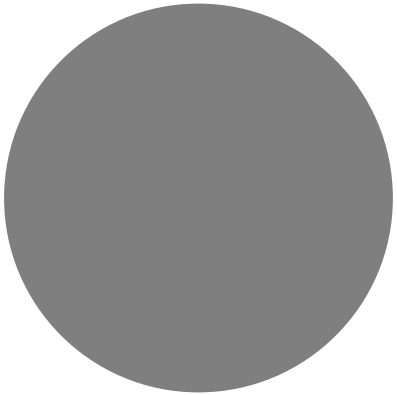
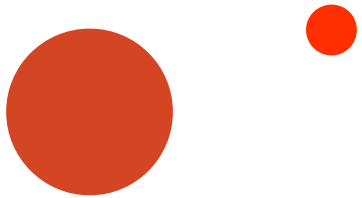
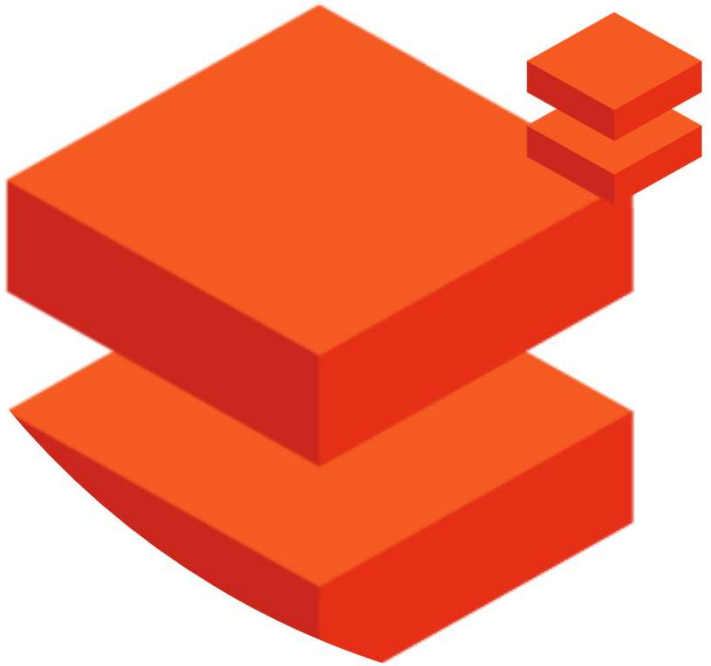
FAQ.pdf





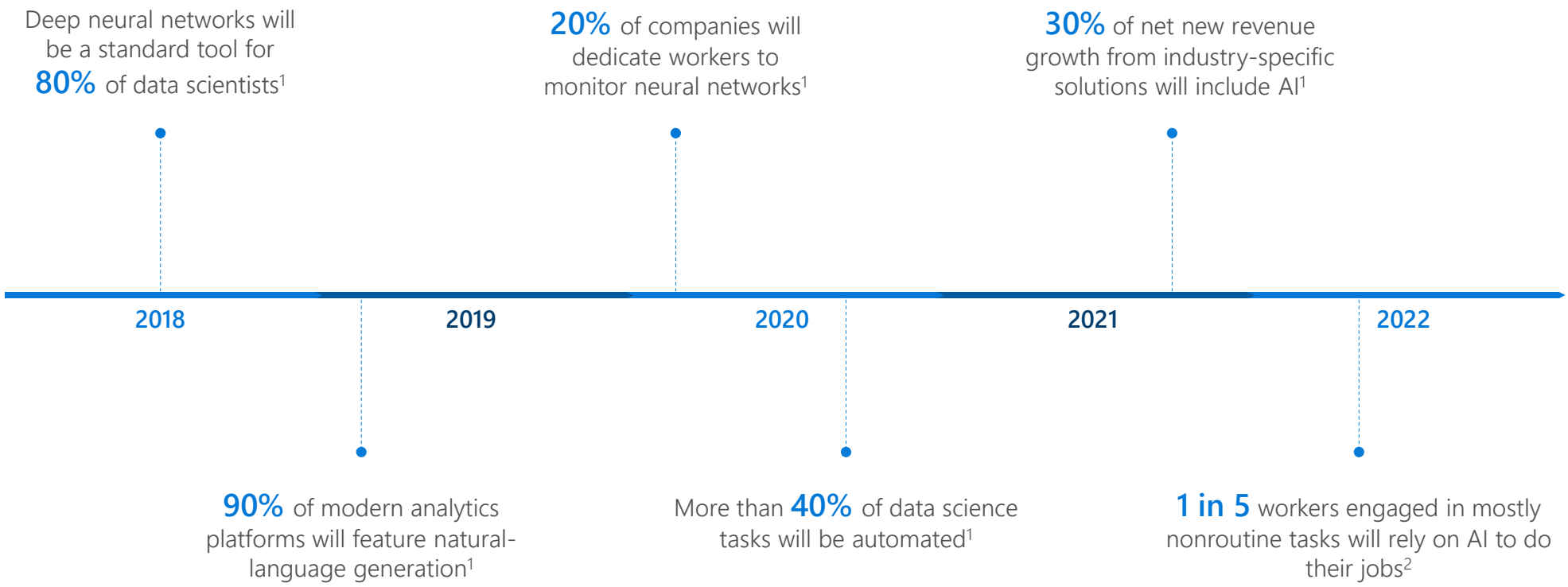


Azure Databricks – Curated

Naveed Hussain

Global CSA for Data & AI in OCP

What are companies looking to do next?



[1] "100 Data and Analytics Predictions Through 2021", Gartner, 2017. [2] "Predicts 2018: AI and the Future of Work", Gartner, 2018.

Hardest Part of AI isn't AI, it's Data

"Hidden Technical Debt in Machine Learning Systems," Google NIPS 2015

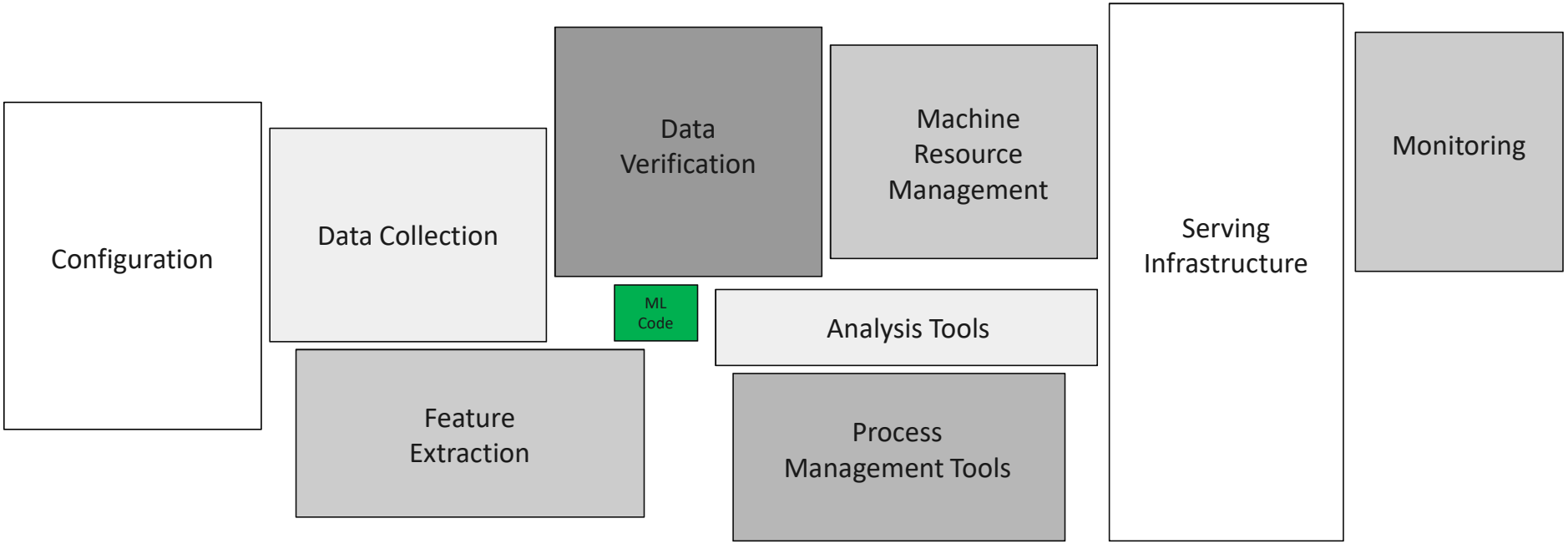


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

Financial services use cases

Effective customer engagement

Customer profiles
Credit history
Transactional data
LTV
Loyalty

Customer analytics

Customer 360 degree evaluation
Customer segmentation
Reduced customer churn
Underwriting, servicing and delinquency handling
Insights for new products

Faster innovation for a better customer experience

Decision services management

Customer segmentation
CRM data
Credit data
Market data

Financial modeling

Commercial/retail banking, securities, trading and investment models
Decision science, simulations and forecasting
Investment recommendations

Improved consumer outcomes and increased revenue

Risk and revenue management

Transaction data
Demographics
Purchasing history
Trends

Risk, fraud, threat detection

Real-time anomaly detection
Card monitoring and fraud detection
Security threat identification
Risk aggregation

Enhanced customer experience with machine learning

Risk and compliance management

CRM
Credit
Risk
Merchant records
Products and services

Credit analytics

Enterprise DataHub
Regulatory and compliance analysis
Credit risk management
Automated credit analytics

Transform growth with predictive analytics

Recommendation engine

Clickstream data
Products
Services
Customer service data

Marketing analytics

Recommendation engine
Predictive analytics and targeted advertising
Fast marketing and multi-channel engagement
Customer sentiment analysis

Improved customer engagement with machine learning

Health and life sciences use cases

DNA sequences	Real world analytics	Image deep learning	Sensor data	Social data listening
FAST-Q	HL7/CCD	MRI	Readings	Social media
BAM	837	X-RAY	Time series	Adverse events
SAM	Pharmacy	CT	Event data	Unstructured
VCF	Registry	Ultrasound		
Expression	EMR			

Genomics and precision medicine

- Single cell sequencing
- Biomarker, genetic, variant and population analytics
- ADAM and HAIL on Databricks

Faster innovation for drug development

Clinical and claims data

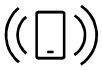
- Claims data warehouse
- Readmission predictions
- Efficacy and comparative analytics
- Prescription adherence
- Market access analysis

Improved outcomes and increased revenue

GPU image processing

- Graphic intensive workloads
- Deep learning using Tensor Flow
- Pattern recognition

Diagnostics leveraging machine learning



IoT device analytics

- Aggregation of streaming events
- Predictive maintenance
- Anomaly detection

Predictive analytics transforms quality of care

Social analytics

- Real-time patient feedback via topic modelling
- Analytics across publication data

Improved patient communications and feedback

Media and entertainment use cases

Personalized recommendations

Customer profiles
Viewing history
Online activity
Content sources
Channels

Content personalization

Personalized viewing and engagement experience
Click-path optimization
Next best content analysis
Improved real time ad targeting

Faster innovation for customer experience

Effective customer retention

Customer profiles
Online activity
Content distribution
Services data

Customer churn prevention

Quality of service and operational efficiency
Market basket analysis
Customer behavior analysis
Click-through analysis

Improved consumer outcomes and increased revenue

Information optimization

Consumption logs
Clickstream and devices
Marketing campaign responses

Recommendation engine

Ad effectiveness
Content monetization
Fraud detection
Information-as-a-service
High value user engagement

Enhance user experience with machine learning

Inventory allocation

Transactions
Subscriptions
Demographics
Credit data

Predictive analytics

Predict audience interests
Network performance and optimization
Pricing predictions
Nielsen ratings and projections
Mobile spatial analytics

Predictive analytics transforms growth

Consumer engagement analysis

Content metadata
Ratings
Comments
Social media activity

Sentiment analysis

Demand-elasticity
Social network analysis
Promotion events time-series analysis
Multi-channel marketing attribution

Improved consumer engagement with machine learning

Retail use cases

Recommendation engine

Customer profiles
Shopping history
Online activity
Social network analysis

Next best and personalized offers

Customer 360/consumer personalization
Right product, promotion, at right time
Multi-channel promotion

Faster innovation for customer experience

Effective customer engagement

Shopping history
Online activity
Floor plans
App data

Store design and ergonomics

Path to purchase
In-store experience
Workforce and manpower optimization

Improved consumer outcomes and increased revenue

Inventory optimization

Demand plans
Forecasts
Sales history
Trends
Local events/weather patterns

Data-driven stock, inventory, ordering

Predict inventory positions and distribution
Fraud detection
Market basket analysis

Omni-channel shopping experience with machine learning

Inventory allocation

Demographics
Buyer perception
Consumer research
Market/competitive analysis

Assortment optimization

Economic modelling
Optimization for foot traffic, Online interactions
Flat and declining categories

Predictive analytics transforms growth

Consumer engagement

Historical sales data
Price scheduling
Segment level price changes

Real-time pricing optimization

Demand-elasticity
Personal pricing schemes
Promotion events
Multi-channel engagement

Improved consumer engagement with machine learning

Advertising and marketing tech use cases

Effective customer engagement

Customer profiles
Online history
Transaction data
Loyalty

Customer value analytics

Customer 360, segmentation aggregation and attribution
Audience modelling/index report
Reduce customer churn
Insights for new products
Historical bid opportunity as a service

Faster innovation for customer growth

Recommendation engine

Customer segmentation
CRM data
Credit data
Market data

Next best and personalized offers

Right product, promotion, at right time
Real time ad bidding platform
Personalized ad targeting
Ad performance reporting

Improved outcomes and increased revenue

Risk and fraud analysis

Transaction data
Demographics
Purchasing history
Trends

Risk and fraud management

Real-time anomaly detection
Fraud prevention
Customer spend and risk analysis
Data relationship maps

Risk management with machine learning

Campaign reporting analytics

CRM
Merchant records
Products
Services
Marketing data

Sales and campaign optimization

Optimizing return on ad spend and ad placement
Multi-channel promotion
Ideal customer traits
Optimized ad placement

Predictive analytics transforms growth

Brand promotion and customer experience

Social media
Online history
Customer service data

Sentiment analysis

Opinion mining/social media analysis
Deeper customer insights
Customer loyalty programs
Shopping cart analysis

Improved customer engagement with machine learning

Oil, gas, and energy use cases

Digital oil field/ oil production

Upstream optimization, maximize well life

Field data
Asset data
Demographics
Production data

- Production optimization
- Integrate exploration and seismic data
- Minimize lease operating expenses
- Decline curve analysis

Faster innovation for revenue growth

Industrial IoT

Grid operations, asset inventory optimization

Sensor stream data
UAVs images
Inventory data
Production data

- Pipeline monitoring
- Preventive maintenance
- Smart grids and microgrids
- Grid operations, field service
- Asset performance as a service

Improved outcomes and increased revenue

Supply-chain optimization

Supply-chain optimization

Transaction data
Demographics
Purchasing history
Trends

- Trade monitoring, optimization
- Retail mobile applications
- Vendor management - construction, transportation, truck and delivery optimization

Optimizing supply-chain with machine learning

Safety and security

Risk optimization

Sensor stream data
Transport
Retail data
Grid production data
Refinery tuning parameters

- Real-time anomaly detection
- Predictive analytics
- Industrial safety
- Environment health and safety

Predictive analytics transforms safety and security

Sales and marketing analytics

Recommendations engine

Clickstream data
Products
Services
Market data
Competitive data
Demographics

- Fast marketing and multi-channel engagement
- Develop new products and monitor acceptance of rates
- Predictive energy trading
- Deep customer insights

Improved customer engagement with machine learning

Security use cases

Security controls to leverage all data

Firewall/network logs
Apps
Data access layers

Intrusion detection and predictive analytics

Prevention of DDoS attacks
Threat classifications
Data loss/anomaly detection in streaming
Cybermetrics and changing use patterns

Prevent complex threats with machine learning

Actionable threat intelligence

Firewall/network logs
Network flows
Authentications

Security intelligence

Real-time data correlation
Anomaly detection
Security context, enrichment
Offence scoring, prioritization
Security orchestration

Faster innovation for threat prevention

Risk and fraud analysis

Firewall/network logs
Web/app logs
Social media content

Fraud detection and prevention

e-Tailing
Inventory monitoring
Social media monitoring
Phishing scams
Piracy protection

Risk management with machine learning

Compliance management

Firewall/network logs
Web
Applications
Devices
OS

Security compliance reporting

Ad-hoc/historic incident reports
SOC/NOC dashboards
Deep OS auditing
Data loss detection in IoT
User behavior analytics

Transform security with improved visibility

Identity and access management for analytics

Files
Tables
Clusters
Reports
Dashboards
Notebooks

Fine-grained data analytics security

Role-based access controls
Auditing and governance
File integrity monitoring
Row level and column level access permissions

Limit malicious insiders to transform growth

Our differentiated value proposition

Accelerate time to value with agile tools and services

Pretrained AI services

Powerful tools

Comprehensive platform

Innovate with AI everywhere – in the cloud, at edge and on-premises

Cloud

Edge

On-premises

Use any language, any development tool and any framework

python PYTORCH ONNX

R TensorFlow Spark

Benefit from industry-leading security, privacy, compliance, transparency, and AI ethics standards

>90% of Fortune 500 companies use Microsoft Cloud

LOOKING ACROSS THE OFFERINGS

Azure HDInsight

What It Is

- Hortonworks distribution as a first party service on Azure
- Big Data engines support – Hadoop Projects, Hive on Tez, Hive LLAP, Spark, HBase, Storm, Kafka, R Server
- Best-in-class developer tooling and Monitoring capabilities
- **Enterprise Features**
  - VNET support (join existing VNETs)
  - Ranger support (Kerberos based Security)
  - Log Analytics via OMS
  - Orchestration via Azure Data Factory
  - Available in most Azure Regions (27) including Gov Cloud and Federal Clouds

Guidance

- Customer needs Hadoop technologies other than, or in addition to Spark
- Customer prefers Hortonworks Spark distribution to stay closer to OSS codebase and/or 'Lift and Shift' from on-premises deployments
- Customer has specific project requirements that are only available on HDInsight

Azure Databricks

What It Is

- Databricks' Spark service as a first party service on Azure
- Single engine for Batch, Streaming, ML and Graph
- Best-in-class notebooks experience for optimal productivity and collaboration
- **Enterprise Features**
- Native Integration with Azure for Security via AAD (OAuth)
- Optimized engine for better performance and scalability
- RBAC for Notebooks and APIs
- Auto-scaling and cluster termination capabilities
- Native integration with SQL DW and other Azure services
- Serverless pools for easier management of resources

Guidance

- Customer needs the best option for Spark on Azure
- Customer teams are comfortable with notebooks and Spark
- Customers need Auto-scaling and
- Customer needs to build integrated and performant data pipelines
- Customer is comfortable with limited regional availability 5 in preview, 8 by GA)

Azure ML

What It Is

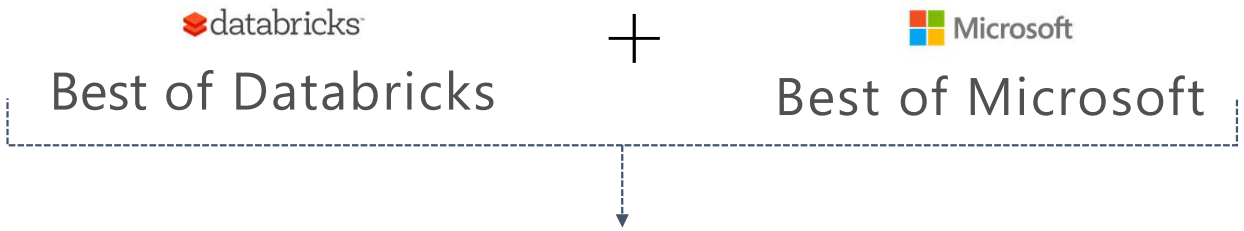
- Azure first party service for Machine Learning
- Leverage existing ML libraries or extend with Python and R
- Targets emerging data scientists with drag & drop offering
- Targets professional data scientists with
  - Experimentation service
  - Model management service
  - Works with customers IDE of choice

Guidance

- Azure Machine Learning Studio is a GUI based ML tool for emerging Data Scientists to experiment and operationalize with least friction
- Azure Machine Learning Workbench is not a compute engine & uses external engines for Compute, including SQL Server and Spark
- AML deploys models to HDI Spark currently
- AML should be able to deploy Azure Databricks in the near future

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, ADLS, Azure Storage, Azure Data Factory, Azure AD, Event Hub, IoT Hub, HDInsight Kafka, SQL DB)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

Azure Databricks

Fast, easy, and collaborative Apache Spark™-based analytics platform

Increase productivity

Build on a secure, trusted cloud

Scale without limits

Built with your needs in mind

- Role-based access controls
- Effortless autoscaling
- Live collaboration
- Enterprise-grade SLAs
- Best-in-class notebooks
- Simple job scheduling

Azure Databricks

A fast, easy, and collaborative Azure service for Apache Spark-based analytics

Productive

Set up **Spark clusters in minutes** with one click from the Azure portal

Use the **language of your choice** with Python, R, Scala, SQL

Improve collaboration through a unified workspace and **collaborate live with notebooks**

Secure

Simplify security and identity control with built-in integration with **Azure Active Directory** and role-based access

Regulate access with fine-grained user permissions to Azure Databricks' notebooks, clusters, jobs and data

Build with confidence on the **trusted cloud** backed by unmatched support, and compliance

Scalable

Autoscale effortlessly

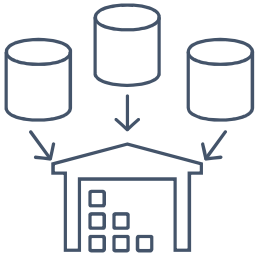
Integrate seamlessly with other **Azure data services**

Benefit from **enterprise-grade SLAs**

Operate at **massive scale**, without limits, globally

Seamlessly integrated with the Azure portfolio

Usage Objectives

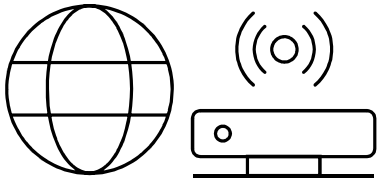


"Want to extend to untapped sources"

Modern Data Warehouse

"Want to use ML and AI to get deeper insights from our data"

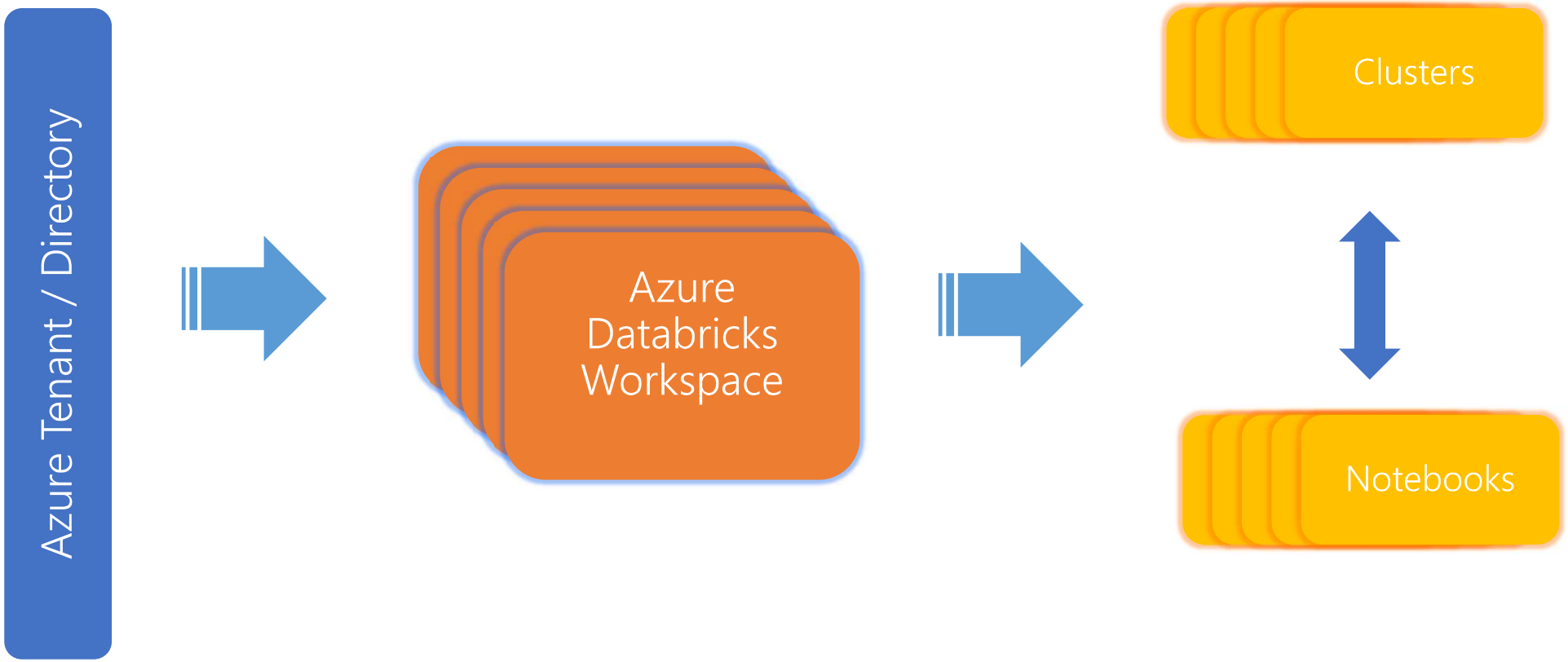
Advanced Analatics



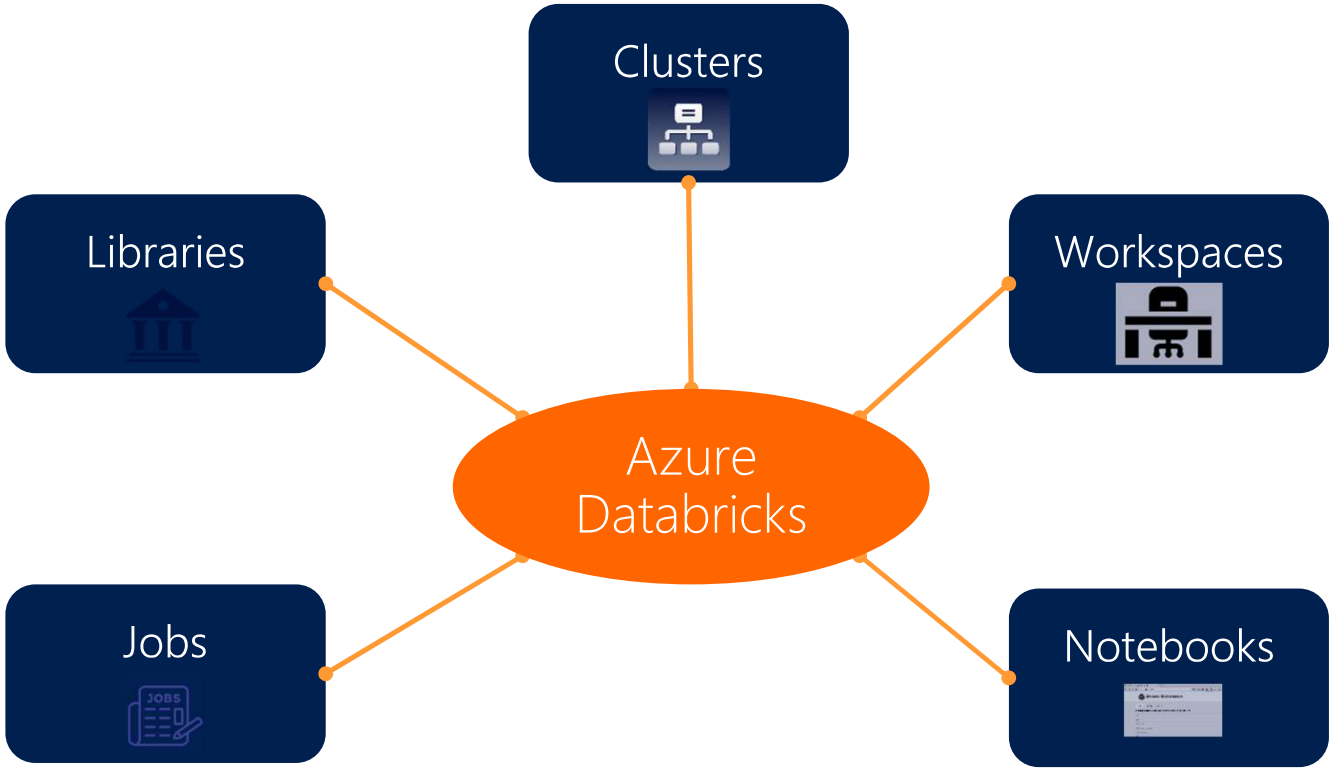
"Want to get insights from our devices in real-time"

Real-time Analytics

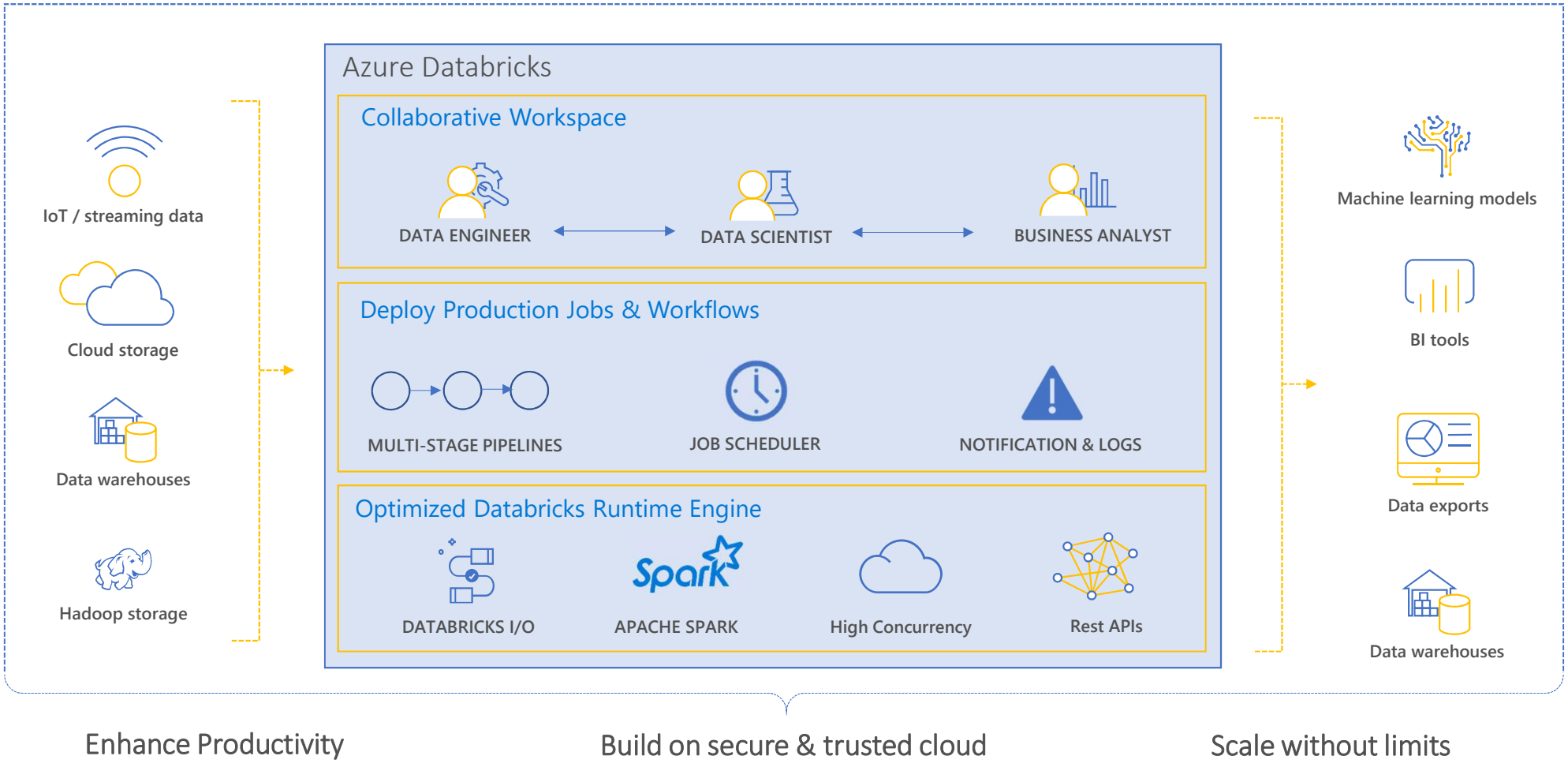
Workspaces, Clusters & Notebooks



AZURE DATABRICKS CORE ARTIFACTS



AZURE DATABRICKS

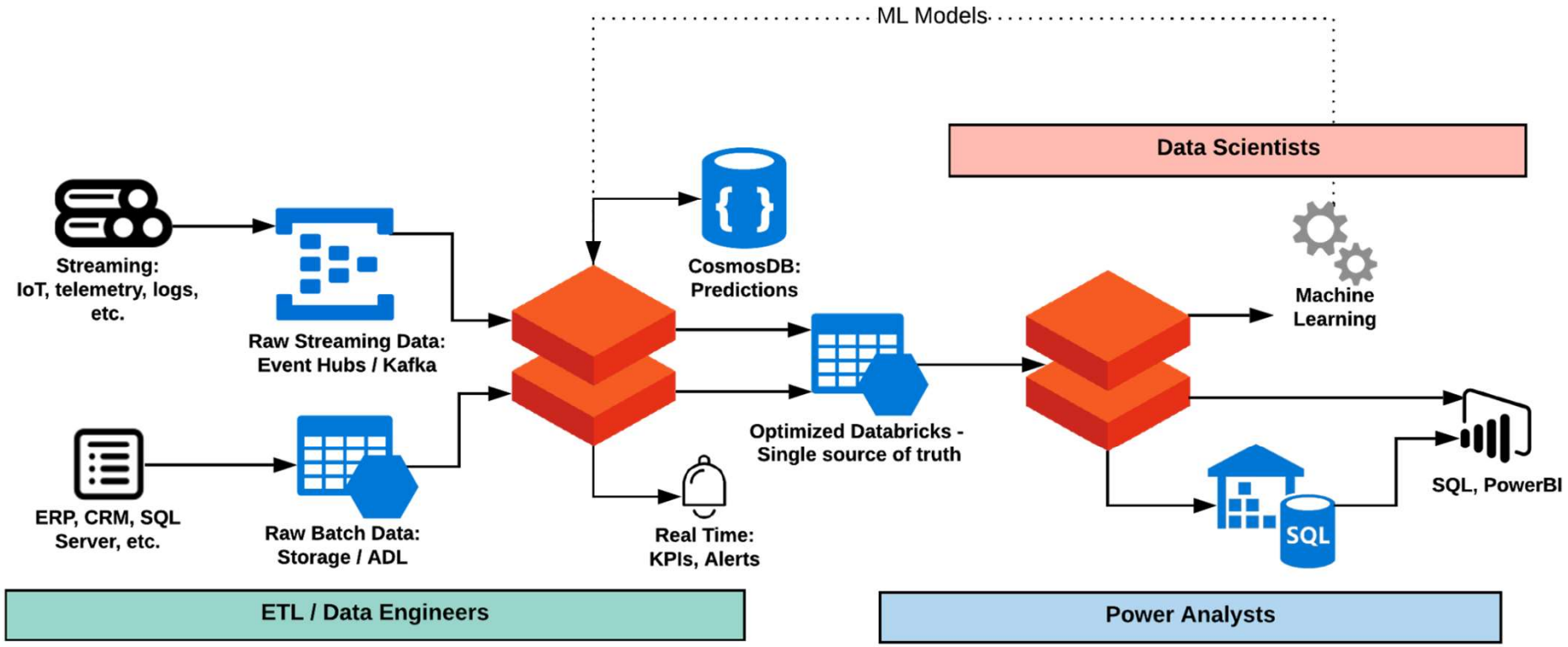


Enhance Productivity

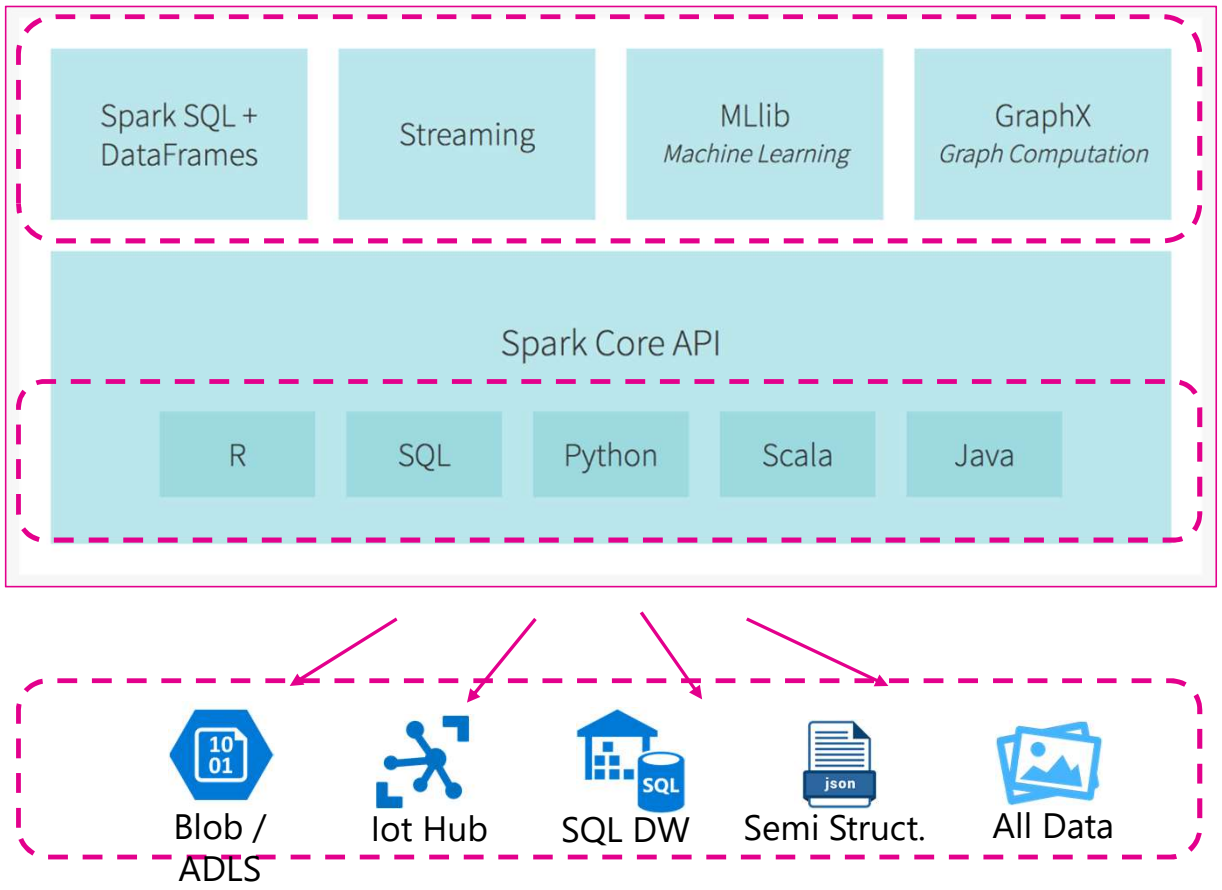
Build on secure & trusted cloud

Scale without limits

Unified Architecture



Apache Spark - Distributed Framework



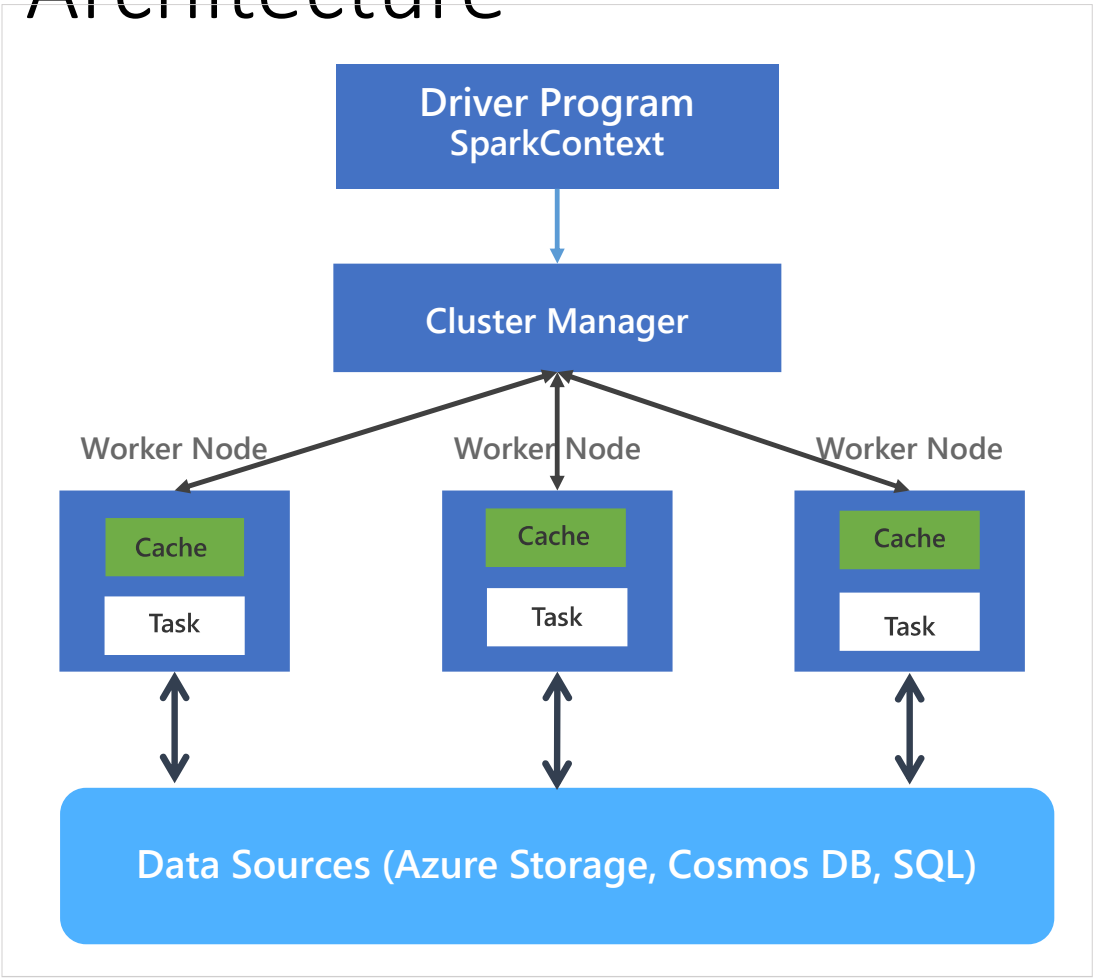
Multiple Use Cases

Multiple Languages

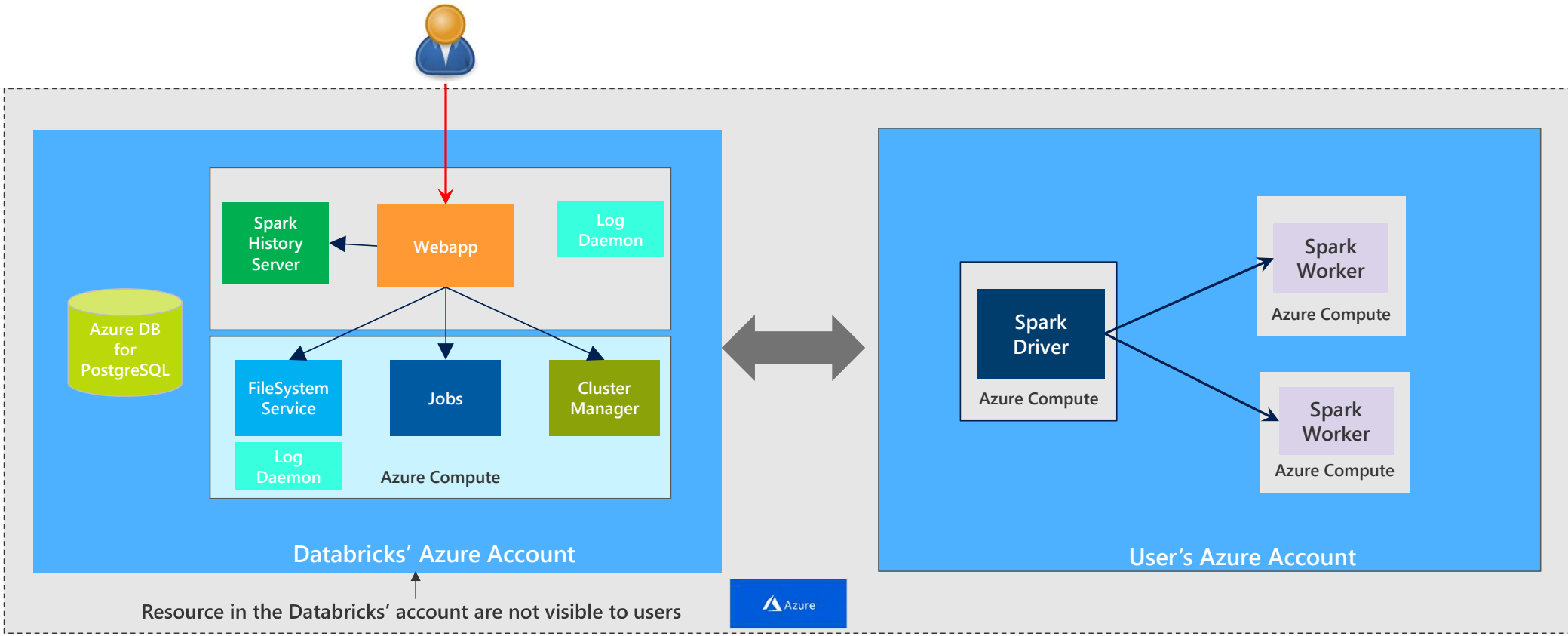
Multiple Data Sources

General Spark Cluster Architecture

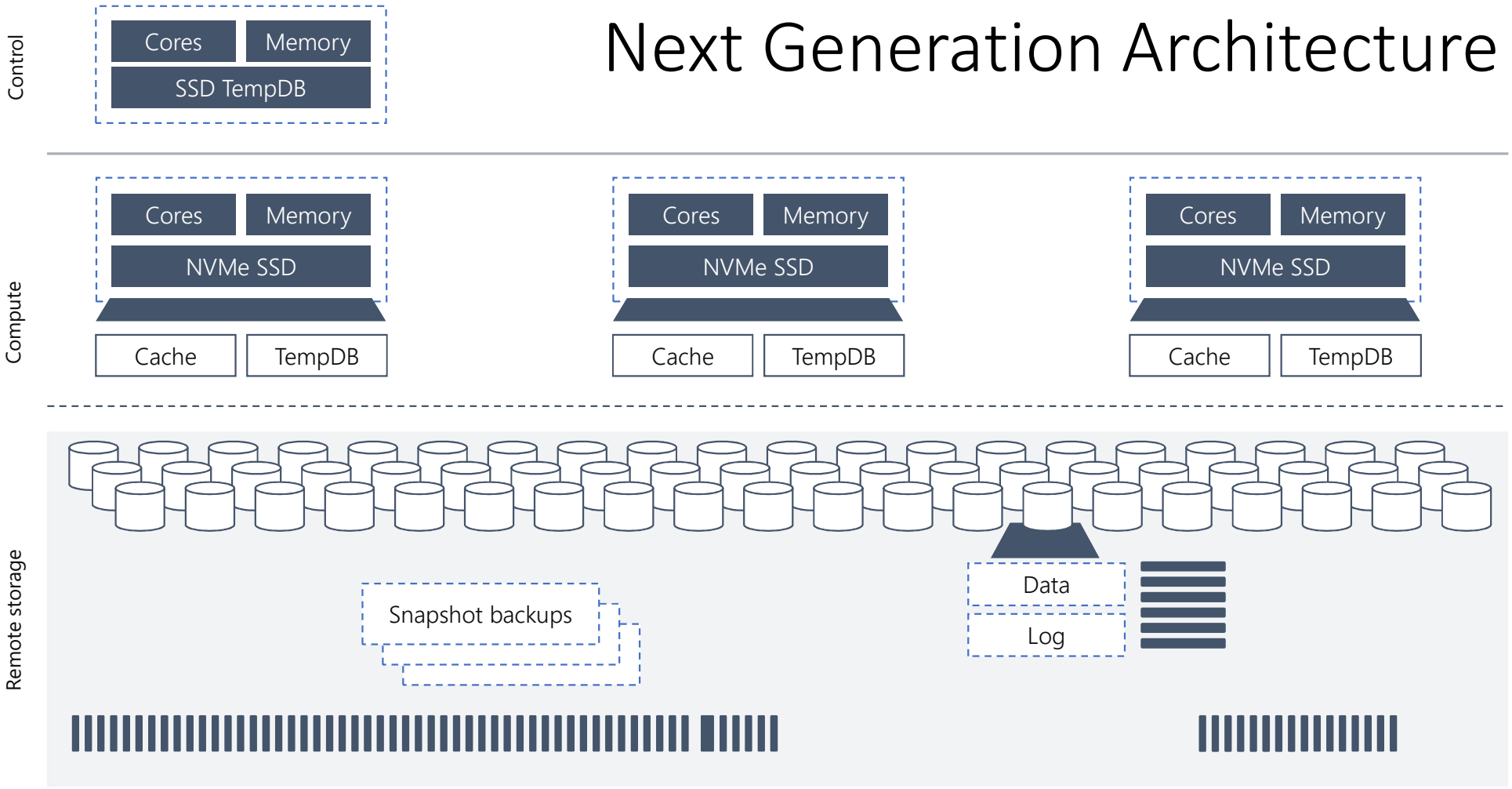
- Spark is designed to run on a Cluster
- A cluster is a set of VMs
- Spark can horizontally scale, bigger workload = Add more VMs
- Azure Databricks can automatically scale up and down
- Data can read from Azure Storage or Azure Datalake Storage



AZURE DATABRICKS CLUSTER ARCHITECTURE



Next Generation Architecture



RDDs, Data Frames and Data Sets

RDDs (2011)

- Distributed collection of JVM Objects
- Functional Operators (map, filters)

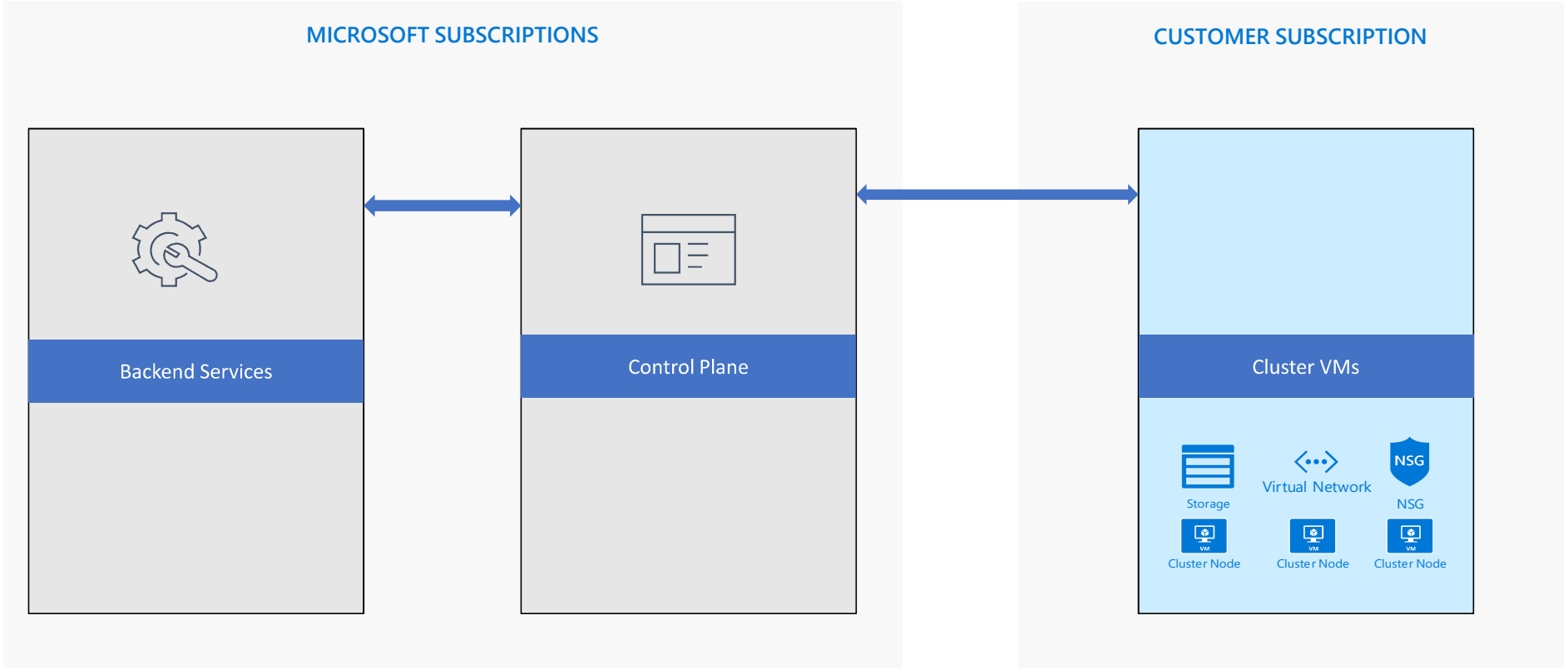
Data Frames (2013)

- Distributed collection of Row Objects
- Expression based operations and UDFs
- Logical plans and optimizer
- Fast/efficient internal representation

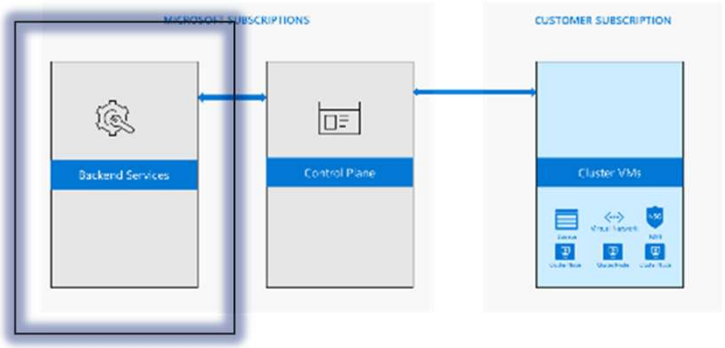
Data Sets (2015)

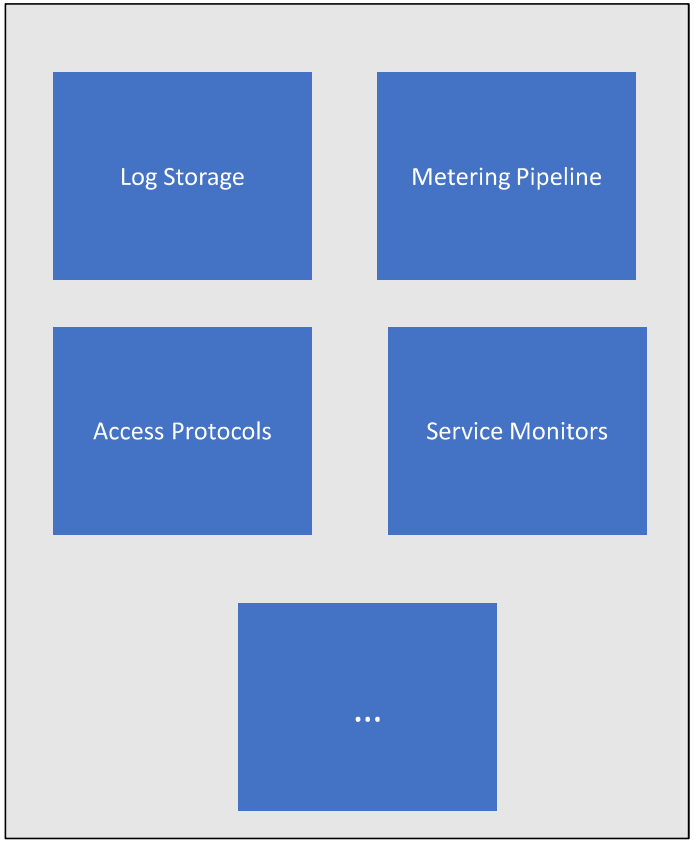
- Internally Rows, externally JVM objects
- Best of both worlds – “type safe + fast”

Getting the Head around

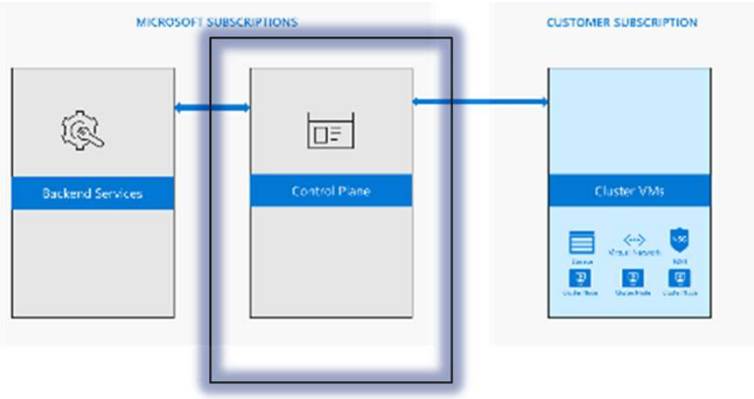


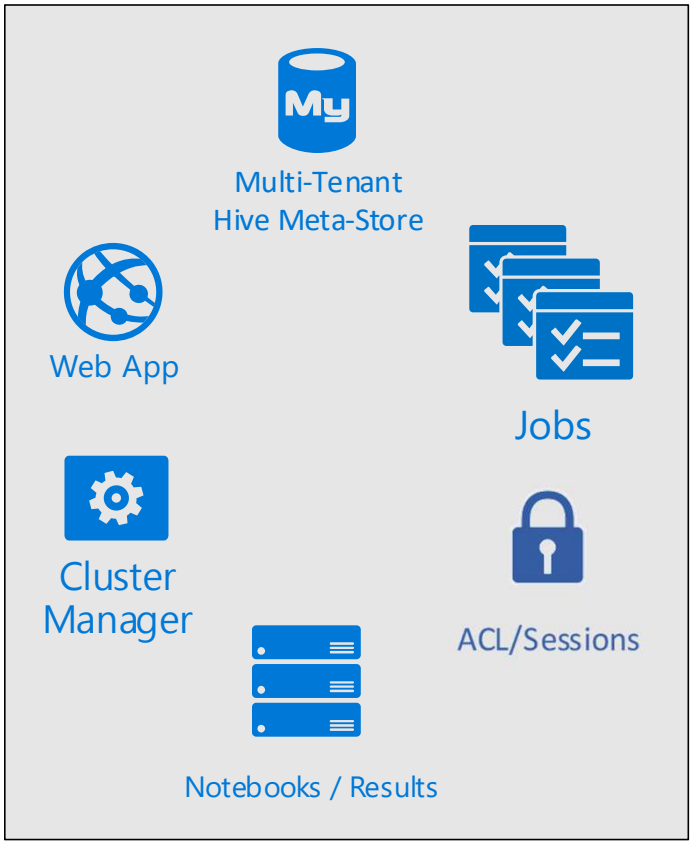
Backend Services



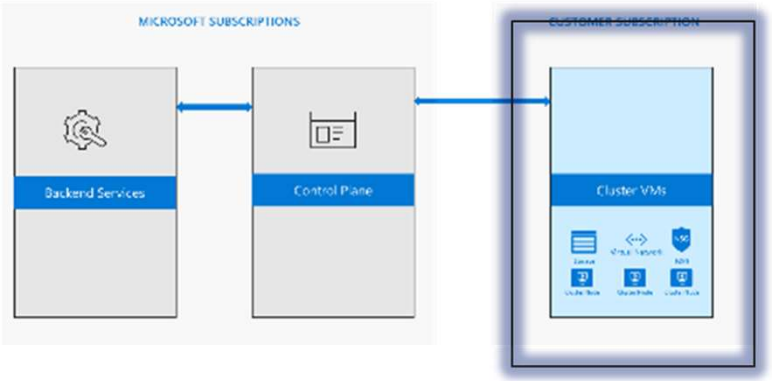


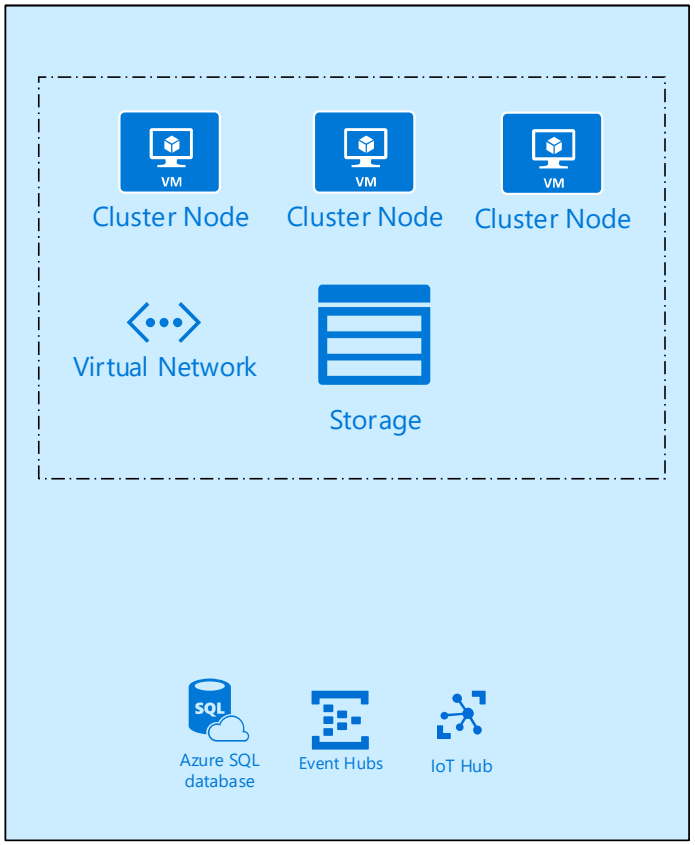
Control Plane



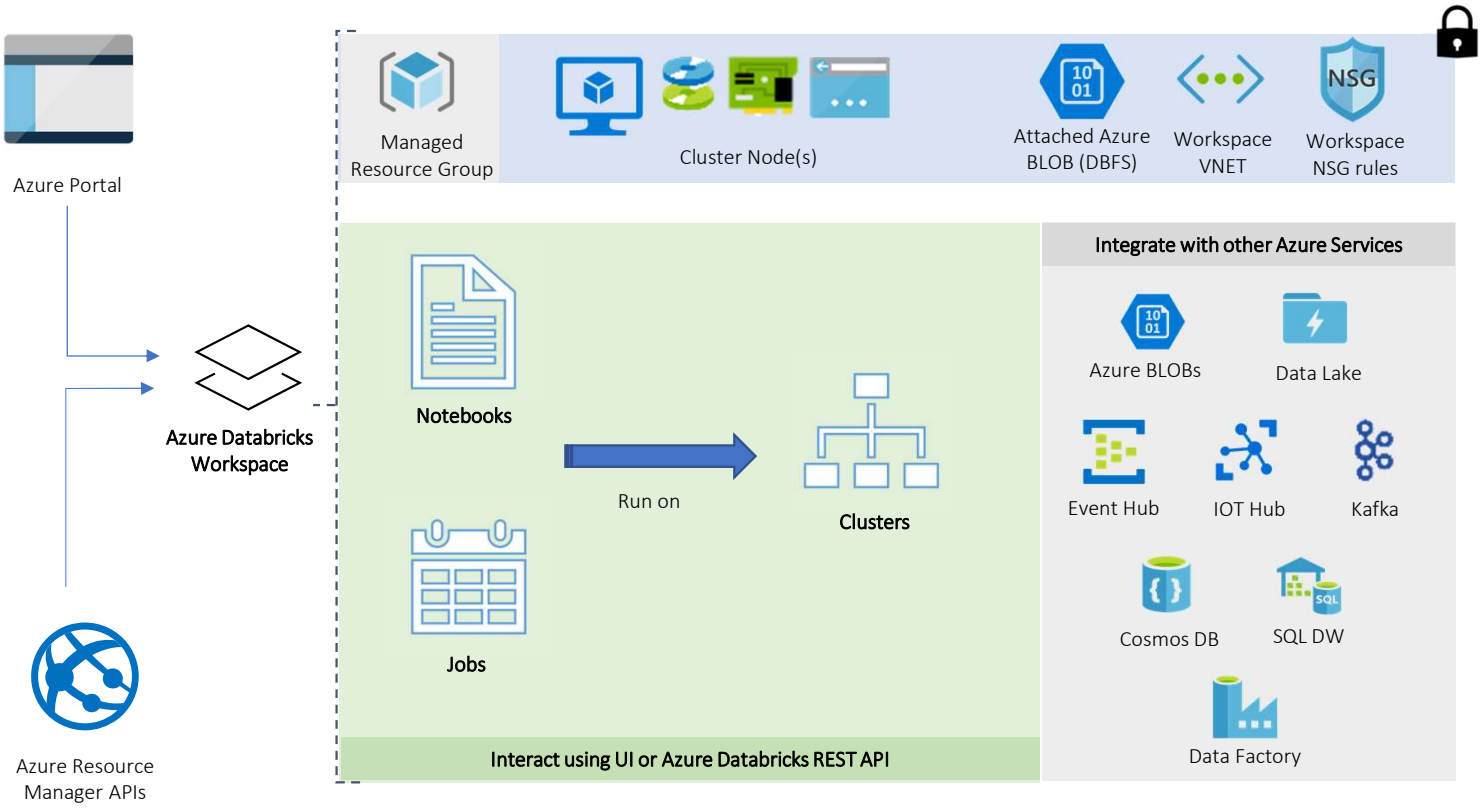


Customer Subscription





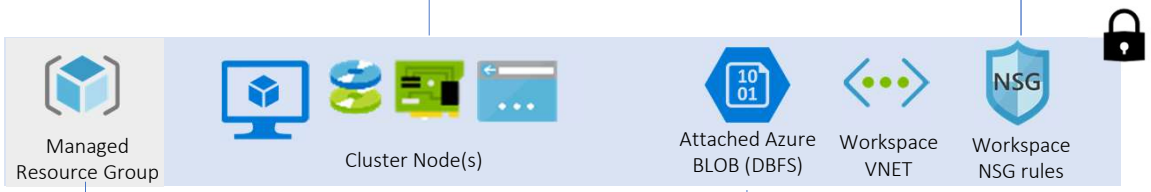
Customer view



Managed resource group

- For each driver or worker node in each cluster, a VM, SSD, Network Interface and IP address resource is created
- Each resource is tagged with
  - "Vendor : Databricks"
  - "ClusterName :" + {Cluster Name}
  - Custom Tags can be added at Cluster creation

- All resources in the RG are in a pre-configured VNET
- VNET allows communication between nodes and with the Control Plane
- NSG is locked contains the Control Plane IP addresses
- VNET & NSG are locked but customers can peer with other VNETs



- A locked RG is created in your subscription when you create the Azure Databricks Workspace
- RG Name = "databricks-rg"+ {workspace} + {random chars}

- Each workspace comes with an attached blob storage, contains workspace related configuration
- This can be access from Notebooks and Jobs as DBFS
- You can mount your own Azure Storage or ADLS to this DBFS

Regional distribution of the control plane

- Available today in 24 Regions / 6 Geographies
- Every geography has a Control Plane & Backend Services
- All dependent services run in that geography
- Data never leaves the geography
- Geographies : https://azure.microsoft.com/en-us/global-infrastructure/geographies/

Who can access to Control Plane ?

Common Scenarios – When deploying a new feature, when making a fix, when adding a new region, automated jobs to read telemetry.

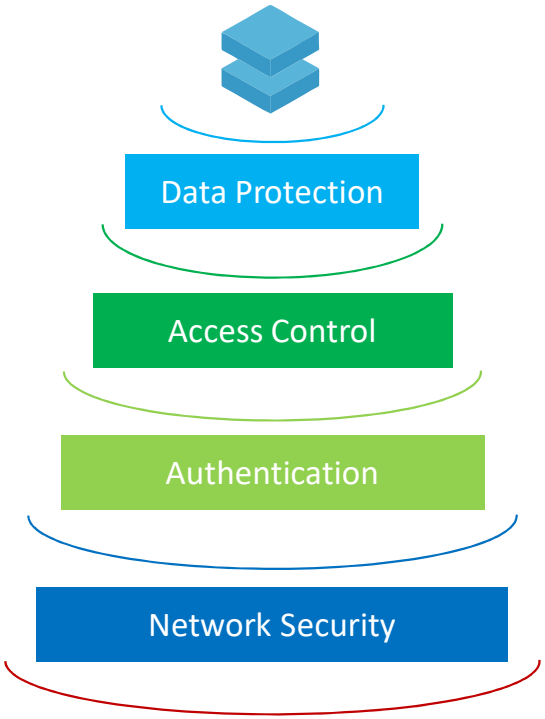
Policies and Procedures

- Follows all Azure Guidelines
- Access only allowed via secure hardware & JIT
- Logged & Audited

Architecture Summary

1. Azure Databricks is a Multi-Tenant Service
2. Control Plane is shared across all customers
3. Controls Planes are scoped to Geographies
4. Clusters are in the Customers Subscription and not shared
5. Access to Control Plane is audited, policies and procedures are enforced from the physical level to the logical level of security

Enterprise Grade Security that is Easy-to Use



- Encryption-at-rest – Service Managed Keys, User Managed Keys
- File/Folder Level ACLs for AAD Users, Groups, Service Principals
- Encryption-in-flight (Transport Layer Security TLS)

- Role Base Access Control for Clusters, Workspaces, Notebooks
- Access Control for Tables

Azure Active Directory Authentication (w/ MFA)
Azure Active Directory Conditional Access

VNET – Managed & Injection*

Defense in Depth

* VNET Injection support in Public Preview

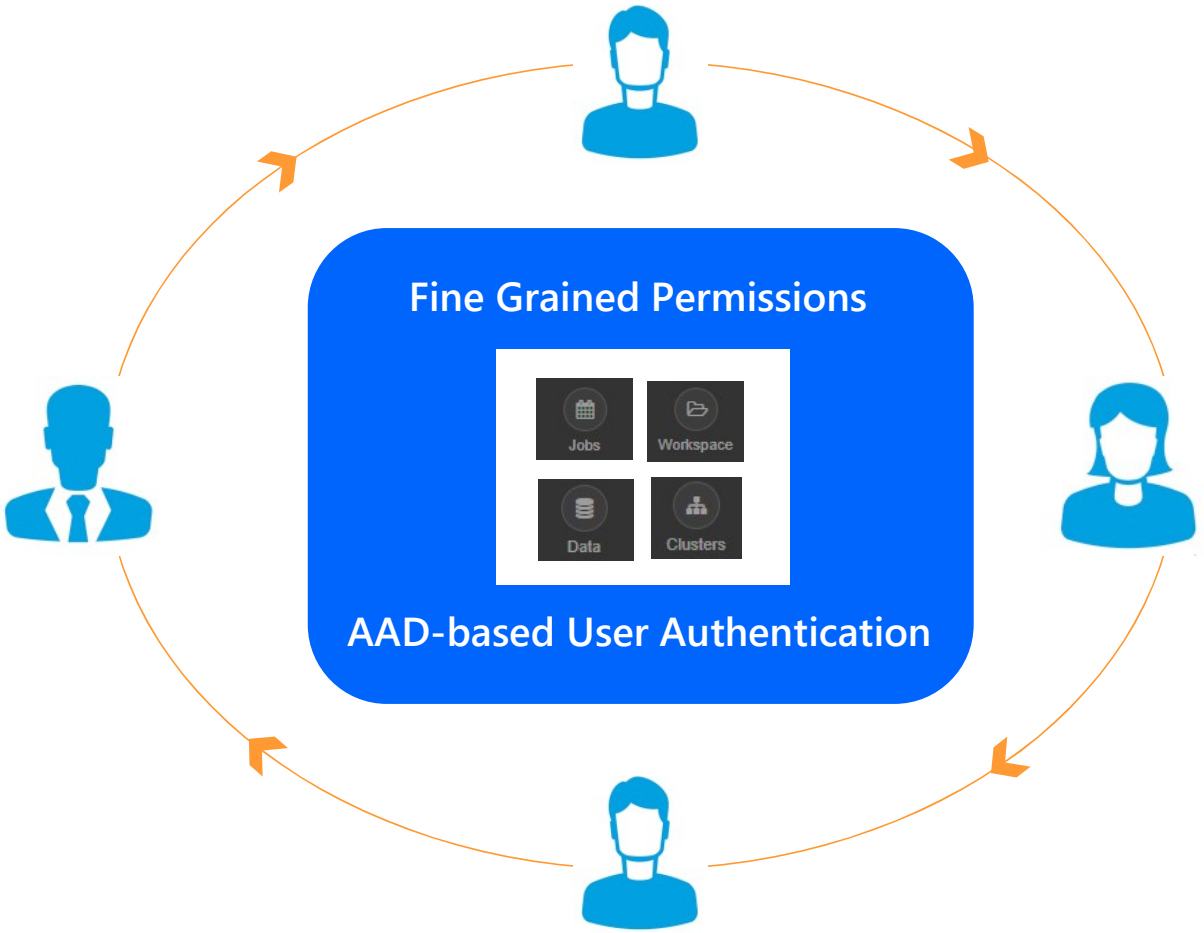
SECURE COLLABORATION

Azure Databricks enables *secure* collaboration between colleagues

- With Azure Databricks colleagues can *securely share* key artifacts such as Clusters, Notebooks, Jobs and Workspaces
- Secure collaboration is enabled through a combination of:

Fine grained permissions: Defines who can do what on which artifacts (access control)

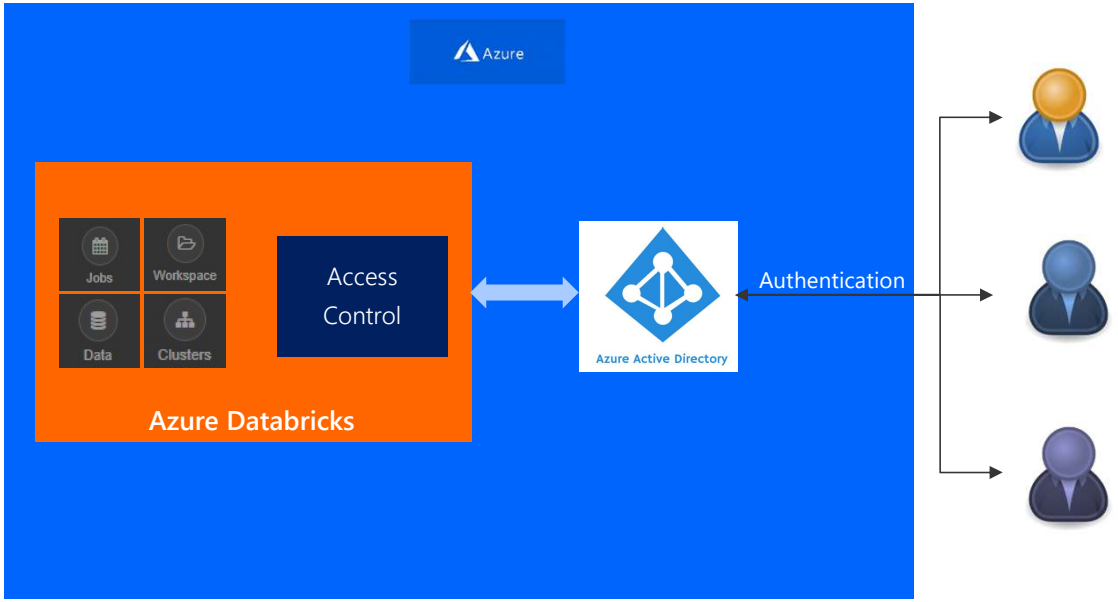
AAD-based authentication: Ensures that users are actually who they claim to be



AZURE DATABRICKS INTEGRATION WITH AAD

Azure Databricks is integrated with AAD—so Azure Databricks users are just regular AAD users

- There is no need to define users—and their access control—separately in Databricks.
- AAD users can be used directly in Azure Databricks for all user-based access control (Clusters, Jobs, Notebooks etc.).
- Databricks has delegated user authentication to AAD enabling single-sign on (SSO) and unified authentication.
- *Notebooks, and their outputs, are stored in the Databricks account*. However, AAD-based access-control ensures that only authorized users can access them.



DATABRICKS ACCESS CONTROL

Access control can be defined at the user level via the Admin Console

Access Control can be defined for Workspaces, Clusters, Jobs and REST APIs

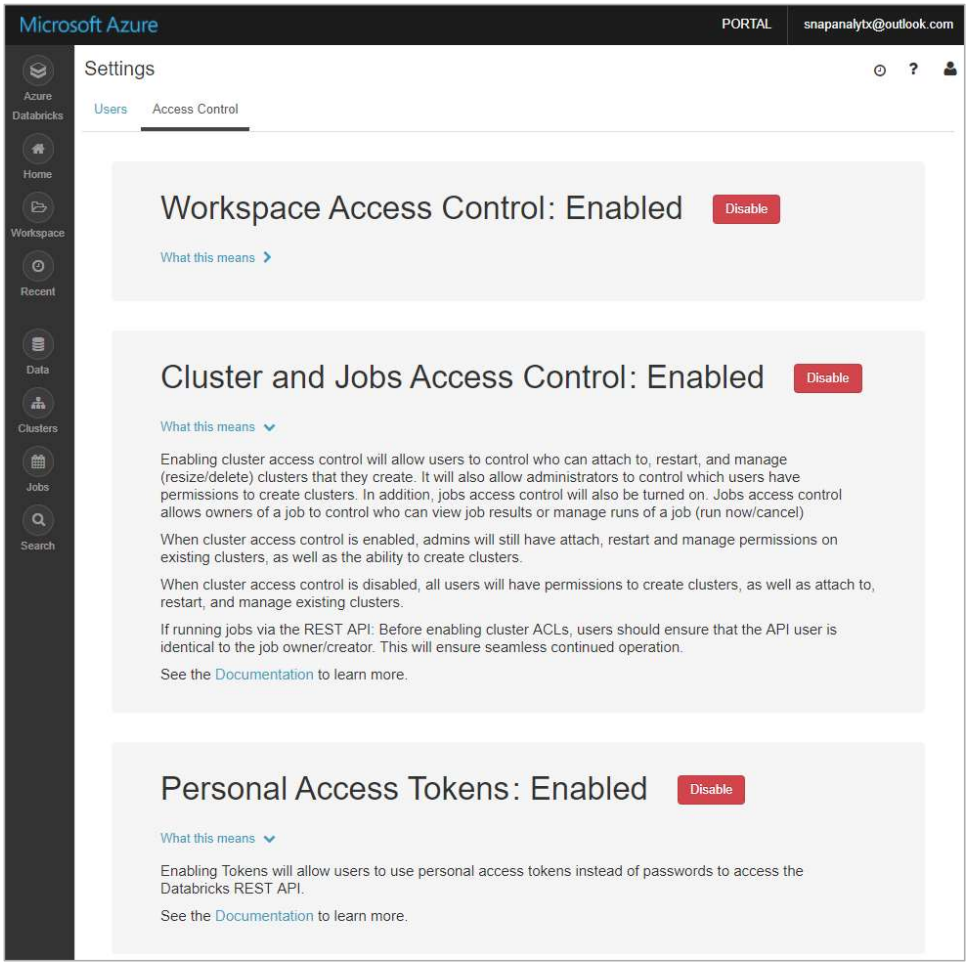
Databricks Access Control

Workspace Access Control	Defines who can who can view, edit, and run notebooks in their workspace
Cluster Access Control	Allows users to who can attach to, restart, and manage (resize/delete) clusters. Allows Admins to specify which users have permissions to create clusters
Jobs Access Control	Allows owners of a job to control who can view job results or manage runs of a job (run now/cancel)
REST API Tokens	Allows users to use personal access tokens instead of passwords to access the Databricks REST API

ENABLE/DISABLE ACCESS CONTROL

Access Control can be selectively enabled or disabled for:

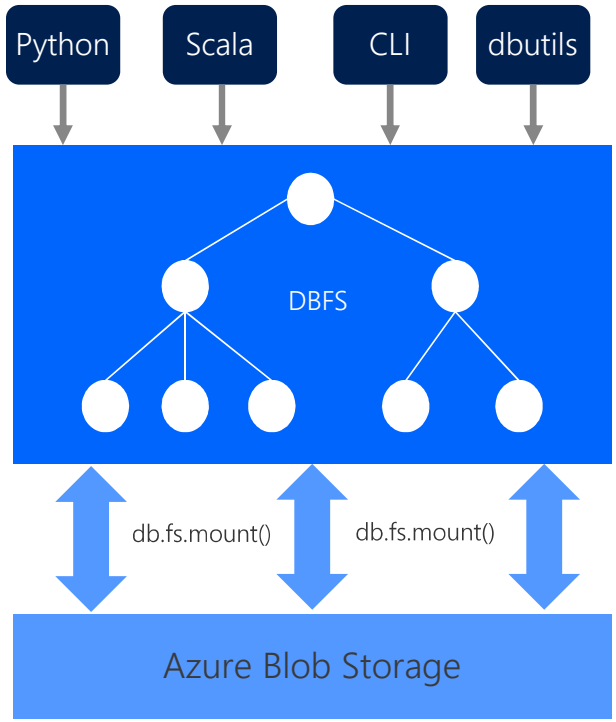
- Workspaces,
- Clusters,
- Jobs
- REST APIs



DATABRICKS FILE SYSTEM (DBFS)

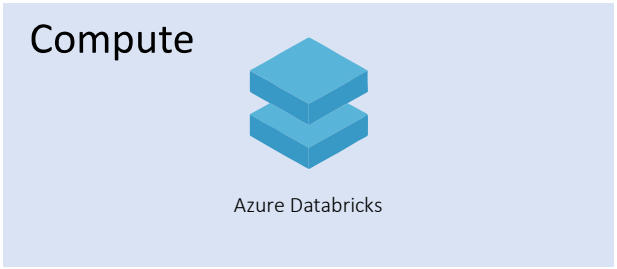
Is a distributed File System (DBFS) that is a layer over Azure Blob Storage

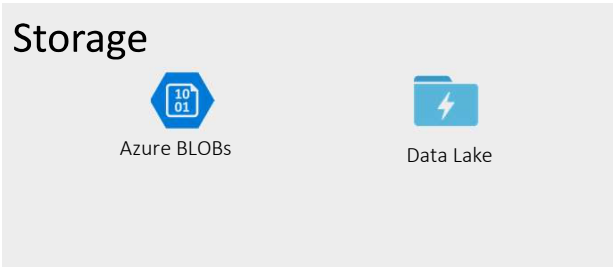
- Azure Storage buckets can be mounted in DBFS so that users can directly access them without specifying the storage keys
- DBFS mounts are created using *dbutils.fs.mount()*
- Azure Storage data can be cached locally on the SSD of the worker nodes
- Available in both Python and Scala and accessible via a DBFS CLI
- Data persist in Azure Blob Storage – is not lost even after cluster termination
- Comes pre-installed on Spark clusters in Databricks



Data Protection | Encryption - Data at rest

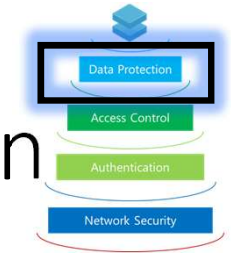
- Azure Databricks has separation of compute and storage



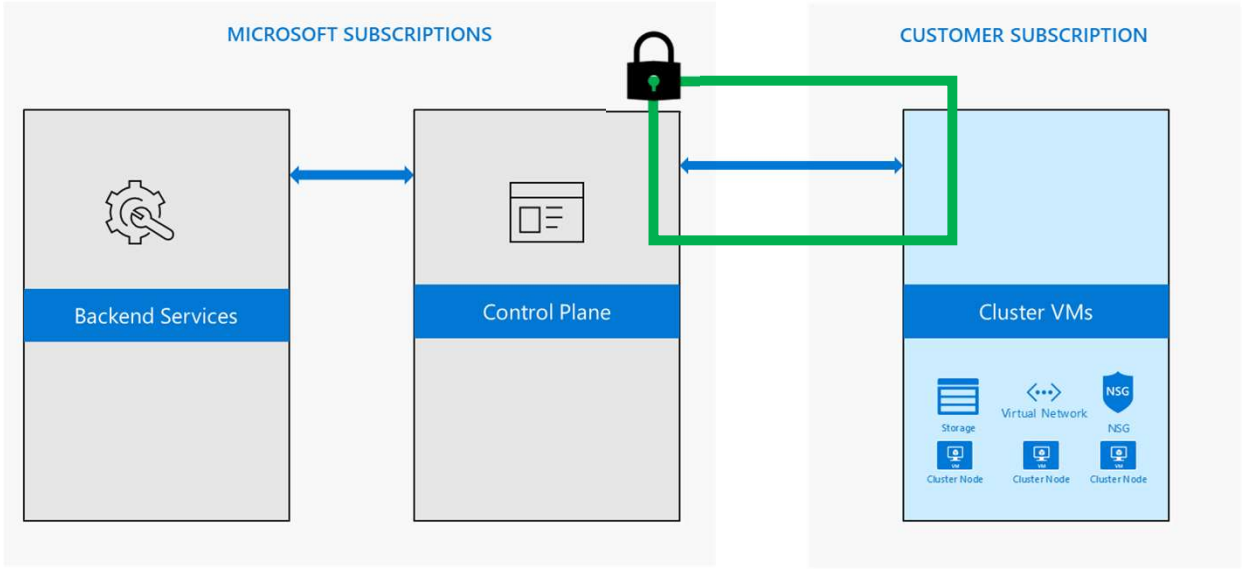


- Storage Services such as Azure Blob Store, Azure Data Lake Storage Provide
  - Encryption of Data
  - Customer Managed Keys
  - File/Folder Level ACLs (Azure Data Lake Storage)
- All Azure Databricks provided data stores are encrypted at rest

Data Protection | Encryption - Data in motion



- All the traffic from the Control Plane to the Clusters in the customer subscription is always encrypted with TLS.

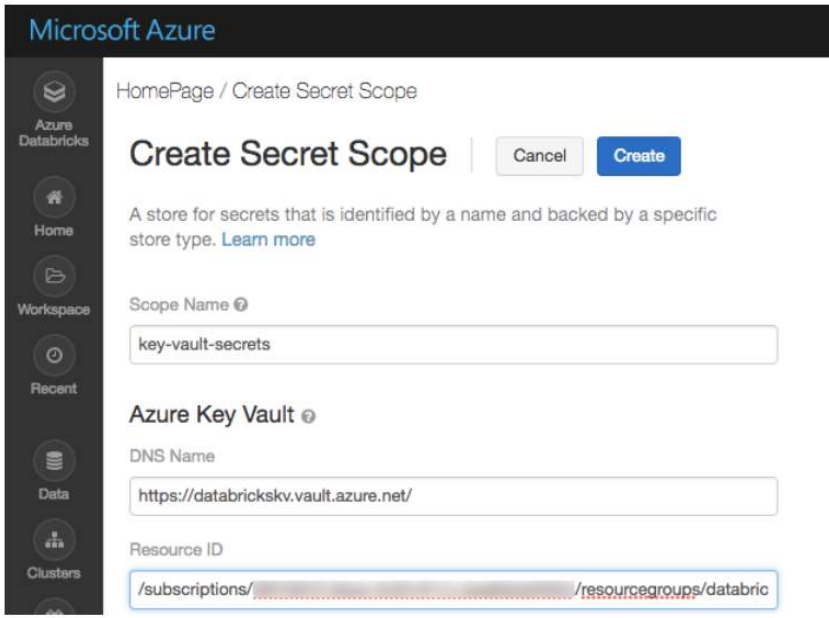


Secrets in Notebooks – Understanding the need

- Customer often connect to other Azure resources such as Azure BLOB Storage, Azure Data Lake, SQL DW from Azure Databricks
- A “Connection String” is required to connect to these services. This string may contain secrets.
- Customers don’t want to store Secrets in the clear

Securing secrets in Notebooks

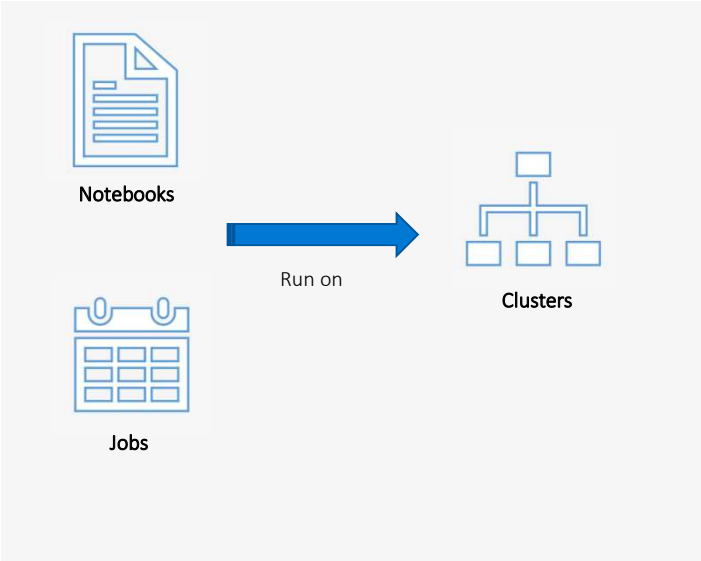
- Using our Secrets APIs, Secrets can be securely stored including in a Key Vault
- Authorized users can consume the secrets to access services but cannot see them.



https://docs.azuredatabricks.net/user-guide/secrets/secret-scopes.html

Access Control

- Many users in the customers organization can use the Service
- Different users have different roles – Admin, Data Scientist, Engineers
- Access Controls lets you limit what users can do

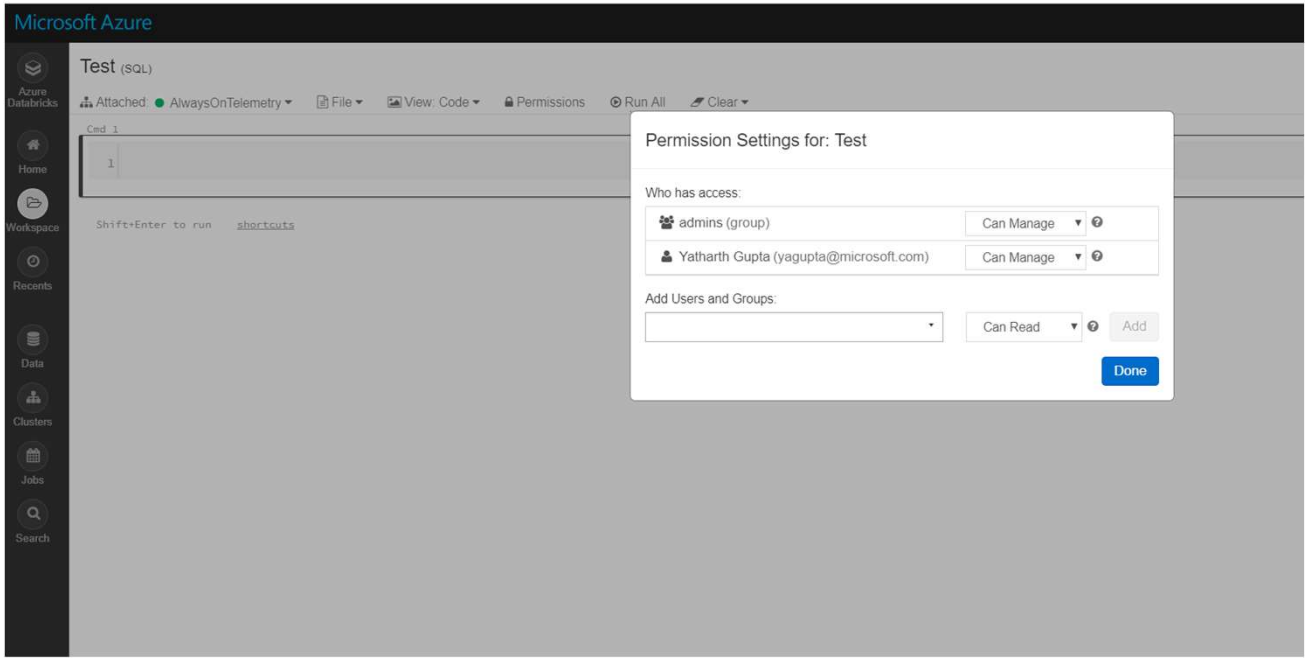


Access Control | Folders

Ability	No Permissions	Read	Run	Edit	Manage
View items		X	X	X	X
Create, clone, import, export items		X	X	X	X
Run commands on notebooks			X	X	X
Attach/detach notebooks			X	X	X
Delete items				X	X
Move/rename items				X	X
Change permissions					X

Access Control | Notebooks

Ability	No Permissions	Read	Run	Edit	Manage
View cells		X	X	X	X
Comment		X	X	X	X
Run commands			X	X	X
Attach/detach notebooks			X	X	X
Edit cells				X	X
Change permissions					X


Microsoft Azure
Azure Databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Test (SQL)
Attached: AlwaysOnTelemetry
File
View: Code
Permissions
Run All
Clear
Cmd 1
1
Shift+Enter to run
shortcuts
Permission Settings for: Test
Who has access:
admins (group)
Can Manage
Yatharth Gupta (yagupta@microsoft.com)
Can Manage
Add Users and Groups:
Can Read
Add
Done


Data Protection
Access Control
Authentication
Network Security

Access Control | Jobs

Ability	No Permissions	Can View	Can Manage Run	Is Owner	Can Manage (admin)
View job details and settings	x	x	x	x	x
View results, Spark UI, logs of a job run		x	x	x	x
Run now			x	x	x
Cancel run			x	x	x
Edit job settings				x	x
Modify permissions				x	x

Access Control | Clusters

Ability	No Permissions	Can Attach To	Can Restart	Can Manage
Attach notebook to cluster		x	x	x
View Spark UI		x	x	x
View cluster metrics		x	x	x
Terminate cluster			x	x
Start cluster			x	x
Restart cluster			x	x
Edit cluster				x
Attach library to cluster				x
Resize cluster				x
Modify permissions				x

Access Control | Tables

Objects

CATALOG | DATABASE | TABLE | VIEW | FUNCTION | ANONYMOUS FUNCTION | ANY FILE

Privileges

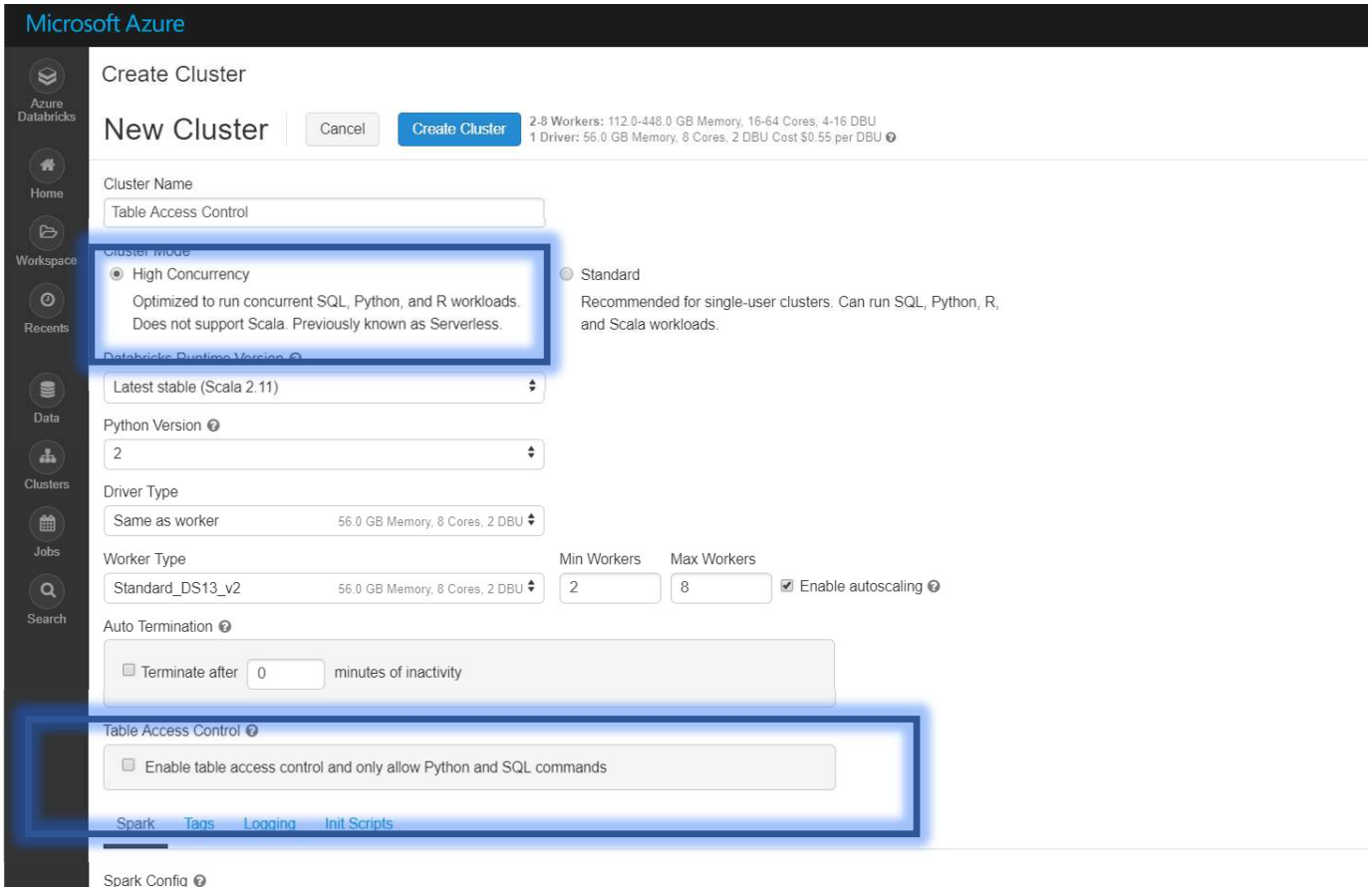
SELECT	- read access to an object
CREATE	- ability to create an object (eg. Table in a Database)
MODIFY	- ability to add/delete/modify data in an Object
READ_METADATA	- ability to read Meta data about an object
ALL_PRIVELEGES	- all of the above

Access Control | Tables


```
[GRANT | DENY]
    ON [OBJECT]
        TO [USER]
            [PRIVELEGE_TYPE]
```

- Access Control on Tables limits to SQL and Python only. This ensures that low level commands cannot be used to bypass these restrictions.
- High concurrence clusters provide isolation between users.

Access Control | Tables



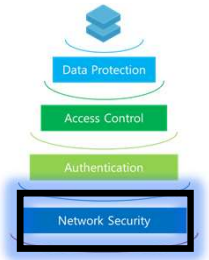
Authentication

- Azure Databricks support Azure Active Directory as an Authentication provider.
- This is pre-configured with zero setup needed. It includes the ability for the organization to enable multi-factor authentication.
- Support for conditional access has been added for additional policies

https://docs.microsoft.com/en-us/azure/active-directory/conditional-access/overview

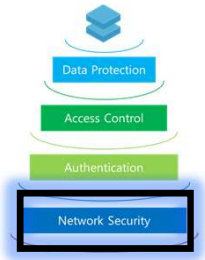
https://docs.azuredatabricks.net/administration-guide/cloud-configurations/azure/conditional-access.html#id1

Network Security | Managed VNETs



- Clusters (VMs) are always deployed in the customer's subscription. We deploy these in a VNET we create for the Customer.
- In this mode, the VNET and accompanying NSG rules are managed by us.
- We allow for Customer's to be able to Peer this with other VNETs

Network Security | Managed VNETs NSG Rules



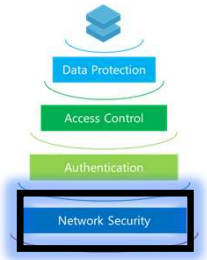
Inbound security rules

PRIORITY	NAME	PORT	PROTOCOL	SOURCE	DESTINATION	ACTION	
100	databricks-control-plane-ssh	22	Any	40.83.178.242	Any	Allow	...
110	databricks-control-plane-worker-pr...	5557	Any	40.83.178.242	Any	Allow	...
200	databricks-worker-to-worker	Any	Any	VirtualNetwork	Any	Allow	...
65000	AllowVnetInBound	Any	Any	VirtualNetwork	VirtualNetwork	Allow	...
65001	AllowAzureLoadBalancerInBound	Any	Any	AzureLoadBalancer	Any	Allow	...
65500	DenyAllInBound	Any	Any	Any	Any	Deny	...

Outbound security rules

PRIORITY	NAME	PORT	PROTOCOL	SOURCE	DESTINATION	ACTION	
65000	AllowVnetOutBound	Any	Any	VirtualNetwork	VirtualNetwork	Allow	...
65001	AllowInternetOutBound	Any	Any	Any	Internet	Allow	...
65500	DenyAllOutBound	Any	Any	Any	Any	Deny	...

Network Security | VNET Injection



You can signup for this today !

https://aka.ms/adbvnet

When to chose this option ?

- You want to use your own VNET
- You want to restrict access to your other services using Service Endpoint
- You want to use User Defined Routes
- You want to use an Application Layer Firewall

Security and Networking Summary

1. Data Protection – With Azure Databricks, storage is always remote and these storage services support encryption, customer managed keys and file/folder ACLs.

2. Access Control – Rich RBAC control available in Azure Databricks for Clusters, Notebooks, Tables etc.

3. Authentication – Always AAD

4. Network Security – Two options, designed for security and flexibility.

Compliance

- ISO 27001
- ISO 27018
- HIPAA
- SOC2, Type 2

Service Level Agreement

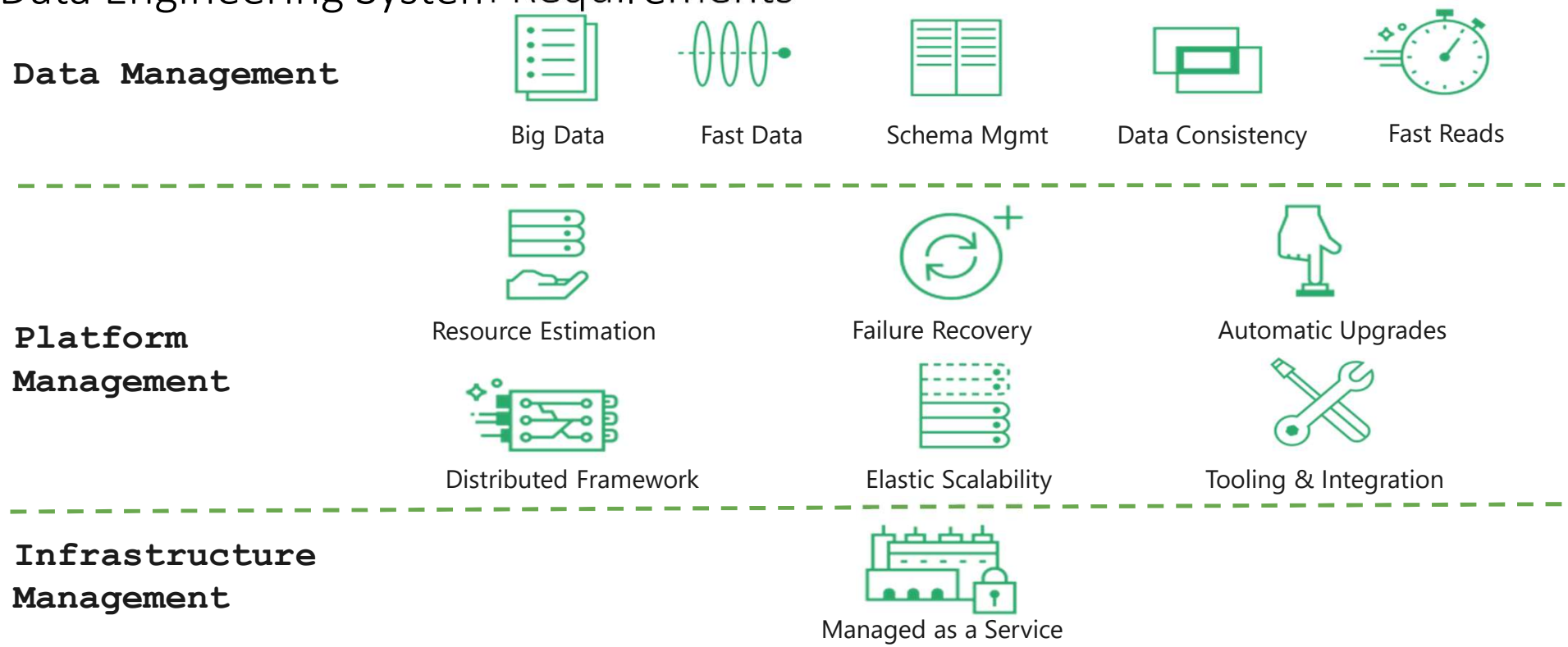
99.95% uptime SLA

MONTHLY UPTIME PERCENTAGE	SERVICE CREDIT
< 99.95%	10%
< 99%	25%

https://azure.microsoft.com/en-us/support/legal/sla/databricks/v1_0/

Databricks for Data Engineers

Data Engineering System Requirements



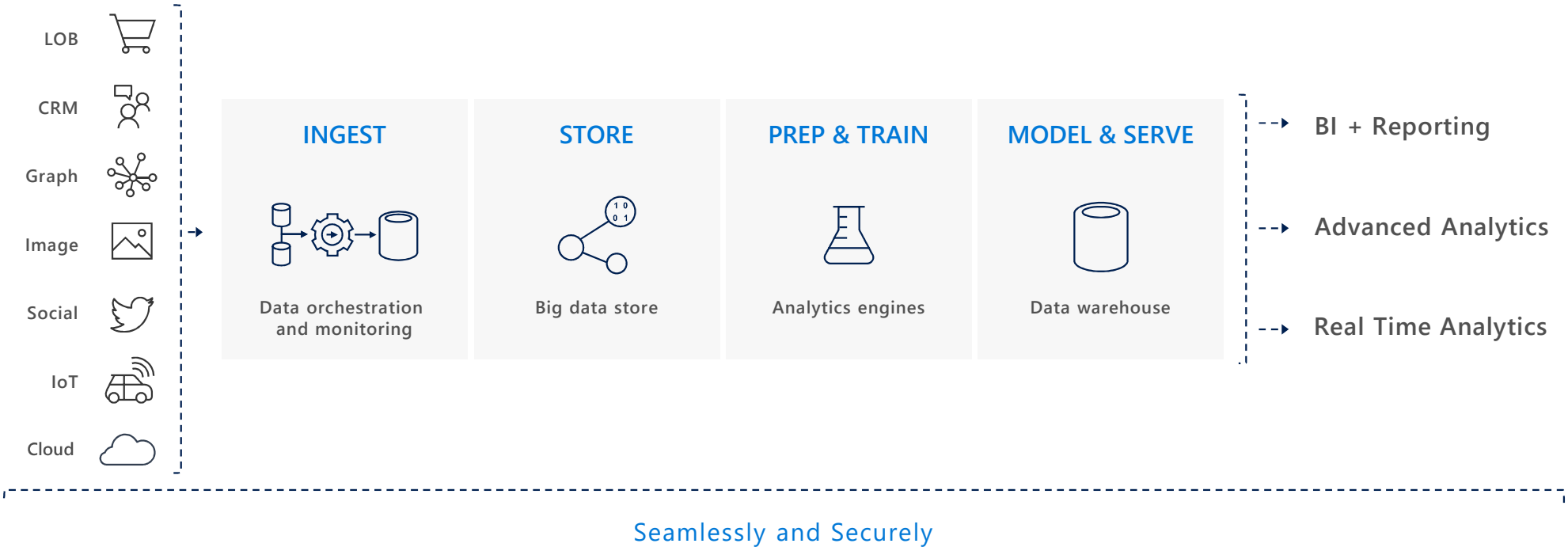
Why Data Engineering is Hard?

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

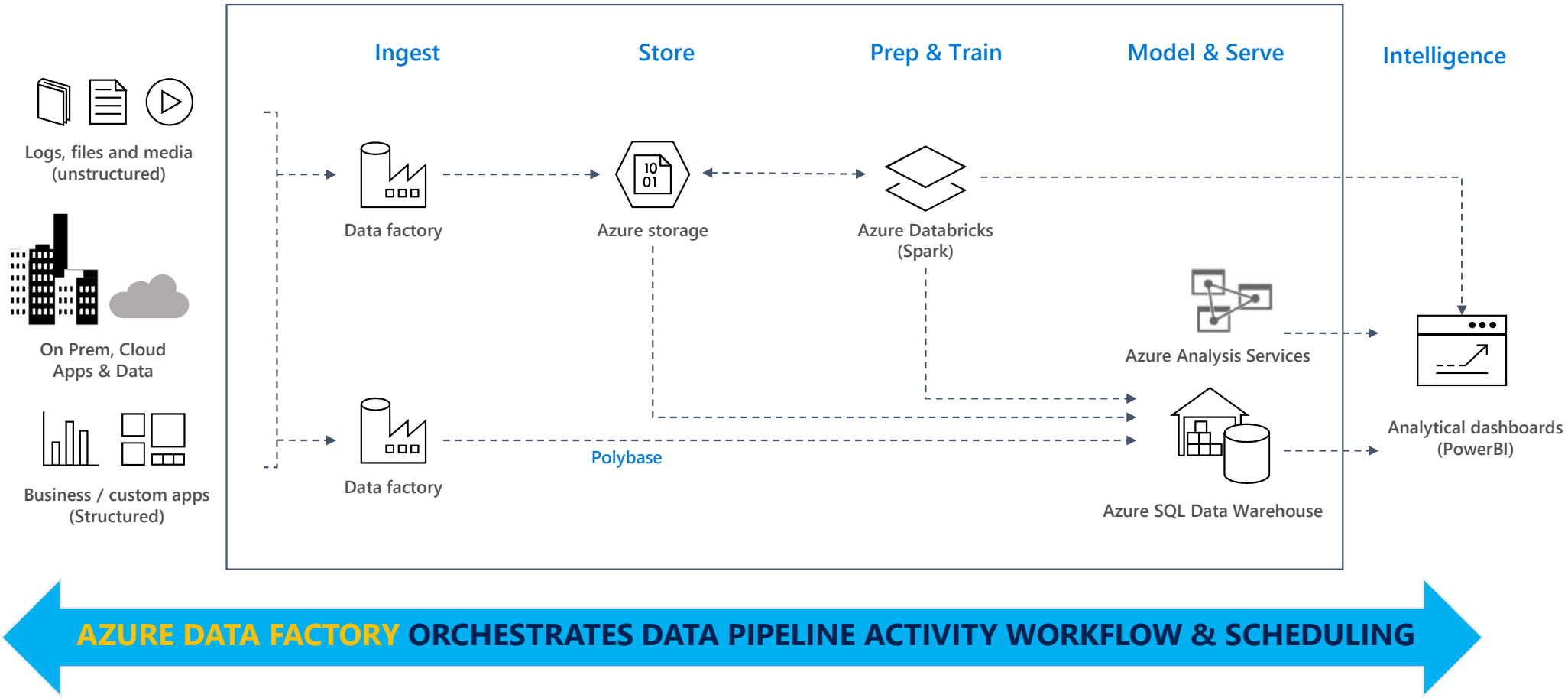
Data Engineering Scenarios

Modern DW for BI
Advance Analytics for Apps

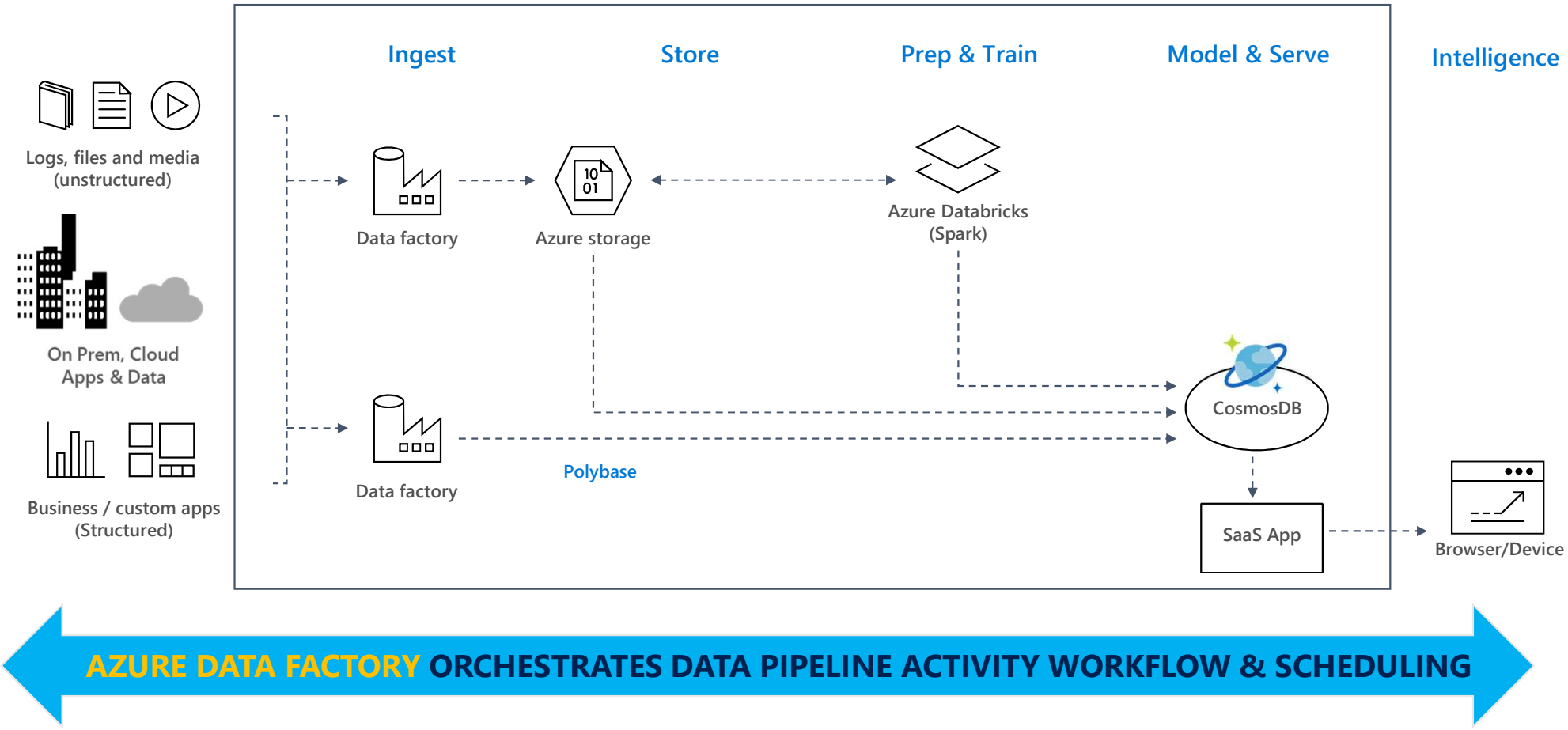
Advanced analytics and AI



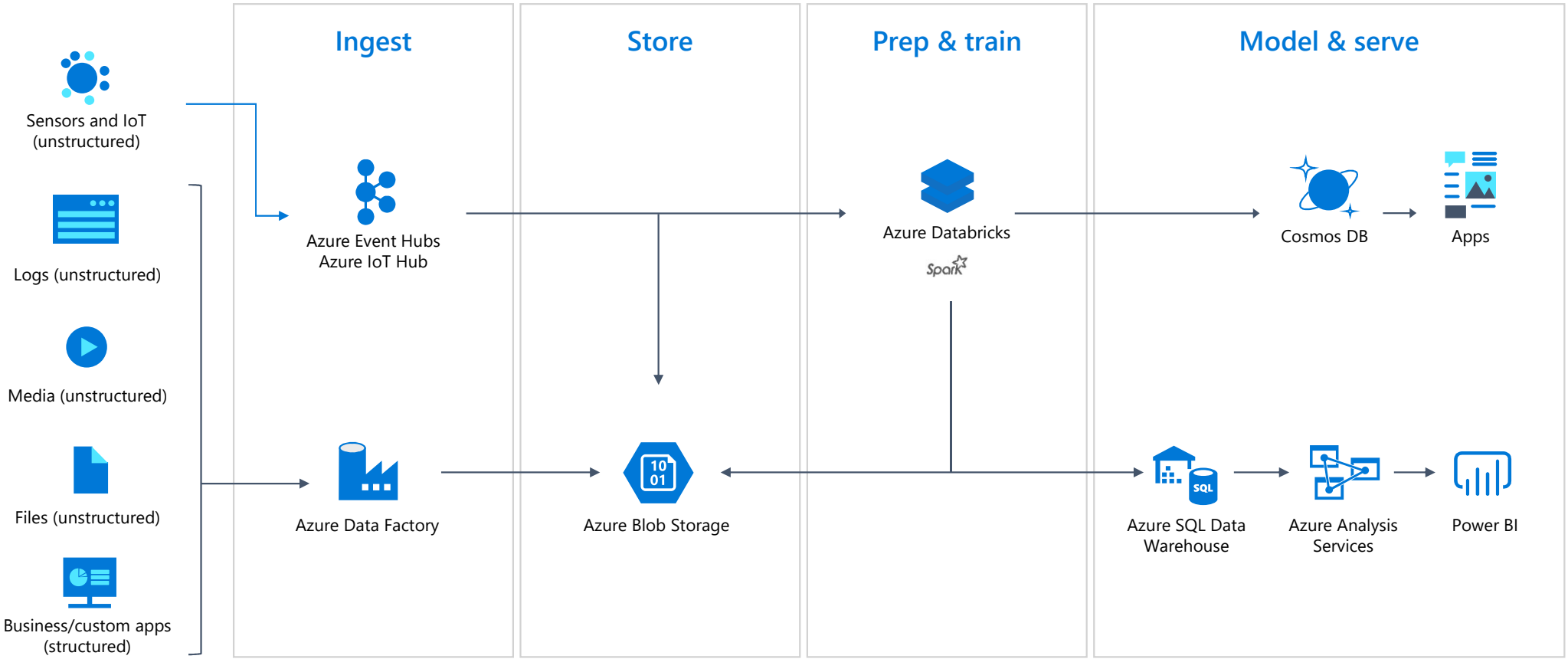
Modern DW for BI



Advance Analytics for Apps



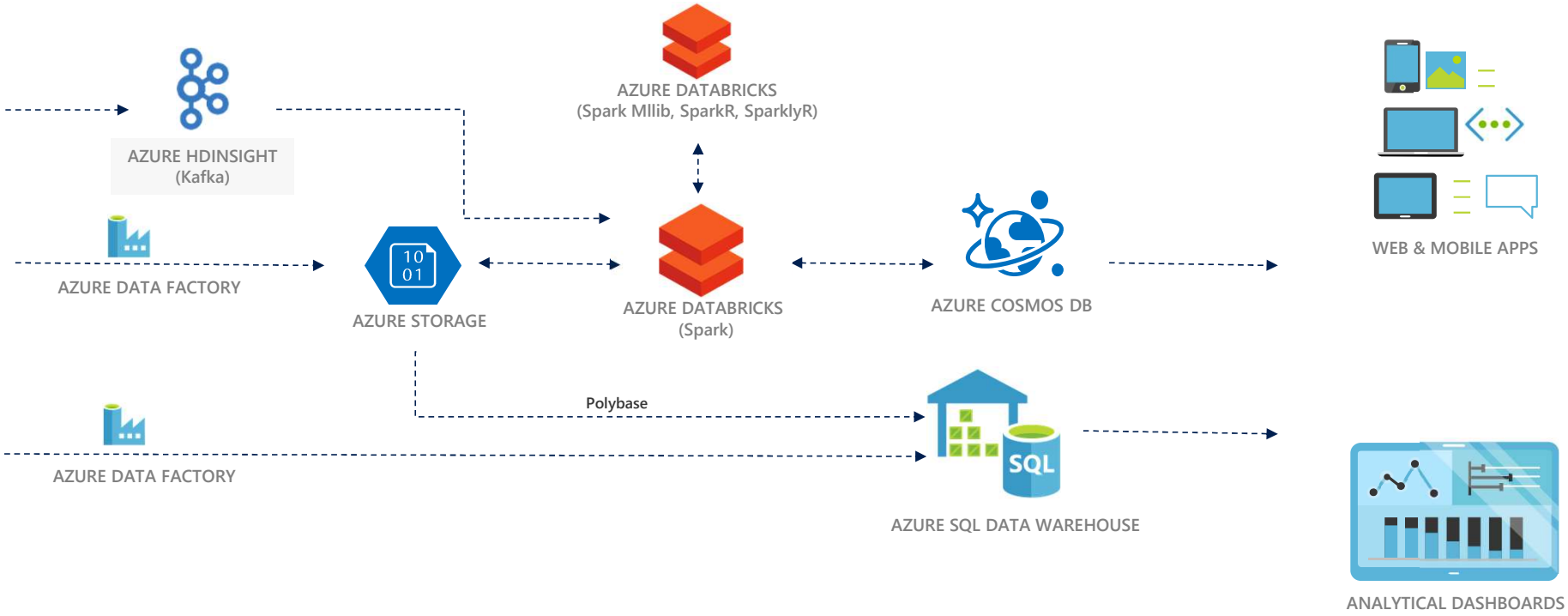
Real-Time Complex / Data Event Processing



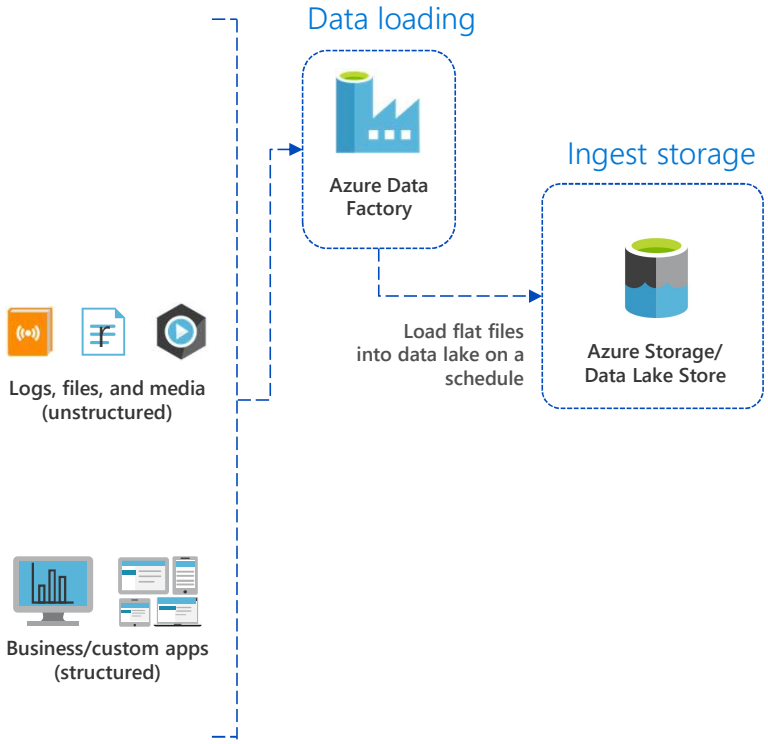
Microsoft Azure also supports other Big Data services like Azure HDInsight and Azure Data Lake to allow customers to tailor the above architecture to meet their unique needs.

Big Data Lambda Architecture

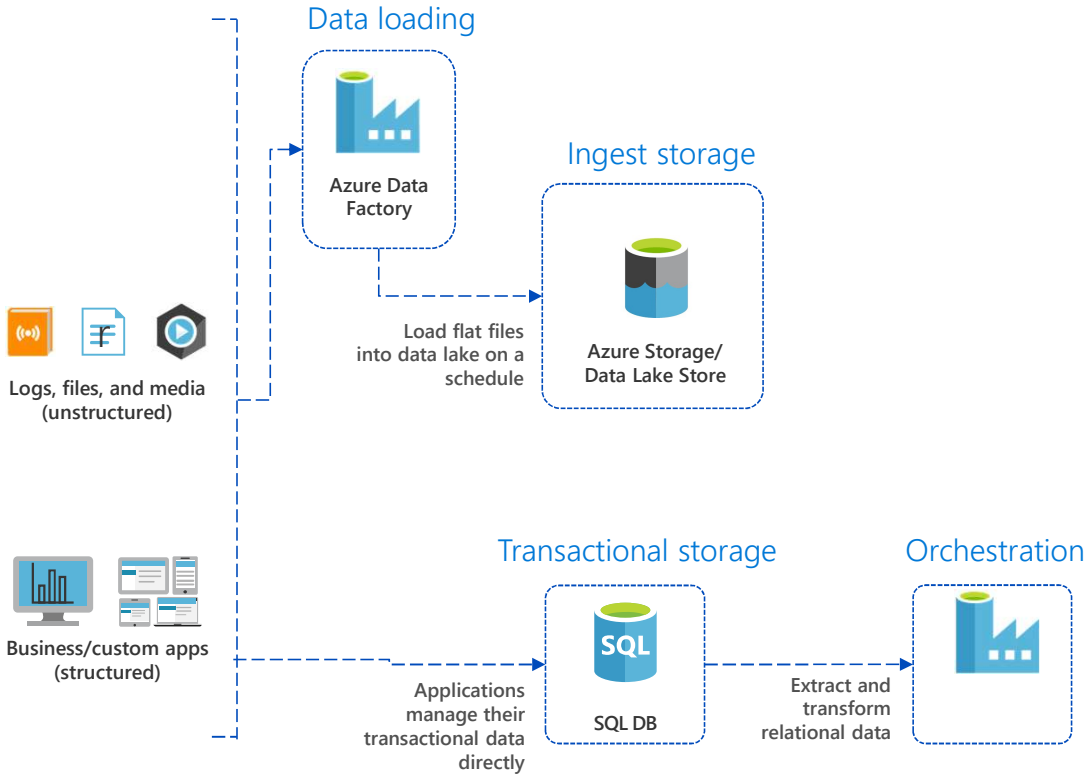
With Azure Databricks



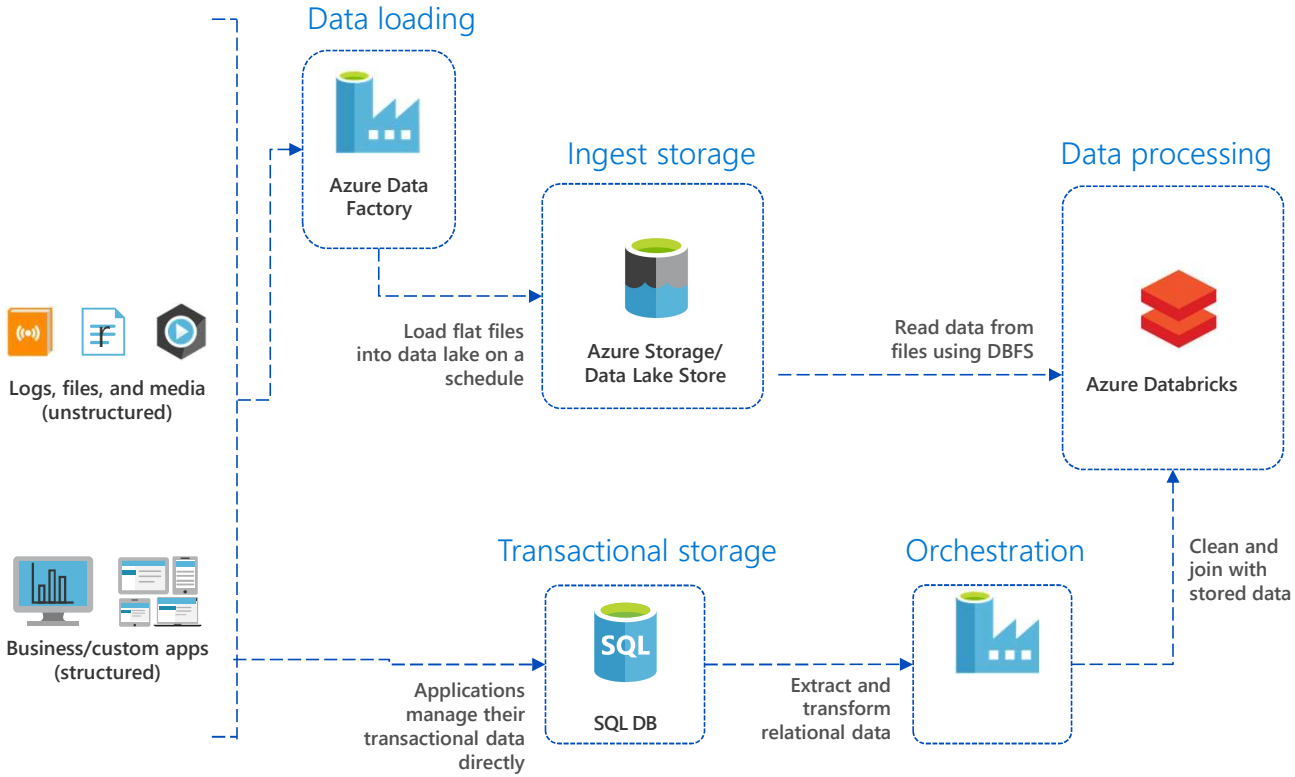
Data Ingestion



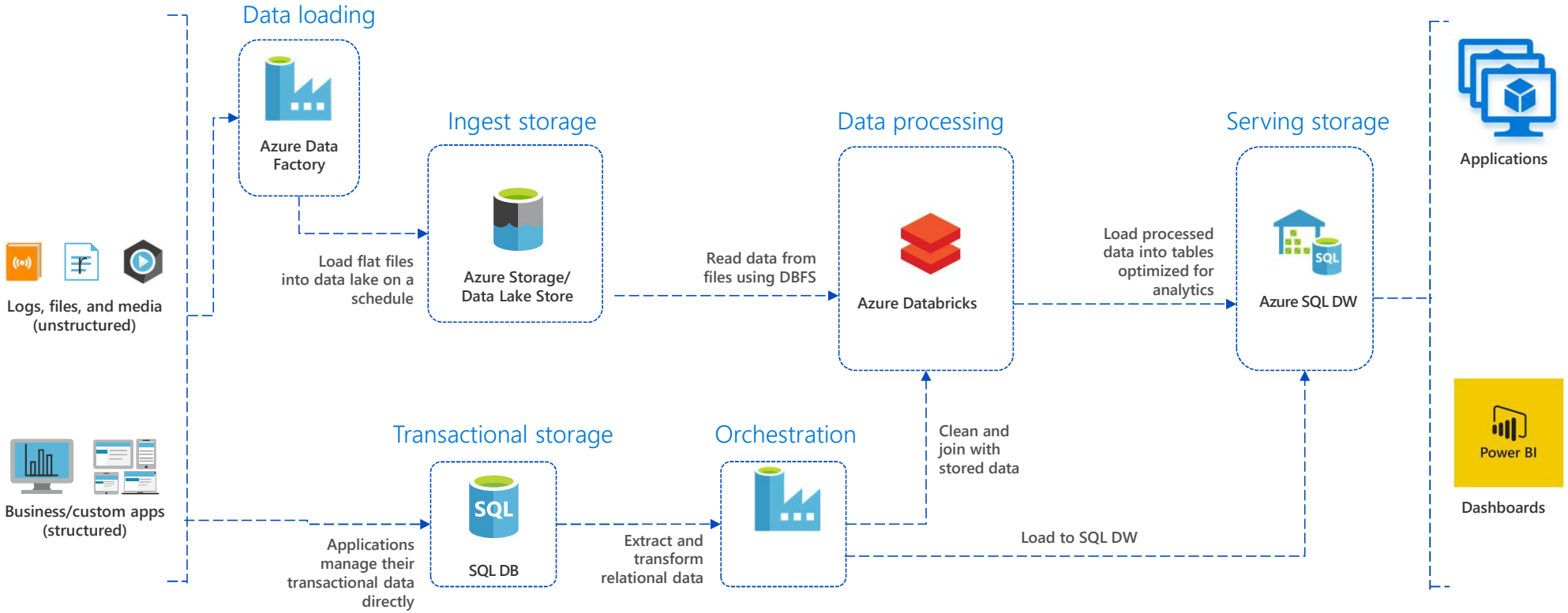
Data Ingestion



Data cleansing and prep



Analysis and Post processing

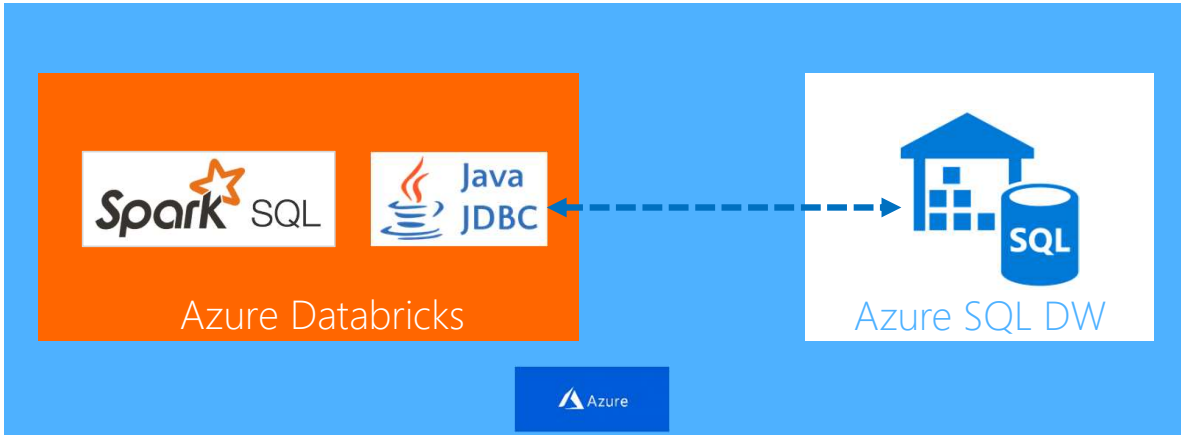


AZURE SQL DW INTEGRATION

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.



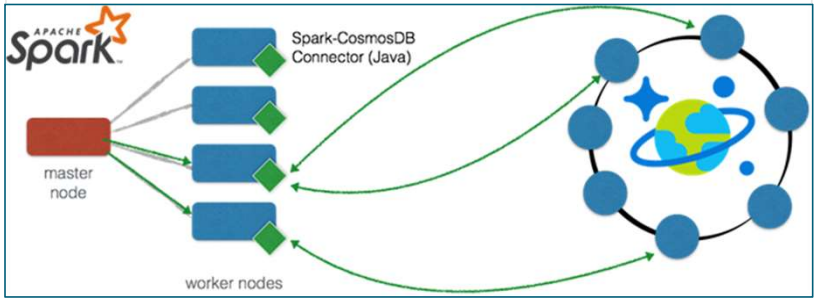
COSMOS DB INTEGRATION

The Spark connector enables real-time analytics over globally distributed data in Azure Cosmos DB

Azure Cosmos DB is Microsoft's globally distributed, multi-model database service for mission-critical applications

- With Spark connector for Azure Cosmos DB, Apache Spark can now interact with all Azure Cosmos DB data models: *Documents, Tables, and Graphs*.
  - efficiently exploits the native Azure Cosmos DB managed indexes and enables updateable columns when performing analytics.
  - utilizes push-down predicate filtering against fast-changing globally-distributed data
- Some use-cases for Azure Cosmos DB + Spark include:
  - Streaming Extract, Transformation, and Loading of data (ETL)
  - Data enrichment
  - Trigger event detection
  - Complex session analysis and personalization
  - Visual data exploration and interactive analysis
  - Notebook experience for data exploration, information sharing, and collaboration

The connector uses the Azure DocumentDB Java SDK and moves data directly between Spark worker nodes and Cosmos DB data nodes



AZURE BLOB STORAGE INTEGRATION

Data can be read from Azure Blob Storage using the Hadoop FileSystem interface. Data can be read from public storage accounts without any additional settings. To read data from a private storage account, you need to set an account key or a Shared Access Signature (SAS) in your notebook

Setting up an account key

```
spark.conf.set ( "fs.azure.account.key.{Your Storage Account Name}.blob.core.windows.net", "{Your Storage Account Access Key}")
```

Setting up a SAS for a given container:

```
spark.conf.set( "fs.azure.sas.{Your Container Name}.{Your Storage Account Name}.blob.core.windows.net", "{Your SAS For The Given Container}")
```

Once an account key or a SAS is setup, you can use standard Spark and Databricks APIs to read from the storage account:

```
val df = spark.read.parquet("wasbs://{Your Container Name}@m{Your Storage Account  name}.blob.core.windows.net/{Your Directory Name}")
dbutils.fs.ls("wasbs://{Your ntainer Name}@{Your Storage Account Name}.blob.core.windows.net/{Your Directory Name}")
```

AZURE DATA LAKE INTEGRATION

To read from your Data Lake Store account, you can configure Spark to use service credentials with the following snippet in your notebook

```
spark.conf.set("dfs.adls.oauth2.access.token.provider.type", "ClientCredential")
spark.conf.set("dfs.adls.oauth2.client.id", "{YOUR SERVICE CLIENT ID}")
spark.conf.set("dfs.adls.oauth2.credential", "{YOUR SERVICE CREDENTIALS}")
spark.conf.set("dfs.adls.oauth2.refresh.url", "https://login.windows.net/{YOUR DIRECTORY ID}/oauth2/token")
```

After providing credentials, you can read from Data Lake Store using standard APIs:

```
val df = spark.read.parquet("adl://{YOUR DATA LAKE STORE ACCOUNT NAME}.azuredatalakestore.net/{YOUR DIRECTORY NAME}")
dbutils.fs.list("adl://{YOUR DATA LAKE STORE ACCOUNT NAME}.azuredatalakestore.net/{YOUR DIRECTORY NAME}")
```

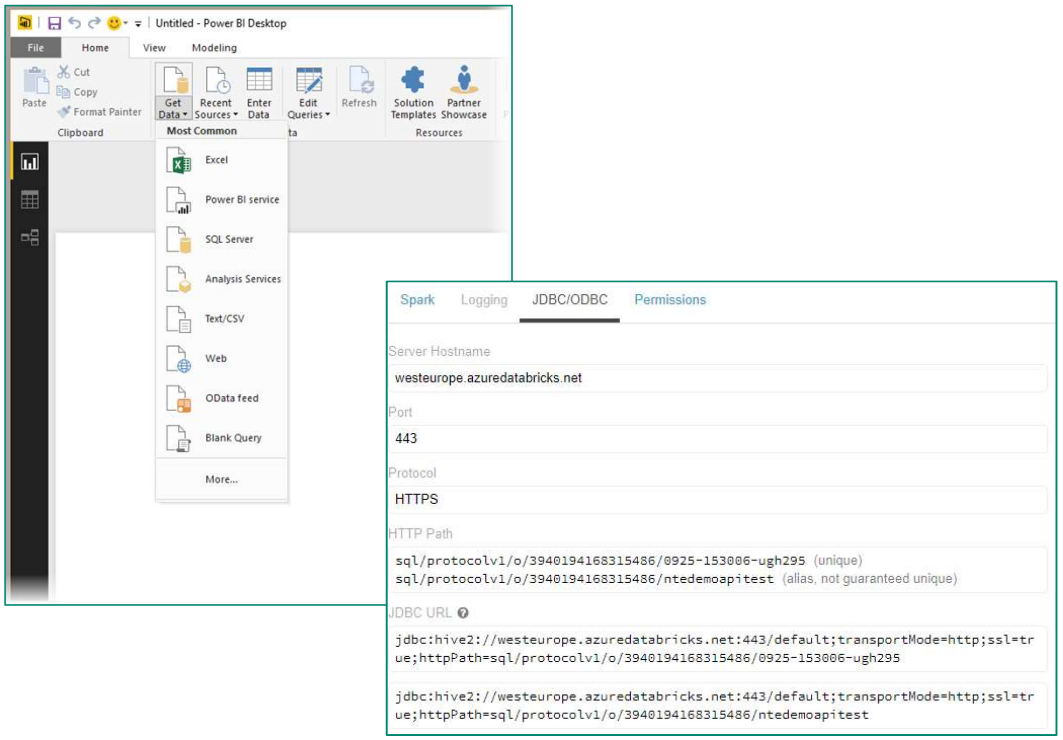
POWER BI INTEGRATION

Enables powerful visualization of data in Spark with Power BI

Power BI is a business analytics tool that provides data Visualization, Report and Dashboard throughout an organization

Power BI Desktop can connect to Azure Databricks clusters to query data using JDBC/ODBC server that runs on the driver node.

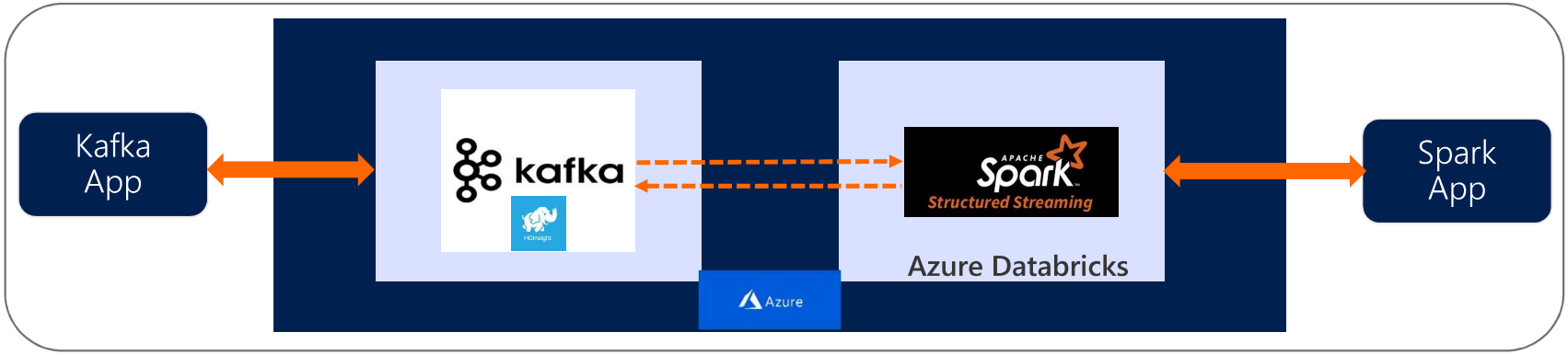
- This server listens on port 10000 and it is not accessible outside the subnet where the cluster is running.
- Azure Databricks uses a public HTTPS gateway
- The JDBC/ODBC connection information can be obtained from the Cluster UI directly as shown in the figure.
- When establishing the connection, you can use a Personal Access Token to authenticate to the cluster gateway. Only users who have attach permissions can access the cluster via the JDBC/ ODBC endpoint.
- In Power BI desktop you can setup the connection by choosing the ODBC data source in the "Get Data" option.



APACHE KAFKA FOR HDINSIGHT INTEGRATION

Azure Databricks Structured Streaming integrates with Apache Kafka for HDInsight

- Apache Kafka for Azure HDInsight is an enterprise grade streaming ingestion service running in Azure.
- Azure Databricks Structured Streaming applications can use Apache Kafka for HDInsight as a data source or sink.
- No additional software (gateways or connectors) are required.
- Setup: Apache Kafka on HDInsight does not provide access to the Kafka brokers over the public internet. So the Kafka clusters and the Azure Databricks cluster must be located in the same Azure Virtual Network.



Note: Azure Databricks Structured Streaming integration with **Azure Event Hubs** is forthcoming

ETL – Databricks Spark

```
spark.read.json("/source/path")
    .filter(...)
  .agg(...)
    .write.mode("append")
    .parquet("/output/path")
```

EXTRACT

TRANSFORM

LOAD

An ETL Query in Apache Spark

```
val csvTable = spark.read.csv("/source/path")
val jdbcTable = spark.read.format("jdbc")
  .option("url", "jdbc:postgresql:...")
  .option("dbtable", "TEST.PEOPLE")
  .load()

csvTable
  .join(jdbcTable, Seq("name"), "outer")
  .filter("id <= 2999")
  .write
  .mode("overwrite")
  .format("parquet")
  .saveAsTable("outputTableName")
```

EXTRACT

TRANSFORM

LOAD

Json: Dealing with Corrupt Records

```
{"a":1, "b":2, "c":3}
{"a":{, b:3}
{"a":5, "b":6, "c":7}
```

```
+--------------+----+----+----+
|_corrupt_record|   a|   b|   c|
+--------------+----+----+----+
|           null|   1|   2|   3|
|   {"a":{, b:3}|null|null|null|
|           null|   5|   6|   7|
+--------------+----+----+----+
```

```
spark.read
 .option("mode", "PERMISSIVE")
 .option("columnNameOfCorruptRecord", "_corrupt_record")
 .json(corruptRecords)
 .show()
```

The default can be configured via
spark.sql.columnNameOfCorruptRecord

Json: Dealing with Corrupt Records

```
{"a":1, "b":2, "c":3}
{"a":{, b:3}
{"a":5, "b":6, "c":7}
```

```
spark.read
  .option("mode", "DROPMALFORMED")
  .json(corruptRecords)
  .show()
```

```
+---+---+---+
|  a|  b|  c|
+---+---+---+
|  1|  2|  3|
|  5|  6|  7|
+---+---+---+
```

Json: Dealing with Corrupt Records

```
{"a":1, "b":2, "c":3}
{"a":{, b:3}
{"a":5, "b":6, "c":7}
```

```
spark.read
 .option("mode", "FAILFAST")
 .json(corruptRecords)
 .show()
```

```
org.apache.spark.sql.catalyst.json
.SparkSQLJsonProcessingException:
Malformed line in FAILFAST mode:
{"a":{, b:3}
```

CSV: Dealing with Corrupt Records

```
year,make,model,comment,blank
"2012","Tesla","S","No comment",
1997,Ford,E350,"Go get one now they",
2015,Chevy,Volt
```

```
spark.read.
  .option("mode", "PERMISSIVE")
  .csv(corruptRecords)
  .show()
```

```
+----+--------+-----+--------------------+-----+
| _c0|     _c1|  _c2|                 _c3|  _c4|
+----+--------+-----+--------------------+-----+
|year|    make|model|             comment|blank|
|2012| "Tesla"|  "S"|        "No comment"| null|
|1997|    Ford| E350|"Go get one now ...| null|
|2015|   Chevy| Volt|                null| null|
+----+--------+-----+--------------------+-----+
```

CSV: Dealing with Corrupt Records

```
year,make,model,comment,blank
"2012","Tesla","S","No comment",
1997,Ford,E350,"Go get one now they",
2015,Chevy,Volt
```

```
spark.read
  .option("header", true)
  .option("mode", "PERMISSIVE")
  .csv(corruptRecords)
  .show()
```

```
+----+-------+-----+--------------------+-----+
|year|   make|model|             comment|blank|
+----+-------+-----+--------------------+-----+
|2012|"Tesla"|  "S"|        "No comment"|  null|
|1997|   Ford| E350|"Go get one now ...|  null|
|2015|  Chevy| Volt|                null|  null|
+----+-------+-----+--------------------+-----+
```

CSV: Dealing with Corrupt Records

```
val schema = "col1 INT, col2 STRING, col3 STRING, col4 STRING, " +
  "col5 STRING, __corrupted_column_name STRING"
spark.read
  .option("header", true)
  .option("mode", "PERMISSIVE")
  .csv(corruptRecords)
  .show()
```

```
+----+--------+-----+--------------------+----+------------------------+
|col1|    col2| col3|                col4|col5|__corrupted_column_name|
+----+--------+-----+--------------------+----+------------------------+
|2012| "Tesla"|  "S"|        "No comment"|null|                    null|
|1997|    Ford| E350| "Go get one now ...|null|                    null|
|2015|   Chevy| Volt|                null|null|      2015, Chevy, Volt|
+----+--------+-----+--------------------+----+------------------------+
```

CSV: Dealing with Corrupt Records

```
year,make,model,comment,blank
"2012","Tesla","S","No comment",
1997,Ford,E350,"Go get one now they",
2015,Chevy,Volt
```

```
spark.read
  .option("mode", ”DROPMALFORMED")
  .csv(corruptRecords)
  .show()
```

```
+----+-----+-----+--------------------+-----+
|year| make|model|             comment|blank|
+----+-----+-----+--------------------+-----+
|2012|Tesla|    S|          No comment| null|
|1997| Ford| E350|Go get one now th...| null|
+----+-----+-----+--------------------+-----+
```

Schema Inference – semi-structured files

```
{"a":1, "b":2, "c":3}
{"e":2, "c":3, "b":5}
{"a":5, "d":7}
```

```
spark.read
  .json("/source/path”)
  .printSchema()
```

```
root
 |-- a: long (nullable = true)
 |-- b: long (nullable = true)
 |-- c: long (nullable = true)
 |-- d: long (nullable = true)
 |-- e: long (nullable = true)
```

Schema Inference – semi-structured files

```
{"a":1, "b":2, "c":3.1}
{"e":2, "c":3, "b":5}
{"a":"5", "d":7}
```

```
spark.read
  .json("/source/path”)
  .printSchema()
```

```
root
 |-- a: string (nullable = true)
 |-- b: long (nullable = true)
 |-- c: double (nullable = true)
 |-- d: long (nullable = true)
 |-- e: long (nullable = true)
```

User-specified Schema

```
{"a":1, "b":2, "c":3}
{"e":2, "c":3, "b":5}
{"a":5, "d":7}
```

```
+----+----+
|   a|   b|
+----+----+
|   1|   2|
|null|   5|
|   5|null|
+----+----+
```

```
val schema = new StructType()
  .add("a", "int")
  .add("b", "int")

spark.read
  .json("/source/path")
  .schema(schema)
  .show()
```

Traditional ETL

Raw, dirty, un/semi-structured is data dumped as files

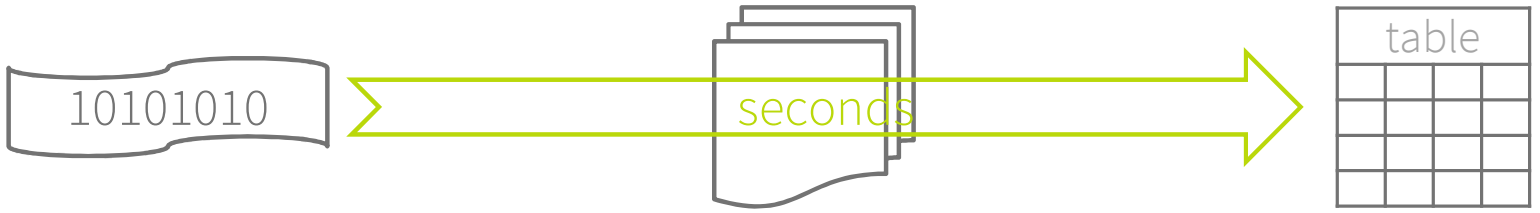
Periodic jobs run every few hours to convert raw data to structured data ready for further analytics

Traditional ETL

Hours of delay before taking decisions on latest data

Unacceptable when time is of essence
[intrusion detection, anomaly detection, etc.]

Streaming ETL w/ Structured Streaming



Structured Streaming enables raw data to be available as structured data as soon as possible

Azure Databricks Delta

Public Preview

Optimize Spark Queries for
faster analytics,
better data reliability guarantees,
and simplified data pipelines

Why Databricks Delta?

Data Engineering Challenges

1. Data Reliability Issues

Inconsistent Data
Evolving Schema

2. Performance Challenges

Slow and Costly
Worse at Scale

LOTS OF NEW DATA

User Behavior Data

Click Streams

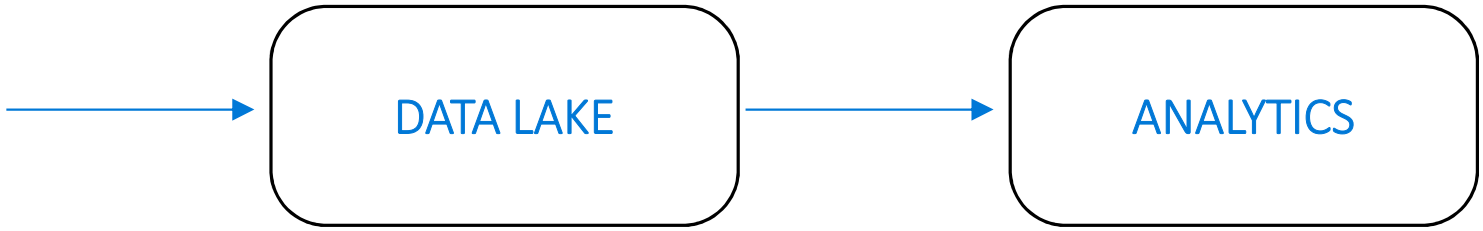
Sensor Data (IoT)

Video/Speech

Usage/Billing Data

Machine Telemetry

Commerce Data



3. System Complexity

Separate batch & streaming
Low-level code for pipelines

Why Databricks Delta?

Data Engineering Challenges

1. Data Reliability

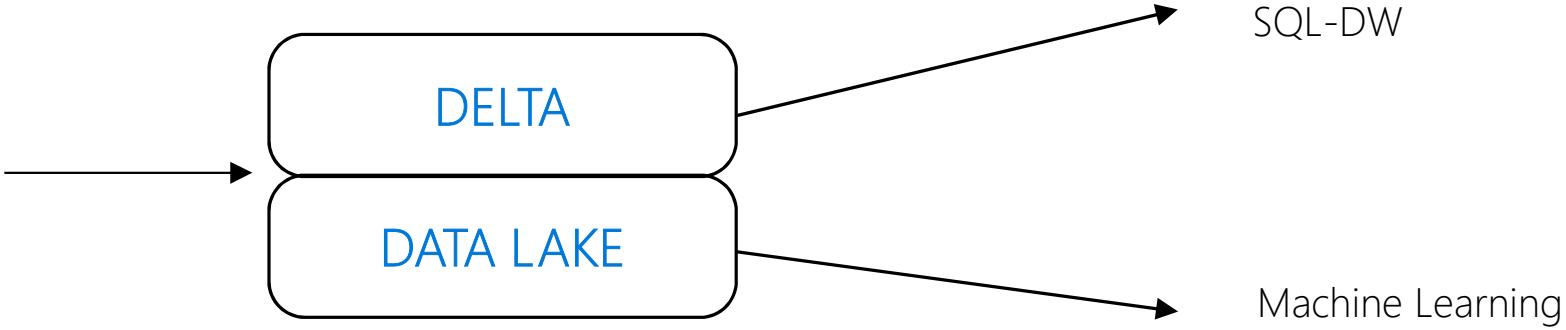
ACID Compliant Transactions
Schema Enforcement & Evolution

2. Spark Query Performance

Fast at Scale (10-100x Faster)
Cheaper to Operate
Indexing, Statistics & Caching

LOTS OF NEW DATA

- User Behavior Data
- Click Streams
- Sensor data (IoT)
- Video/Speech
- Usage/Billing data
- Machine Telemetry
- Commerce Data



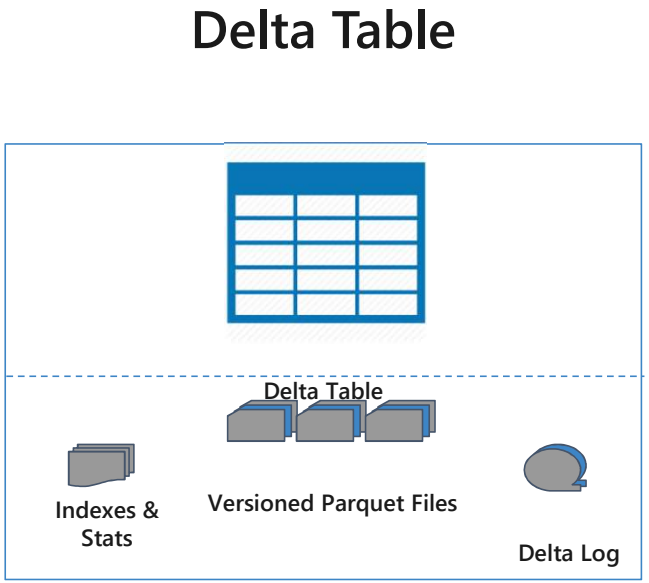
3. Simplified Architecture

Unify batch & streaming
Early data availability for analytics

Azure Databricks Delta Architecture

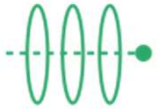
Delta Table = Parquet + Transaction Log

- Linear history of atomic changes
- Optimistic Concurrency Control
- Log checkpoint is stored as Parquet
- Lazy GC = Free Snapshot Isolation



Azure Databricks Delta

Handle terabytes & petabytes of data



Low latency streaming ingestion

Avoid corrupt & messy data while reading & writing

Control on how to adapt to changing schema

Enable scientists & analysts to read data quickly for interactive analysis - Indexing

Azure Databricks Delta – Fast Reads

Data format

Compaction

Partitioning

Indexing

Caching

An ETL Query In Spark

```
val csvTable = spark.read.csv("/source/path")

val jdbcTable = spark.read.format("jdbc")
  .option("url", "jdbc:postgresql:...")
  .option("dbtable", "TEST.PEOPLE")
  .load()

csvTable
  .join(jdbcTable, Seq("name"), "outer")
  .filter("id <= 2999")
  .write
  .mode("overwrite")
  .format("delta")
  .saveAsTable("outputTableName")
```

Read Table 1

Read Table 2

Join

Filter

Write

Performance Tips

- Land data in Blob Store/ADLS partitioned into separate directory
  - Avoid high list cost on large directories
- Parallelization of Azure Databricks streaming is driven by number of partitions in Eventhub
- For best query performance use Delta table. Alternatively, use regular Spark table backed by Parquet
  - ORC OK if customer has a preference
- Avoid small files. File size 100s MB – 1GB preferred
  - Delta supports compaction

ETL – Spark Databricks

Azure AI

AI apps & agents

Azure Bot Service
Azure Cognitive Services

Machine learning

Azure Databricks
Azure Machine Learning

Knowledge mining

Azure Cognitive Search

Azure AI

AI apps & agents

Azure Bot Service
Azure Cognitive Services

Machine learning

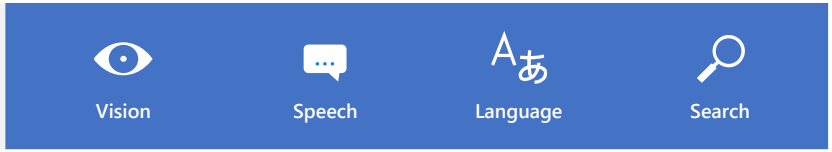
Azure Databricks
Azure Machine Learning

Knowledge mining

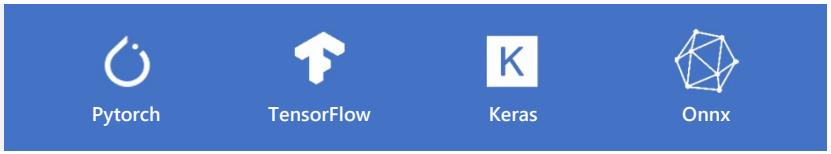
Azure Cognitive Search

Machine Learning on Azure

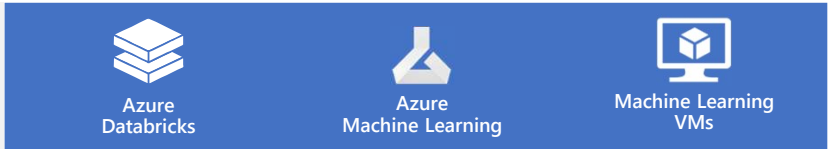
Sophisticated pretrained models
To simplify solution development



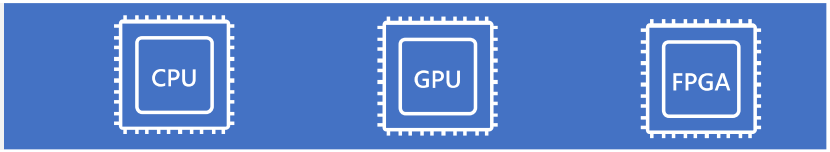
Popular frameworks
To build advanced deep learning solutions



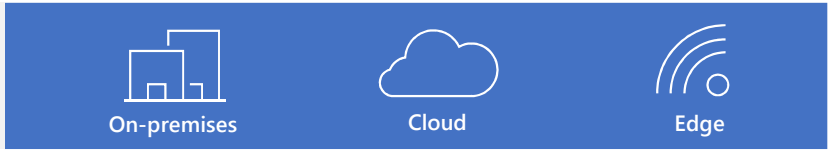
Productive services
To empower data science and development teams



Powerful infrastructure
To accelerate deep learning

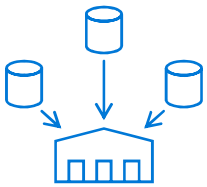


Flexible deployment
To deploy and manage models on intelligent cloud and edge



How companies are transforming

Serving business users and end users with **intelligent** and **dynamic** applications

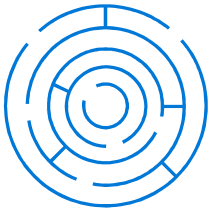


Build a unified and usable data pipeline

Train ML and DL models to derive insights

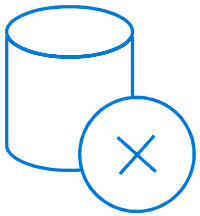
Operationalize models and distribute insights at scale

But there's a lot to consider



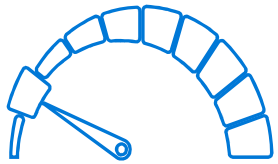
Complexity of solutions

Many options in the marketplace



Data silos

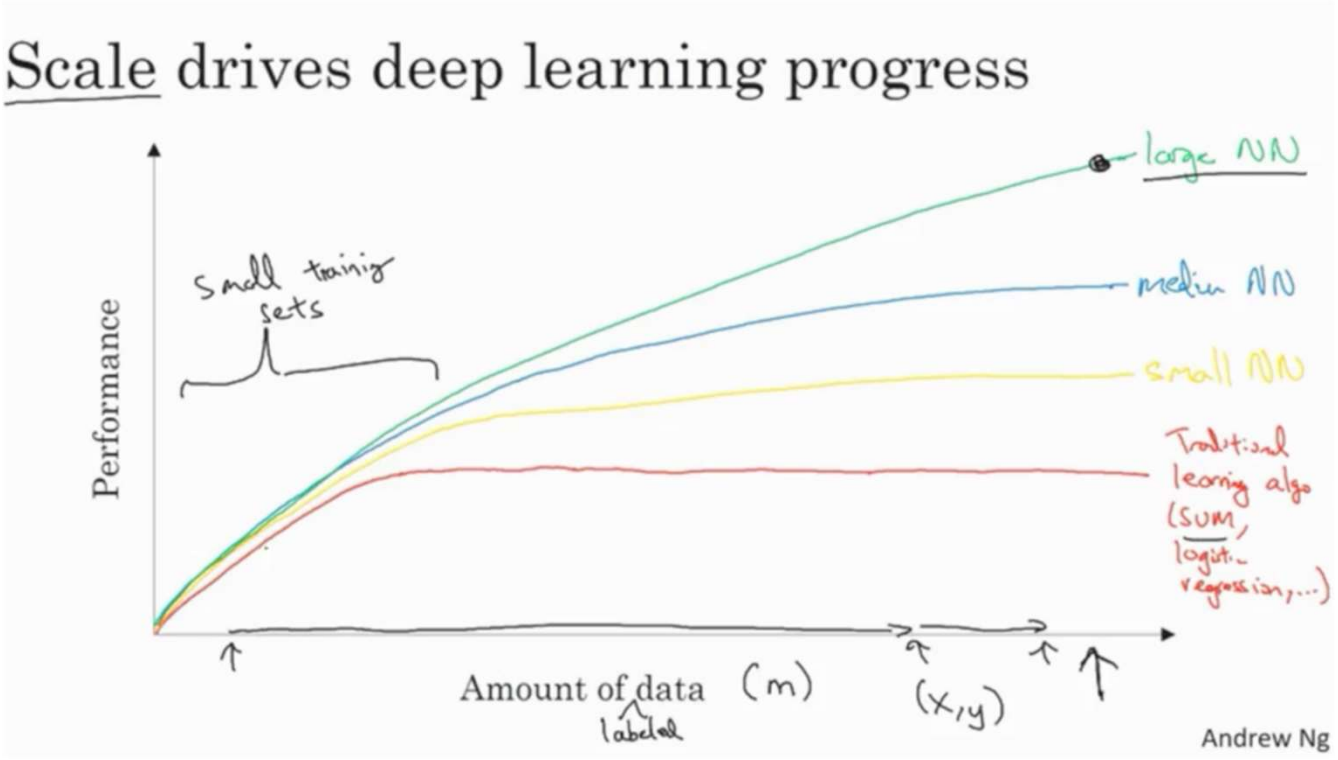
Incongruent data types



Difficult to scale effectively

Performance constraints

DL model accuracy goes up w/ data size!



From Andrew Ng's Coursera course

Challenges for Data Scientists

- Infrastructure management
- Data exploration and visualization at scale
- Time to value - From model iterations to intelligence
- Integrating with various ML tools to stitch a solution together
- Operationalize ML models to integrate them into applications

Typical Data Science Projects

DATA EXPLORATION

ANALYSIS AND MODELING

COMMUNICATE RESULTS

OPERATIONALIZE MODELS

What do Data Scientists care about?

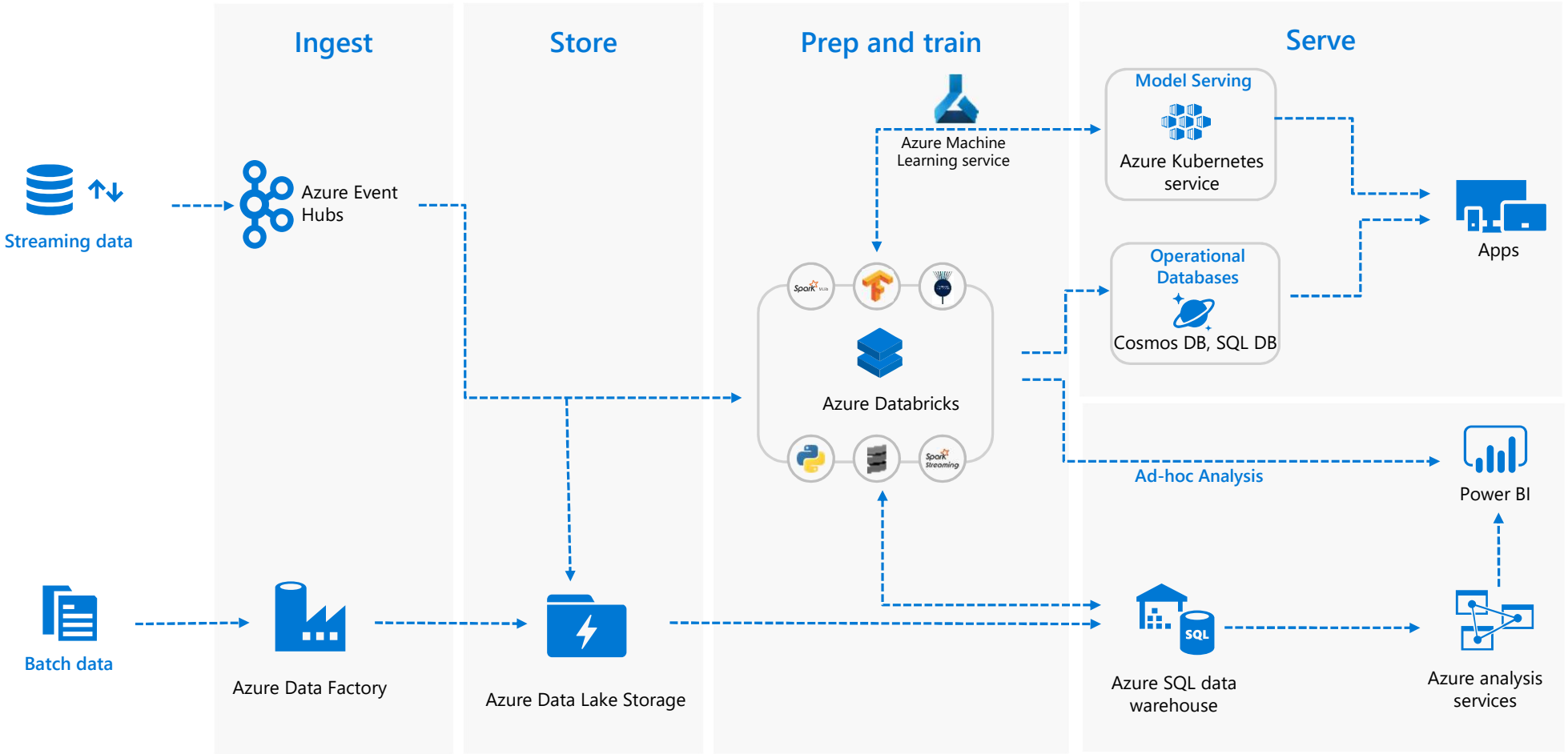
RAPID EXPERIMENTATION

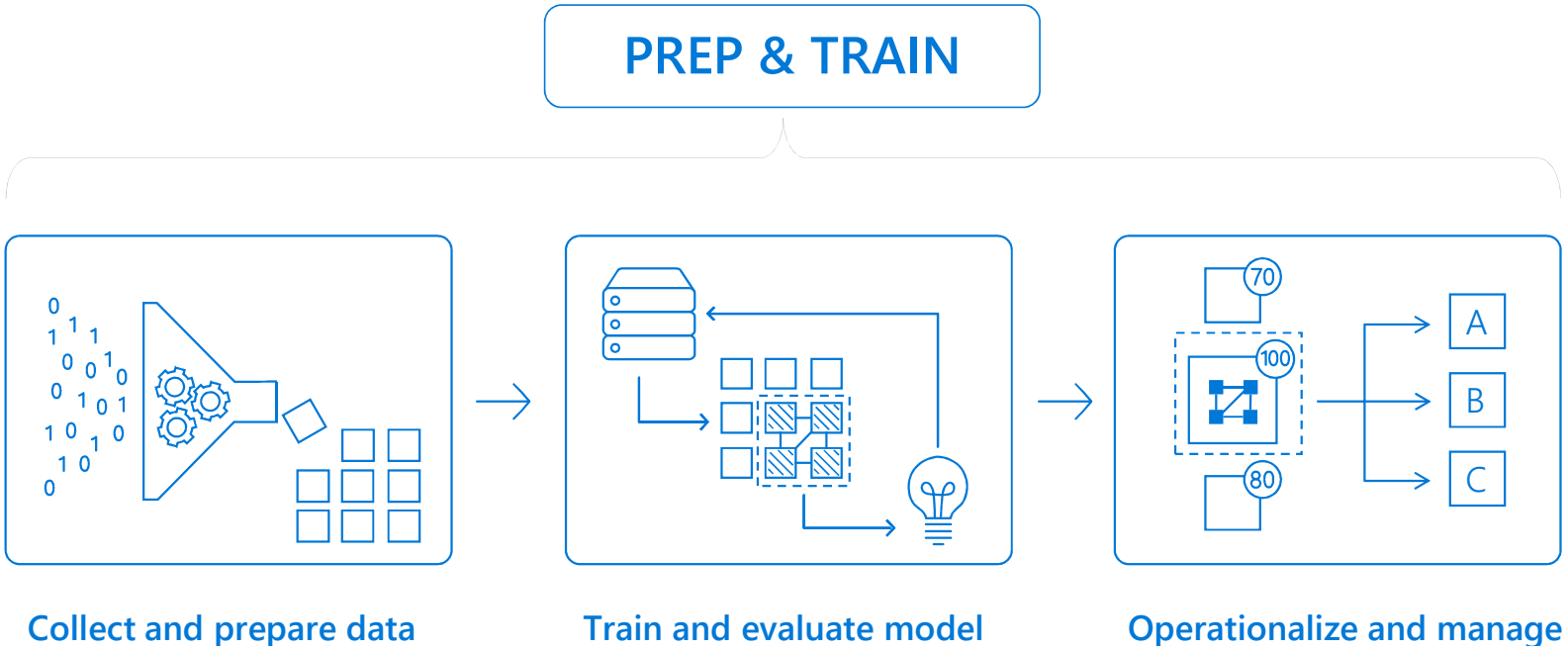
DATA VISUALIZATION

CROSS-TEAM COLLABORATION

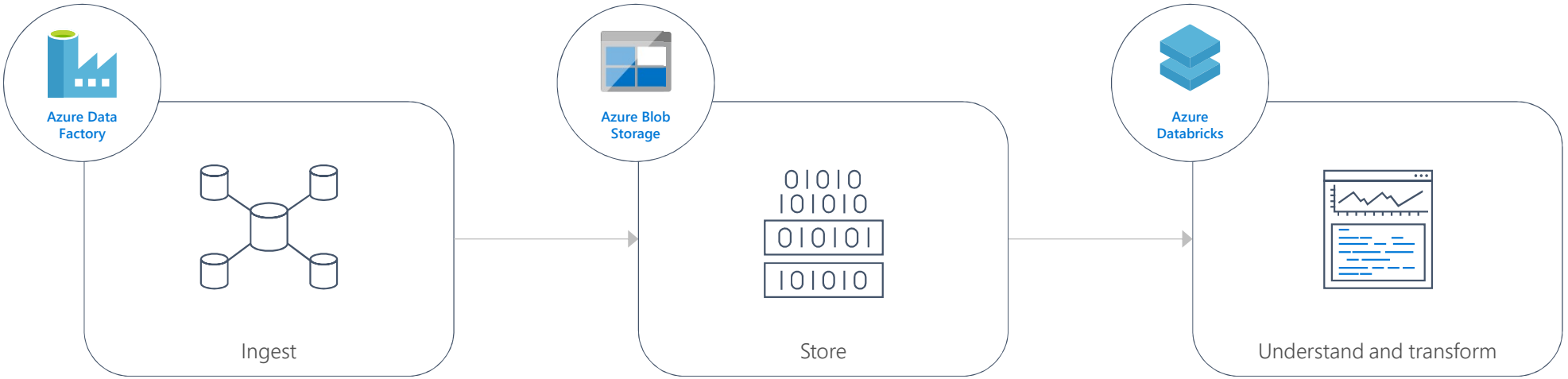
EASY SHARING OF INSIGHTS

Recommended architecture to build e2e ML solutions




PREP & TRAIN
Collect and prepare data
Train and evaluate model
Operationalize and manage
70
100
80
A
B
C

Collect and prepare all of your data at scale



Connect to data from any source

- Integrate with all of your data sources
- Create hybrid pipelines
- Orchestrate in a code-free environment

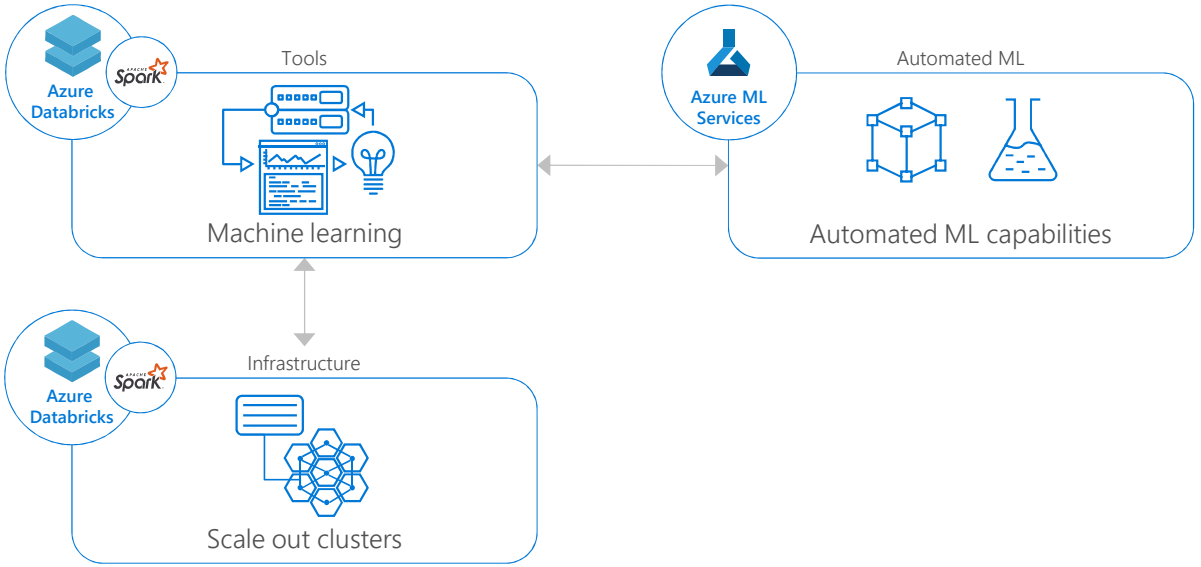
Leverage best-in-class analytics capabilities

- Leverage open source technologies
- Collaborate within teams
- Use ML on batch streams

Scale without limits

- Build in the language of your choice
- Leverage scale out topology
- Scale compute and storage separately

Train and evaluate machine learning models



Simplify model development

- Collaborate in interactive workspaces
- Access a library of battle-tested models
- Automate job execution

Scale compute resources to meet your needs

- Easily scale up or scale out
- Autoscale on serverless infrastructure
- Leverage commodity hardware

Quickly determine the right model for your data

- Determine the best algorithm
- Tune hyperparameters to optimize models
- Rapidly prototype in agile environments

3 Ways for Machine Learning

#1 Scalable Machine Learning with Spark MLlib

- Goal is to make practical machine learning extremely ***scalable*** and ***easy***
- Common Algorithms, Featurization, Pipelines, and Utilities need for ML
- Subset of all ML techniques, but ***extremely scalable***

#2 Single Machine Data Science on Big Data with Azure Databricks

- Use ADB to query “Big Data” stored on ADLS or Blob
- Use Spark to Aggregate, Sample “Big Data” to make it “small data”
- Collect this “small data” back to the driver for normal smaller data ML tools, R, Scikit-learn, etc

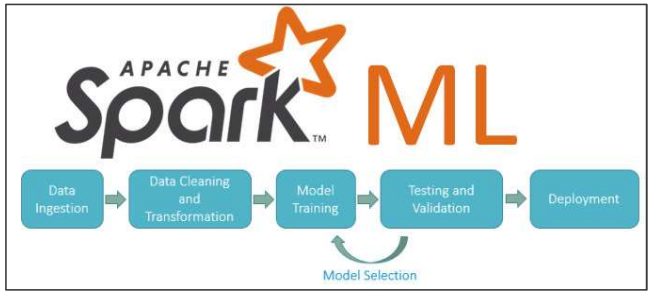
#3 Scale Out / Parallelization for Single Machine Data Science

- Combination of the above two
- Use Databricks for cross validation, training a bunch of small models, etc
- Apply user defined functions from R and Python

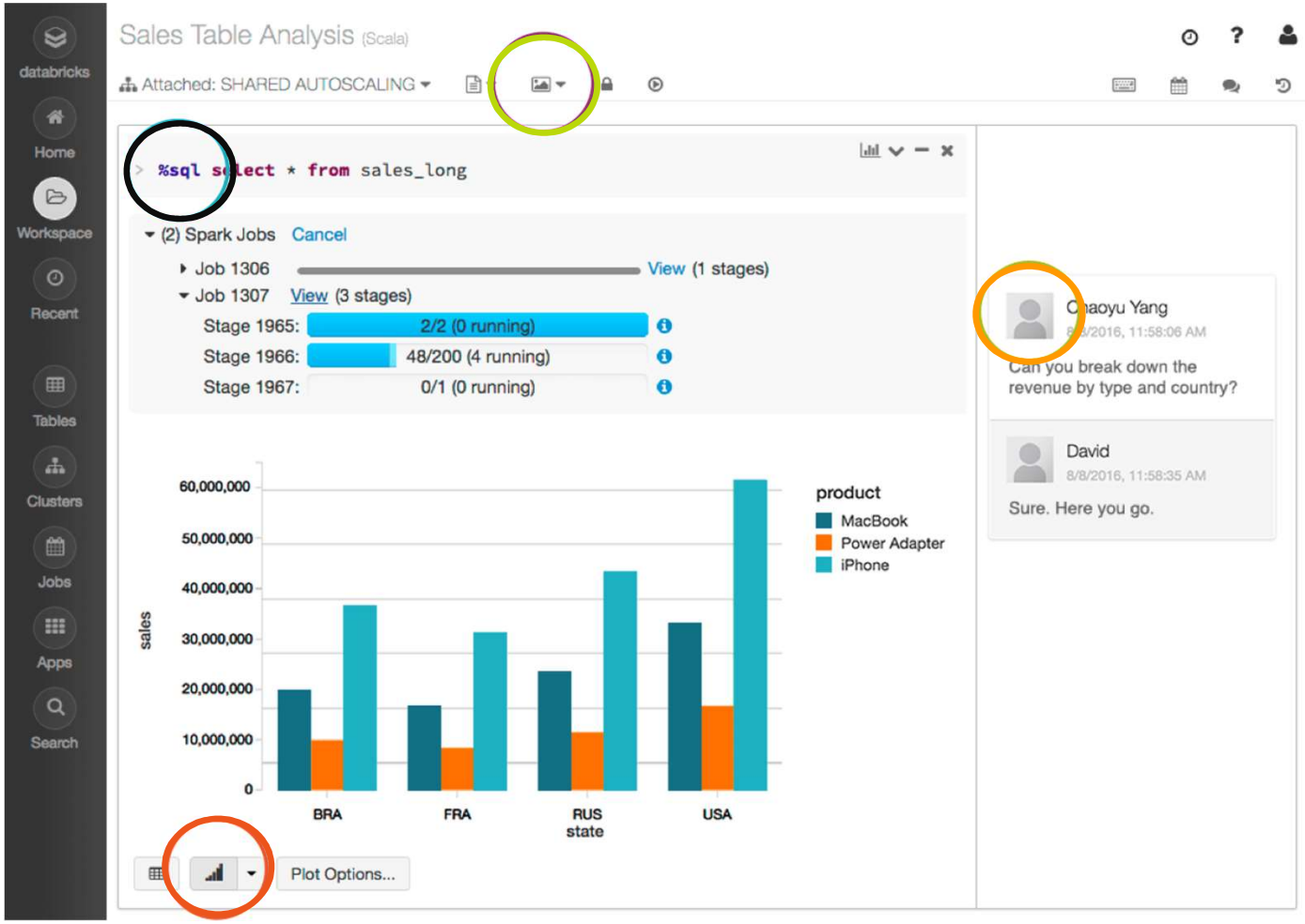
SPARK MACHINE LEARNING(ML) OVERVIEW

Enables Parallel, Distributed ML for large datasets on Spark Clusters

- Offers a set of parallelized machine learning algorithms (see next slide)
- Supports Model Selection (hyperparameter tuning) using Cross Validation and Train-Validation Split.
- Supports Java, Scala or Python apps using DataFrame-based API (as of Spark 2.0). Benefits include:
  - An uniform API across ML algorithms and across multiple languages
  - Facilitates ML pipelines (enables combining multiple algorithms into a single pipeline).
  - Optimizations through Tungsten and Catalyst
- Spark MLlib comes pre-installed on Azure Databricks
- 3rd Party libraries supported include: H20 Sparkling Water, SciKit-learn and XGBoost



Interactive Data Science



SPARK ML ALGORITHMS

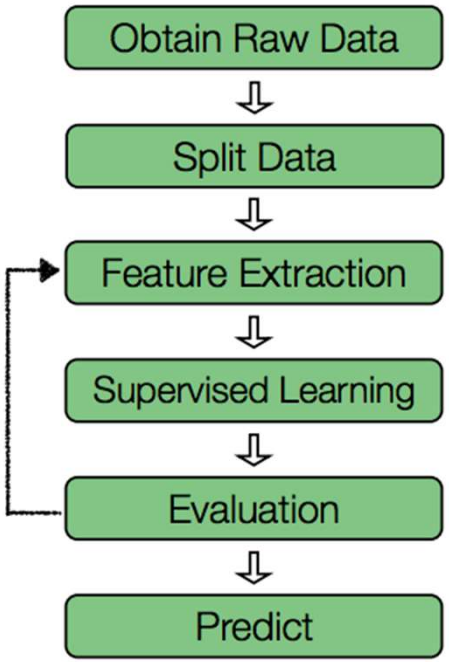
Spark ML Algorithms

Classification and Regression	• Linear Models (SVMs, logistic regression, linear regression) • Naïve Bayes • Decision Trees • Ensembles of trees (Random Forest, Gradient-Boosted Trees) • Isotonic regression
Clustering	• k-means and streaming k-means • Gaussian mixture • Power iteration clustering (PIC) • Latent Dirichlet allocation (LDA)
Collaborative Filtering	• Alternating least squares (ALS)
Dimensionality Reduction	• SVD • PCA
Frequent Pattern Mining	• FP-growth • Association rules
Basic Statistics	• Summary statistics • Correlations • Stratified sampling • Hypothesis testing • Random data generation

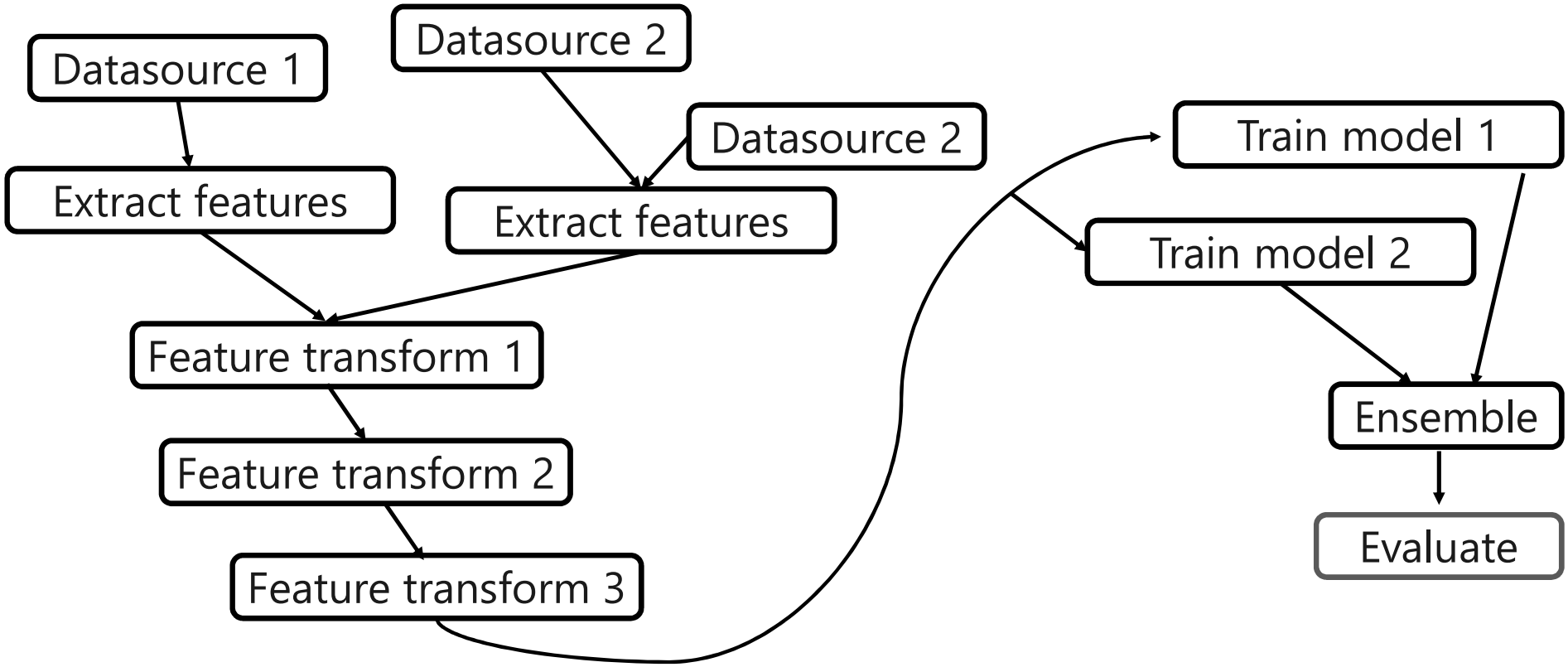
Why use Azure Databricks for Machine learning?

- Complete platform in one (Data ingestion, exploration, transformation, featurization, model building, model tuning, and even model serving).
- No need to copy the data in our system to do ML on it.
- DataScientists like the ease of use of our platform.
- Deep learning algorithms are now available!
- Productionization Features built in.

ML Pipelines



ML Pipelines



Estimator

- More complex feature processing, and/or prediction
- Transformer that is initialized with data (eg – convert a column into %tile rep and would require to initialize it)
- Actual models that be trained and tuned so we can use for prediction
  - Logistic Regression (classification)
  - Decision Tree
  - Random Forest

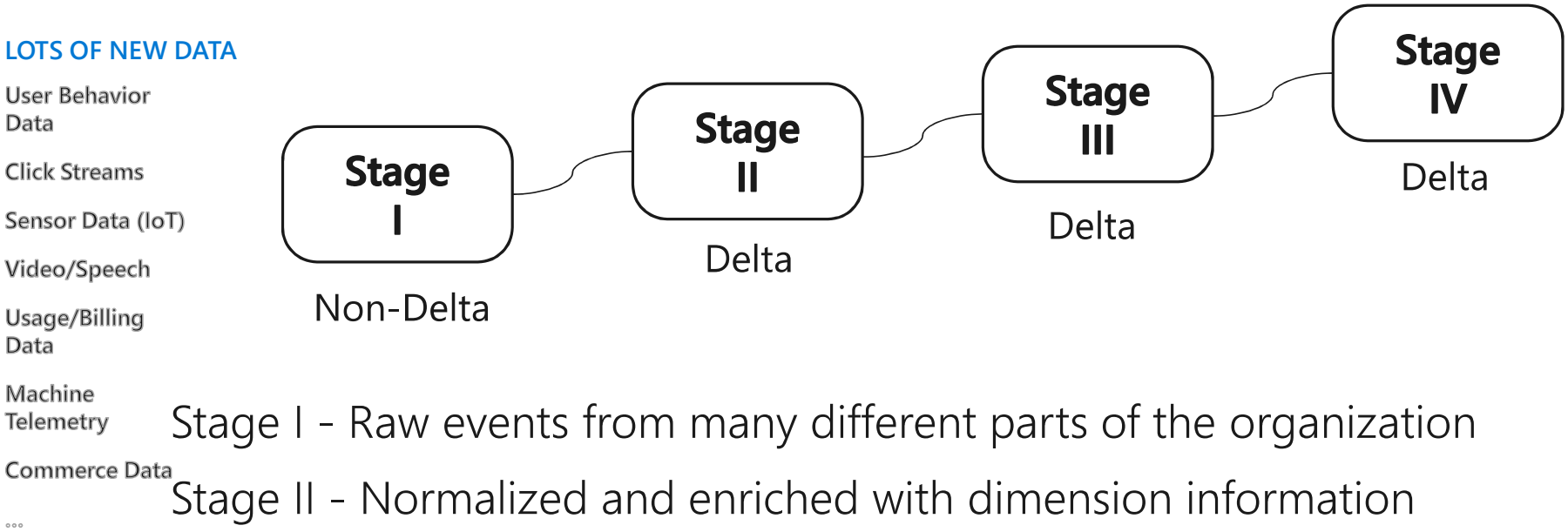
Evaluator

- Evaluate the models
- Evaluate how a given estimator performs according to criteria like ROC curve, RMSE
- Select best model based on the test

Pipeline

- Represents composition of
  - various Transformers' .transform methods
  - various Estimators' .fit and result's .transform methods
- Transformations and Estimators are specified together
- Pipeline allows to set up a dataflow of the relevant transformations, ending with an estimator that is automatically tuned according to your specifications resulting in a tuned model ready for a production use case

Multi Hop Data Pipelines



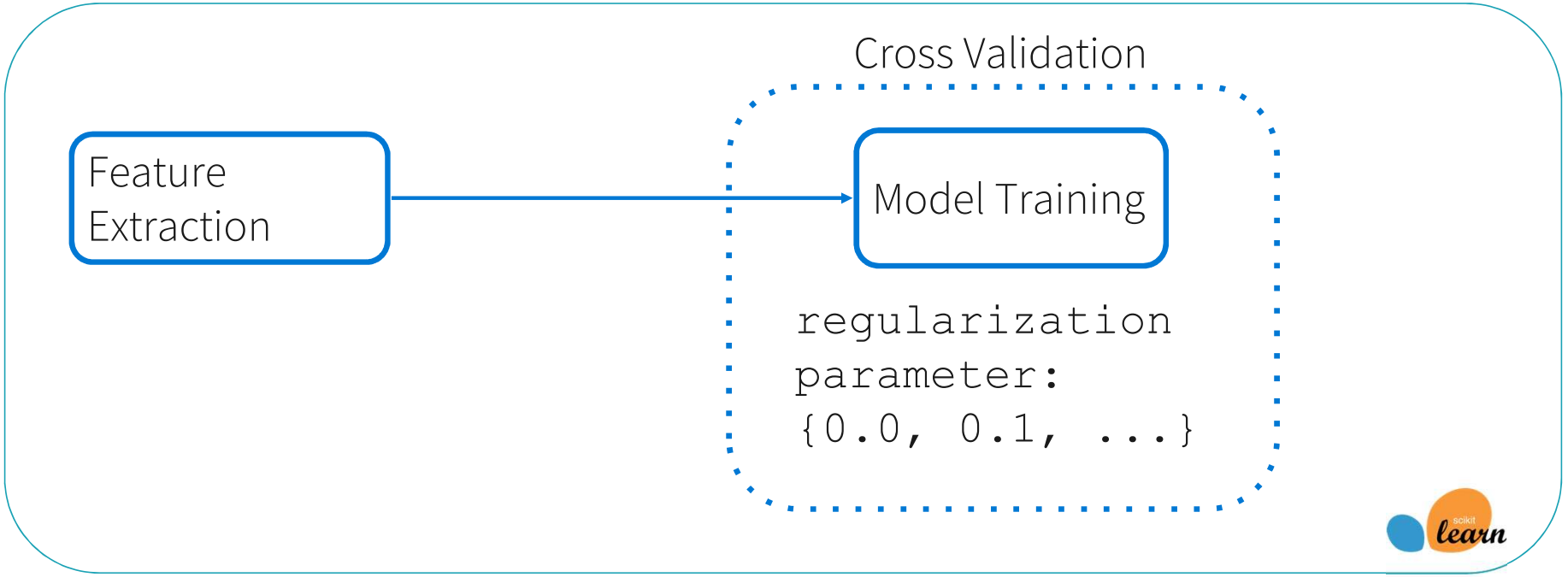
Stage I - Raw events from many different parts of the organization

Stage II - Normalized and enriched with dimension information

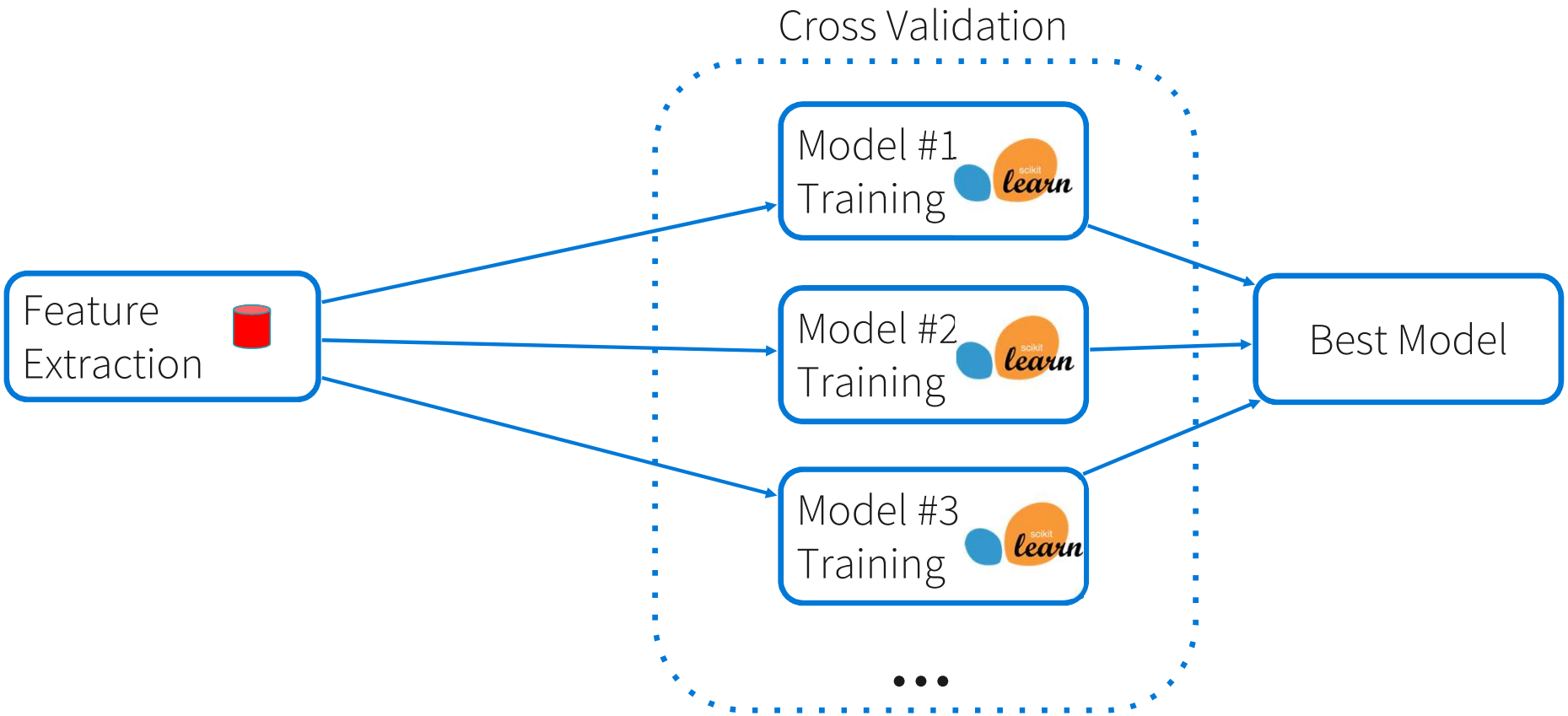
Stage III - Filtered down and aggregated for particular business objective.

Stage IV - High-level summaries of key business metrics.

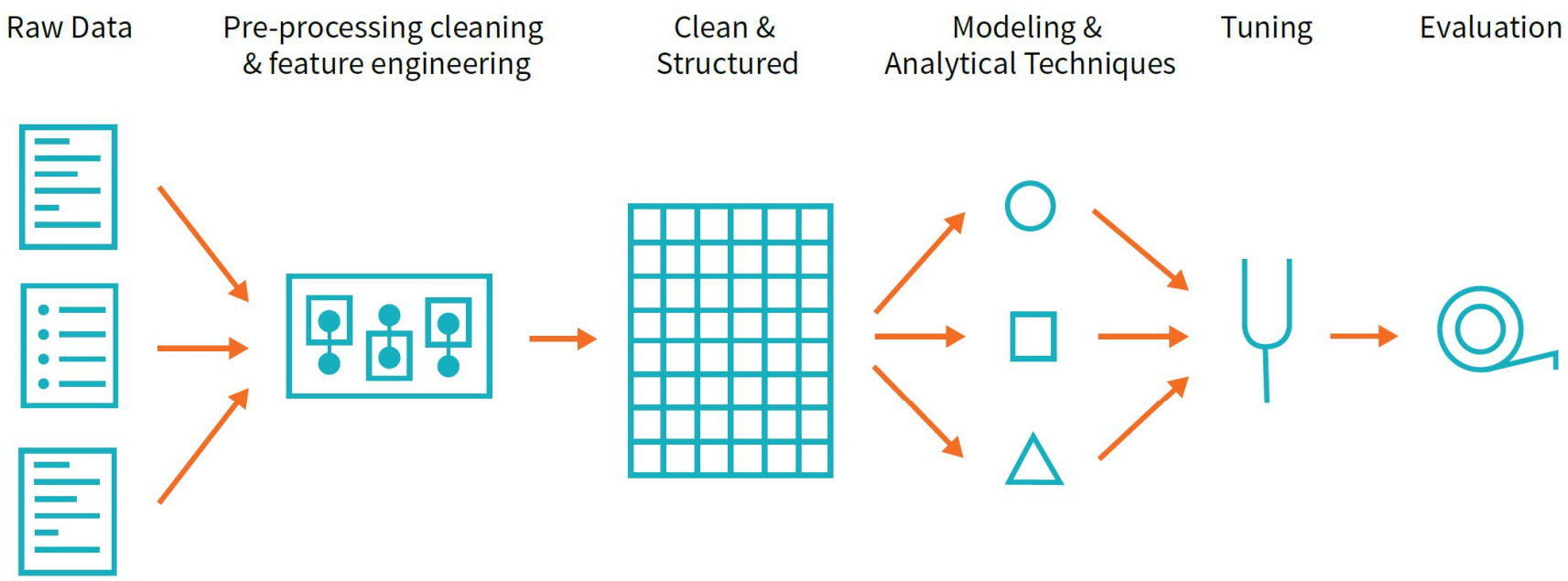
Cross Validation and Tuning



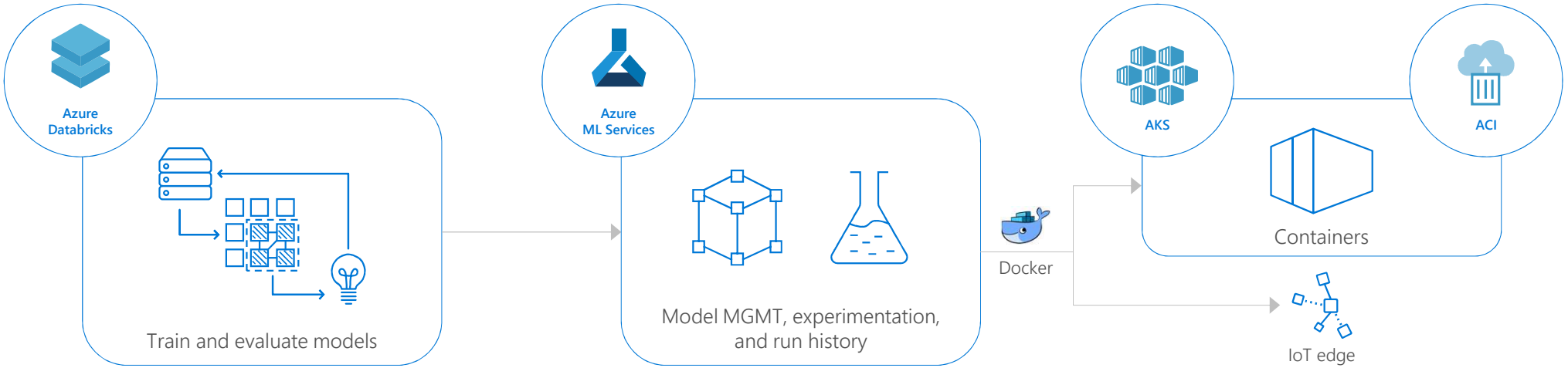
Model Selection



Advanced Analytics: Pipeline



Operationalize and manage models with ease



Bring models to life quickly

- Build and deploy models in minutes
- Iterate quickly on serverless infrastructure
- Easily change environments

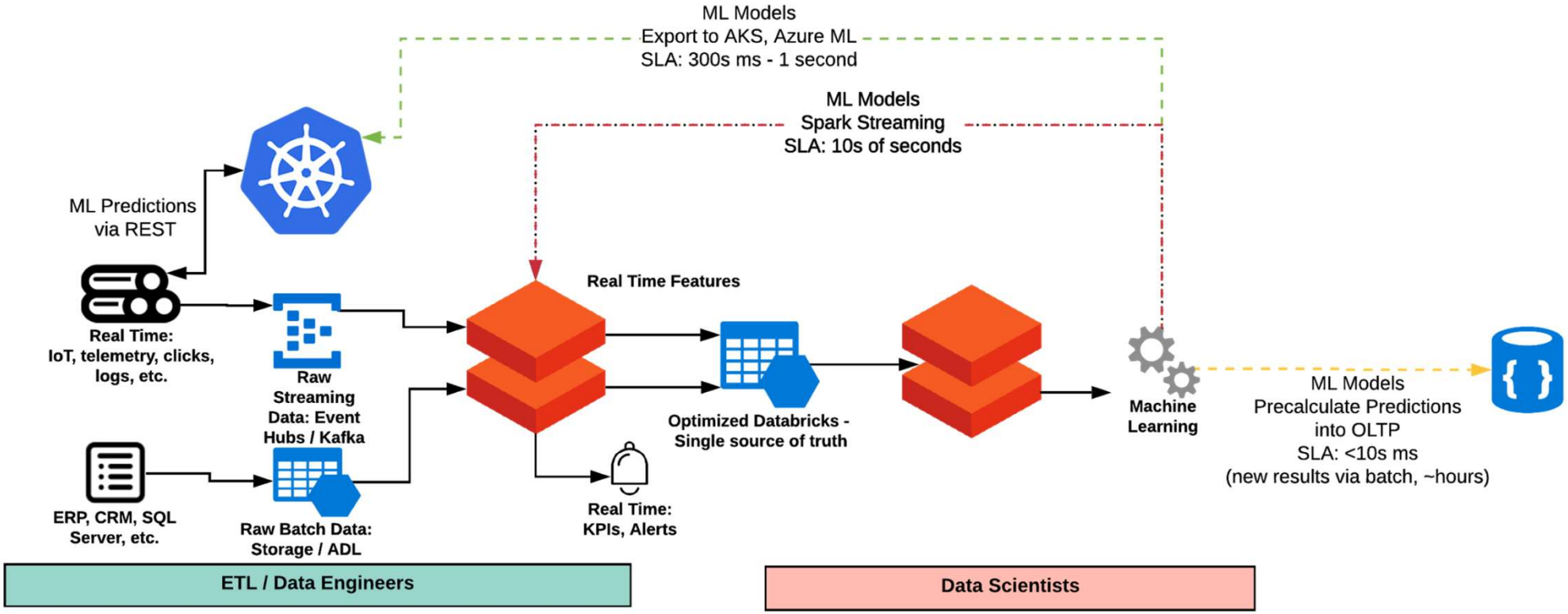
Proactively manage model performance

- Identify and promote your best models
- Capture model telemetry
- Retrain models with APIs

Deploy models closer to your data

- Deploy models anywhere
- Scale out to containers
- Infuse intelligence into the IoT edge

Operationalize Models



Azure Machine Learning service

Bring AI to everyone with an end-to-end, scalable, trusted platform

Boost your data science productivity

Increase your rate of experimentation

Deploy and manage your models everywhere

Built with your needs in mind

- GPU-enabled virtual machines
- Low-latency predictions at scale
- Integration with popular Python IDEs
- Role-based access controls
- Model versioning
- Automated model retraining

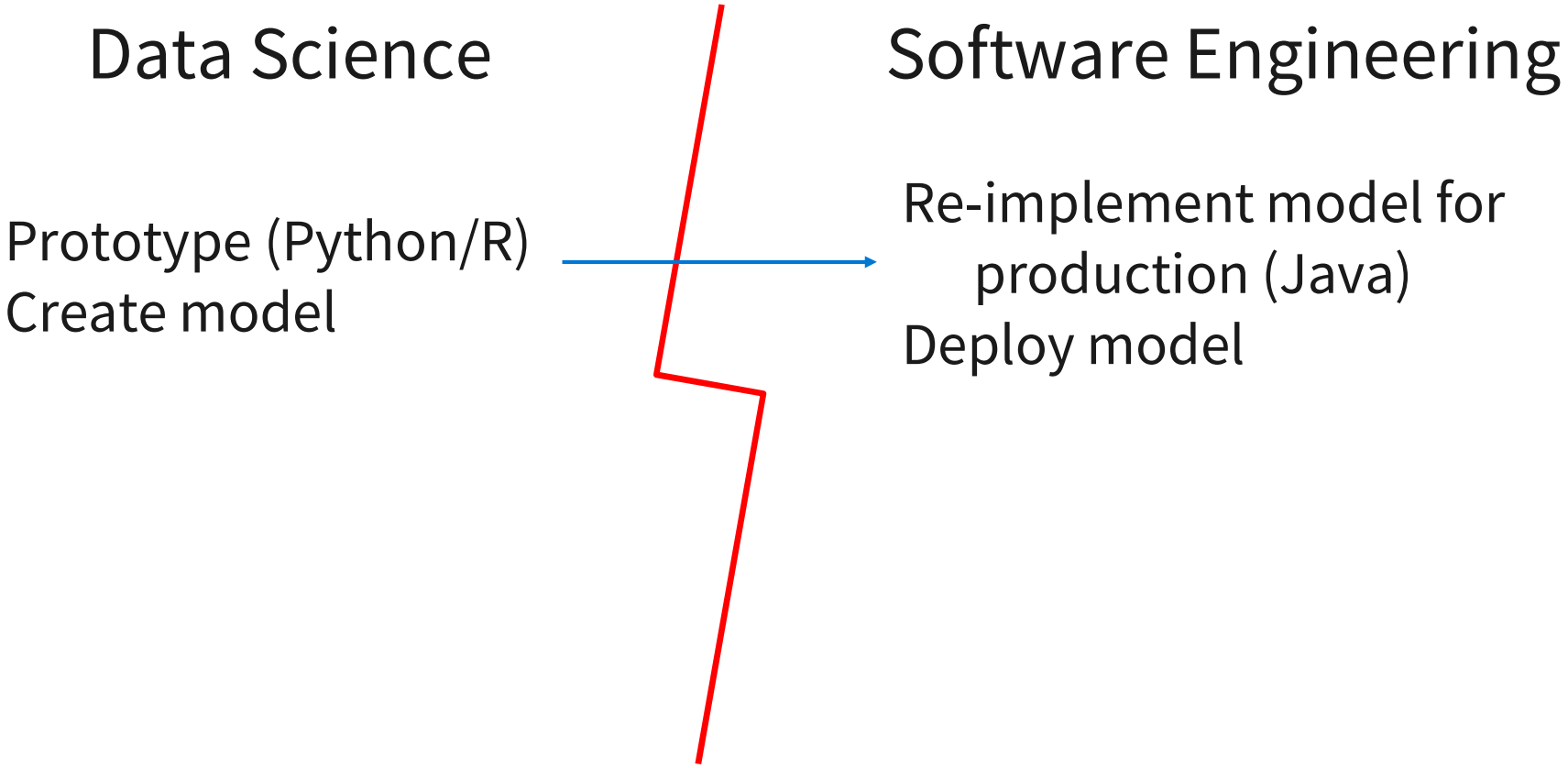
Seamlessly integrated with the Azure Portfolio

ML Export

- ML Model Export allows you to export models and full ML pipelines
  - Then imported into Spark and non-Spark platforms to do scoring, make predictions
  - Targeted at low-latency, lightweight ML-powered applications

- **We recommend using MLeap**, an open source solution for ML Model Export that works well in Azure Databricks

Why ML persistence?

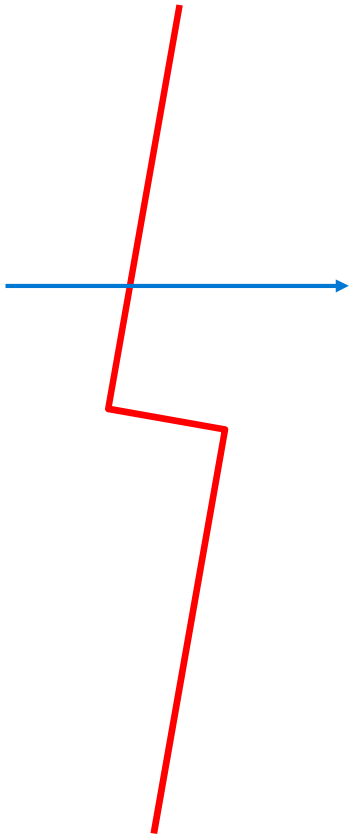


Why ML persistence?

Data Science

Prototype (Python/R)
Create Pipeline

- Extract raw features
- Transform features
- Select key features
- Fit multiple models
- Combine results to make prediction



Software Engineering

Re-implement Pipeline for production (Java)
Deploy Pipeline

- Extra implementation work
- Different code paths
- Synchronization overhead

With ML persistence...

Data Science

Prototype (Python/R)
Create Pipeline

Persist model or Pipeline:

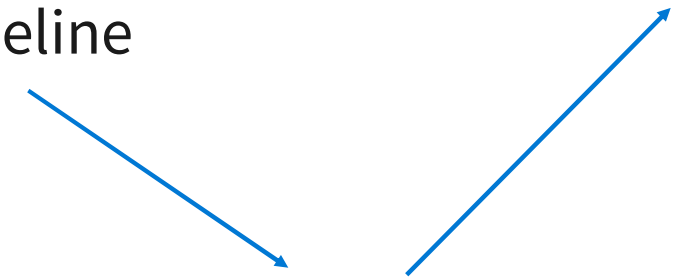
```
model.writeBundle.save()
```

Software Engineering

Load Pipeline (Scala/Java)

```
bundle.loadSparkBundle()
```

Deploy in production



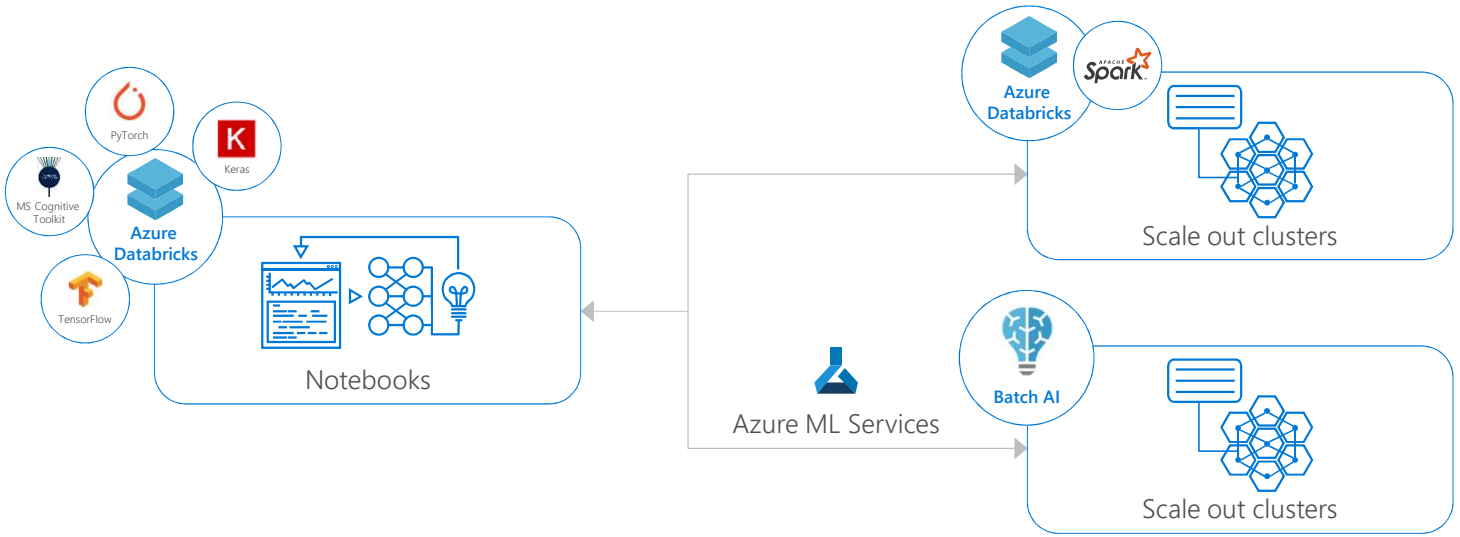
Workflow

1) Fit ML model in Databricks using Apache Spark MLlib.

2) Export model in Databricks
   - MLeap calls these “bundles”

3) Deploy model in external system
   - Load model from “bundle” at runtime
   - Make predictions in real time

Build and deploy deep learning models



Streamline AI development efforts

- Leverage popular deep learning toolkits
- Develop your language of choice

Scale compute resources in any environment

- Choose VMs for your modeling needs
- Process video using GPU-based VMs

Quickly evaluate and identify the right model

- Run experiments in parallel
- Provision resources automatically

Azure Databricks for deep learning modeling

Ready-to-use clusters with Azure Databricks Runtime for ML

Tools

Use HorovodEstimator via a native runtime to enable build deep learning models with a few lines of code

Load images natively in Spark DataFrames to automatically decode them for manipulation at scale with distributed DNN training on Spark

Simultaneously collaborate within notebooks environments to streamline model development

Frameworks

Full Python and Scala support for transfer learning on images

Seamlessly use TensorFlow, Microsoft Cognitive Toolkit, Caffe2, Keras, and more

Use built-in hyperparameter tuning via Spark MLLib to quickly optimize the model

Infrastructure

Leverage powerful GPU-enabled VMs pre-configured for deep neural network training

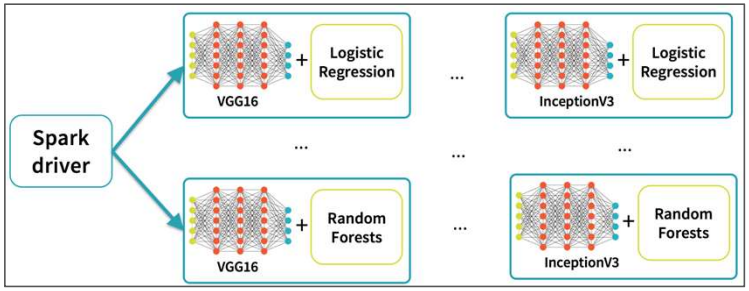
Automatically store metadata in Azure Database with geo-replication for fault tolerance

Improve performance 10x-100x over traditional Spark deployments with an optimized environment

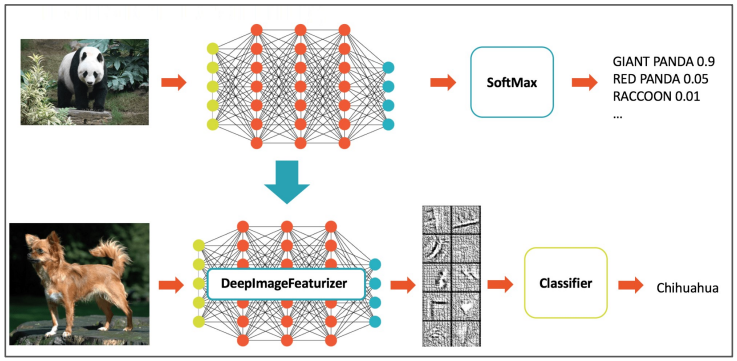
DEEP LEARNING

Azure Databricks supports and integrates with a number of Deep Learning libraries and frameworks to make it easy to build and deploy Deep Learning applications

- Supports Deep Learning Libraries/frameworks including:
  - Microsoft Cognitive Toolkit (CNTK).
    - Article explains how to install CNTK on Azure Databricks.
  - TensorFlowOnSpark
  - BigDL
- Offers Spark Deep Learning Pipelines, a suite of tools for working with and processing images using deep learning using transfer learning. It includes high-level APIs for common aspects of deep learning so they can be done efficiently in a few lines of code:
  - Image loading
  - Applying pre-trained models as transformers in a Spark ML pipeline
  - Transfer learning
  - Distributed hyperparameter tuning
  - Deploying models in DataFrames and SQL



Distributed Hyperparameter Tuning



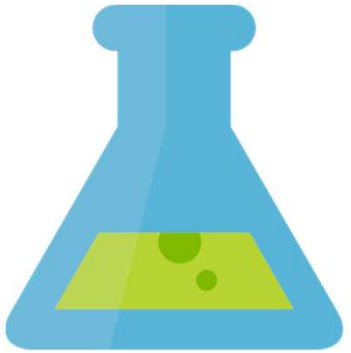
Transfer Learning

MMLSPARK

Microsoft Machine Learning Library for Apache Spark (MMLSpark) lets you easily create scalable machine learning models for large datasets.

It includes integration of SparkML pipelines with the Microsoft Cognitive Toolkit and OpenCV, enabling you to:

- Ingress and pre-process image data
- Featurize images and text using pre-trained deep learning models
- Train and score classification and regression models using implicit featurization



Leverage deep learning services and frameworks

AZURE DATABRICKS

Accelerate processing with the fastest Spark engine

Integrate natively with Azure services

Access enterprise-grade Azure security

AZURE ML service

Bring AI to the edge

Increase your rate of experimentation

Deploy and manage your models everywhere

Leverage your favorite deep learning frameworks

TensorFlow | MS Cognitive Toolkit | PyTorch | Scikit-Learn | ONNX | Caffe2 | MXNet | Chainer

Powerful infrastructure

Accelerate deep learning

CPUs

General purpose machine learning

D, F, L, M, H Series

GPUs

Deep learning

N Series

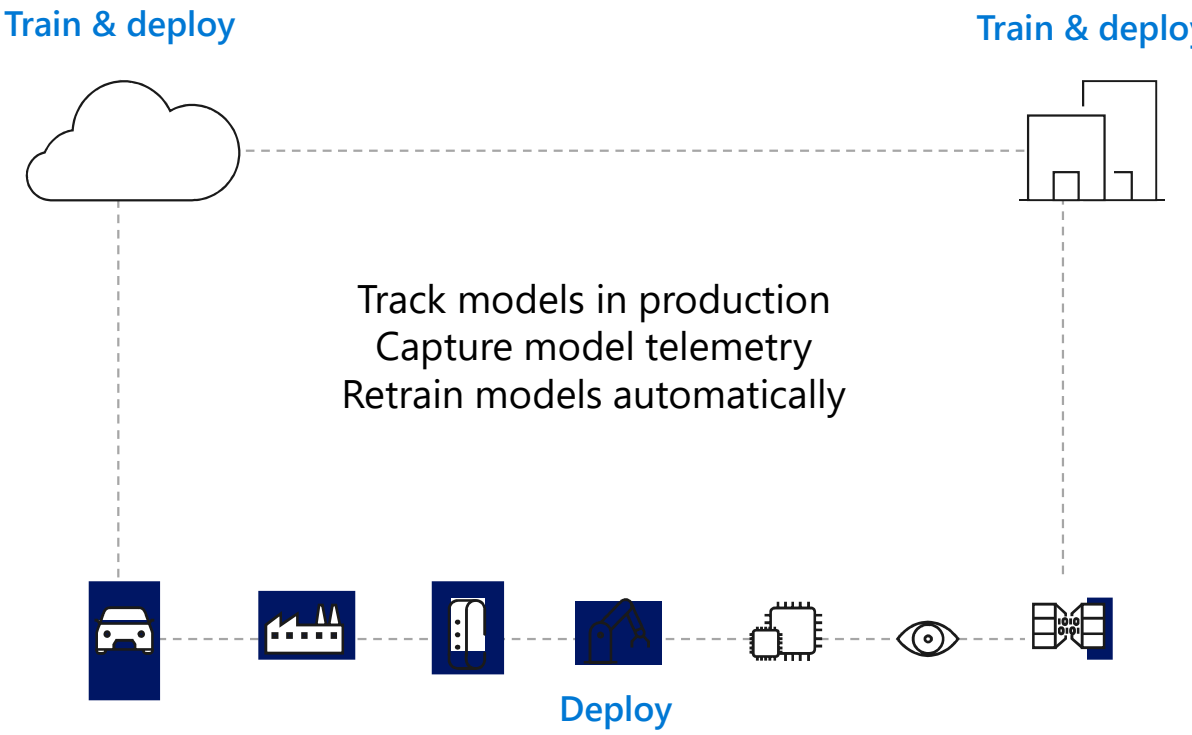
FPGAs

Specialized hardware accelerated deep learning

Project Brainwave

Flexible deployment

Deploy and manage models on intelligent cloud and edge



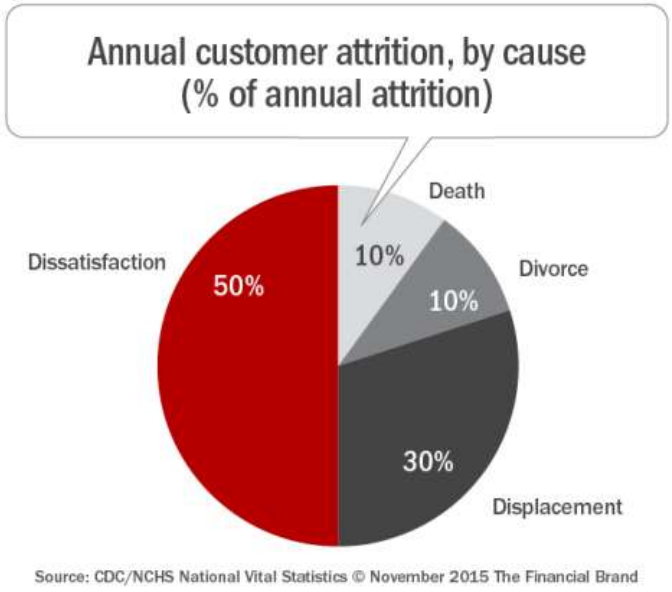
paypal-churn-sparkml.pdf

Use Case – Churn Analytics

Customer Churn: A Key Performance Indicator for Banks

- Customer churn and engagement is one of the top issues for most banks
- Costs significantly more to acquire new customers than retain existing ones
- and it costs far more to re-acquire deflected customers.

The key issue: knowing the customer and predicting churn

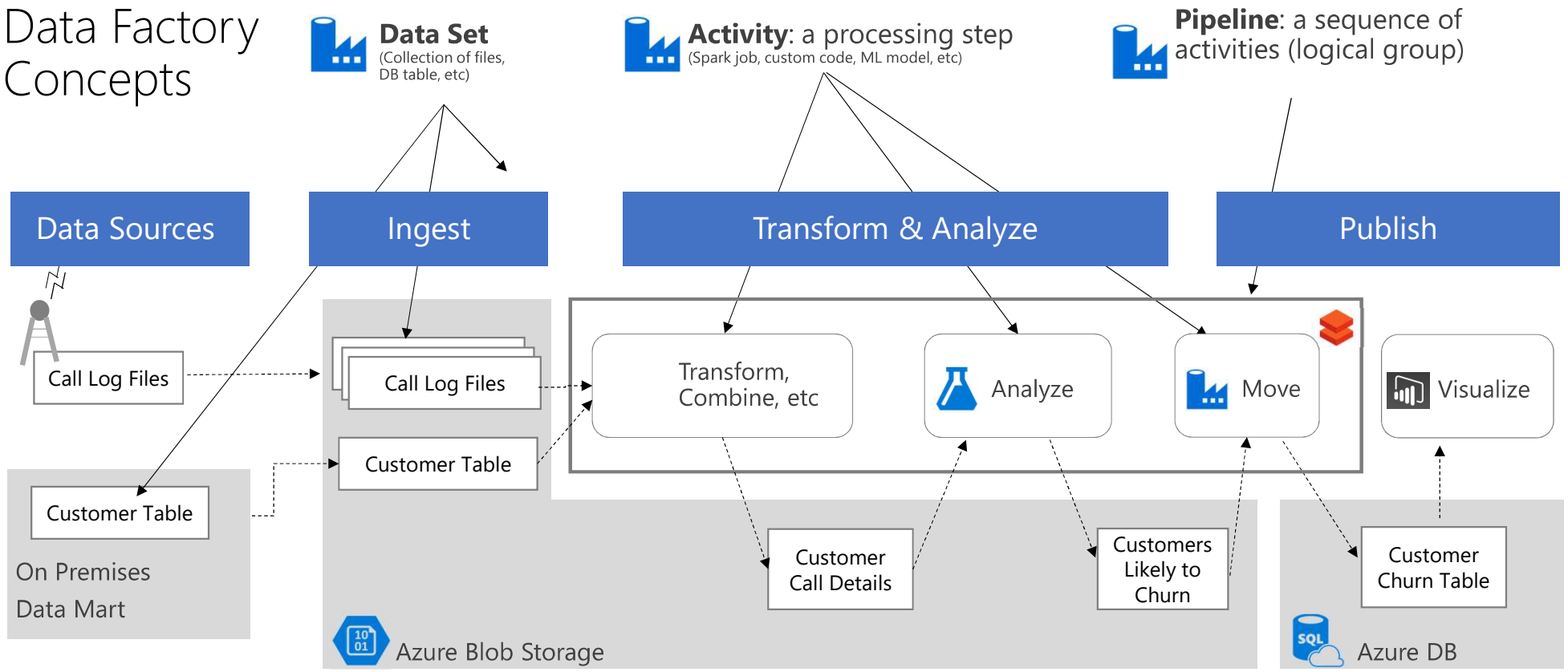


Building a churn prediction model

What to look for – Opportunities?

- Death – One way to reduce the impact of death on your portfolio is to influence the age distribution of your customer base
- Divorce – One way to stem attrition from divorce is to provide education on account ownership options and to ensure staff members are properly trained
- Displacement – Strong mobile and online banking capabilities will help you keep customers who are comfortable using mobile or online banking even if they move out of your market area
- Dissatisfaction – Sentiment and QoS Analysis

Conceptual Architecture



Resources


Democratization-of-AI-and-Deep-Learning.pdf


Data-Scientists-Guide-to-Apache-Spark.pdf


The-Data-Engineers-Guide-to-Apache-Spark.pdf


FY18_AA_Databricks_e-book_FINAL_032018.pdf


A-Gentle-Introduction-to-Apache-Spark.pdf


Three practical use cases with Azure Databrick - Aug 2018.pdf