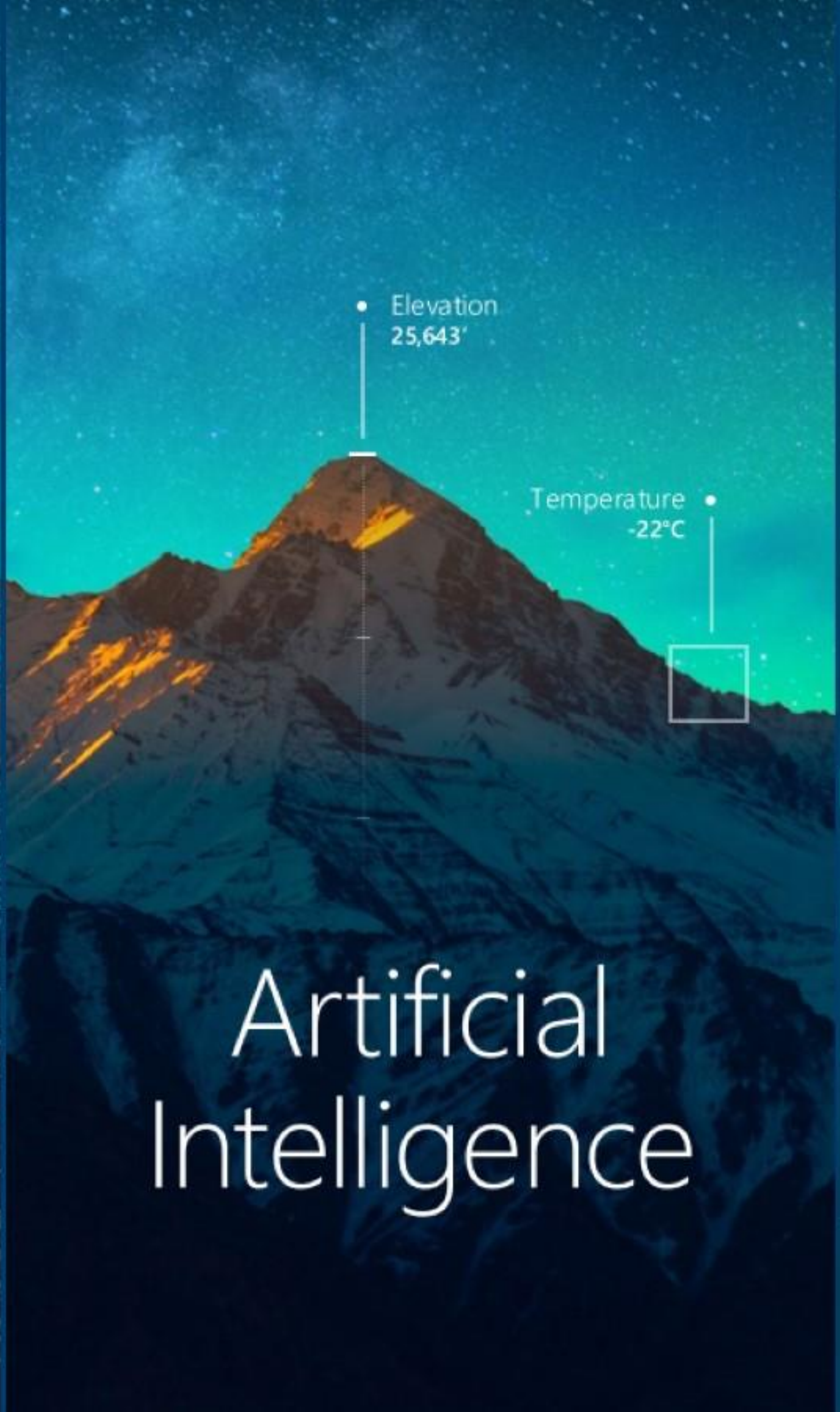
Microsoft Azure Databricks

Csom Gergely
Data & AI megoldásszakértő
Microsoft

Mixed Reality



Artificial Intelligence

Quantum Computing

Linear flow

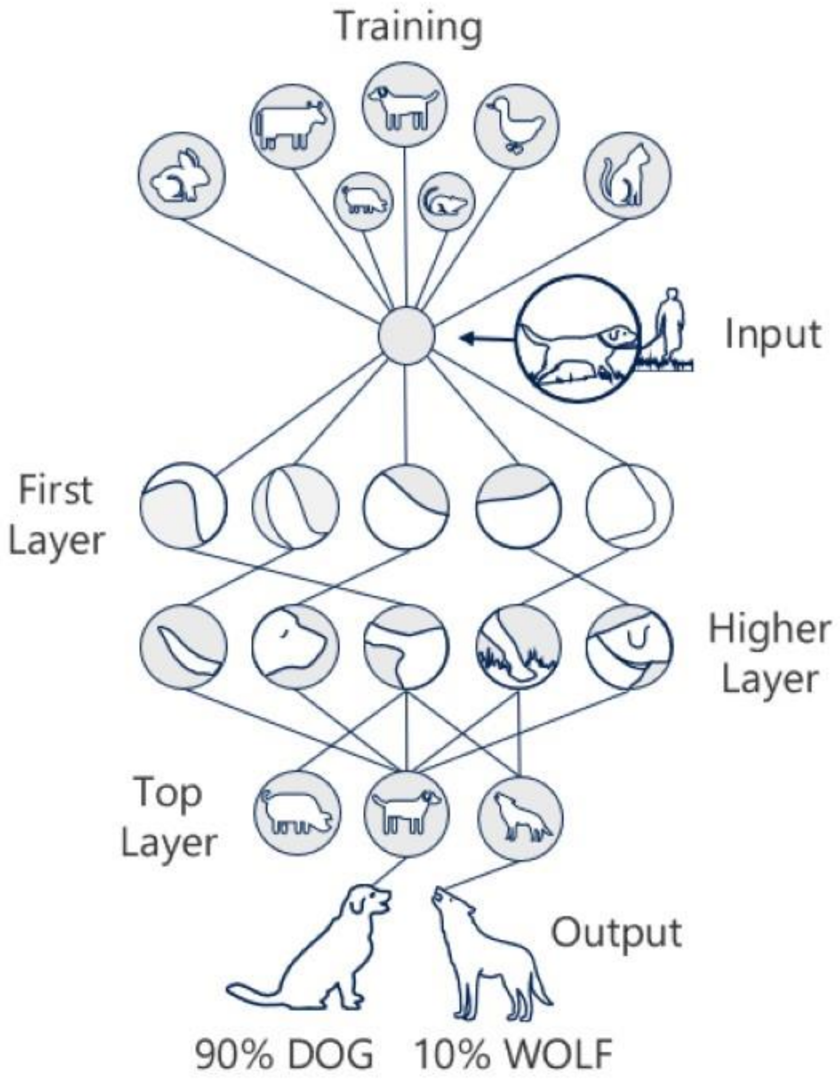
```
switch (animal)
{
    case doberman_pinscher:
    case german_shepard:
    case labrador_retriever:
    // Need to list every dog type :(
        animaltype = "dog”;
        break;

    // Managing different poses, ouch
    // What about dogs that I missed?
    // How about other animals?
    // This is tedious

    default:
        break;
}
```

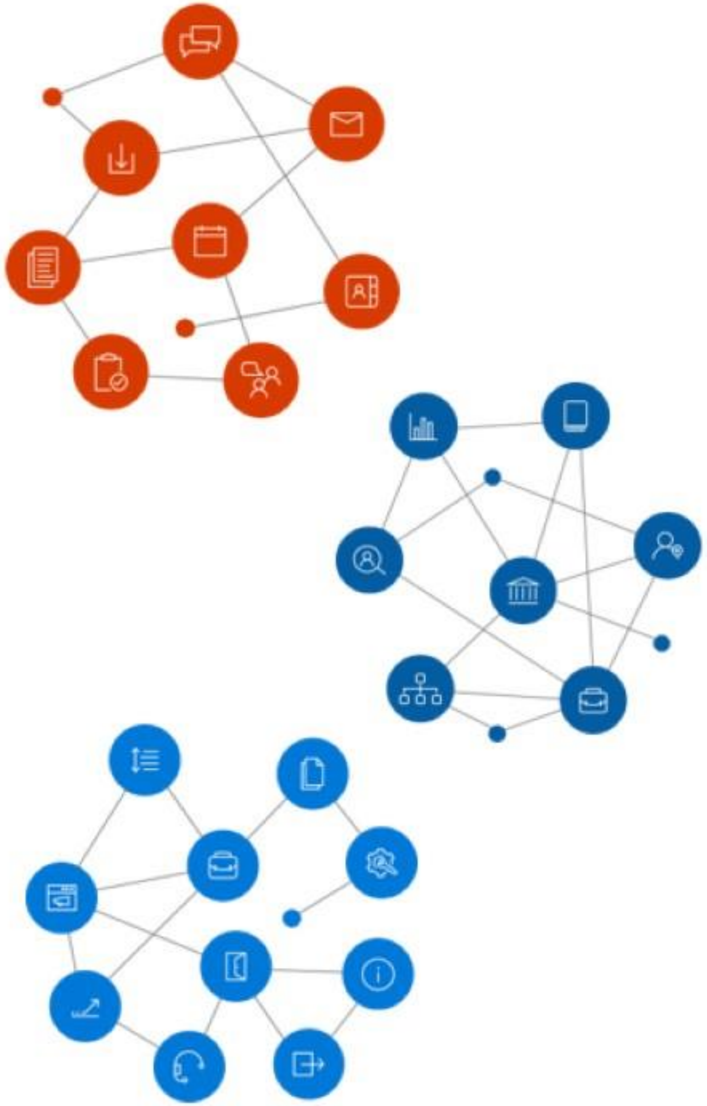
AI development

Probabilistic flow

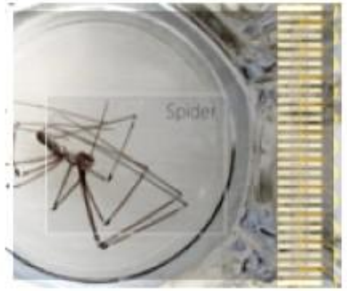


Cloud

Massive data



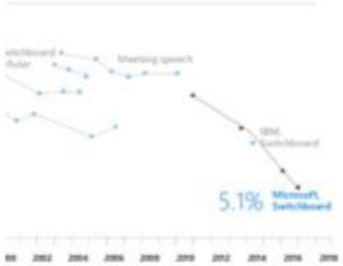
AI innovation



Vision

152 layer DNN

Wins ImageNet



Speech

5.1% Error Rate

New milestone

Wins Switchboard

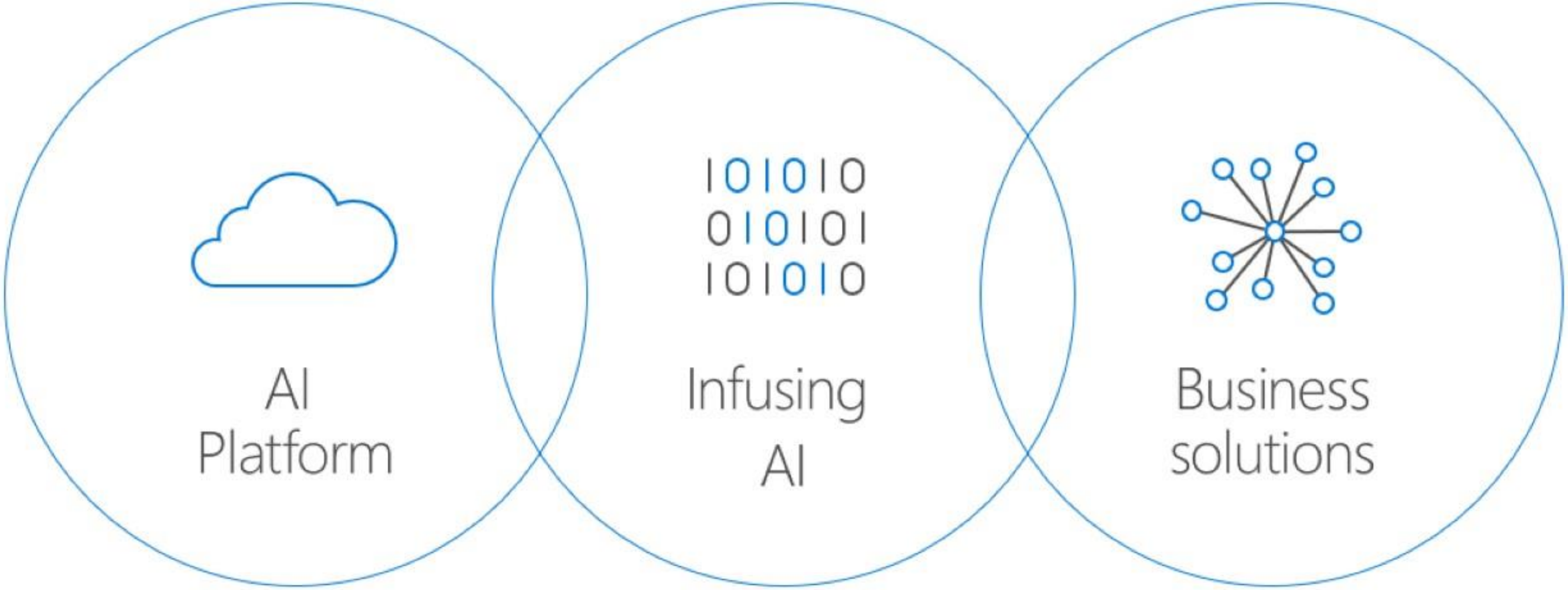
Reading

Comprehension Test

100,000+ Q&A Pairs

#1 SQuAD Leader

Microsoft AI Investment Areas

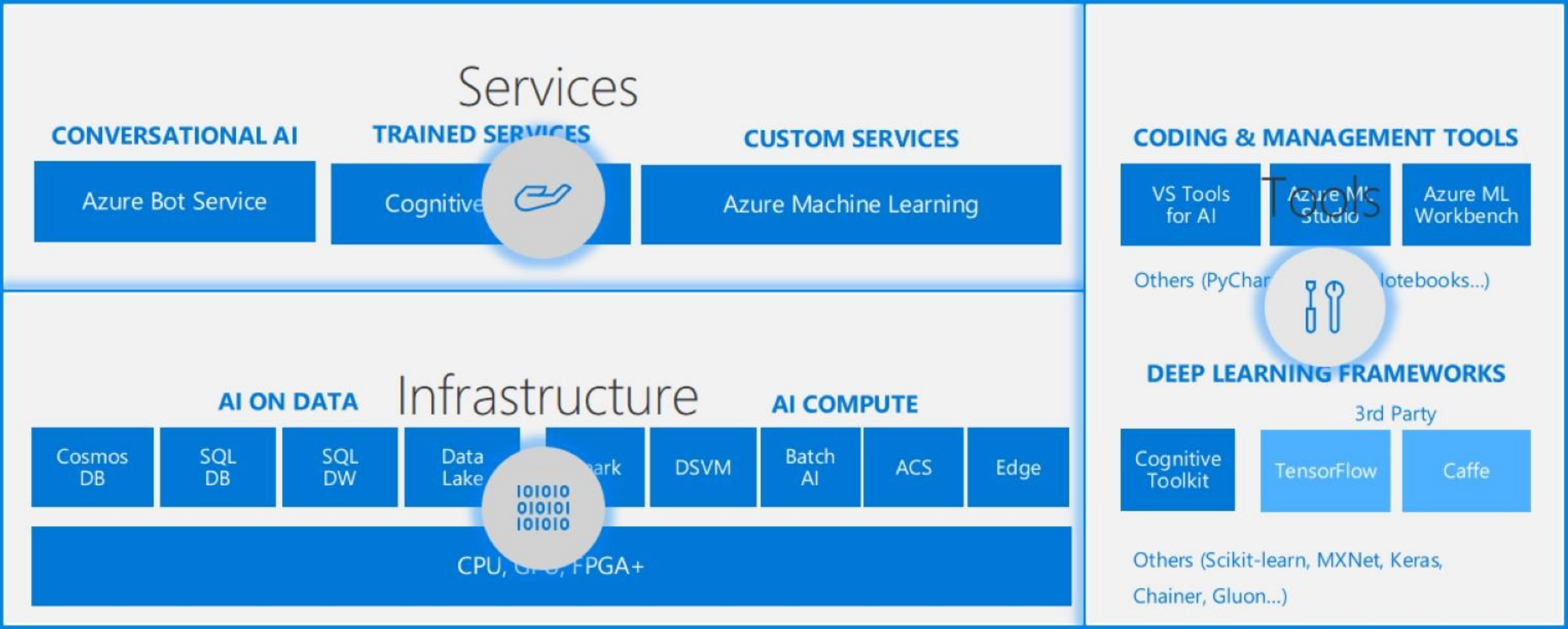


AI Platform

Infusing AI

Business solutions

The Microsoft AI platform



Services

CONVERSATIONAL AI

Azure Bot Service

TRAINED SERVICES

Cognitive

CUSTOM SERVICES

Azure Machine Learning

Infrastructure

AI ON DATA

Cosmos DB

SQL DB

SQL DW

Data Lake

ark

AI COMPUTE

DSVM

Batch AI

ACS

Edge

CPU, GPU, FPGA+

Tools

CODING & MANAGEMENT TOOLS

VS Tools for AI

Azure ML Studio

Azure ML Workbench

Others (PyChar lotebooks...)

DEEP LEARNING FRAMEWORKS

3rd Party

Cognitive Toolkit

TensorFlow

Caffe

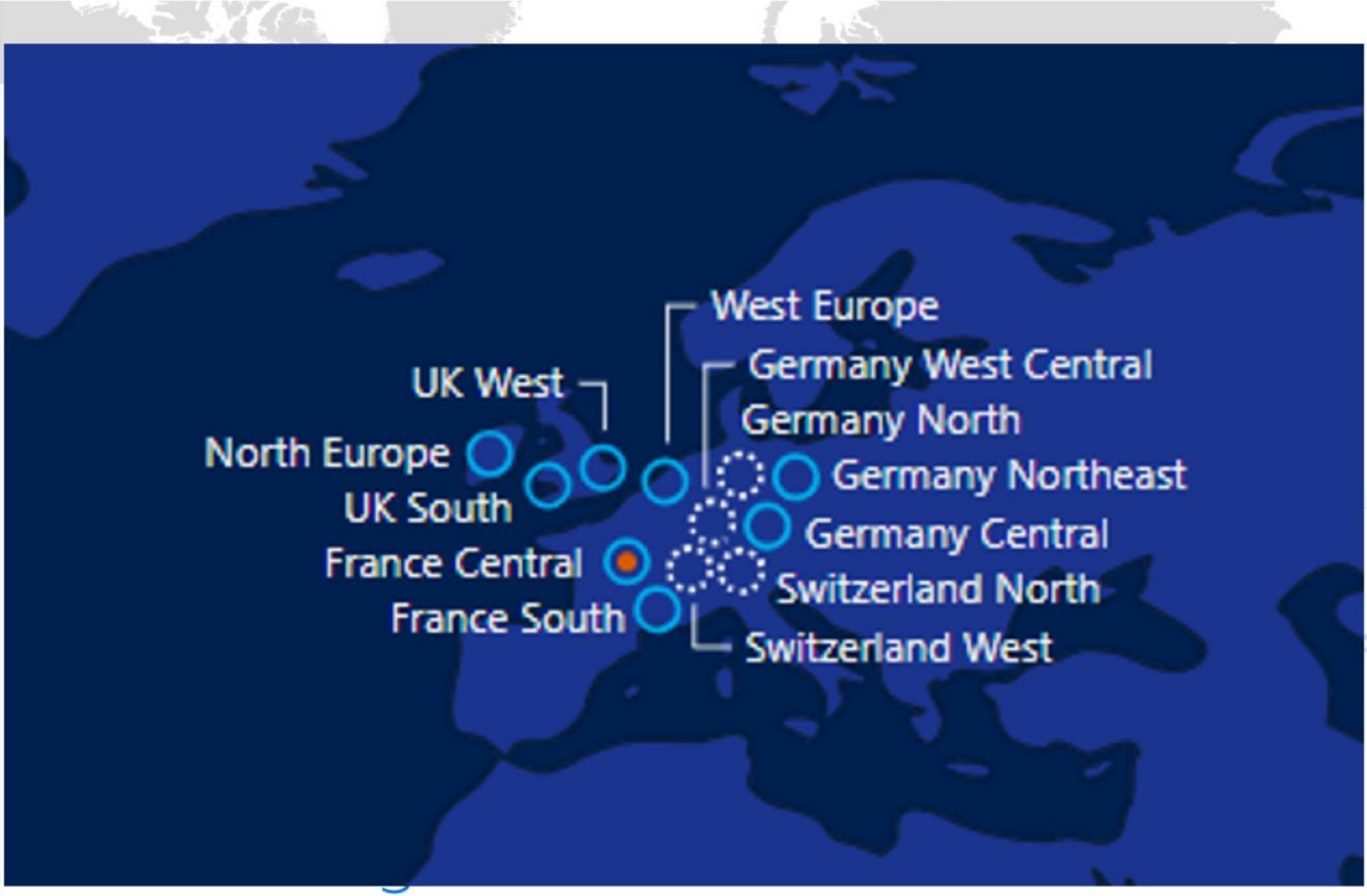
Others (Scikit-learn, MXNet, Keras, Chainer, Gluon...)

Productive

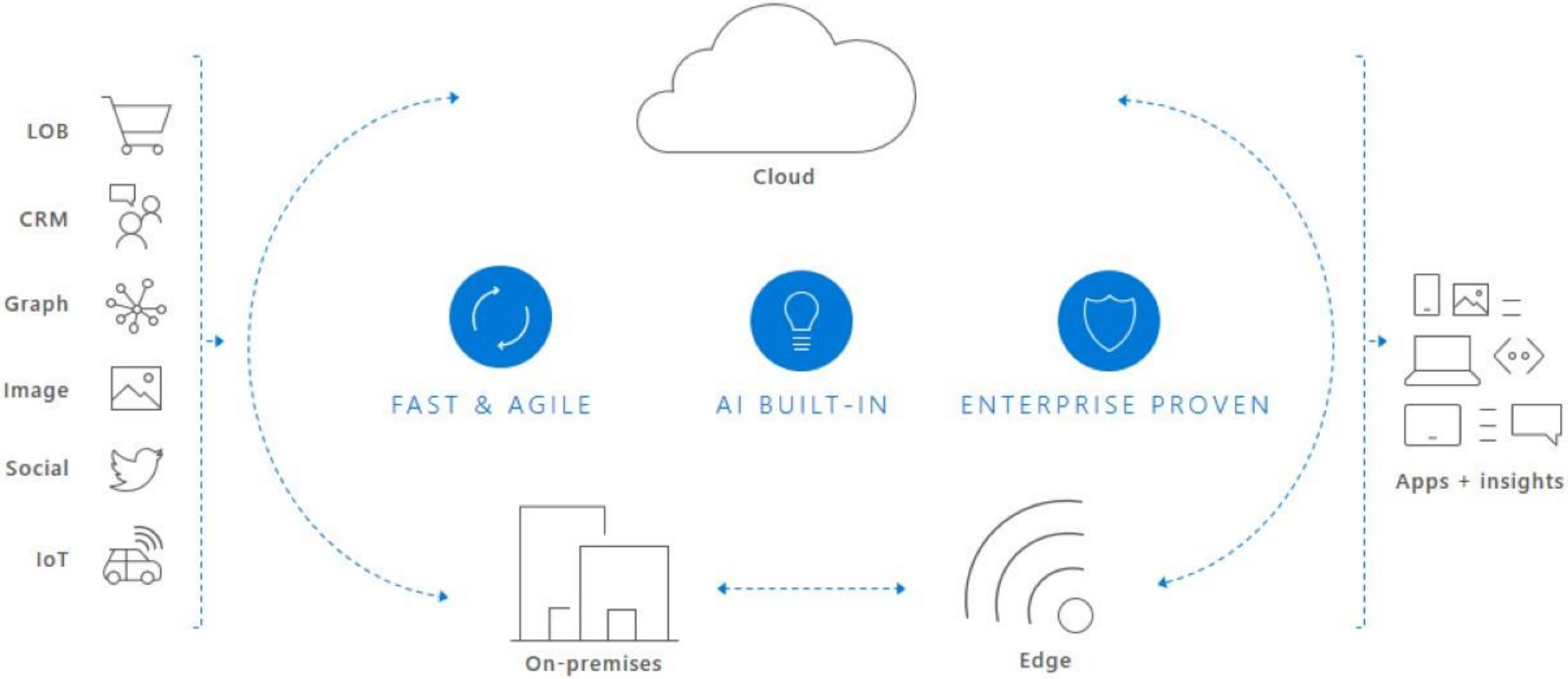
Hybrid

Intelligent

Trusted

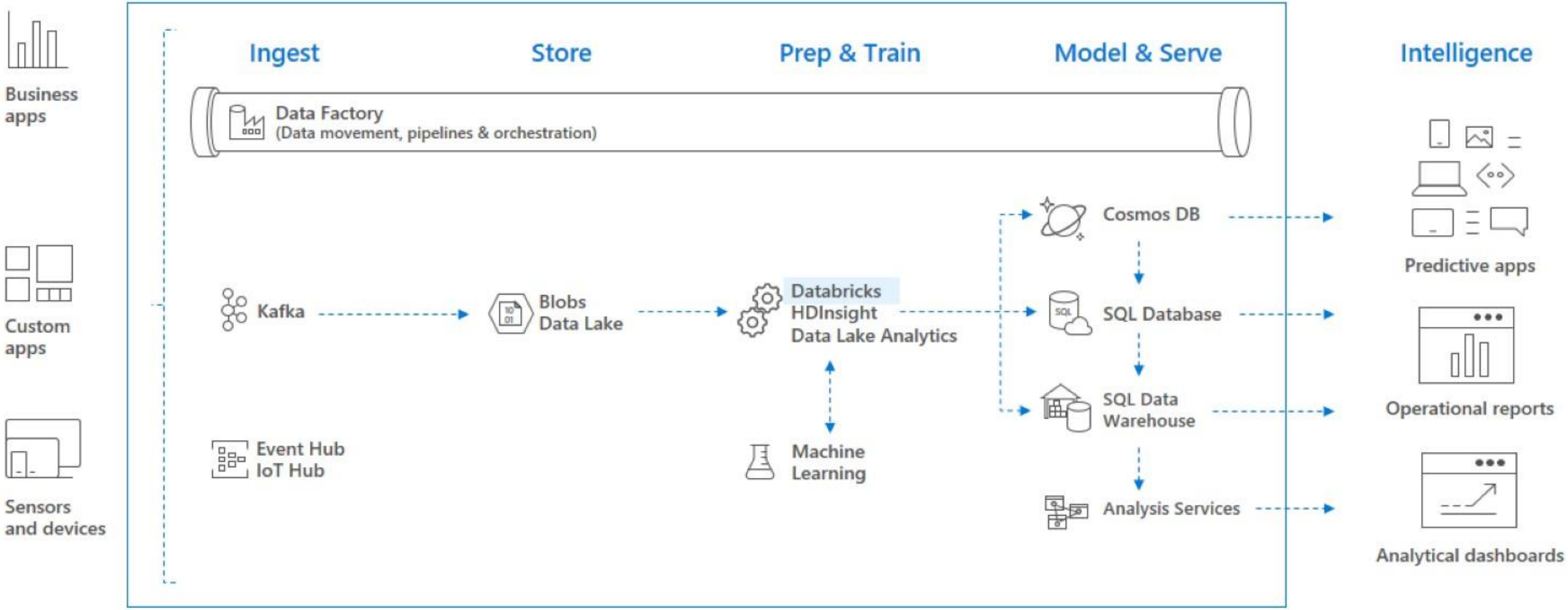


MICROSOFT AI PLATFORM



Big Data & Advanced Analytics in Azure

BIG DATA & ADVANCED ANALYTICS AT A GLANCE



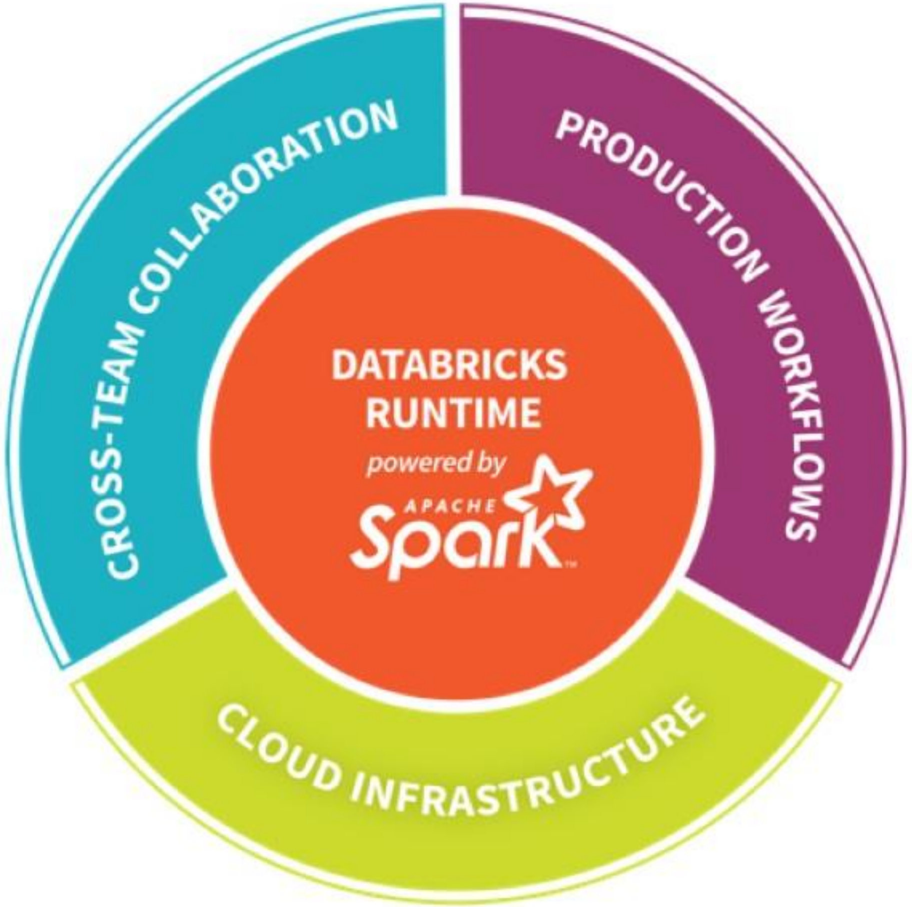
Azure Databricks
Powered by Apache Spark

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)

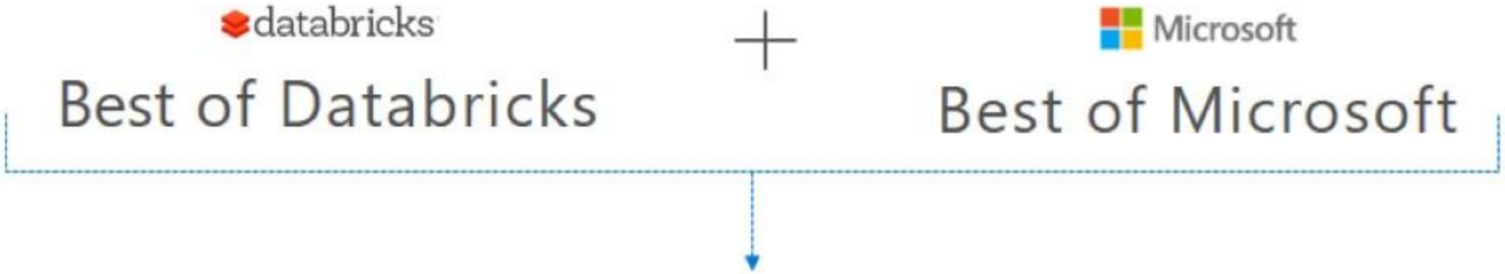
DATABRICKS - COMPANY OVERVIEW

- Founded in late 2013
- By the creators of Apache Spark, original team from UC Berkeley AMPLab
- Largest code contributor code to Apache Spark
- Level 2/3 support partnership with
  - Hortonworks
  - MapR
  - DataStax
- Provides certifications such as Databricks Certified Application, Databricks Certified Distribution and Databricks Certified Developer
- Main Product: The Unified Analytics Platform
- In Oct 2017, introduced Databricks Delta (currently in private preview).



What is Azure Databricks?

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



databricks

Best of Databricks

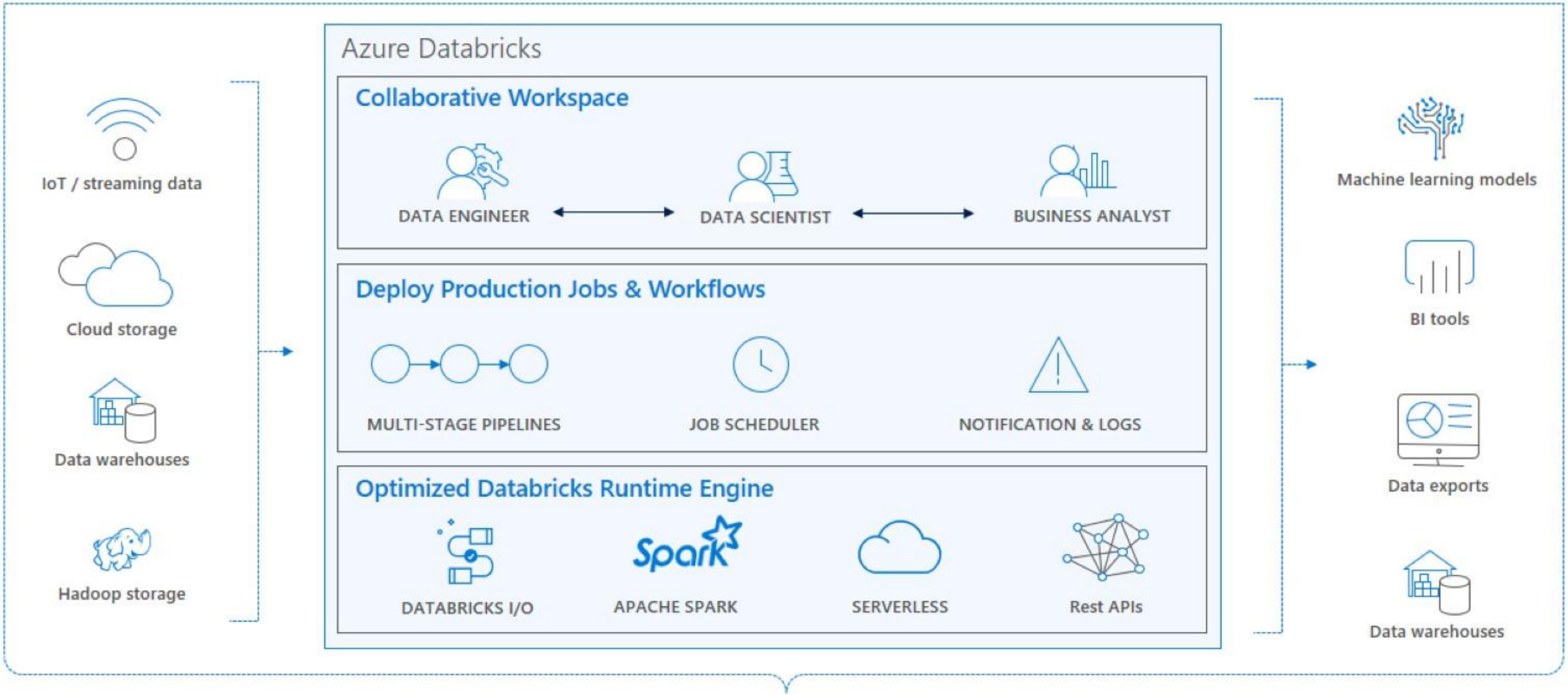
+

Microsoft

Best of Microsoft

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

Azure Databricks



IoT / streaming data

Cloud storage

Data warehouses

Hadoop storage

Azure Databricks

Collaborative Workspace

DATA ENGINEER ⟷ DATA SCIENTIST ⟷ BUSINESS ANALYST

Deploy Production Jobs & Workflows

MULTI-STAGE PIPELINES

JOB SCHEDULER

NOTIFICATION & LOGS

Optimized Databricks Runtime Engine

DATABRICKS I/O

APACHE SPARK

SERVERLESS

Rest APIs

Machine learning models

BI tools

Data exports

Data warehouses

Enhance Productivity

Build on secure & trusted cloud

Scale without limits

Collaborative Workspace

GET STARTED IN SECONDS

Single click to launch your new Spark environment

INTERACTIVE EXPLORATION

Explore data using interactive notebooks with support for multiple programming languages including R, Python, Scala, and SQL

COLLABORATION

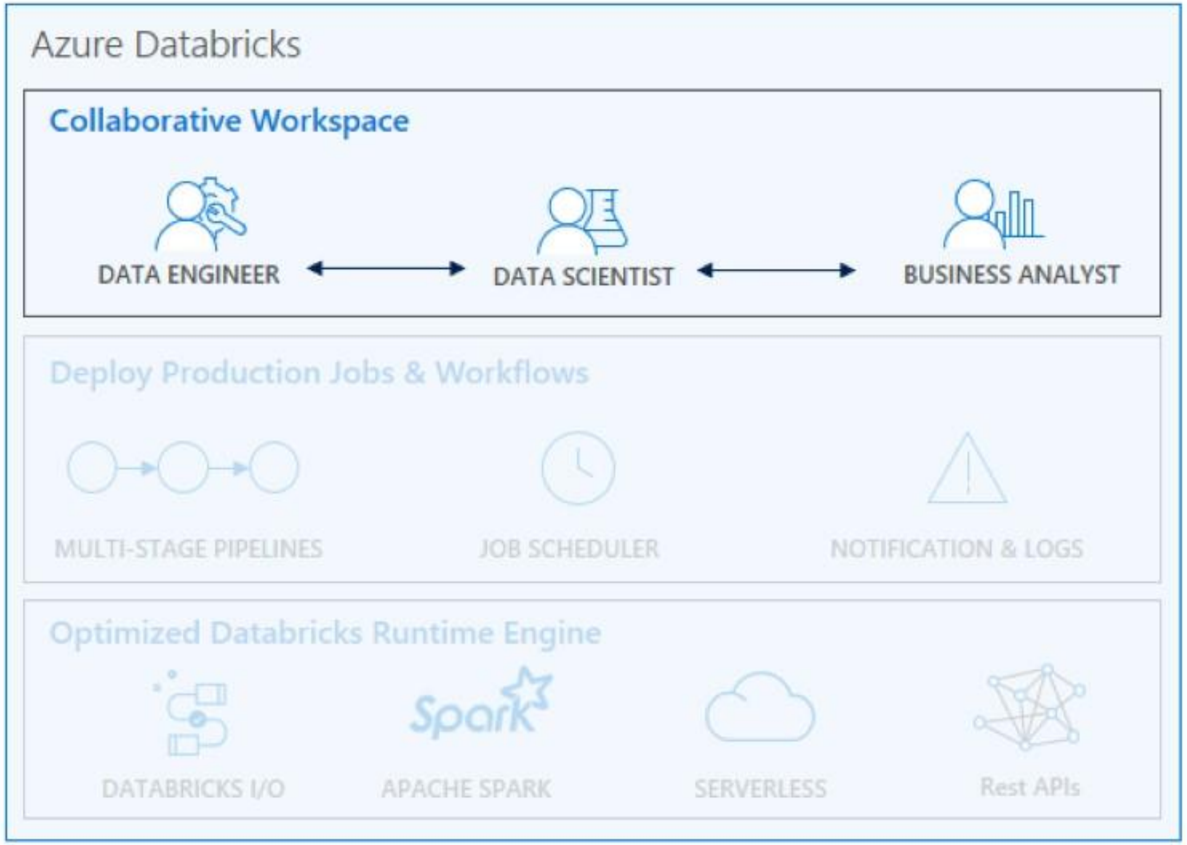
Work on the same notebook in real-time while tracking changes with detailed revision history, GitHub, or Bitbucket

VISUALIZATIONS

Visualize insights through a wide assortment of point-and-click visualizations. Or use powerful scriptable options like matplotlib, ggplot, and D3

DASHBOARDS

Rich integration with PowerBI to discover and share your insights in powerful new ways



Deploy Production Jobs & Workflows

JOBS SCHEDULER

Execute jobs for production pipelines on a specific schedule

NOTEBOOK WORKFLOWS

Create multi-stage pipelines with the control structures of the source programming language

RUN NOTEBOOKS AS JOBS

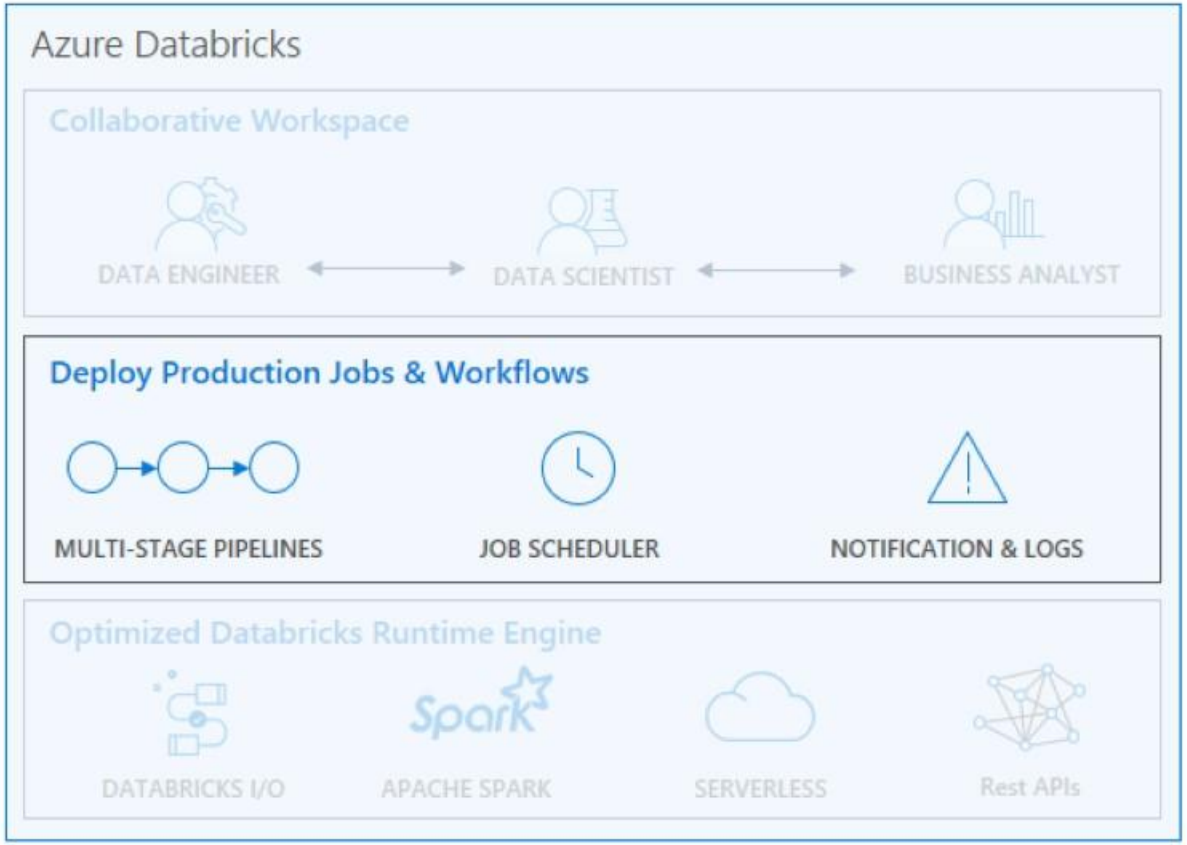
Turn notebooks or JARs into resilient Spark jobs with a click or an API call

NOTIFICATIONS AND LOGS

Set up alerts and quickly access audit logs for easy monitoring and troubleshooting

INTEGRATE NATIVELY WITH AZURE SERVICES

Deep integration with Azure SQL Data Warehouse, Cosmos DB, Azure Data Lake Store, Azure Blob Storage, and Azure Event Hub



Optimized Databricks Runtime Engine

OPTIMIZED I/O PERFORMANCE

The Databricks I/O module (DBIO) takes processing speeds to the next level — significantly improving the performance of Spark in the cloud

FULLY-MANAGED PLATFORM ON AZURE

Reap the benefits of a fully managed service and remove the complexity of big data and machine learning

SERVERLESS INFRASTRUCTURE

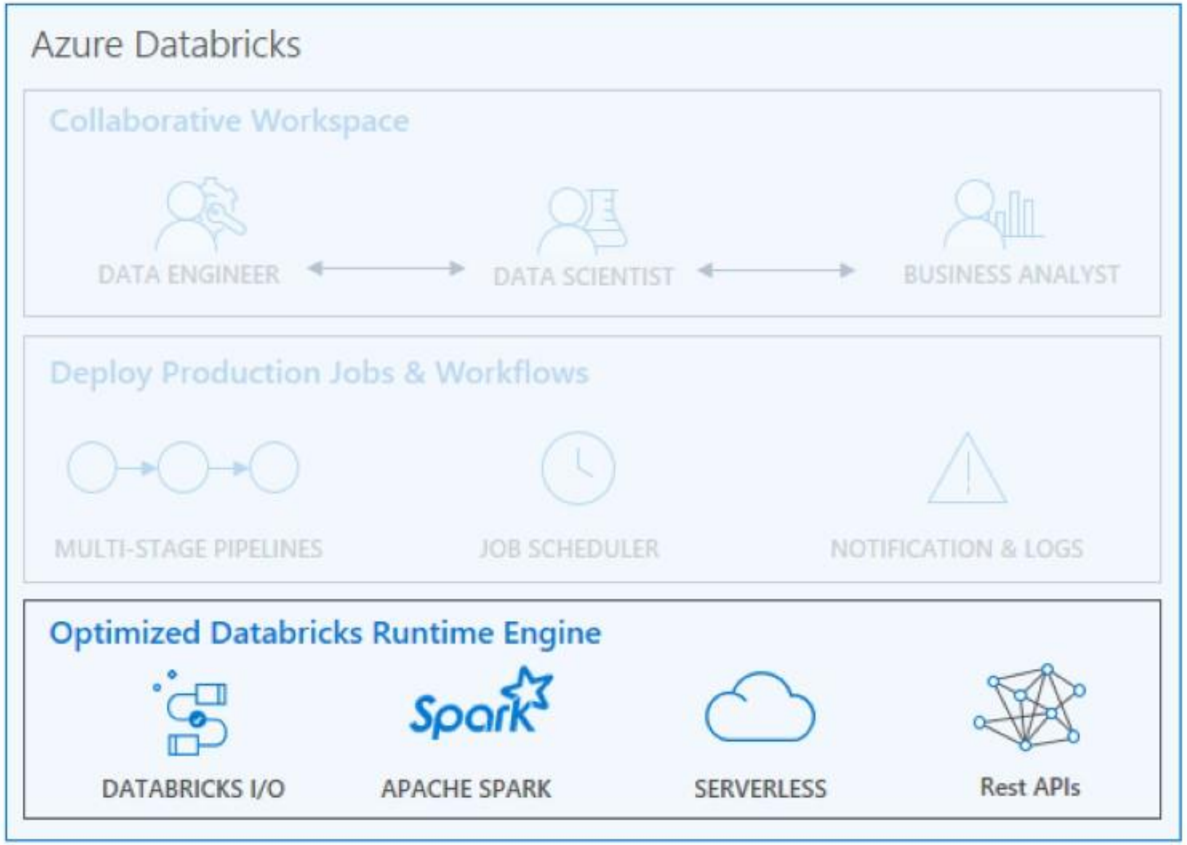
Databricks' serverless and highly elastic cloud service is designed to remove operational complexity while ensuring reliability and cost efficiency at scale

OPERATE AT MASSIVE SCALE

Without limits globally

SUPPORT FOR GPU ENABLED VMS

Specialized compute for your deep learning needs

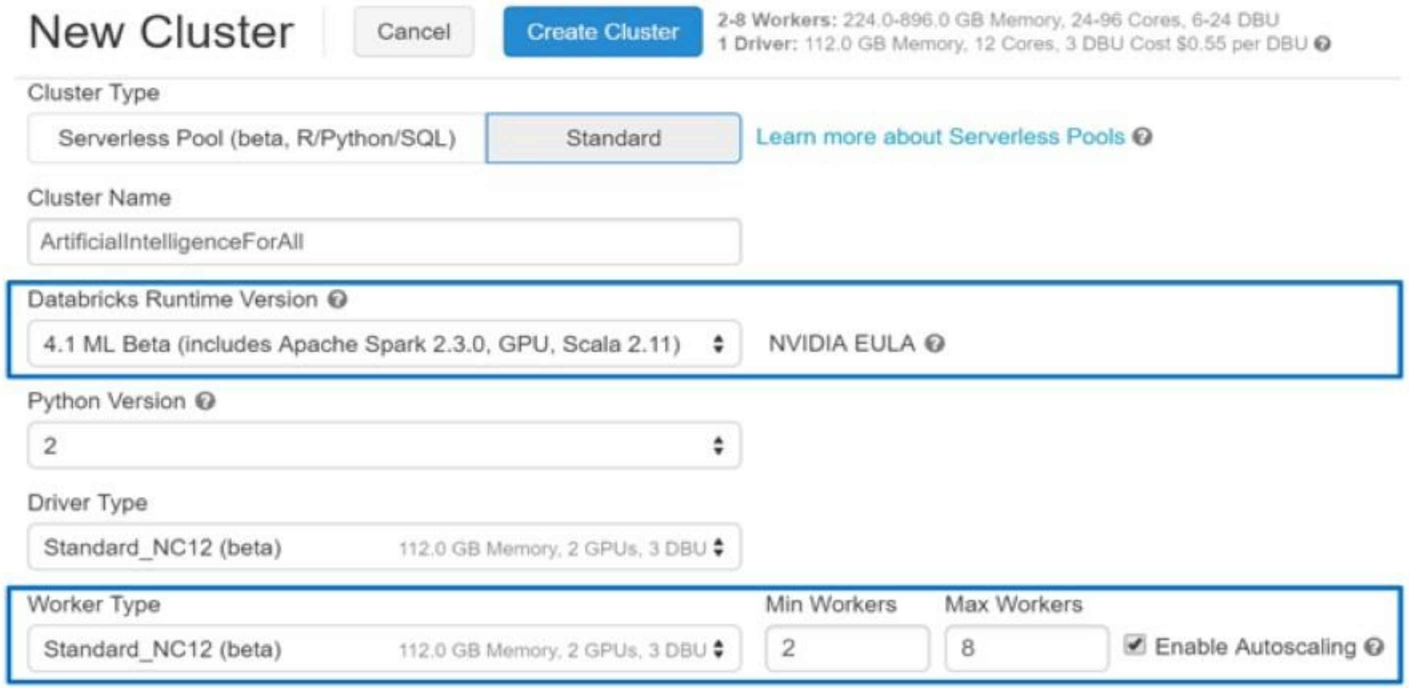


Azure Databricks Runtime for Machine Learning

AZURE Databricks Runtime for Machine Learning

- Pre-installed packages for machine learning like **Tensorflow, Keras, Horovod and XGBoost**
- Pre-configured HorovodEstimator for seamless integration of Horovod with the Spark DataFrames
- Support for GPU enabled VMs for specialized compute for your deep learning needs
- Multi-GPU trainings of deep neural networks using Horovod
- Unlock complex machine learning and deep learning scenarios with a few lines of code

AZURE DATABRICKS RUNTIME FOR MACHINE LEARNING



2-8 Workers: 224.0-896.0 GB Memory, 24-96 Cores, 6-24 DBU
1 Driver: 112.0 GB Memory, 12 Cores, 3 DBU Cost $0.55 per DBU

Cluster Type: Serverless Pool (beta, R/Python/SQL) | Standard

Learn more about Serverless Pools

Cluster Name: ArtificialIntelligenceForAll

Databricks Runtime Version: 4.1 ML Beta (includes Apache Spark 2.3.0, GPU, Scala 2.11)

NVIDIA EULA

Python Version: 2

Driver Type: Standard_NC12 (beta) 112.0 GB Memory, 2 GPUs, 3 DBU

Worker Type: Standard_NC12 (beta) 112.0 GB Memory, 2 GPUs, 3 DBU

Min Workers: 2

Max Workers: 8

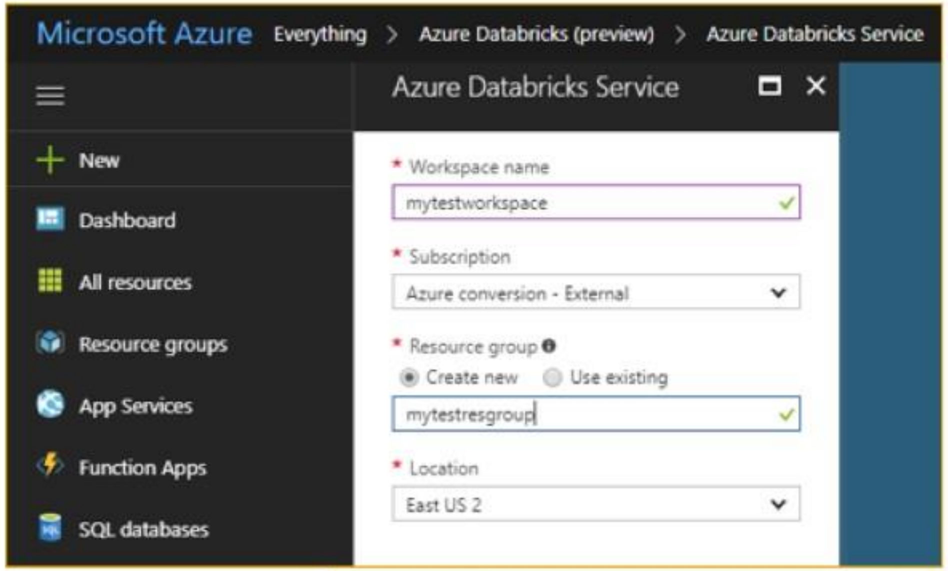
☑ Enable Autoscaling

Azure Databricks

Core Concepts

PROVISIONING AZURE DATABRICKS WORKSPACE

- Azure Databricks is provisioned directly from the Azure Portal like any other Azure service
  - In contrast, with other clouds, it has to be provisioned through the Databricks portal.
  - With Azure Databricks, the Azure Portal offers a unified portal to provision and administer Azure Databricks as well as other Azure services.
- Any Azure user with the appropriate subscription and authorization can provision Azure Databricks service*.
  - There is no need for a separate Databricks account



* Workspace name: mytestworkspace ✓

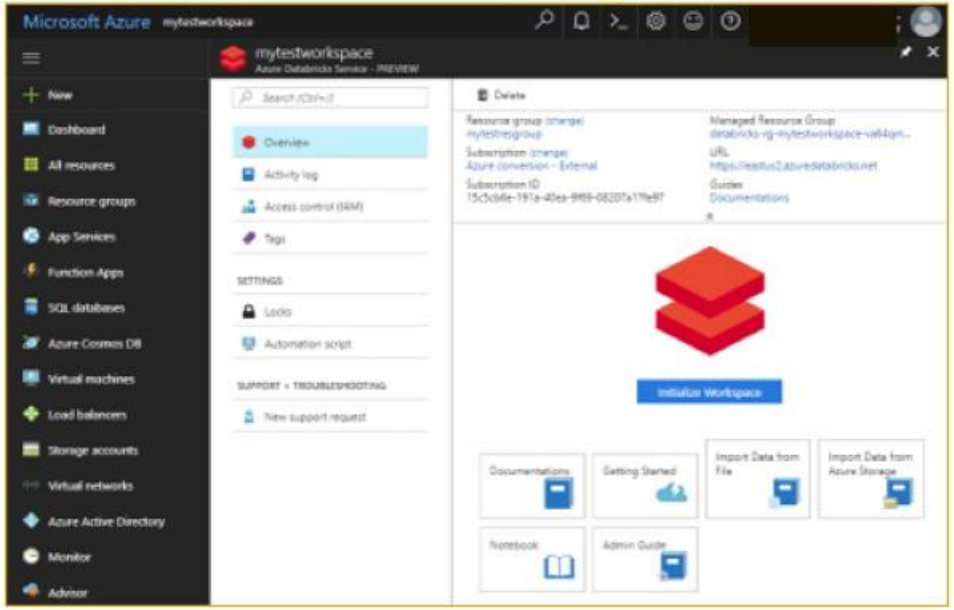
* Subscription: Azure conversion - External

* Resource group: ● Create new ○ Use existing

mytestresgroup ✓

* Location: East US 2

Provisioning the Azure Databricks Service



[redacted]

mytestworkspace
Azure Databricks Service - PREVIEW

Resource group (change): mytestresgroup

Subscription (change): Azure conversion - External

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

Managed Resource Group: databricks-rg-mytestworkspace-va64qm...

URL: https://eastus2.azuredatabricks.net

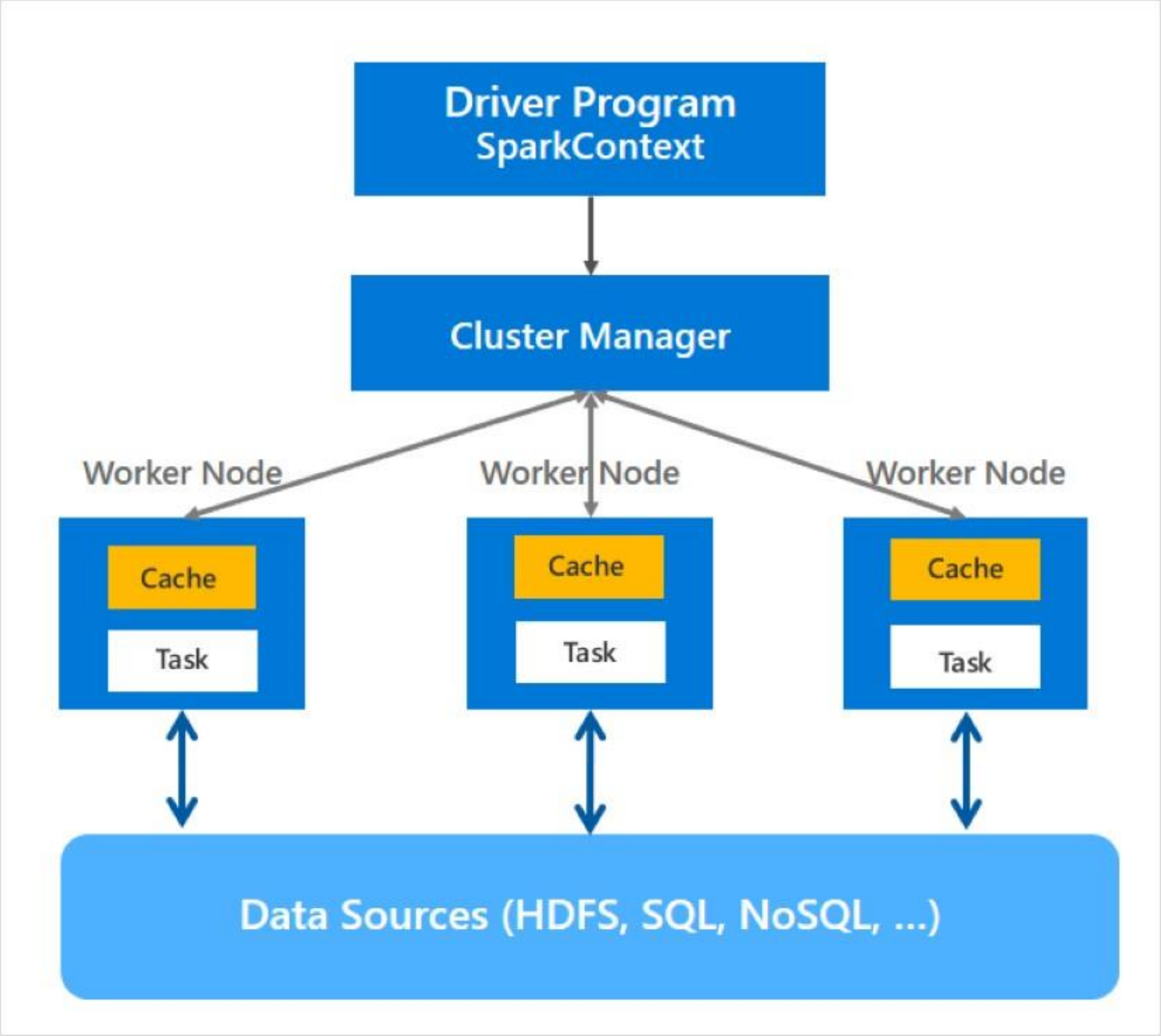
Guides: Documentations

After provisioning the is complete

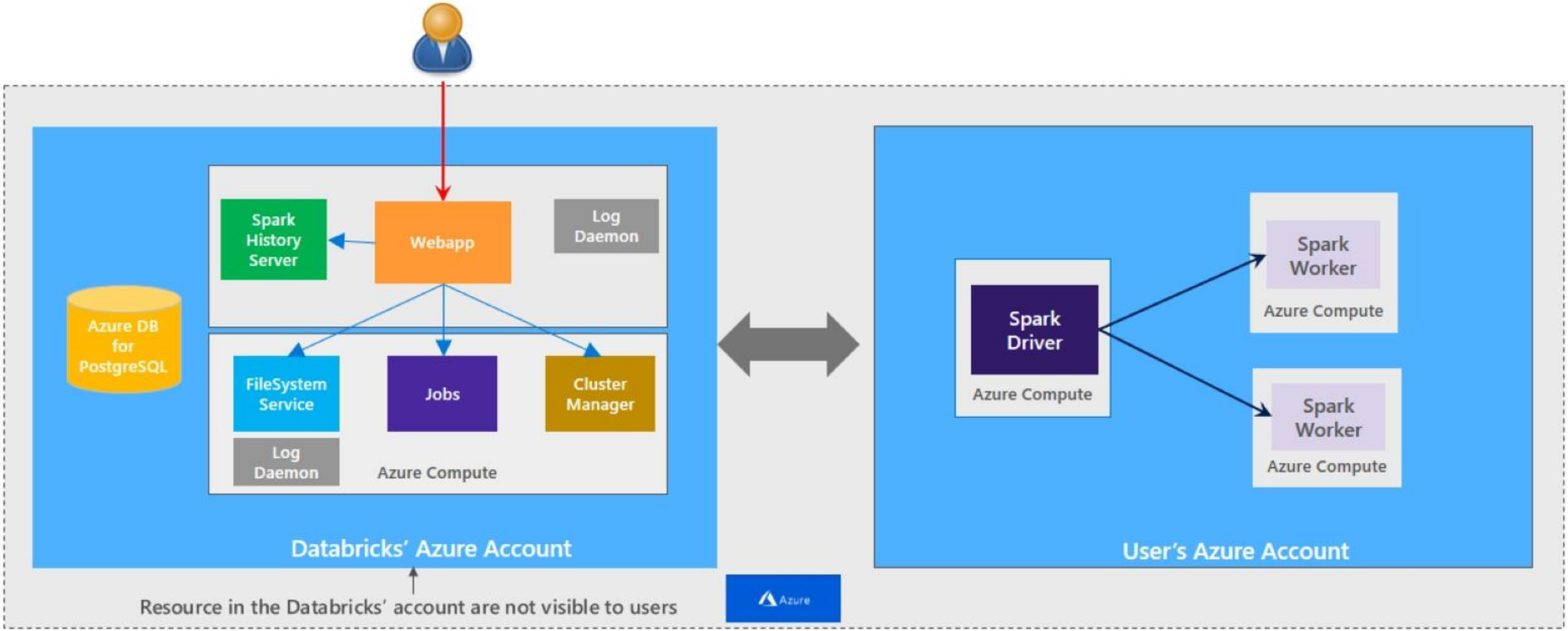
** During the current preview phase, the subscription has to be whitelisted.*

GENERAL SPARK CLUSTER ARCHITECTURE

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



AZURE DATABRICKS CLUSTER ARCHITECTURE

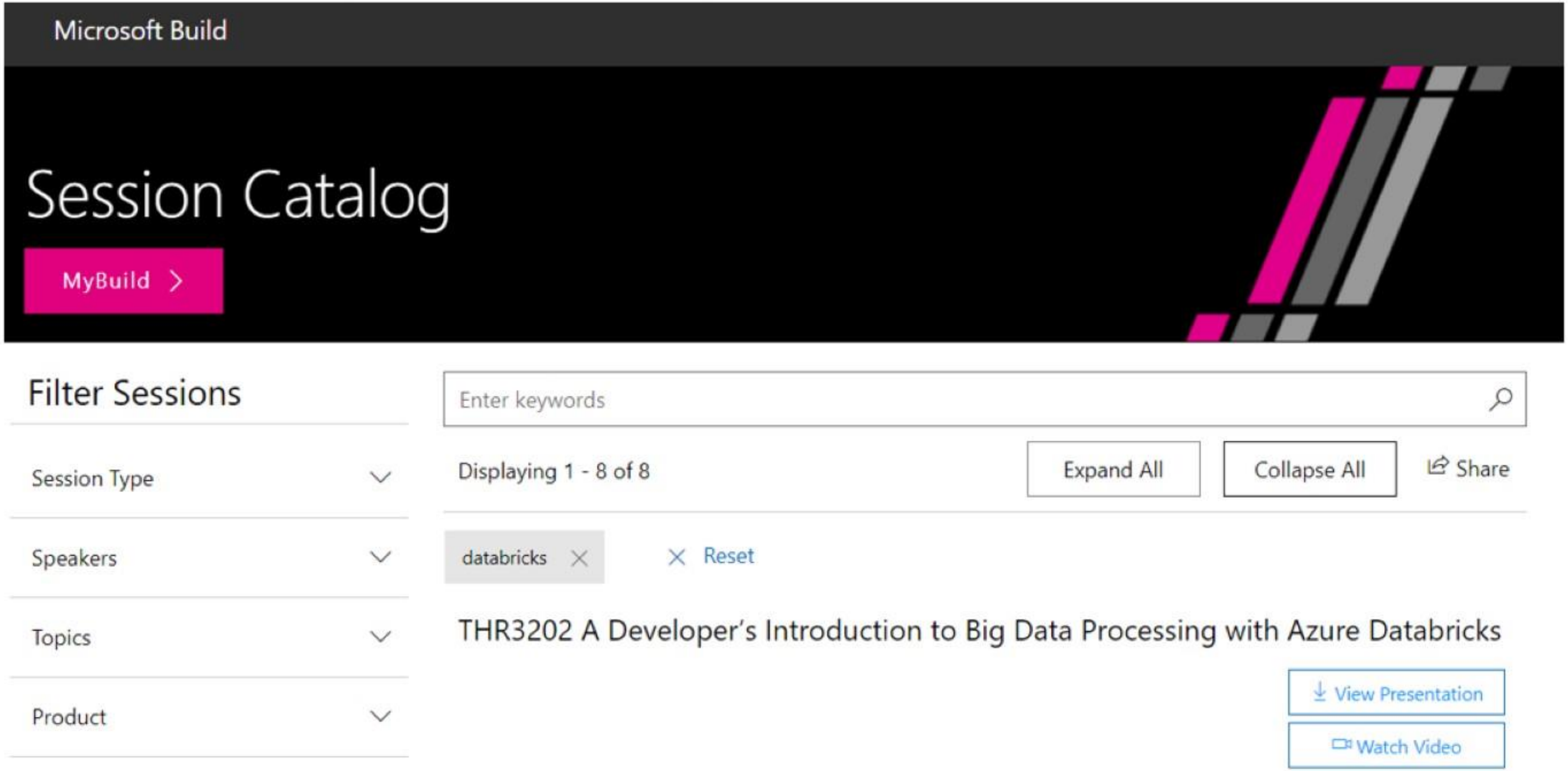


Resource in the Databricks' account are not visible to users

Demo

Továbblépés

https://mybuild.microsoft.com/sessions



Microsoft Build

Session Catalog

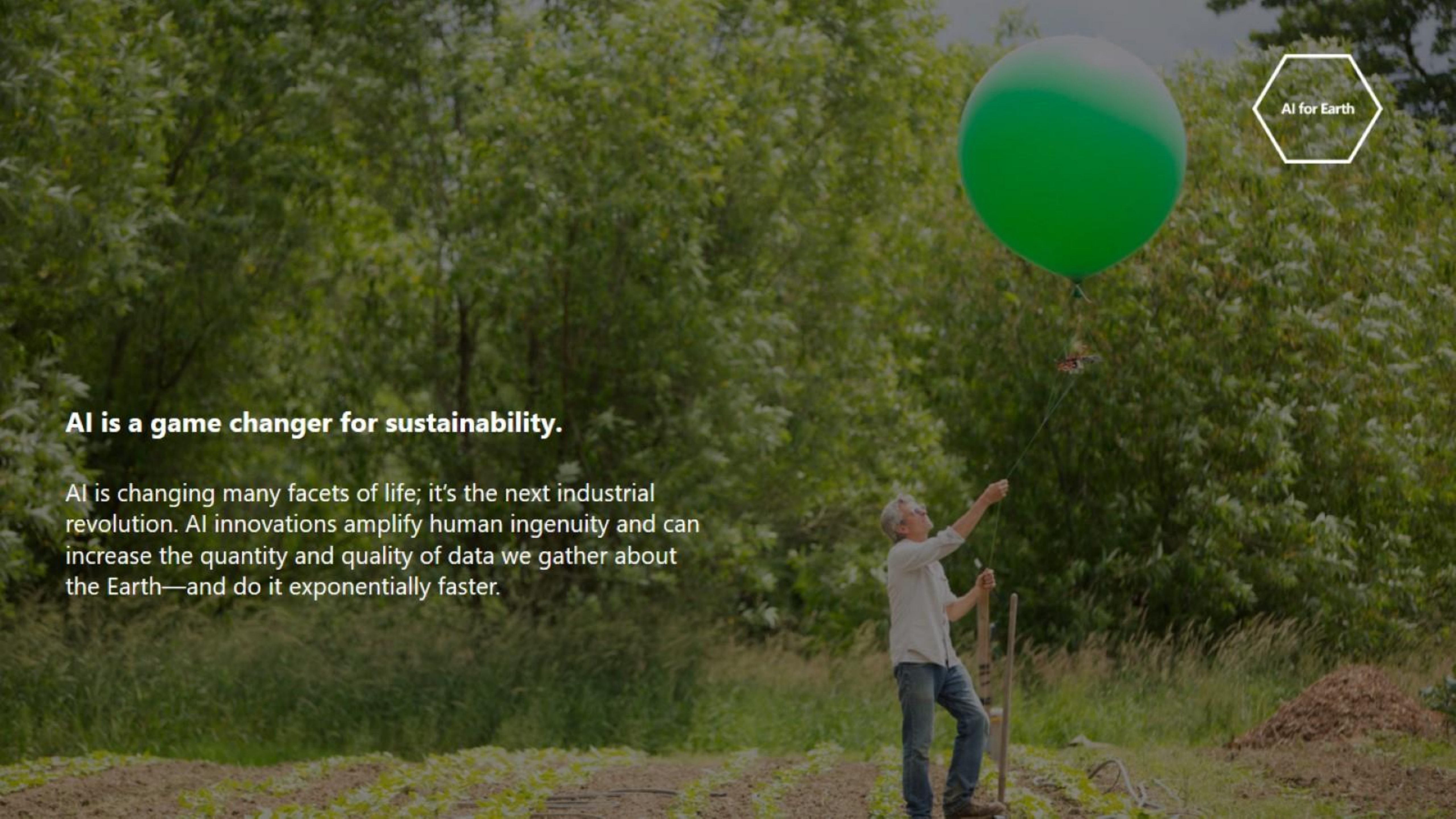
Filter Sessions

Enter keywords

Displaying 1 - 8 of 8

databricks

THR3202 A Developer's Introduction to Big Data Processing with Azure Databricks



AI for Earth

AI is a game changer for sustainability.

AI is changing many facets of life; it's the next industrial revolution. AI innovations amplify human ingenuity and can increase the quantity and quality of data we gather about the Earth—and do it exponentially faster.

The “AI for Earth” Vision

Microsoft invests $50 million in AI for Earth

Microsoft President Brad Smith makes announcement in Paris

AI for Earth

http://www.microsoft.com/AIforEarth

AGRICULTURE

In order to feed the world's rapidly growing population, farmers must produce more food, on less arable land, and with lower environmental impact.

WATER

In less than two decades, demand for fresh water (for human consumption, agriculture and hygiene) is projected to dramatically outpace supply.

BIODIVERSITY

Species are going extinct beyond the natural rate by orders of magnitude, driving the decay of key ecosystem services, like pollination, that humans depend upon.

CLIMATE CHANGE

An increasingly variable climate, extreme weather events, rising sea levels, higher global temperatures, and increased ocean acidity threaten human health, infrastructure, and the natural systems we rely on for life itself.

Microsoft