Introduction to Azure Databricks

James Serra

Big Data Evangelist

Microsoft

JamesSerra3@gmail.com

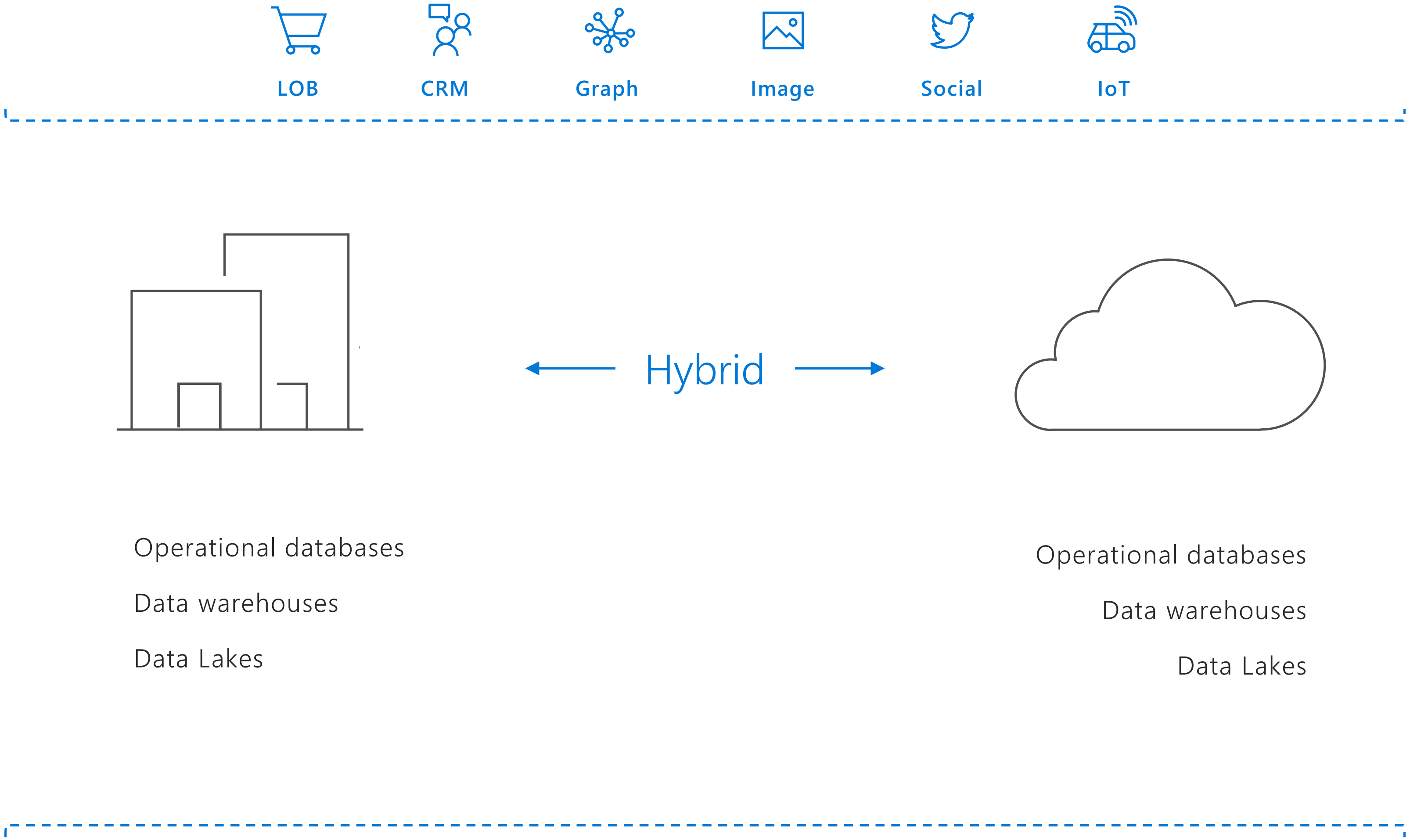
About Me

- Microsoft, Big Data Evangelist
- In IT for 30 years, worked on many BI and DW projects
- Worked as desktop/web/database developer, DBA, BI and DW architect and developer, MDM architect, PDW/APS developer
- Been perm employee, contractor, consultant, business owner
- Presenter at PASS Business Analytics Conference, PASS Summit, Enterprise Data World conference
- Certifications: MCSE: Data Platform, Business Intelligence; MS: Architecting Microsoft Azure Solutions, Design and Implement Big Data Analytics Solutions, Design and Implement Cloud Data Platform Solutions
- Blog at JamesSerra.com
- Former SQL Server MVP
- Author of book "Reporting with Microsoft SQL Server 2012"

Agenda

- Big Data Architectures
- Why data lakes?
- Top-down vs Bottom-up
- Data lake defined
- Hadoop as the data lake
- Modern Data Warehouse
- Federated Querying
- Solution in the cloud
- SMP vs MPP

THE MODERN DATA ESTATE



LOB

CRM

Graph

Image

Social

IoT

Hybrid

Operational databases

Data warehouses

Data Lakes

Operational databases

Data warehouses

Data Lakes

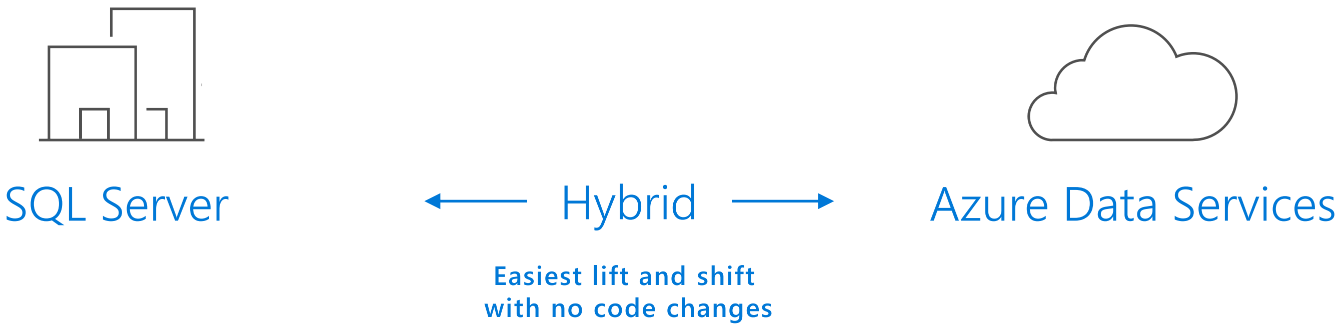
Reason over any data, anywhere

Flexibility of choice

Security and performance

THE MICROSOFT OFFERING

LOB CRM Graph Image Social IoT



SQL Server ← Hybrid → Azure Data Services

Easiest lift and shift
with no code changes

Industry leader 2 years in a row	Operational databases
#1 TPC-H performance	Data warehouses
T-SQL query over any data	Data lakes

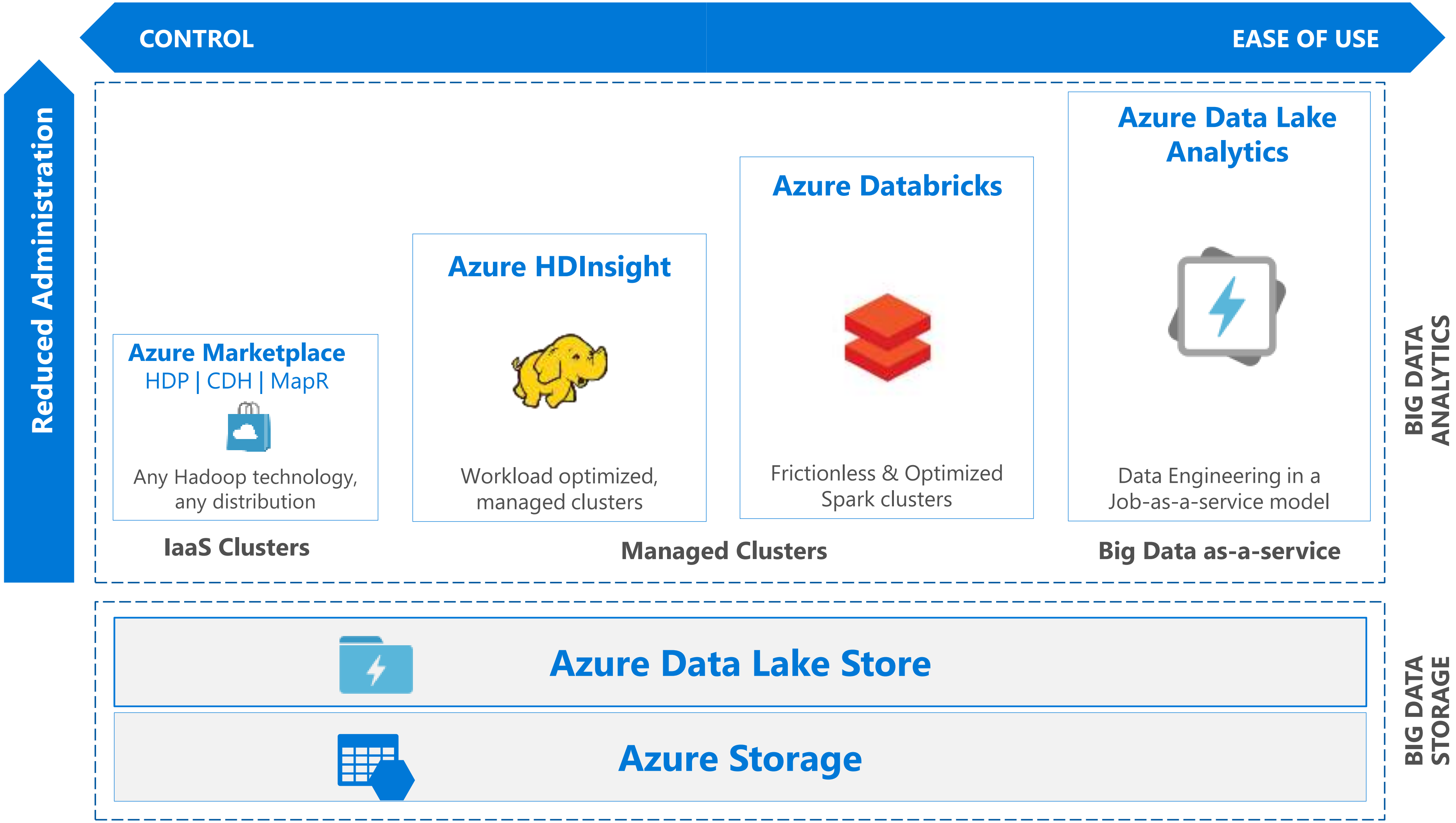
Operational databases	**70% faster than Aurora**
Data warehouses	**2x global reach than Redshift**
Data lakes	**No Limits Analytics with 99.9% SLA**

AI built-in | Most secure | Lowest TCO

Reason over any data, anywhere Flexibility of choice Security and performance

Big Data & Advanced Analytics in Azure

KNOWING THE VARIOUS BIG DATA SOLUTIONS



CONTROL

EASE OF USE

Reduced Administration

Azure Marketplace
HDP | CDH | MapR

Any Hadoop technology, any distribution

IaaS Clusters

Azure HDInsight

Workload optimized, managed clusters

Azure Databricks

Frictionless & Optimized Spark clusters

Managed Clusters

Azure Data Lake Analytics

Data Engineering in a Job-as-a-service model

Big Data as-a-service

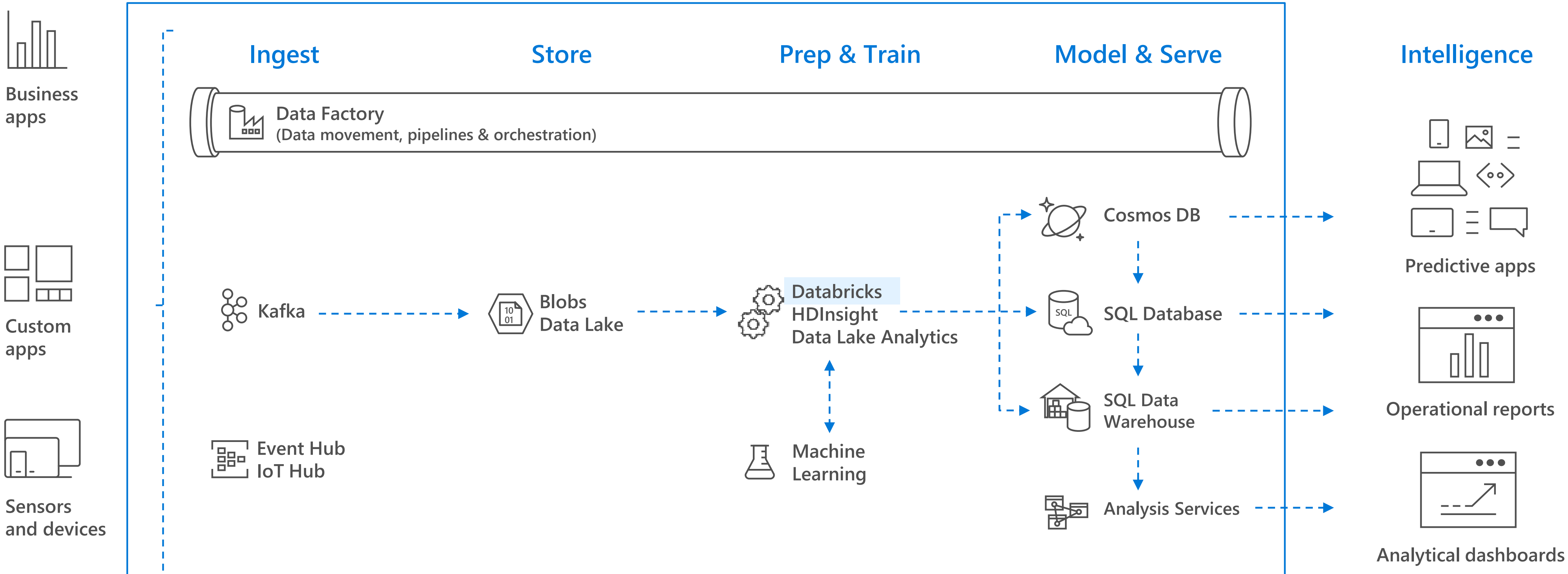
BIG DATA ANALYTICS

Azure Data Lake Store

Azure Storage

BIG DATA STORAGE

BIG DATA & ADVANCED ANALYTICS AT A GLANCE



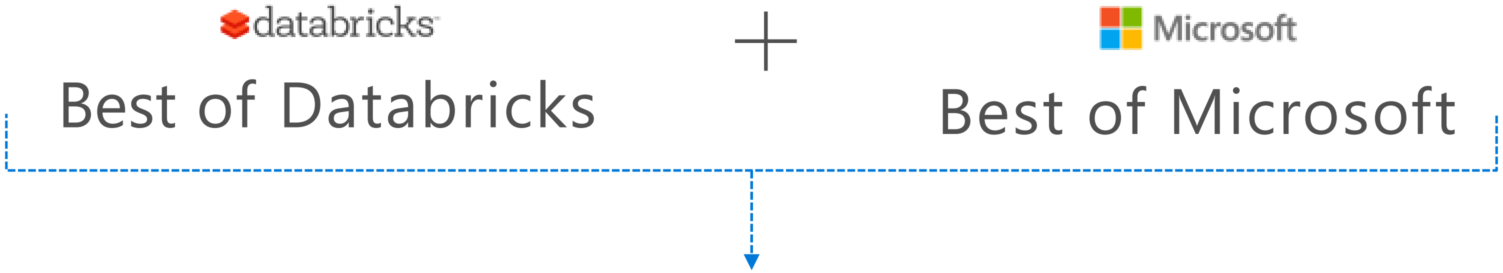
Azure Databricks
Powered by Apache Spark

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)

What is Azure Databricks?

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



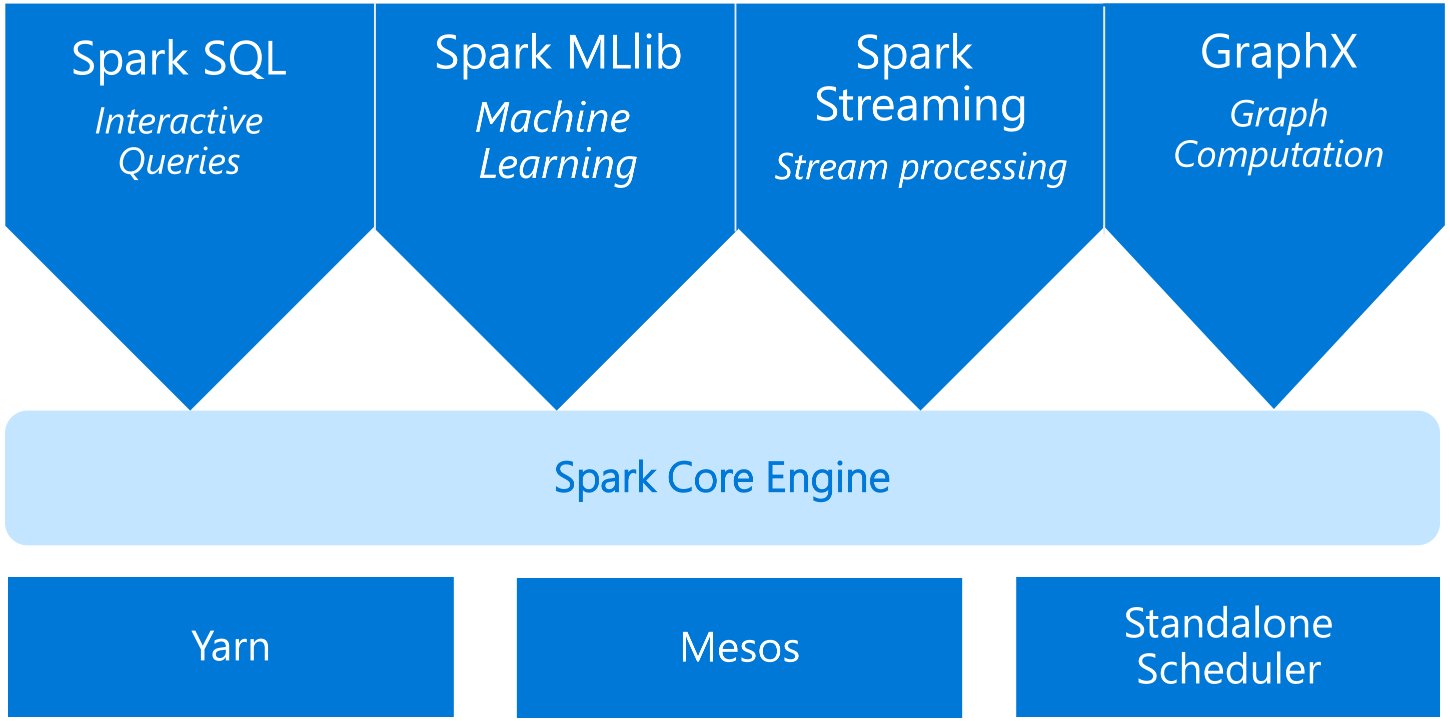
- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)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

APACHE SPARK

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

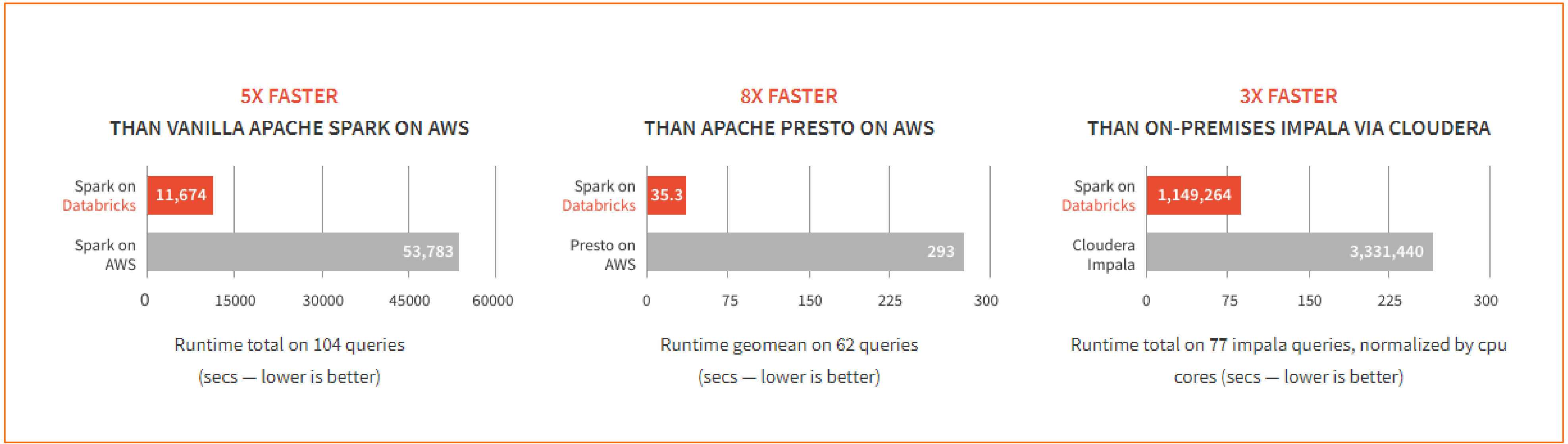
Spark Unifies:

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



DATABRICKS SPARK IS FAST

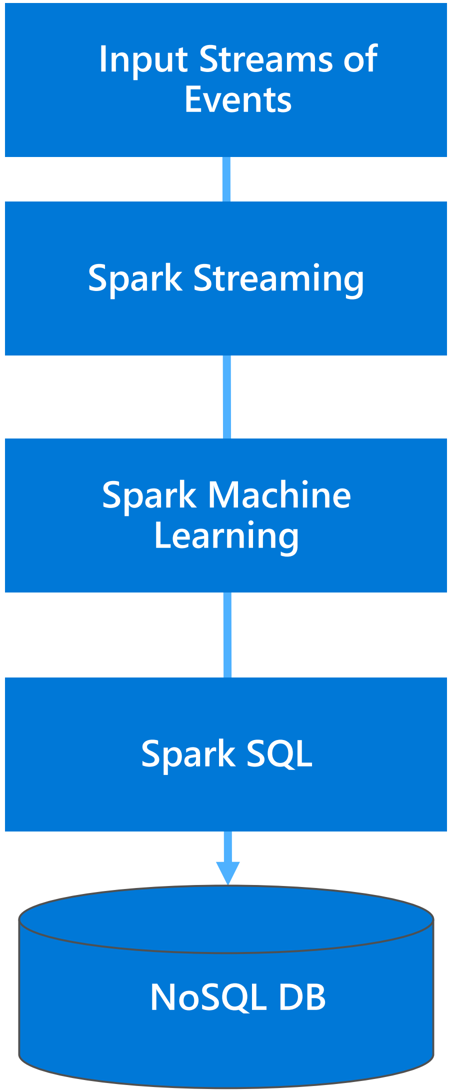
Benchmarks have shown Databricks to often have better performance than alternatives



SOURCE: Benchmarking Big Data SQL Platforms in the Cloud

ADVANTAGES OF A UNIFIED PLATFORM

- Improves developer productivity—a single consistent set of APIs
- All different systems in Spark share the same abstraction – RDDs (Resilient Distributed Datasets)
- Developers can mix and match different kind of processing in the same application. This is a common requirement for many big data pipelines.
- Performance improves because unnecessary movement of data across engines is eliminated. In many pipelines, data exchange between engines is the dominant cost



Differentiated experience on Azure

ENHANCE PRODUCTIVITY

Get started quickly by launching your new Spark environment with one click.

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

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.

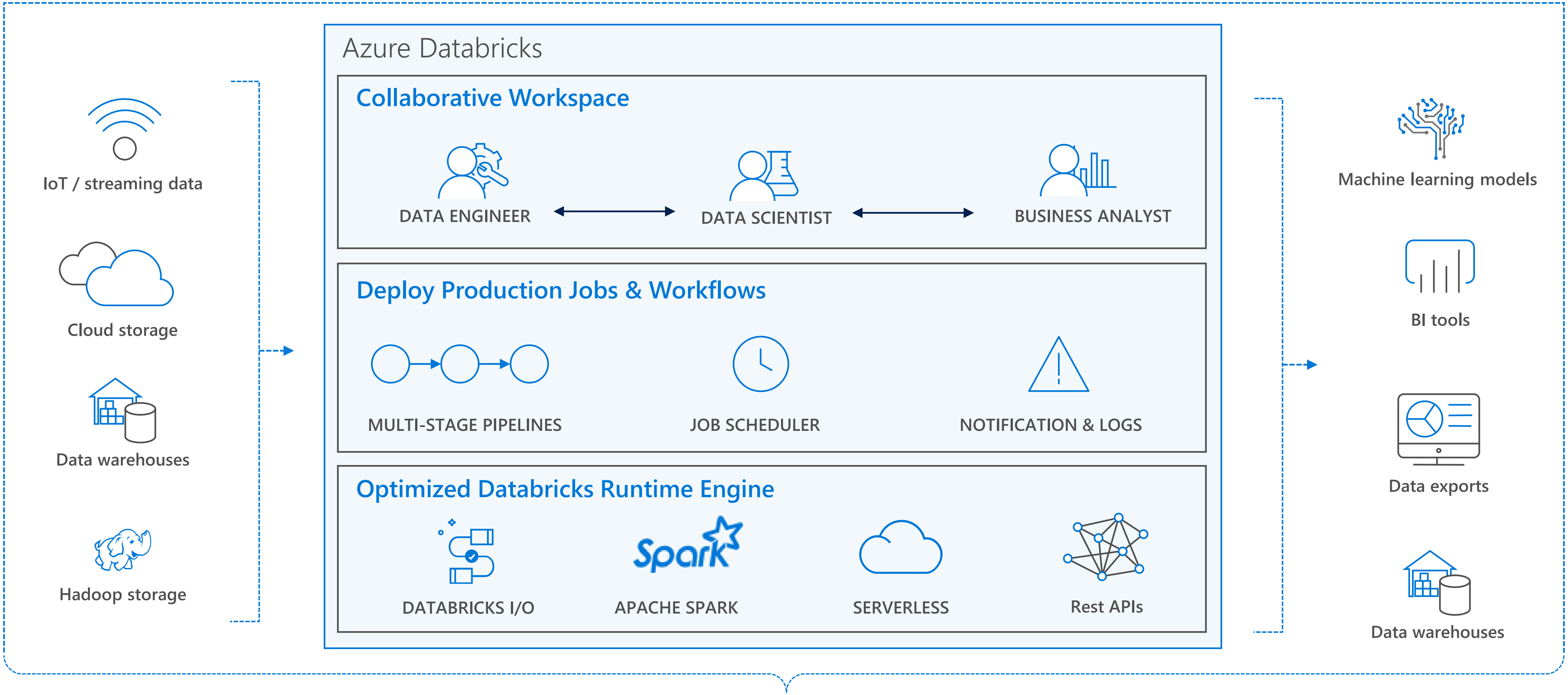
Build with confidence on the trusted cloud backed by unmatched support, compliance and SLAs.

SCALE WITHOUT LIMITS

Operate at massive scale without limits globally.

Accelerate data processing with the fastest Spark engine.

Azure Databricks



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

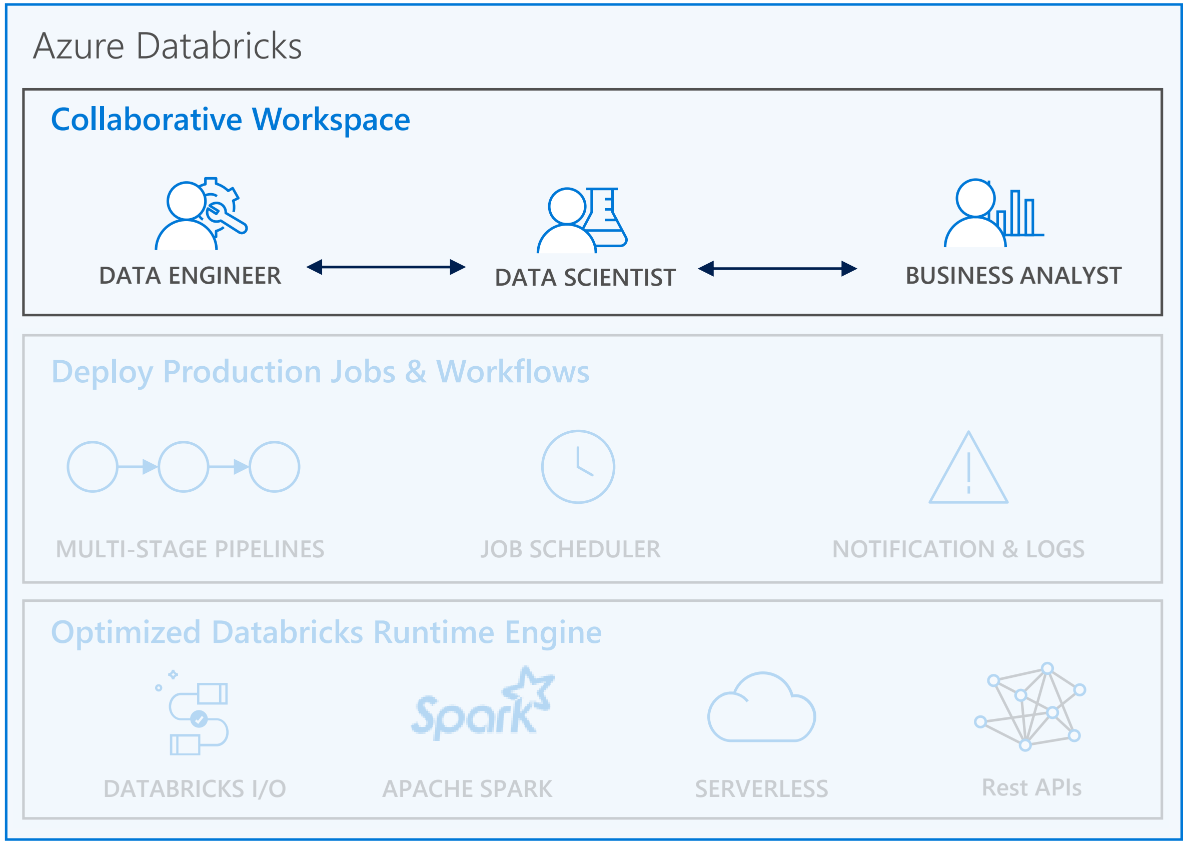
Work on the same notebook in real-time while tracking changes with detailed revision history, GitHub, or Bitbucket

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

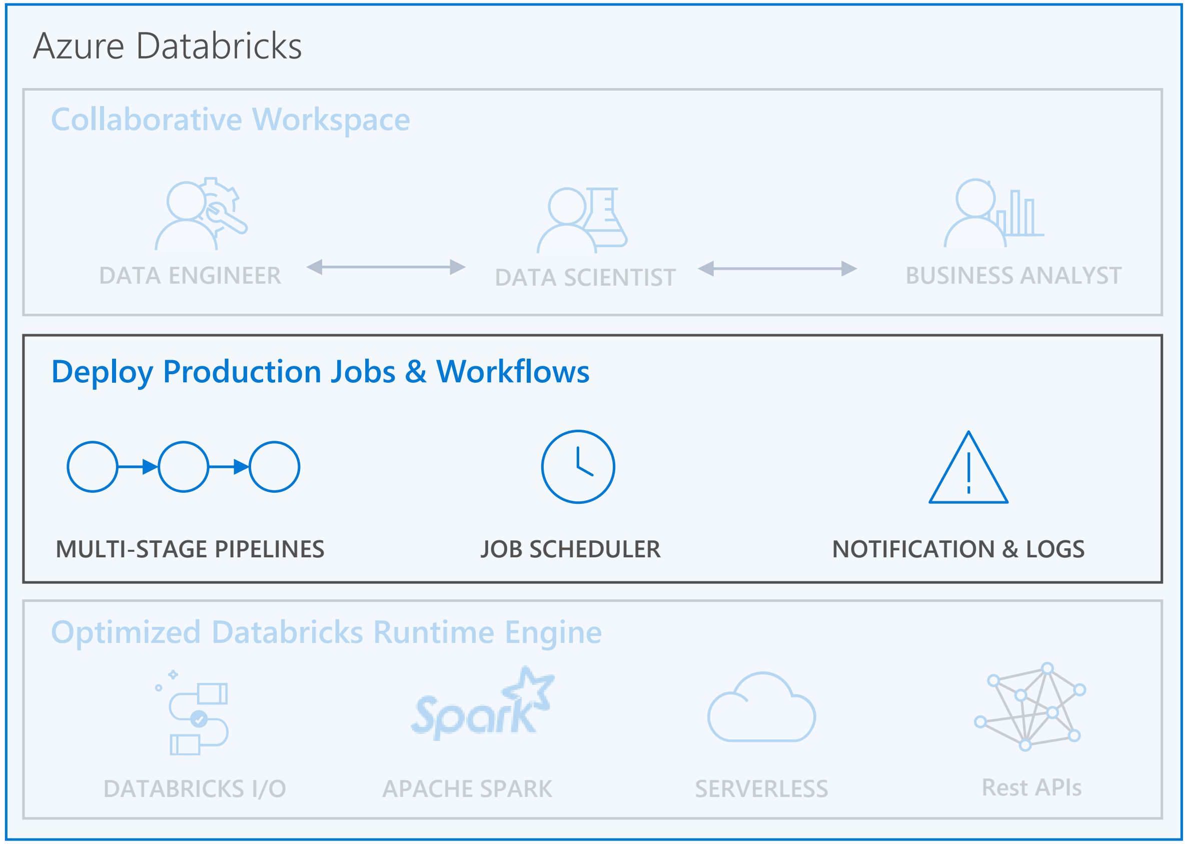
Turn notebooks or JARs into resilient Spark jobs with a click or an API call

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

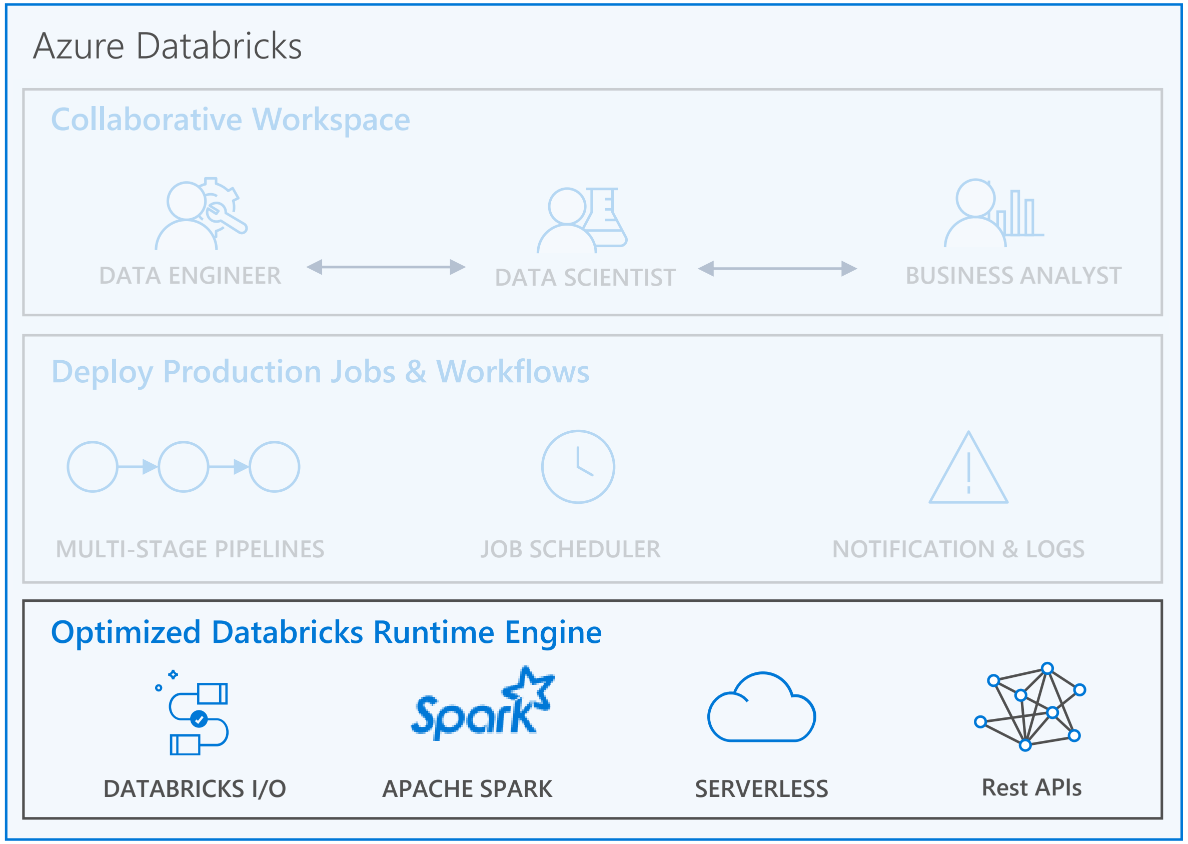
Reap the benefits of a fully managed service and remove the complexity of big data and machine learning

SERVERLESS INFRASTRUCTURE

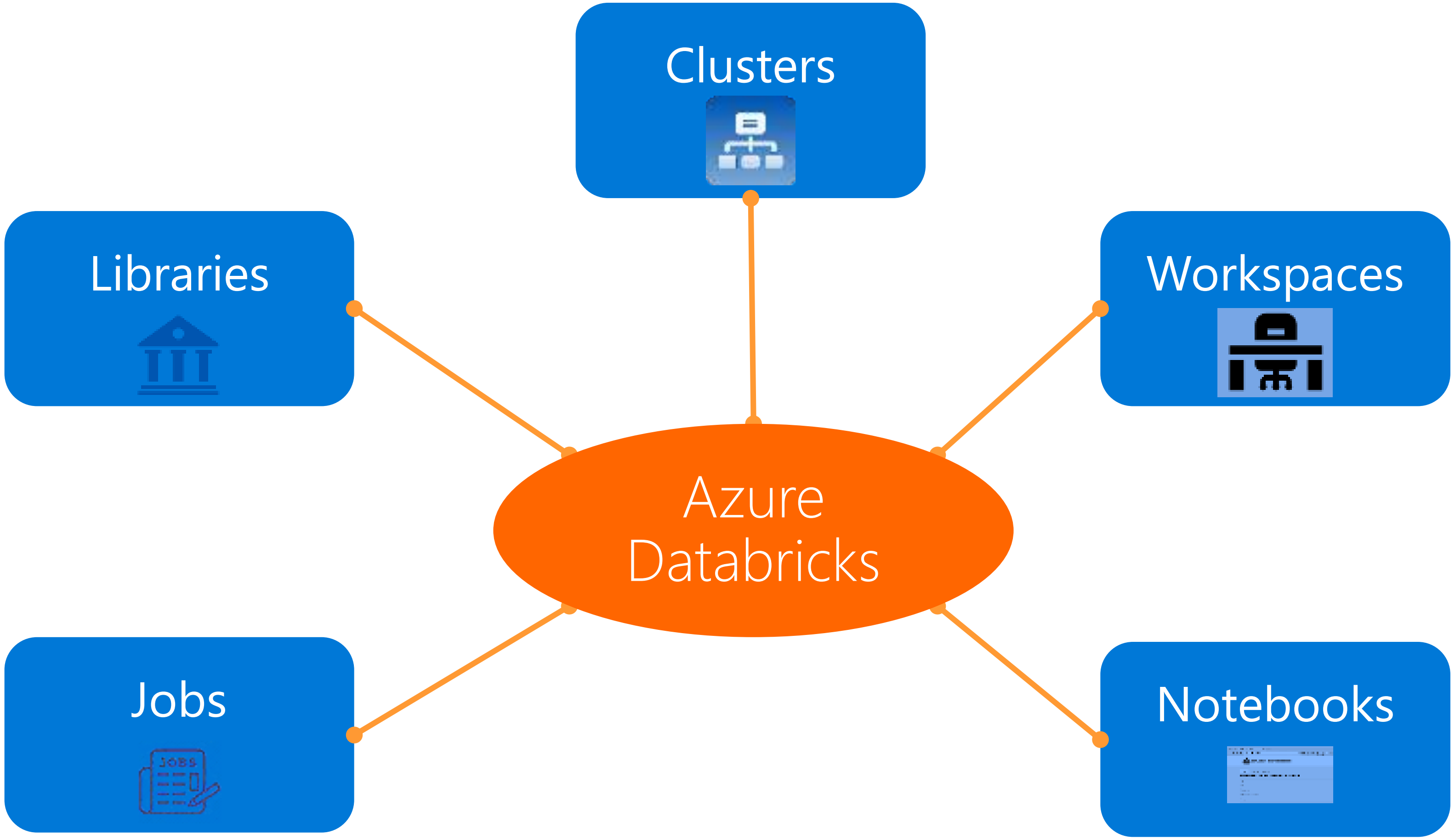
Databricks' serverless and highly elastic cloud service is designed to remove operational complexity while ensuring reliability and cost efficiency at scale

OPERATE AT MASSIVE SCALE

Without limits globally

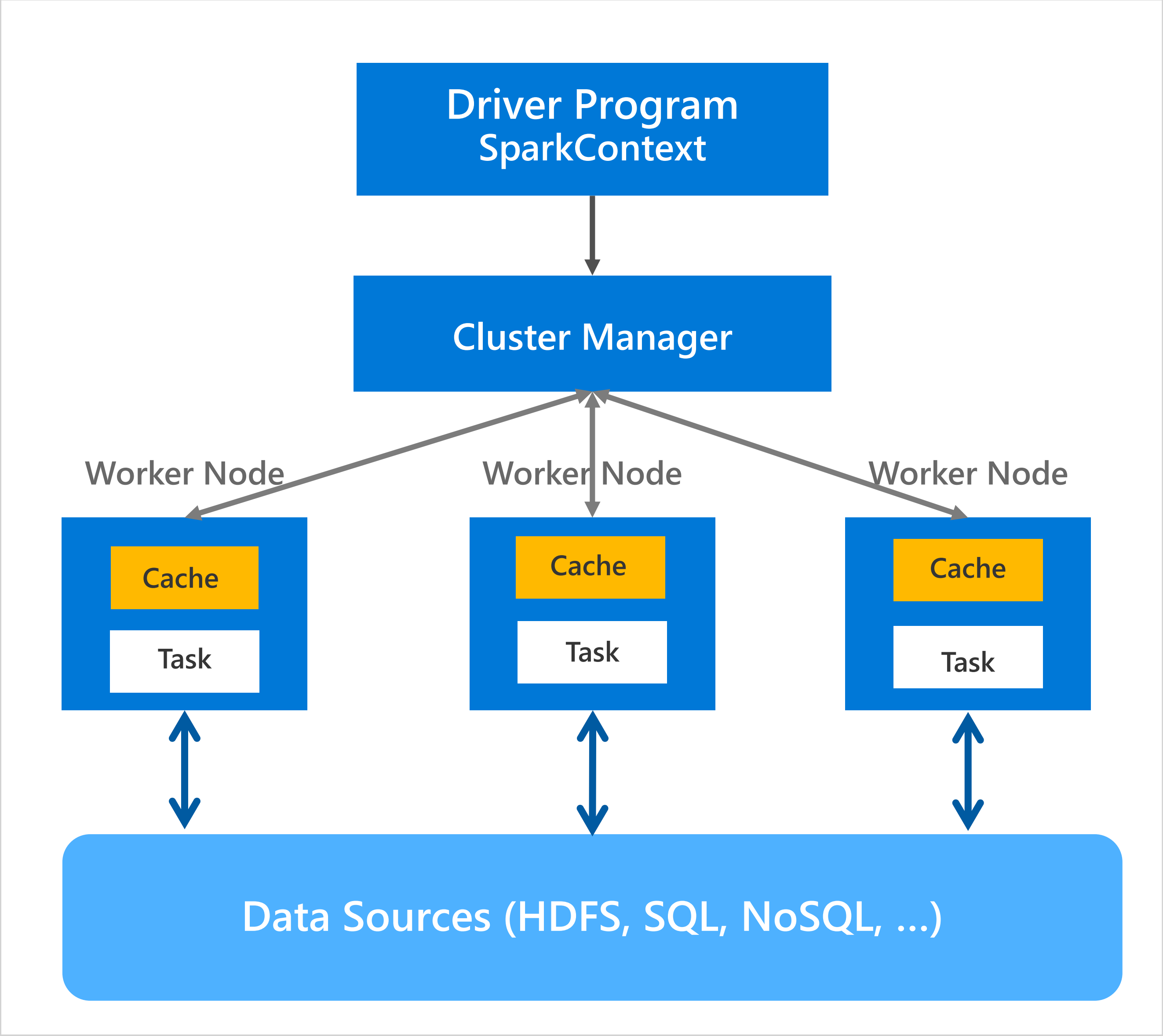


AZURE DATABRICKS CORE ARTIFACTS



GENERAL SPARK CLUSTER ARCHITECTURE

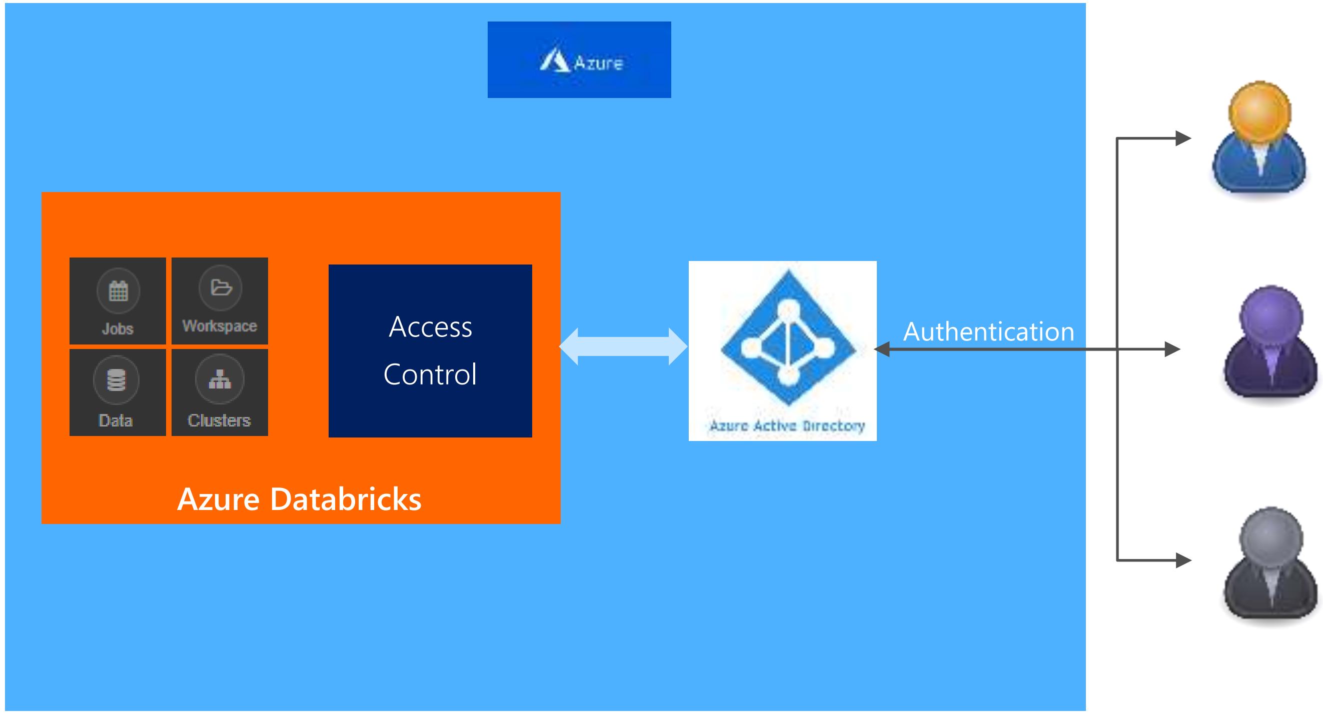
- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs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



AZURE DATABRICKS 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.



CLUSTERS: AUTO SCALING AND AUTO TERMINATION

Simplifies cluster management and reduces costs by eliminating wastage

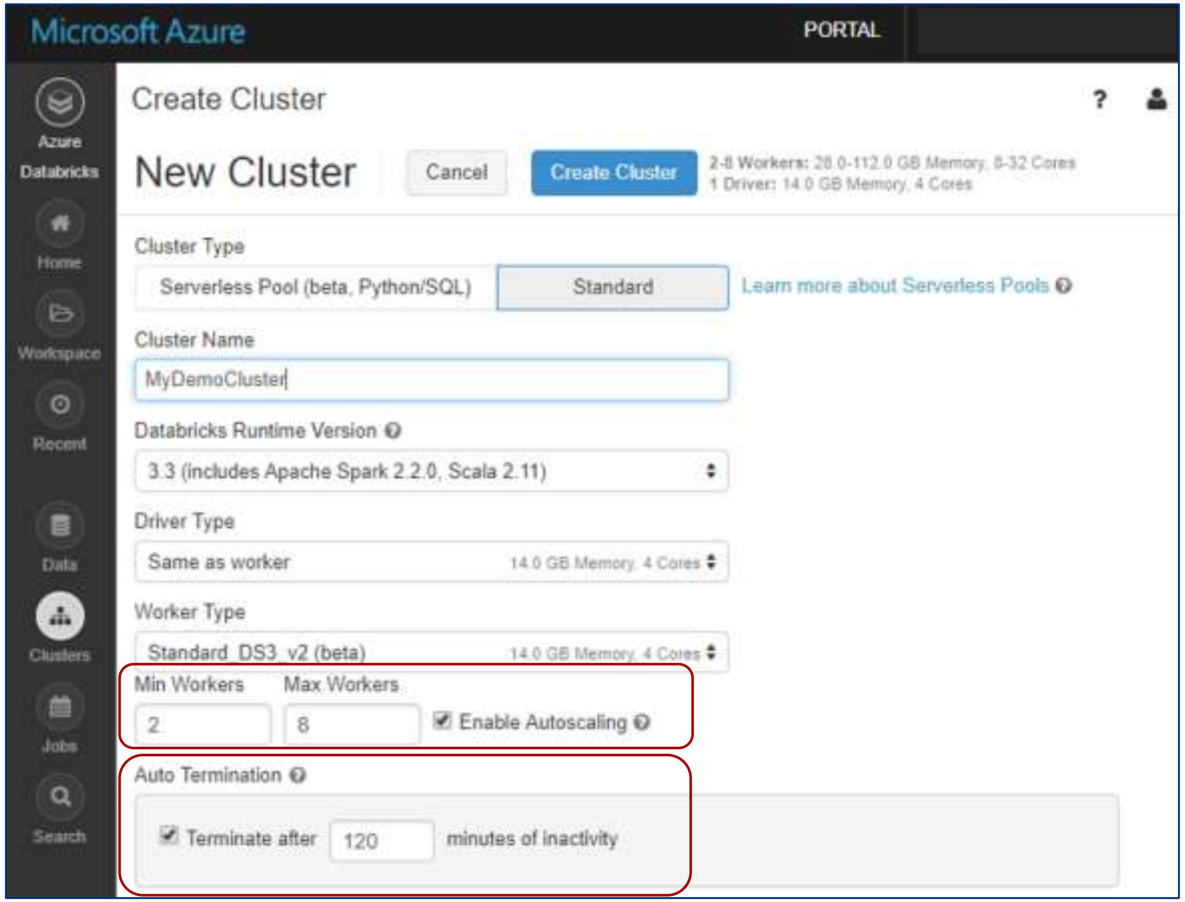
When creating Azure Databricks clusters you can choose Autoscaling and Auto Termination options.

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

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

Benefits:

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



Create Cluster

New Cluster

2-8 Workers: 28.0-112.0 GB Memory, 8-32 Cores
1 Driver: 14.0 GB Memory, 4 Cores

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

Learn more about Serverless Pools

Cluster Name: MyDemoCluster

Databricks Runtime Version: 3.3 (includes Apache Spark 2.2.0, Scala 2.11)

Driver Type: Same as worker — 14.0 GB Memory, 4 Cores

Worker Type: Standard_DS3_v2 (beta) — 14.0 GB Memory, 4 Cores

Min Workers: 2 Max Workers: 8 ☑ Enable Autoscaling

Auto Termination

☑ Terminate after 120 minutes of inactivity

JOBS

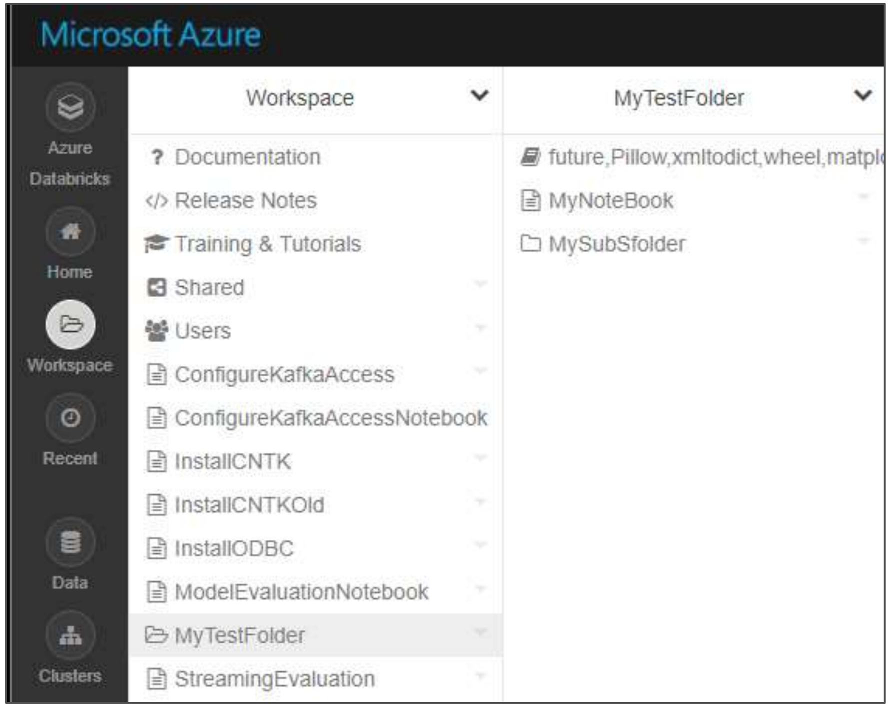
Jobs are the mechanism to submit Spark application code for execution on the Databricks clusters

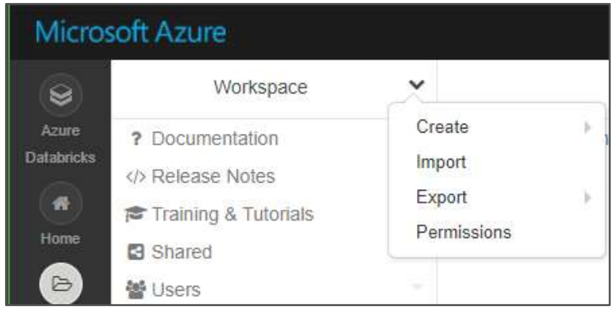
- Spark application code is submitted as a ‘Job’ for execution on Azure Databricks clusters
- Jobs execute either ‘Notebooks’ or ‘Jars’
- Azure Databricks provide a comprehensive set of graphical tools to create, manage and monitor Jobs.

WORKSPACES

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

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

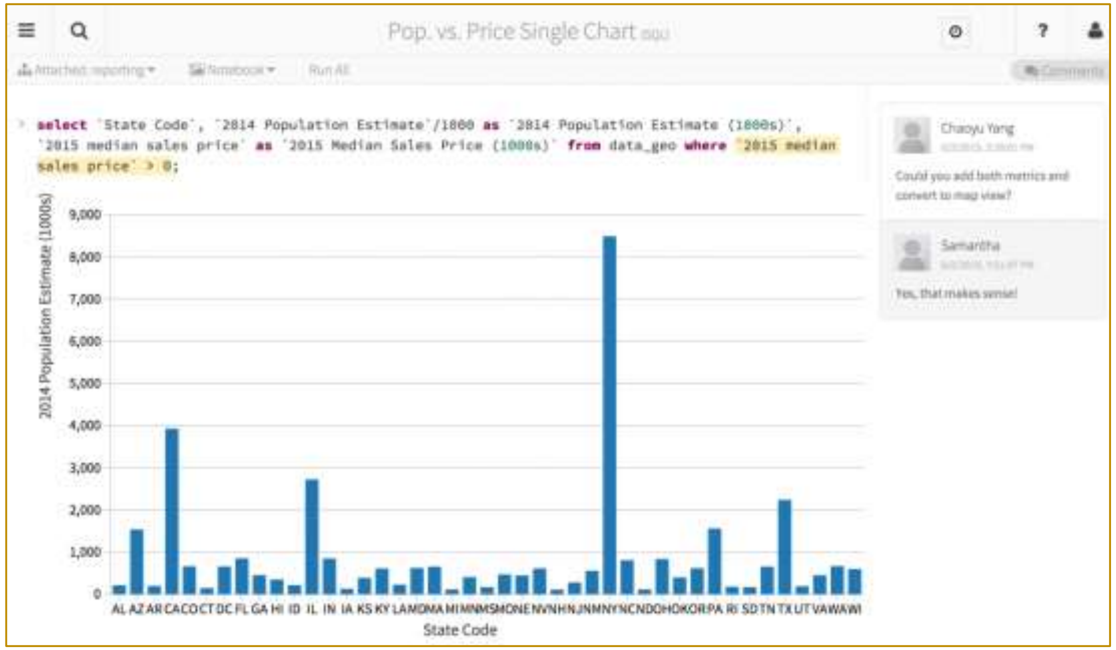




AZURE DATABRICKS NOTEBOOKS OVERVIEW

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

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


Pop. vs. Price Single Chart
Attached: reporting
Notebook
Run All
Comments
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
2014 Population Estimate (1000s)
AL AZ AR CA CO CT DC FL GA HI ID IL IN IA KS KY LA MD MA MI MN MS MO NE NV NH NJ NM NY NC ND OH OK OR PA RI SD TN TX UT VA WA WI
State Code

```
select `State Code`, `2014 Population Estimate`/1000 as `2014 Population Estimate (1000s)`,
`2015 median sales price` as `2015 Median Sales Price (1000s)` from data_geo where `2015 median
sales price` > 0;
```

Chaoyu Yang

Could you add both metrics and convert to map view?

Samantha

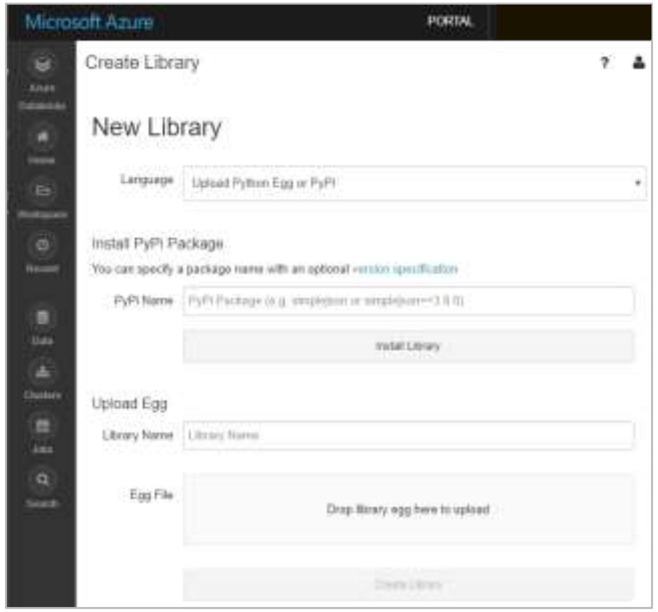
Yes, that makes sense!

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

LIBRARIES OVERVIEW

Enables external code to be imported and stored into a Workspace

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



Create Library

New Library

Language: Upload Python Egg or PyPI

Install PyPI Package

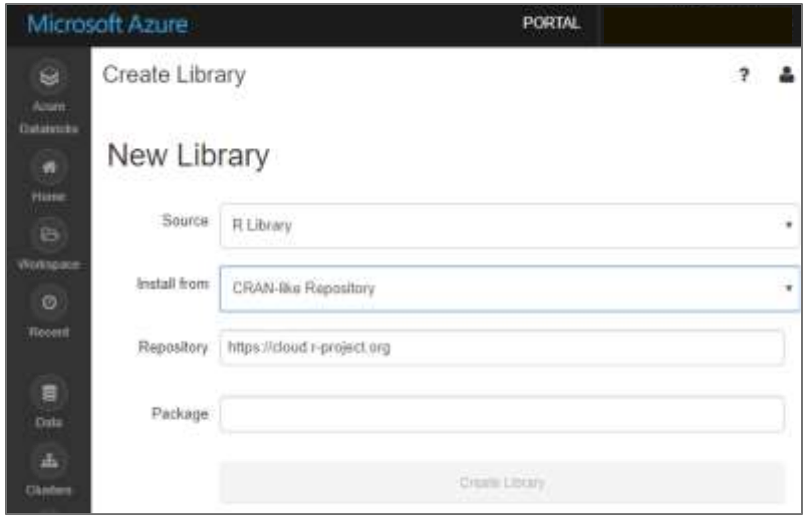
You can specify a package name with an optional version specification

PyPI Name: PyPI Package (e.g. simplejson or simplejson==3.8.0)

Upload Egg

Library Name: Library Name

Egg File: Drop library egg here to upload



Create Library

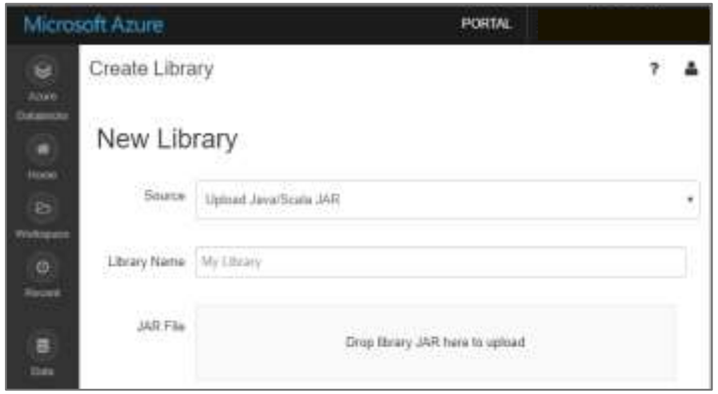
New Library

Source: R Library

Install from: CRAN-like Repository

Repository: https://cloud.r-project.org

Package:



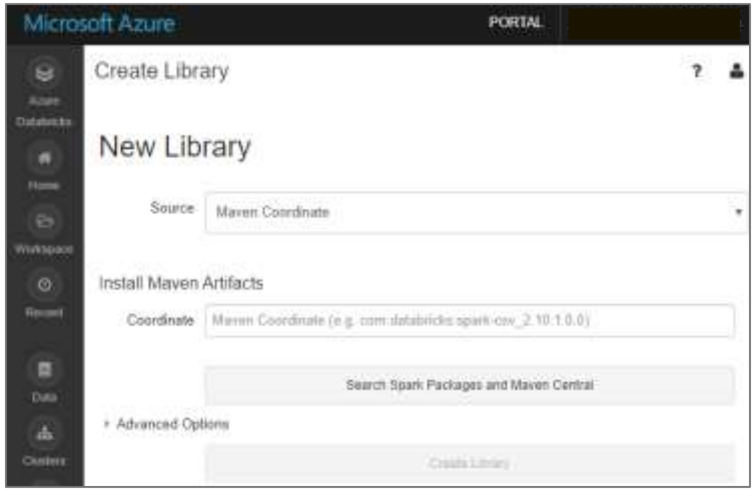
Create Library

New Library

Source: Upload Java/Scala JAR

Library Name: My Library

JAR File: Drop library JAR here to upload



Create Library

New Library

Source: Maven Coordinate

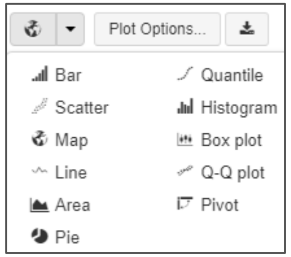
Install Maven Artifacts

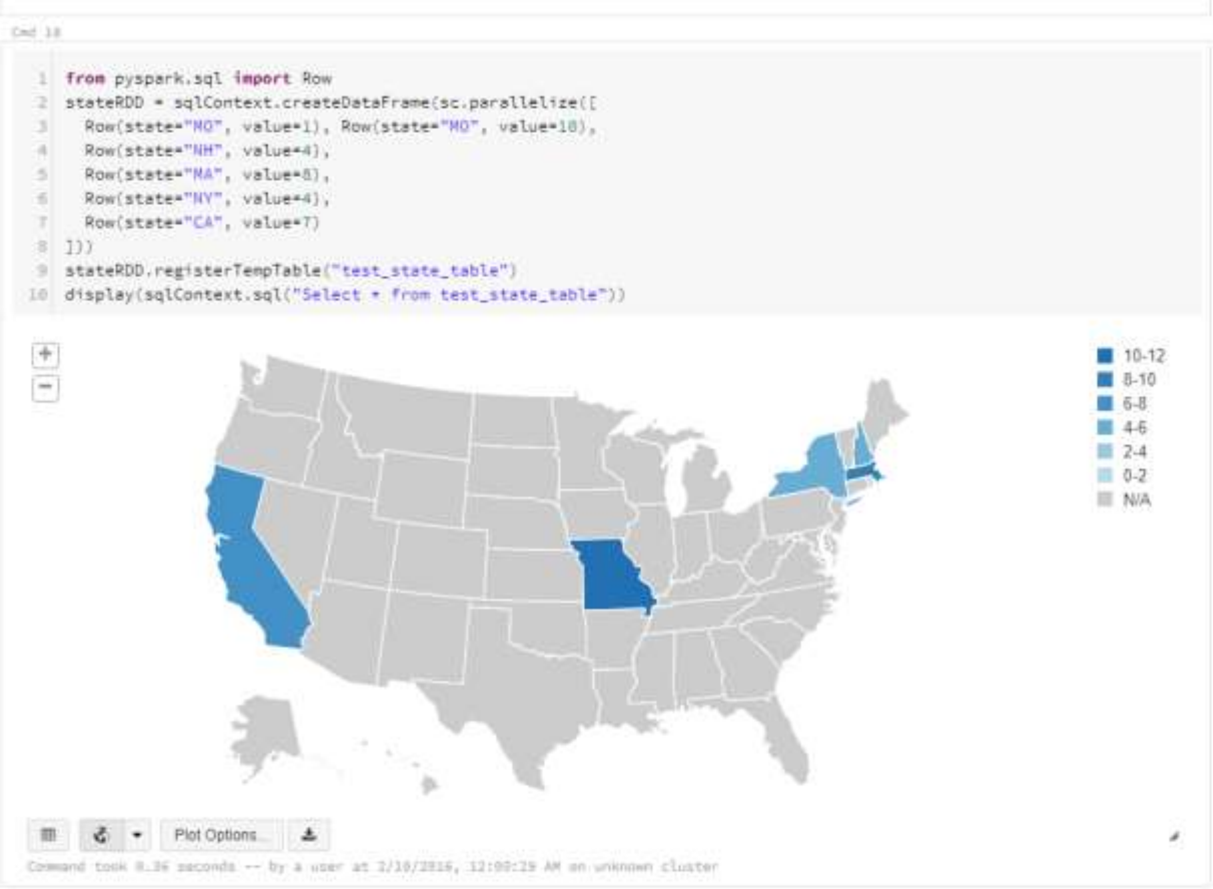
Coordinate: Maven Coordinate (e.g. com.databricks:spark-csv_2.10:1.0.0)

VISUALIZATION

Azure Databricks supports a number of visualization plots out of the box

- All notebooks, *regardless of their language*, support Databricks visualizations.
- When you run the notebook the visualizations are rendered inside the notebook in-place
- The visualizations are written in HTML.
  - You can save the HTML of the entire notebook by exporting to HTML.
  - If you use Matplotlib, the plots are rendered as images so you can just right click and download the image
- You can change the plot type just by picking from the selection





Cmd 18

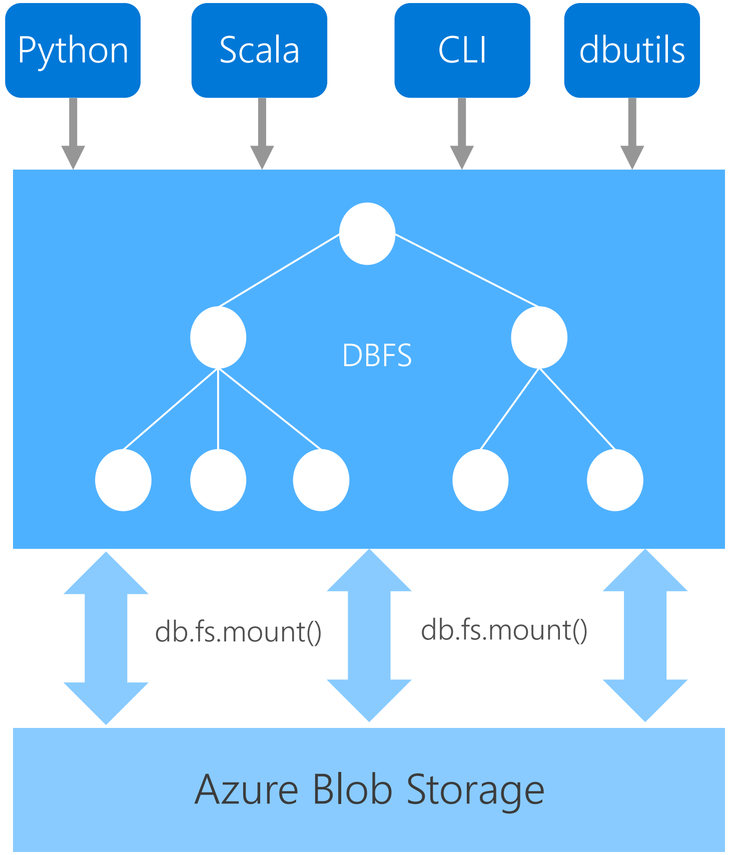
```
from pyspark.sql import Row
stateRDD = sqlContext.createDataFrame(sc.parallelize([
  Row(state="NO", value=1), Row(state="MO", value=10),
  Row(state="NH", value=4),
  Row(state="MA", value=8),
  Row(state="NY", value=4),
  Row(state="CA", value=7)
]))
stateRDD.registerTempTable("test_state_table")
display(sqlContext.sql("Select * from test_state_table"))
```

Command took 0.36 seconds -- by a user at 3/10/2016, 12:00:29 AM on unknown cluster

DATABRICKS FILE SYSTEM (DBFS)

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

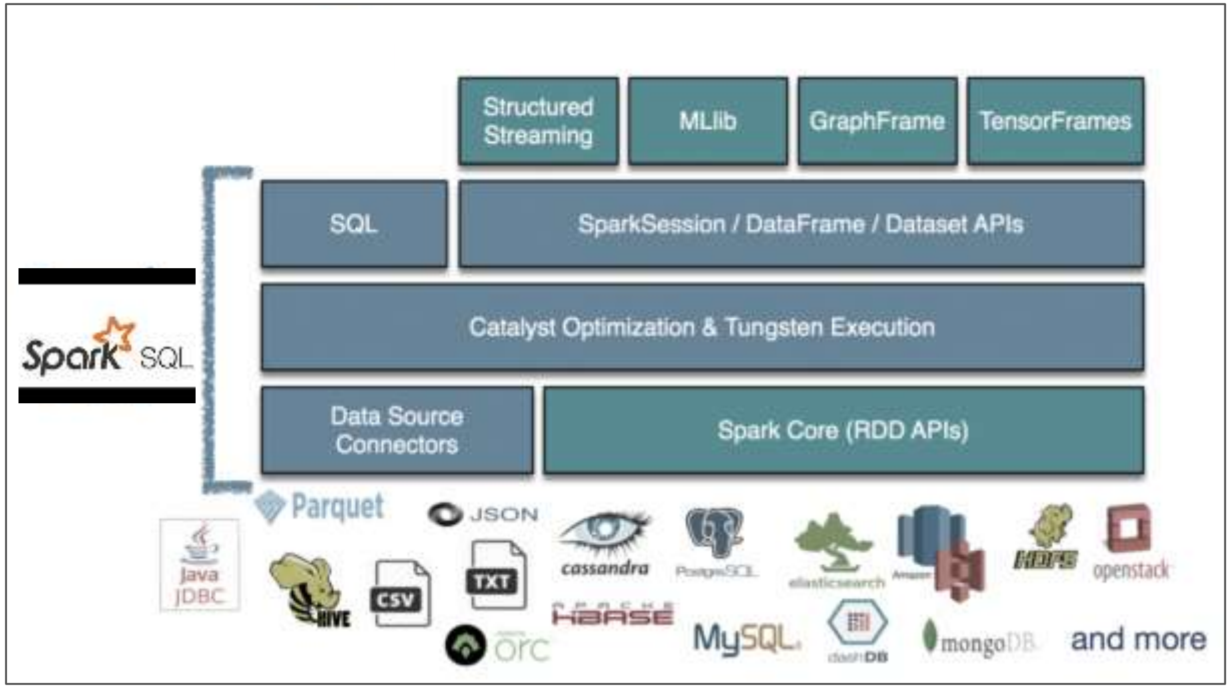
- Azure Storage buckets can be mounted in DBFS so that users can directly access them without specifying the storage keys
- DBFS mounts are created using *dbutils.fs.mount()*
- Azure Storage data can be cached locally on the SSD of the worker nodes
- Available in both Python and Scala and accessible 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



SPARK SQL OVERVIEW

Spark SQL is a distributed SQL query engine for processing structured data

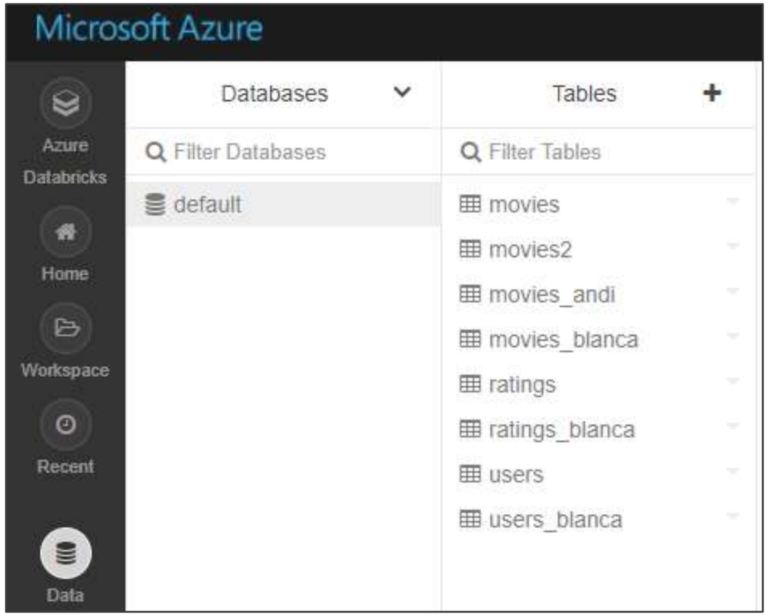
- Can query data stored in wide variety of data sources—external databases, structured data files, Hive tables and more.
- Data can be queried using either SQL or HiveQL
- Has bindings in Python, Scala and Java
- Has built-in support for structured streaming.
- Built using the [Catalyst optimizer]and [Tungsten execution]



DATABASES AND TABLES OVERVIEW

Tables enable data to be structured and queried using Spark SQL or any of the Spark's language APIs

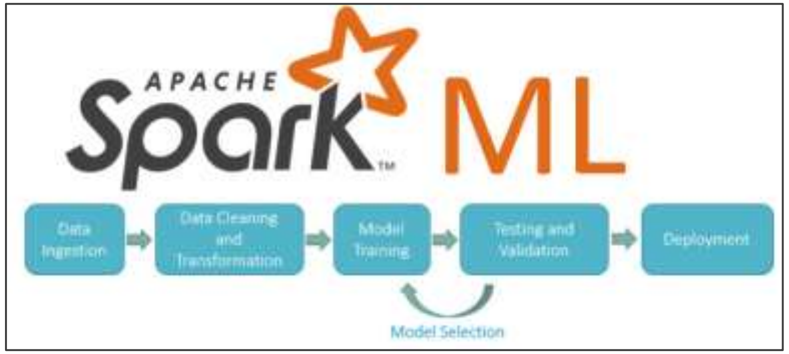
- Databases are a collection of related tables
- Tables are defined using the GUI in the console or programmatically using APIs or Notebooks
- Databricks uses the Hive metastore to manage tables, and supports all file formats and Hive data sources.
- There are multiple ways to create tables (see next slide).
- Like Apache Spark DataFrames, any Spark operation can be applied to Tables (including caching, filtering).
- Partitioned Tables and Partition Pruning: Spark SQL is able to dynamically generate partitions at the file storage level to provide partition columns for tables. When the table is scanned, Spark pushes down the filter predicates involving the partitionBy keys for partition pruning.



SPARK MACHINE LEARNING(ML) OVERVIEW

Enables Parallel, Distributed ML for large datasets on Spark Clusters

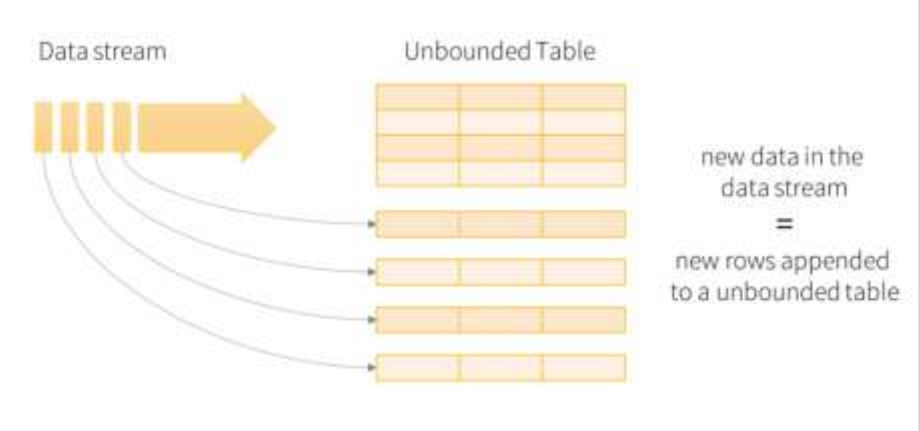
- Offers a set of parallelized machine learning algorithms (MMLSpark, Spark ML, Deep Learning, SparkR)
- Supports Model Selection (hyperparameter tuning) using Cross Validation and Train-Validation Split.
- Supports Java, Scala or Python apps using DataFrame-based API (as of Spark 2.0). Benefits include:
  - 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



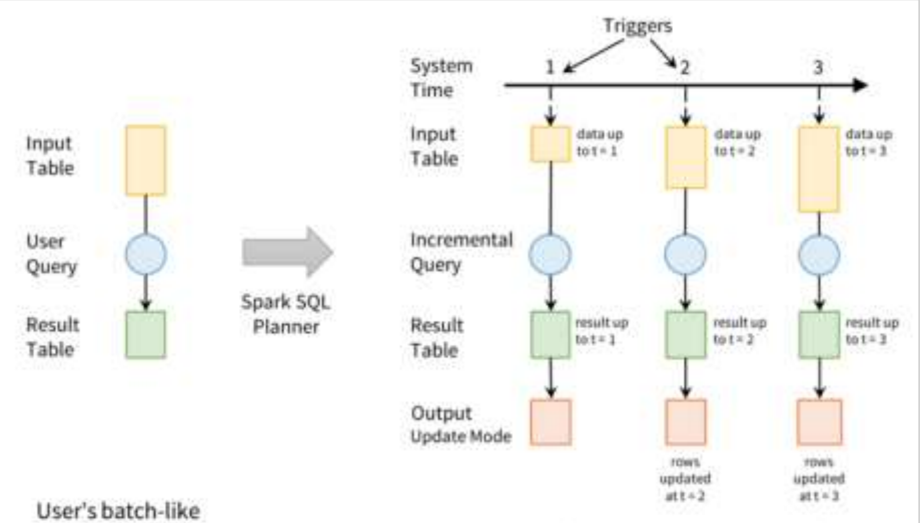
SPARK STRUCTURED STREAMING OVERVIEW

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)


Data stream
Unbounded Table
new data in the data stream
=
new rows appended to a unbounded table

Data stream as an unbounded table



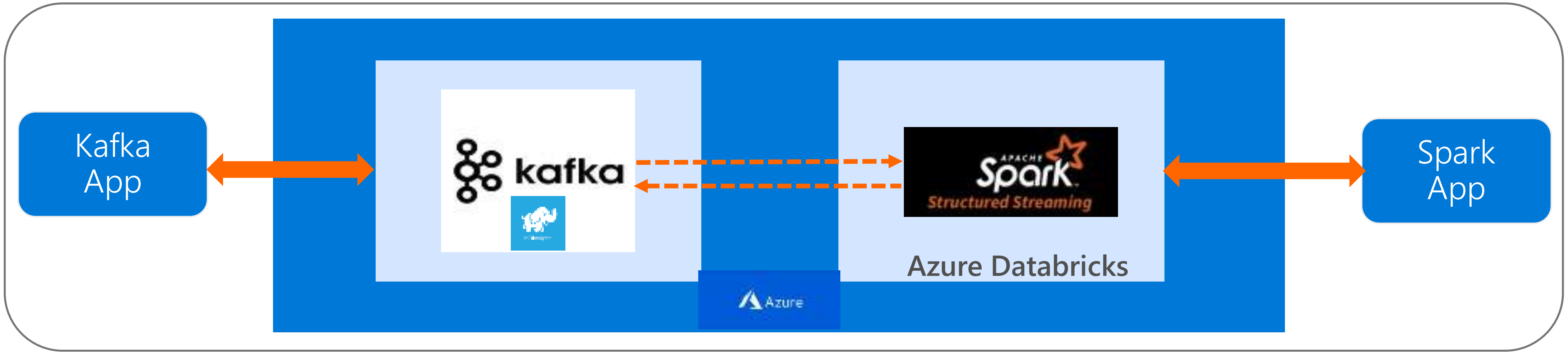
User's batch-like query on input table

Incremental execution on streaming data

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

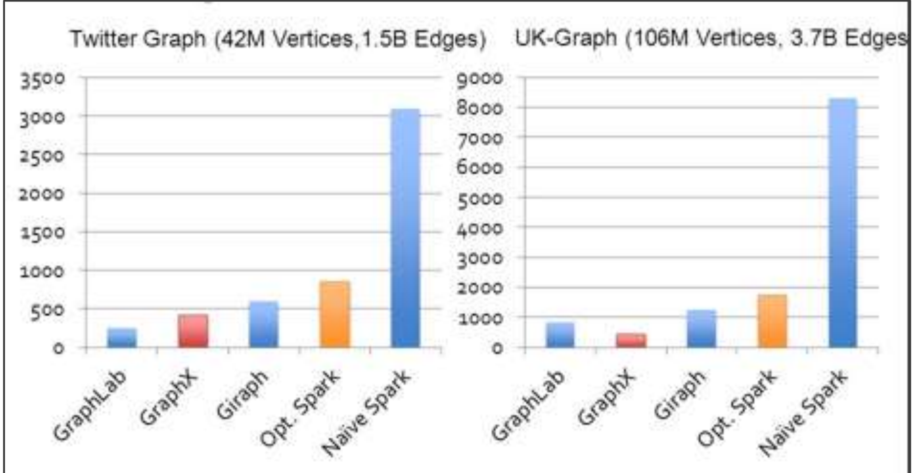
A set of APIs for graph and graph-parallel computation.

- 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

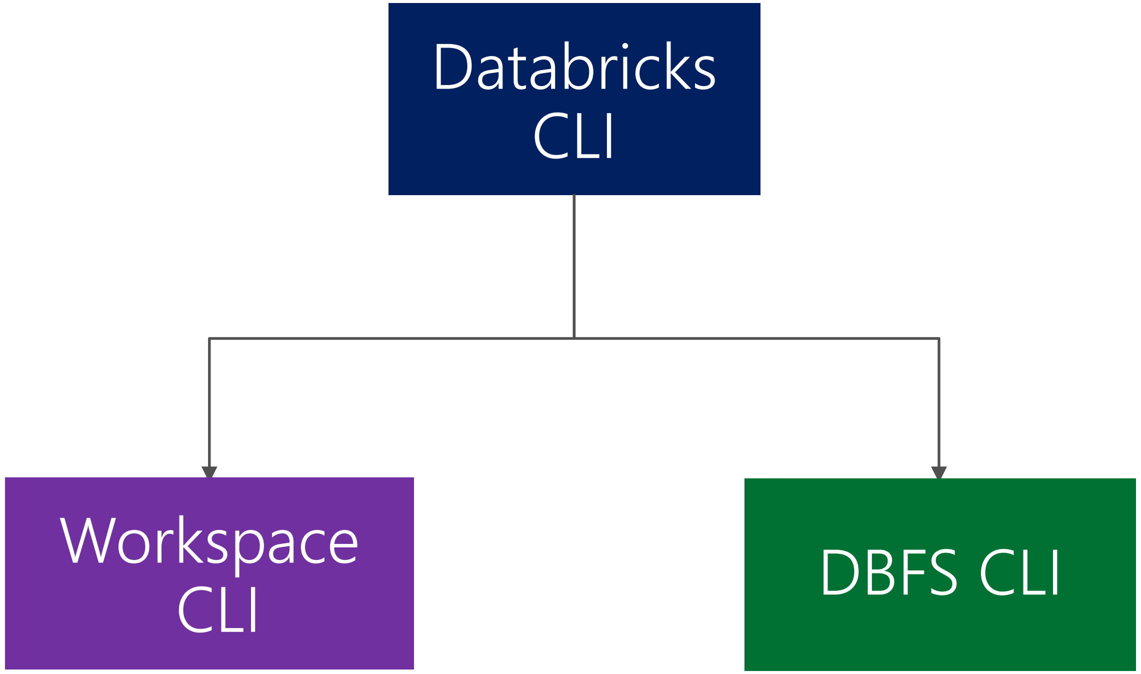


GraphX performs comparably to state-of-the-art graph processing systems.

Source: AMPLab

DATABRICKS CLI

An easy to use interface built on top of the Databricks REST API



Currently, the CLI fully implements the DBFS API and the Workspace API

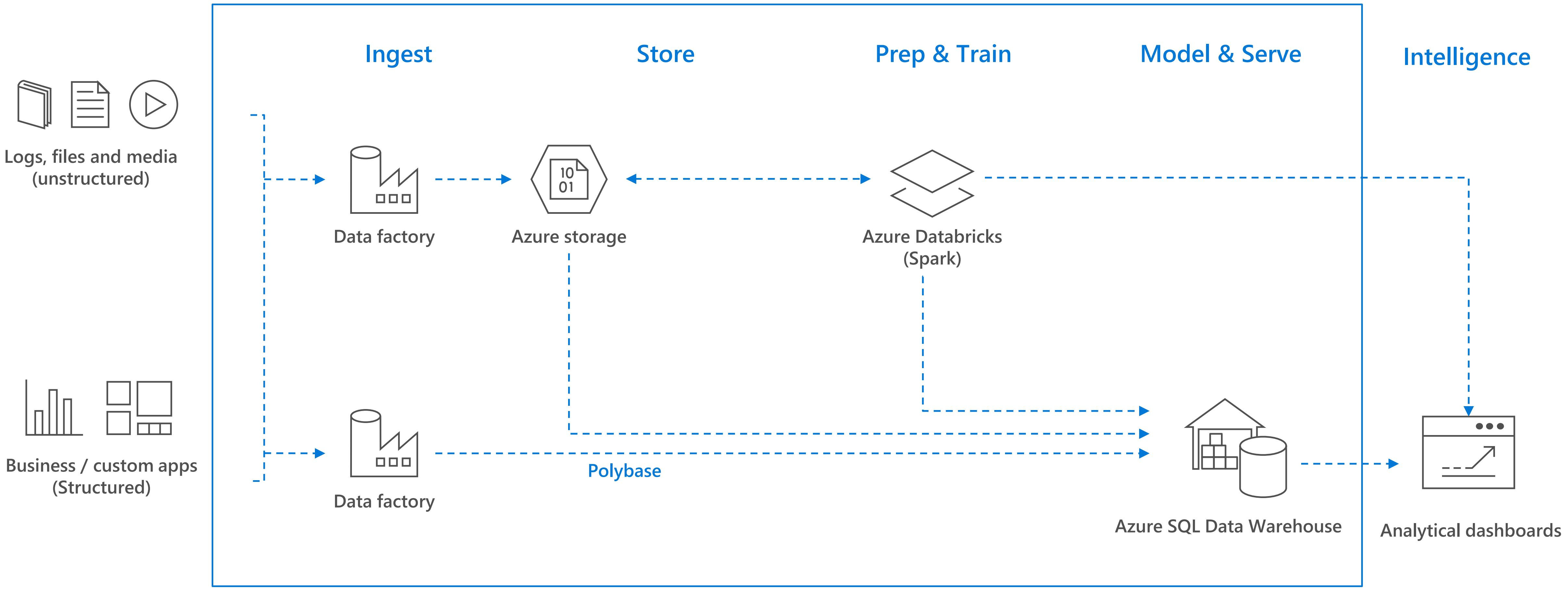
DATABRICKS REST API

Databricks REST API

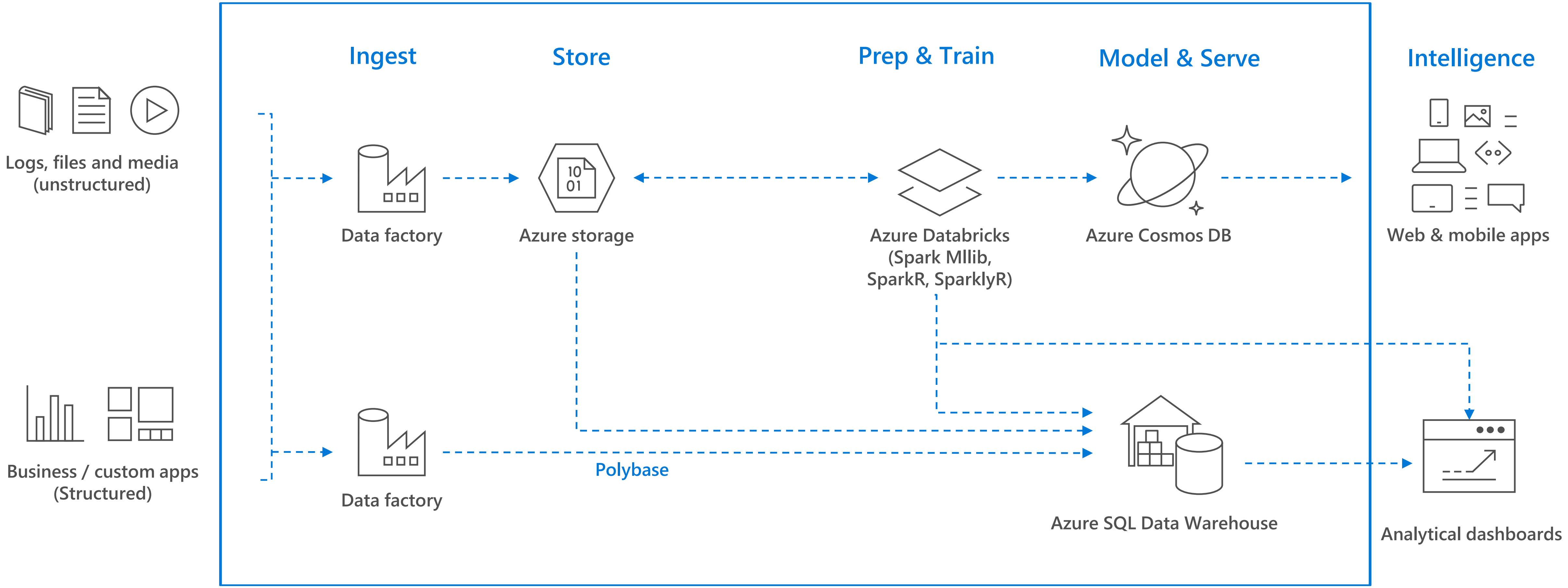
Cluster API	Create/edit/delete clusters
DBFS API	Interact with the Databricks File System
Groups API	Manage groups of users
Instance Profile API	Allows admins to add, list, and remove instances profiles that users can launch clusters with
Job API	Create/edit/delete jobs
Library API	Create/edit/delete libraries
Workspace API	List/import/export/delete notebooks/folders

Use Cases

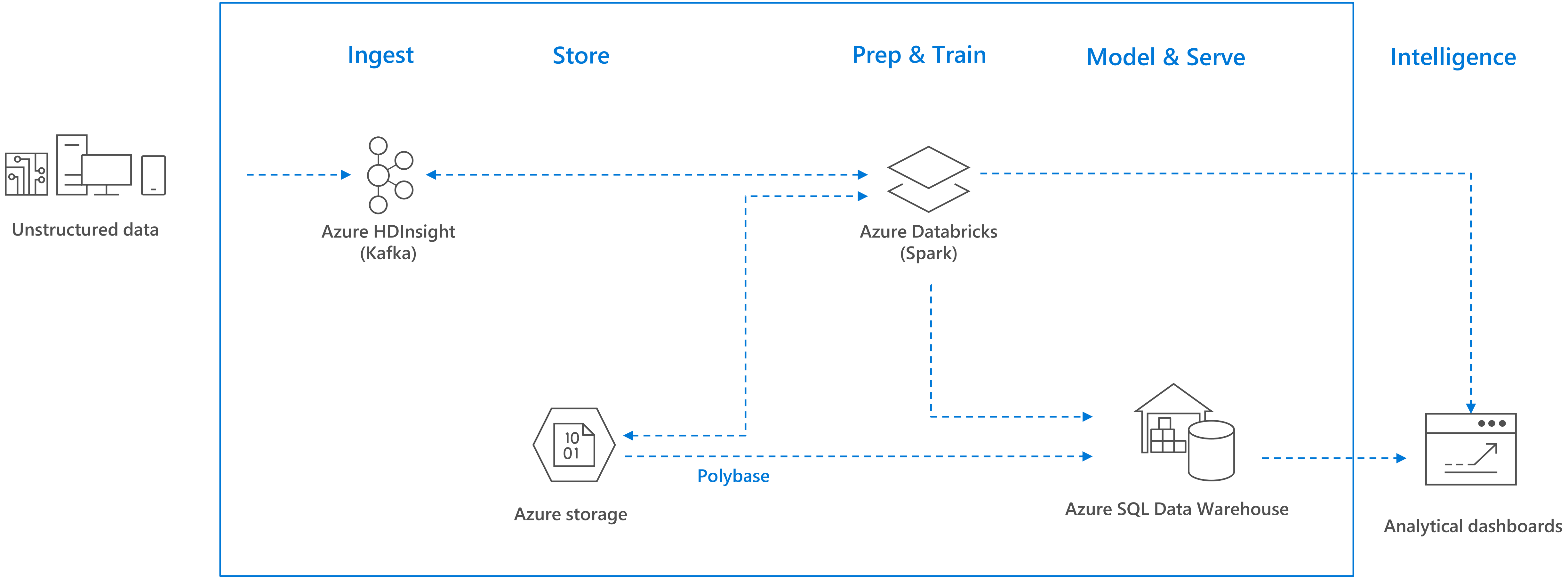
Modern Big Data Warehouse



Advanced Analytics on Big Data



Real-time analytics on Big Data



Pricing & Product Guidance

Big Data OSS - Comparison

Azure HDInsight (1st party + Support)

What it is

- **Hadoop** (Hortonworks' Distribution) as a managed service supporting a variety of open-source analytics engines such as Apache Spark, Hive LLAP, Storm, Kafka, HBase.
- Security via Ranger (Kerberos based)

Pricing

- Priced to compete with AWS EMR. Standard offering.

Use When

- Customer prefers a **PaaS** like experience to address big data use cases by working with different OSS analytics engines to address big data use cases. Cost sensitive.

Azure Databricks (1st party + Support)

What it is

- Databricks **Spark**, the most popular open-source analytics engine, as a managed service providing an easy and fast way to unlock big data use cases. Offers best-in-class notebooks experience for productivity and collaboration as well integration with Azure Data Warehouse, Power BI, etc
- Security via native Azure AD integration

Pricing

- Priced to match Databricks on AWS. Premium offering.

Use When

- Customer prefers **SaaS** like experience to address big data use cases and values Databricks' ease of use, productivity & collaboration features.

3rd Party Offerings

What it is

Hadoop distributions from Cloudera, MapR & Hortonworks available on Azure Marketplace as IaaS VMs.

Pricing

- N/A. Vendor prices their products.

Use When

- Customer wants to move their on premises Hadoop distribution to Azure IaaS using their existing licenses.

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 (3 in preview, 8 by GA)

Azure ML

What It Is

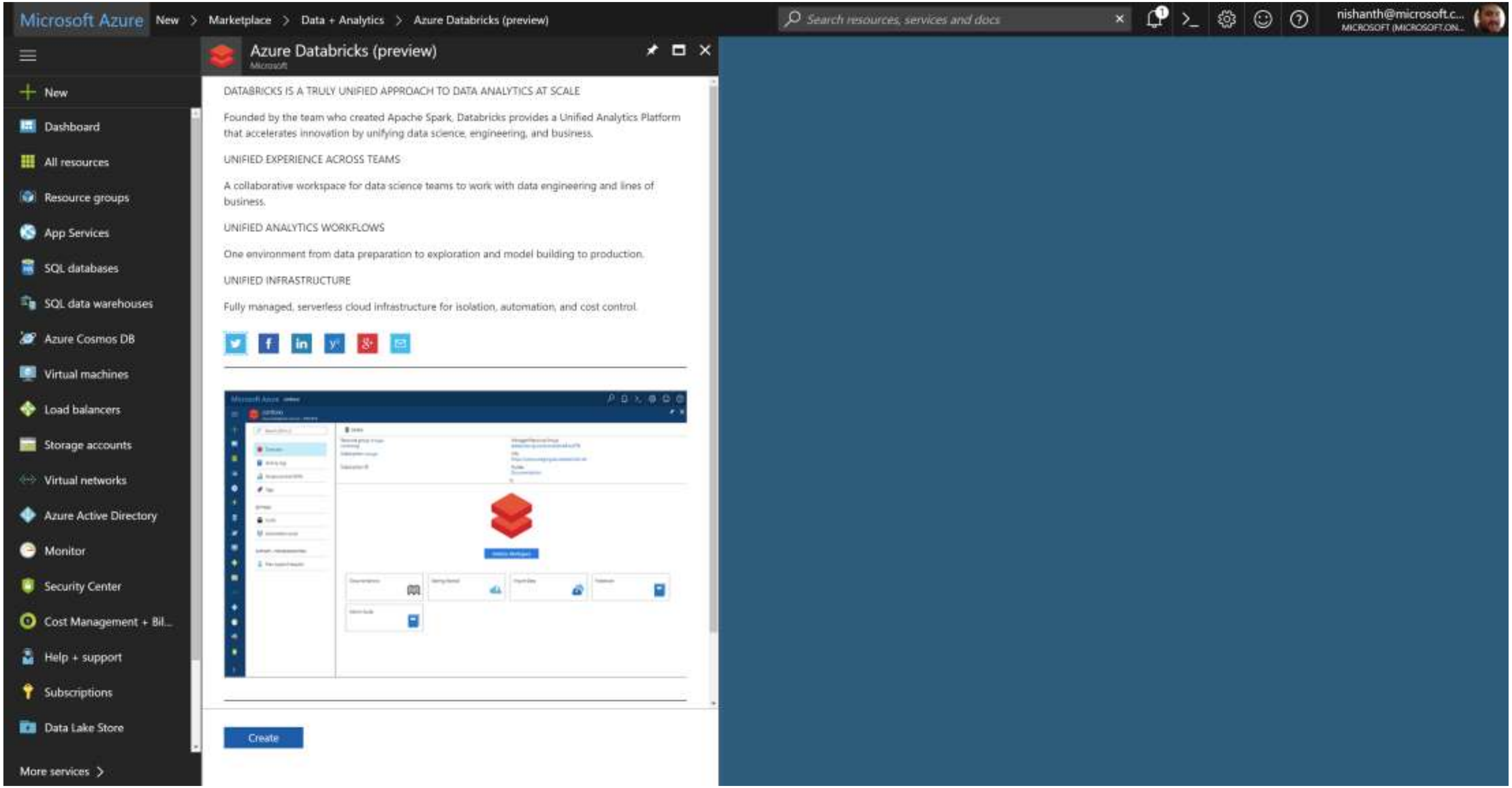
- Azure first party service for Machine Learning
- Leverage existing ML libraries or extend with Python and R
- Targets emerging data scientists with drag & drop offering
- Targets professional data scientists with
  - Experimentation service
  - Model management service
  - Works with customers IDE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

Demo

Azure Databricks – service home page



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.

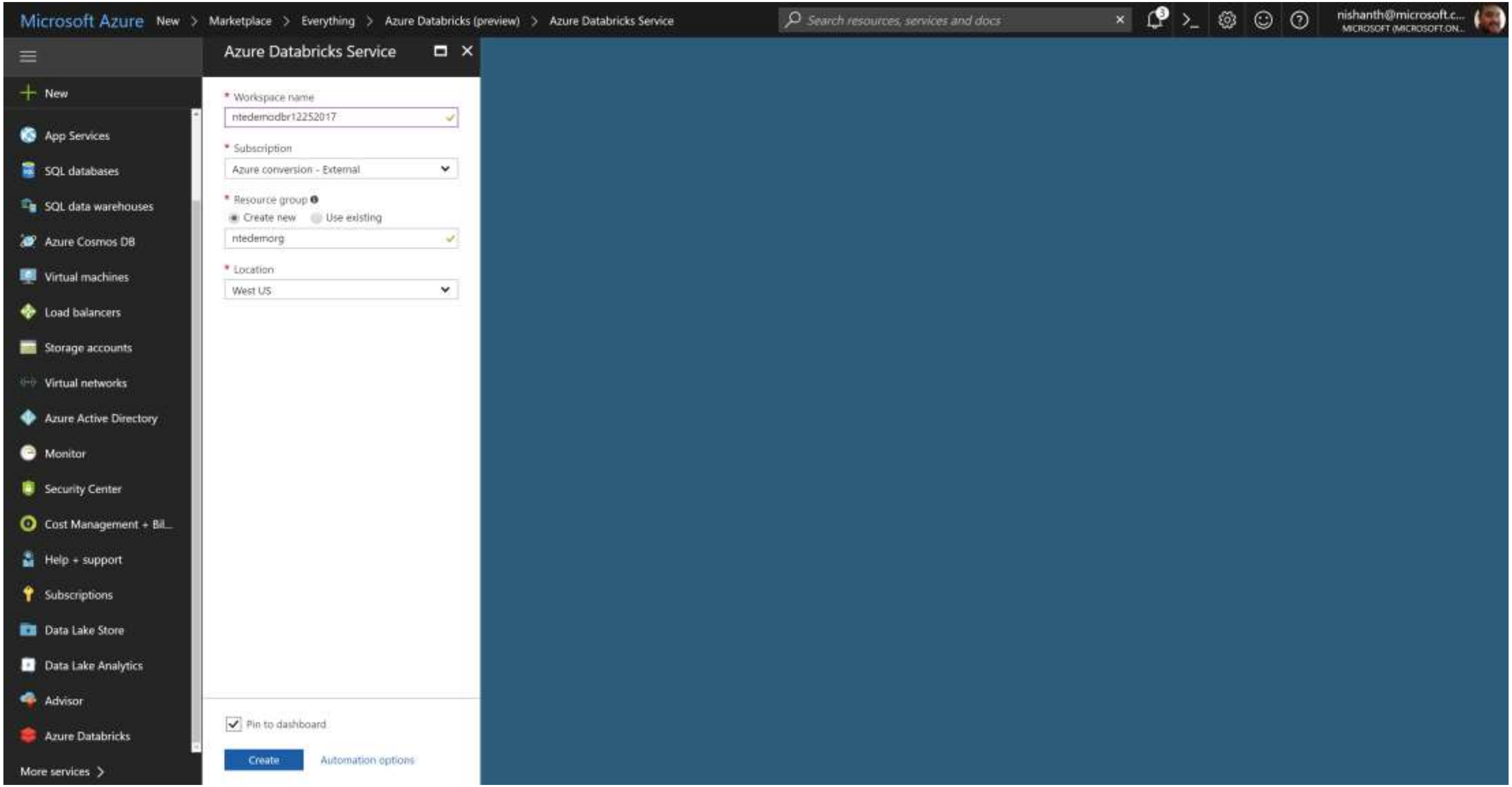
UNIFIED ANALYTICS WORKFLOWS

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.

Azure Databricks – creating a workspace



Azure Databricks Service

* Workspace name: ntedemodbr12252017

* Subscription: Azure conversion - External

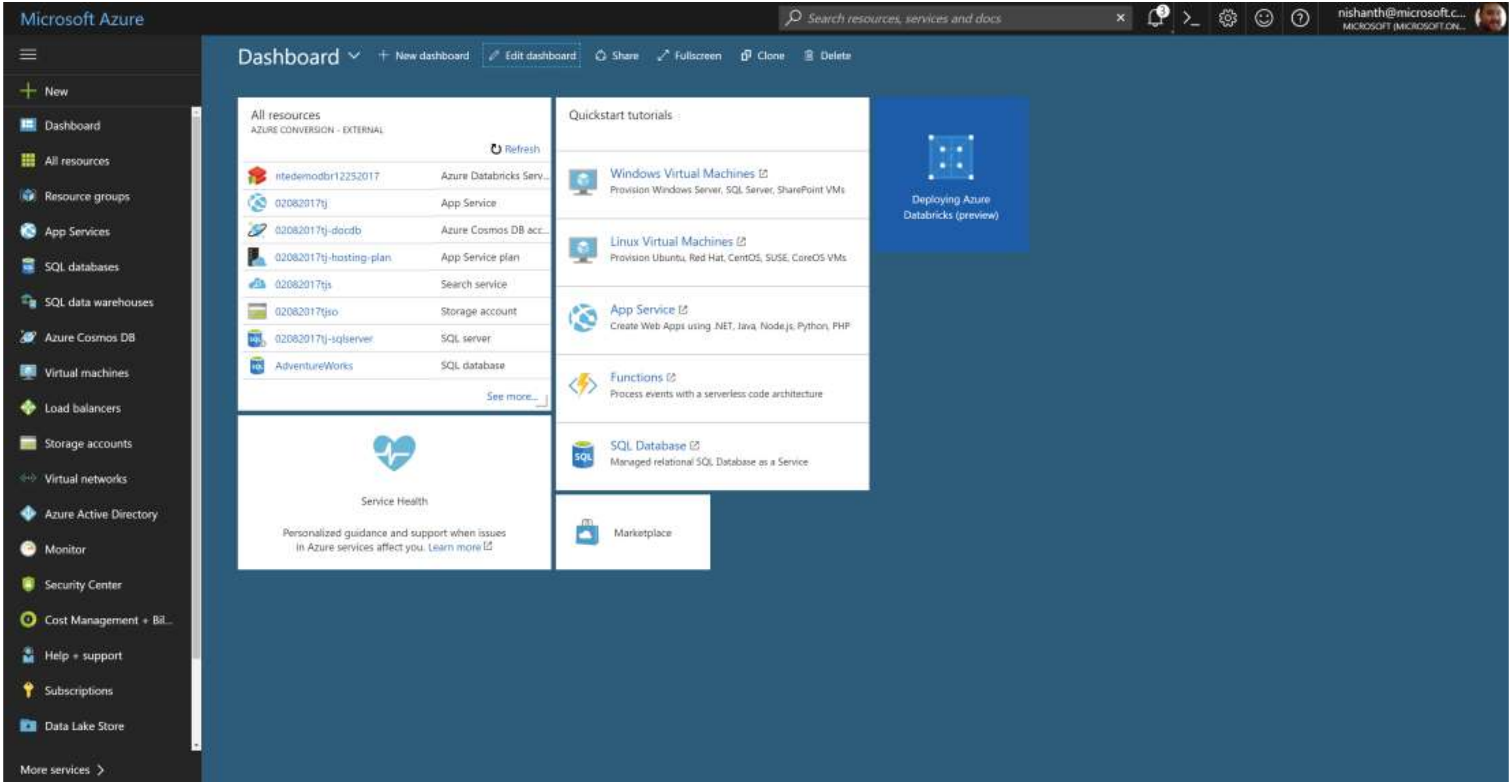
* Resource group: ● Create new ○ Use existing

ntedemorg

* Location: West US

☑ Pin to dashboard

Azure Databricks – workspace deployment



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...

Service Health

Personalized guidance and support when issues in Azure services affect you. Learn 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

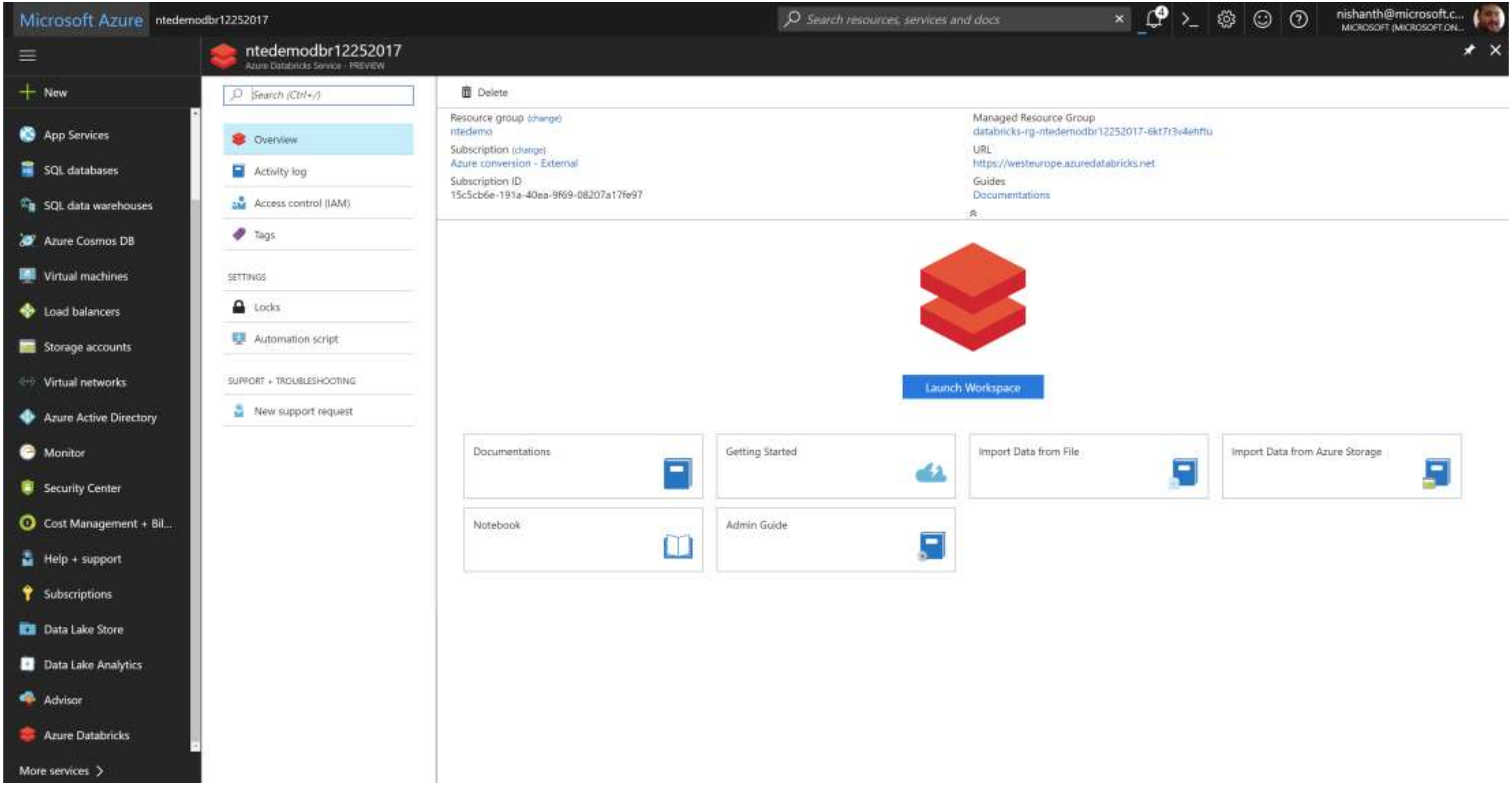
Functions
Process events with a serverless code architecture

SQL Database
Managed relational SQL Database as a Service

Marketplace

Deploying Azure Databricks (preview)

Azure Databricks – launching the workspace



Resource group (change)
ntedemo

Subscription (change)
Azure conversion - External

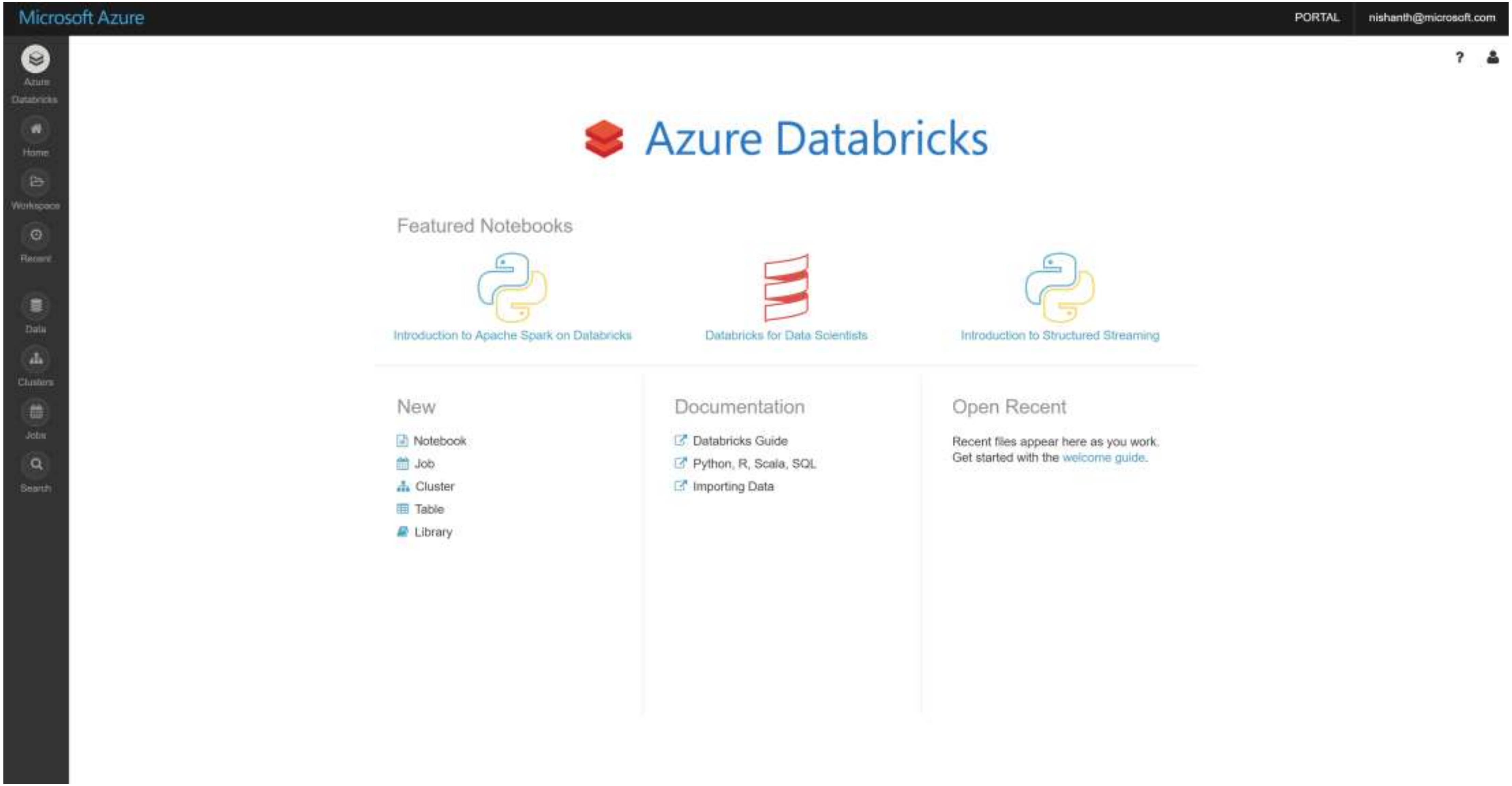
Subscription ID
15c5cb6e-191a-40ea-9f69-08207a17fe97

Managed Resource Group
databricks-rg-ntedemodbr12252017-6kt7r3v4ehftu

URL
https://westeurope.azuredatabricks.net

Guides
Documentations

Azure Databricks – workspace home page



Azure Databricks

Featured Notebooks

Introduction to Apache Spark on Databricks

Databricks for Data Scientists

Introduction to Structured Streaming

New

- Notebook
- Job
- Cluster
- Table
- Library

Documentation

- Databricks Guide
- Python, R, Scala, SQL
- Importing Data

Open Recent

Recent files appear here as you work.
Get started with the welcome guide.

How to get started

How to get started

Sign up for preview at http://databricks.azurewebsites.net

Engage Microsoft experts for a workshop to help identify high impact scenarios

Learn more about Azure Databricks www.azure.com/databricks

Q & A

James Serra, Big Data Evangelist

Email me at: JamesSerra3@gmail.com

Follow me at: @JamesSerra

Link to me at: www.linkedin.com/in/JamesSerra

Visit my blog at: JamesSerra.com (where this slide deck is posted under the “Presentations” tab)