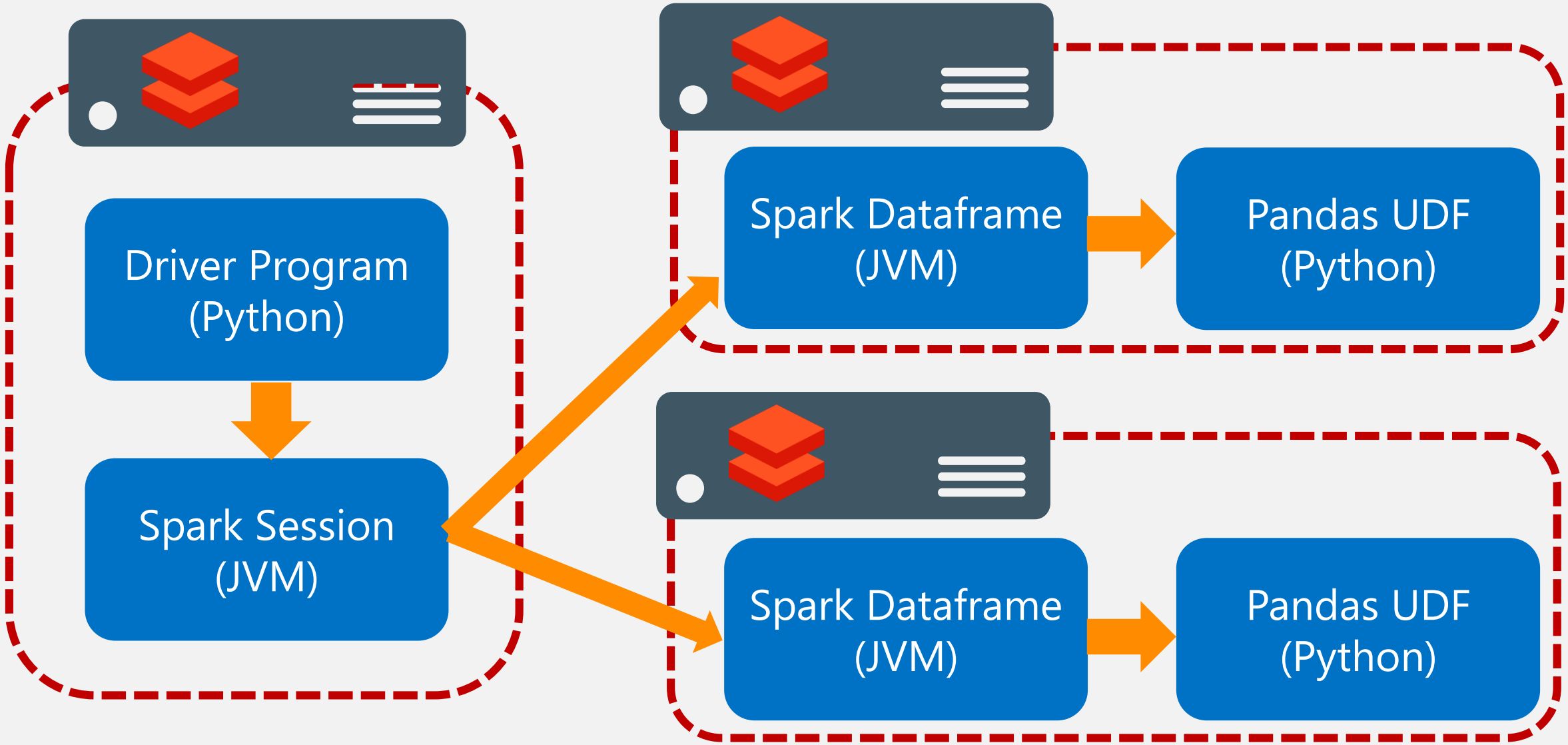
UDFs


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

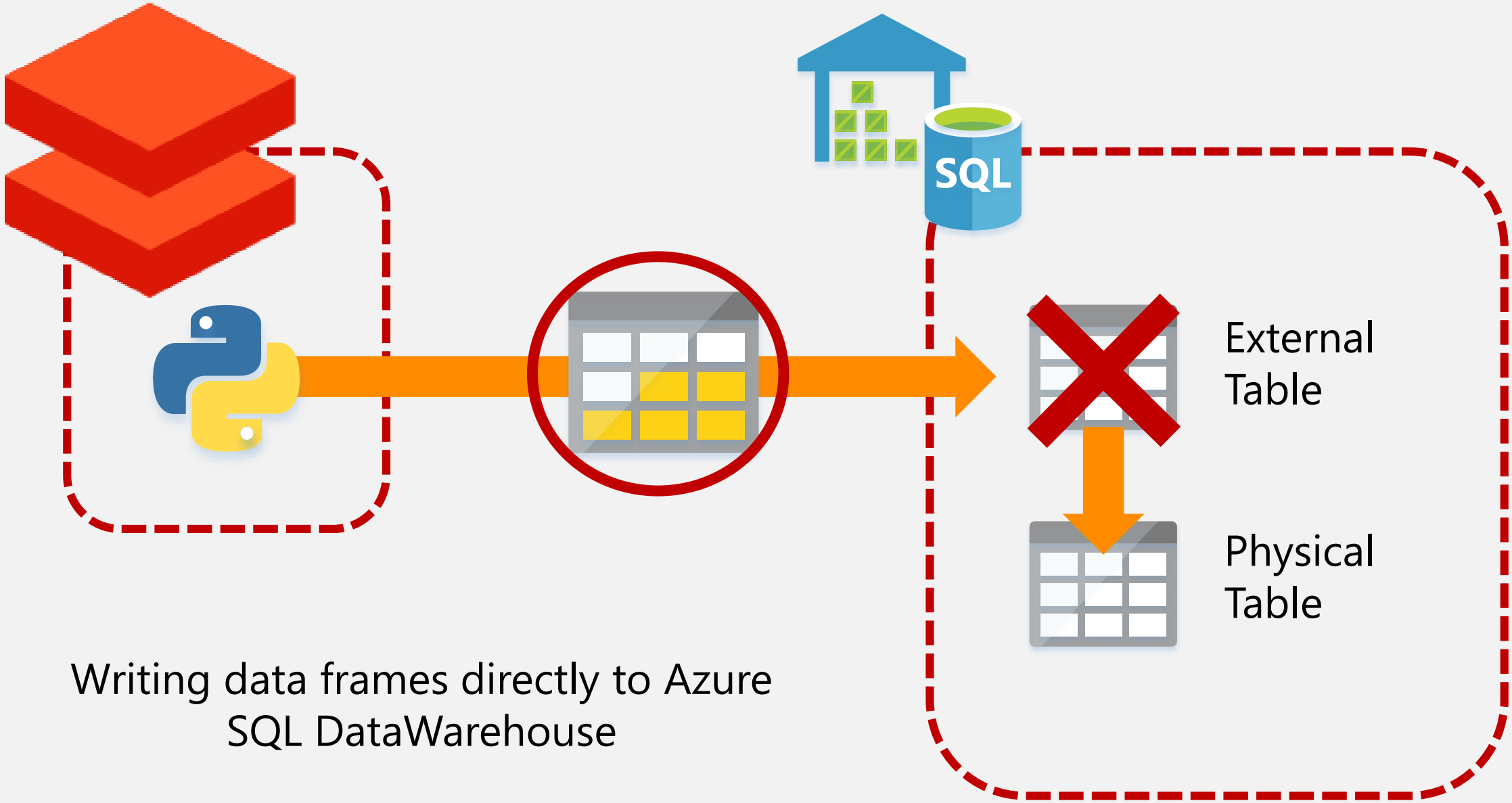

Driver Program
(Python)
Spark Session
(JVM)
Spark Dataframe
(JVM)
Pandas UDF
(Python)
Spark Dataframe
(JVM)
Pandas UDF
(Python)

adatis

Performance comparison of different UDF methods in Databricks

https://bit.ly/2CAXkVI

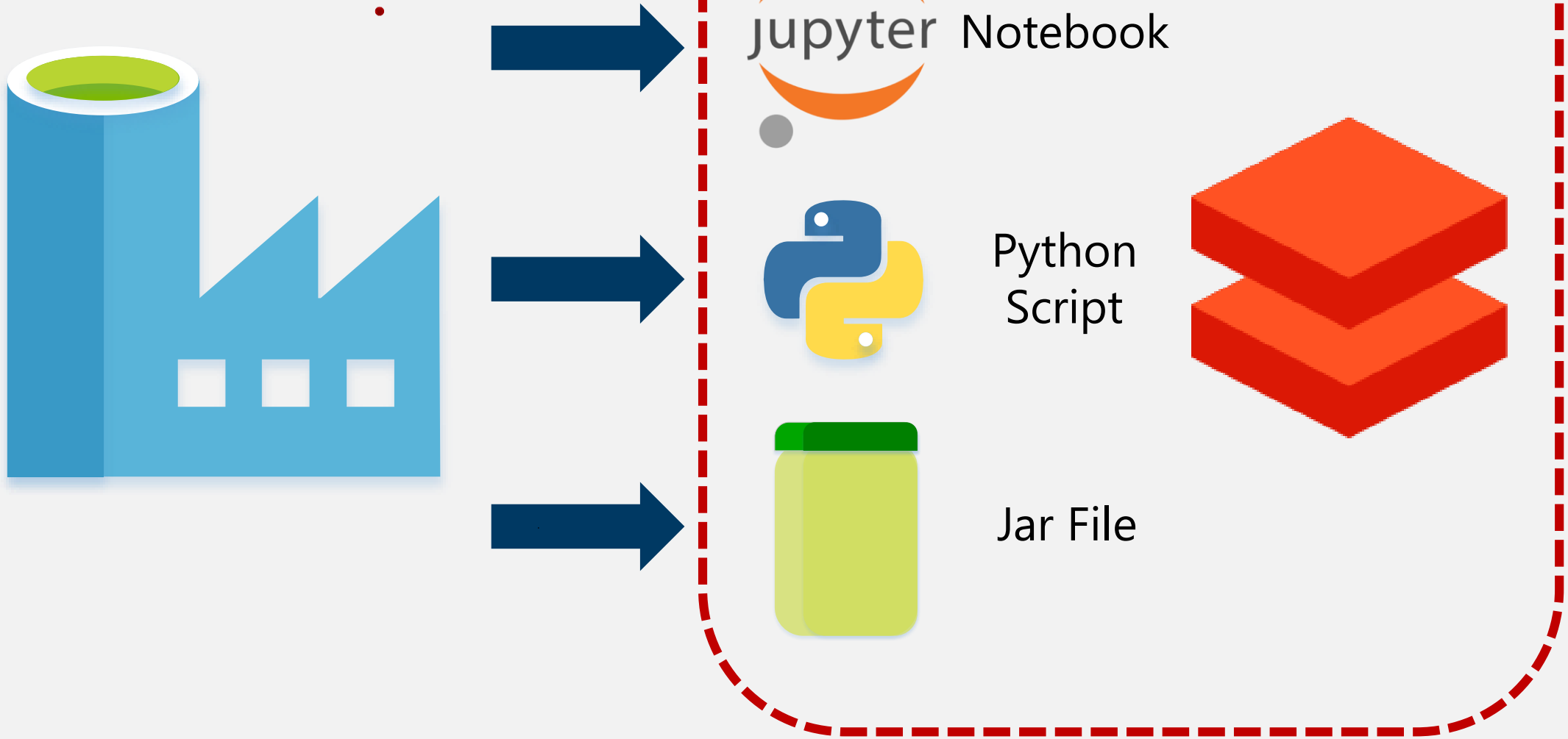
adatis



Writing data frames directly to Azure SQL DataWarehouse

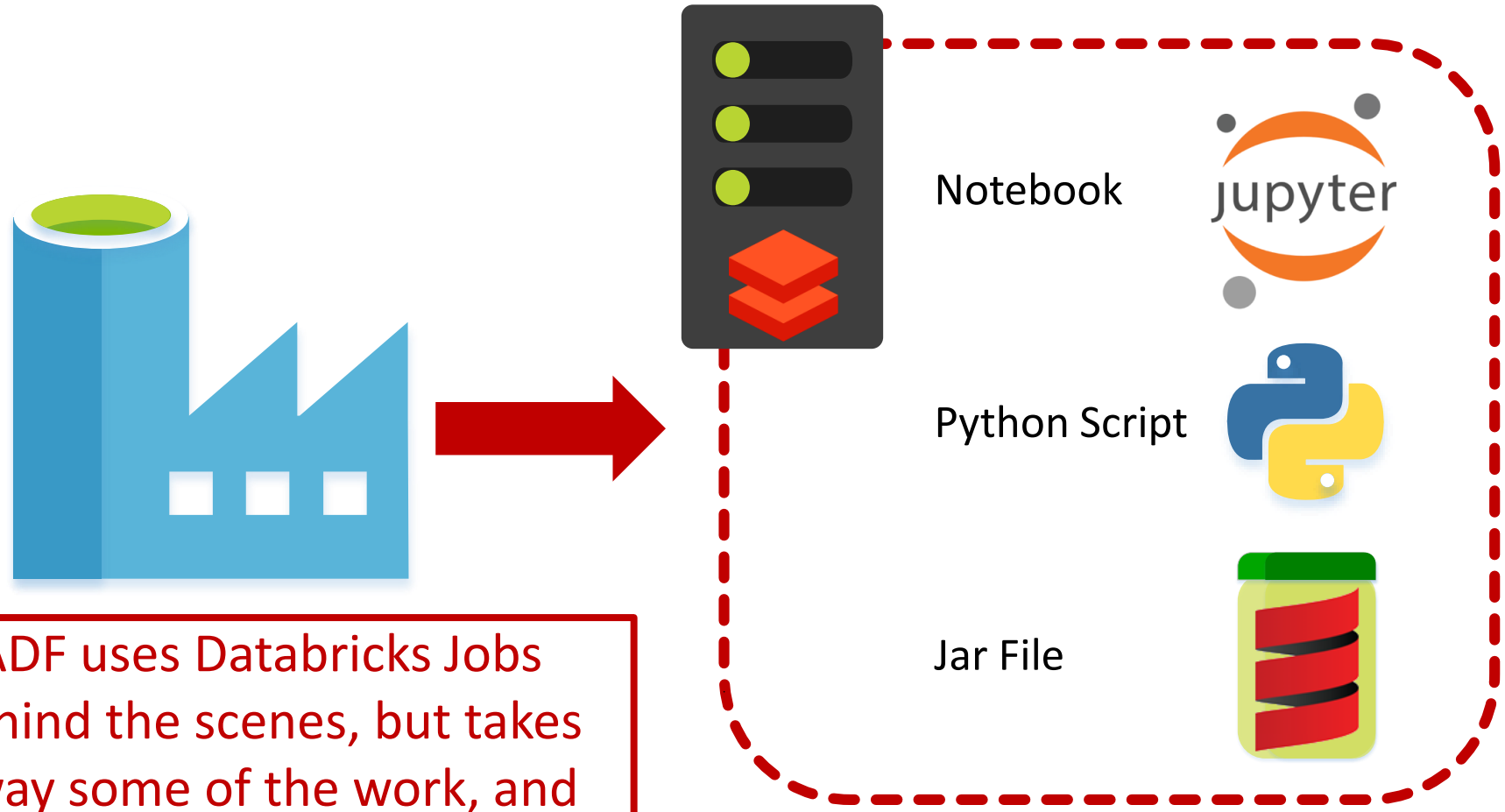
adatis

Orchestration



adatis

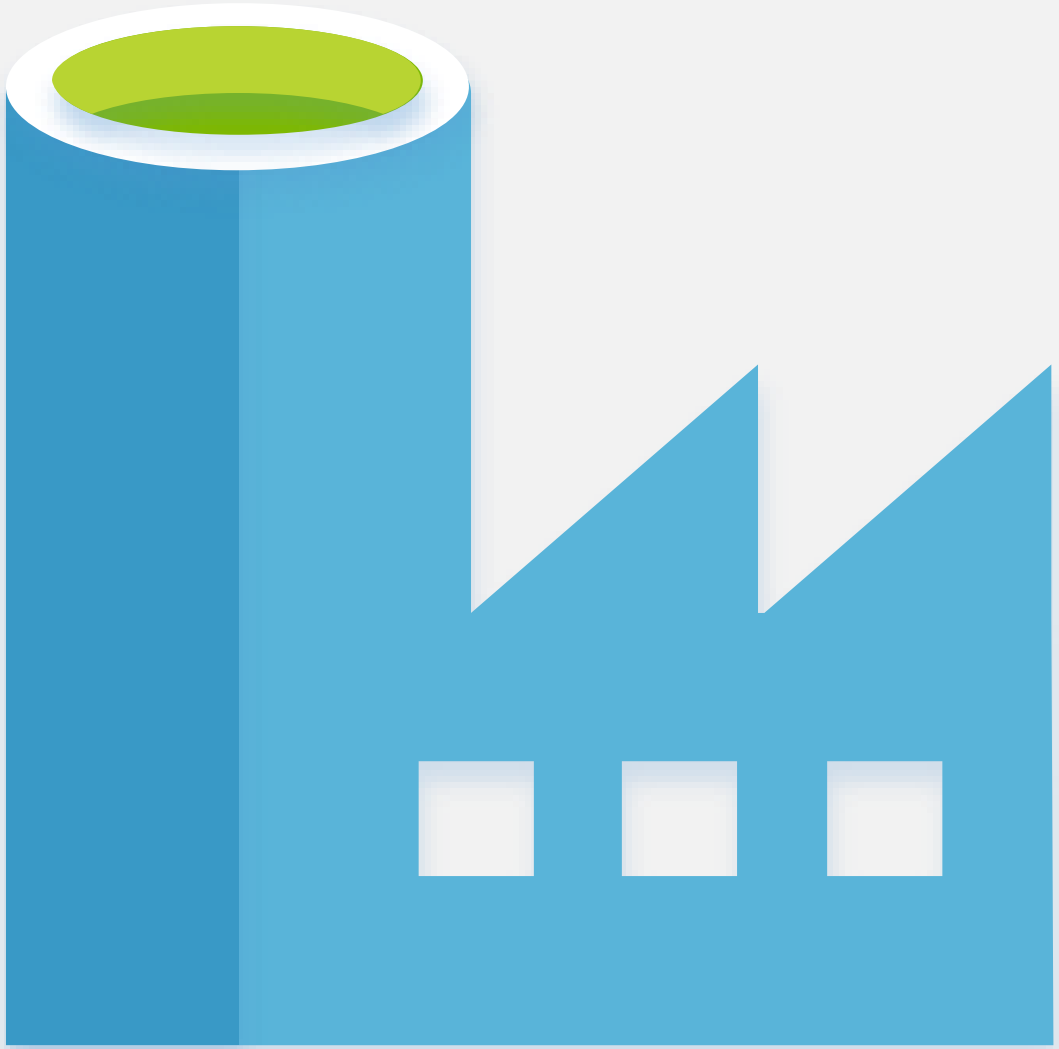
Azure Data Factory



ADF uses Databricks Jobs behind the scenes, but takes away some of the work, and means you can orchestrate your notebooks with other tasks!

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

adatis

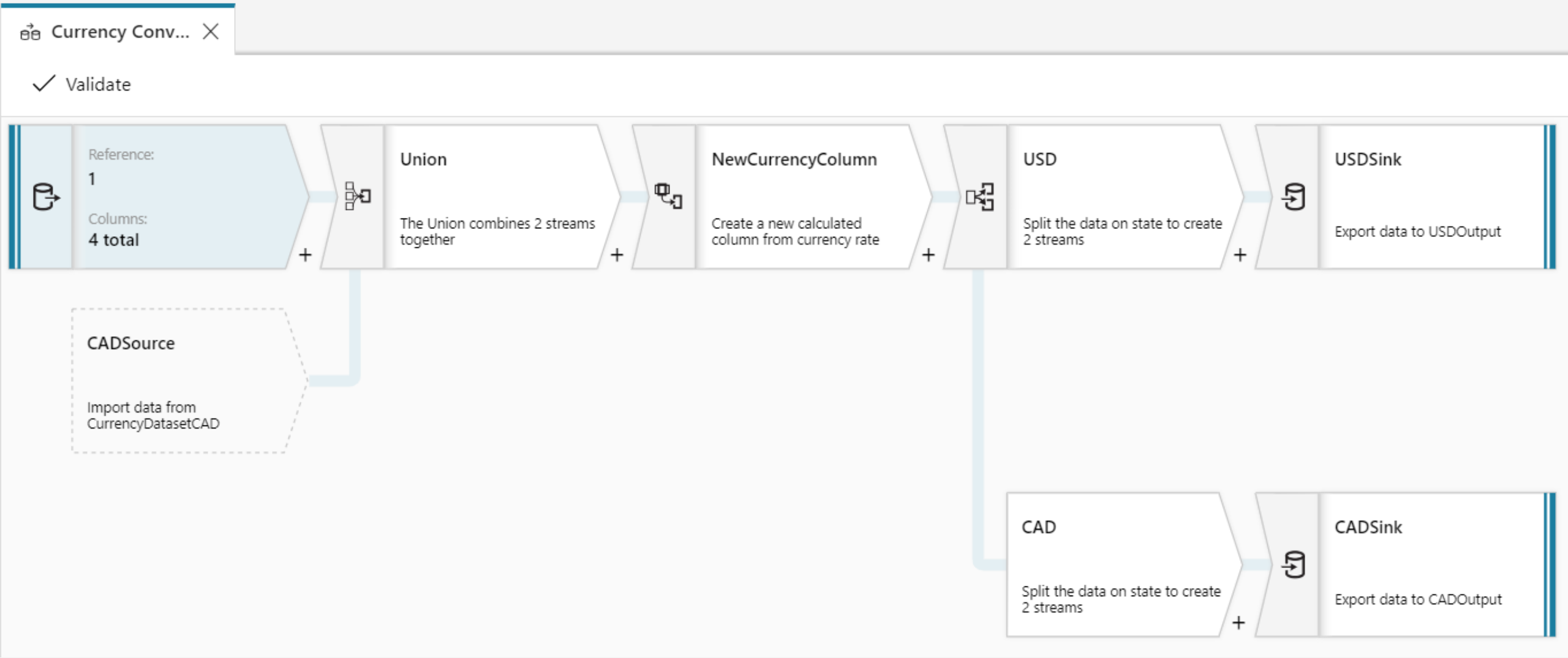


Azure Data Factory

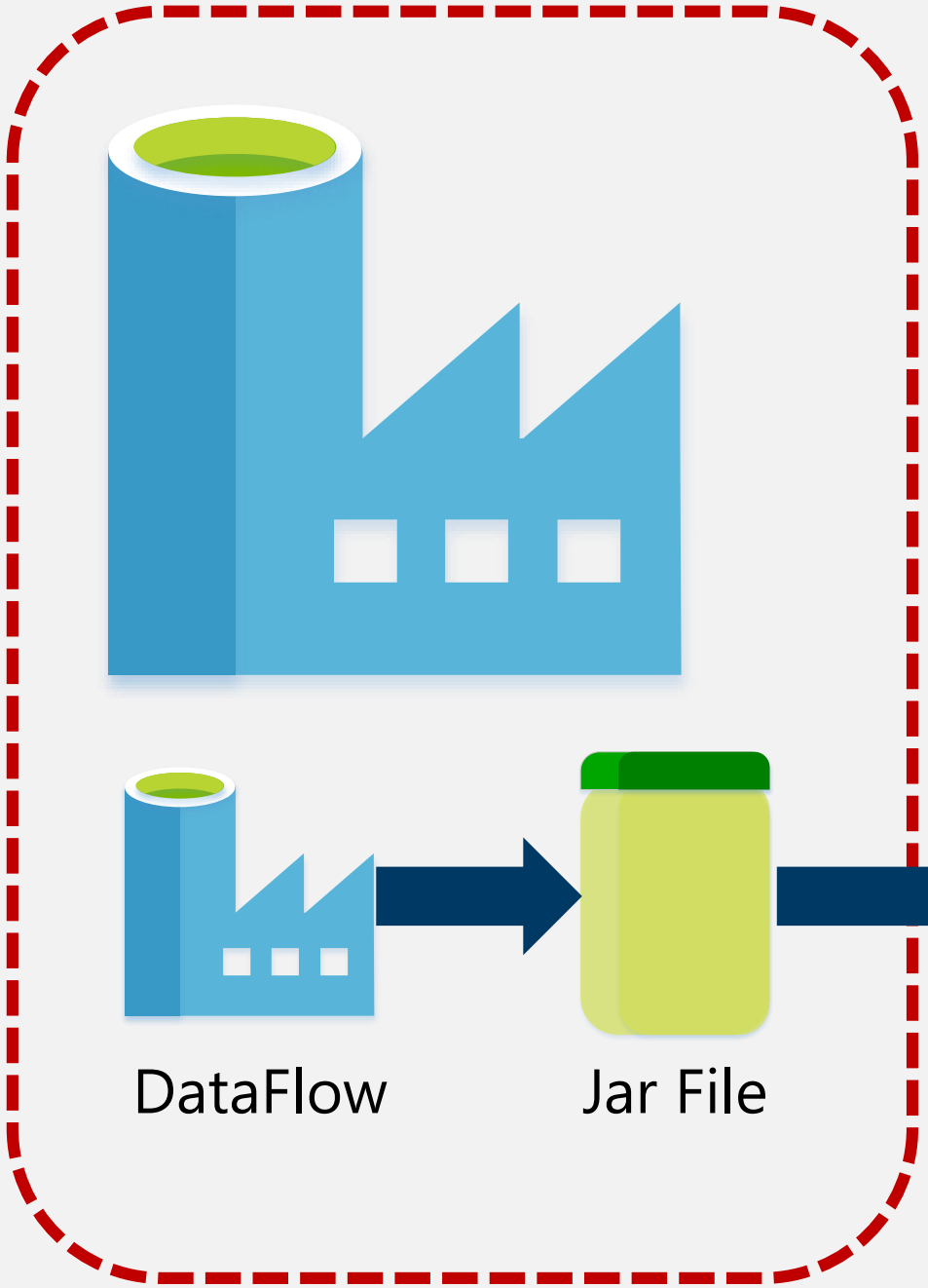
Mapping Data Flows

adatis

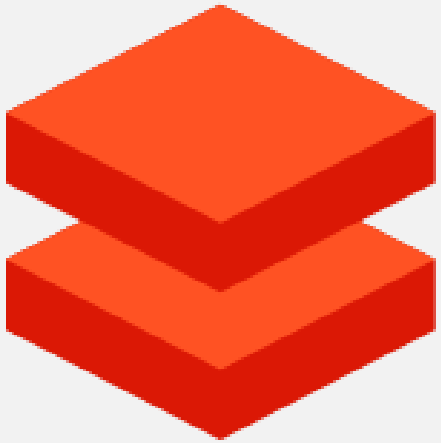
New Data Factory DataFlows can write Databricks processing packages for you!!



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