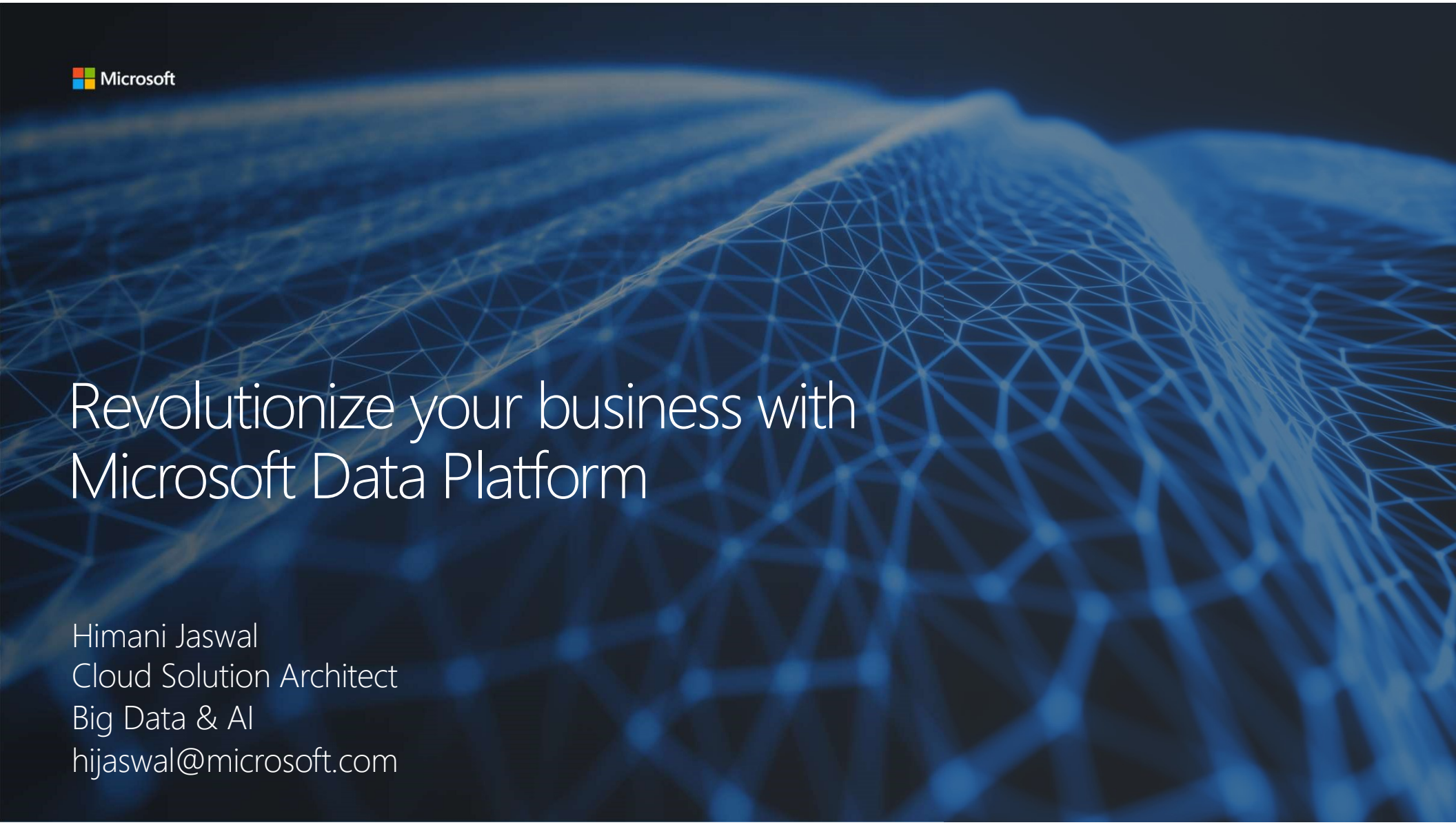


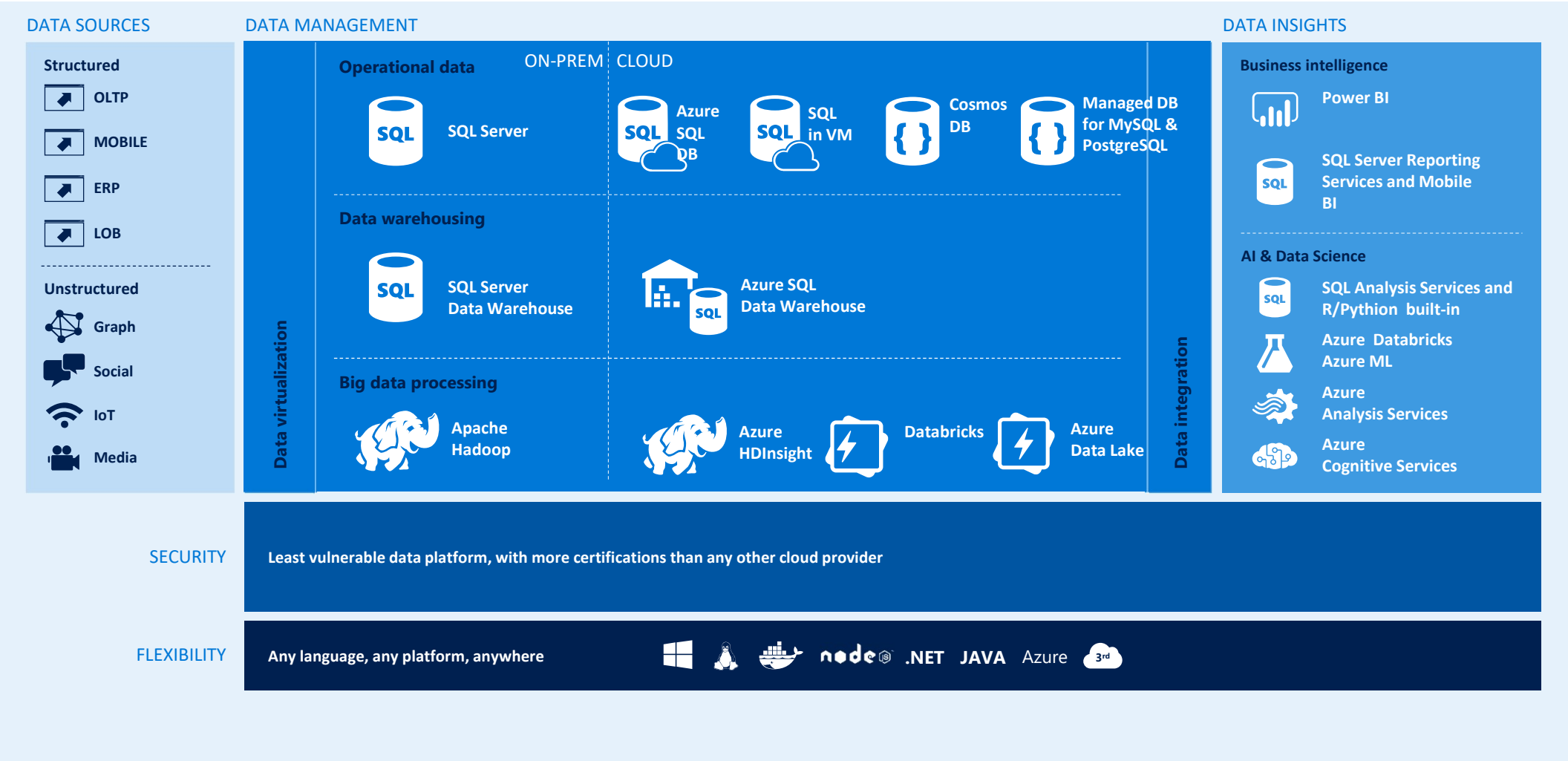
Revolutionize your business with Microsoft Data Platform

Himani Jaswal
Cloud Solution Architect
Big Data & AI
hijaswal@microsoft.com

Agenda

- 8:30-9:00-Kickoff
- 9:00-9:45-Overview of Data platform and ADF concepts
- 9:45-11:00-Lab 1 - Setting up environment, Q&A
- 11:00-11:15-Break
- 11:15-12:00-Azure Datawarehouse concepts
- 12:00-1:00-Lunch
- 1:00-3:00-Lab2 - End to End pipeline batch pipeline, Q&A
- 3:00-3:15- Break
- 3:15-4:00-Stream Analytics and Databricks Delta
- 4:00-5:00-Lab 3, Q&A

Microsoft View on modern data estate



Recent Gartner Magic Quadrants

Microsoft is the only vendor presented in all 3 leader magic quadrants

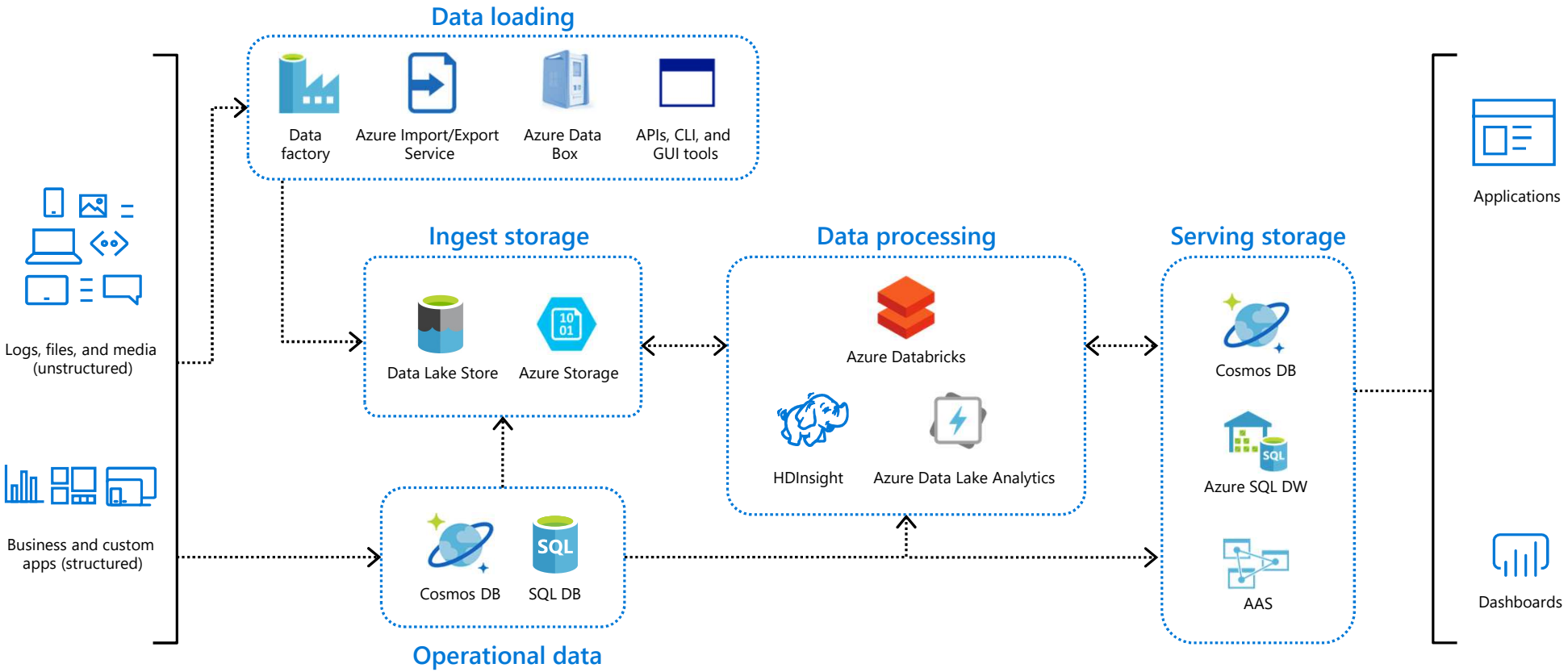
Operational Data

BI and Analytics Platform

Data Management for Analytics (DW)

Analytics in Azure

Loading and preparing data for analysis with a data warehouse



Introduction to Azure Data Factory

Azure Data Factory

A fully-managed data integration service in the cloud

PRODUCTIVE

- ✓ Drag & Drop UI
- ✓ Codeless Data Movement

HYBRID

- ✓ Orchestrate where your data lives
- ✓ Lift SSIS packages to Azure

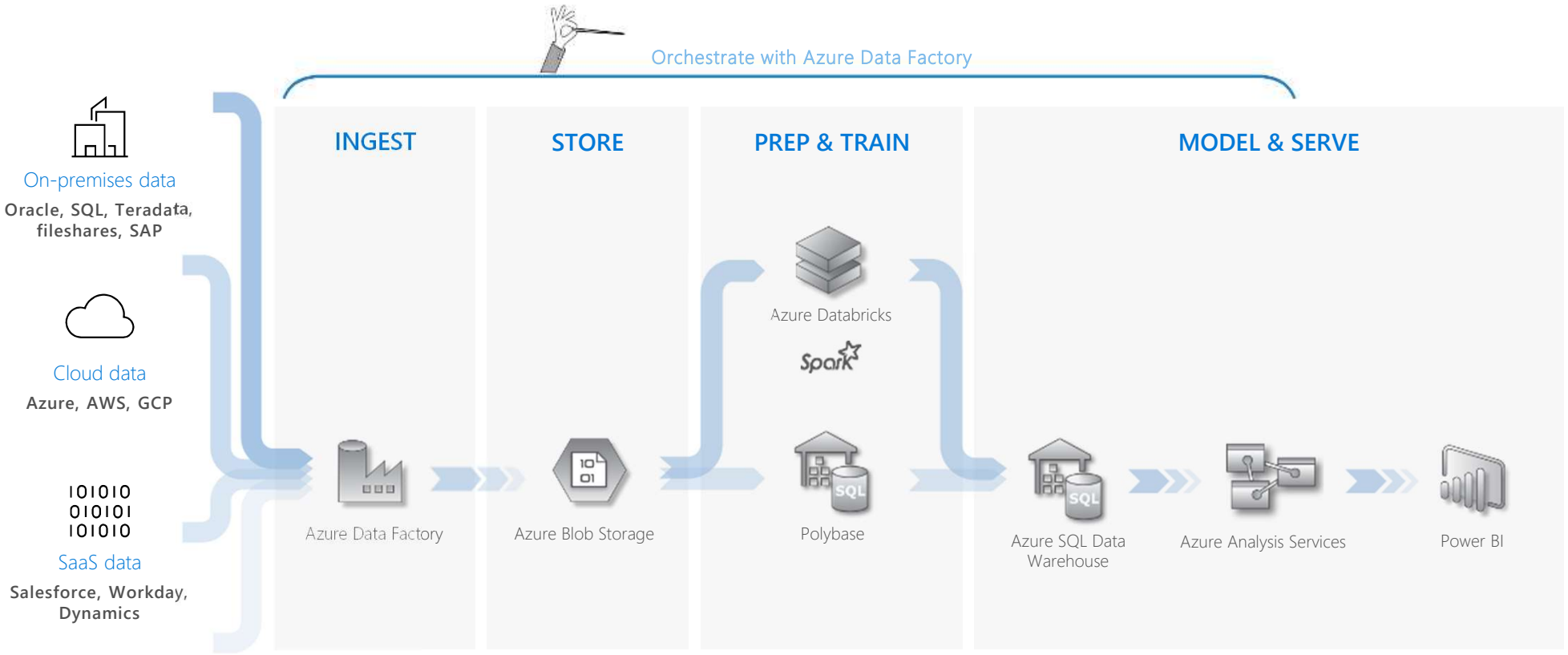
SCALABLE

- ✓ Serverless scalability with no infrastructure to manage

TRUSTED

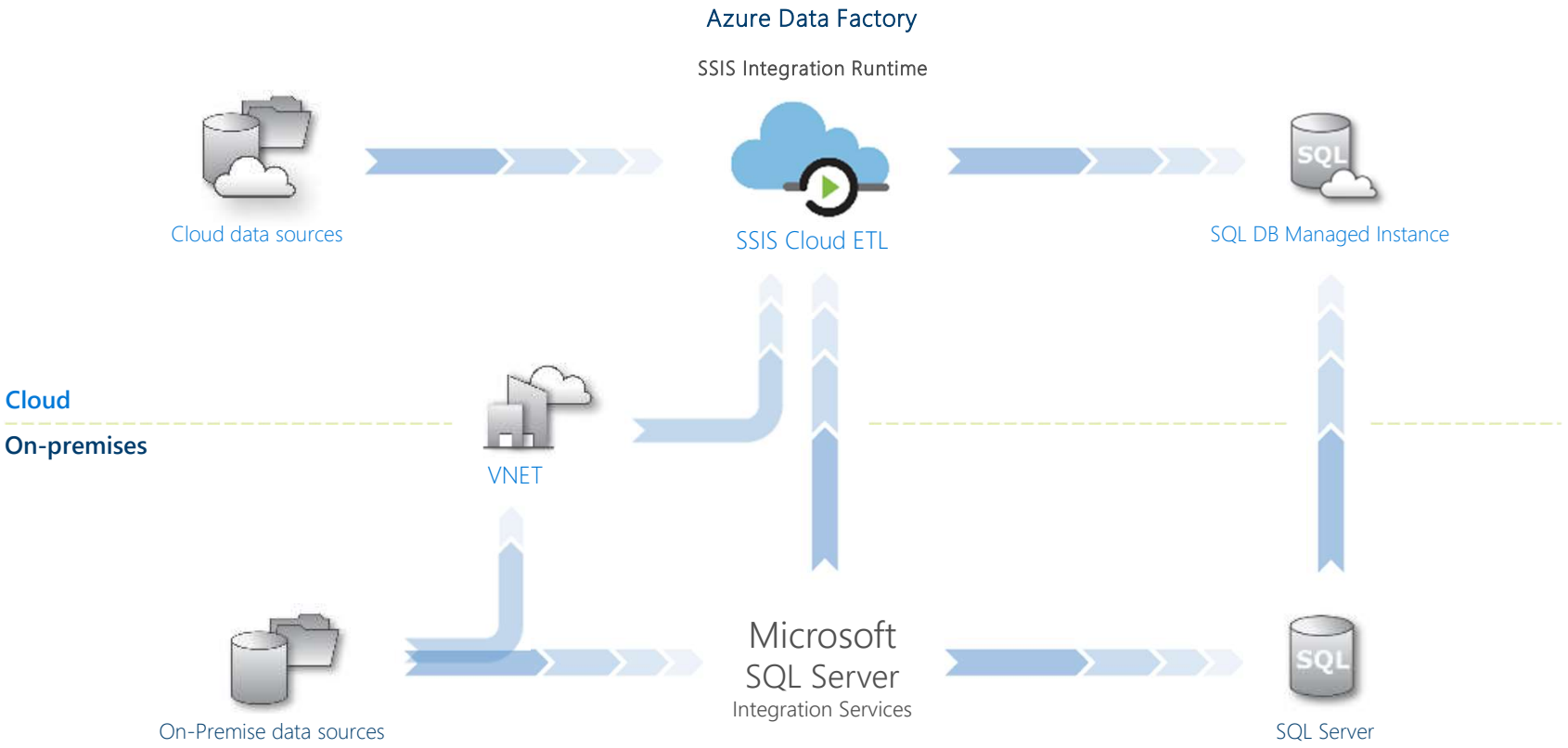
- ✓ Certified compliant Data Movement

Modernize your enterprise data warehouse at scale

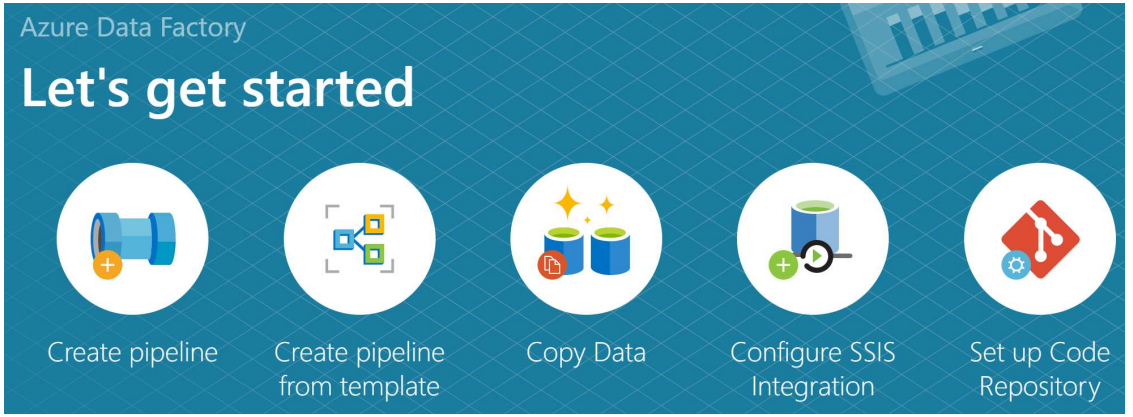


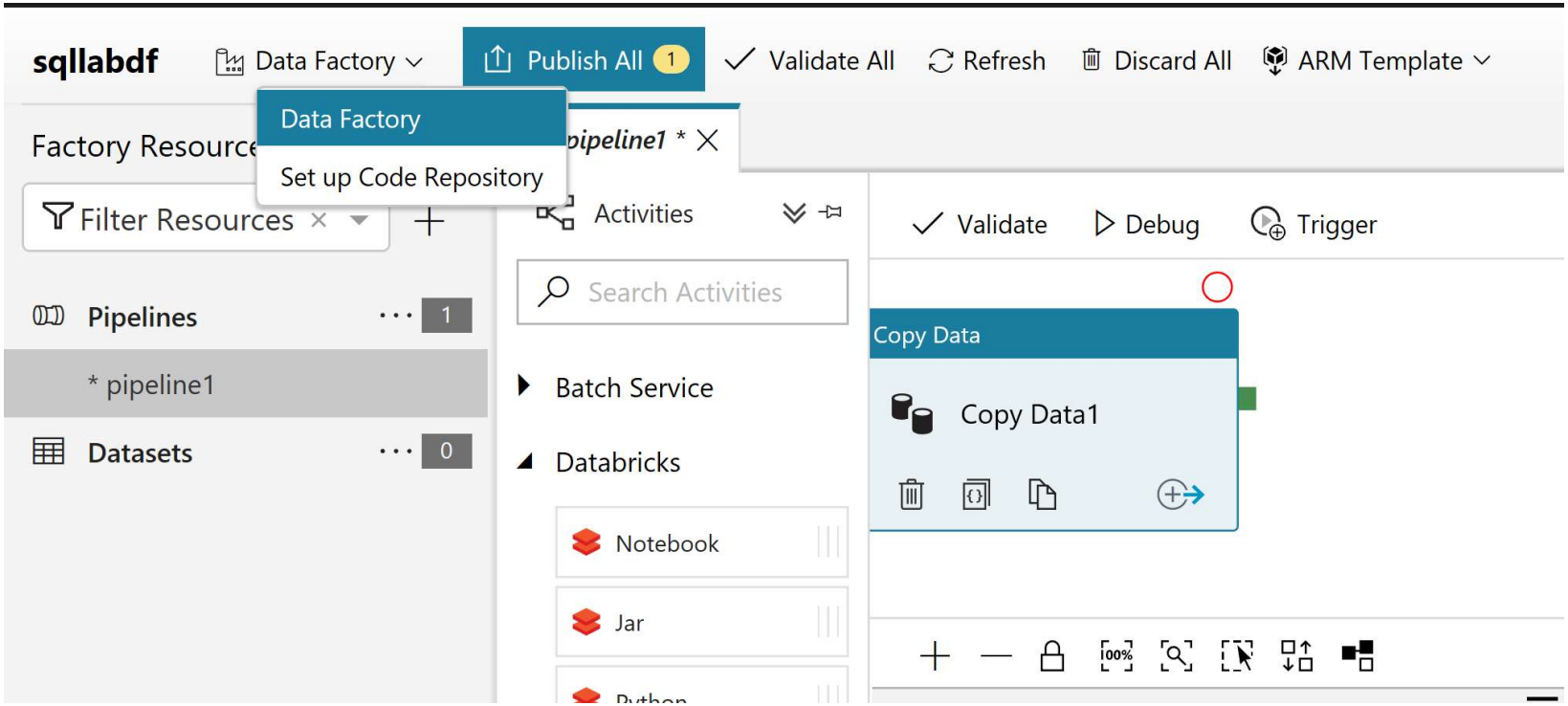
Microsoft Azure also supports other **Big Data** services like **Azure HDInsight**, **Azure SQL Database** and **Azure Data Lake** to allow customers to tailor the above architecture to meet their unique needs.

Lift your SQL Server Integration Services (SSIS) packages to Azure

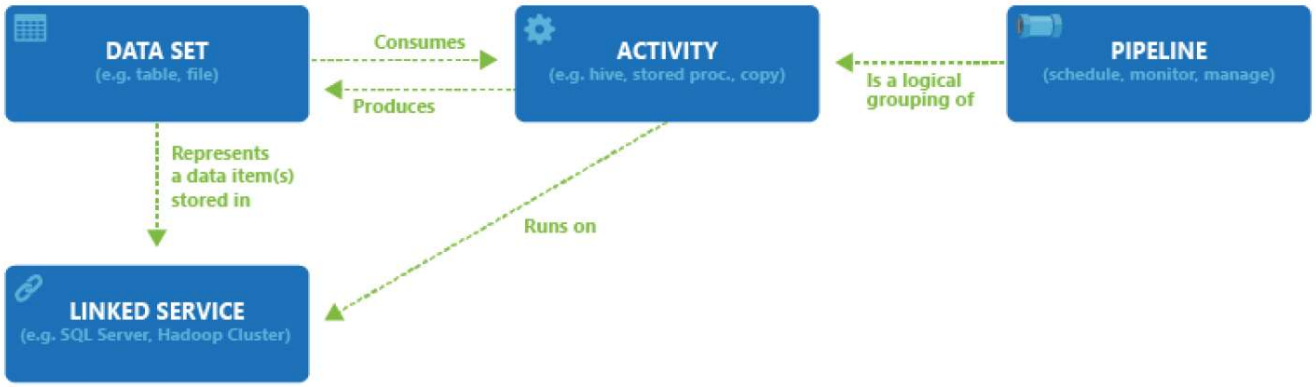


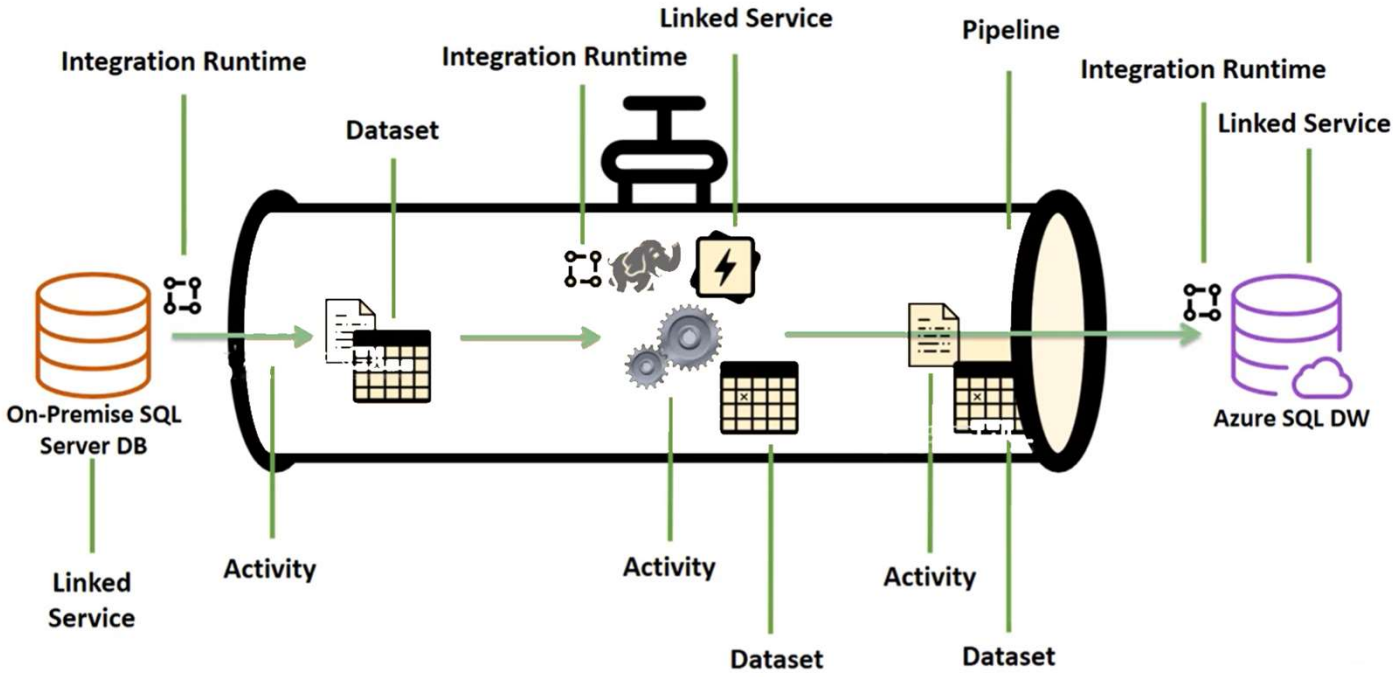
Azure Data Factory





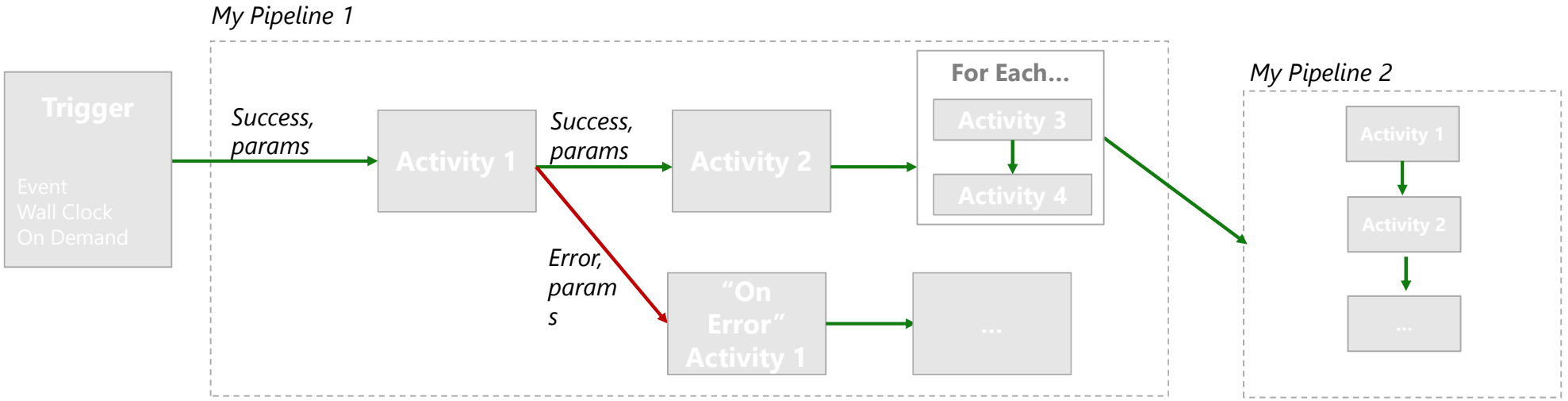
Azure Data Factory

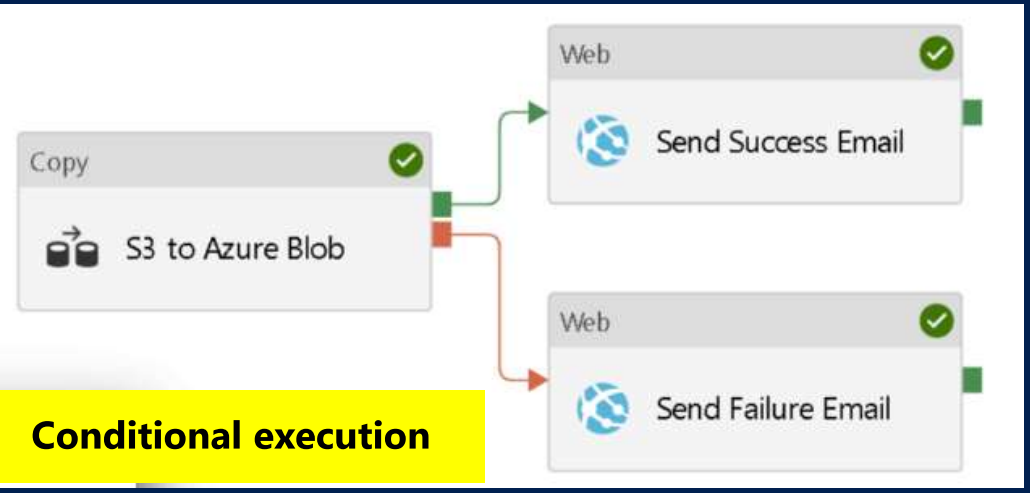




Control Flow Introduced in Azure Data Factory

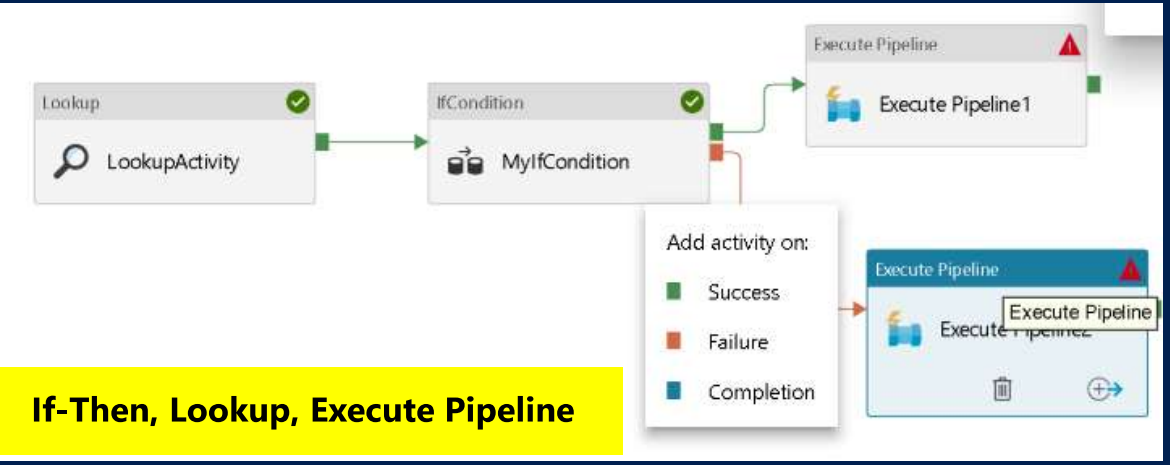
Coordinate pipeline activities into finite execution steps to enable looping, conditionals and chaining while separating data transformations into individual data flows



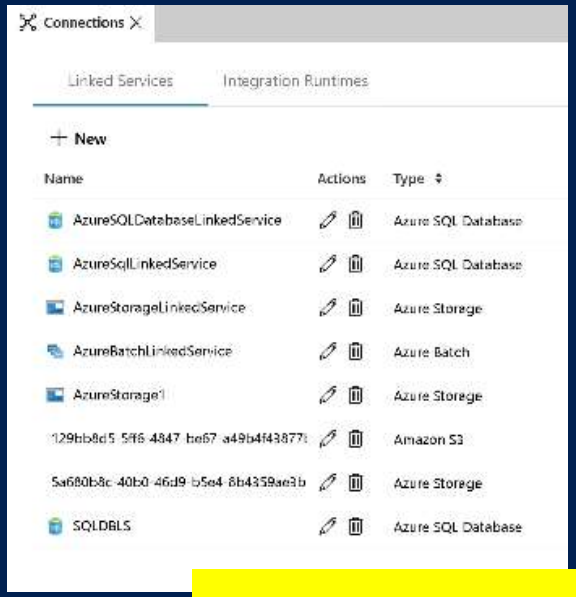


Conditional execution

Incremental Delta Data Copy

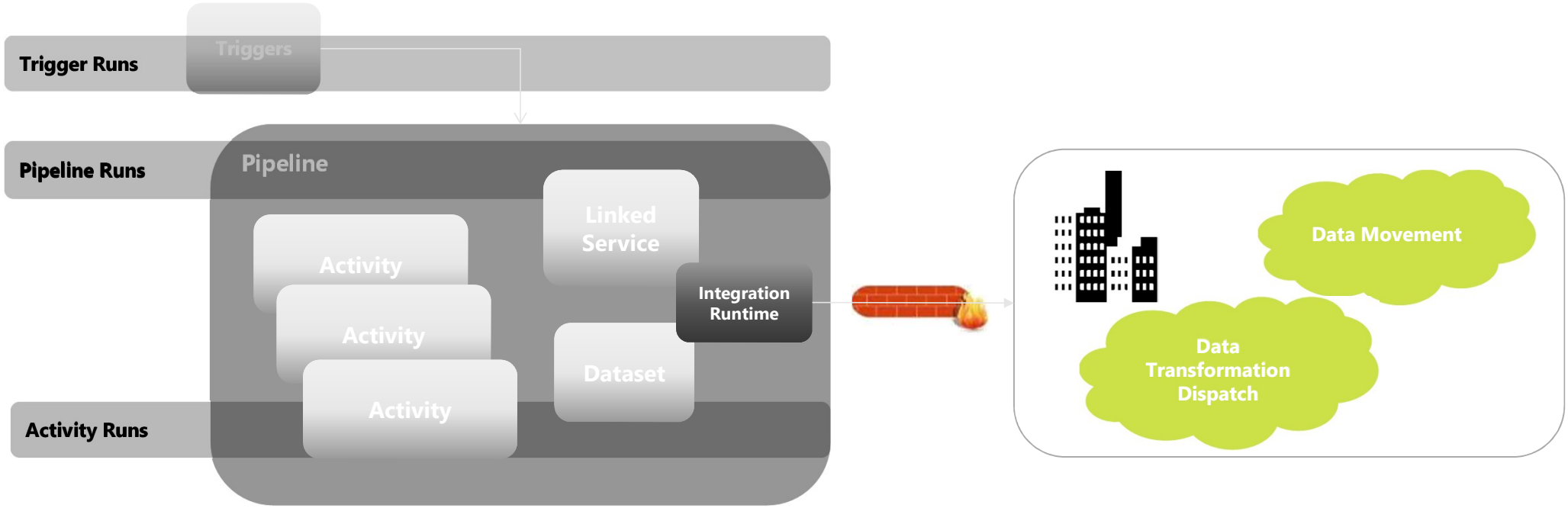


If-Then, Lookup, Execute Pipeline

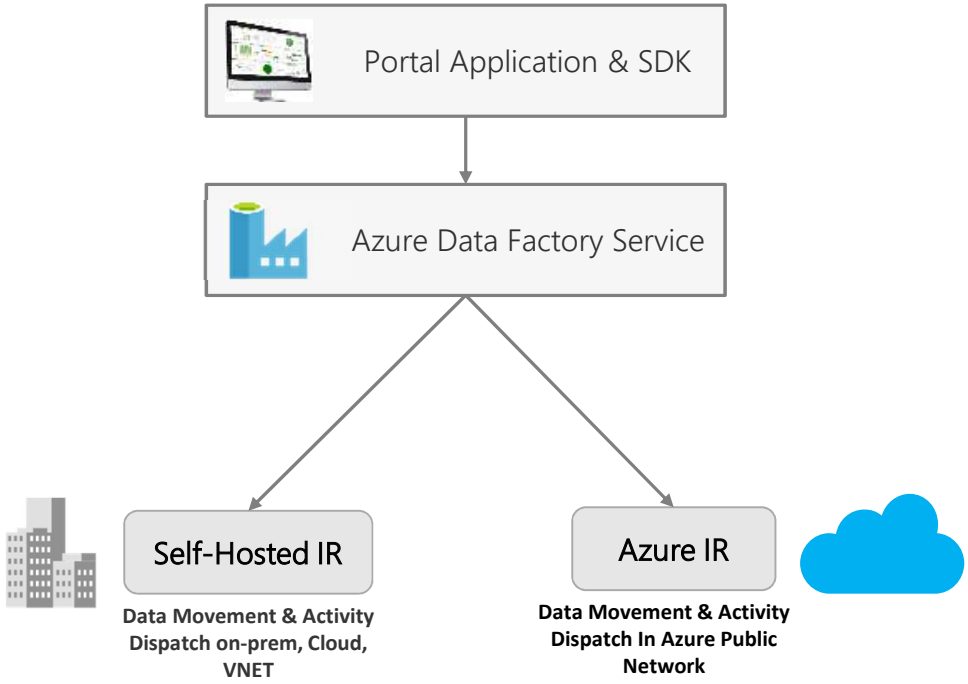


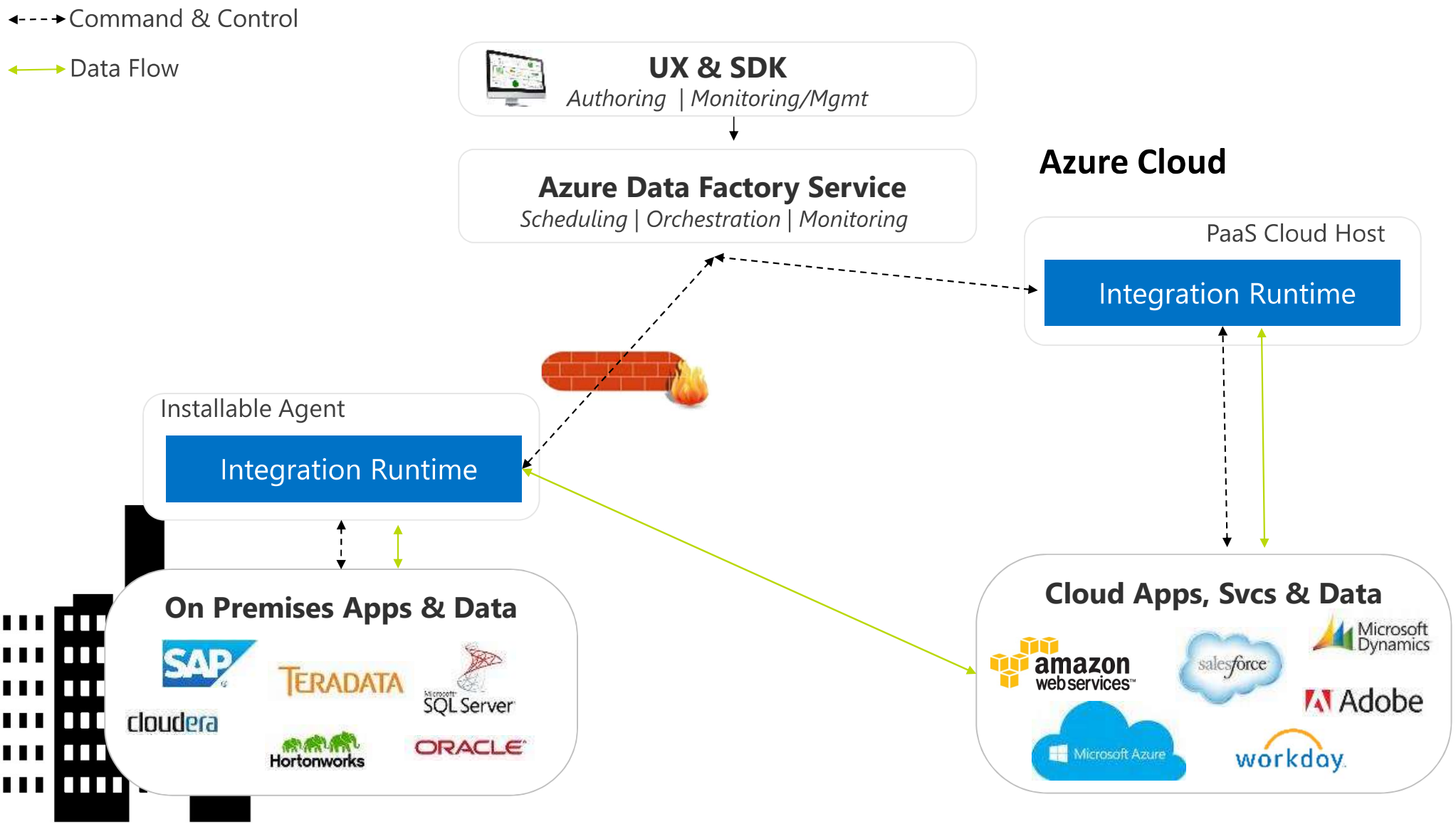
Connection Managers

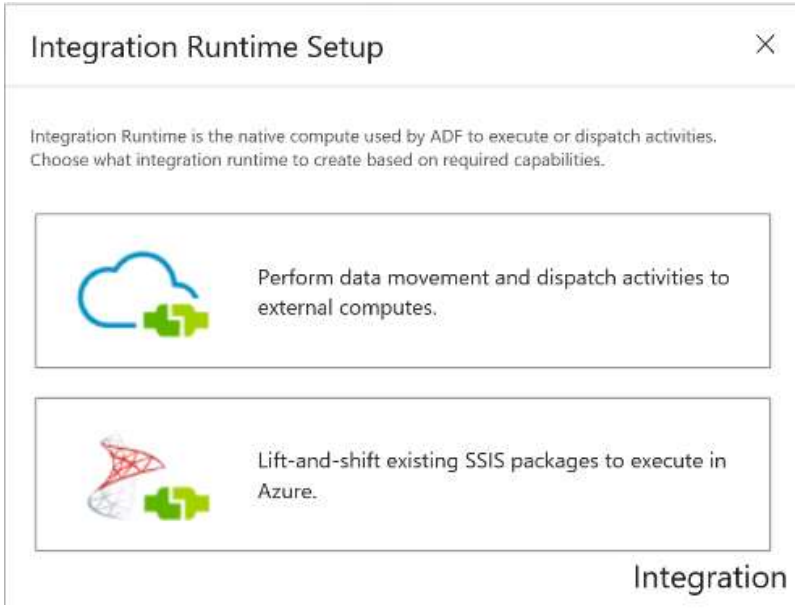
Azure Data Factory Updated Flexible Application Model

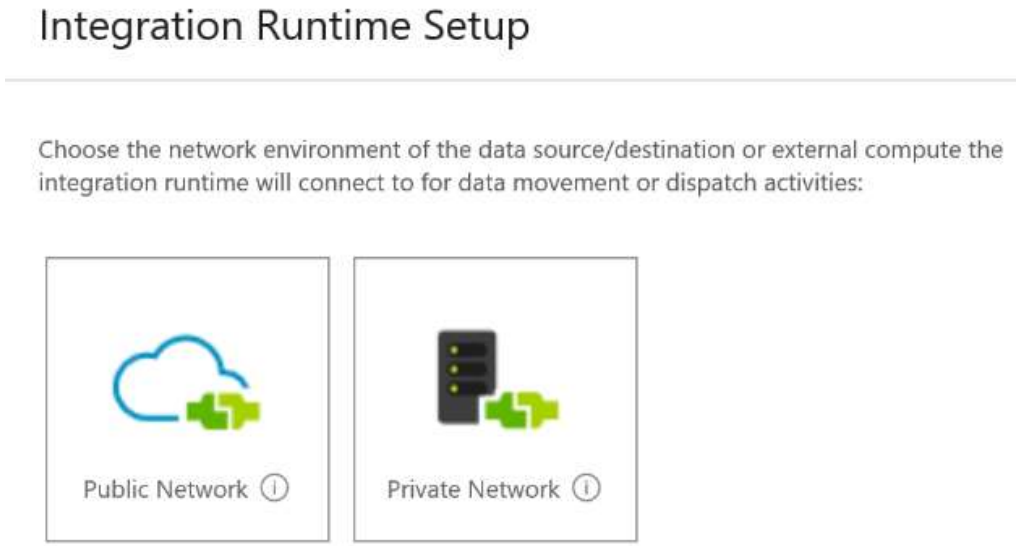


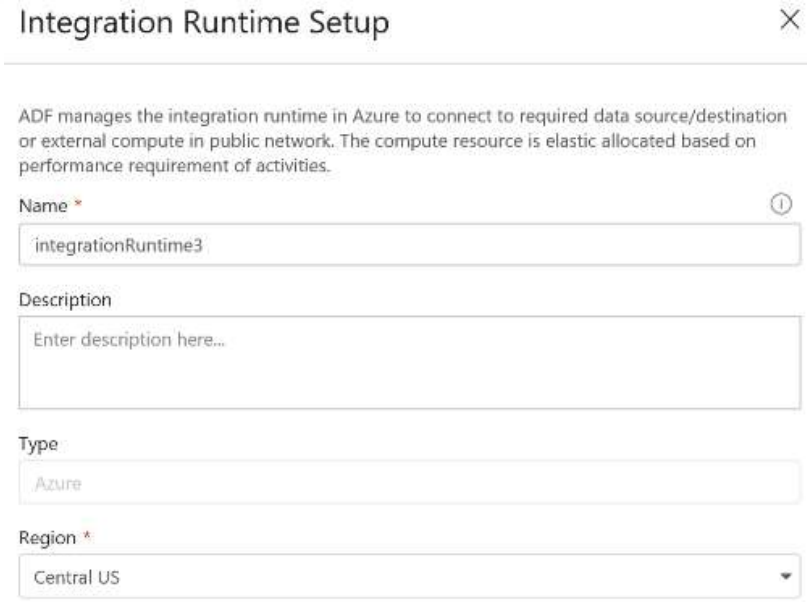
ADF Integration Runtime (IR)




Command & Control
Data Flow
UX & SDK
Authoring | Monitoring/Mgmt
Azure Data Factory Service
Scheduling | Orchestration | Monitoring
Azure Cloud
PaaS Cloud Host
Integration Runtime
Installable Agent
Integration Runtime
On Premises Apps & Data
SAP
TERADATA
SQL Server
cloudera
Hortonworks
ORACLE
Cloud Apps, Svcs & Data
amazon web services
salesforce
Microsoft Dynamics
Adobe
Microsoft Azure
workday


Integration Runtime Setup
Integration Runtime is the native compute used by ADF to execute or dispatch activities. Choose what integration runtime to create based on required capabilities.
Perform data movement and dispatch activities to external computes.
Lift-and-shift existing SSIS packages to execute in Azure.


Integration Runtime Setup
Choose the network environment of the data source/destination or external compute the integration runtime will connect to for data movement or dispatch activities:
Public Network
Private Network


Integration Runtime Setup
ADF manages the integration runtime in Azure to connect to required data source/destination or external compute in public network. The compute resource is elastic allocated based on performance requirement of activities.
Name *
integrationRuntime3
Description
Enter description here...
Type
Azure
Region *
Central US

New ADF V2 Concepts

Concept	Description	Sample
Control Flow	Orchestration of pipeline activities that includes chaining activities in a sequence, branching, conditional branching based on an expression, parameters that can be defined at the pipeline level and arguments passed while invoking the pipeline on demand or from a trigger. Also includes custom state passing and looping containers, I.e. For-each, Do-Until iterators.	{ "name":"MyForEachActivityName", "type":"ForEach", "typeProperties":{ "isSequential":"true", "items":"@pipeline().parameters.mySinkDatasetFolderPathCollection", "activities":[{ "name":"MyCopyActivity", "type":"Copy", "typeProperties": …
Runs	A Run is an instance of the pipeline execution. Pipeline Runs are typically instantiated by passing the arguments to the parameters defined in the Pipelines. The arguments can be passed manually or properties created by the Triggers.	POST https://management.azure.com/subscriptions/<subId>/resourceGroups/<resourceGroupName>/providers/Microsoft.DataFactory/factories/<dataFactoryName>/pipelines/<pipelineName>/createRun?api-version=2017-03-01-preview
Activity Logs	Every activity execution in a pipeline generates activity start and activity end logs event	
Integration Runtime	Replaces DMG as a way to move & process data in Azure PaaS Services, self-hosted or on prem or IaaS Works with VNETs Enables SSIS package execution	Set-AzureRmDataFactoryV2IntegrationRuntime -Name $integrationRuntimeName -Type SelfHosted

Scheduling	Flexible Scheduling
	Wall-clock scheduling
	Event-based triggers
		"type": "ScheduleTrigger",
		"typeProperties": {
		"recurrence": {
		"frequency": <<Minute, Hour, Day, Week, Year>>,
		"interval": <<int>>, // optional, how often to fire (default to 1)
		"startTime": <<datetime>>,
		"endTime": <<datetime>>,
		"timeZone": <<default UTC>>
		"schedule": { // optional (advanced scheduling specifics)
		"hours": [<<0-24>>],
		"weekDays": ": [<<Monday-Sunday>>],
		"minutes": [<<0-60>>],
		"monthDays": [<<1-31>>],
		"monthlyOccurences": [
		{
		"day": <<Monday-Sunday>>,
		"occurrence": <<1-5>>

New ADF V2 Concepts

Concept	Description	Sample
On-Demand Execution	Instantiate a pipeline by passing arguments as parameters defined in a pipeline and execute from script / REST / API.	Invoke-AzureRmDataFactoryV2PipelineRun -DataFactory $df -PipelineName "Adfv2QuickStartPipeline" -ParameterFile .\PipelineParameters.json
Parameters	Name-value pairs defined in the pipeline. Arguments for the defined parameters are passed during execution from the run context created by a Trigger or pipeline executed manually. Activities within the pipeline consume the parameter values. A **Dataset** is a strongly typed parameter and a reusable/referenceable entity. An activity can reference datasets and can consume the properties defined in the Dataset definition A **Linked Service** is also a strongly typed parameter containing the connection information to either a data store or a compute environment. It is also a reusable/referenceable entity.	Accessing parameters of other activities Using expressions @parameters("{name of parameter}") @activity("{Name of Activity}").output.RowsCopied

Incremental Data Loading	Leverage parameters and define your high-water mark for delta copy while moving dimension or reference tables from a relational store either on premises or in the cloud to load the data into the lake	name": "LookupWaterMarkActivity",
		"type": "Lookup",
		"typeProperties": {
		"source": {
		"type": "SqlSource",
		"sqlReaderQuery": "select * from watermarktable"
		}

On-Demand Spark	Support for on-demand HDI Spark clusters, similar to on-demand Hadoop activities in V1	"type": "HDInsightOnDemand", "typeProperties": { "clusterSize": 2, "clusterType": "spark", "timeToLive": "00:15:00",
SSIS Runtime	Lift & shift, deploy, manage, monitor SSIS packages in the cloud with SSIS Azure IR Service in Azure Data Factory	Start-AzureRmDataFactoryV2IntegrationRuntime -DataFactoryName $DataFactoryName -Name
Code-free UI	Build end-to-end data pipeline solutions for ADF without writing code or JSON	

Graphical user interface tools

A side-by-side comparison of the capabilities and features

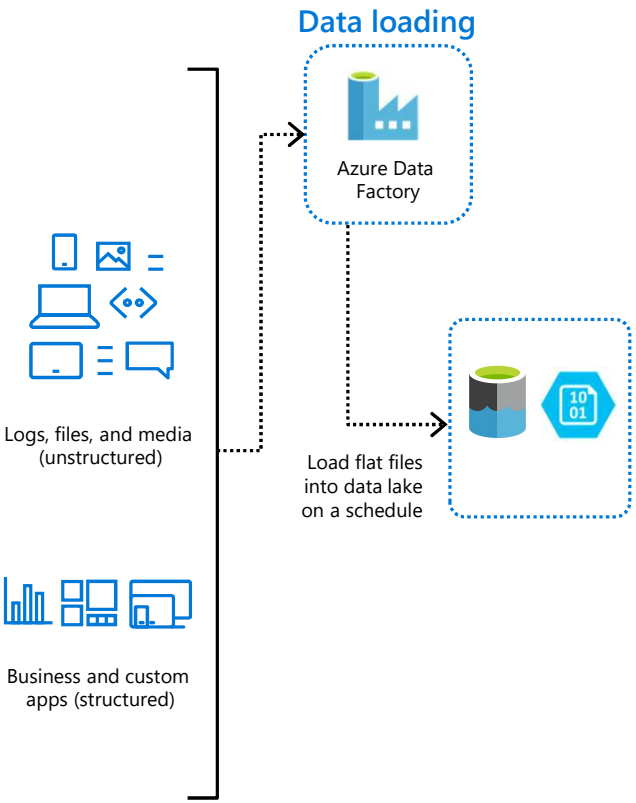
	Azure Storage Explorer	Azure Portal*	Azure Data Factory
Copy to and from relational database	No	No	Yes
Copy to blob storage	Yes	No	Yes
Copy from blob storage	Yes	No	Yes
Copy to data lake store	No	No	Yes
Copy from data lake store	No	No	Yes
Upload to blob storage	Yes	Yes	Yes
Upload to data lake store	Yes	Yes	Yes
Orchestrate data transfers	No	No	Yes
Custom data transformations	No	No	Yes

* Azure Portal in this case means using the web-based exploration tools for blob storage and data lake store. This excludes using the portal for other services, such as Azure Data Factory.

Introduction to Azure Storage

Storage in Azure

Loading data into ingest storage



File storage

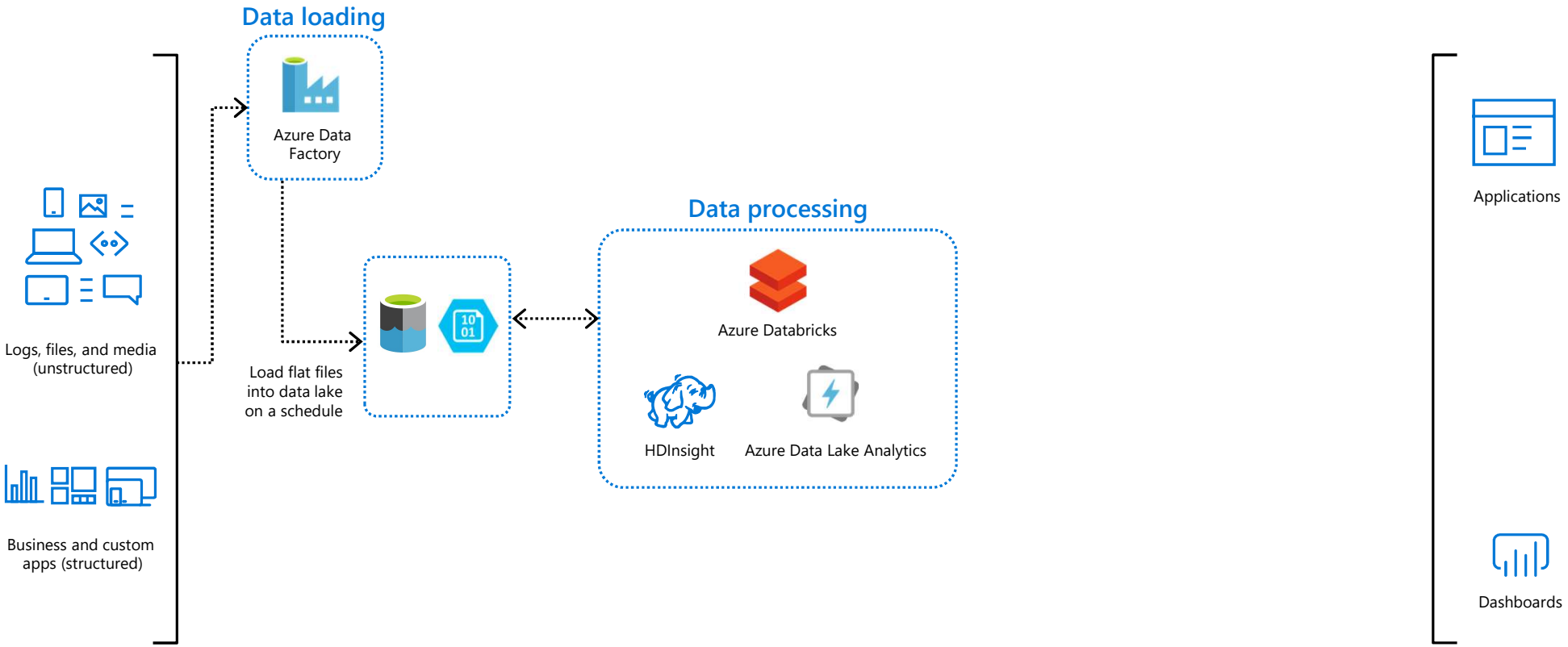
A side-by-side comparison of the capabilities and features

	Azure Data Lake Store	Azure Blob Storage containers
Purpose	Optimized storage for Big Data analytics workloads	General purpose object store for a wide variety of storage scenarios
Structure	Hierarchical file system	Object store with flat namespace
API	REST API over HTTPS	REST API over HTTP/HTTPS
Analytics workload performance	Optimized performance for parallel analytics workloads, high throughput and IOPS	Not optimized for analytics workloads
Size limits	No limits on account sizes, file sizes, or number of files	Max 500 TB per account and 4.75 TB per file
Geo-redundancy	Locally-redundant (multiple copies of data in one Azure region)	Locally redundant (LRS), globally redundant (GRS), and read-access globally redundant (RA-GRS). See Azure Storage replication for more information
Service state	Generally available	Generally available
Regional availability	Some regions	All regions

Introduction to Azure Databricks

Processing in Azure

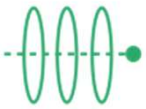
Loading data into ingest storage



Data Engineering system requirement

Data Management


Big Data


Fast Data


Schema Mgmt


Data Consistency


Fast Reads

Platform Management


Resource Estimation


Failure Recovery


Automatic Upgrades


Distributed Framework


Elastic Scalability

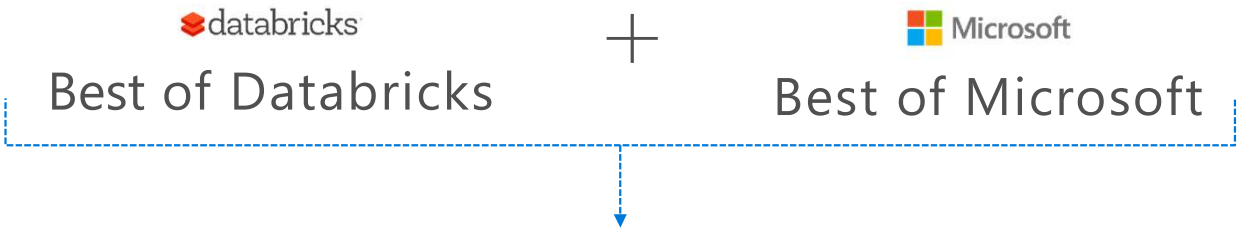

Tooling & Integration

Infrastructure Management


Managed as a Service

What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



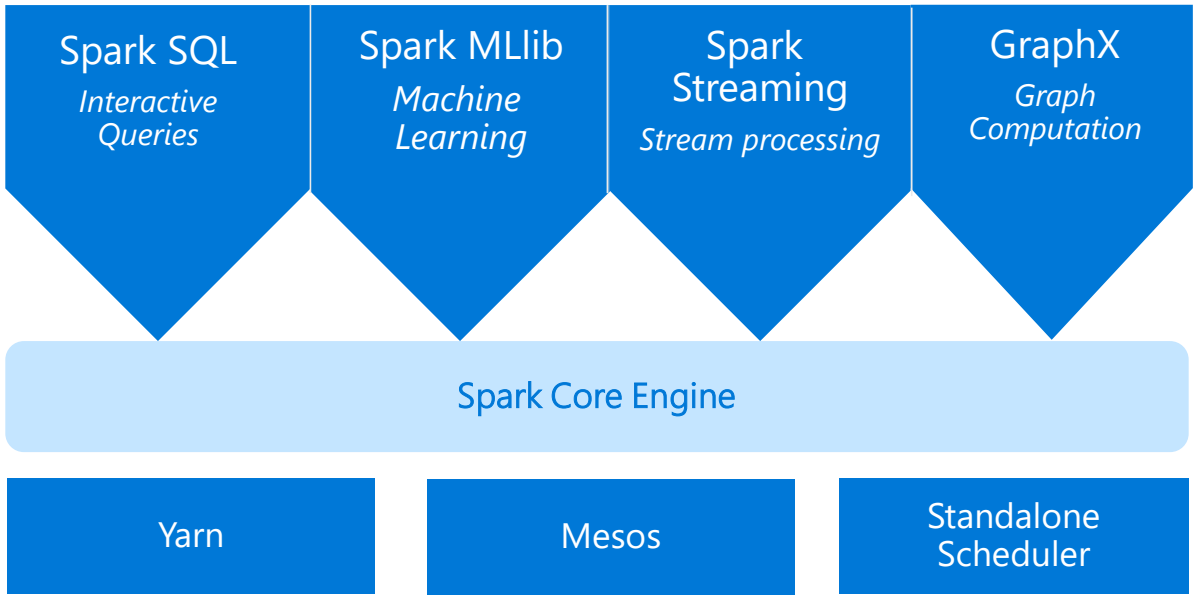
- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

Apache Spark

An unified, open source, parallel, data processing framework for Big Data Analytics

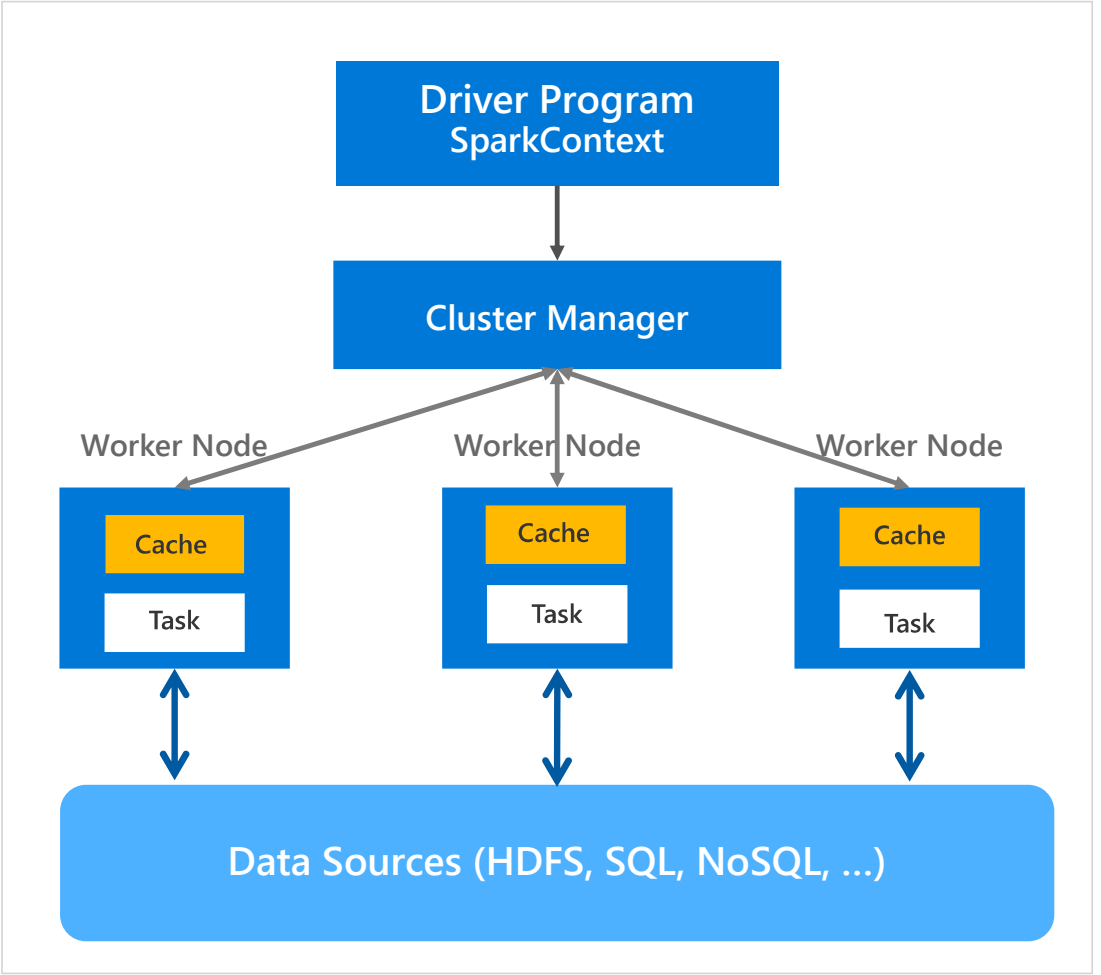
Spark Unifies:

- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing

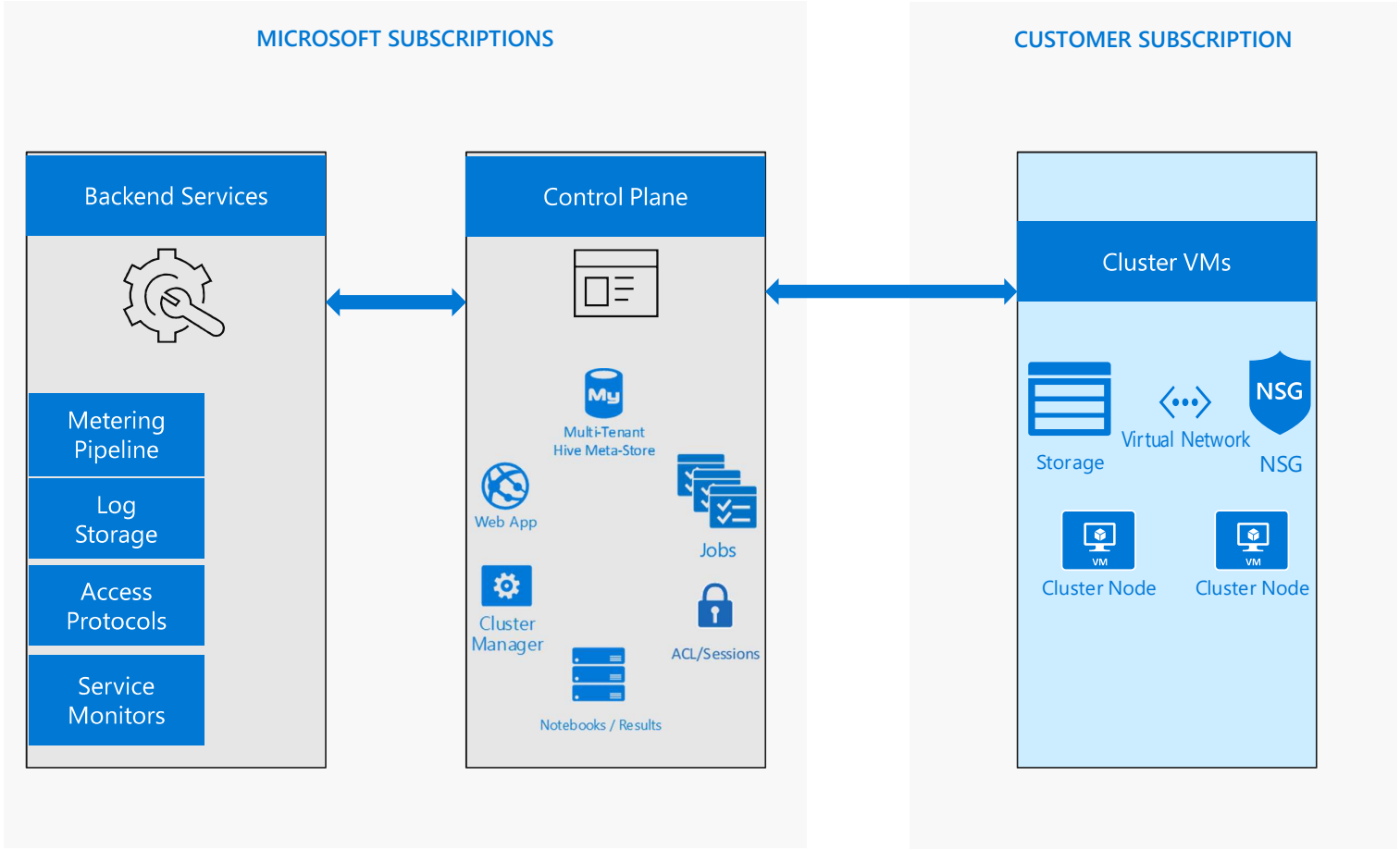


General Spark Cluster Architecture

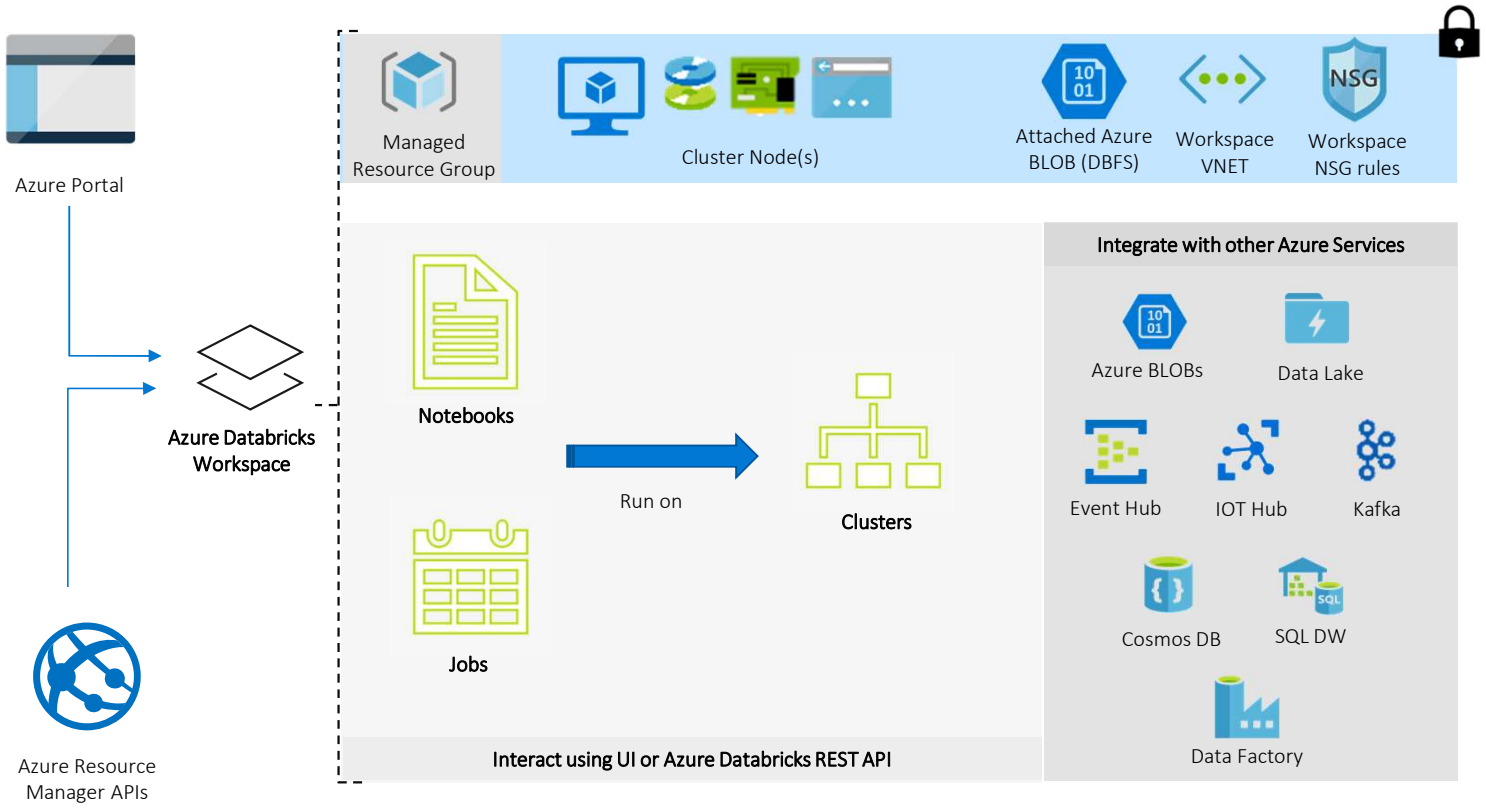
- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).



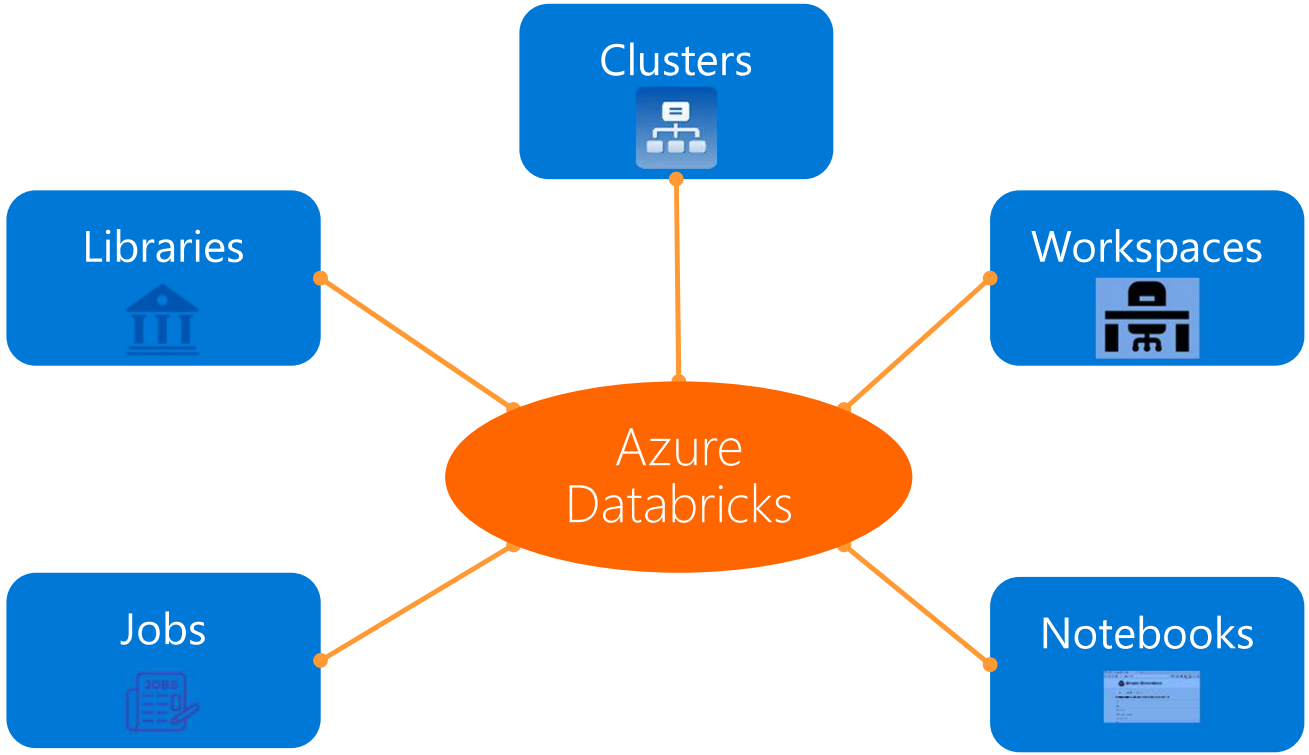
High Level Concepts



Azure Databricks – Customer view

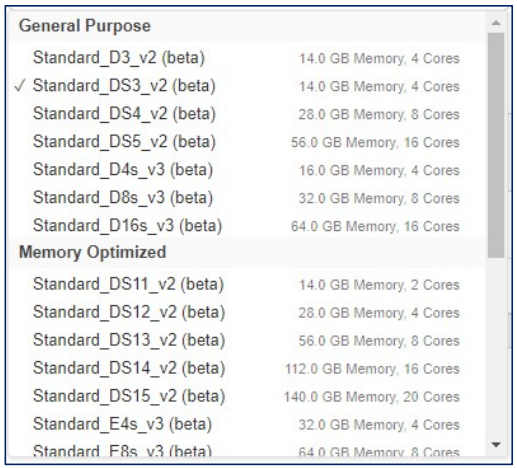


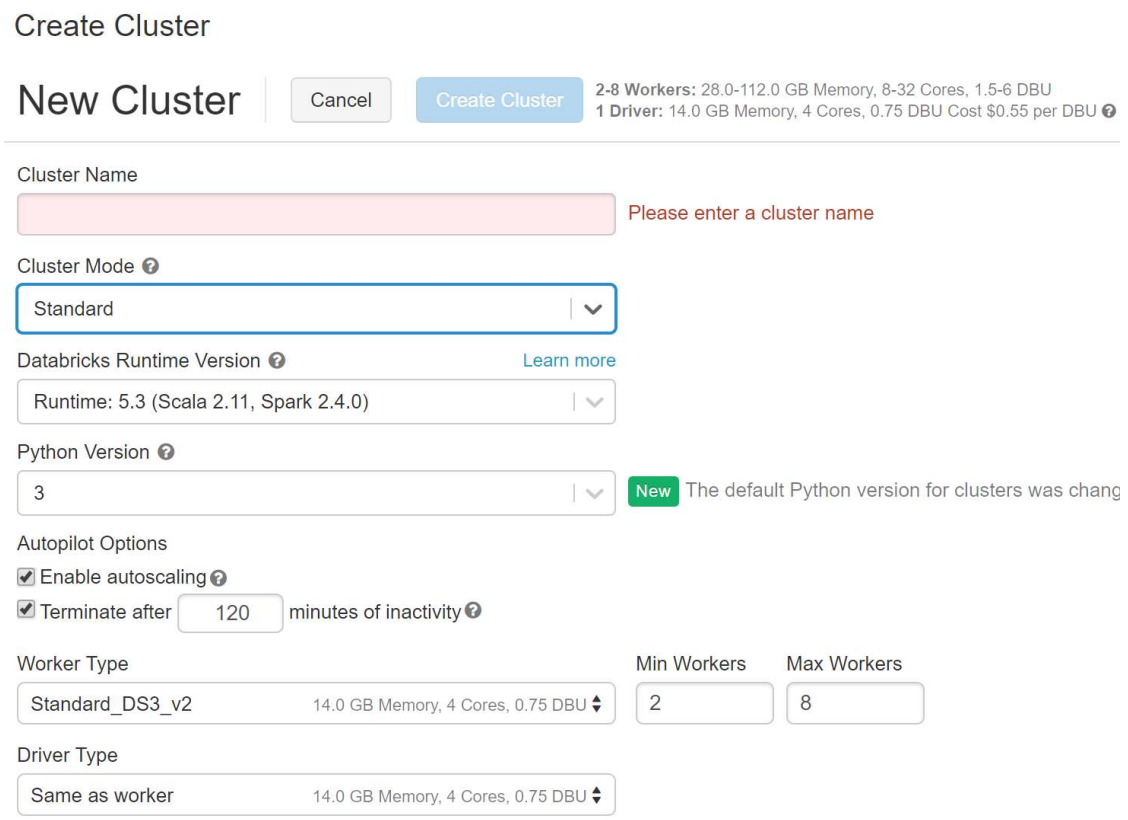
Azure Databricks Core Artifacts



Cluster Creation

- You can create two types of clusters – *Standard* and *High Performance*
- While creating a cluster you can specify:
  - Number of nodes
  - Autoscaling and Auto Termination policy
  - Auto Termination policy
  - Spark Configuration details
  - The Azure VM instance types for the Driver and Worker Nodes





Clusters: Auto Scaling and Auto Termination

Simplifies cluster management and reduces costs by eliminating wastage

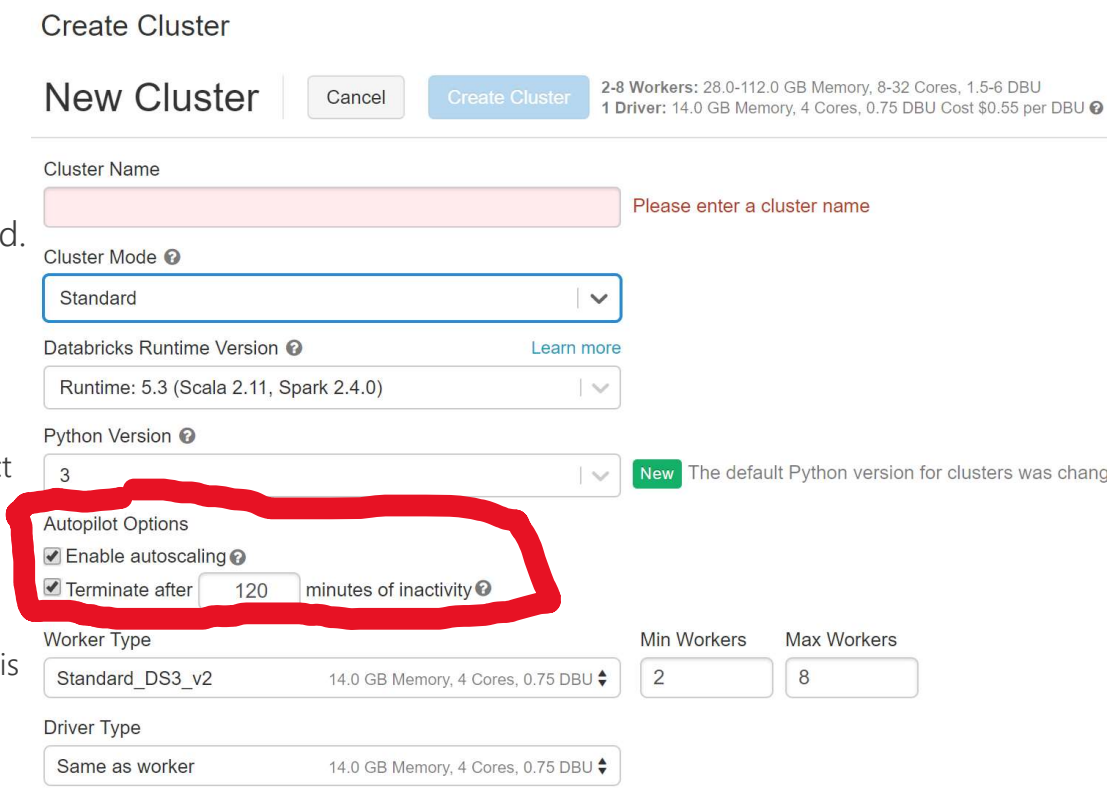
When creating Azure Databricks clusters you can choose Autoscaling and Auto Termination options.

Autoscaling: Just specify the min and max number of clusters. Azure Databricks automatically scales up or down based on load.

Auto Termination: After the specified minutes of inactivity the cluster is automatically terminated.

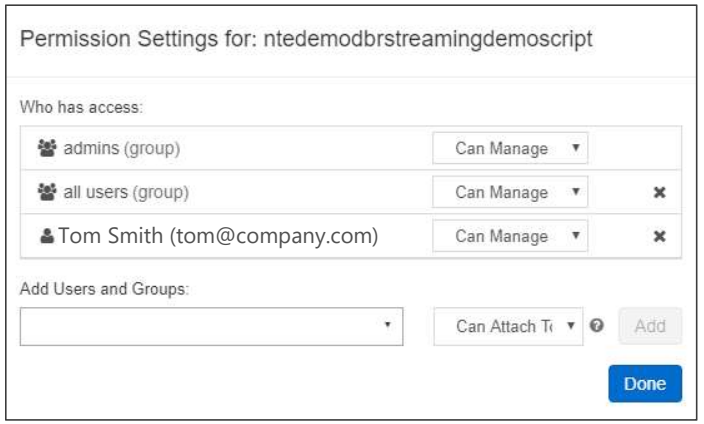
Benefits:

- You do not have to guess, or determine by trial and error, the correct number of nodes for the cluster
- As the workload changes you do not have to manually tweak the number of nodes
- You do not have to worry about wasting resources when the cluster is idle. You only pay for resource when they are actually being used
- You do not have to wait and watch for jobs to complete just so you can shutdown the clusters



Cluster Access Control

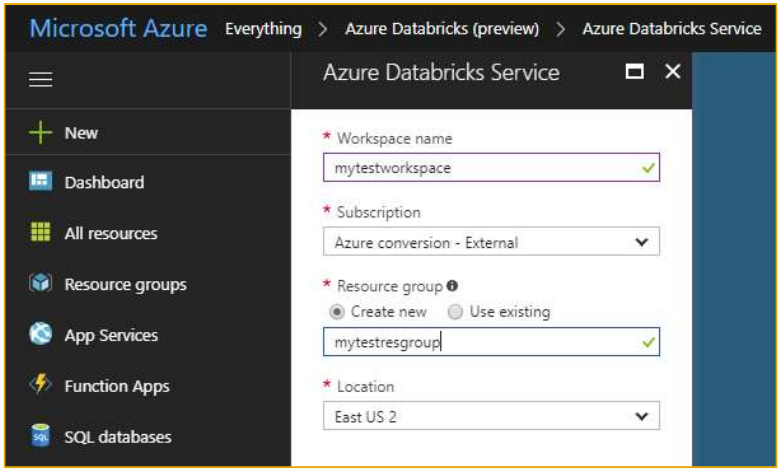
- There are two configurable types of permissions for Cluster Access Control:
  - *Individual Cluster Permissions* - This controls a user's ability to attach notebooks to a cluster, as well as to restart/resize/terminate/start clusters.
  - *Cluster Creation Permissions* - This controls a user's ability to create clusters
- Individual permissions can be configured on the Clusters Page by clicking on Permissions under the 'More Actions' icon of an existing cluster
- There are 4 different individual cluster permission levels: *No Permissions*, *Can Attach To*, *Can Restart*, *and Can Manage*. Privileges are shown below



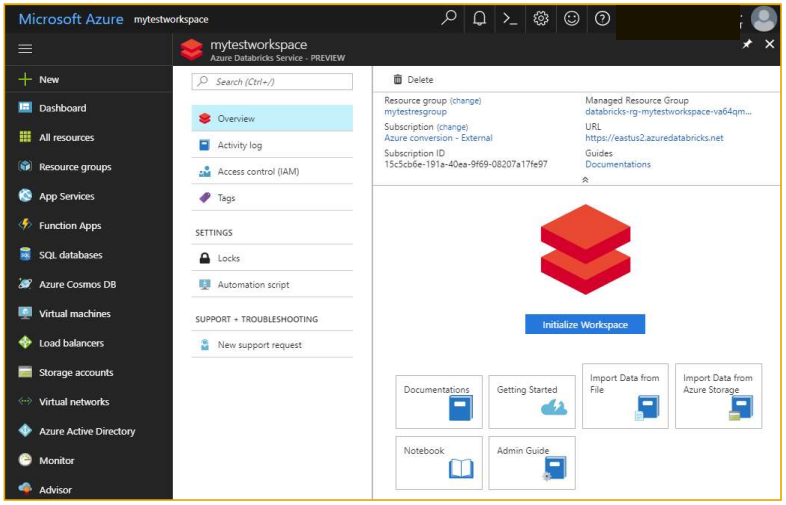
Abilities	No Permissions	Can Attach To	Can Restart	Can Manage
Attach notebooks to cluster		x	x	x
View Spark UI		x	x	x
View cluster metrics (Ganglia)		x	x	x
Terminate cluster			x	x
Start cluster			x	x
Restart cluster			x	x
Resize cluster				x
Modify permissions				x

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

** During the current preview phase, the subscription has to be whitelisted.*

Jobs

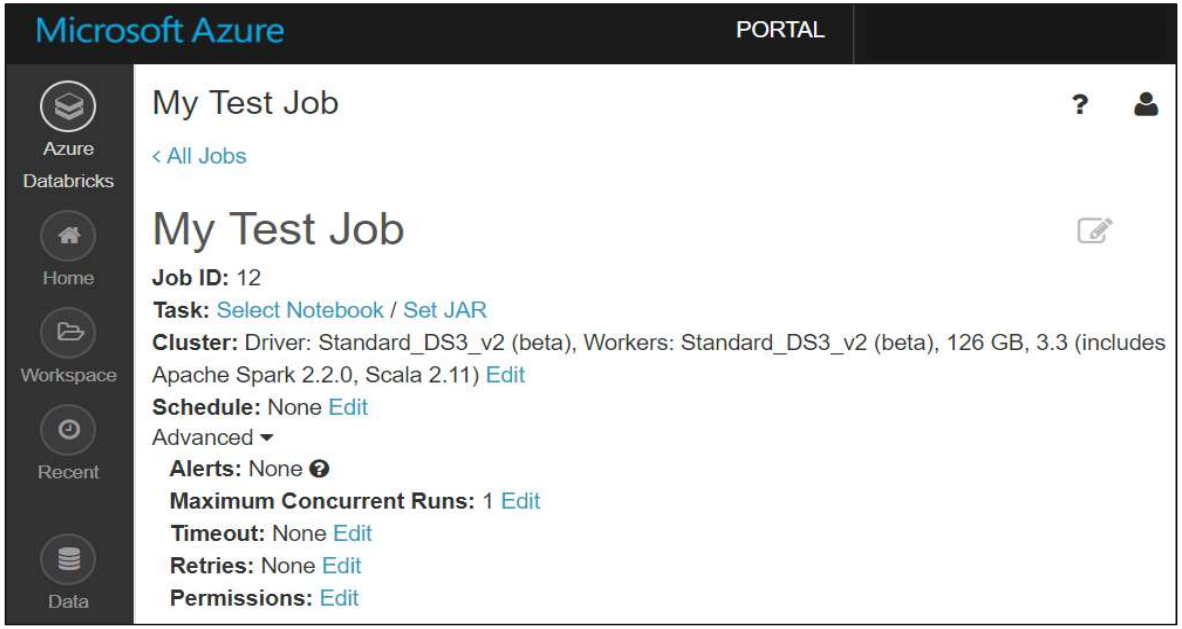
Jobs are the mechanism to submit Spark application code for execution on the Databricks clusters

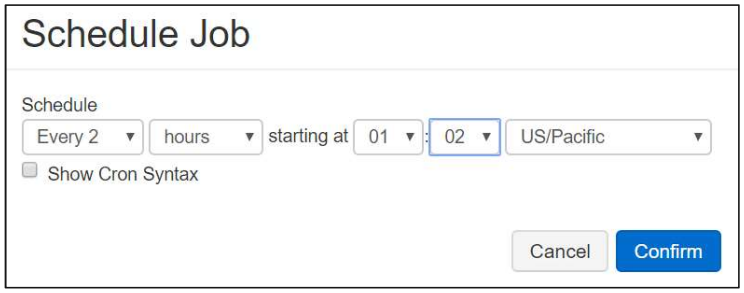
- Spark application code is submitted as a 'Job' for execution on Azure Databricks clusters
- Jobs execute either 'Notebooks' or 'Jars'
- Azure Databricks provide a comprehensive set of graphical tools to create, manage and monitor Jobs.

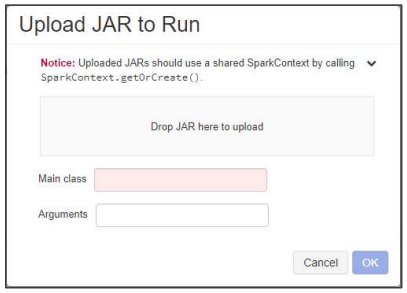
Creating and Running Jobs (1/2)

When you create a new Job you have to specify:

- The Notebook or Jar to execute
- Cluster: The cluster on which the Job execute. This could be an exiting or new cluster.
- Schedule i.e. how often the Job runs. Jobs can also be run one time right away.



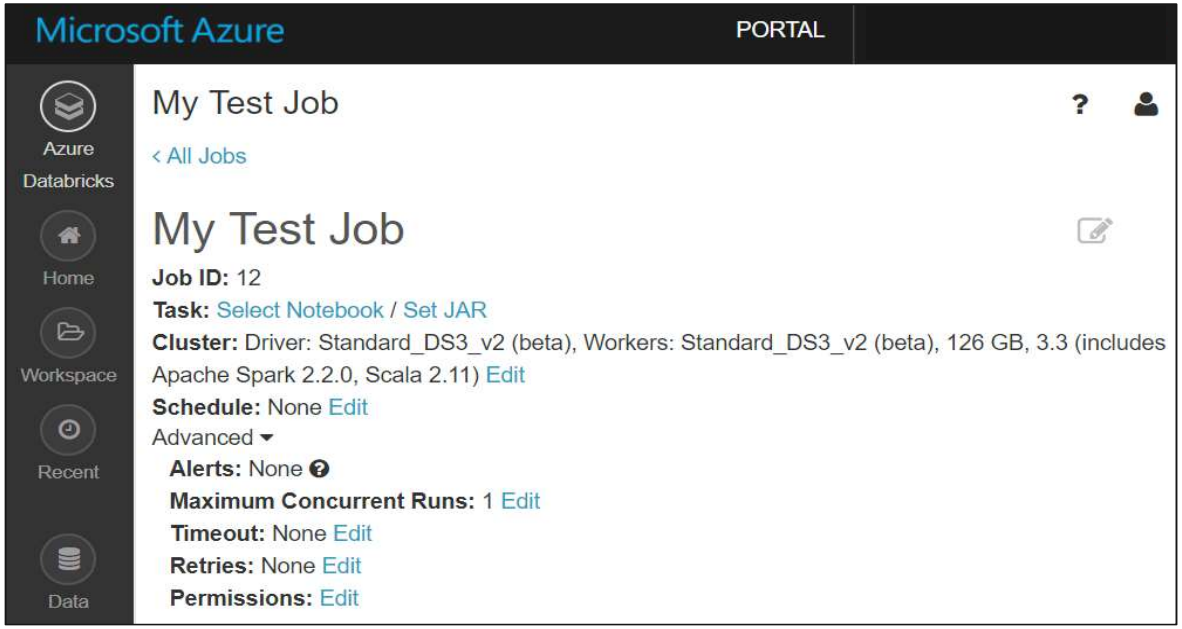


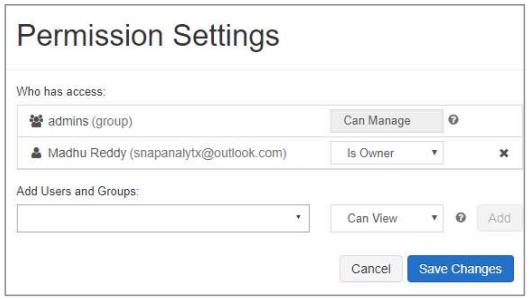


Creating and Running Jobs (2/2)

When you create a new job you can optionally specify advanced options:

- Maximum number of concurrent runs of the Job
- Timeout: Jobs still running beyond the specified duration are automatically killed
- Retry Policy: Specifies if—and when—failed jobs will be retried
- Permissions: Who can do what with jobs. This allows for Job definition and management to be *securely shared* with others (see next slide)

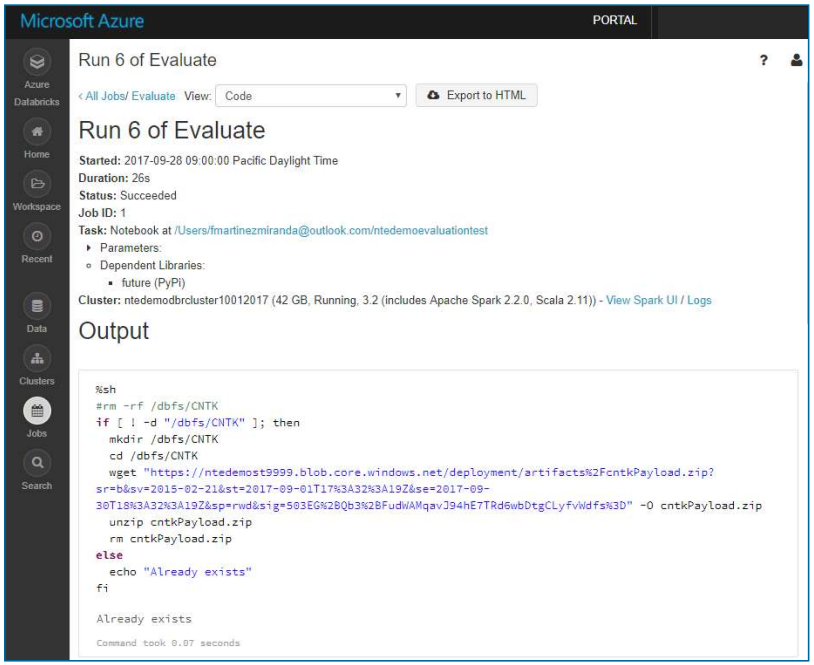


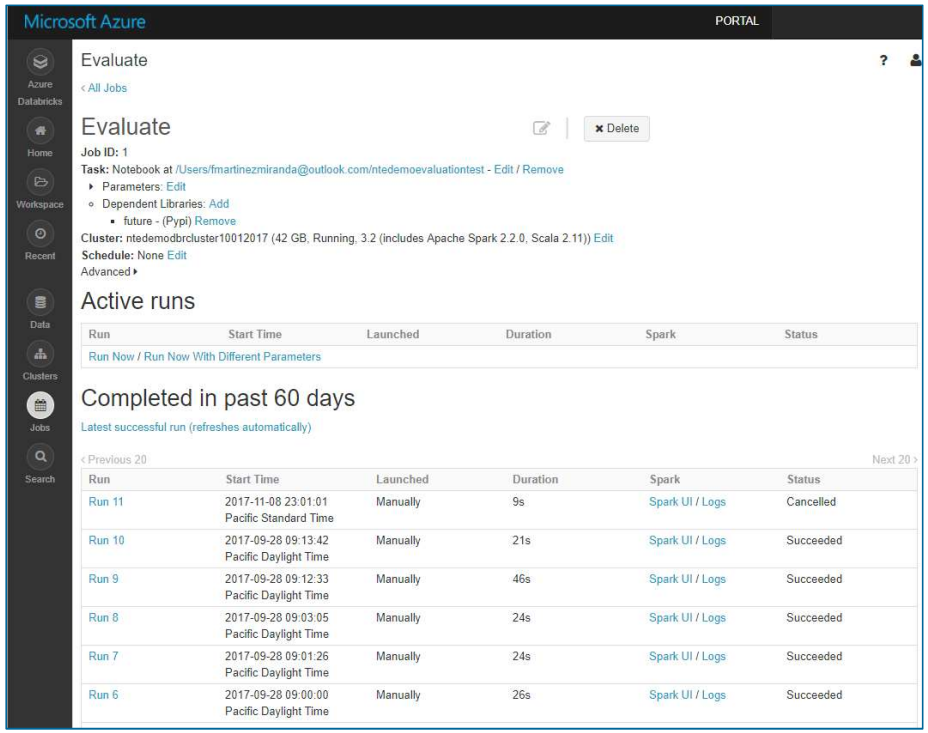


Viewing Jobs History

In the Azure Databricks Jobs Portal you can view:

- The list of currently running (Active) Jobs
- History of old Job runs (for up to 60 days)
- The output of a particular Job run (including standard error, standard output, Spark UI logs)

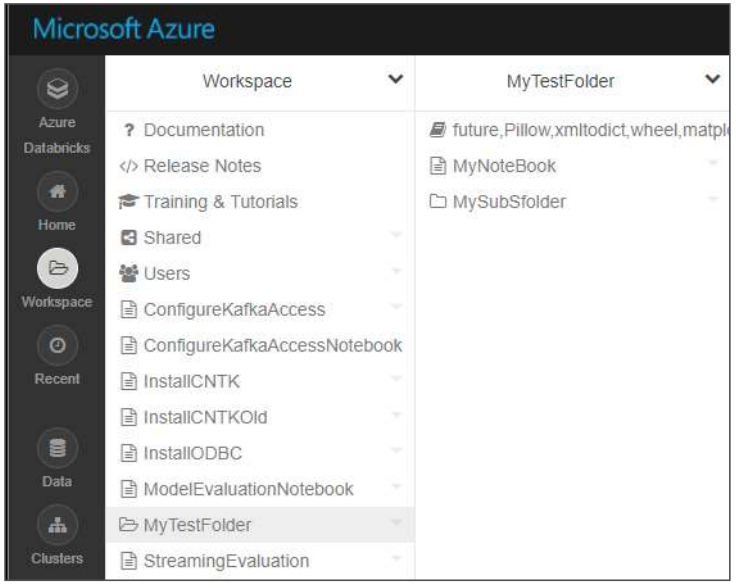


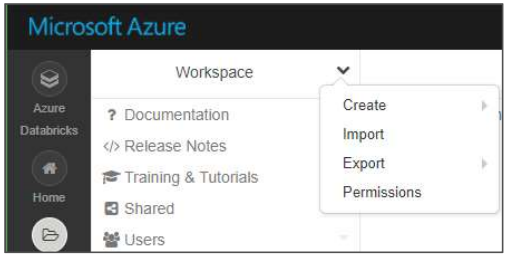


Workspaces

Workspaces enables users to organize—and share—their Notebooks, Libraries and Dashboards

- Workspaces—sort of like Directories— are a convenient way to organize an user's Notebook, Libraries and Dashboards.
- Everything in a workspace is organized into hierarchical folders. Folders can hold Libraries, Notebooks, Dashboard or more (sub) folders.
  - Icons indicate the type of the object contained in a folder
- Every user has one directory that is private and unshared.
  - By default, the workspace and all its contents are available to users.
- Fine grained access control can be defined on workspaces (next slide) to enable *secure collaboration with colleagues*.



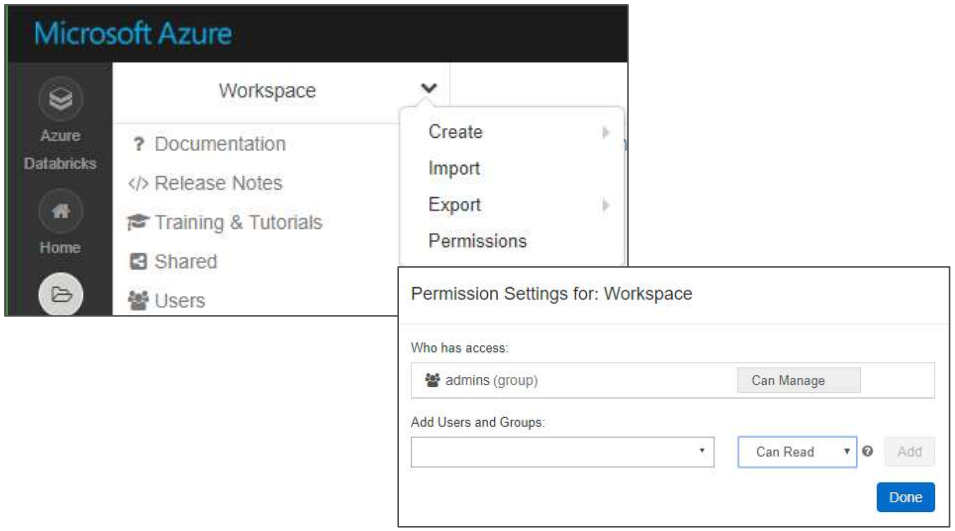


Workspace Operations

You can search the entire Databricks workspace

In the Azure Databricks Portal, via the Workspaces drop down menu, you can:

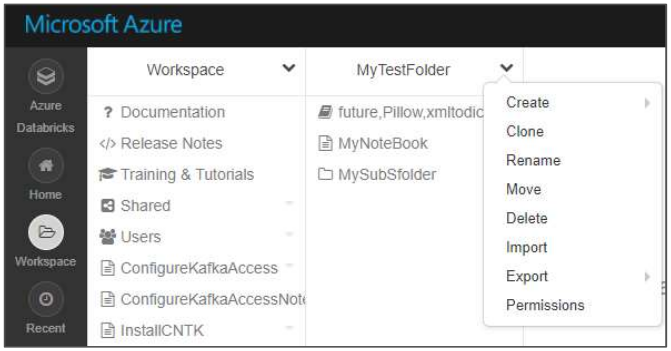
- Create Folders, Notebooks and Libraries
- Import Notebooks into the Workspace
- Export the Workspace to a database archive
- Set Permissions. You can grant 4 levels of permissions
  - Can Manage
  - Can Read
  - Can Edit
  - Can Run



Folder Operations & Access Control

In the Azure Databricks Portal, via the Folder drop down menu, you can:

- Create Folders, Notebooks and Libraries within the folder
- Clone the folder to create a deep copy of the folder
- Rename or delete the folder
- Move the folder to another location
- Export a folder to save it and its contents as a Databricks archive
- Import a saved Databricks archive into the selected folder
- Set Permissions for the folder. As with Workspaces you can set 5 levels of permissions: *No Permissions, Can Manage, Can Read, Can Edit, Can Run*



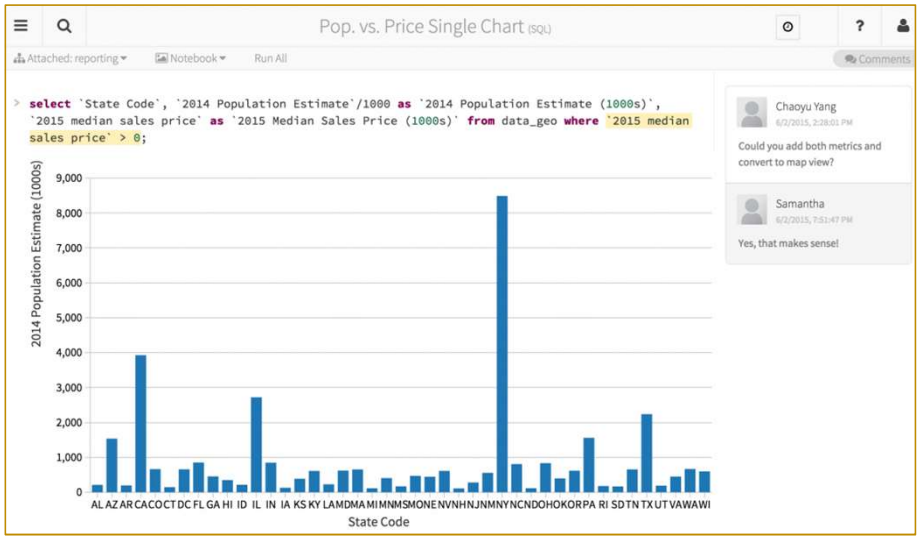
Abilities	No Permissions	Read	Run	Edit	Manage
Create items					✅
Delete items					✅
Move/rename items					✅
Change permissions					✅

Abilities associated with each permission level

Azure Databricks Notebooks Overview

Notebooks are a popular way to develop, and run, Spark Applications

- Notebooks are not only for authoring Spark applications but can be *run/executed directly* on clusters
  - Shift+Enter
  - click the ▶ at the top right of the cell in a notebook
  - Submit via Job
- Notebooks support fine grained permissions—so they can be *securely shared* with colleagues for collaboration (see following slide for details on permissions and abilities)
- Notebooks are well-suited for prototyping, rapid development, exploration, discovery and iterative development



Notebooks typically consist of code, data, visualization, comments and notes

Mixing Languages in Notebooks

You can mix multiple languages in the same notebook

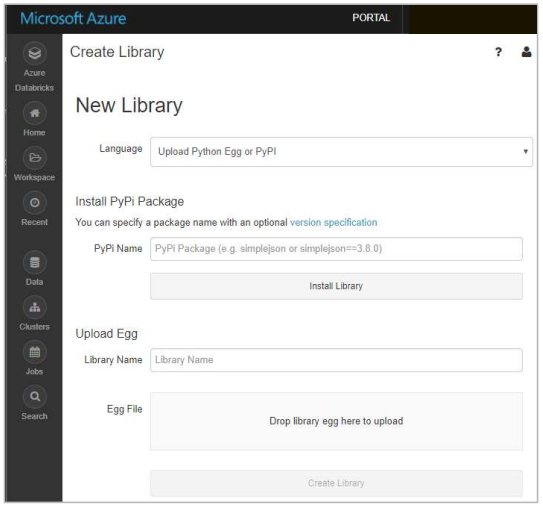
Normally a notebook is associated with a specific language. However, with Azure Databricks notebooks, you can mix multiple languages in the same notebook. This is done using the language magic command:

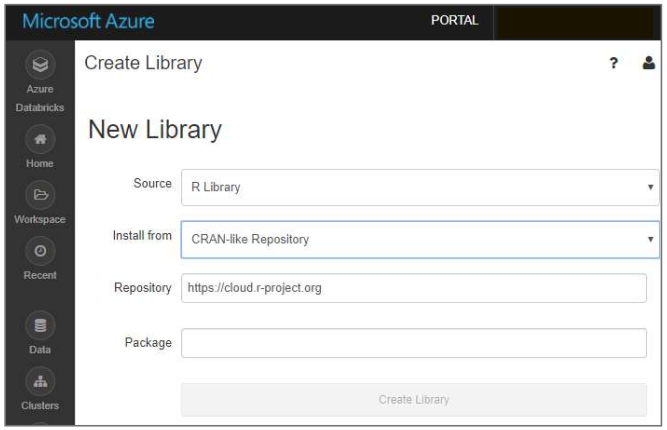
- %python Allows you to execute python code in a notebook (even if that notebook is not python)
- %sql Allows you to execute sql code in a notebook (even if that notebook is not sql).
- %r Allows you to execute r code in a notebook (even if that notebook is not r).
- %scala Allows you to execute scala code in a notebook (even if that notebook is not scala).
- %sh Allows you to execute shell code in your notebook.
- %fs Allows you to use Databricks Utilities - dbutils filesystem commands.
- %md To include rendered markdown

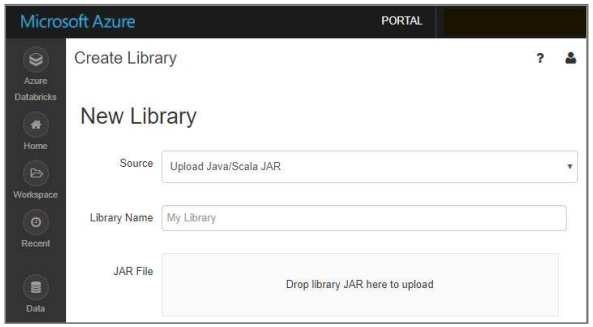
Libraries Overview

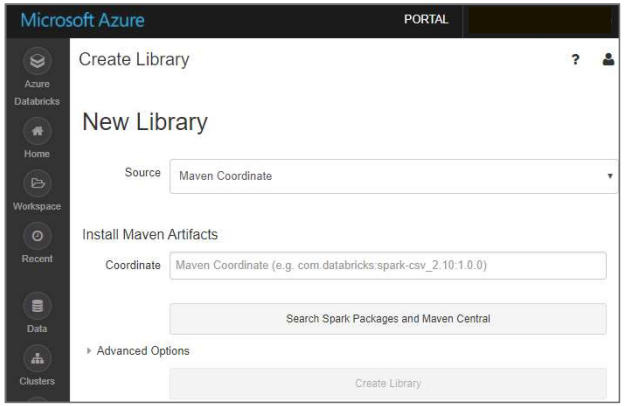
Enables external code to be imported and stored into a Workspace

- Libraries are containers to hold all your *Python, R, Java/Scala* libraries.
- Libraries resides within workspaces or folders.
- Libraries are created by importing the source code
- After importing libraries are immutable—can be deleted or overwritten only.
- You can customize installation of libraries via Init Scripts by writing custom UNIX scripts
- Libraries can also be managed via the Library API



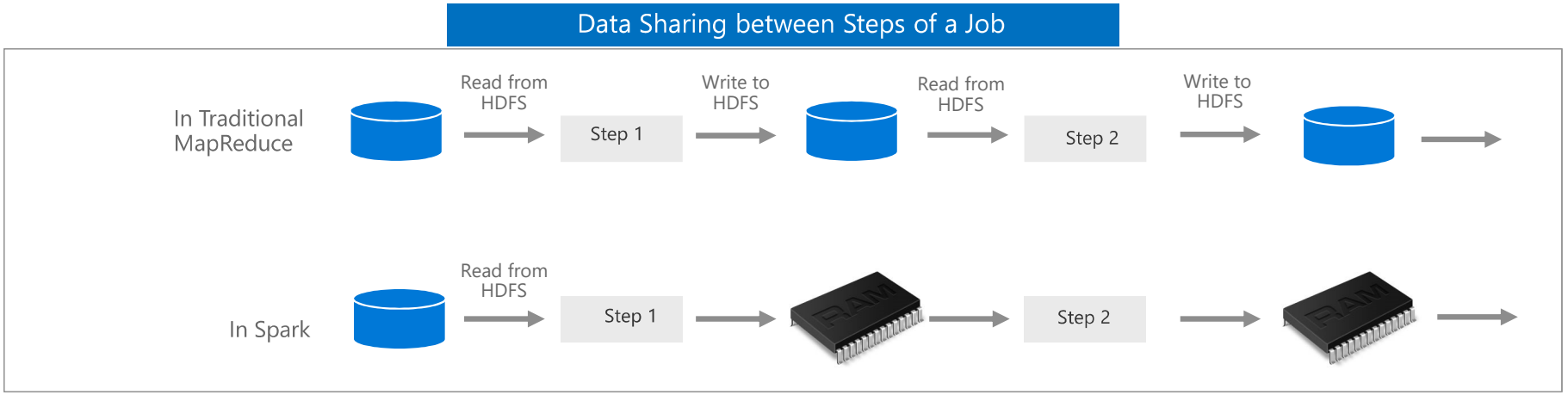




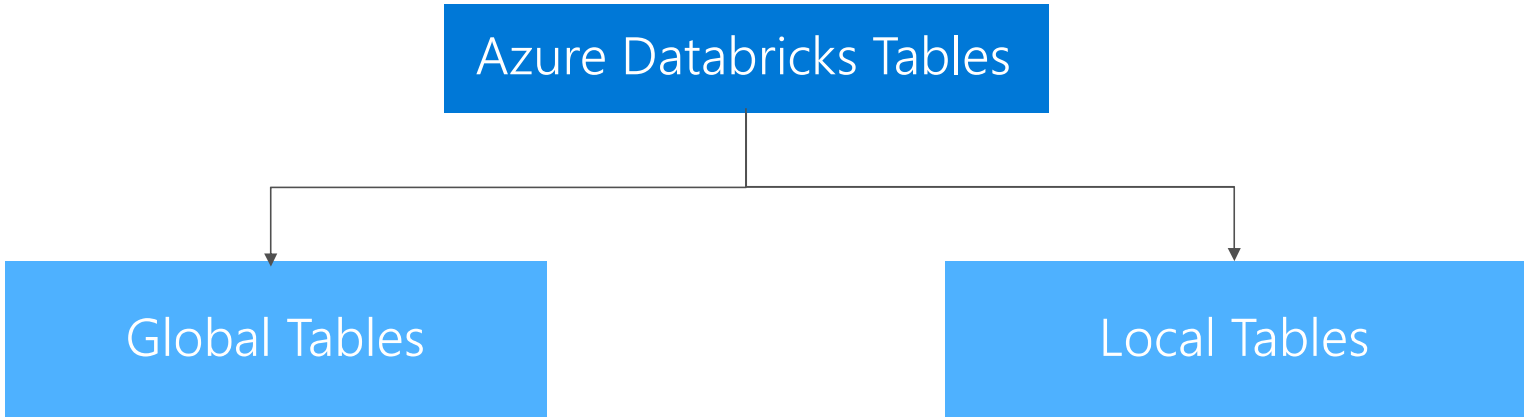


What Makes Spark Fast ?

- **In-memory cluster computing:** Spark provides primitives for *in-memory* cluster computing. A Spark job can *load and cache* data into memory and query it repeatedly (iteratively) much quicker than disk-based systems.
- **Scala Integration:** Spark integrates into the Scala programming language, letting you manipulate distributed datasets like local collections. No need to structure everything as map and reduce operations
- **Faster Data-sharing:** Data-sharing between operations is faster as data is in-memory:
  - In (traditional) Hadoop data is shared through HDFS which is expensive. HDFS maintains three replicas.
  - Spark stores data in-memory *without any replication*.



Local & Global Tables



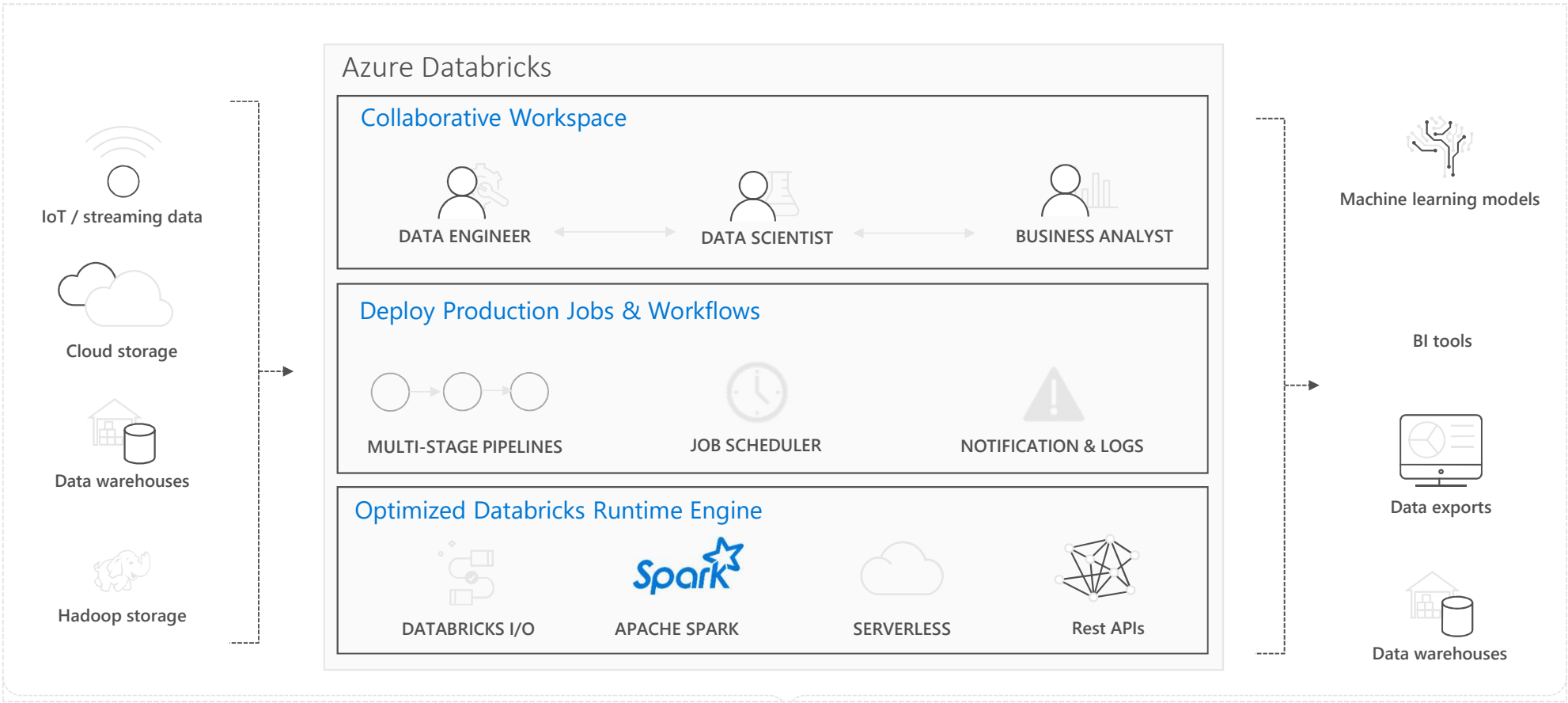
Databricks registers global tables to the Hive metastore and makes them available across all clusters.

Only global tables are visible in the Tables pane

Databricks does not registers local tables in the Hive metastore and are only available within one cluster.

Also known as *temporary* tables

Databricks



Enhance Productivity

Build on secure & trusted cloud

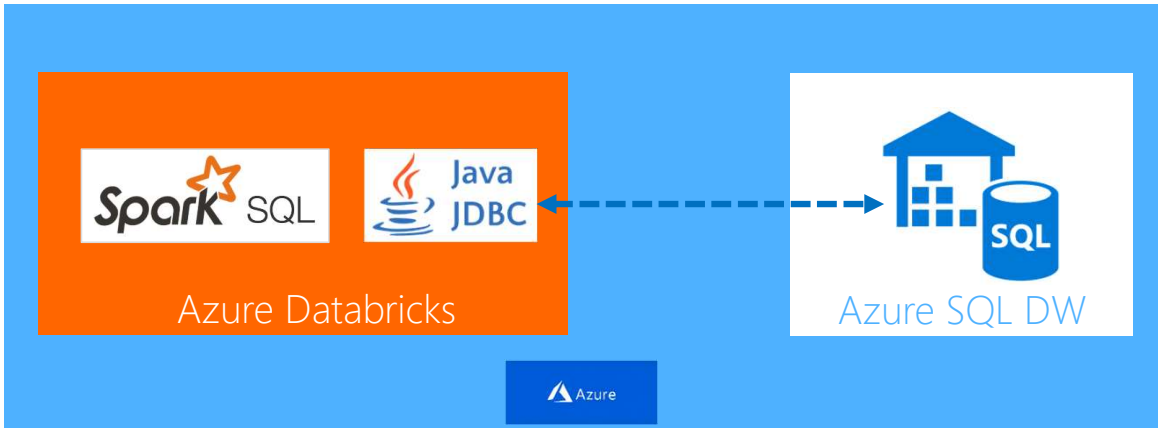
Scale without limits

Azure SQL DW Integration

Integration enables structured data from SQL DW to be included in Spark Analytics

Azure SQL Data Warehouse is a SQL-based fully managed, petabyte-scale cloud solution for data warehousing

- You can bring in data from Azure SQL DW to perform advanced analytics that require both structured and unstructured data.
- Currently you can access data in Azure SQL DW via the JDBC driver. From within your spark code you can access just like any other JDBC data source.
- If Azure SQL DW is authenticated via AAD then Azure Databricks user can seamlessly access Azure SQL DW.



Batch data processing

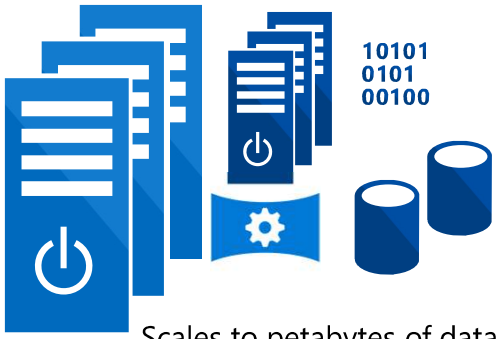
A side-by-side comparison of general capabilities and features

	Azure Data Lake Analytics	HDInsight	Azure Databricks
Is a managed service	Yes	Yes	Yes
Auto-scaling	No	No	Yes
Supports pausing compute	No	No	Yes
Programmability	U-SQL	Python, Scala, Java, R, SQL	Python, Scala, Java, SQL, R
Programming paradigm	Mixture of declarative and imperative	Mixture of declarative and imperative	Mixture of declarative and imperative
Pricing model	Per job (by job run per hour times analytics unit used)	By cluster hour	By cluster hour

Introduction to Azure SQL data warehouse

Azure SQL Data Warehouse

Elastic scale & performance



Scales to petabytes of data

Massively Parallel Processing

Instant-on compute scales in seconds

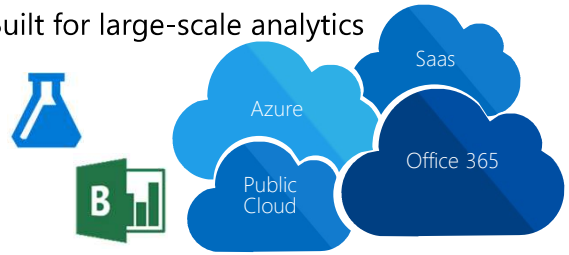
Query Relational and Non-Relational data

Relational batch processing

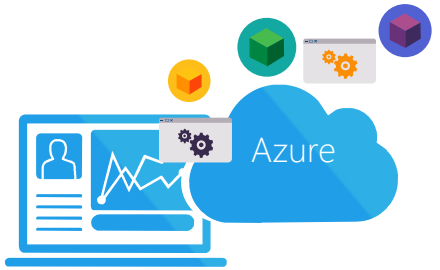
Query large datasets in minutes

Full hub-and-spoke support

Built for large-scale analytics



Market Leading Price & Performance



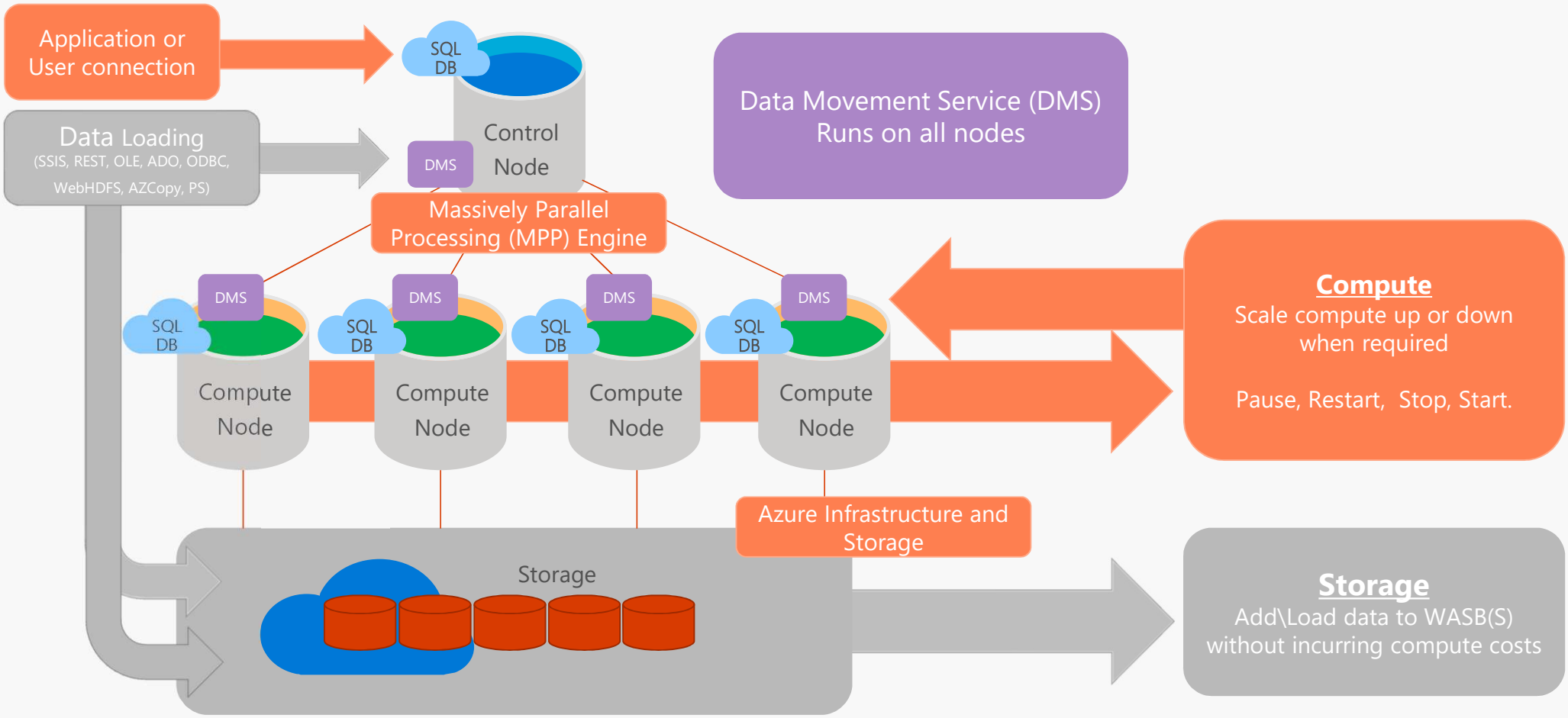
Simple billing compute & storage

Pay for what you need, when you need it with dynamic pause

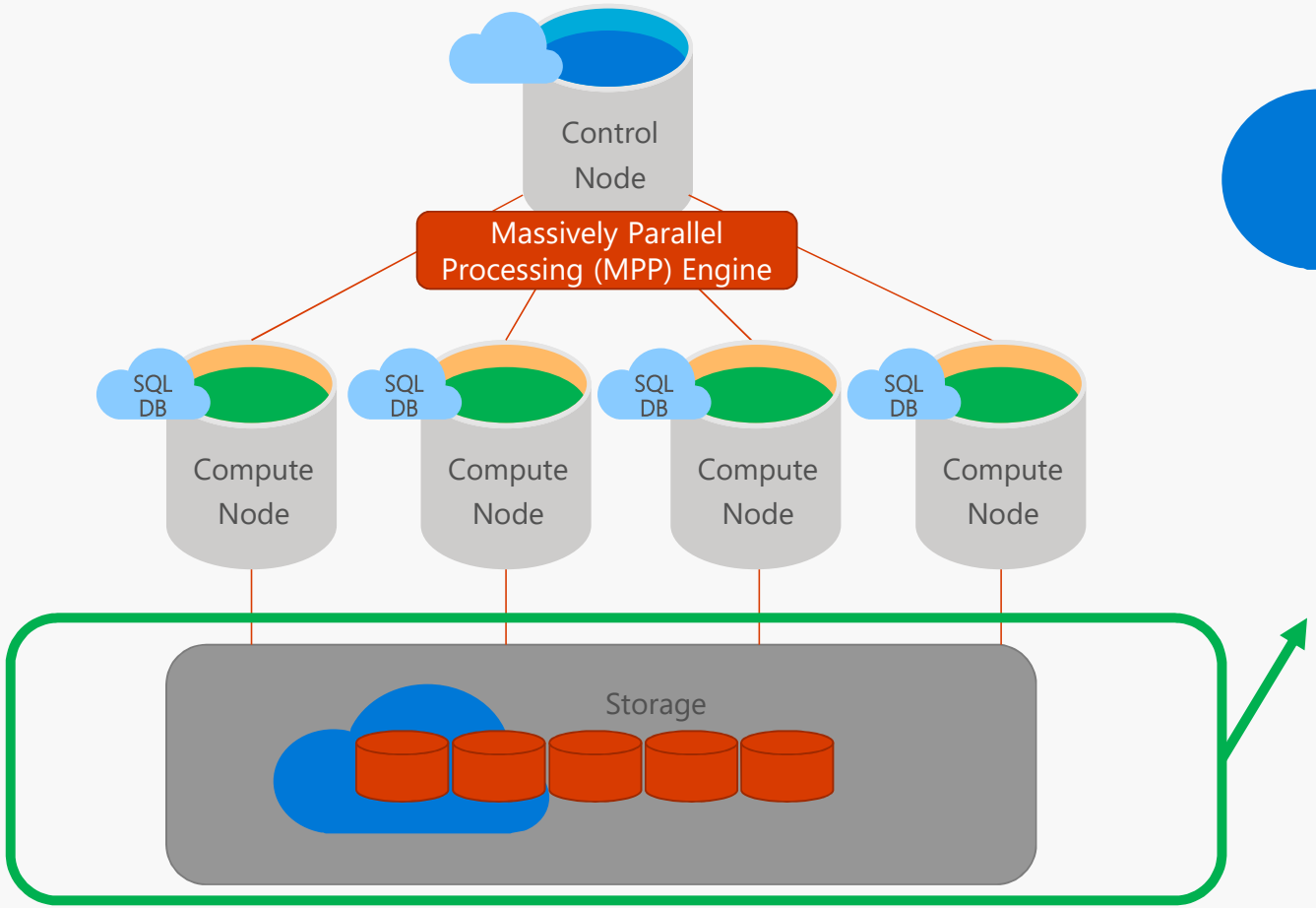
Sets it different..

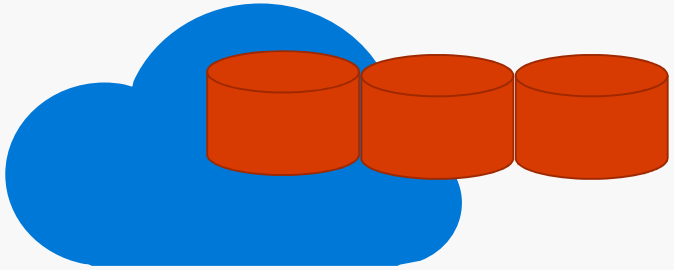
- Separate compute from storage
- Independently scale compute
- Pause and resume workload
- Supports Rowstore & Columnstore
- Scale out = Distributed tables
- Local or Geo-redundancy
- Parallel Loading with PolyBase

Azure SQL Data Warehouse Architecture



Azure SQL Data Warehouse – premium storage





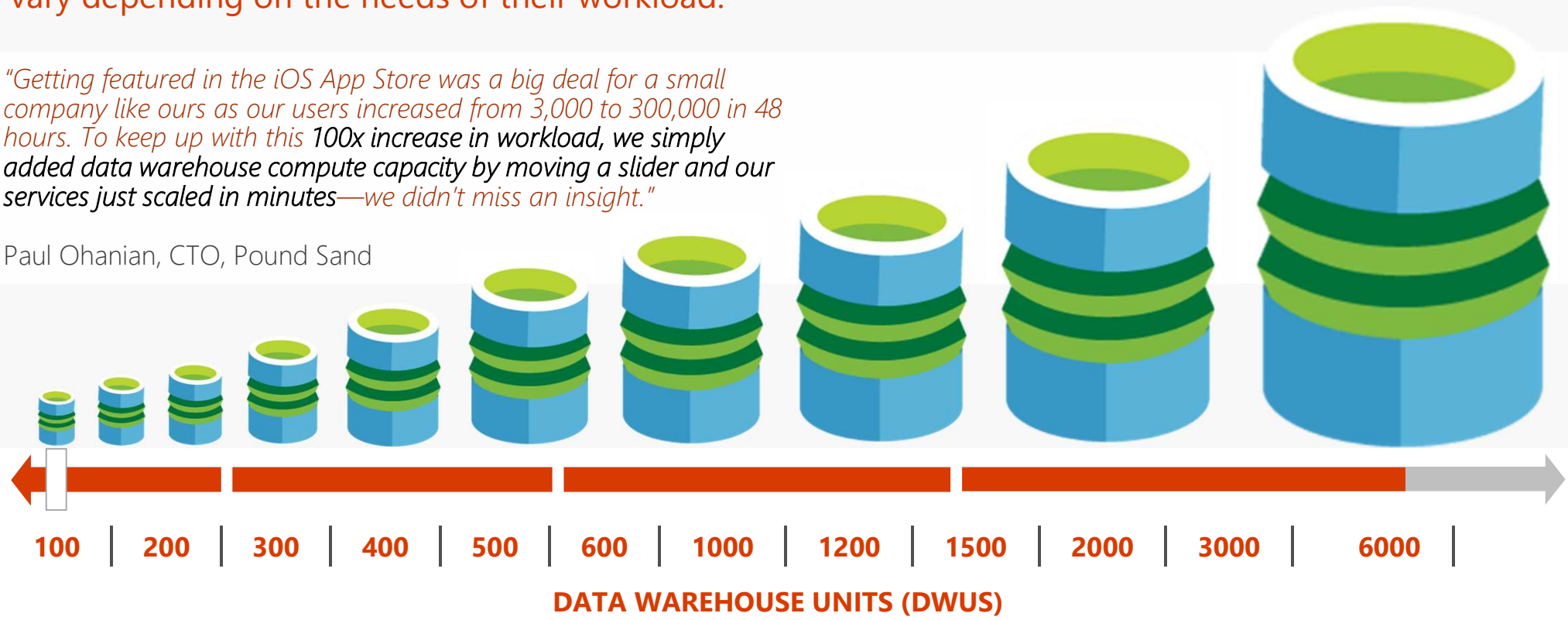
- Premium RA-GRS storage
- SSD-based
- Better latency
- Reliable I/O – 400 MB/sec per node
- Persist the data without incurring compute costs

Scale compute on the fly when you need it

Data Warehouse Units (DWUs) are a measure of reserved compute performance or 'power.' A customer's DWU needs can vary depending on the needs of their workload.

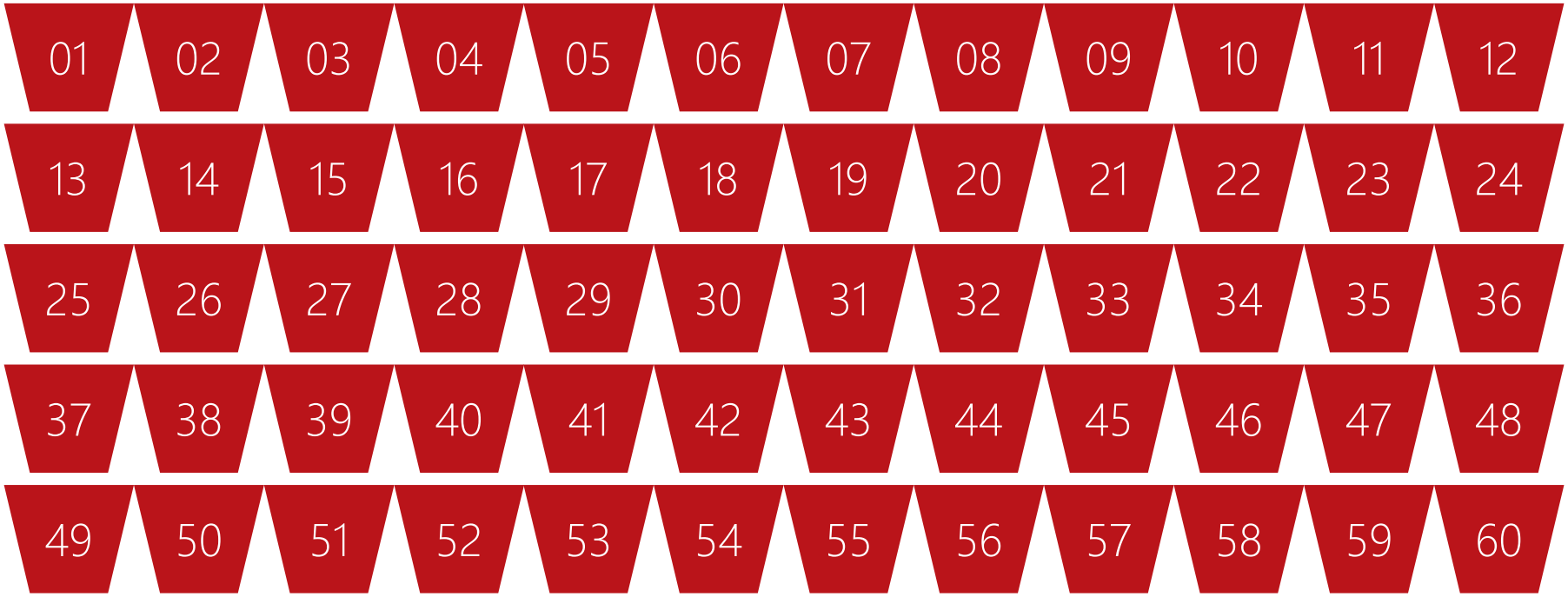
"Getting featured in the iOS App Store was a big deal for a small company like ours as our users increased from 3,000 to 300,000 in 48 hours. To keep up with this ***100x increase in workload, we simply added data warehouse compute capacity by moving a slider and our services just scaled in minutes****—we didn't miss an insight."*

Paul Ohanian, CTO, Pound Sand

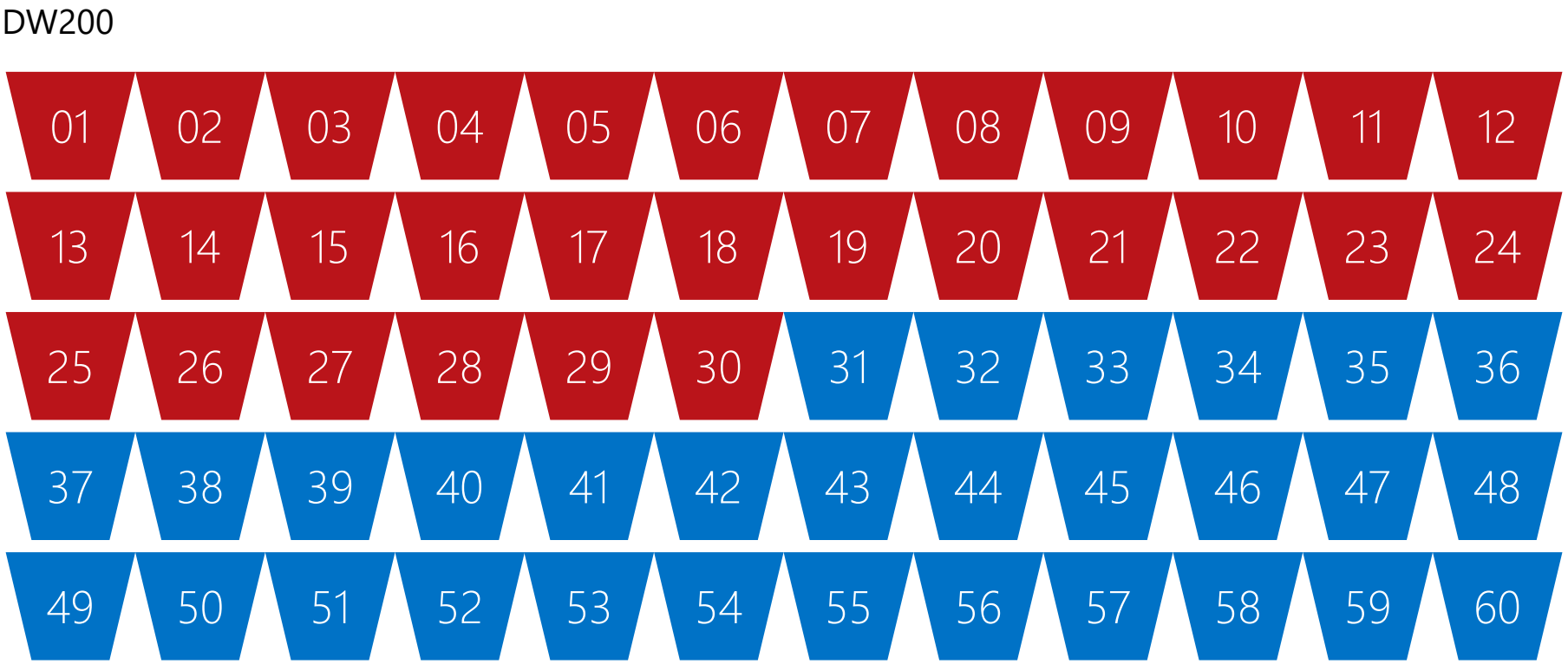


Mapping Compute in SQLDW

DW100



Mapping Compute in SQLDW



Distributed queries

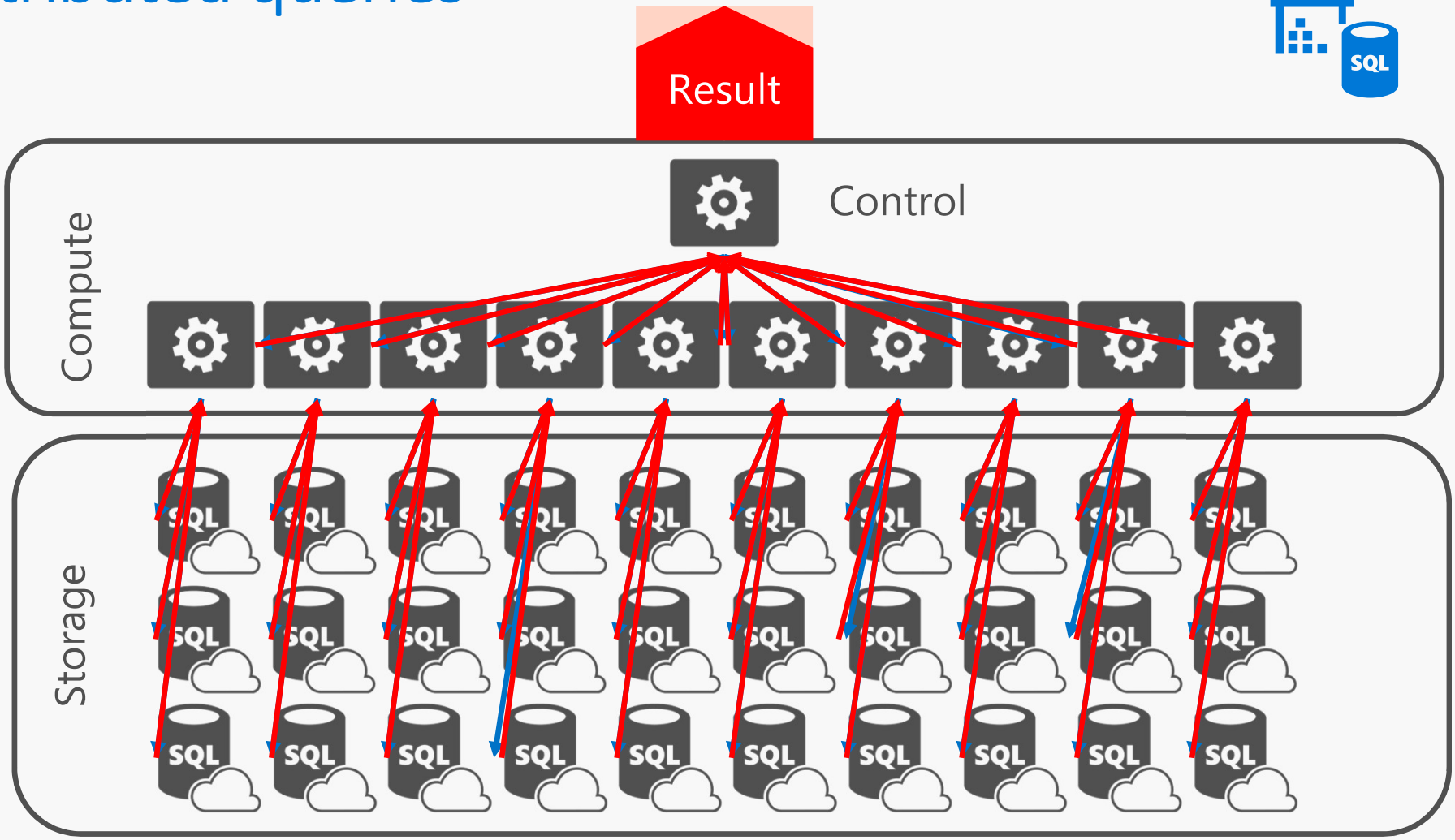


Table Distribution Options

Hash Distributed	Round Robin (Default)	Replicated
Data divided across nodes based on hashing algorithm Same value will always hash to same distribution Single column only	Data distributed evenly across nodes Easy place to start, don't need to know anything about the data Simplicity at a cost	Data repeated on every node Simplifies many query plans and reduces data movement Best with joining hash table
Check for Data Skew, NULLS, -1	Will incur more data movement at query time	Consumes more space Joining two Replicated Table runs on one node

The most secure and compliant cloud data warehouse

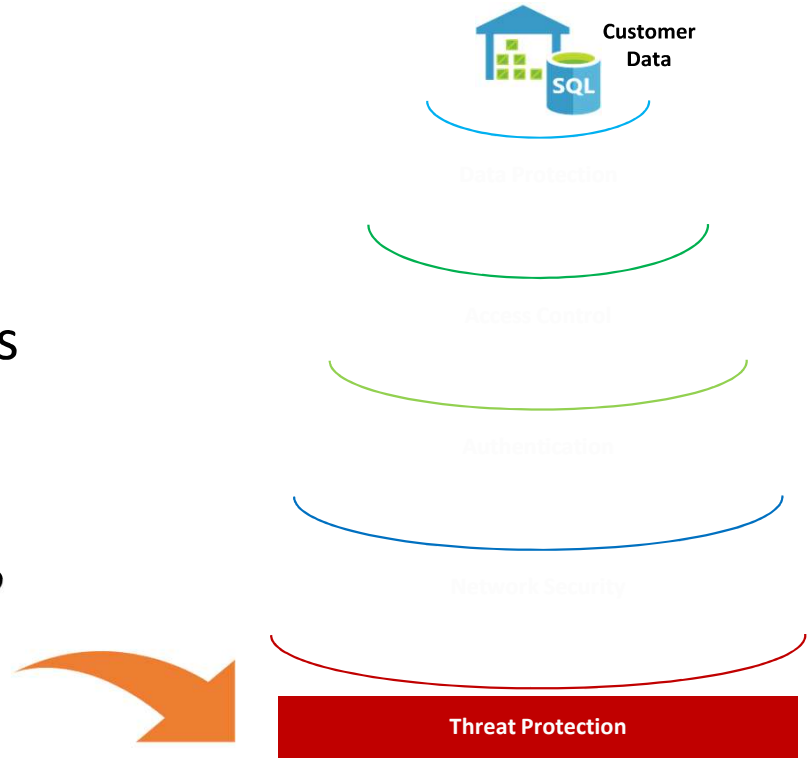
Enterprise Grade Security

Industry Leading Security

Category	Feature	SQL Data Warehouse	Amazon Redshift	Snowflake	Google Big Query
Data Protection	Data In Transit	Yes	Yes	Yes	No
	Data encryption at rest (Service & User Managed Keys)	Yes	Yes	Yes	Yes
	Data In Use (Always Encrypted)	No	No	No	No
	Data Discovery and Classification	Yes (March 2019)	No	No	No
Access Control	Native Row Level Security	Yes	No	No	No
	Table and View Security (GRANT / DENY)	Yes	Yes	Yes	Yes
	Column Level Security	Yes	No	No	No
Authentication	SQL Authentication	Yes	Yes	Yes	No
	Native Azure Active Directory	Yes	No	Yes	No
	Integrated Security	Yes	Yes	Yes	Yes
	Multi-Factor Authentication	Yes	Yes	Yes	Yes
Network Security	Virtual Network (VNET)	Yes	Yes	Yes	Yes
	SQL Firewall (server)	Yes	Yes	Yes	No
Threat Protection	SQL Threat Detection	Yes	Yes	No	No
	SQL Auditing	Yes	Yes	No	Yes
	Vulnerability Assessment	Yes	Yes	No	No

Business Requirements

- How do I discover potential database vulnerabilities?
- How do my auditors know if my data is compliant?
- How do I monitor malicious activities?



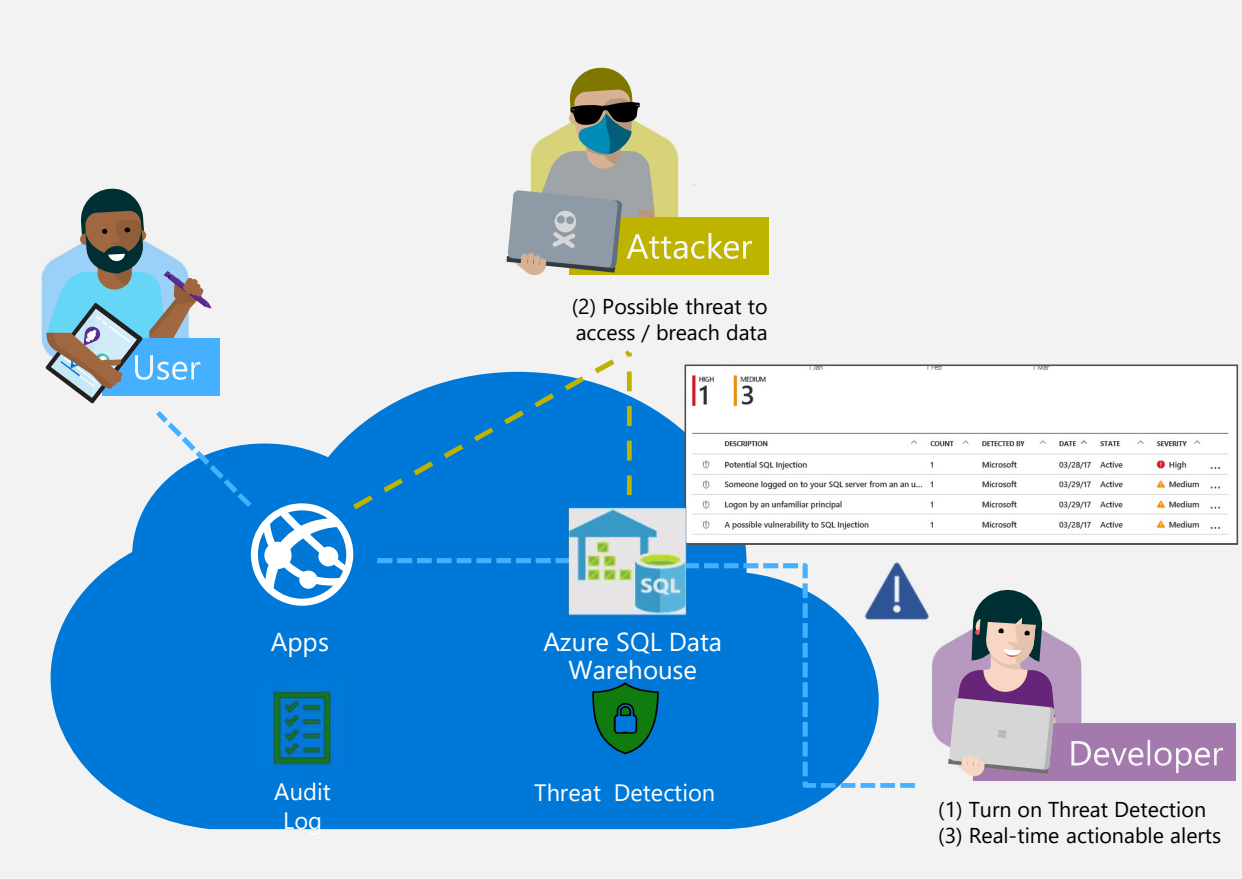
SQL Vulnerability Assessment

Discover, track, and remediate security misconfigurations

- ✓ Identify security misconfigurations
- ✓ Actionable remediation steps
- ✓ Security baseline tuned to your environment
- ✓ Manual/periodic scans
- ✓ Coherent reports for auditors

SQL Threat Detection

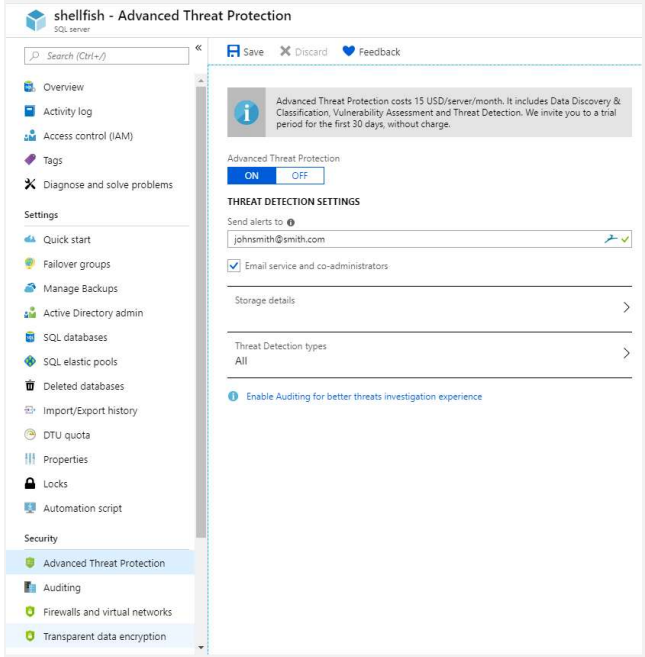
Detect unusual and harmful attempts to breach your data



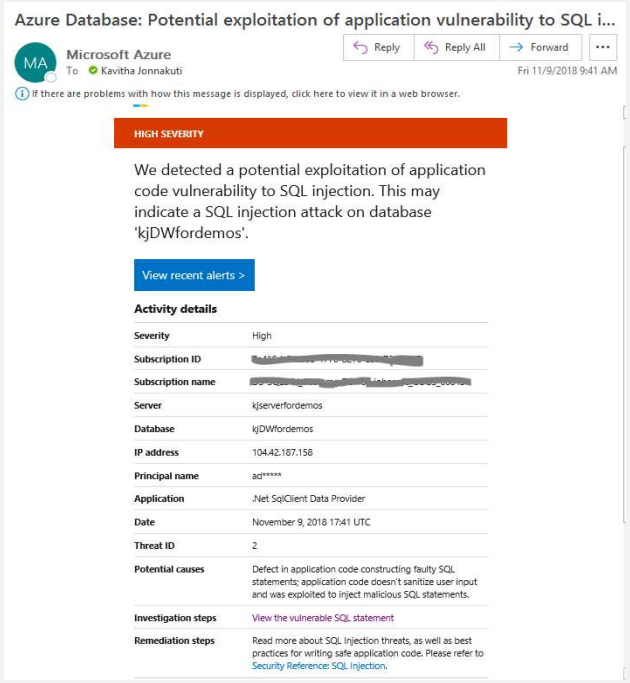
- ✓ Detects potential SQL injection attacks
- ✓ Detects unusual access & data exfiltration activities
- ✓ Actionable alerts to investigate & remediate
- ✓ View alerts for your entire Azure tenant using Azure Security Center

How Threat Detection Works

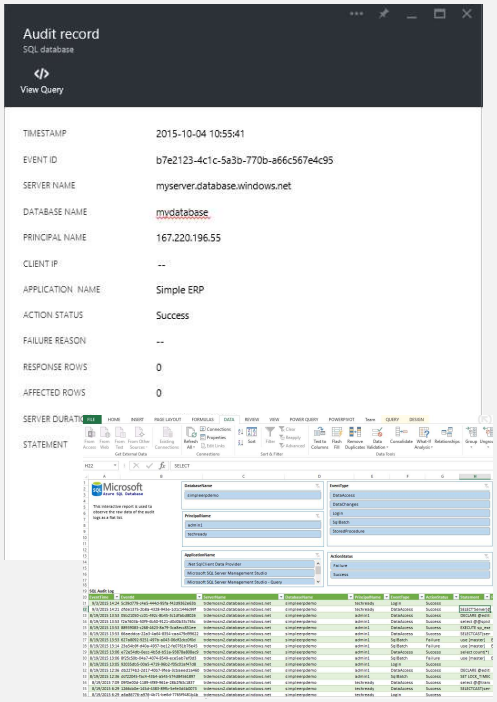
Set up



Alert



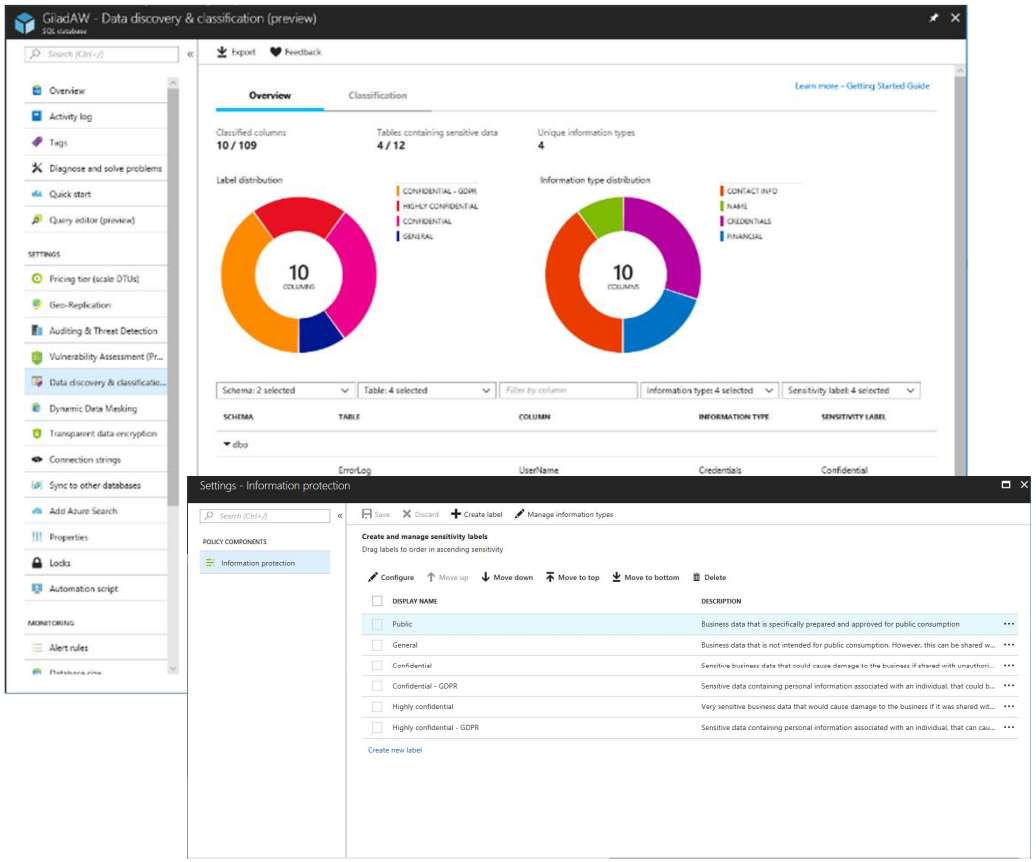
Explore



Threat Protection

Public Preview
Mar'2019

SQL Data Classification, Discover, Classify, Protect & Track Access to Sensitive Data

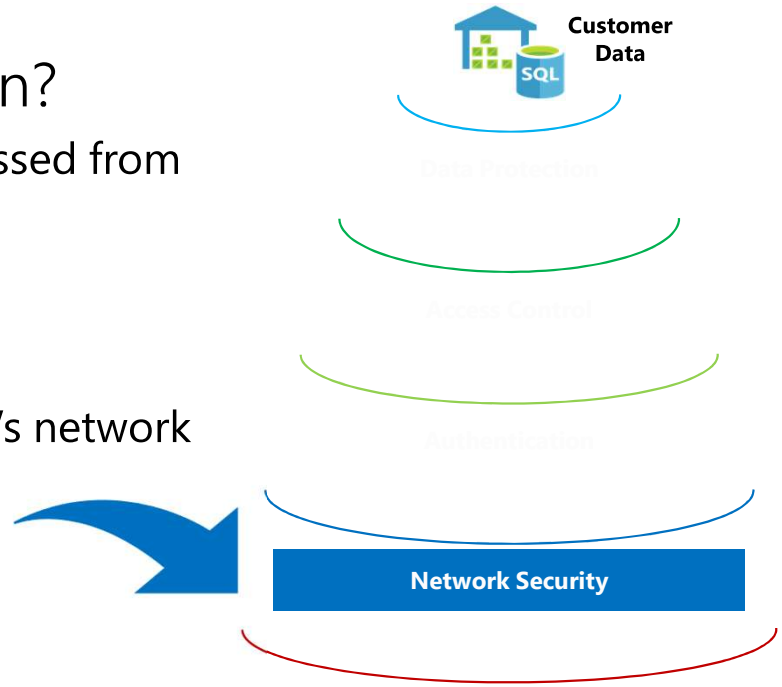


- ✓ Automatic discovery of columns with sensitive data
- ✓ Add persistent sensitive data labels
- ✓ Audit and detect access to the sensitive data
- ✓ Manage labels for your entire Azure tenant using Azure Security Center

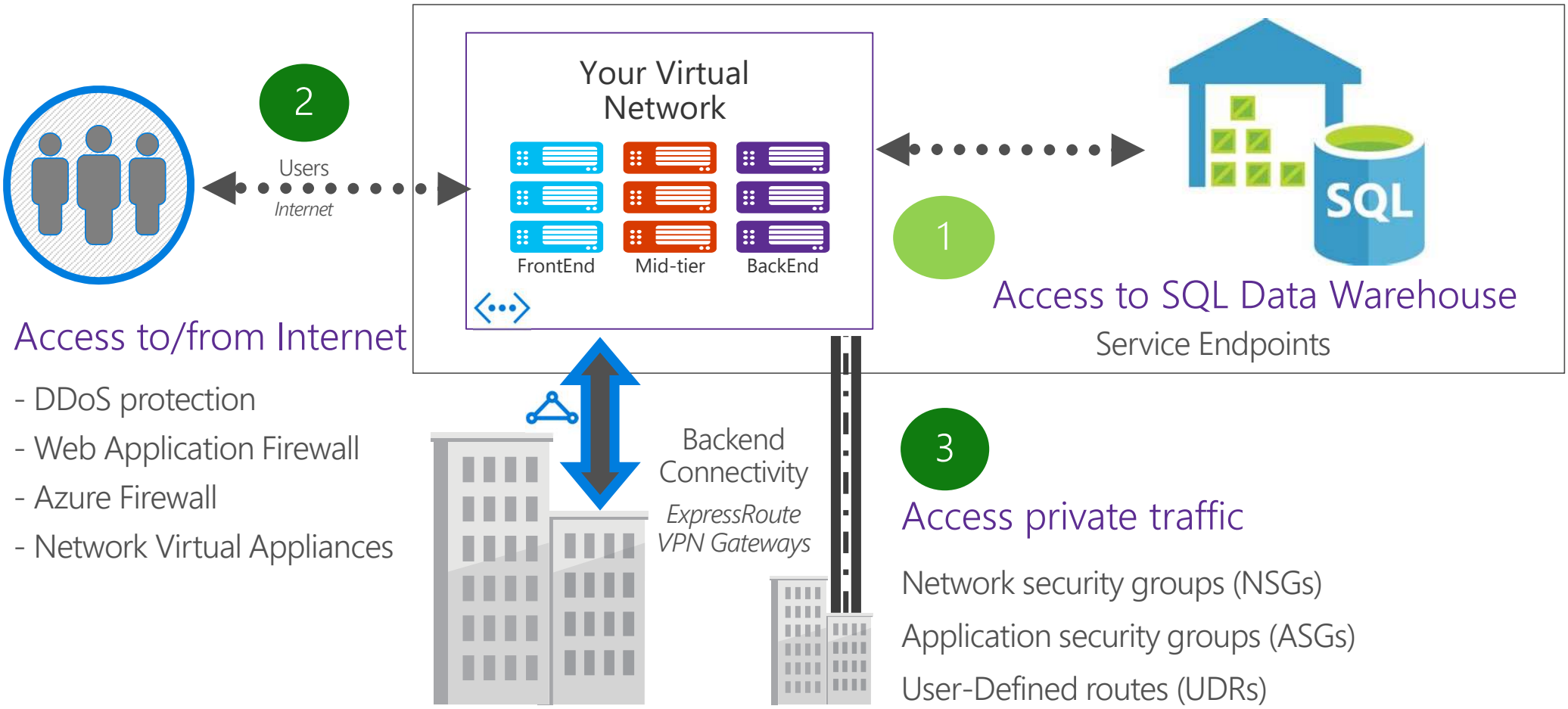
Threat Protection

Business Requirements

- How do we implement network isolation?
  - Data at different levels of security needs to be accessed from different locations.
- How do we achieve separation?
  - Disallowing Access to entities outside the company's network security boundary

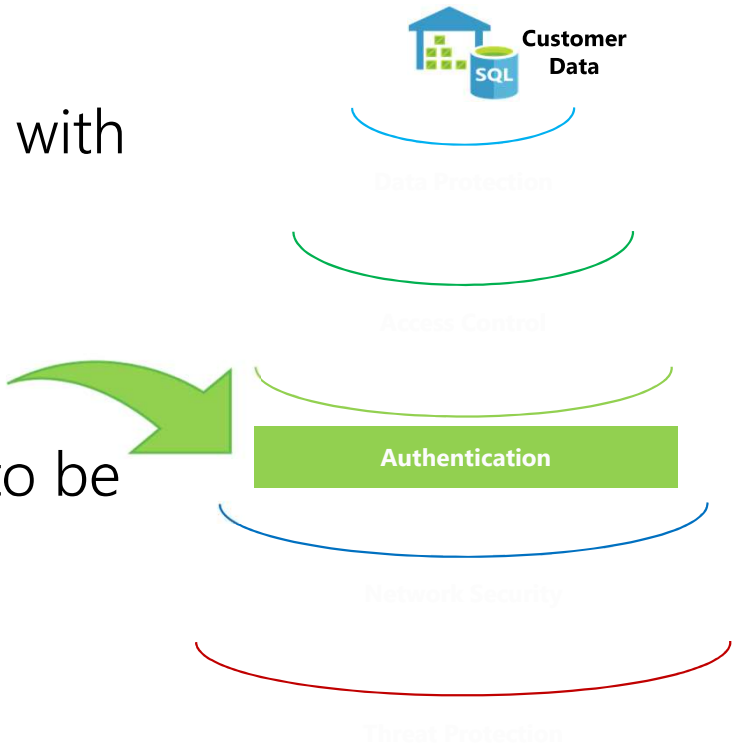


Azure Networking – Application Access Patterns



Business Requirements

- Ho do I configure Azure Active Directory with Azure SQL Data Warehouse?
  - I want additional control in the form of Multi-factor authentication
- How do I allow non-Microsoft accounts to be able to authenticate?



Authentication

Azure Active Directory Authentication

Overview

Manage user identities in one location

Enable access to Azure SQL Data Warehouse and other Microsoft services with Azure Active Directory user identities and groups

Benefits

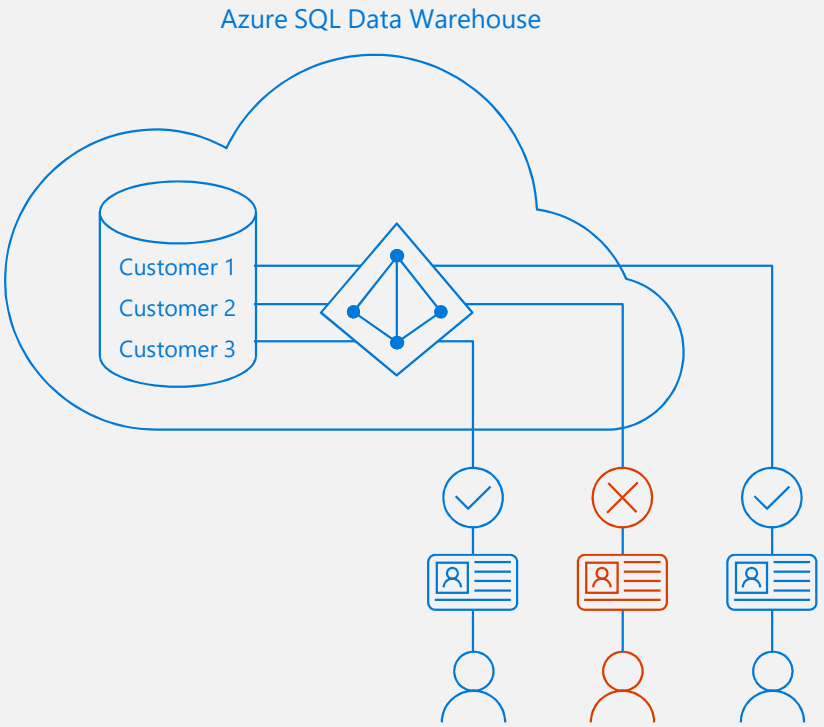
Alternative to SQL Server authentication

Limits proliferation of user identities across databases

Allows password rotation in a single place

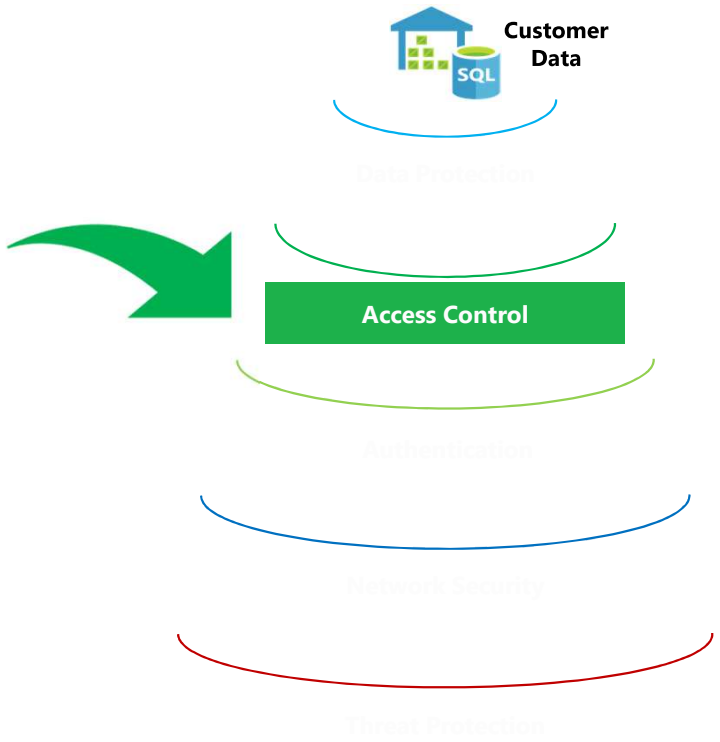
Enables management of database permissions by using external Azure Active Directory groups

Eliminates the need to store passwords



Business Requirements

- How do I restrict access to sensitive data to specific database users?
- How do I ensure users only have access to relevant data?
  - For example, in a hospital only medical staff should be allowed to see patient data that is relevant to them and not every patient's data



Access Control

Column-Level Security

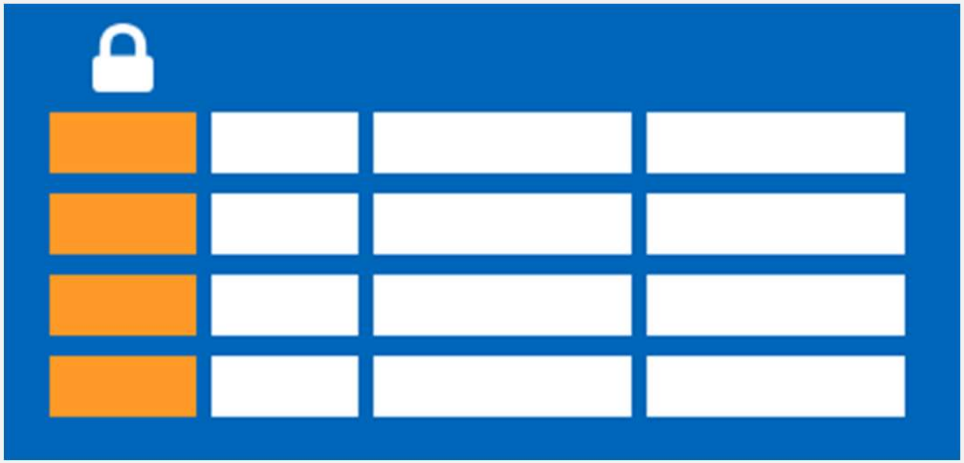
Overview

Control access of specific columns in a database table based on customer's group membership or execution context

Simplifies the design and implementation of security by putting restriction logic in database tier as opposed to application tier

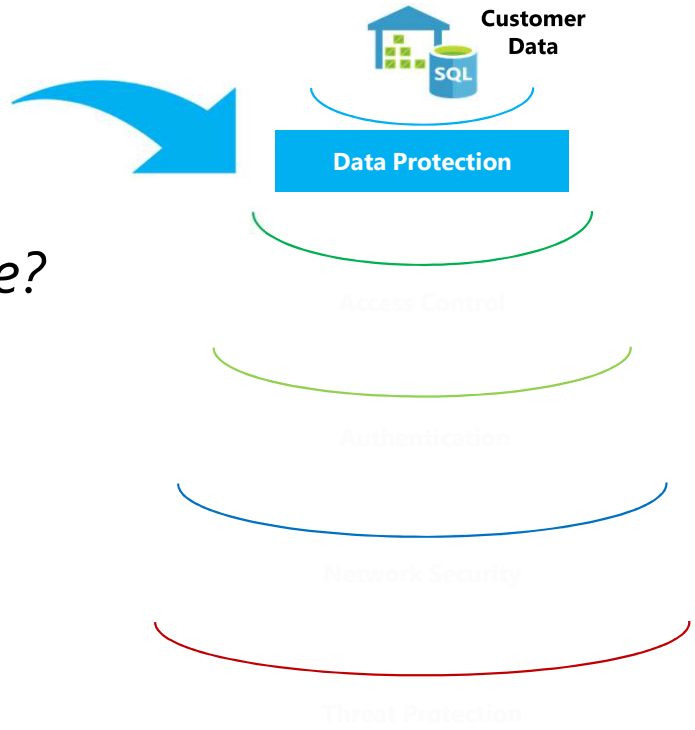
Administer via GRANT T-SQL statement

Both Azure Active Directory (AAD) and SQL authentication are supported



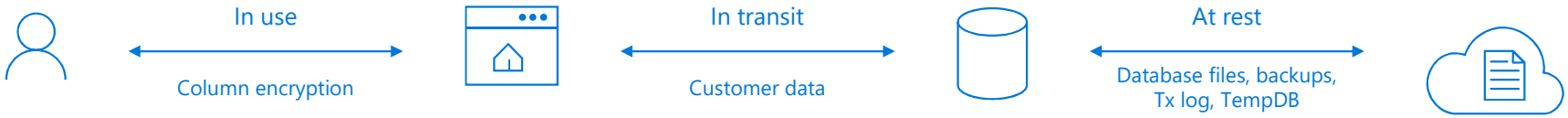
Business Requirements

- How do I protect sensitive data against unauthorized (high-privileged) users?
  - *What key management options do I have?*



Types of Data Encryption

Data encryption	Encryption technology	Customer value
In transit	Transport Layer Security (TLS) from the client to the server TLS 1.2	Protects data between client and server against snooping and man-in-the-middle attacks
At rest	Transparent Data Encryption (TDE) for Azure SQL Data Warehouse	Protects data on the disk User or Service Managed key management is handled by Azure, which makes it easier to obtain compliance



Disaster Recovery and High Availability built-in

Snapshots and Restores

Overview

Automatic copies of the data warehouse state

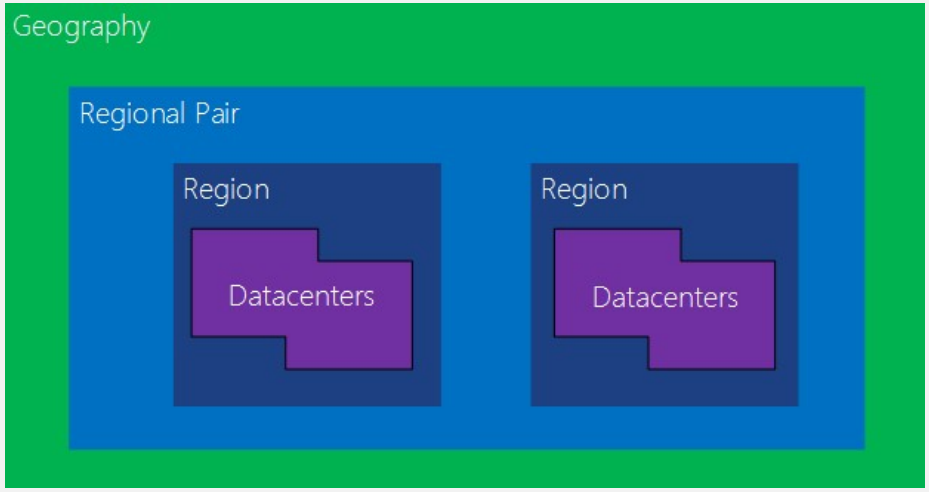
Taken throughout the day, or triggered manually

Available for up to 7 days, even after DW deletion

Regional restore in under 20 minutes, no matter data size

Snapshots and geo-backups allow cross-region restores

Automatic snapshots and geo-backups on by default



Benefits

Snapshots protect against data corruption and deletion

Restore to quickly create dev/test copies of data

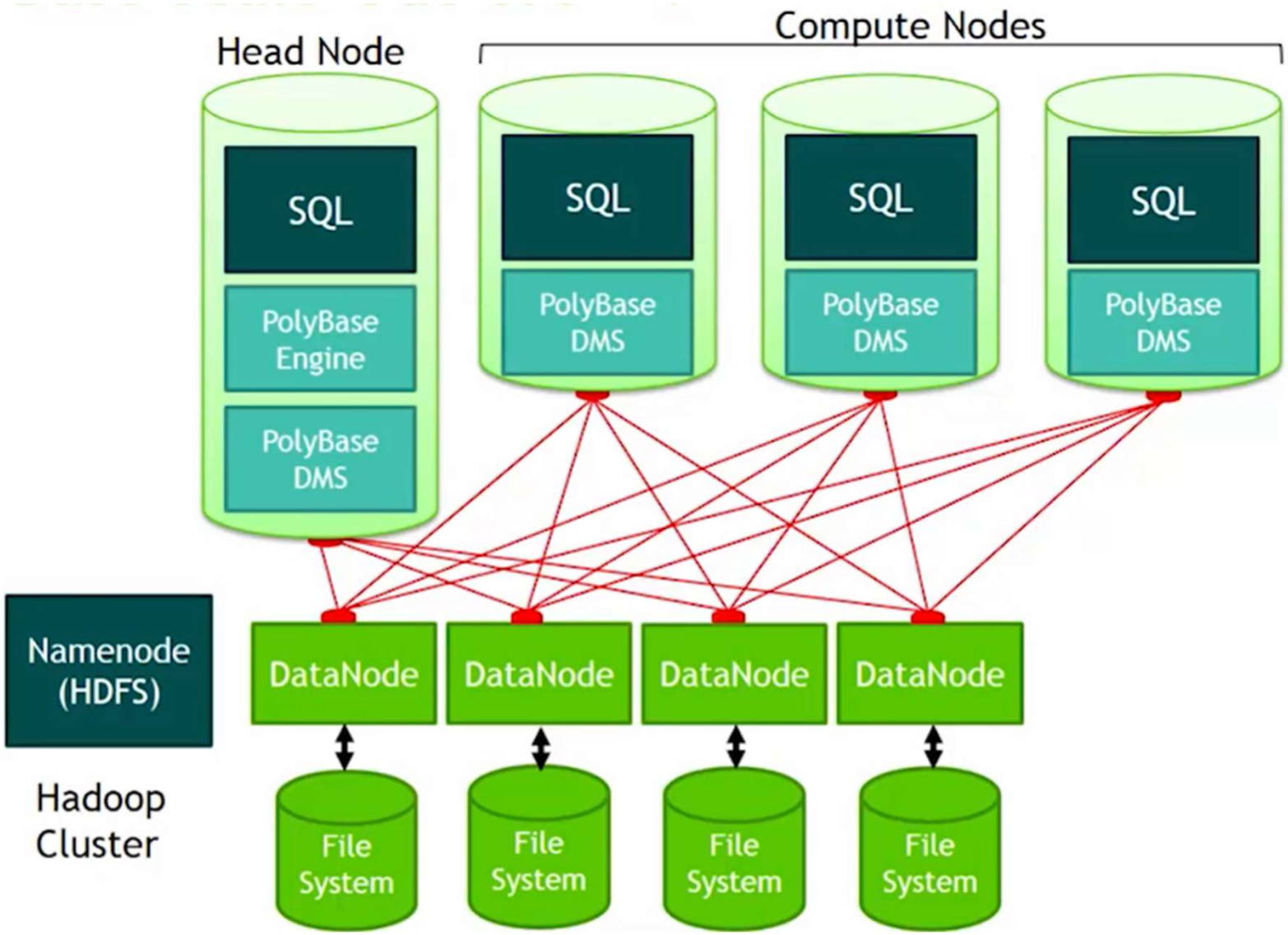
Manual snapshots protect large modifications

Geo-backups ensure data security in region loss

```
--View most recent snapshot time
SELECT top 1 *
FROM sys.pdw_loader_backup_runs
ORDER BY run_id DESC;
```

Data ingestion

Polybase



Polybase – External File Access

Overview

Polybase supports querying files stored in a Hadoop File System (HDFS), Azure Blob storage, or Azure Data Lake Store

To query files, users create three objects:

External data source, external file format, external table

```
-- Create Azure DataLake Gen2 Storage reference
CREATE EXTERNAL DATA SOURCE AzureStorage with
(
TYPE = HADOOP,
LOCATION='abfss://<containername>@<storageaccount>.blob.core
.windows.net',
CREDENTIAL = AzureStorageCredential
);
-- Type of format in Hadoop (CSV, RCFILE , ORC, PARQUET).
CREATE EXTERNAL FILE FORMAT TextFileFormat WITH
(
FORMAT_TYPE = DELIMITEDTEXT,
FORMAT_OPTIONS (FIELD_TERMINATOR ='|', USE_TYPE_DEFAULT =
TRUE)
)
-- LOCATION: path to file or directory that contains data
CREATE EXTERNAL TABLE [dbo].[CarSensor_Data]
(
[SensorKey] int NOT NULL,
[Speed] float NOT NULL,
[YearMeasured] int NOT NULL
)
WITH (LOCATION='/Demo/', DATA_SOURCE = AzureStorage,
FILE_FORMAT = TextFileFormat
);
```

Create Table As Select (CTAS)

Overview

The CTAS statement is a parallel operation that creates a new table in based the output of a SELECT statement.

It is the simplest and fastest way to create a copy of a table. Additionally, it allows for ingesting data stored in an external source into managed data warehouse tables.

```
-- Ingest external table data into data
warehouse
CREATE TABLE [dbo].[FactInternetSales]
WITH
(
    DISTRIBUTION = ROUND_ROBIN
,   CLUSTERED COLUMNSTORE INDEX
)
AS
SELECT  *
FROM    [staging].[FactInternetSales]
;
```

Copy Command

Overview

- Simplifies loading data into data warehouse
- Accesses directly from external sources
  - Data Lake Store Gen 2
  - Blob Storage
- Highly parallelized, scales with cluster compute
- No dependency on managed objects such as External Tables

```
--Create destination table in SQL DW
CREATE TABLE [dbo].[weatherTable]
(
  [ObservationTypeCode] [nvarchar](5) NOT NULL,
  [ObservationTypeName] [nvarchar](100) NOT NULL,
  [ObservationUnits] [nvarchar](5) NULL
)
WITH (
DISTRIBUTION = ROUND_ROBIN, HEAP);


--Copy files in parallel directly into datawarehouse table
COPY INTO [dbo].[weatherTable]
FROM 'abfss://<storageaccount>.blob.core.windows.net/<filepath>'
WITH (
FILE_FORMAT = 'DELIMITEDTEXT’,
SECRET = CredentialObject);
```

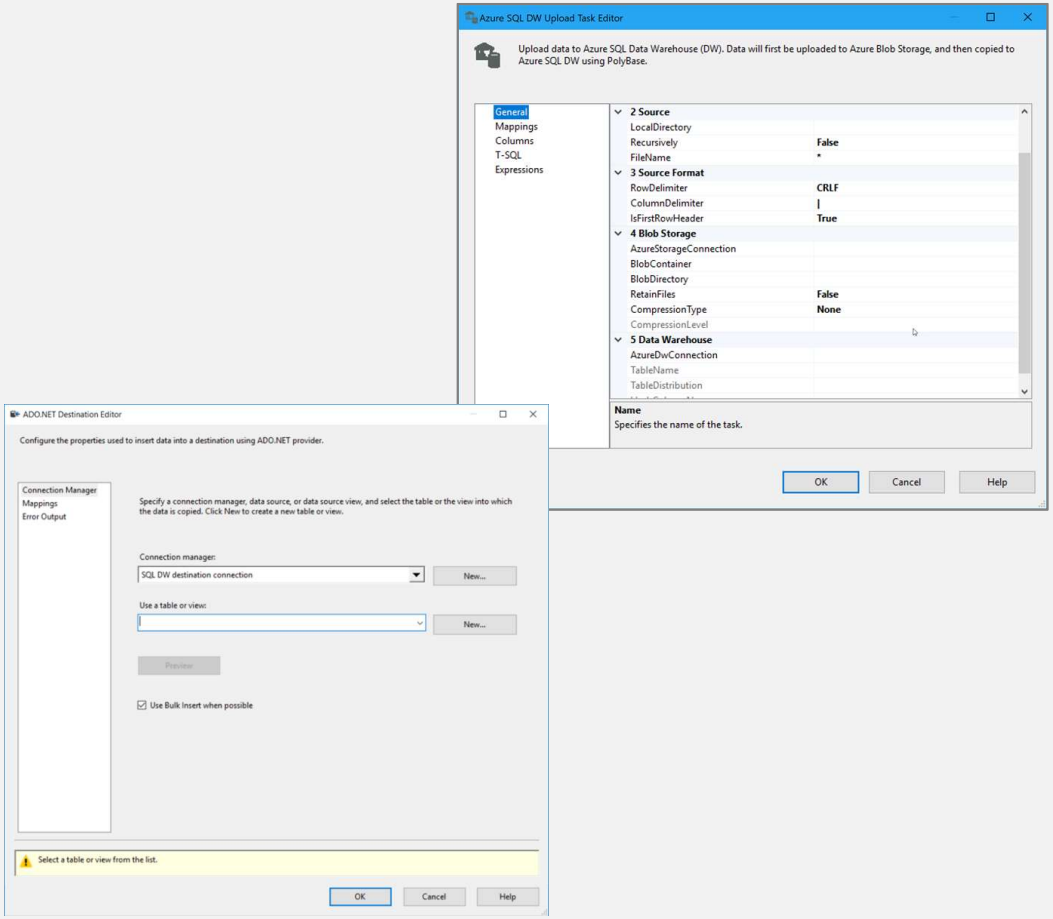
SQL Server Integration Services (SSIS)

Overview

SQL Server Integration Services is used to extract, transform data and load data from a variety of sources into Azure SQL Data Warehouse.

There are 2 options for loading data into SQL Data Warehouse with SSIS:

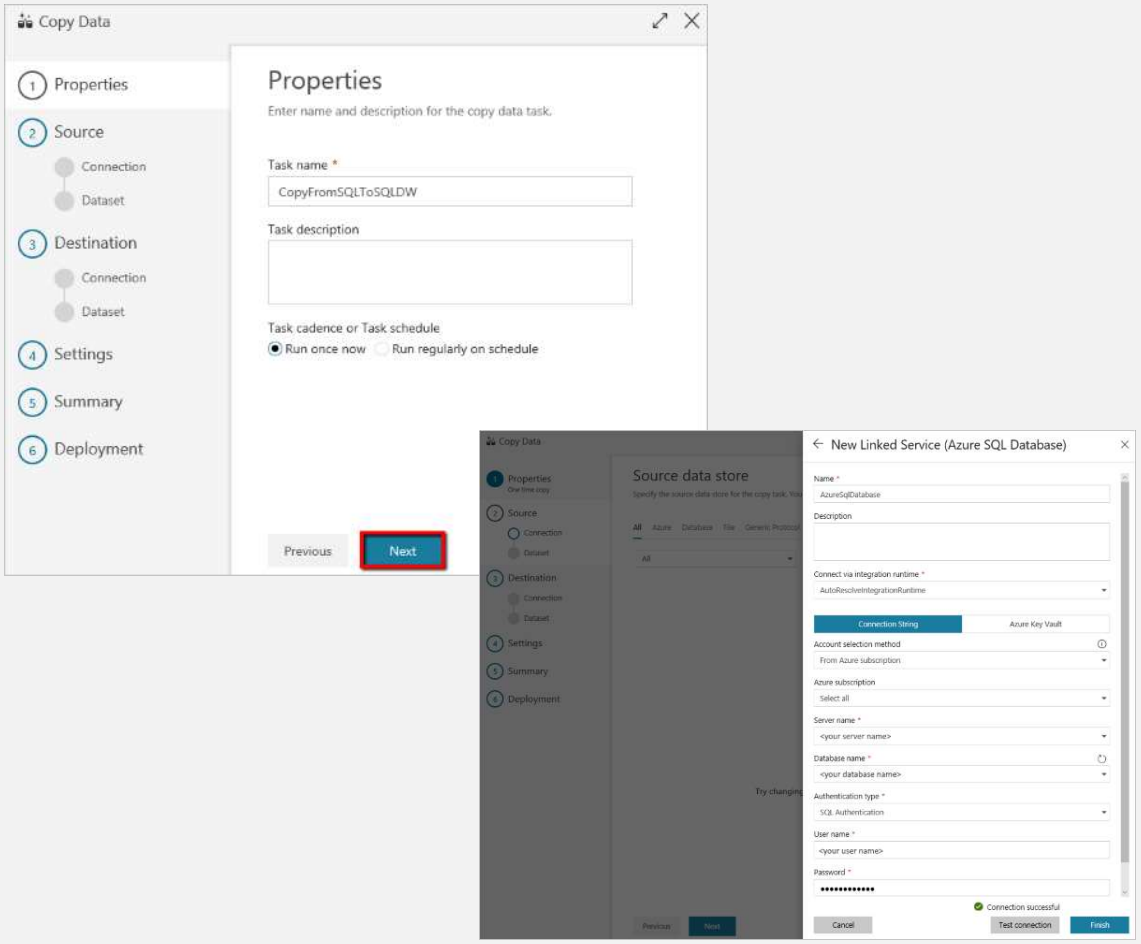
- **Azure SQL DW Upload Task**: provides best performance but assumes source data is in delimited text file format.
- **Data Flow Task**: slower than SQL DW upload task but supports a wider range of data sources.



Azure Data Factory Copy Data Tool

Overview

The Azure Data Factory Copy tool in the provides an intuitive wizard that allows you to copy data from a variety of data sources into Azure SQL Data Warehouse



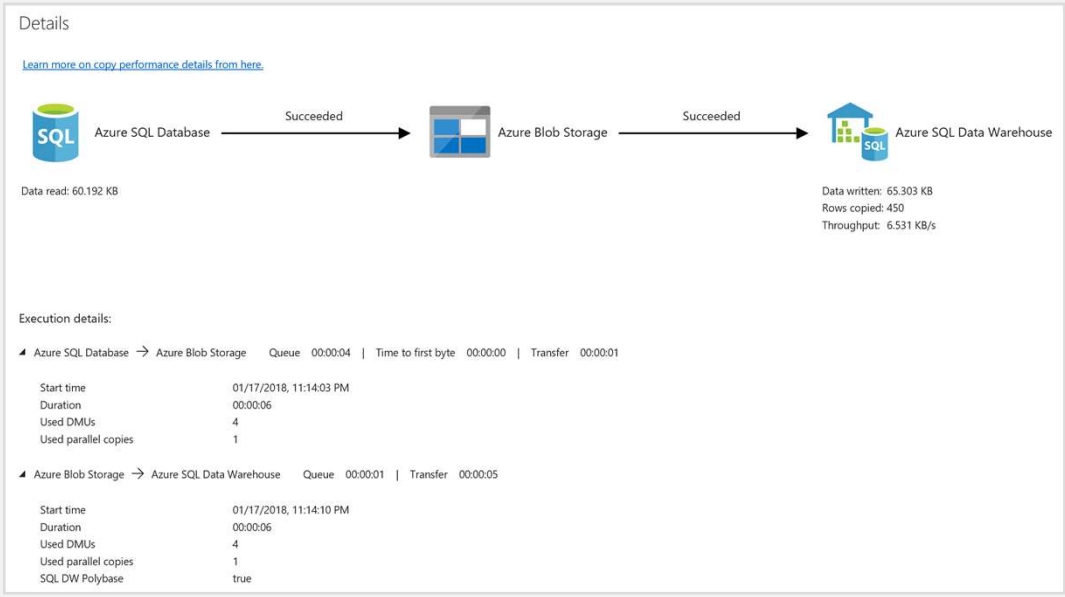
Azure Data Factory Copy Activity

Overview

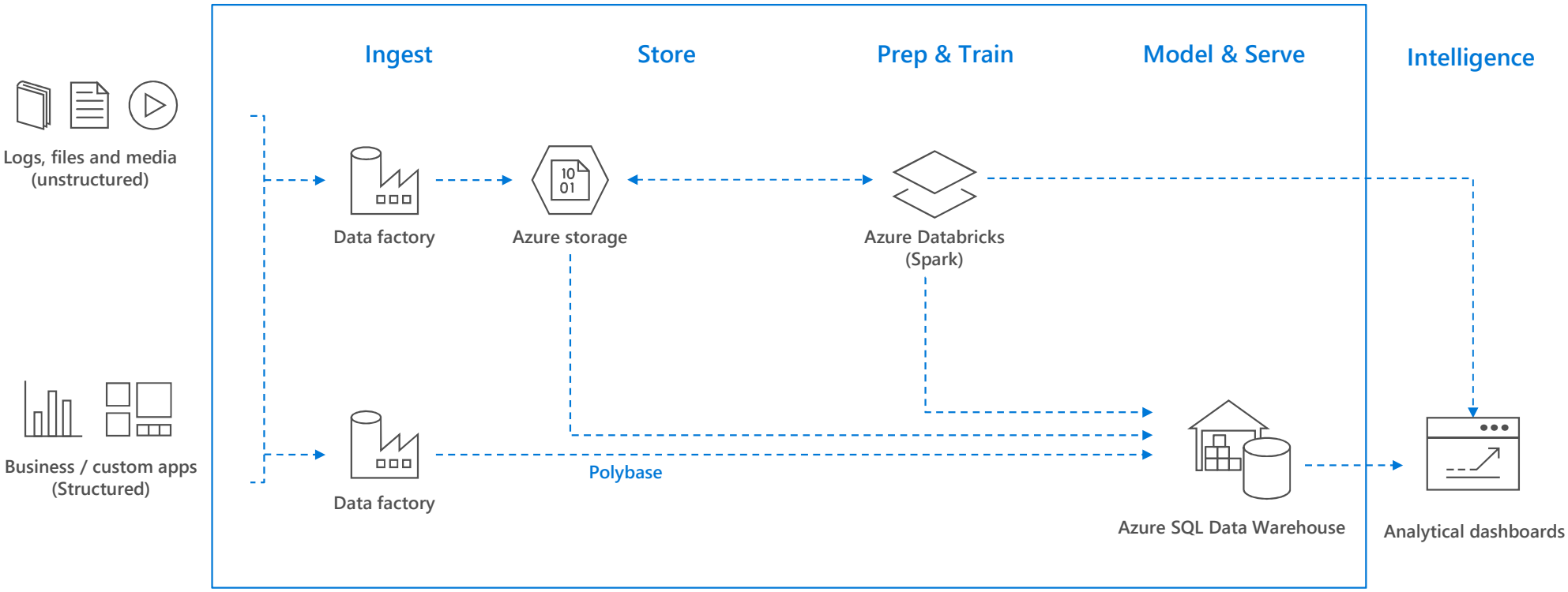
The Azure Data Factory Copy activity allows copy to and from Azure SQL Data Warehouse from any supported data store.

The copy activity also supports retrieving data from a SQL source by using a SQL query or stored procedure. Authentication can be via:

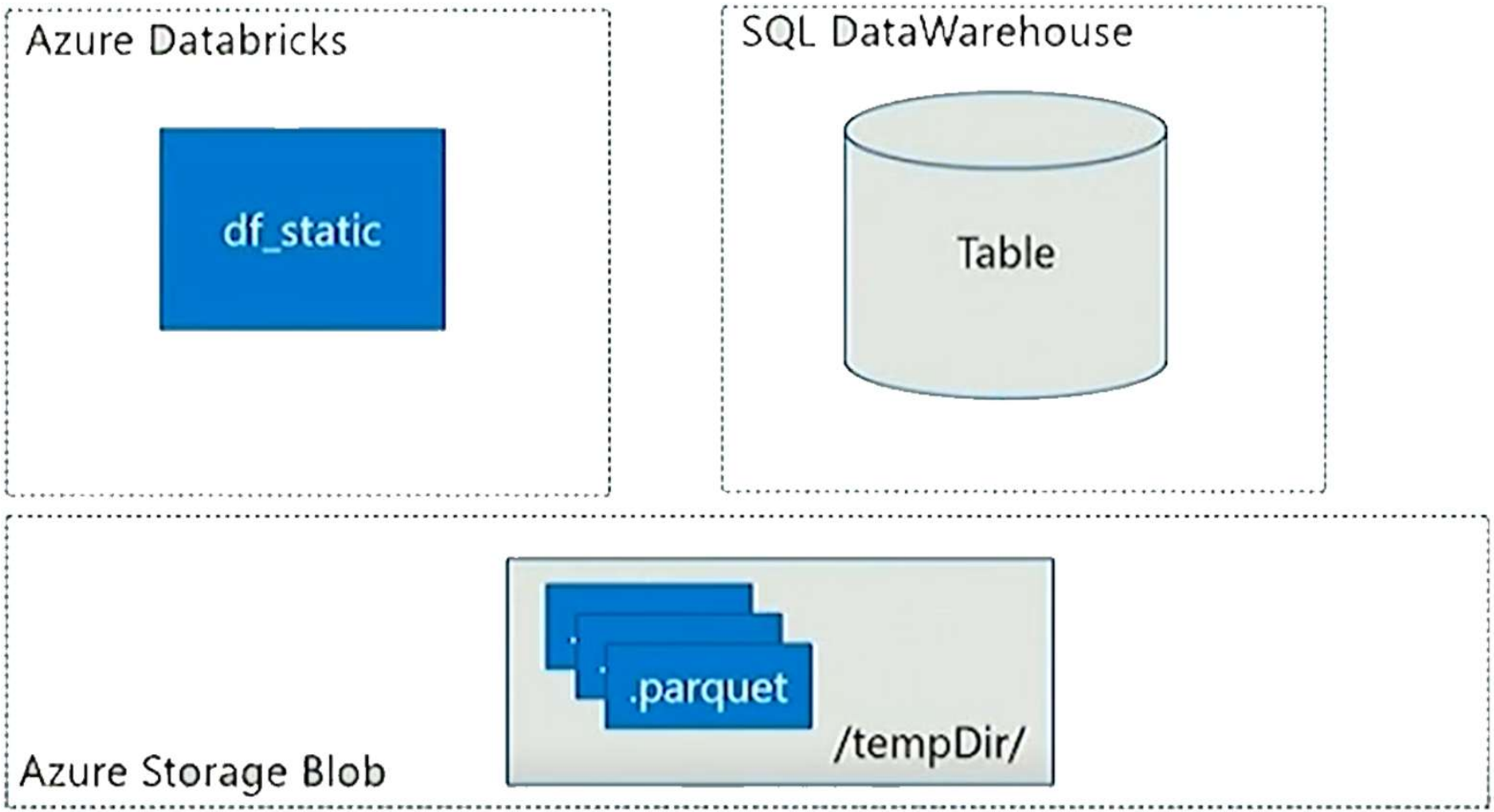
- SQL Authentication
- Service principal token authentication
- Managed identity token authentication



Modern Big Data Warehouse



Batch Data into Datawarehouse



Python for Batch Data Load

```
df.write \
  .format("com.databricks.spark.sqldw") \
  .option("url", "jdbc-url" ) \
  .option("tempDir", tempdir ) \
  .option("forwardSparkAzureStorageCredentials", "true") \
  .option("dbTable", "my_table_in_dw") \
  .save()
```

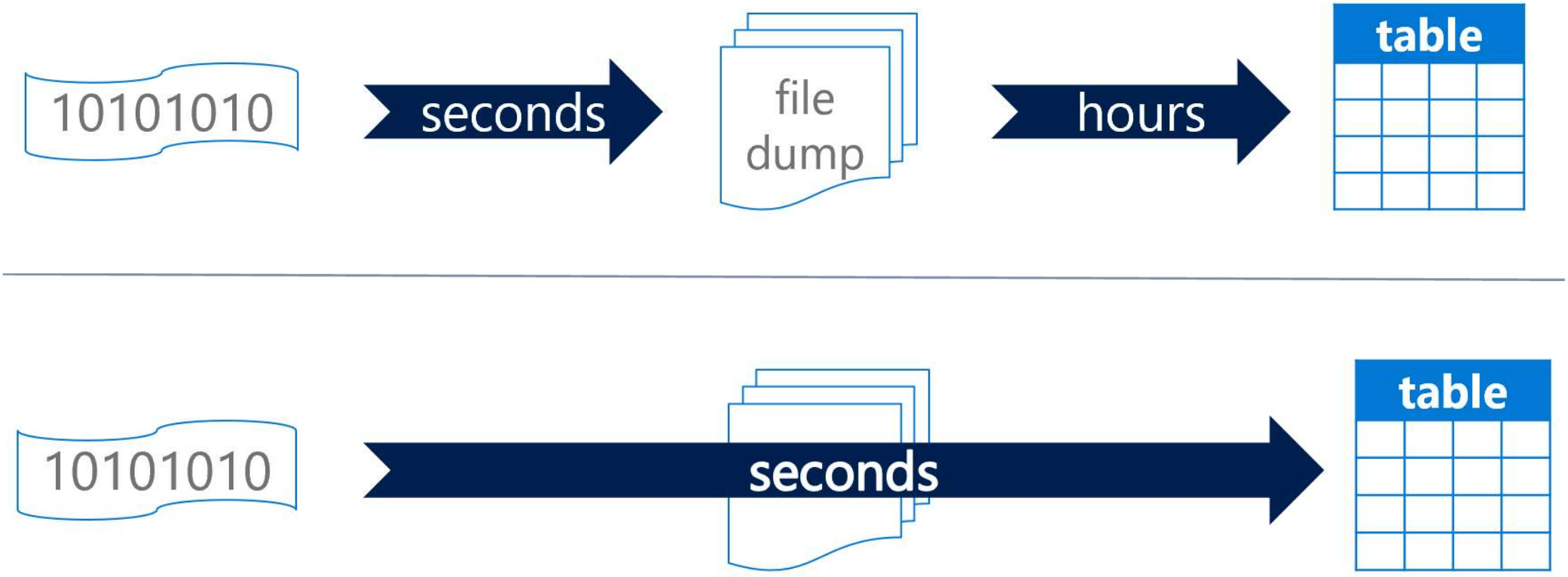
JDBC Connection String

Azure Storage Blob Directory

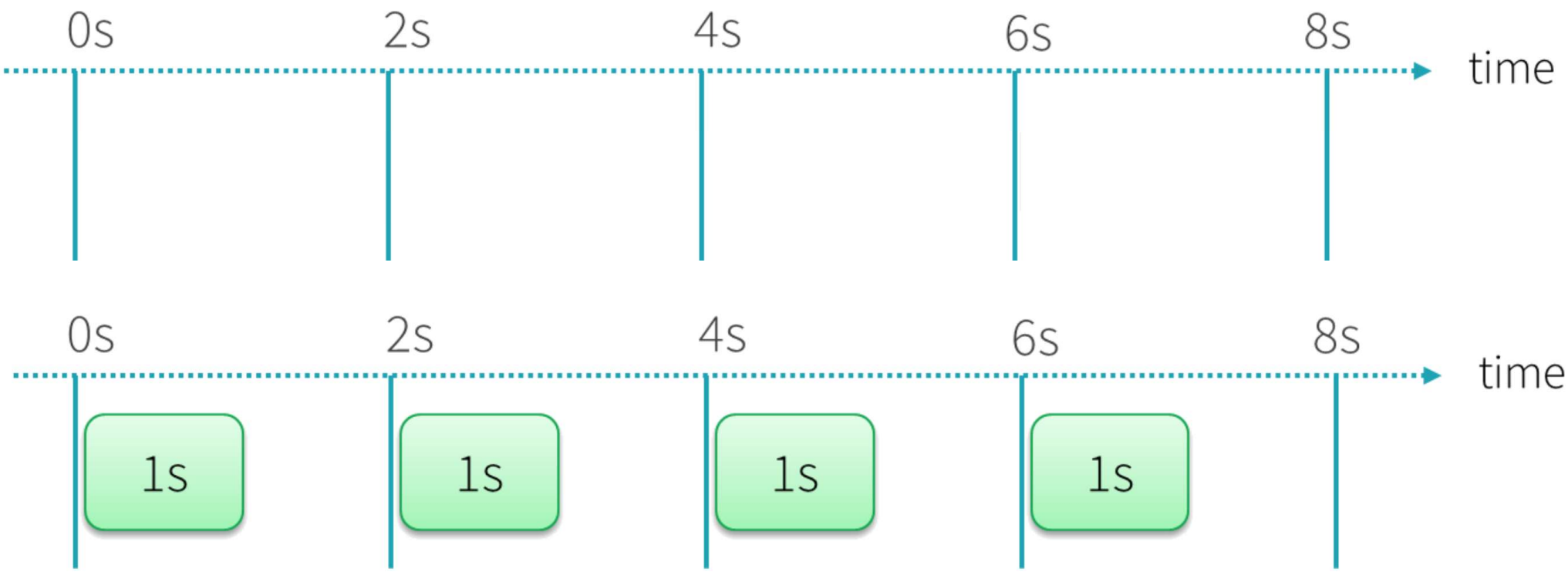
Introduction to Structure Streaming

The Problem ????

Streaming ETL w/ Structured Streaming



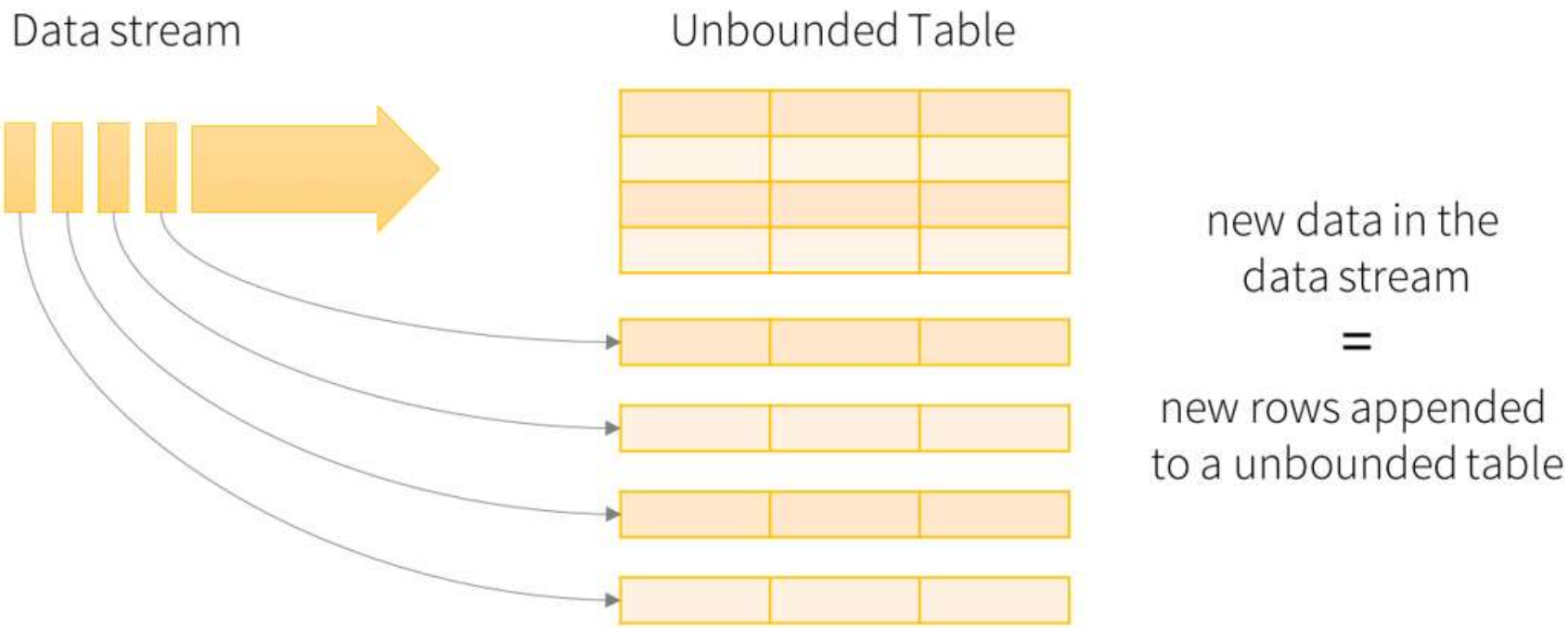
Micro Batch Model



Questions?

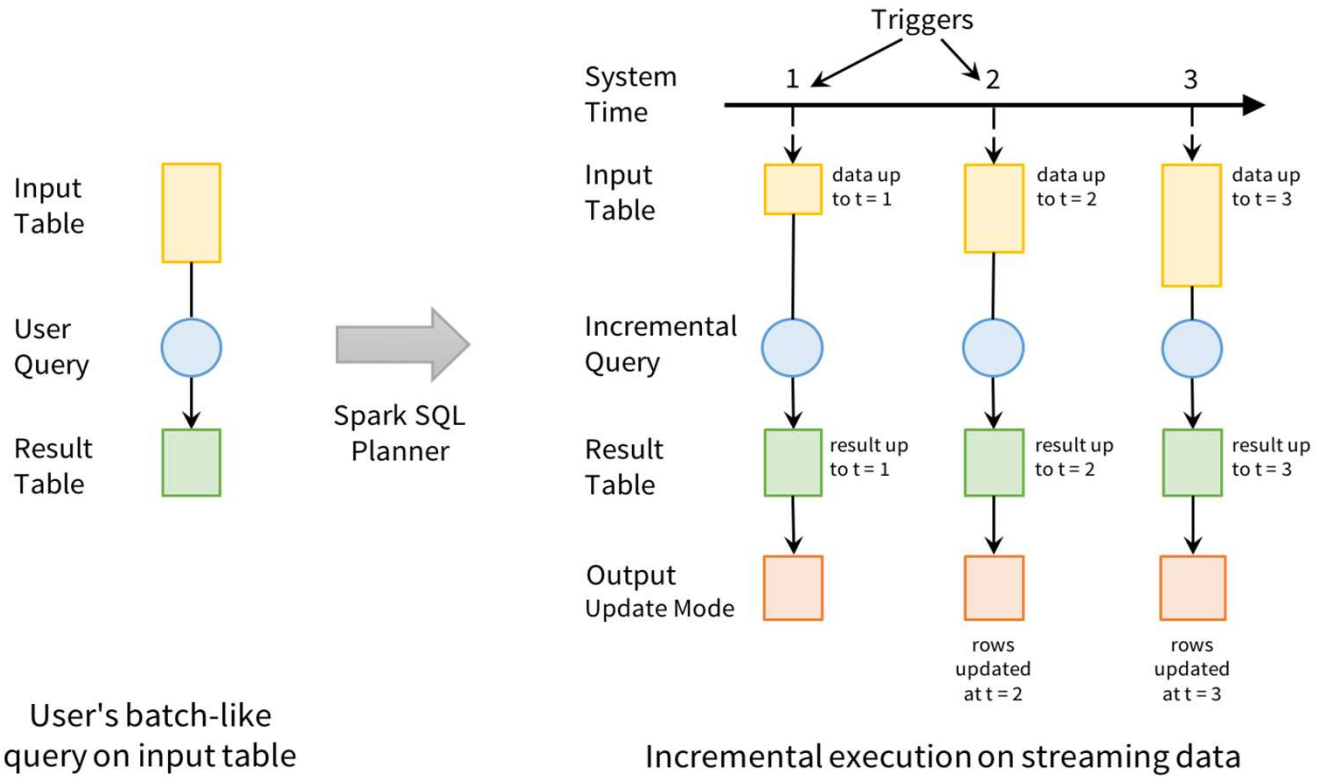
- **What happens if we don't process the data fast enough when reading from a TCP/IP Stream?**
- **What happens if we don't process the data fast enough when reading from a pubsub system like Kafka?**

From Micro Batch Model to Table



Data stream as an unbounded table

What is structured streaming?



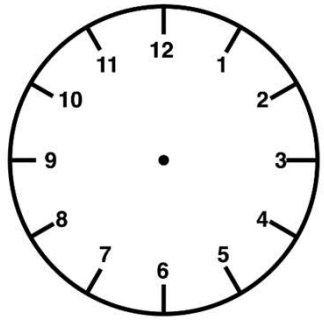
Structured Streaming Processing Model

Users express queries using a batch API; Spark incrementalizes them to run on streams

Question?

- **What problems might you encounter if you aggregate over a stream's entire dataset?**

Solution



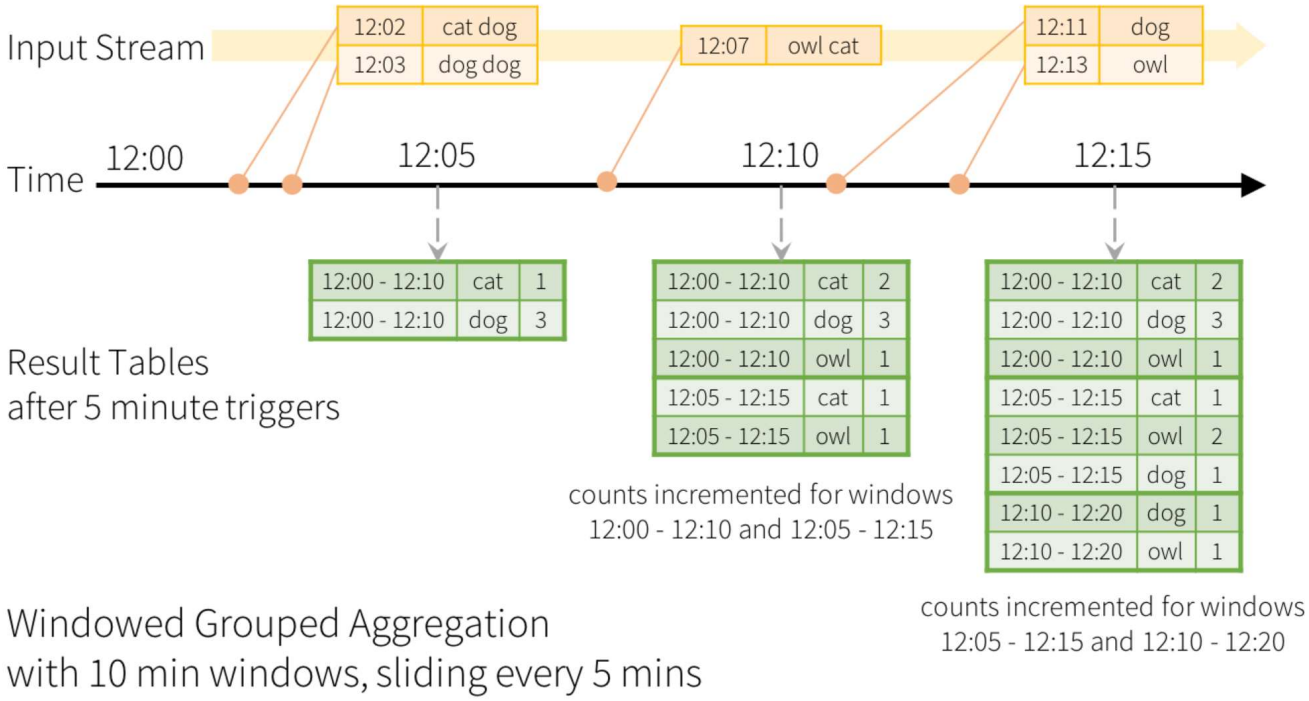
Working with time

Windowing Concepts

Output at the end of each window

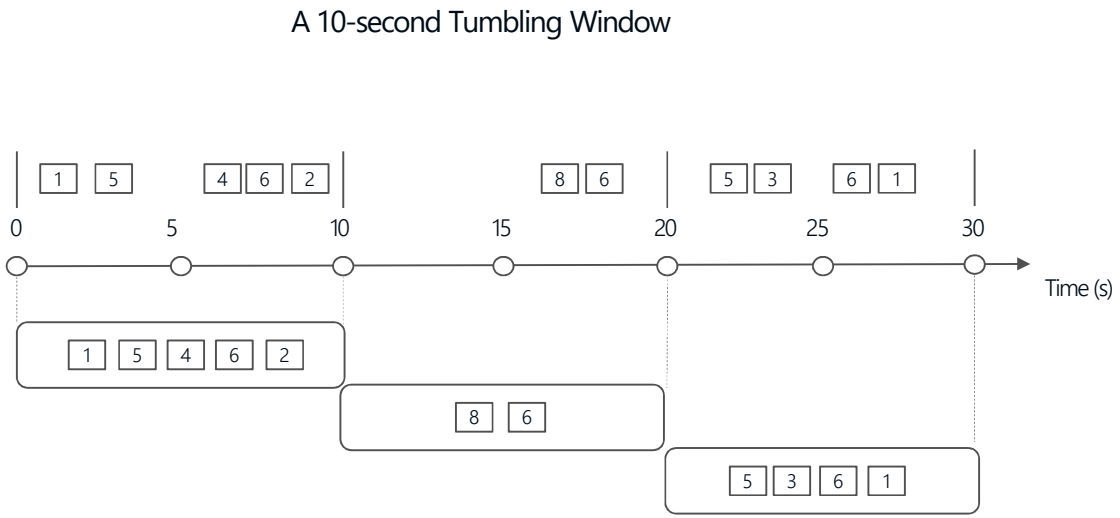
Windows are fixed length

Used in a GROUP BY clause



Tumbling Windows

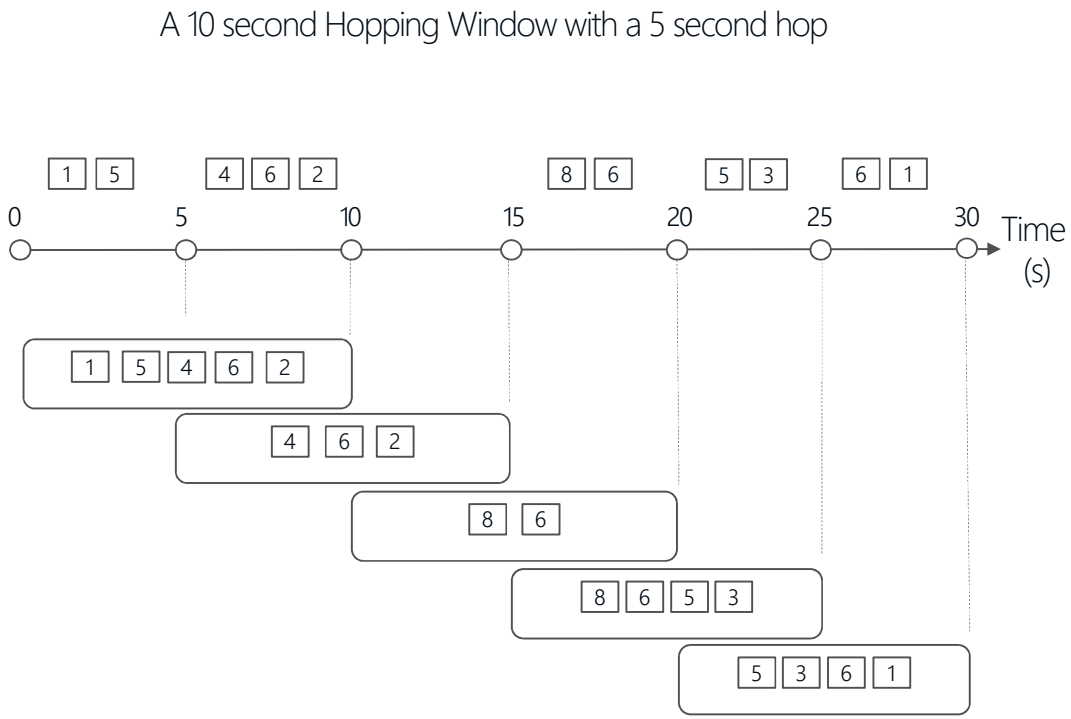
Every 10 seconds give me the count of vehicles entering each toll booth over the last 10 seconds




```
SELECT TollId, Count(*)
FROM EntryStream TIMESTAMP BY EntryTime
GROUP BY TollId, TumblingWindow(second, 10)
```

Hopping Windows

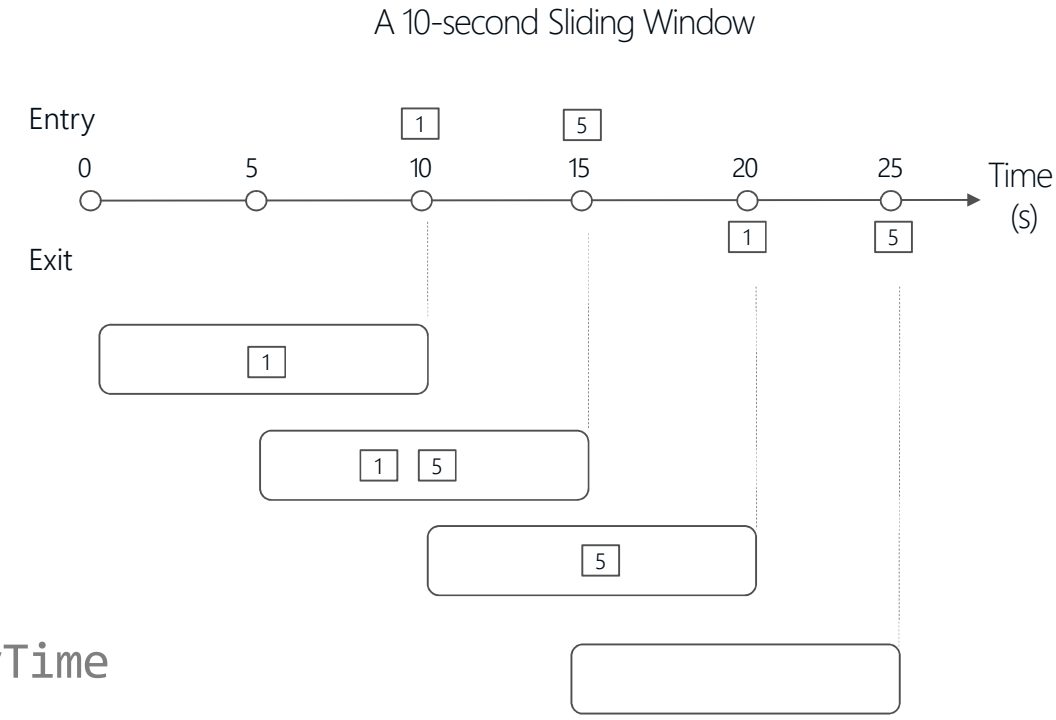
Every 5 seconds give me the count of vehicles entering each toll booth over the last 10 seconds




```
SELECT TollId, Count(*)
FROM EntryStream TIMESTAMP BY EntryTime
GROUP BY TollId, HoppingWindow(second, 10, 5)
```

Sliding Windows

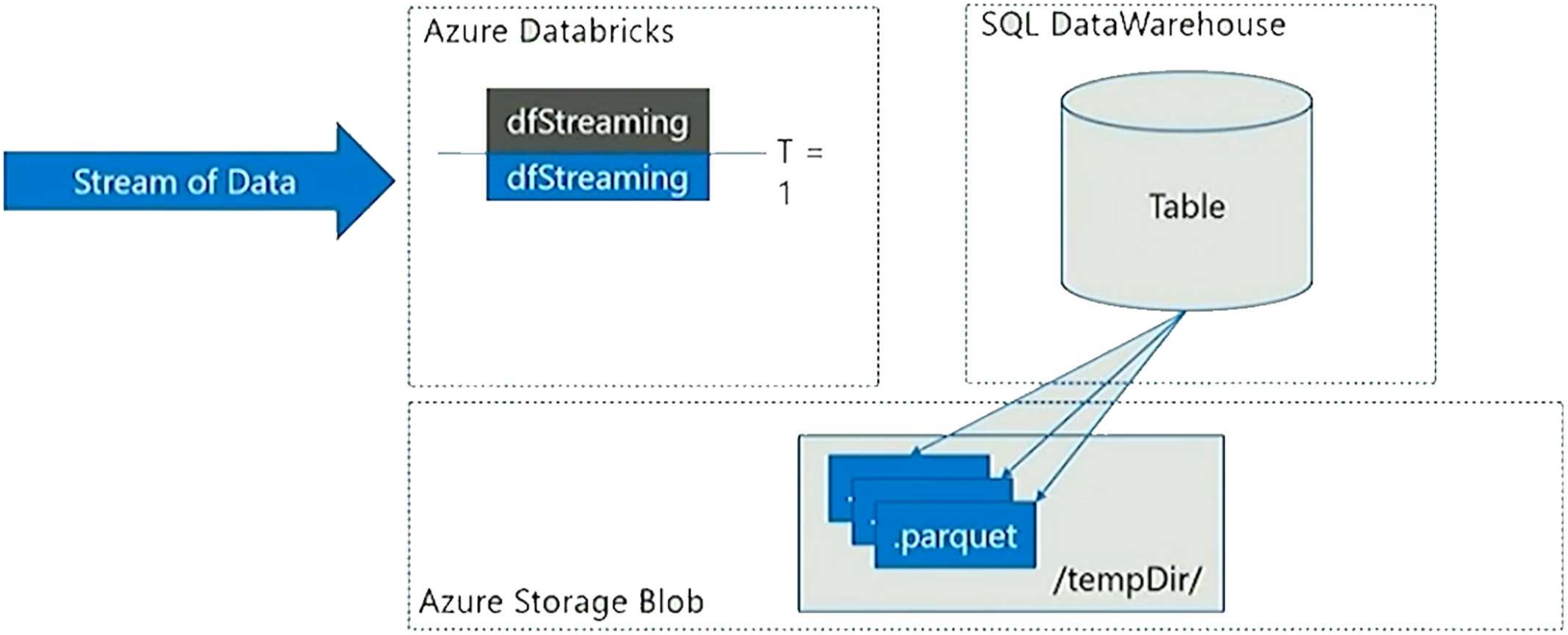
Find all toll booths that have served more than 10 vehicles in the last 20 seconds




```
SELECT TollId, Count(*)
FROM EntryStream TIMESTAMP BY EntryTime
GROUP BY TollId, SlidingWindow(second, 20)
HAVING Count(*) > 10
```

An output is generated whenever an event either enters/leaves the system

Structured streaming into Data Warehouse



Python for structured streaming load

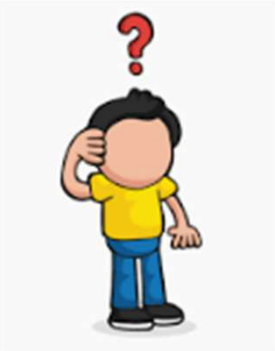
```
df.writeStream \
  .format("com.databricks.spark.sqldw") \
  .option("url", "jdbc-url" ) \
  .option("tempDir", tempdir ) \
  .option("forwardSparkAzureStorageCredentials", "true") \
  .option("dbTable", "my_table_in_dw") \
  .option("chkpointLoc", "/tmp_loc") \
  .save()
```

JDBC Connection String

Azure Storage Blob Directory

Introduction to Databricks Deltas

Services in Azure



Errr….
What are all these?
Why so many?
When do I use these?
Wait, now you have this new thing called *Delta*?
I am dizzy and ready to pass out!

MPP database

Azure SQL Datawarehouse

Object store

Azure Data Lake Store

Azure Blob Storage

Processing Engine

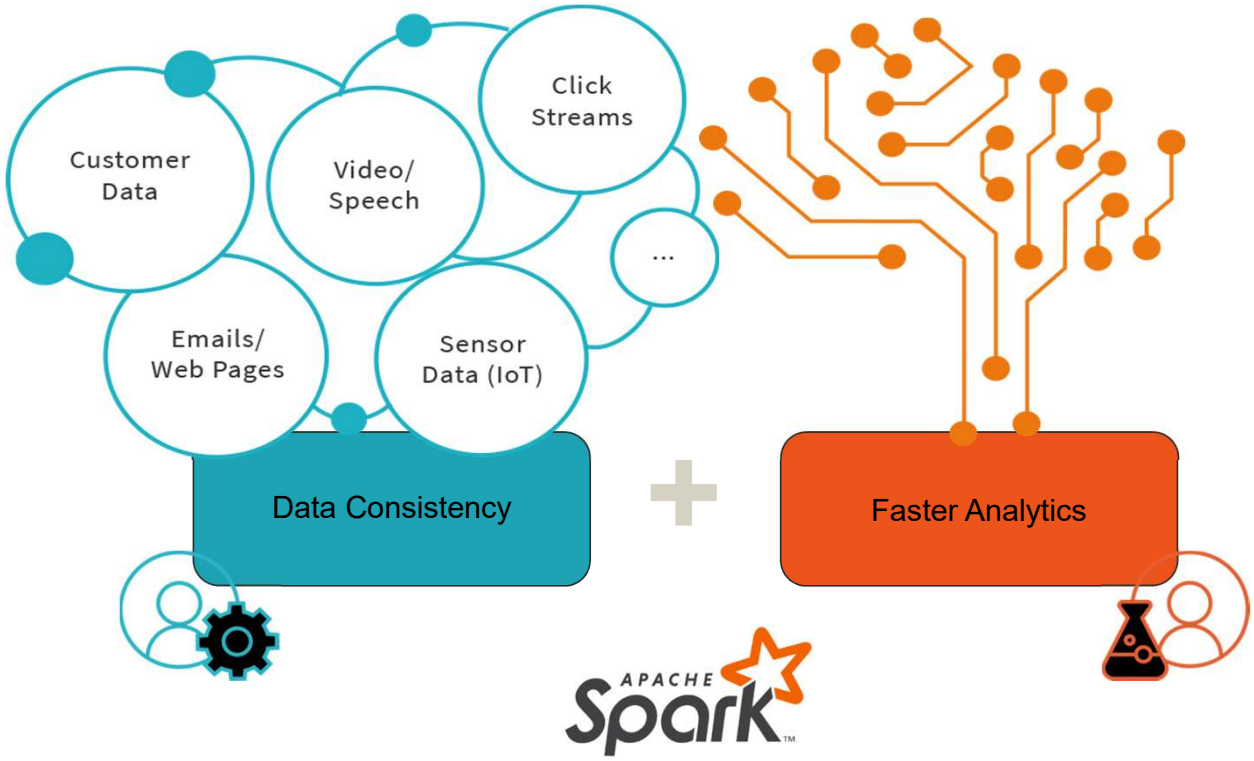
Azure Databricks

HDInsight

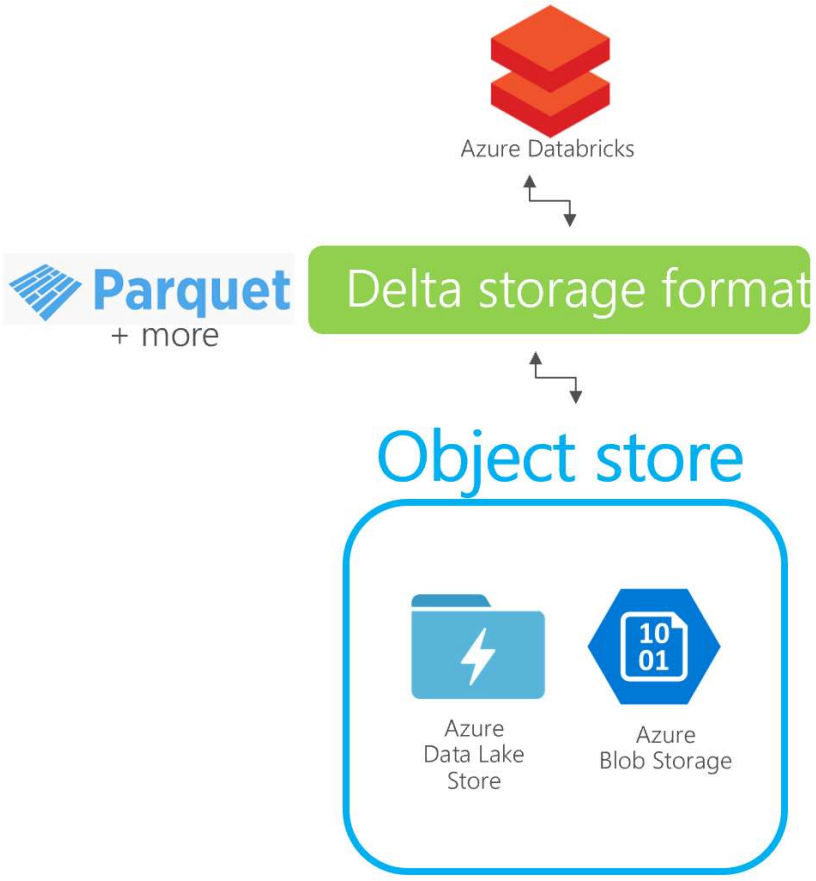
Azure Data Lake Analytics

Databricks Delta

Extends Apache Spark to simplify data reliability and boost performance



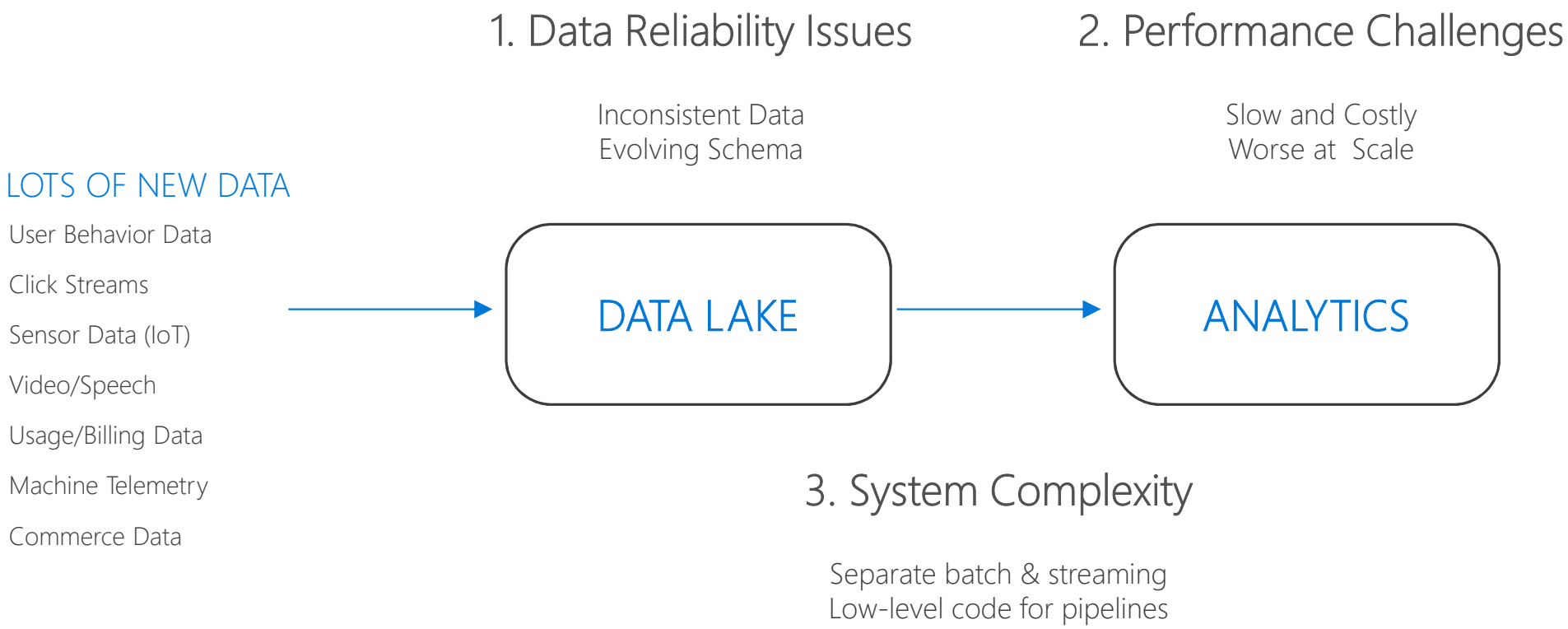
Databricks Delta fundamentally?



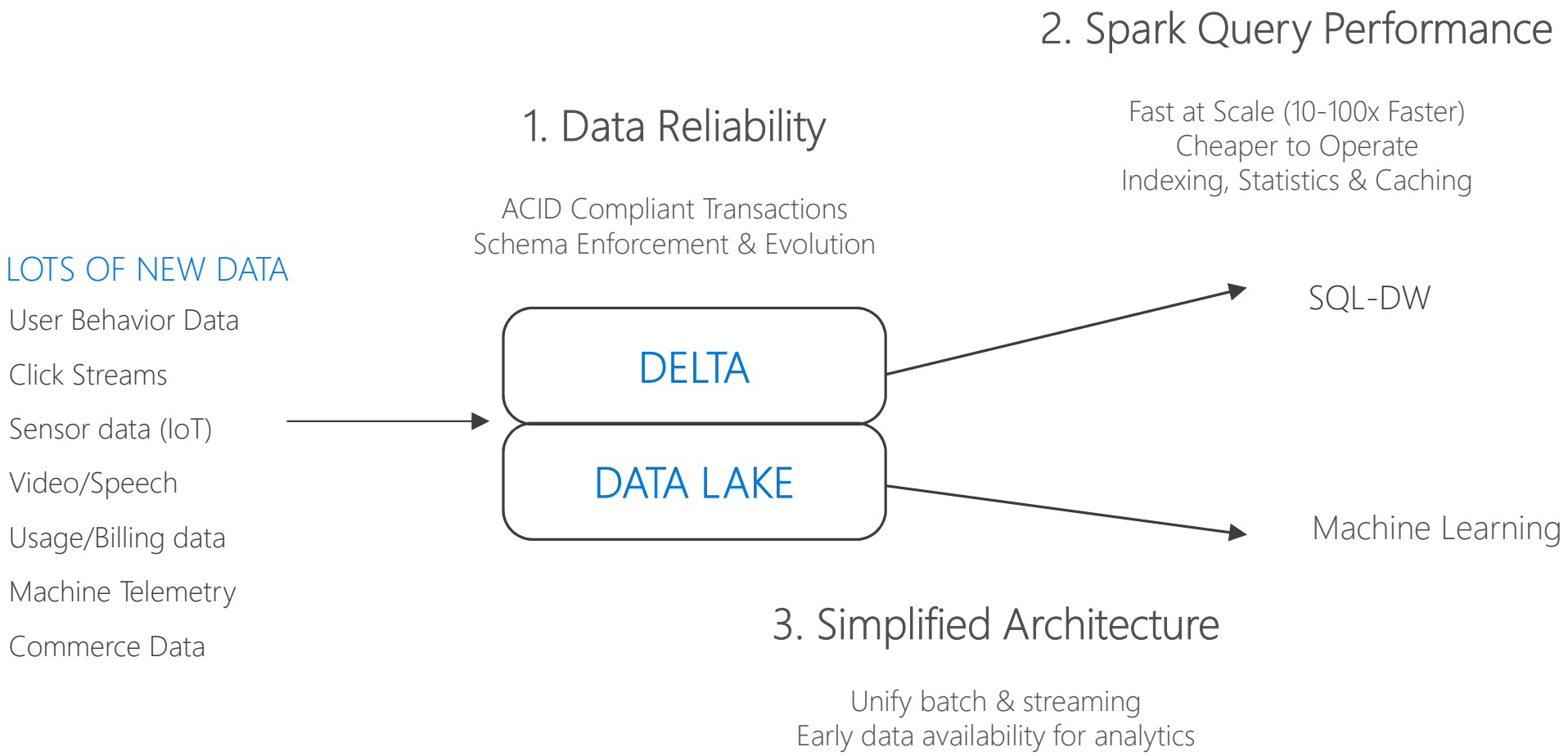
ONLY ON Databricks – not on OSS Spark

Databricks Delta – Unified data management system

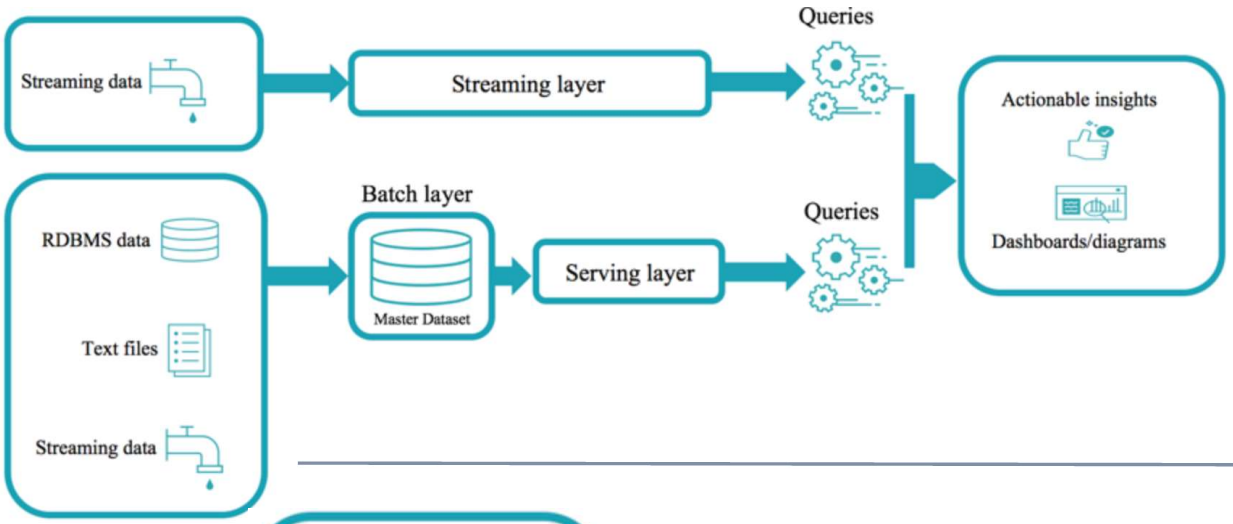
Why Databricks Delta?

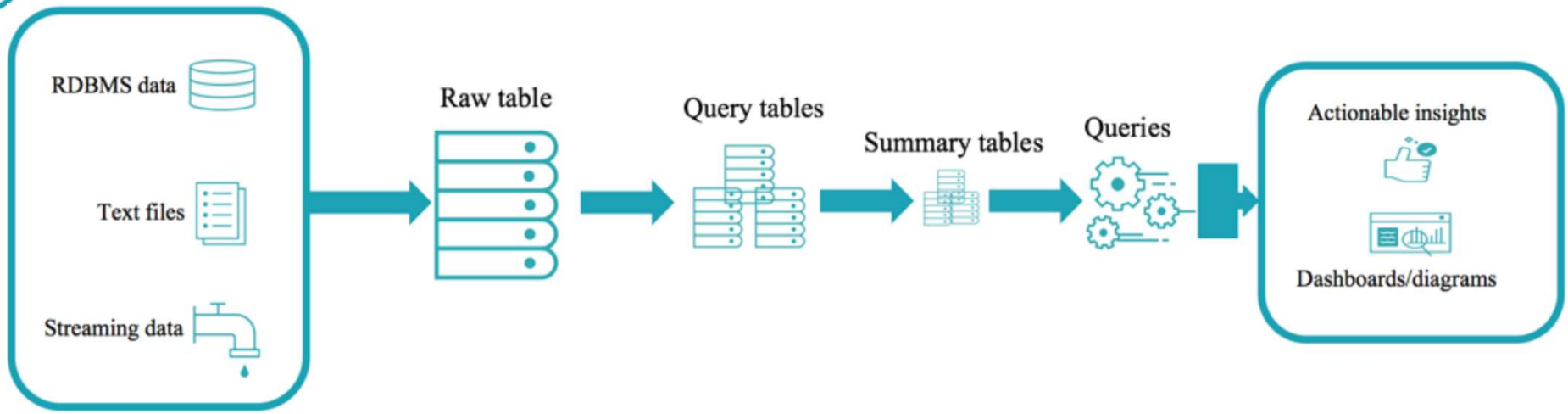


Problems Solved



Lambda architecture vs Databricks Delta





Multi-Hop Data Pipelines

LOTS OF NEW DATA

User Behavior Data

Click Streams

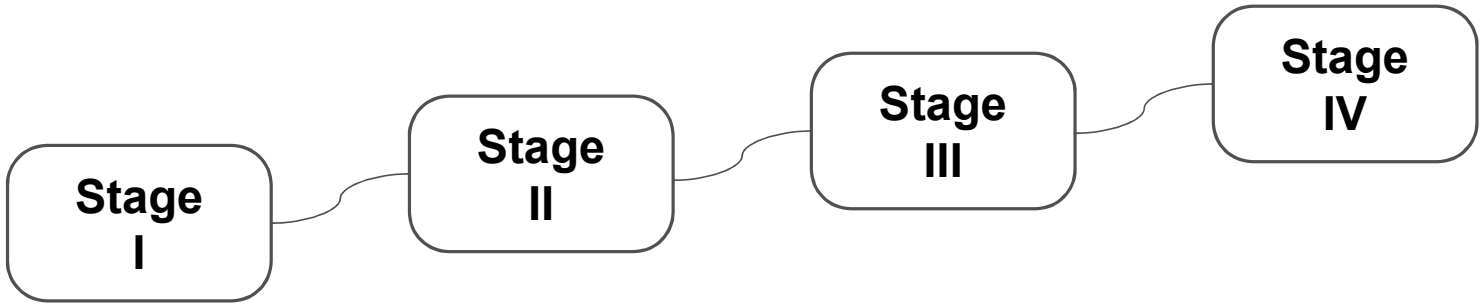
Sensor Data (IoT)

Video/Speech

Usage/Billing Data

Machine Telemetry

Commerce Data



- Stage I - Raw events from many different parts of the organization
- Stage II - Normalized and enriched with dimension information
- Stage III - Filtered down and aggregated for particular business objective.
- Stage IV - High-level summaries of key business metrics.

Challenges solved

Problem:
Failed production jobs leave data in corrupt state requiring tedious recovery

Solution:
Failed write jobs do not update the commit log, hence partial / corrupt files not DELTA visible to readers

Problem :
Lack of consistency makes it almost impossible to mix appends, deletes, upserts and get consistent reads

Solution:
All reads have full snapshot consistency
All successful writes are consistent
In practice, most writes don't conflict
Tunable isolation levels

Challenges solved

Challenge :
Lack of schema enforcement creates inconsistent and low quality data

Solution:
Schema recorded in the log
Fails attempts to commit data with incorrect schema
Allows explicit schema evolution
Allows invariant and constraint checks (high data quality)

Challenge:
Too many small files increase resource usage significantly

Solution:
Transactionally performed compaction using OPTIMIZE

Questions?