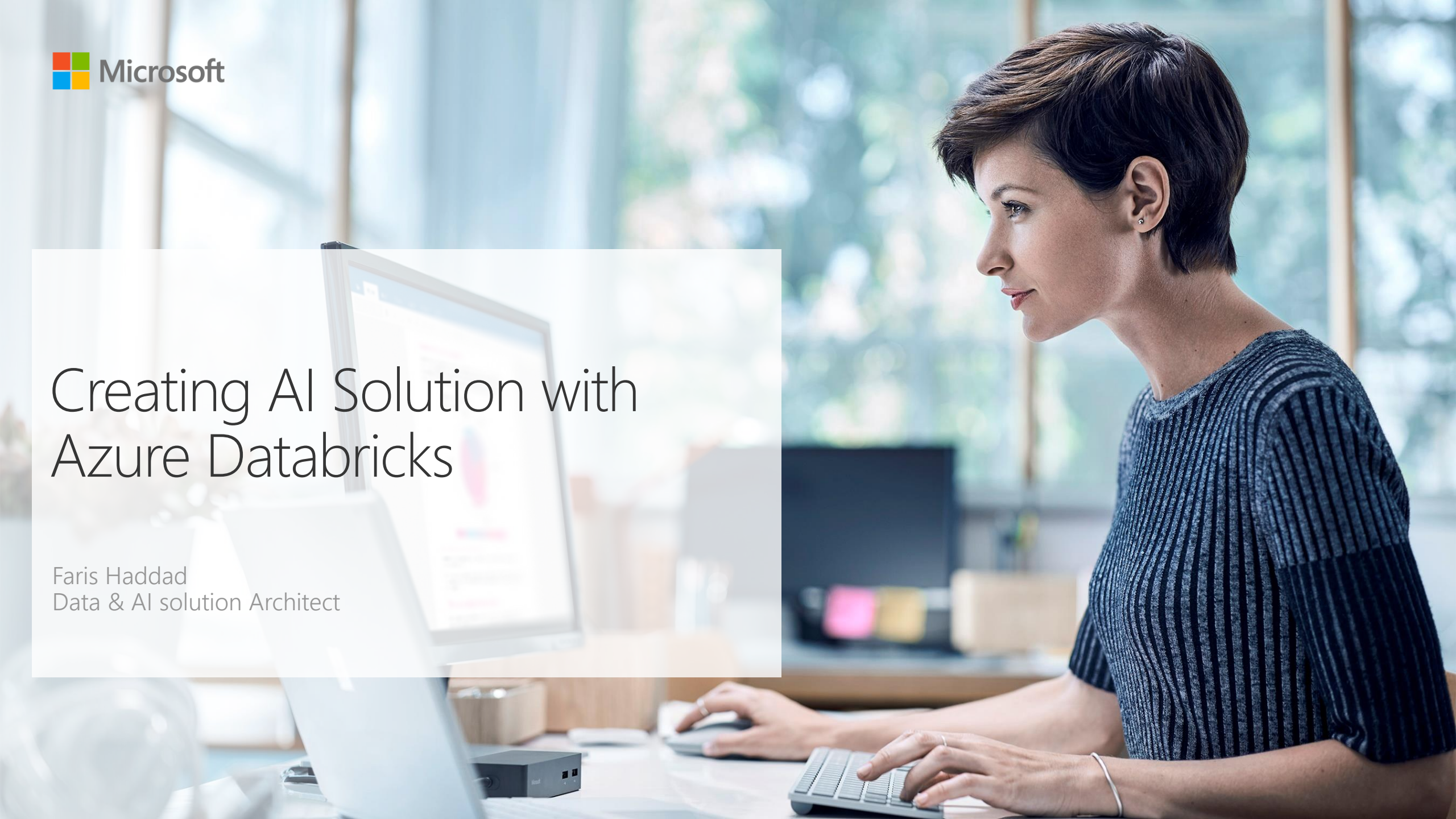




# Creating AI Solution with Azure Databricks

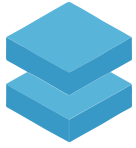
Faris Haddad  
Data & AI solution Architect



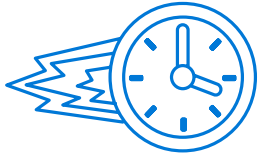
# Azure Databricks

*Fast, Easy, Collaborative – Powered by Apache Spark*

# What is Azure Databricks?



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



**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

Seamlessly integrated with the Azure Portfolio

# Differentiated experience on Azure

## ENHANCE PRODUCTIVITY

**Get started quickly** by launching your new Spark environment with one click from **Azure portal**.

**Share your insights in powerful ways** through rich integration with Power BI.

**Improve collaboration** amongst your analytics team through a unified workspace.

**Innovate faster** with native integration with rest of Azure platform like Azure SQL Data Warehouse, Azure Cosmos DB, Azure IoT hub, Azure Data Factory and more.

## BUILD ON THE MOST COMPLIANT CLOUD

**Simplify security and identity control** with built-in integration with Active Directory.

**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, compliance and SLAs.

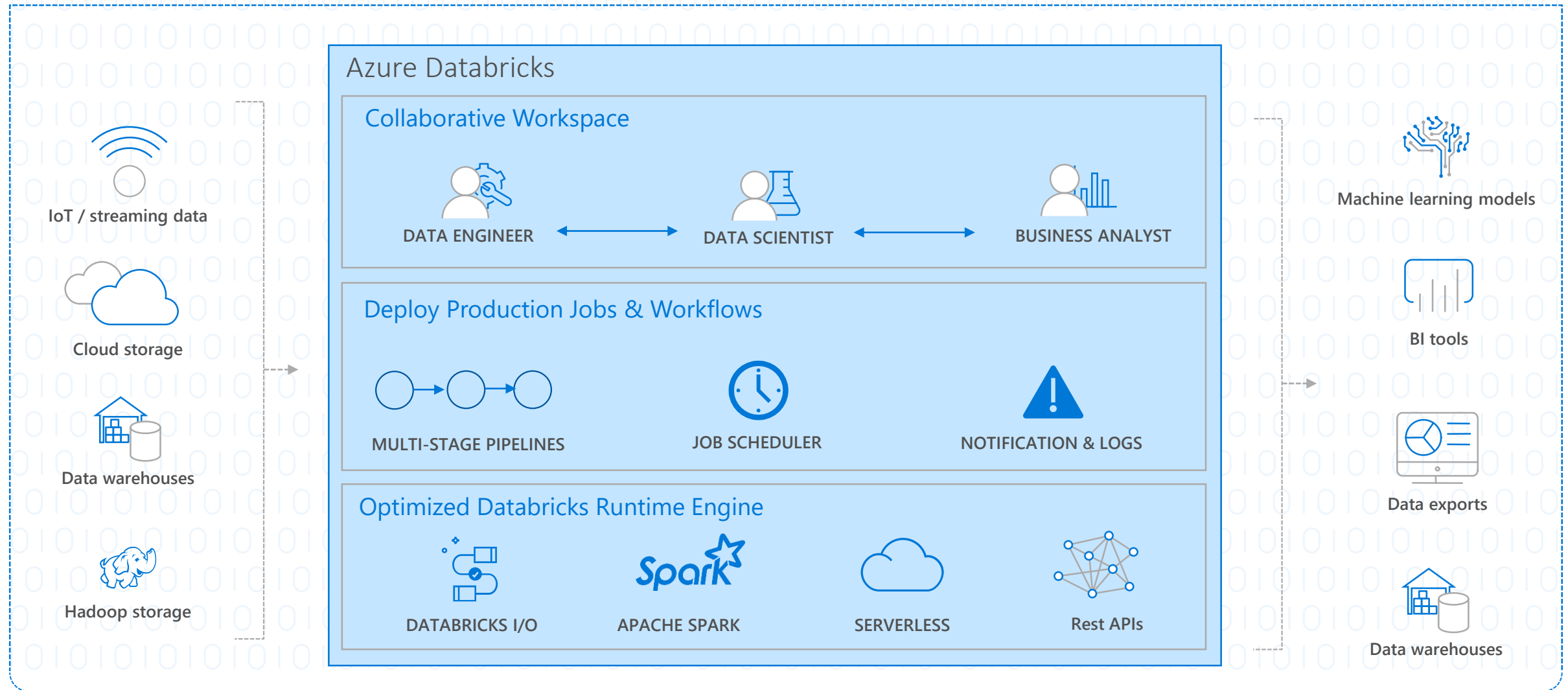
Get support and a **single bill from Azure** billing without the need to create a separate account with Databricks.

## SCALE WITHOUT LIMITS

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

**Accelerate data processing** with the fastest Spark engine.

# A Z U R E   D A T A B R I C K S



Enhance Productivity

Build on secure & trusted cloud

Scale without limits



# Why Spark?



- Open-source data processing engine built around **speed, ease of use, and sophisticated analytics**
- In memory engine that is up to **100 times faster than Hadoop**
- **Largest open-source data project** with 1000+ contributors
- **Highly extensible** with support for Scala, Java and Python alongside Spark SQL, GraphX, Streaming and Machine Learning Library (MLlib)

## Spark Unifies:

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

# Product Target

# Azure Databricks key audiences & benefits



## Data scientist

Integrated workspace

Easy data exploration

Collaborative experience

Interactive dashboards

Faster insights

- Best Spark & serverless
- Databricks managed Spark



## Data engineer

Improved ETL performance

- Zero management clusters, serverless

Easy to schedule jobs

Automated workflows

Enhanced monitoring & troubleshooting

- Automated alerts & easy access to logs

Zero Management Spark

Cluster democratization (serverless)



## Technology Decision Makers (CDO, CTO, VP of Analytics)

Fast, collaborative analytics platform accelerating time to market

No dev-ops required

Enterprise grade security

- Encryption
- End-to-end auditing
- Role-based control
- Compliance



# Data Scientists & Data Engineers

## Key Personas



### Data Scientist

Responsible for analyzing data to uncover patterns and make future predictions

#### PAIN POINTS/CONCERNS

- Often spends too much time on accessing/ingesting data. Exploration at scale is difficult

#### Azure Databricks Opportunity

- Get to tool in their hands ASAP, it increases their productivity
- Azure + Spark + Databricks = great resume builder
- Can be your best champion
- Be careful of devs & data engineers “rebranding” as data scientists
- Trouble accessing budget, focus on finding value



### Data Engineer

Responsible for turning raw data into usable form for non-technical end-users through ETL/cleansing

#### PAIN POINTS/CONCERNS

- Difficult to do fast and reliably enough to support the business when dealing with scale, and variety of data sources and types. Painful to access and ETL data.

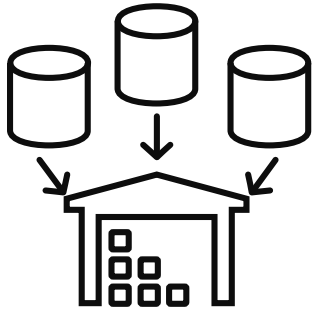
#### Azure Databricks Opportunity

- Easier and faster data access and ETL, cost effective and zero-maintenance infrastructure
- Very careful about production grade deployments
- They want programmable control of the platform
- Focus on APIs, performance and reliability
- Can be very cheap, focus on finding value.

# 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

# Our customers have three common objectives



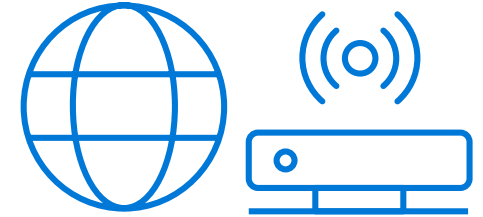
*"We want to extend to  
untapped sources"*

Modern Data Warehouse



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

Advanced Analytics



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

Real-time Analytics

# Scenarios

- E-mails – Classify e-mails as spam or not spam (Classification)
- Customer churn analysis (Classification)
- Predict sales using historical sales data (Regression)
- Movie recommendation
- Anomaly detection (Unsupervised Learning)

# 3 Ways for Machine Learning

## #1 Scalable Machine Learning with Spark MLlib - [example notebook](#)

- 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 - [example notebook \(see Part 1\)](#)

- 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

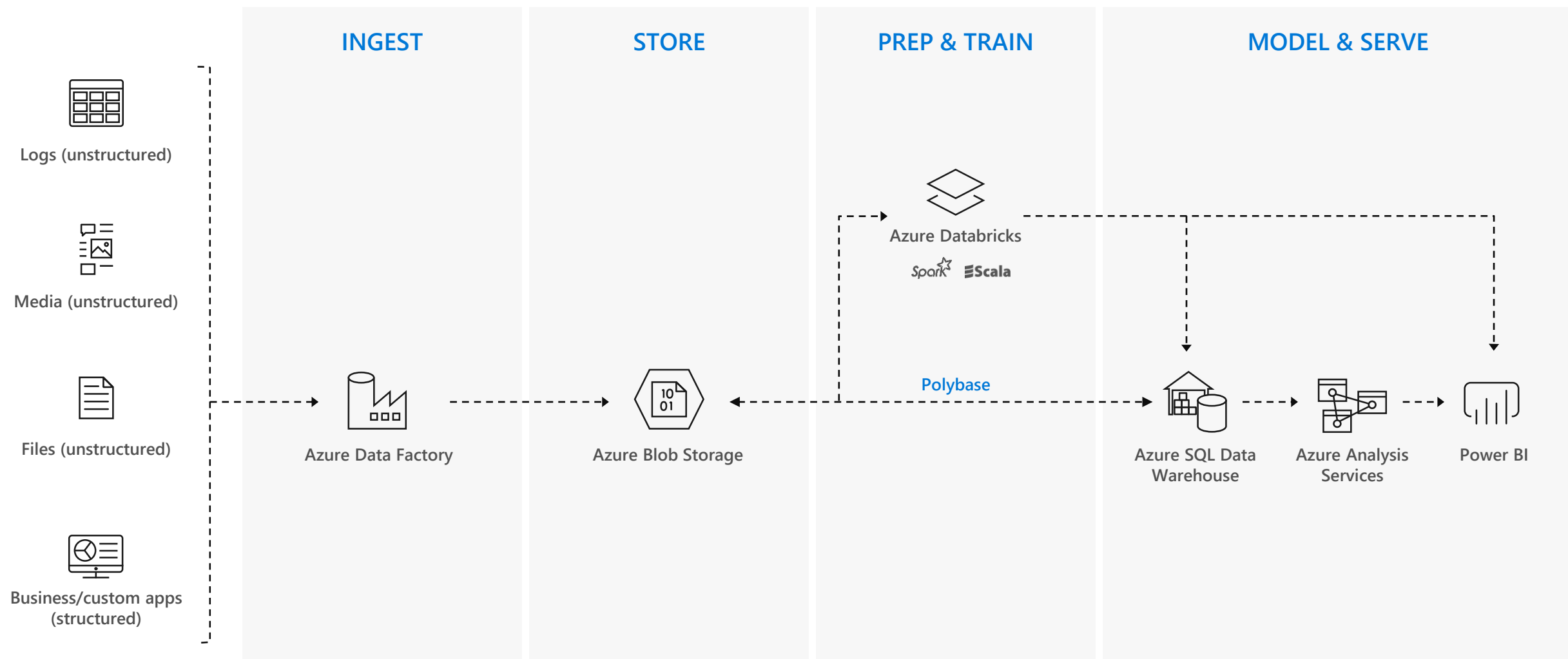
## #3 Scale Out / Parallelization for Single Machine Data Science - [example notebook \(see Part 2\)](#)

- 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

# Use Cases

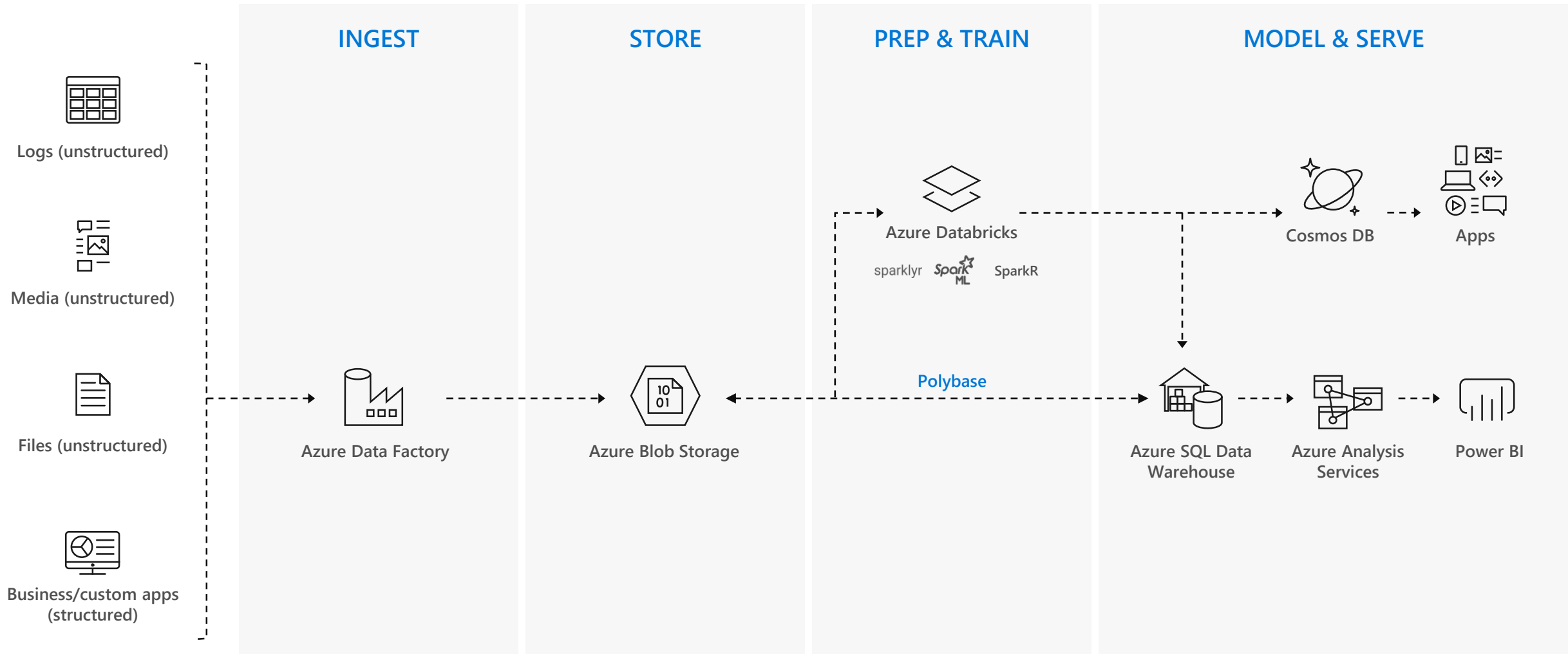


# Modern Data Warehouse



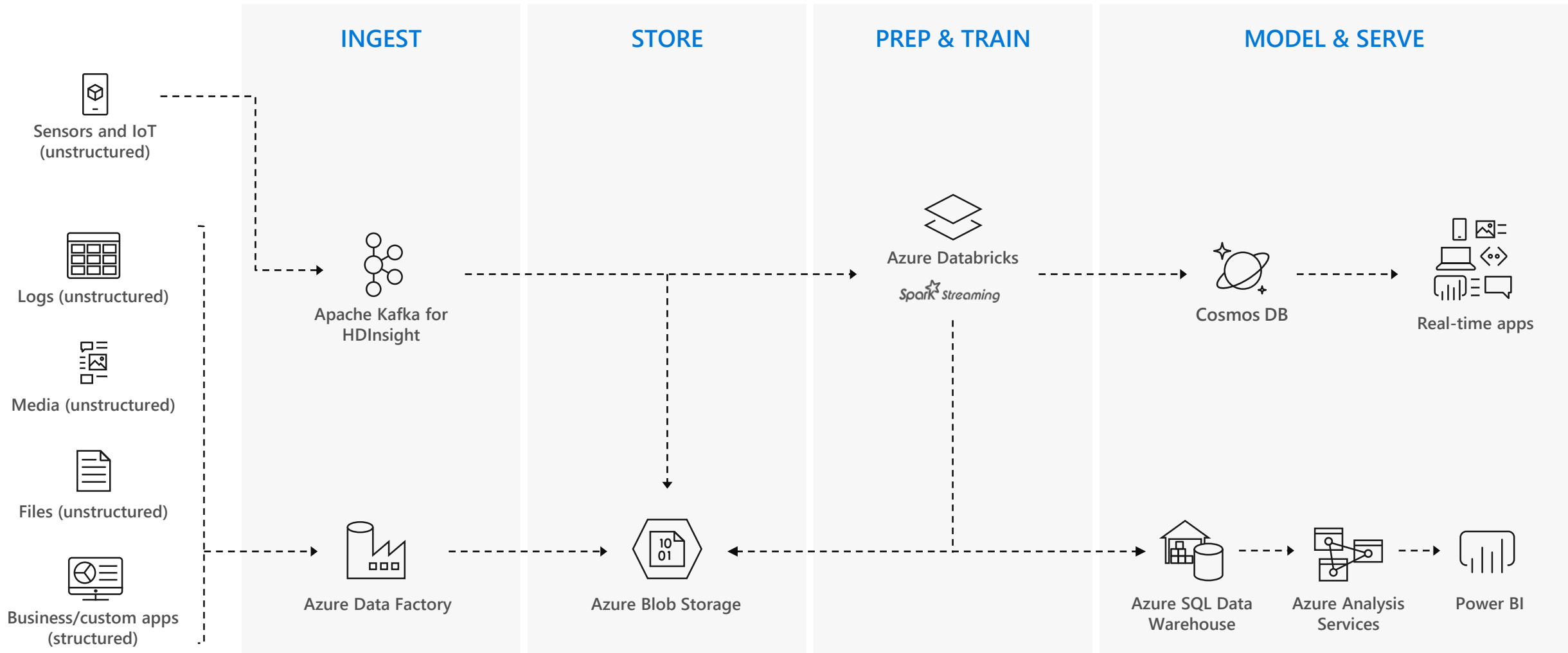
Microsoft Azure also supports other Big Data services like Azure HDInsight and Azure Data Lake to allow customers to tailor the above architecture to meet their unique needs.

# Advanced Analytics on Big data



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

# Real time analytics



Microsoft Azure also supports other Big Data services like Azure IoT Hub, Azure Event Hubs, Azure Machine Learning and Azure Data Lake to allow customers to tailor the above architecture to meet their unique needs.

# Renewables.AI – Demand forecasting maximizes revenue

## Problem

Renewable energy sources such as solar power are challenging to integrate with global energy markets, as they are generated inconsistently from a variety of sources

## Solution

Leverage AI to forecast energy production, integrate with power transmission grids and align energy supply with the optimal energy markets

## Results



Streamline product development



Increase revenue and savings



Drive adoption of renewable energy

# FINANCIAL SERVICES

## Use cases

### Effective customer engagement

Customer profiles  
Credit history  
Transactional data  
LTV  
Loyalty



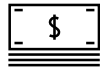
### Customer analytics

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






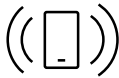

### Marketing analytics

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

**Improved customer  
engagement with  
machine learning**

# HEALTH & LIFE SCIENCES






Use cases

<div>DNA sequences</div> <div>FAST-Q BAM SAM VCF Expression</div> <div></div> <div>Genomics and precision medicine</div> <div>Single cell sequencing Biomarker, genetic, variant and population analytics ADAM and HAIL on Databricks</div> <div>Faster innovation for drug development</div>	<div>Real world analytics</div> <div>HL7/CCD 837 Pharmacy Registry EMR</div> <div></div> <div>Clinical and claims data</div> <div>Claims data warehouse Readmission predictions Efficacy and comparative analytics Prescription adherence Market access analysis</div> <div>Improved outcomes and increased revenue</div>	<div>Image deep learning</div> <div>MRI X-RAY CT Ultrasound</div> <div></div> <div>GPU image processing</div> <div>Graphic intensive workloads Deep learning using Tensor Flow Pattern recognition</div> <div>Diagnostics leveraging machine learning</div>	<div>Sensor data</div> <div>Readings Time series Event data</div> <div></div> <div>IoT device analytics</div> <div>Aggregation of streaming events Predictive maintenance Anomaly detection</div> <div>Predictive analytics transforms quality of care</div>	<div>Social data listening</div> <div>Social media Adverse events Unstructured</div> <div></div> <div>Social analytics</div> <div>Real-time patient feedback via topic modelling Analytics across publication data</div> <div>Improved patient communications and feedback</div>
--	--	--	---	---








# MEDIA & ENTERTAINMENT

Use cases

<div>Personalized recommendations</div> <div>Customer profiles Viewing history Online activity Content sources Channels</div> <div></div> <div>Content personalization</div> <div>Personalized viewing and engagement experience Click-path optimization Next best content analysis Improved real time ad targeting</div> <div>Faster innovation for customer experience</div>	<div>Effective customer retention</div> <div>Customer profiles Online activity Content distribution Services data</div> <div></div> <div>Customer churn prevention</div> <div>Quality of service and operational efficiency Market basket analysis Customer behavior analysis Click-through analysis</div> <div>Improved consumer outcomes and increased revenue</div>	<div>Information optimization</div> <div>Consumption logs Clickstream and devices Marketing campaign responses</div> <div></div> <div>Recommendation engine</div> <div>Ad effectiveness Content monetization Fraud detection Information-as-a-service High value user engagement</div> <div>Enhance user experience with machine learning</div>	<div>Inventory allocation</div> <div>Transactions Subscriptions Demographics Credit data</div> <div></div> <div>Predictive analytics</div> <div>Predict audience interests Network performance and optimization Pricing predictions Nielsen ratings and projections Mobile spatial analytics</div> <div>Predictive analytics transforms growth</div>	<div>Consumer engagement analysis</div> <div>Content metadata Ratings Comments Social media activity</div> <div></div> <div>Sentiment analysis</div> <div>Demand-elasticity Social network analysis Promotion events time-series analysis Multi-channel marketing attribution</div> <div>Improved consumer engagement with machine learning</div>
---	---	--	---	--

# RETAIL

## Use cases

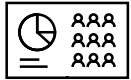
Recommendation engine	Effective customer engagement	Inventory optimization	Inventory allocation	Consumer engagement
Customer profiles Shopping history Online activity Social network analysis	Shopping history Online activity Floor plans App data	Demand plans Forecasts Sales history Trends Local events/weather patterns	Demographics Buyer perception Consumer research Market/competitive analysis	Historical sales data Price scheduling Segment level price changes
				
<b>Next best and personalized offers</b>	<b>Store design and ergonomics</b>	<b>Data-driven stock, inventory, ordering</b>	<b>Assortment optimization</b>	<b>Real-time pricing optimization</b>
Customer 360/consumer personalization Right product, promotion, at right time Multi-channel promotion	Path to purchase In-store experience Workforce and manpower optimization	Predict inventory positions and distribution Fraud detection Market basket analysis	Economic modelling Optimization for foot traffic, Online interactions Flat and declining categories	Demand-elasticity Personal pricing schemes Promotion events Multi-channel engagement
<b>Faster innovation for customer experience</b>	<b>Improved consumer outcomes and increased revenue</b>	<b>Omni-channel shopping experience with machine learning</b>	<b>Predictive analytics transforms growth</b>	<b>Improved consumer engagement with machine learning</b>

# 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

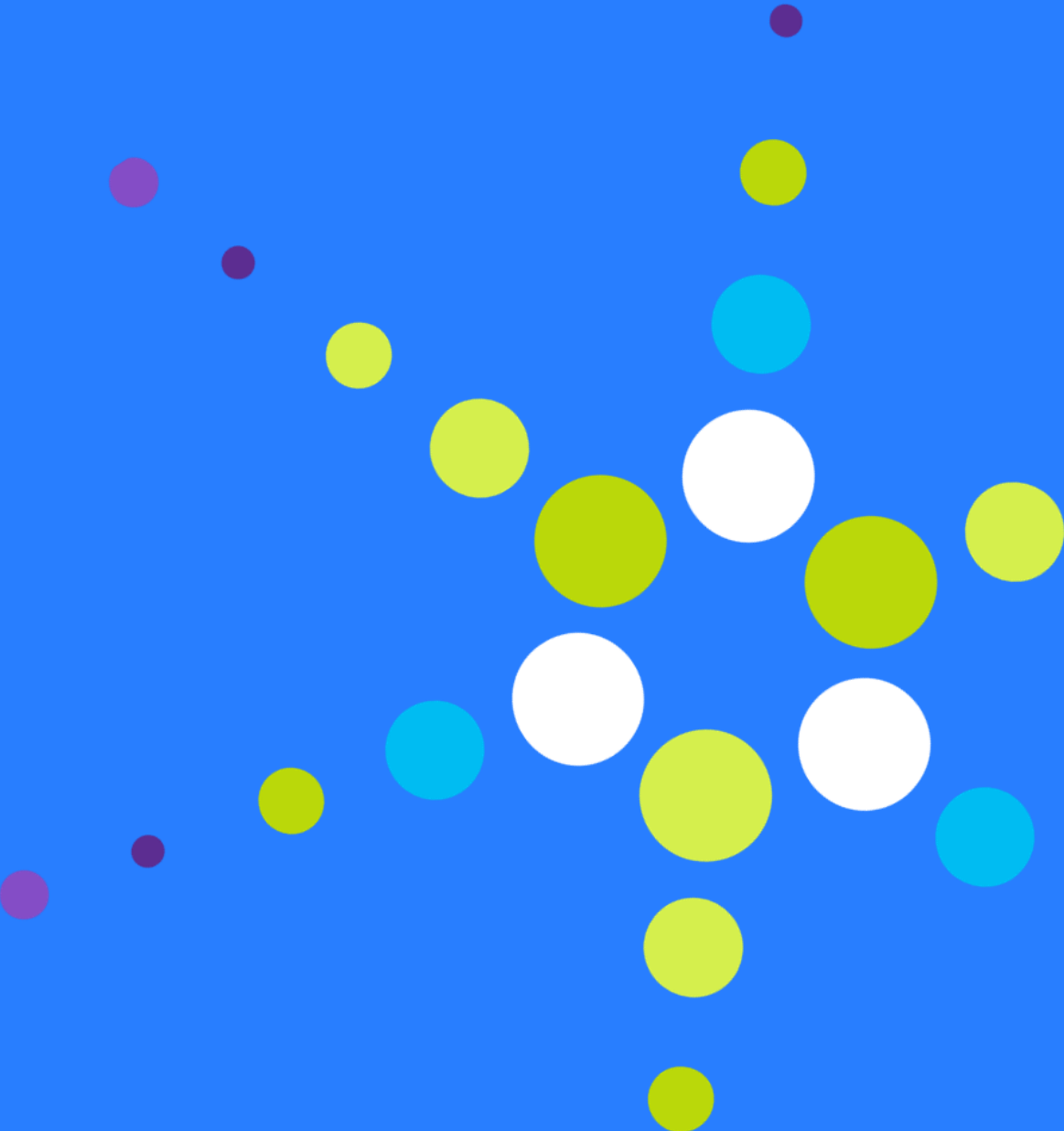
<div>Upstream optimization, maximize well life</div> <div>Field data Asset data Demographics Production data</div> <div></div> <div>Digital oil field/ oil production</div> <div>Production optimization Integrate exploration and seismic data Minimize lease operating expenses Decline curve analysis</div> <div>Faster innovation for revenue growth</div>	<div>Grid operations, asset inventory optimization</div> <div>Sensor stream data UAVs images Inventory data Production data</div> <div></div> <div>Industrial IoT</div> <div>Pipeline monitoring Preventive maintenance Smart grids and microgrids Grid operations, Field Service Asset performance as a Service</div> <div>Improved outcomes and increased revenue</div>	<div>Supply-chain optimization</div> <div>Transaction data Demographics Purchasing history Trends</div> <div></div> <div>Supply-chain optimization</div> <div>Trade monitoring, optimization Retail mobile applications Vendor management - construction, transportation, truck &amp; delivery optimization</div> <div>Optimizing supply- chain with machine learning</div>	<div>Risk optimization</div> <div>Sensor stream data Transport Retail data Grid production data Refinery tuning parameters</div> <div></div> <div>Safety and security</div> <div>Real-time anomaly detection Predictive analytics Industrial safety Environment health and safety</div> <div>Predictive analytics transforms safety and security</div>	<div>Recommendations engine</div> <div>Clickstream data Products Services Market data Competitive data Demographics</div> <div></div> <div>Sales and marketing analytics</div> <div>Fast marketing and multi-channel engagement Develop new products and monitor acceptance of rates Predictive energy trading Deep customer insights</div> <div>Improved customer engagement with machine learning</div>
---	--	--	---	--

# SECURITY

## Use cases

Security controls to leverage all data	Actionable threat intelligence	Risk and fraud analysis	Compliance management	Identity and access management for analytics
Firewall/network logs Apps Data access layers	Firewall/network logs Network flows Authentications	Firewall/network logs Web/app logs Social media content	Firewall/network logs Web Applications Devices OS	Files Tables Clusters Reports Dashboards Notebooks
				
<b>Intrusion detection and predictive analytics</b>	<b>Security intelligence</b>	<b>Fraud detection and prevention</b>	<b>Security compliance reporting</b>	<b>Fine-grained data analytics security</b>
Prevention of DDoS attacks Threat classifications Data loss/anomaly detection in streaming Cybermetrics and changing use patterns	Real-time data correlation Anomaly detection Security context, enrichment Offence scoring, prioritization Security orchestration	e-Tailing Inventory monitoring Social media monitoring Phishing scams Piracy protection	Ad-hoc/historic incident reports SOC/NOC dashboards Deep OS auditing Data loss detection in IoT User behavior analytics	Role-based access controls Auditing and governance File integrity monitoring Row level and column level access permissions
<b>Prevent complex threats with machine learning</b>	<b>Faster innovation for threat prevention</b>	<b>Risk management with machine learning</b>	<b>Transform security with improved visibility</b>	<b>Limit malicious insiders to transform growth</b>

Demo





# Azure Databricks – service home page

The screenshot displays the Microsoft Azure portal interface. The top navigation bar shows the path: Microsoft Azure > New > Marketplace > Data + Analytics > Azure Databricks (preview). A search bar is present on the right, and the user's email (nishanth@microsoft.c...) is visible in the top right corner.

The left sidebar contains a list of services: New, Dashboard, All resources, Resource groups, App Services, SQL databases, SQL data warehouses, Azure Cosmos DB, Virtual machines, Load balancers, Storage accounts, Virtual networks, Azure Active Directory, Monitor, Security Center, Cost Management + Billing, Help + support, Subscriptions, and Data Lake Store. A "More services" link is at the bottom.

The main content area is titled "Azure Databricks (preview)" and features the following text:

- DATABRICKS IS A TRULY UNIFIED APPROACH TO DATA ANALYTICS AT SCALE**
- Founded by the team who created Apache Spark, Databricks provides a Unified Analytics Platform that accelerates innovation by unifying data science, engineering, and business.
- UNIFIED EXPERIENCE ACROSS TEAMS**
- A collaborative workspace for data science teams to work with data engineering and lines of business.
- UNIFIED ANALYTICS WORKFLOWS**
- One environment from data preparation to exploration and model building to production.
- UNIFIED INFRASTRUCTURE**
- Fully managed, serverless cloud infrastructure for isolation, automation, and cost control.

Below the text are social media icons for Twitter, Facebook, LinkedIn, YouTube, Google+, and Email. A "Create" button is located at the bottom left of the main content area.

An inset image shows a preview of the Databricks workspace interface, which includes a sidebar with navigation options like Overview, Activity log, Access control (IAM), Tags, Settings, and Support. The main workspace area displays the Databricks logo and a "Get started" button.

# Azure Databricks – creating a workspace

Microsoft Azure New > Marketplace > Everything > Azure Databricks (preview) > Azure Databricks Service

Search resources, services and docs

nishanth@microsoft.c... MICROSOFT (MICROSOFT.ON...)

### Azure Databricks Service

**Workspace name**  
ntedemodbr12252017 ✓

**Subscription**  
Azure conversion - External ▼

**Resource group**  
☒ Create new ☐ Use existing  
ntedemorg ✓

**Location**  
West US ▼

☒ Pin to dashboard

[Create](#) [Automation options](#)

**Left sidebar (Services):**

- New
- App Services
- SQL databases
- SQL data warehouses
- Azure Cosmos DB
- Virtual machines
- Load balancers
- Storage accounts
- Virtual networks
- Azure Active Directory
- Monitor
- Security Center
- Cost Management + Bil...
- Help + support
- Subscriptions
- Data Lake Store
- Data Lake Analytics
- Advisor
- Azure Databricks
- More services >

# Azure Databricks – workspace deployment

The screenshot displays the Microsoft Azure portal dashboard. The left sidebar contains a navigation menu with the following items: New, Dashboard, All resources, Resource groups, App Services, SQL databases, SQL data warehouses, Azure Cosmos DB, Virtual machines, Load balancers, Storage accounts, Virtual networks, Azure Active Directory, Monitor, Security Center, Cost Management + Billing, Help + support, Subscriptions, and Data Lake Store. The main content area is titled "Dashboard" and includes a search bar, a "New dashboard" button, and an "Edit dashboard" button. Below the search bar, there are several sections: "All resources" (listing various Azure services like Azure Databricks, App Service, Azure Cosmos DB, App Service plan, Search service, Storage account, SQL server, and SQL database), "Quickstart tutorials" (listing tutorials for Windows Virtual Machines, Linux Virtual Machines, App Service, Functions, and SQL Database), and "Service Health" (providing personalized guidance and support). A "Marketplace" button is also visible. The top right corner shows the user's profile and email address: nishanth@microsoft.c... MICROSOFT (MICROSOFT.ON...).

Microsoft Azure

Search resources, services and docs

Dashboard

+ New dashboard

Edit dashboard

Share

Fullscreen

Clone

Delete

All resources

AZURE CONVERSION - EXTERNAL

Refresh

ntedemodbr12252017 Azure Databricks Serv...

02082017tj App Service

02082017tj-docdb Azure Cosmos DB acc...

02082017tj-hosting-plan App Service plan

02082017tjs Search service

02082017tjso Storage account

02082017tj-sqlserver SQL server

AdventureWorks SQL database

See more...

Quickstart tutorials

Windows Virtual Machines

Provision Windows Server, SQL Server, SharePoint VMs

Linux Virtual Machines

Provision Ubuntu, Red Hat, CentOS, SUSE, CoreOS VMs

App Service

Create Web Apps using .NET, Java, Node.js, Python, PHP

Functions

Process events with a serverless code architecture

SQL Database

Managed relational SQL Database as a Service

Marketplace

Service Health

Personalized guidance and support when issues in Azure services affect you. [Learn more](#)

Deploying Azure Databricks (preview)

# Azure Databricks – launching the workspace

The screenshot displays the Microsoft Azure portal interface for the resource group 'ntedemodbr12252017'. The left sidebar lists various Azure services, with 'Azure Databricks' at the bottom. The main content area shows the 'Overview' tab for the 'ntedemo' resource group. It includes details such as the Managed Resource Group 'databricks-rg-ntedemodbr12252017-6kt7r3v4ehftu', the URL 'https://westeurope.azure.databricks.net', and the Subscription ID '15c5cb6e-191a-40ea-9f69-08207a17fe97'. A large red Databricks logo is centered, with a 'Launch Workspace' button below it. At the bottom, there are six tiles for 'Documentations', 'Getting Started', 'Import Data from File', 'Import Data from Azure Storage', 'Notebook', and 'Admin Guide'.

Microsoft Azure ntedemodbr12252017 Search resources, services and docs nishanth@microsoft.c... MICROSOFT (MICROSOFT.ON...

ntedemodbr12252017 Azure Databricks Service - PREVIEW

Search (Ctrl+/,)

Overview

Activity log

Access control (IAM)

Tags

SETTINGS

Locks

Automation script

SUPPORT + TROUBLESHOOTING

New support request

Delete

Resource group (change) ntedemo

Subscription (change) Azure conversion - External

Subscription ID 15c5cb6e-191a-40ea-9f69-08207a17fe97

Managed Resource Group databricks-rg-ntedemodbr12252017-6kt7r3v4ehftu

URL https://westeurope.azure.databricks.net

Guides Documentations

Launch Workspace

Documentations

Getting Started

Import Data from File

Import Data from Azure Storage

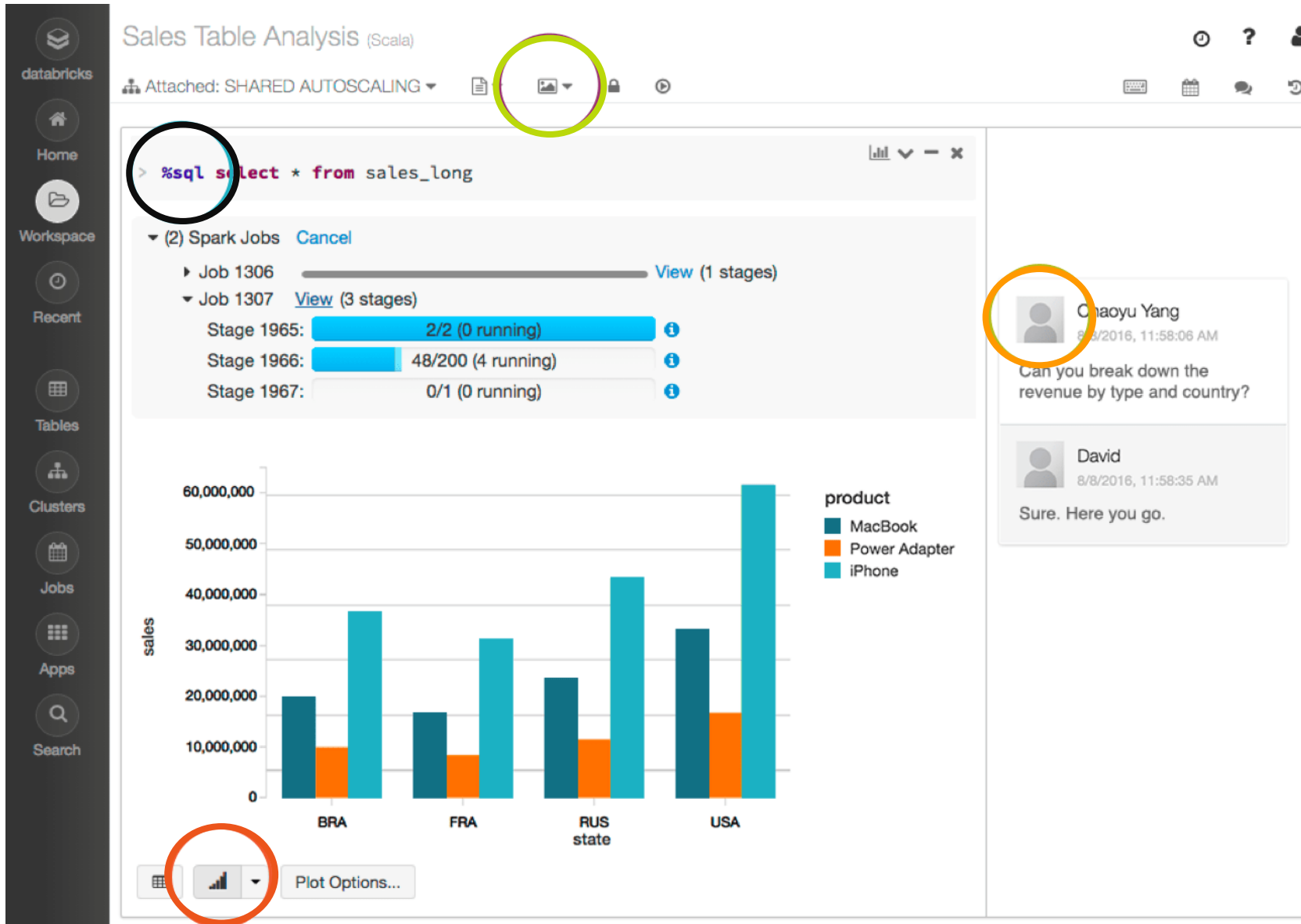
Notebook

Admin Guide

# Azure Databricks – workspace home page

The image shows the Azure Databricks portal interface. At the top, there is a dark header bar with the 'Microsoft Azure' logo on the left, the word 'PORTAL' in the center, and the user email 'nishanth@microsoft.com' on the right. Below the header is a vertical sidebar on the left containing icons and labels for 'Azure', 'Databricks', 'Home', 'Workspace', 'Recent', 'Data', 'Clusters', 'Jobs', and 'Search'. The main content area has a large 'Azure Databricks' logo at the top. Below the logo, there is a section titled 'Featured Notebooks' which contains three notebook thumbnails: 'Introduction to Apache Spark on Databricks', 'Databricks for Data Scientists', and 'Introduction to Structured Streaming'. Each thumbnail has a Python logo icon. Below this section, there are three columns. The first column is titled 'New' and lists 'Notebook', 'Job', 'Cluster', 'Table', and 'Library' with corresponding icons. The second column is titled 'Documentation' and lists 'Databricks Guide', 'Python, R, Scala, SQL', and 'Importing Data' with external link icons. The third column is titled 'Open Recent' and contains the text 'Recent files appear here as you work. Get started with the welcome guide.' with a link to the 'welcome guide'.

# Interactive Data Science



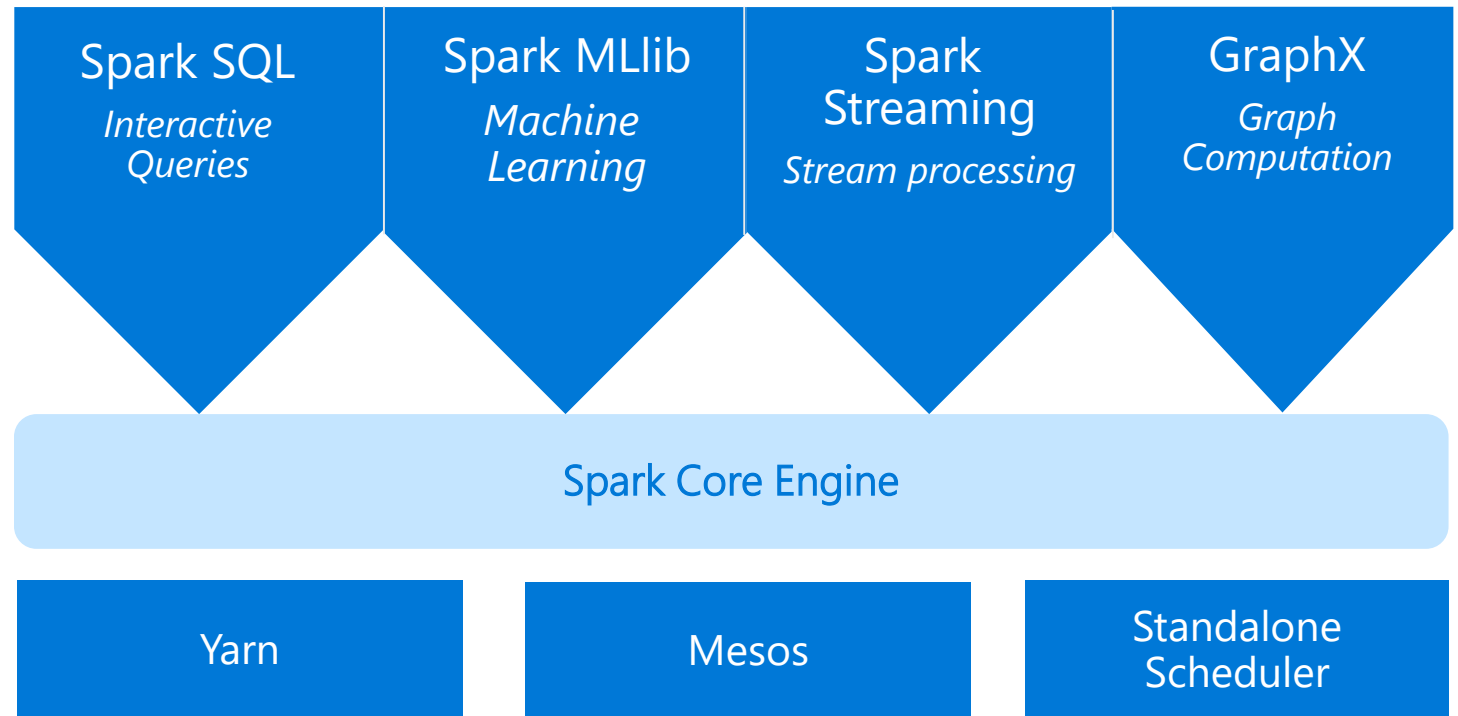


# A P A C H E   S P A R K

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



# SPARK - BENEFITS

## Performance

Using in-memory computing, Spark is considerably faster than Hadoop (100x in some tests).  
Can be used for batch and real-time data processing.

## Developer Productivity

Easy-to-use APIs for processing large datasets.  
Includes 100+ operators for transforming.

## Unified Engine

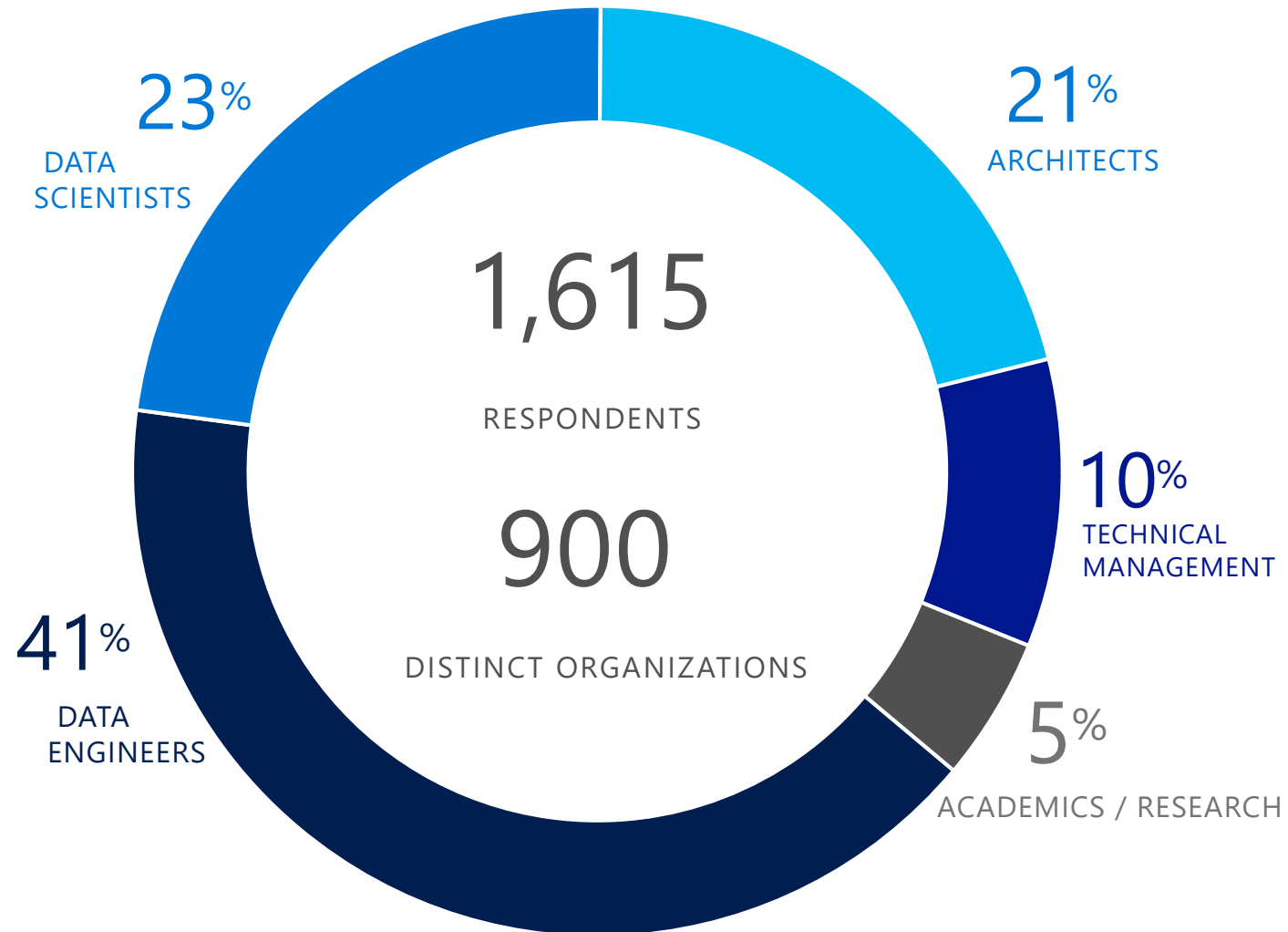
Integrated framework includes higher-level libraries for interactive SQL queries, Stream Analytics, ML and graph processing.  
A single application can combine all types of processing

## Ecosystem

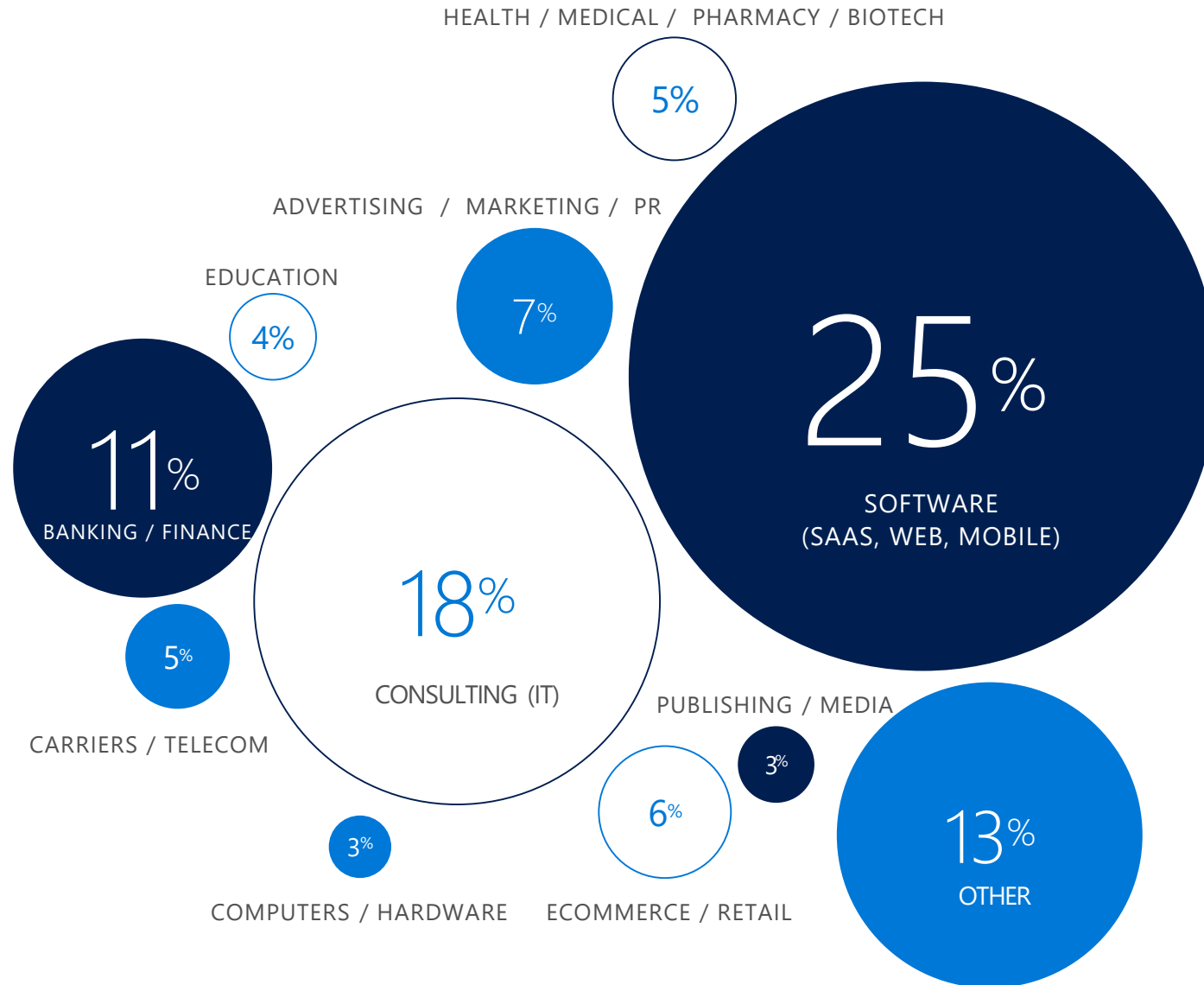
Spark has built-in support for many data sources, rich ecosystem of ISV applications and a large dev community.  
  
Available on multiple public clouds (AWS, Google and Azure) and multiple on-premises distributors

# APACHE SPARK SURVEY

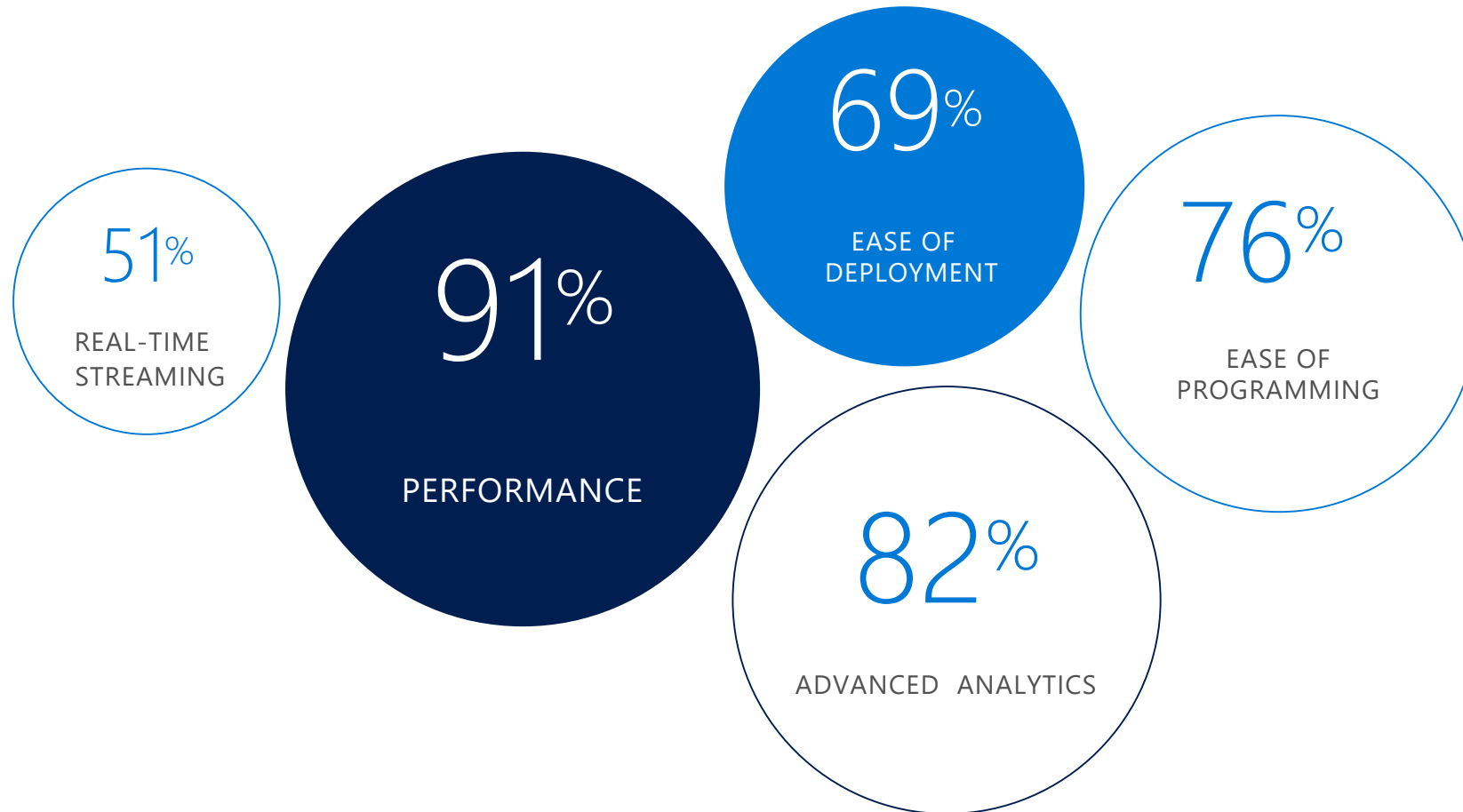
Identifying how organizations use Apache Spark.



# KEY INDUSTRIES USING SPARK



# FEATURES CONSIDERED IMPORTANT



# RECOMMENDATION ENGINES DELIGHT CUSTOMERS

## Problem

The average size of a single cart has decreased as customers spend more money in general but spend less per retailer

## Solution

Convert shoppers into buyers by providing personalized digital content

## Results



Engage customers



Deliver relevant content



Increase cart size



ASOS delivers 15.4 million personalized experiences  
with 33 orders per second

# PREDICTIVE MAINTENANCE OPTIMIZES OPERATIONS

## Problem

Reactive maintenance results in cost over-runs and unnecessary down-time

## Solution

Determine the condition of in-service equipment in order to predict when maintenance should be performed

## Results



Maximize asset availabilities



Minimize the production hours lost to maintenance



Minimize the cost of spare parts and supplies



Hybrid solution predicts onboard water usage,  
saving \$200k/ship/year

# DEMAND FORECASTING MAXIMIZES REVENUE

## Problem

Renewable energy sources such as solar power are challenging to integrate with global energy markets, as they are generated inconsistently from a variety of sources

## Solution

Leverage AI to forecast energy production, integrate with power transmission grids and align energy supply with the optimal energy markets

## Results



Streamline product development



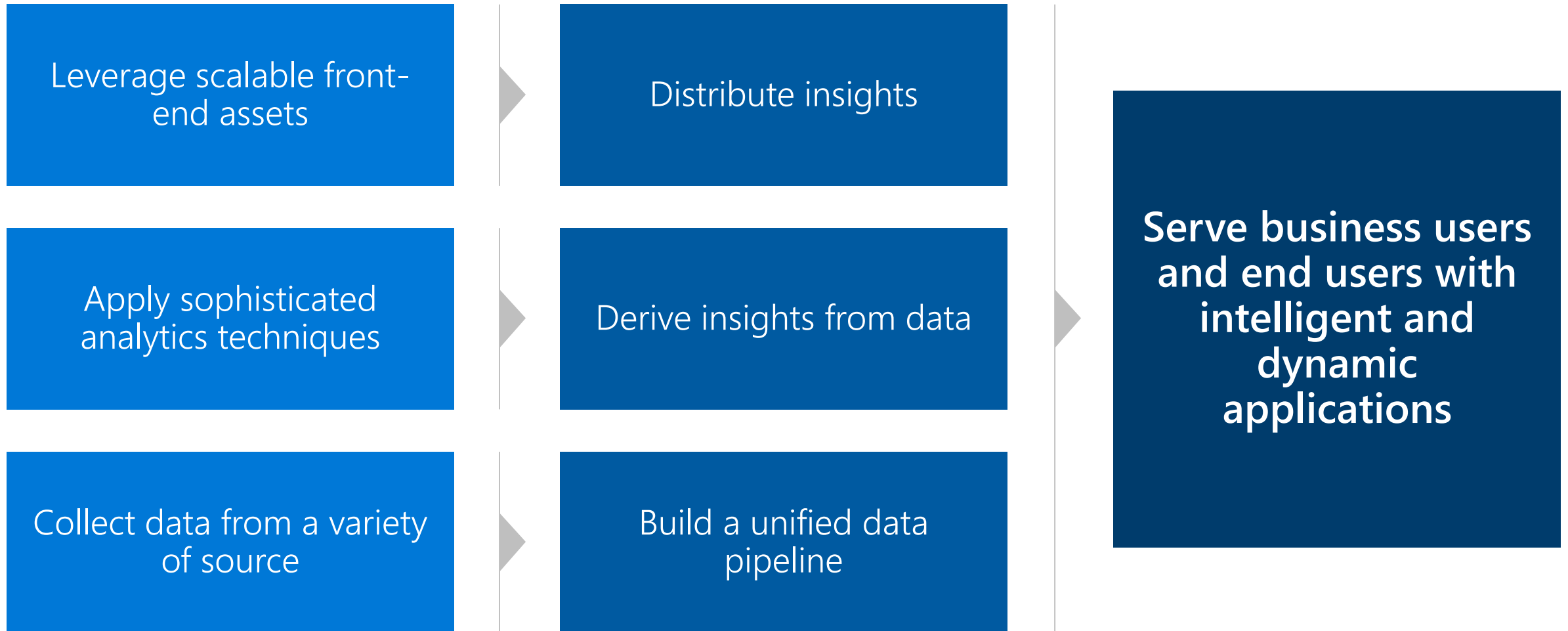
Increase revenue and savings



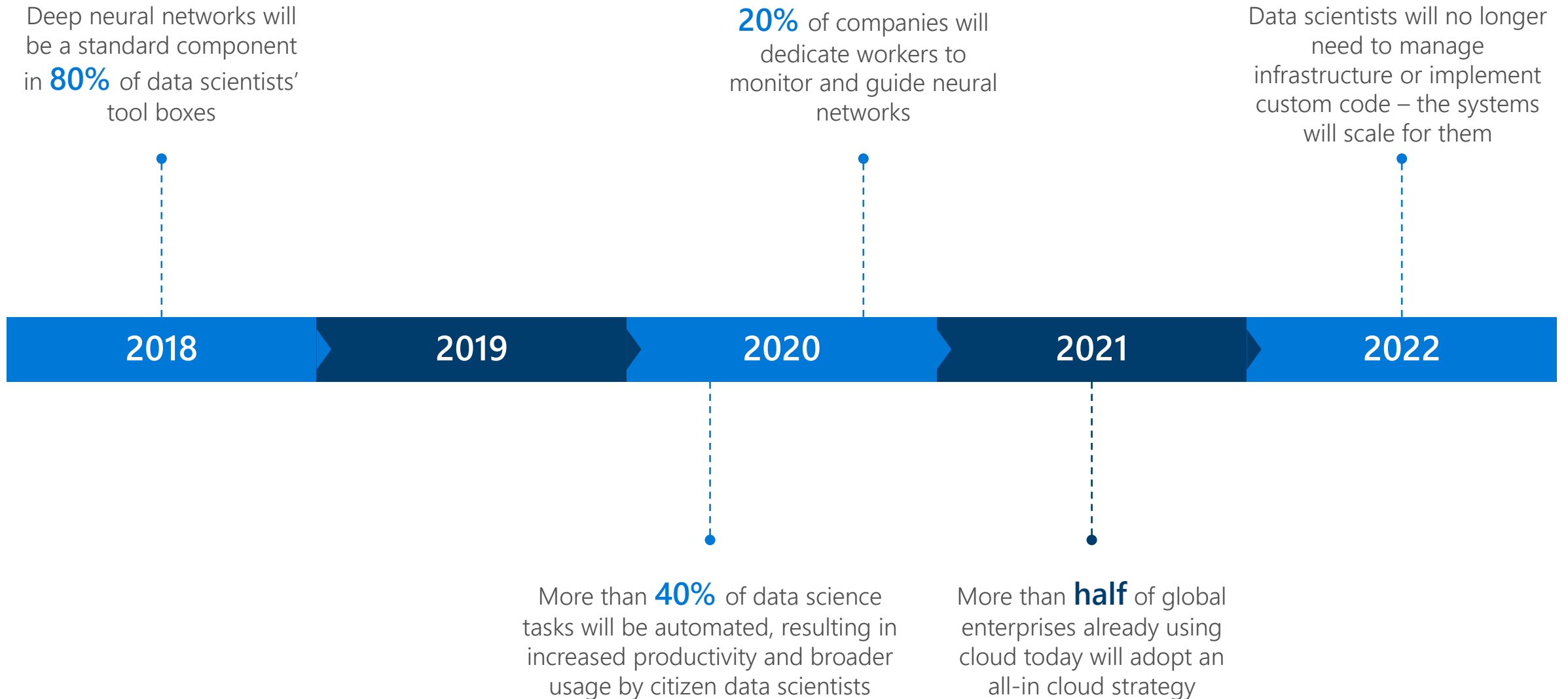
Drive adoption of renewable energy



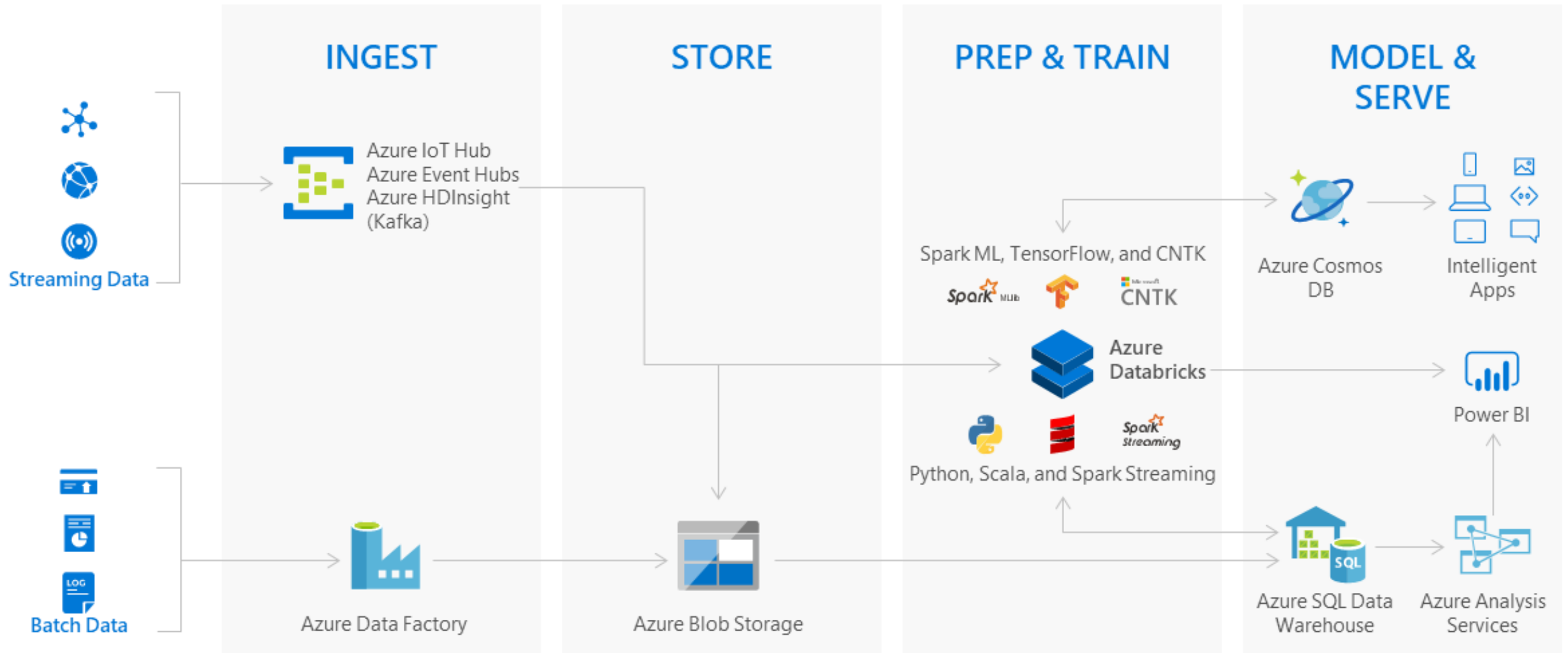
# How are these businesses transforming?



# What are these customers looking to do next?



# LOGICAL ARCHITECTURE

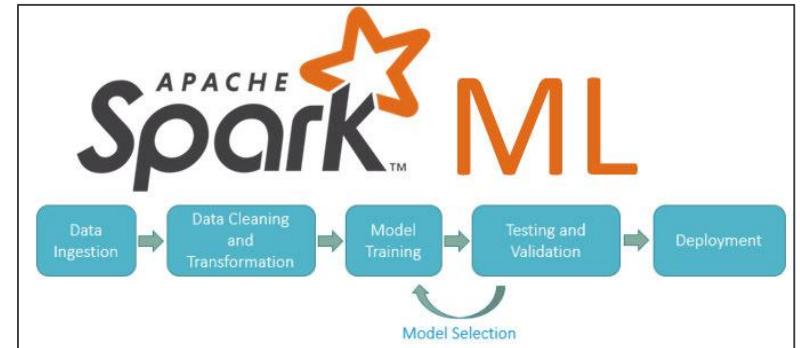


# Machine Learning and Deep Learning with Azure Databricks

# SPARK MACHINE LEARNING (ML) OVERVIEW

Enables Parallel, Distributed ML for large datasets on Spark Clusters

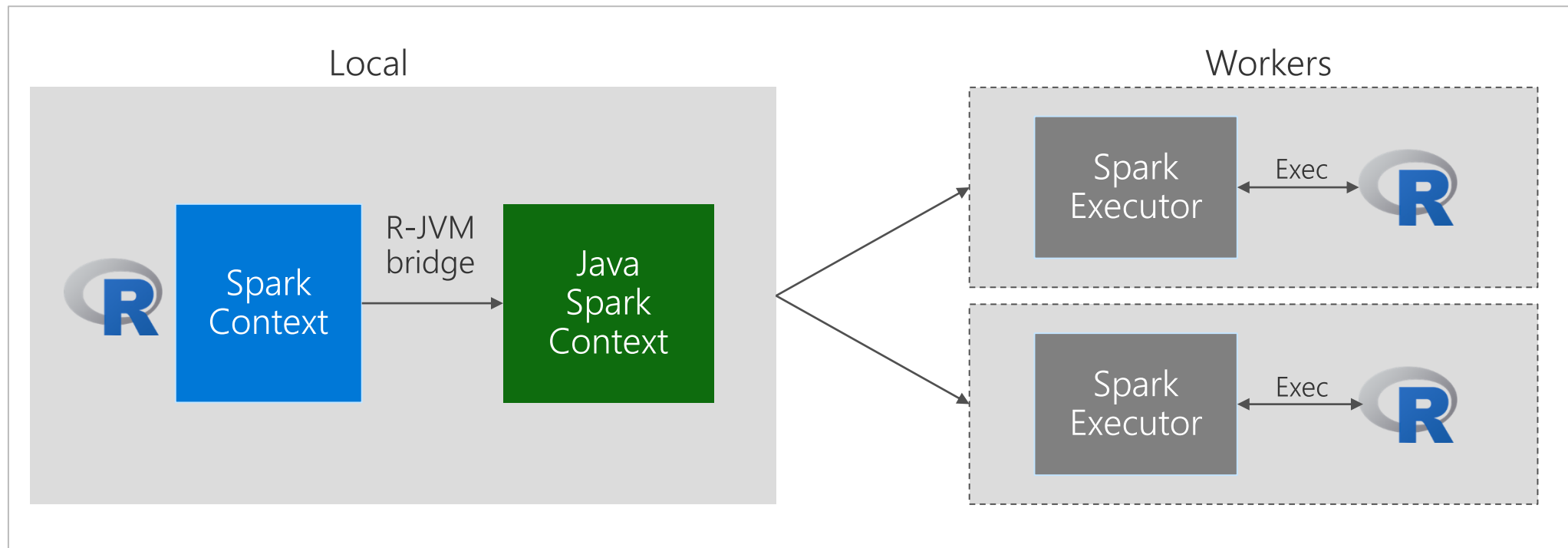
- Offers a set of parallelized machine learning algorithms (see next slide)
- Supports [Model Selection](#) (hyperparameter tuning) using [Cross Validation](#) and [Train-Validation Split](#).
- Supports Java, Scala or Python apps using [DataFrame](#)-based API (as of Spark 2.0). Benefits include:
  - 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
- 3<sup>rd</sup> Party libraries supported include: [H2O Sparkling Water](#), [SciKit-learn](#) and [XGBoost](#)



# SPARKR OVERVIEW

An R package that provides a light-weight frontend to use Apache Spark from R

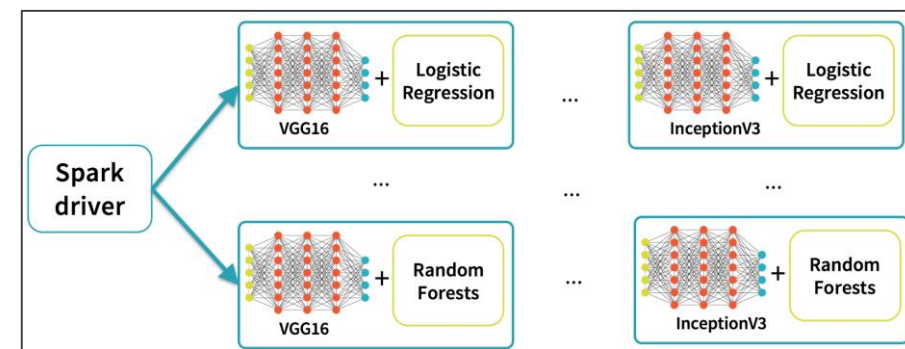
- Provides a distributed DataFrame implementation that supports operations like selection, filtering, aggregation etc (similar to R data frames, dplyr)
- Supports distributed machine learning using Spark MLlib.
- R programs can connect to a Spark cluster from RStudio, R shell, Rscript or other R IDEs.



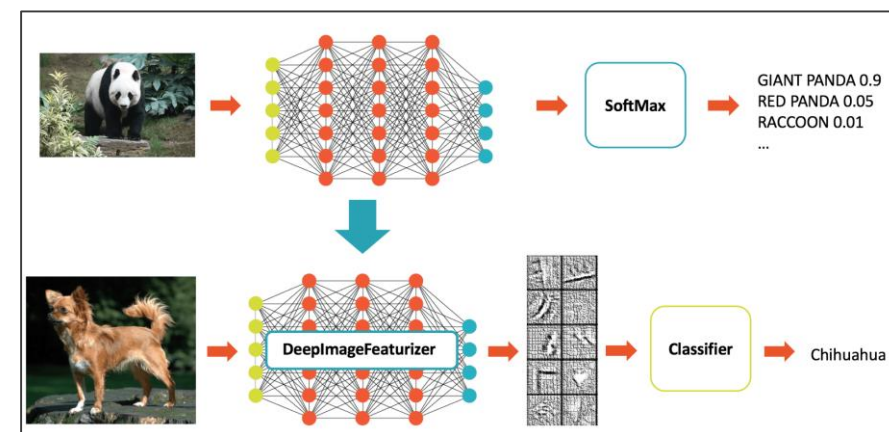
# DEEP LEARNING

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



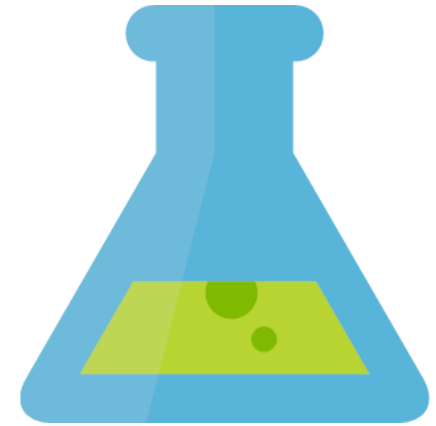
Transfer Learning

# M M L S P A R K

[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:

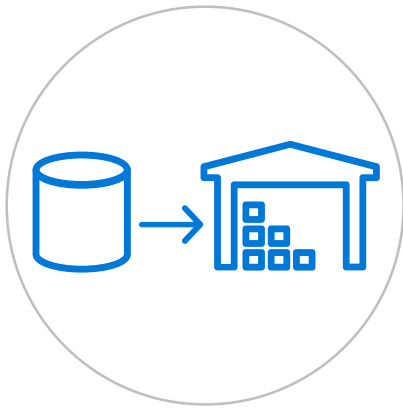
- Ingress and pre-process image data
- Featurize images and text using pre-trained deep learning models
- Train and score classification and regression models using implicit featurization





Azure Databricks has a proven end-to-end process for Custom AI

# CREATE AND DEPLOY MODELS IN THREE STEPS



---

Collect and prepare data  
for training



---

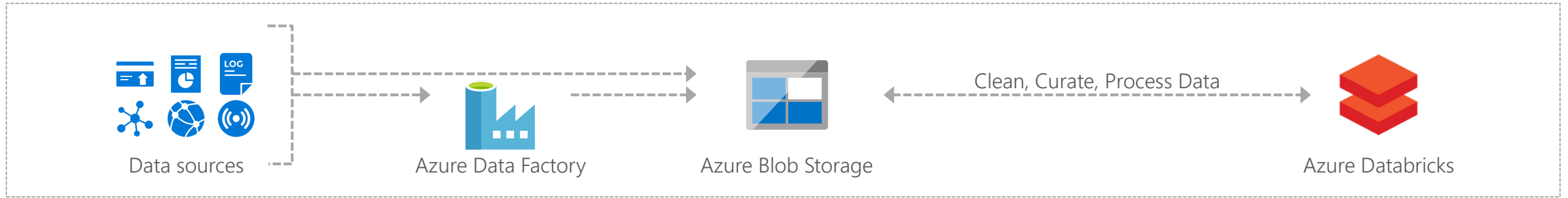
Train and evaluate AI  
and ML models



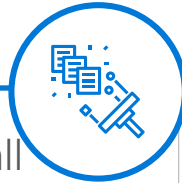
---

Operationalize and  
manage models

# COLLECT AND PREPARE DATA AT SCALE

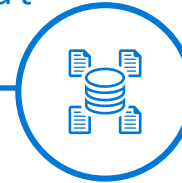


## Connect to data from any source



- Seamlessly integrate with all of your data sources
- Create hybrid pipelines
- Ingest and orchestrate with a serverless, code-free environment

## Store and process without limits



- Process any data, of any size, and at any speed with enterprise-grade security
- Scale up or scale out to serve your analytics needs
- Scale compute and storage separately to manage TCO

## Leverage best-in-class analytics capabilities

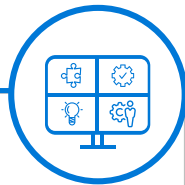


- Choose from the best of open source and proprietary technologies
- Use machine learning on batch streams to create unified data pipelines
- Collaborate within teams to accelerate time-to-insight

# Train and evaluate Machine Learning models



## Simplify model development



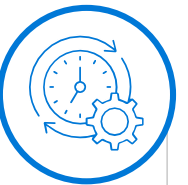
- Easily collaborate in interactive workspaces
- Leverage a library of battle-tested models

## Scale compute resources to meet your needs



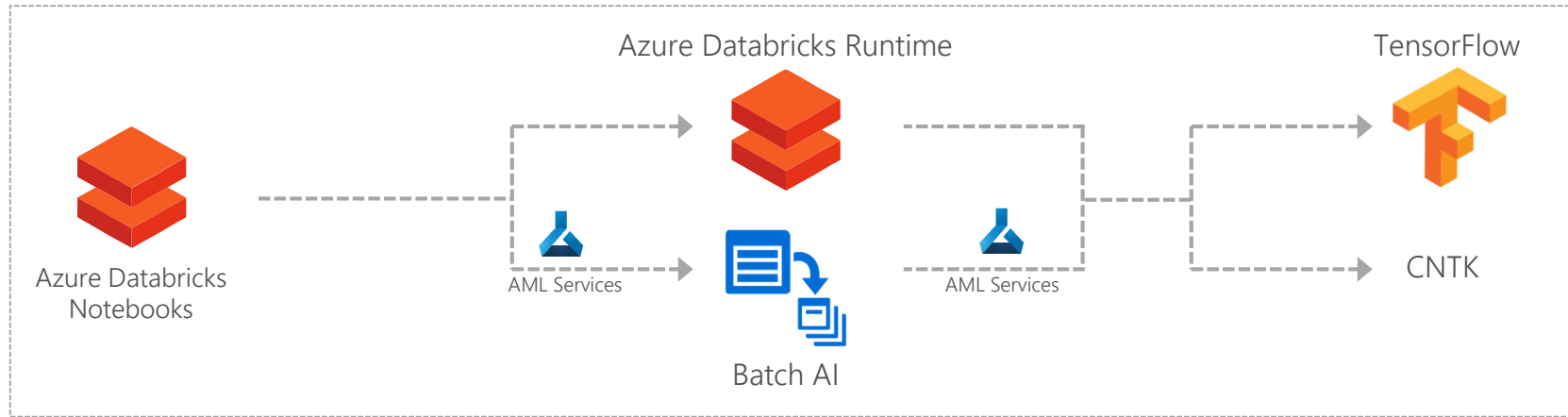
- Easily scale up on VMs or scale out on clusters
- Pay only for what you use in the cloud
- Leverage commodity hardware to reduce TCO

## Quickly determine the right model for your data



- Test multiple models in a single experiment and quickly compare results
- Rapidly prototype in agile environments

# Train and evaluate AI and DL models



## Streamline AI development efforts



- Build models using popular deep learning toolkits
- Leverage out-of-the-box capabilities across many common use cases
- Develop in the languages and tools of your choice

## Scale compute resources in any environment



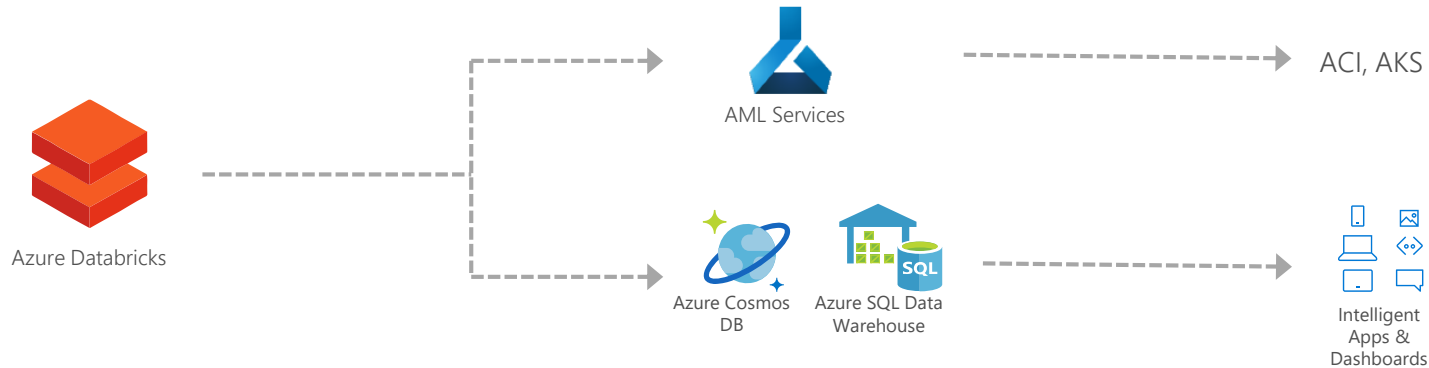
- Choose VM types based on your modeling requirements
- Process images and videos using GPU-based and FPGA-based VMs

## Quickly evaluate and identify the right model



- Save time by running experiments in parallel
- Focus on your workload by automating resource provisioning and management

# Deploy and manage models with ease



## Bring models to life quickly



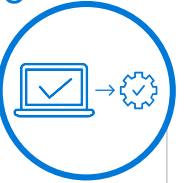
- Take models from inception to production in minutes
- Deploy models to the destination of your choice
- Easily change environments for agile experimentation

## Proactively manage model performance



- Identify and promote your best performing models
- Capture model telemetry for actionable insights
- Retrain models and update deployments with programmatic APIs

## Empower data scientists to drive powerful analytics



- Develop in the languages and tools of your choice
- Leverage pre-configured VMs for data science
- Build in a serverless, drag-and-drop environment

# Demo

*Seeing Custom AI with Azure Databricks in action*

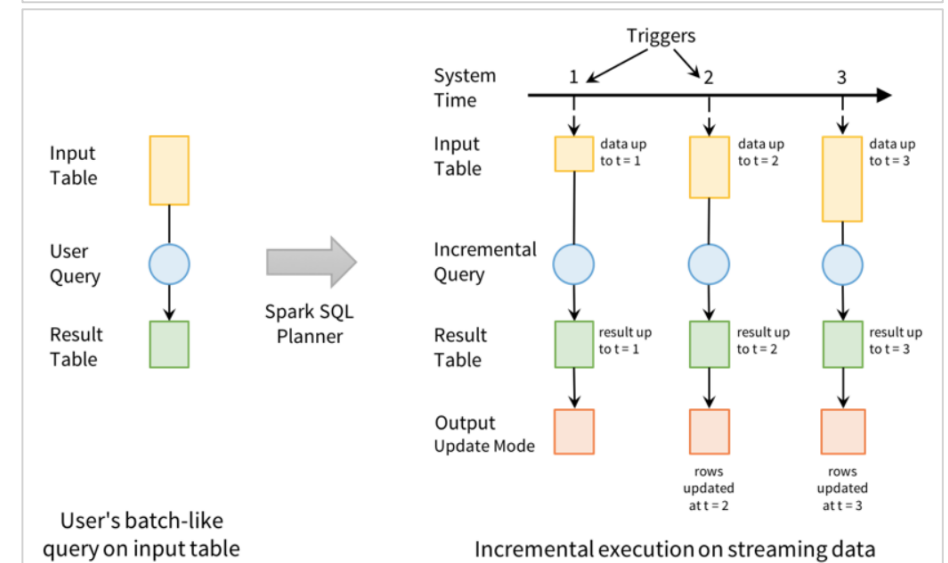
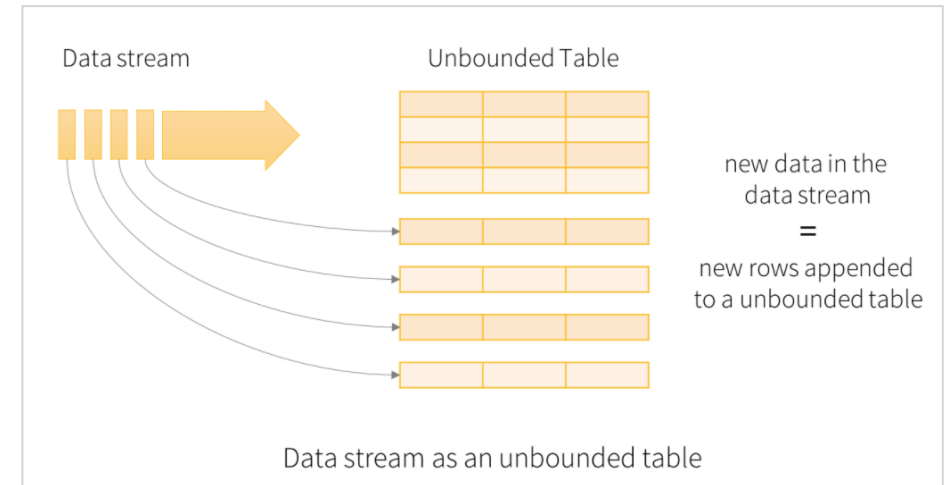
# Stream Analytics & Graph Processing



# SPARK STRUCTURED STREAMING OVERVIEW

A unified system for end-to-end fault-tolerant, exactly-once stateful stream processing

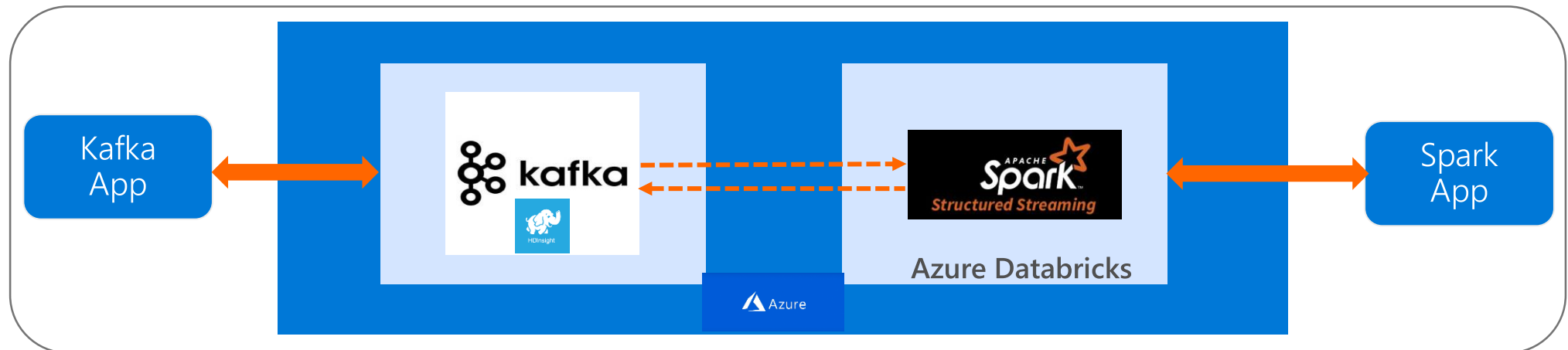
- Unifies streaming, interactive and batch queries—a single API for both static bounded data and streaming unbounded data.
- Runs on Spark SQL. Uses the Spark SQL [Dataset/DataFrame](#) API used for batch processing of static data.
- Runs incrementally and continuously and updates the results as data streams in.
- Supports app development in Scala, Java, Python and R.
- Supports streaming aggregations, event-time windows, windowed grouped aggregation, stream-to-batch joins.
- Features streaming deduplication, multiple output modes and APIs for managing/monitoring streaming queries.
- Built-in sources: Kafka, File source (json, csv, text, parquet)



# 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

# SPARK GRAPHX OVERVIEW

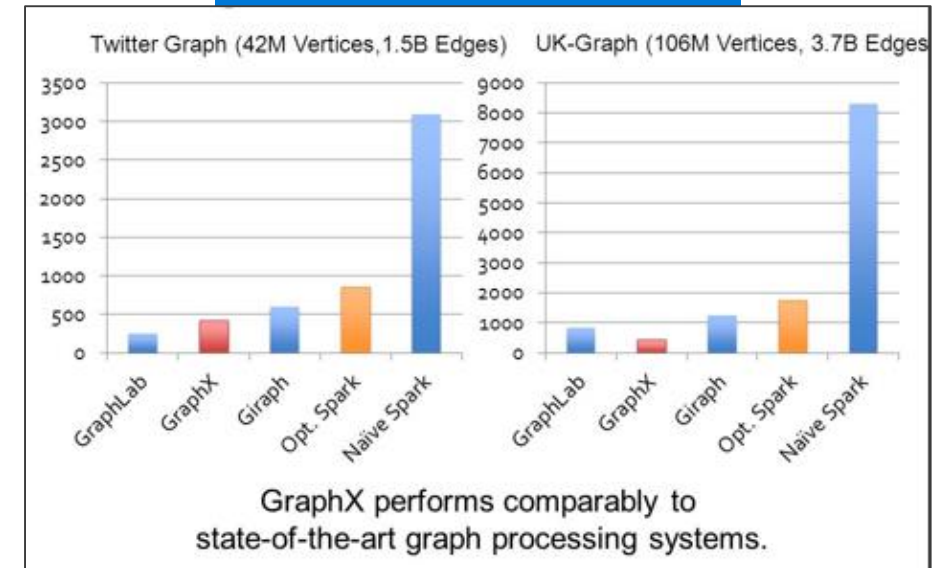
A set of APIs for graph and graph-parallel computation.

- Unifies ETL, exploratory analysis, and iterative graph computation within a single system.
- Developers can:
  - [view](#) the same data as both graphs and collections,
  - [transform](#) and [join](#) graphs with RDDs, and
  - write custom iterative graph algorithms using the [Pregel API](#).
- Currently only supports using the Scala and RDD APIs.

## Algorithms

- PageRank
- Connected components
- Label propagation
- SVD++
- Strongly connected components
- Triangle count

## PageRank Benchmark

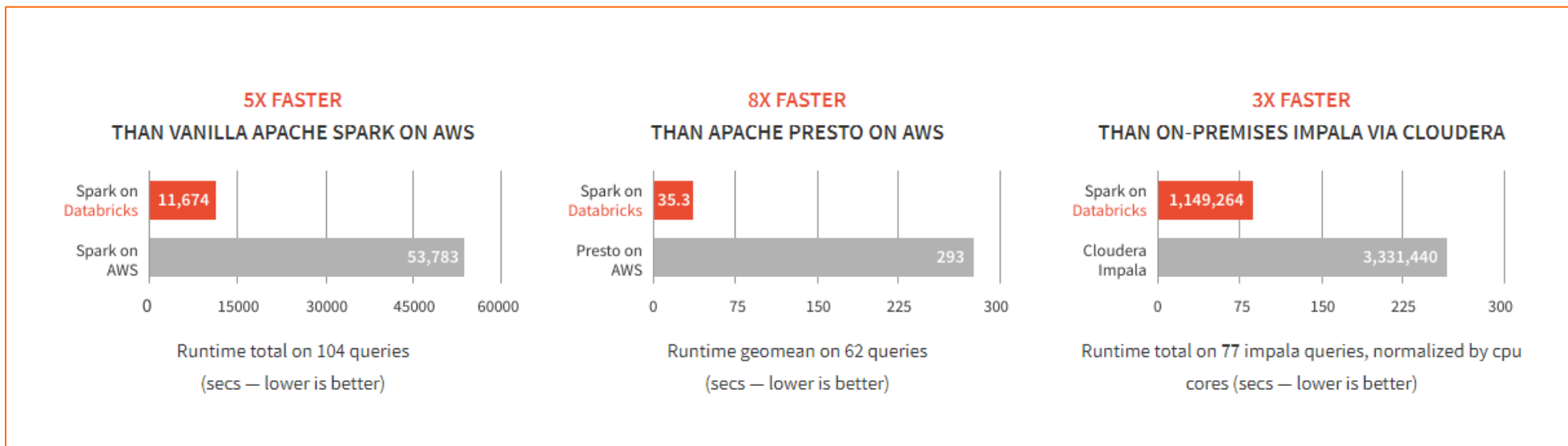


Source: [AMPLab](#)

# Azure Databricks Performance

# DATABRICKS SPARK IS FAST

Benchmarks have shown Databricks to often have better performance than alternatives



**SOURCE:** [Benchmarking Big Data SQL Platforms in the Cloud](#)

# Turn ideas into solutions faster with Microsoft



## Productive

Accelerate time to market



## Hybrid

Optimize your infrastructure



## Intelligent

Innovate at scale



## Trusted

Develop with confidence

# Azure Databricks Security Architecture

# Azure Security Best Practices

Databricks designed the security infrastructure and configurations based on Azure security best practices including but not limited to the following:

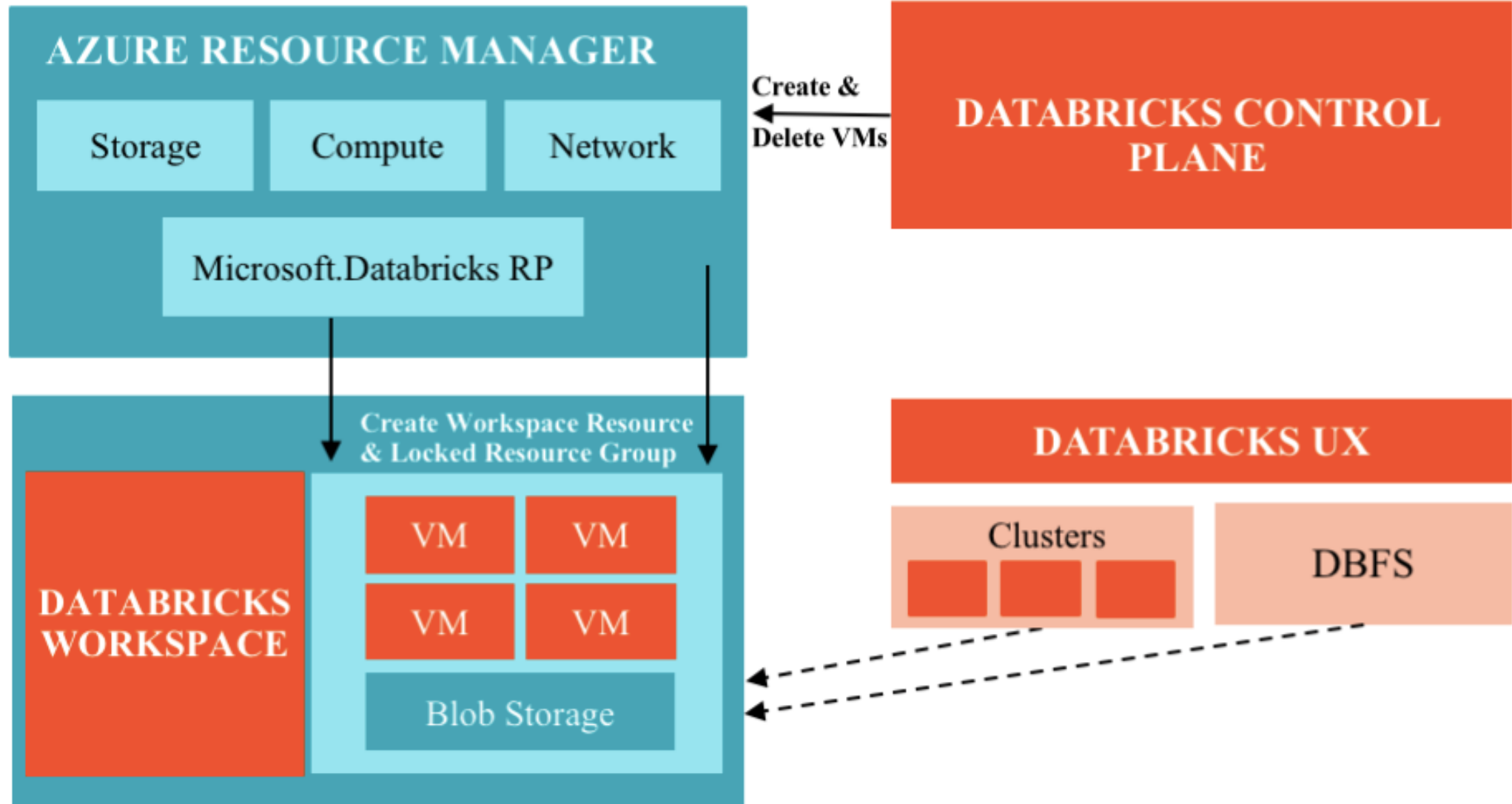
- Azure Management Console access is controlled with IdP (e.g. OKTA) SSO with Multifactor-Authentication.
- Data stores are encrypted at rest using native Azure storage encryption.
- Customers are isolated using managed resource groups and VNETs.



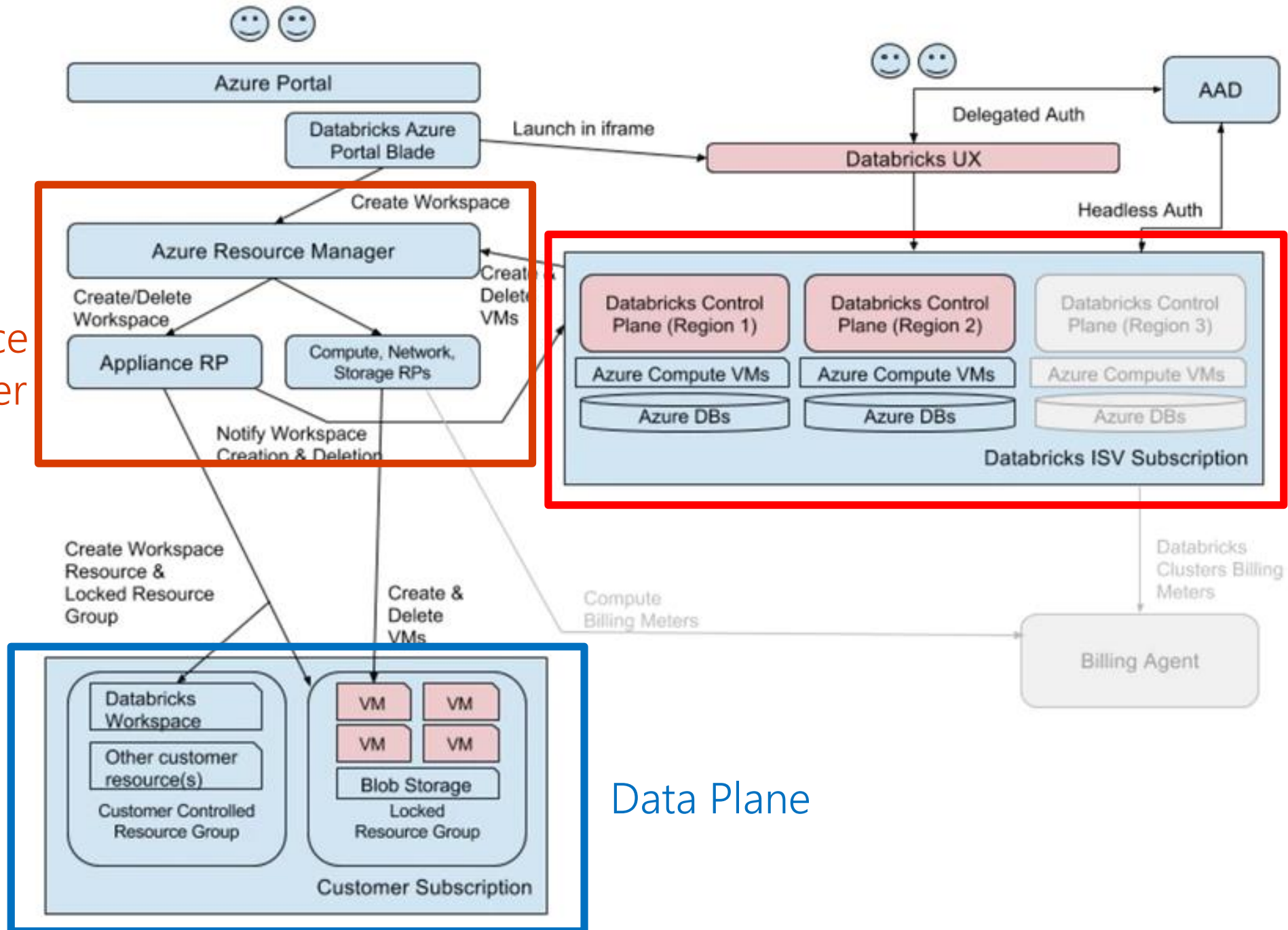
# Architecture

- Azure Databricks is a multi-tenant deployment.
- Upon sign-up, the Databricks launches a Control Plane in Databricks' Azure account and the Data Plane will be launched in Customers' Azure account.
- The Control plane and the Data plane are be connected via vnet peering (VNET peering targeted for GA).

# AZURE PORTAL



Azure  
Resource  
Manager




Control  
Plane

Data Plane



# Azure Databricks Architecture

**demo**  
Azure Databricks Service - PREVIEW

Search (Ctrl+ /)

Overview

Activity log

Access control (IAM)

Tags


SETTINGS

Locks

Automation script

SUPPORT + TROUBLESHOOTING

New support request

 Delete

Resource group [\(change\)](#)  
demo

Subscription [\(change\)](#)  
Databricks Development worker


Subscription ID  
36f75872-9ace-4c20-911c-aea8eba2945c

Managed Resource Group  
databricks-rq-demo-cizog3x24bcpi


URL  
<https://eastus2.azuredatabricks.net>


Guides  
[Documentations](#)


⤴




Loading...

Documentations

Getting Started

Import Data from File

Import Data from Azure Storage



# Azure Databricks Architecture



databricks-rg-demo-cjzoq3x24bcpi

Resource group



Search (Ctrl+/)



Overview



Activity log



Access control (IAM)



Tags

## SETTINGS



Quickstart



Resource costs



Deployments



Add



Assign Tags



Columns



Delete resource group



Refresh



Move

Essentials



Subscription name [\(change\)](#)

Databricks Development worker

Subscription ID

36f75872-9ace-4c20-911c-aea8eba2945c

Deployments

1 Succeeded

Filter by name...

All types



All locations



No grouping



3 items



NAME

TYPE

LOCATION



dbstorage4pqprhkhpdgoa

Storage account

East US 2



workers-sg

Network security gro...

East US 2



workers-vnet

Virtual network

East US 2



# Control Plane

- Notebooks, jobs, clusters, users, and ACLs are managed by control plane services.
- These services store their data in dedicated databases in Databricks' Azure subscription.
- Access to Control Plane VNET is limited to Databricks Shared Services through security groups. These services are not Internet facing, and access is only provided via a proxy server.
- The web application UI and API is accessible via the Internet.
- Access to Cluster Manager is restricted to the webapp via security groups.

# Data Plane

Spark Clusters are deployed in a customer Azure subscription. Each workspace and its associated clusters reside within a dedicated VNET, which separates it from other workspaces.

- The customer deployment is isolated at a VNET level. In a VNET, public IPs are assigned to nodes, but access is restricted via Azure security groups.
- Workspaces are launched within managed security groups.
- Managed groups allow connections from the control plane and allow workers to communicate with each other.

# Databricks File System (DBFS)

- The Databricks File System or DBFS is a distributed file system that comes installed on Spark Clusters in Databricks. It is a layer over Azure Blob Storage, which allows customers to Mount containers to make them available to users in your workspace.



# Application Security




# Authentication

- Databricks authentication is delegated to Azure Active Directory (AAD).

# Access Control and Permissions

Databricks has powerful permissions settings for stricter control over what users can perform what actions. This includes access controls for clusters and workspaces. These controls are quintessential for larger organizations with many users; Databricks makes managing permissions easy.

Who has access:

 admins (group)		
 Alice (alice@mycompany.com)	<div>No Permissions Can Read Can Run Can Edit ✓ Can Manage</div>	✕
 Bob (bob@mycompany.com)	Can Manage ▾	✕



# Cluster Access Control

- Using this feature, control which users can:
  - Attach notebooks to clusters
  - Terminate clusters
  - Start clusters
  - Restart clusters
  - Resize clusters
  - Modify permissions

# Workspace Access Control

- For each Notebook, control which users can:
  - View Cells
  - Comment
  - Run Commands
  - Attach or Detach notebooks
  - Edit cells
  - Change Permissions
- For each Folder, control which users can:
  - Create or Delete items
  - Move or Rename Items
  - Change Permissions

<i>Abilities</i>	No Permissions	Read	Run	Edit	Manage
View cells		✓	✓	✓	✓
Comment		✓	✓	✓	✓
Run Commands			✓	✓	✓
Attach/detach notebooks			✓	✓	✓
Edit cells				✓	✓
Change permissions					✓

<i>Abilities</i>	No Permissions	Read	Run	Edit	Manage
Create items					✓
Delete items					✓
Move/rename items					✓
Change permissions					✓

# Jobs Access Control

- For each scheduled job, control which users can:
  - View job details and settings
  - View results, Spark UI, logs of a job run
  - Run now
  - Cancel run
  - Edit job settings
  - Modify permissions
  - Delete job
  - Change owner

# Jobs Access Control

Abilities	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
Delete job				x	x
Change owner					x



# Tables Access Control

- Databricks supports fine-grained access control via the Spark SQL interface. In this context, access can be restricted for any securable objects, e.g., tables, views, databases or functions.
- Fine-grained level access control (i.e. on rows or columns matching specific conditions) can be accomplished via access control on derived views that can contain arbitrary queries. These access control policies are enforced by the Databricks SQL query analyzer at runtime.

# Structured Data (Tables) Access Control

## Privileges

- `SELECT` privilege – gives read access to an object.
- `CREATE` privilege – gives ability to create an object (e.g., a table in a database).
- `MODIFY` privilege – gives ability to add/delete/modify data to/from an object (e.g., a table).
- `READ_METADATA` privilege – gives ability to view an object and its metadata.
- `CREATE_NAMED_FUNCTION` privilege – gives ability to create a named UDF in an existing catalog or database.
- `ALL PRIVILEGES` – gives all privileges (gets translated into all the above privileges).

## Objects

These privileges apply to the following class of objects:

- `CATALOG` - controls access to the entire data catalog.
- `DATABASE` - controls access to a database.
- `TABLE` - controls access to a managed or external table.
- `VIEW` - controls access to SQL views.
- `FUNCTION` - controls access to a named function.
- `ANONYMOUS FUNCTION` - controls access to anonymous or temporary functions.
- `ANY FILE` - controls access to the underlying filesystem.

# HOW TO GET STARTED

- [Already using Azure? try Azure Databricks now](#) or
- Create a [free Azure account to start using Azure Databricks](#)
- **Engage** Microsoft experts for a workshop to help identify high impact scenarios
- **Learn more** about Azure Databricks  
[www.azure.com/databricks](https://www.azure.com/databricks)

## Customer Ready Resources

- [14-Day Free Trial and Custom Account page](#)
- [Azure Databricks app on Azure portal](#)
- [Accelerate innovation with Azure Databricks webinar](#)
- [Azure Databricks Get Started Documentation](#)

## External Product Sites:

- [Azure Databricks website](#)
- [Databricks website](#)
- [Announcement blog post](#)





© 2018 Microsoft Corporation. All rights reserved. Microsoft, Windows, and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.