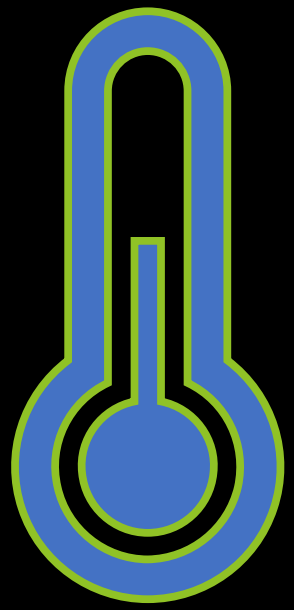
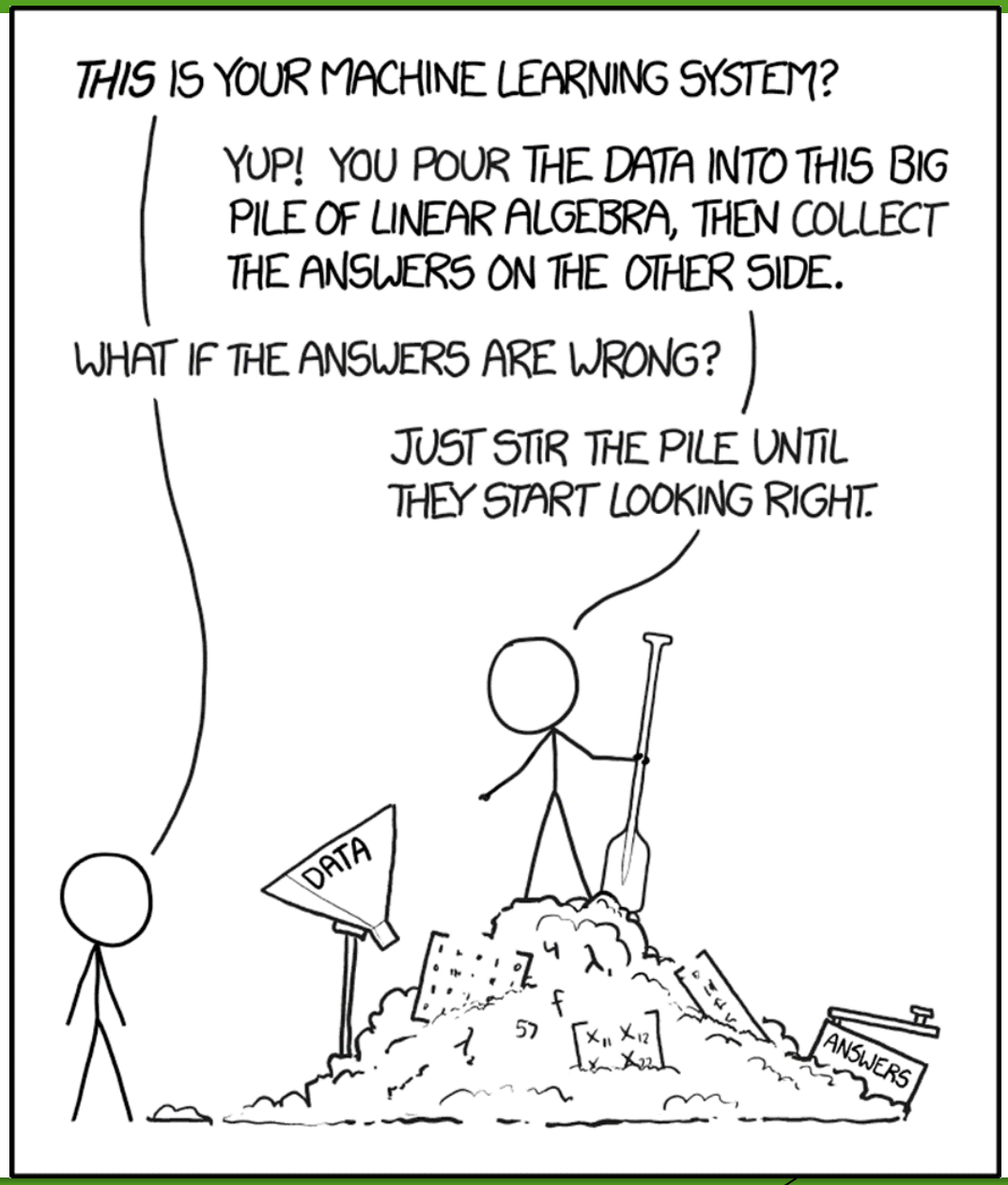
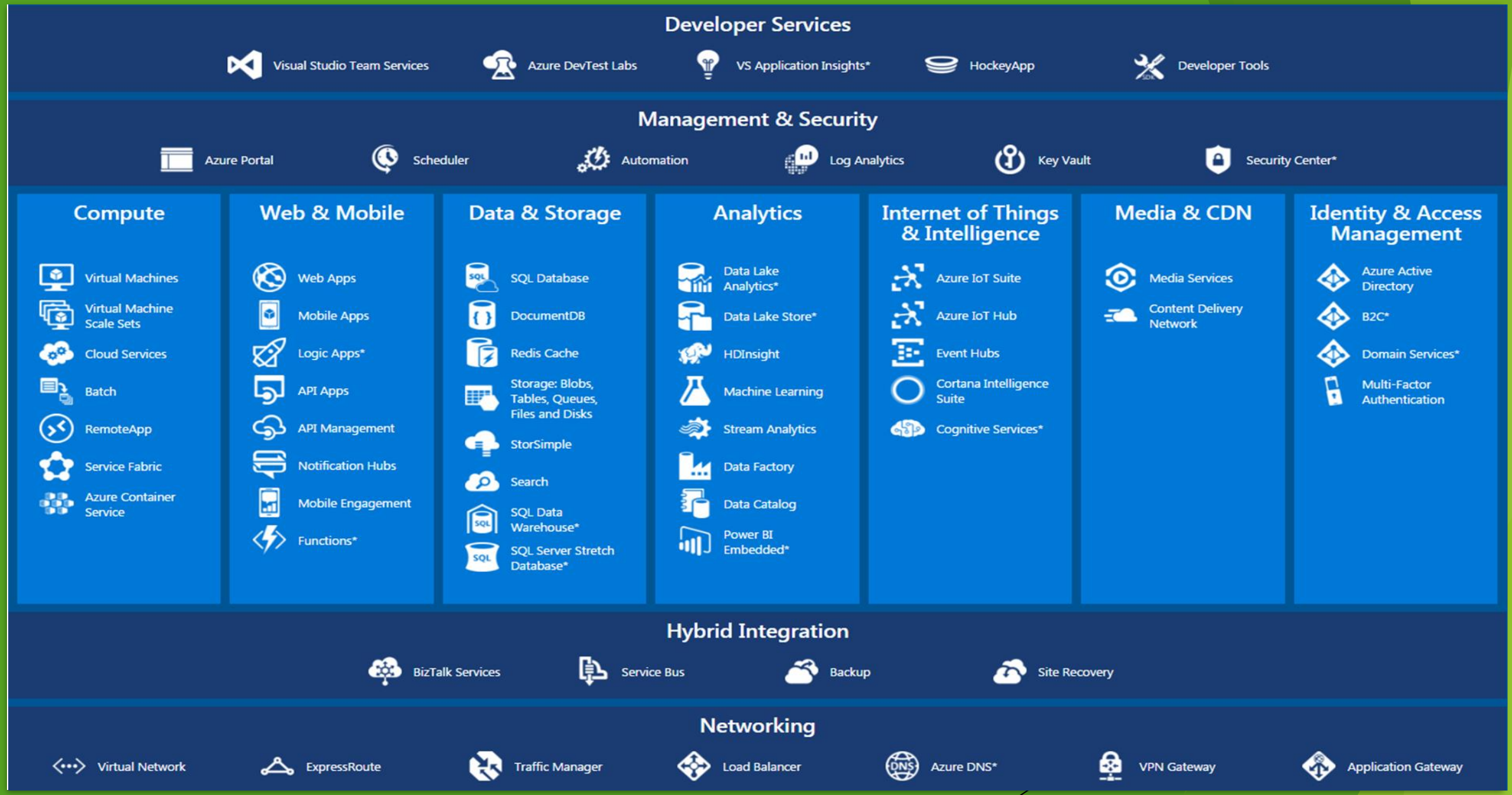


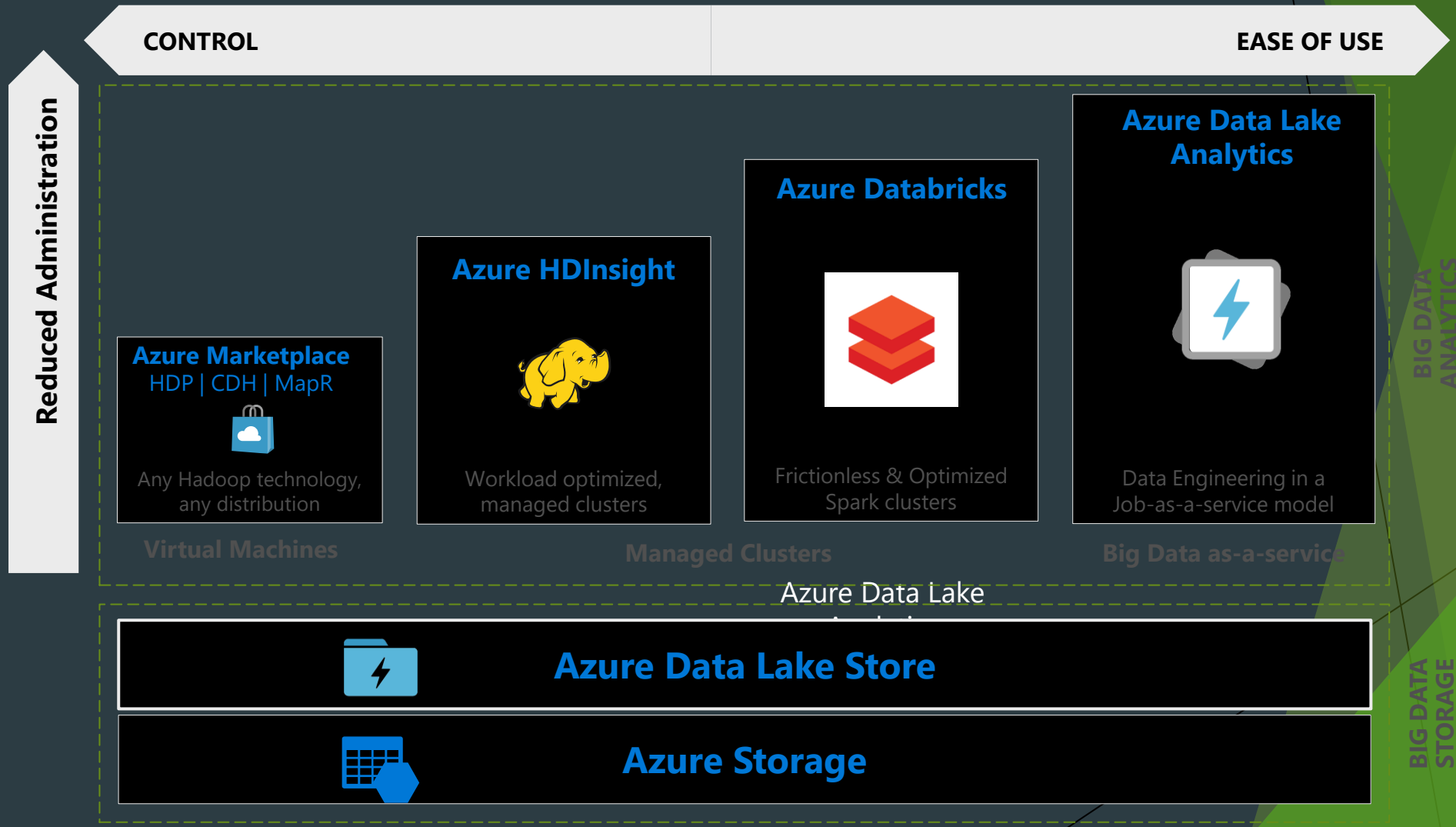
SERVICIUL AZURE DATABRICKS

CONF.DR. CRISTIAN KEVORCHIAN
UNIVERSITATEA DIN BUCUREȘTI
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

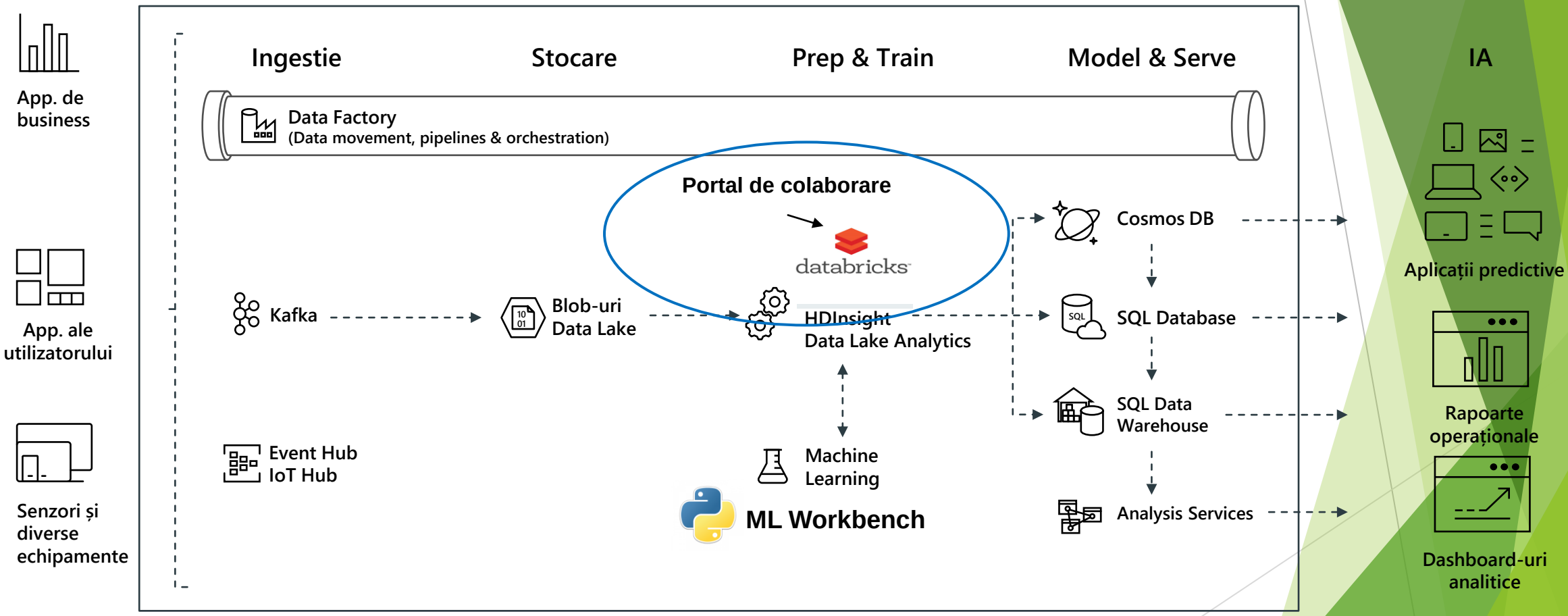

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


Developer Services
Visual Studio Team Services
Azure DevTest Labs
VS Application Insights*
HockeyApp
Developer Tools
Management & Security
Azure Portal
Scheduler
Automation
Log Analytics
Key Vault
Security Center*
Compute
Virtual Machines
Virtual Machine Scale Sets
Cloud Services
Batch
RemoteApp
Service Fabric
Azure Container Service
Web & Mobile
Web Apps
Mobile Apps
Logic Apps*
API Apps
API Management
Notification Hubs
Mobile Engagement
Functions*
Data & Storage
SQL Database
DocumentDB
Redis Cache
Storage: Blobs, Tables, Queues, Files and Disks
StorSimple
Search
SQL Data Warehouse*
SQL Server Stretch Database*
Analytics
Data Lake Analytics*
Data Lake Store*
HDInsight
Machine Learning
Stream Analytics
Data Factory
Data Catalog
Power BI Embedded*
Internet of Things & Intelligence
Azure IoT Suite
Azure IoT Hub
Event Hubs
Cortana Intelligence Suite
Cognitive Services*
Media & CDN
Media Services
Content Delivery Network
Identity & Access Management
Azure Active Directory
B2C*
Domain Services*
Multi-Factor Authentication
Hybrid Integration
BizTalk Services
Service Bus
Backup
Site Recovery
Networking
Virtual Network
ExpressRoute
Traffic Manager
Load Balancer
Azure DNS*
VPN Gateway
Application Gateway

BIG DATA ÎN CLOUD(AZURE)



BIG DATA & SISTEME DE ANALITICE



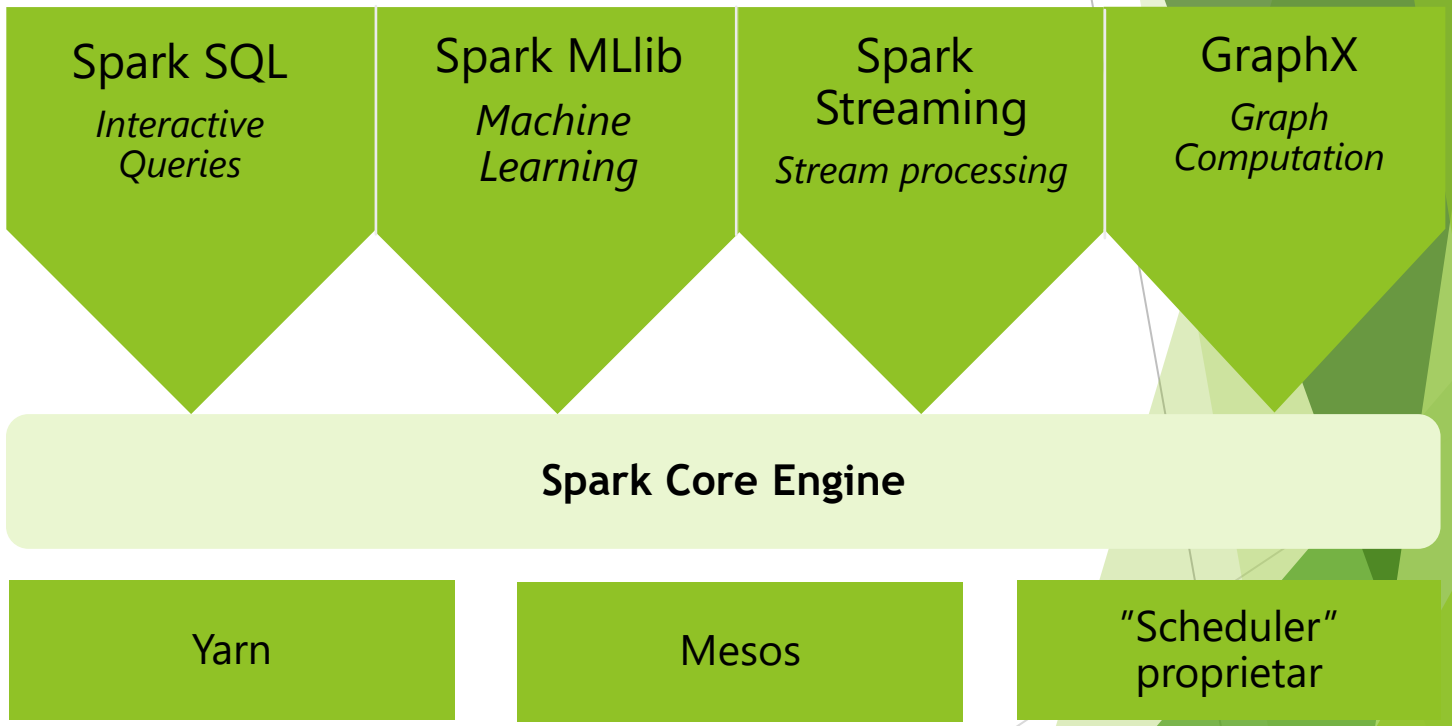
Azure Databricks și Apache Spark

APACHE SPARK

Sistem open source, paralel incluzând un framework pentru procesare date pentru ”Big Data Analytics”

Spark unifică:

- **Procesări Batch**
- **SQL Interactiv**
- **Procesare real-time**
- **Machine Learning**
- **Deep Learning**
- **Graph Processing**

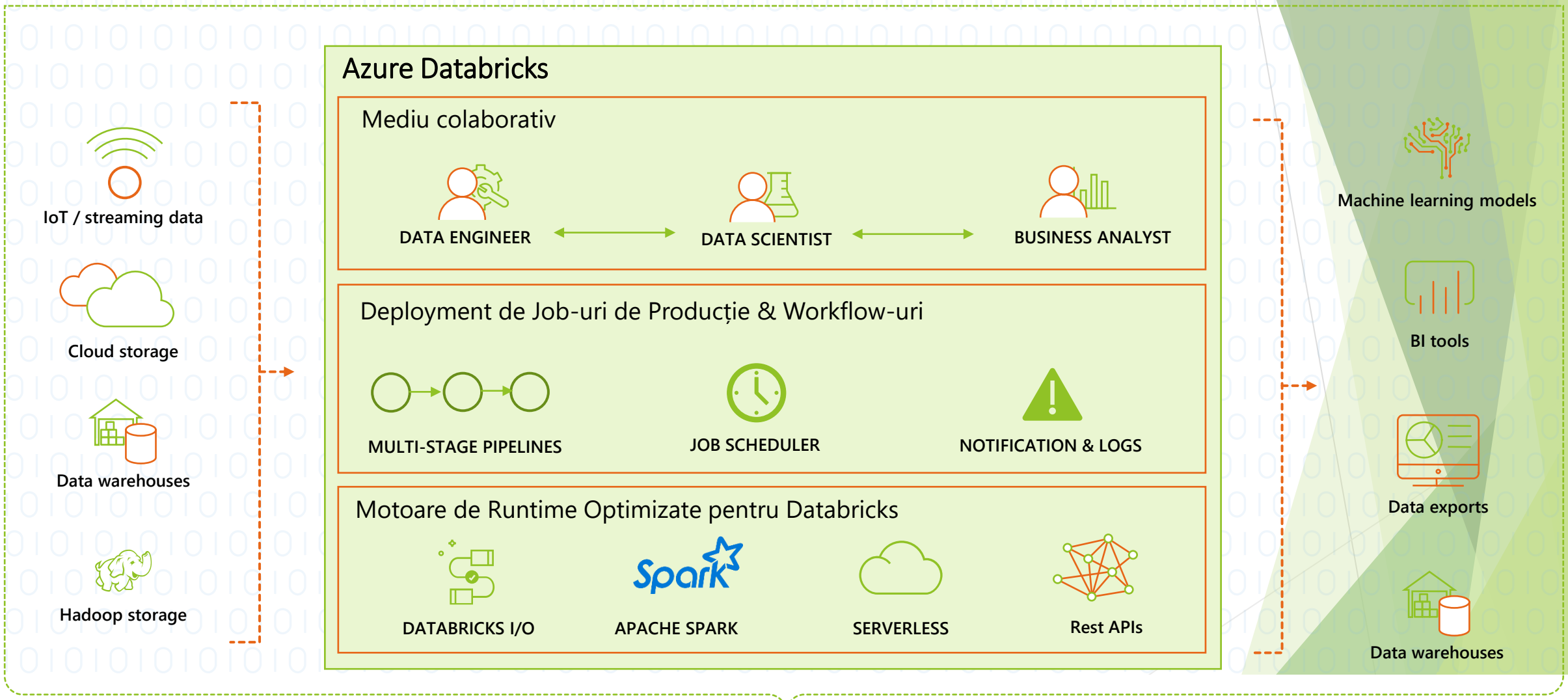


DATABRICKS

Azure Databricks este un serviciu în Azure.

- Bazele de date Azure sunt integrate cu serviciile Azure Databricks
- Serviciile de stocare Azure permit accesarea datelor din Azure Blob Storage și Azure Data Lake Store
- Azure Active Directory permite autentificarea utilizatorilor, eliminând necesitatea menținerii a două familii de utilizări separate în Databricks și Azure.
- Azure SQL DW și Azure Cosmos DB permit combinarea datelor structurate și nestructurate pentru lucrul cuanalitice
- Apache Kafka pentru HDInsight permite lucrul cu fluxuri de date
- Azure Power BI pentru vizualizarea datelor

STRUCTURA AZURE DATABRICKS



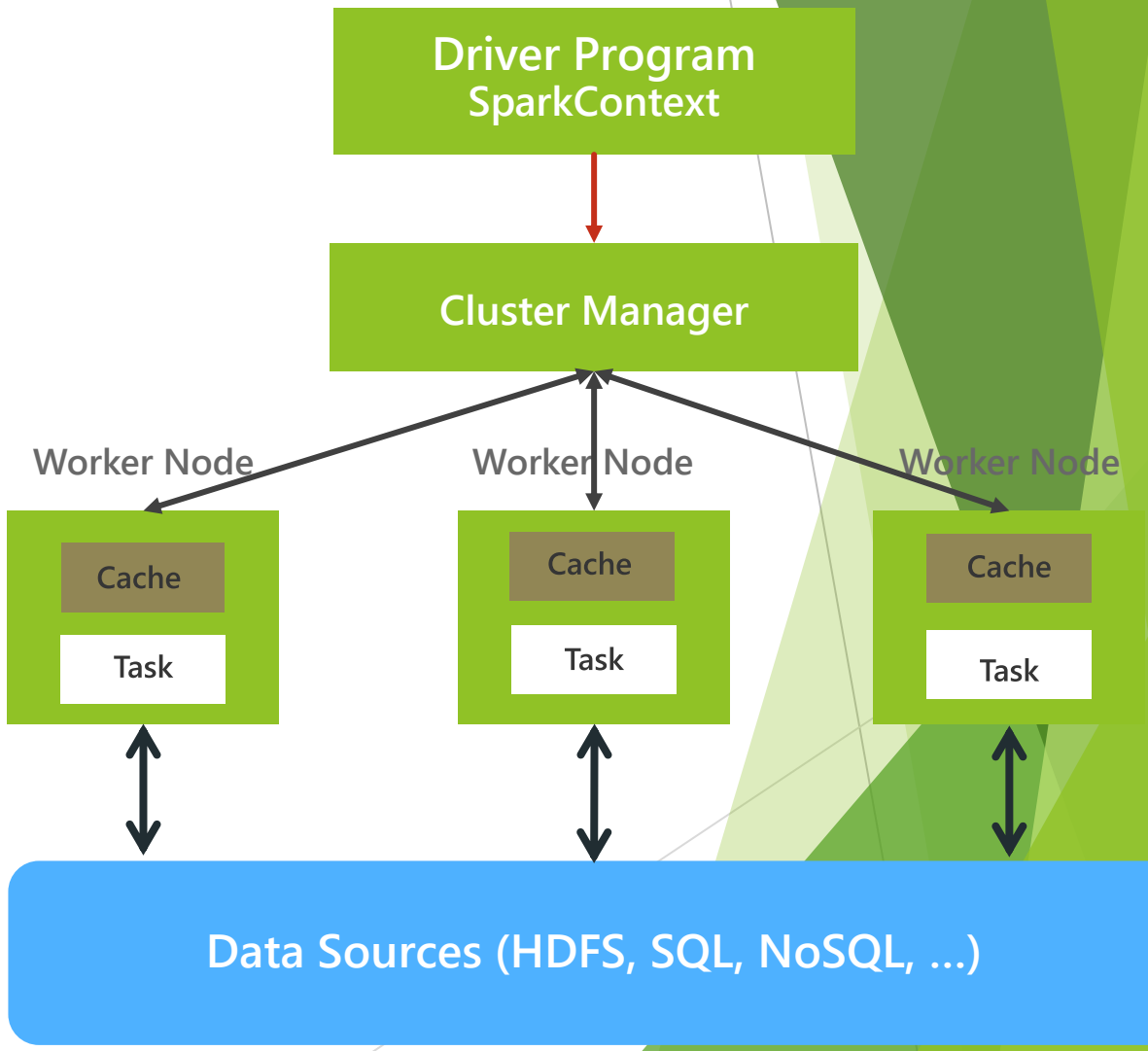
Productivitate îmbunătățită

Build securizat

Scalare fara limite

- Spark Driveste un proces JVM care găzduiește SparkContext pentru o app Spark.
- Spark Context este end-point-ul serviciului Spark(execution engine) și componenta centrală a unei app Spark
- Rezultatele operațiilor sunt colectate de driver
- ”Worker node”-urile citesc și scriu date din/în Sursele de Date incluzând HDFS.
- Un ”worker node” (cache) transformă datele în RDD-uri (Resilient Distributed Data sets).
- ”Worker node”-urile și ”Driver Node”-urile sunt executate ca MV in cloud public (AWS, Google and Azure).

ARHITECTURA CLUSTERULUI SPARK



Spark Context

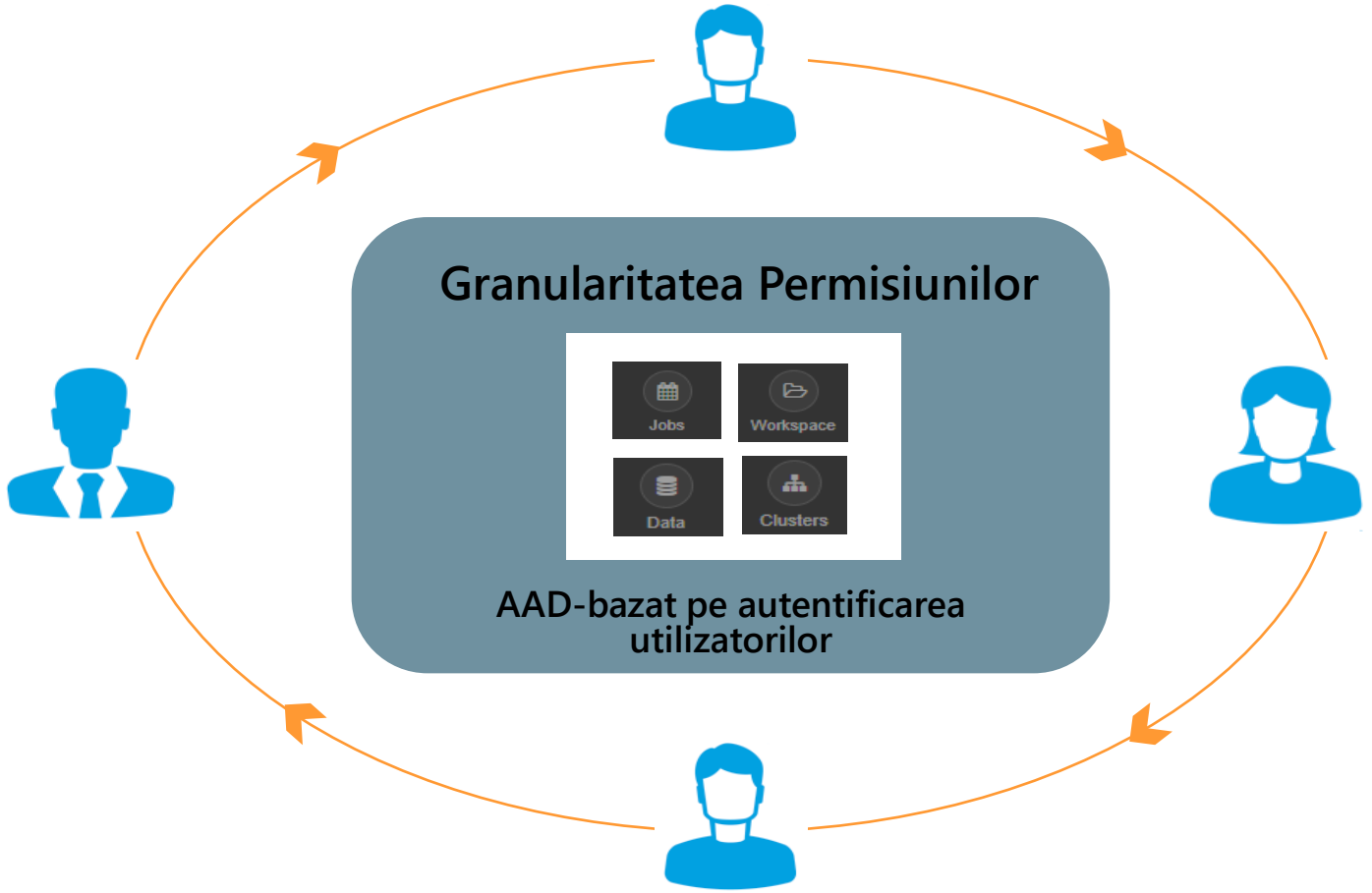
Cmd 11

```
dataPath = "/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv"
diamonds = spark.read.format("com.databricks.spark.csv")\
  .option("header","true")\
  .option("inferSchema", "true")\
  .load(dataPath)

# inferSchema means we will automatically figure out column types
# at a cost of reading the data more than once
```

SparkContext reprezintă punctul de intrare al funcționalităților în Spark. Cel mai important pas al oricărei aplicații driver în Spark este de a generarea SparkContext. Acesta, permite aplicației utilizatorului să acceseze cluster-ul Sparc cu ajutorul Resource Manager. Managerul de resurse poate fi unul dintre : Spark Standalone, YARN, Apache Mesos.

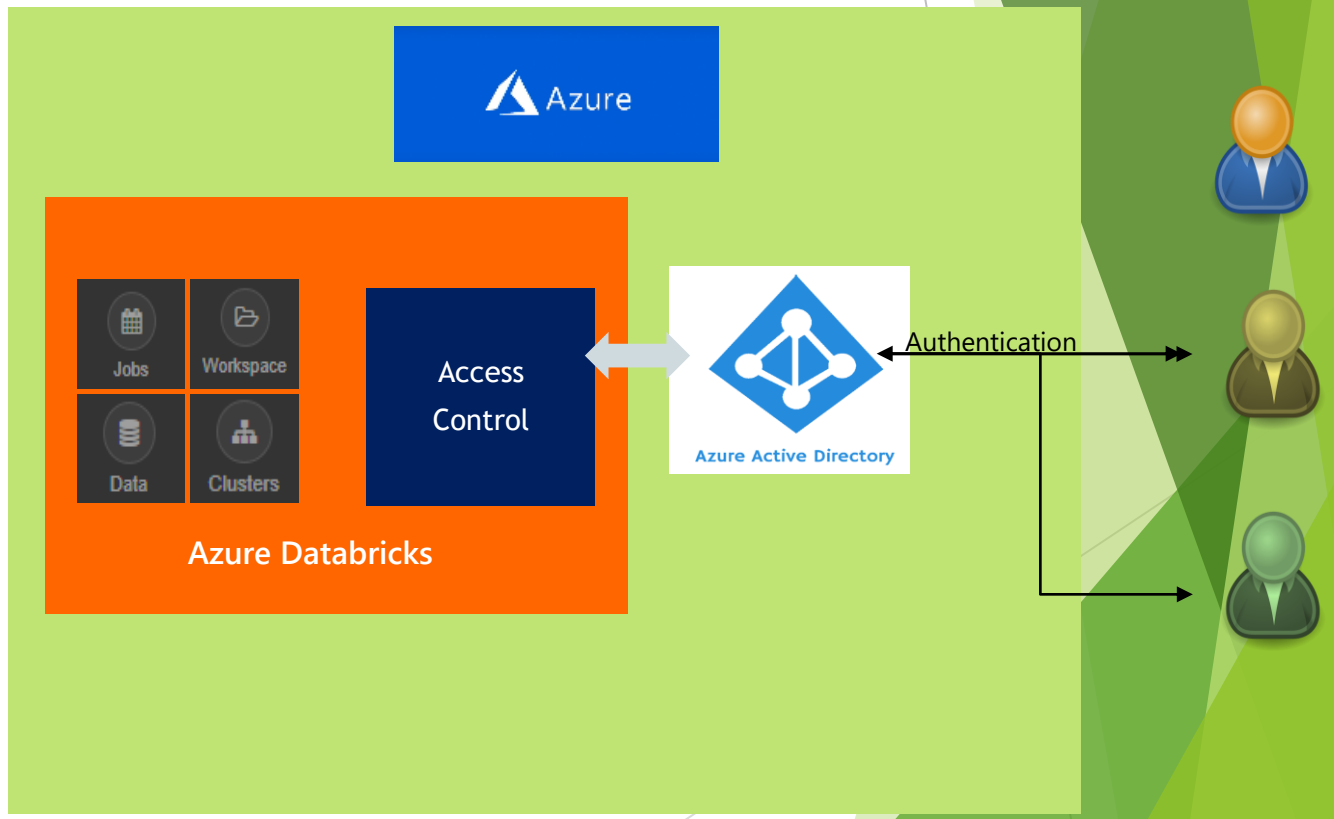
În Azure Databricks se pot partaja în deplină siguranță artefacte cum ar fi Clustere, Notebook-uri, Job-uri și Workspace-uri



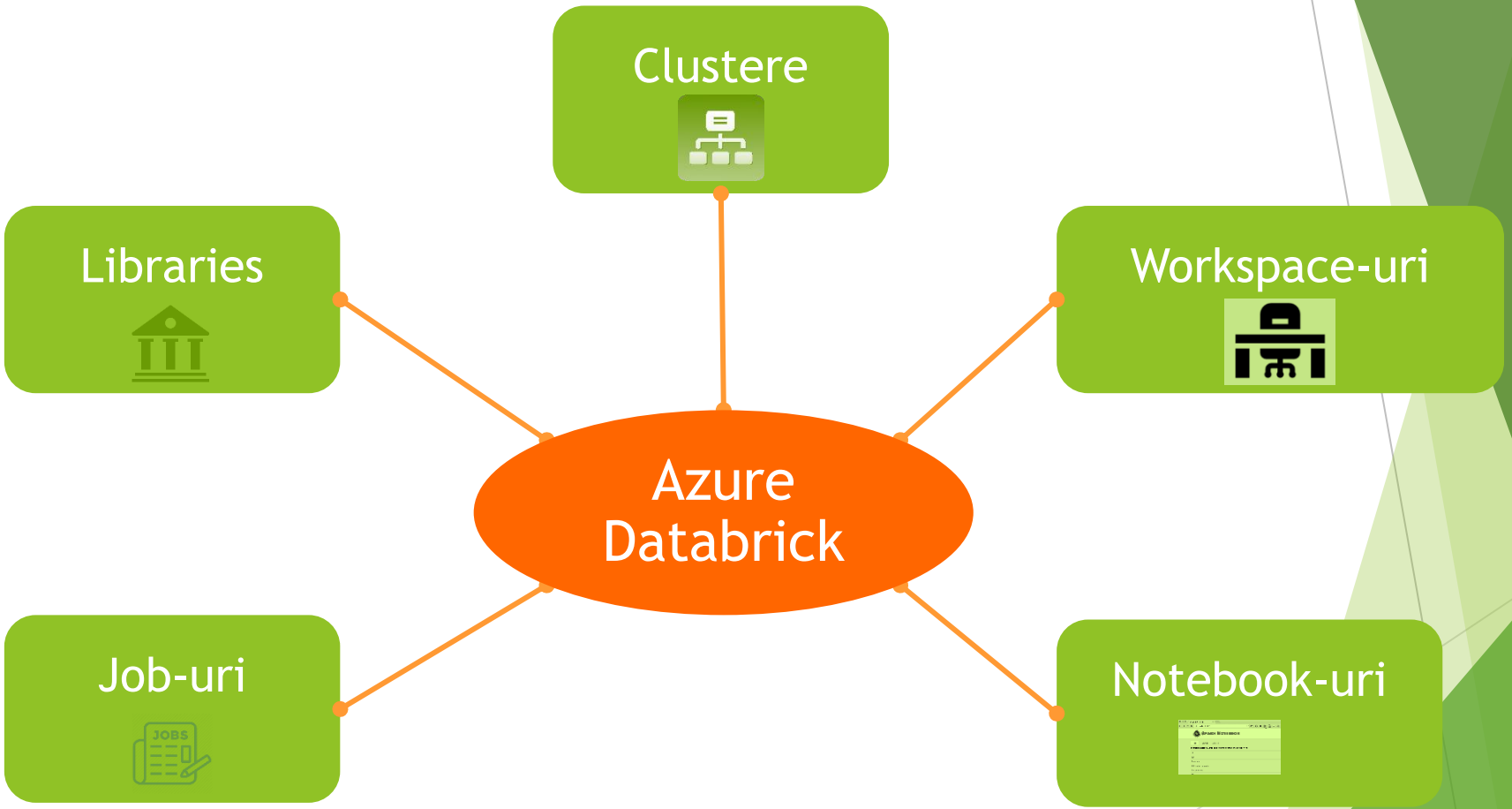
Azure Databricks este un mediu sigur pentru colaborare

AZURE DATABRICKS SE INTEGREAZĂ CU AAD

- Nu este nevoie de definit utilizatori —nivelul permisiunilor la nivelul Databricks nu este necesar.
- Databricks a realizat delegarea și autentificarea catre AAD prin SSO(single-sign on).
- *Notebook-urile, și output-urile asociate, sunt încărcate în contul Databricks*. Totuși, asigură faptul că numai utilizatorii autorizati pot avea access.



ARTEFACTE SPECIFICE AZURE DATABRICKS



Workspace Colaborativ

UN MEDIU DE LUCRU INTUITIV

Access la date prin intermediul notebook-urilor interactive bazate pe limbaje cum ar fi R, Python, Scala, and SQL

COLABORARE

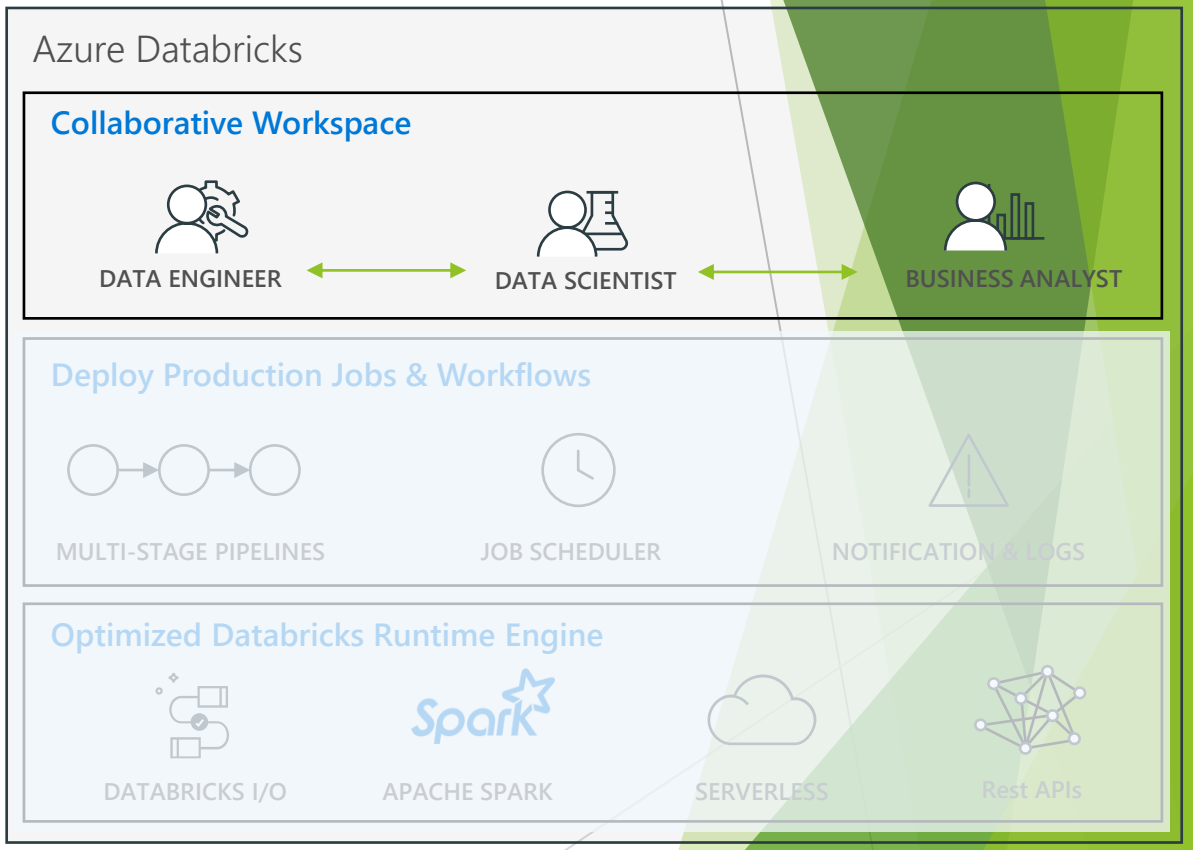
Notebook-ul poate fi modificat în timp real cu posibilitatea de a urmări modificările prin istoricul detaliat al variantelor în GitHub sau Bitbucket

VISUALIZĂRI

Vizualizarea statisticilor printr-o gamă largă de instrumente de tip point-and-click sau prin utilizarea opțiunilor bazate pe scripturi precum matplotlib, ggplot și D3

DASHBOARD

Integrare cu PowerBI pentru a analiza și disemina cunoștințele sintetizate în pattern-uri.



Job-uri de producție & Workflow-uri

PROGRAMATOR DE JOB-URI

Executa job-uri pentru pipeline-urile de producție pentru o programare data

WORKFLOW-URI ASOCIATE NOTEBOOK-URILOR

Creaza pipeline-uri multi-stage cu controlul structurii sursei limbajului de programare

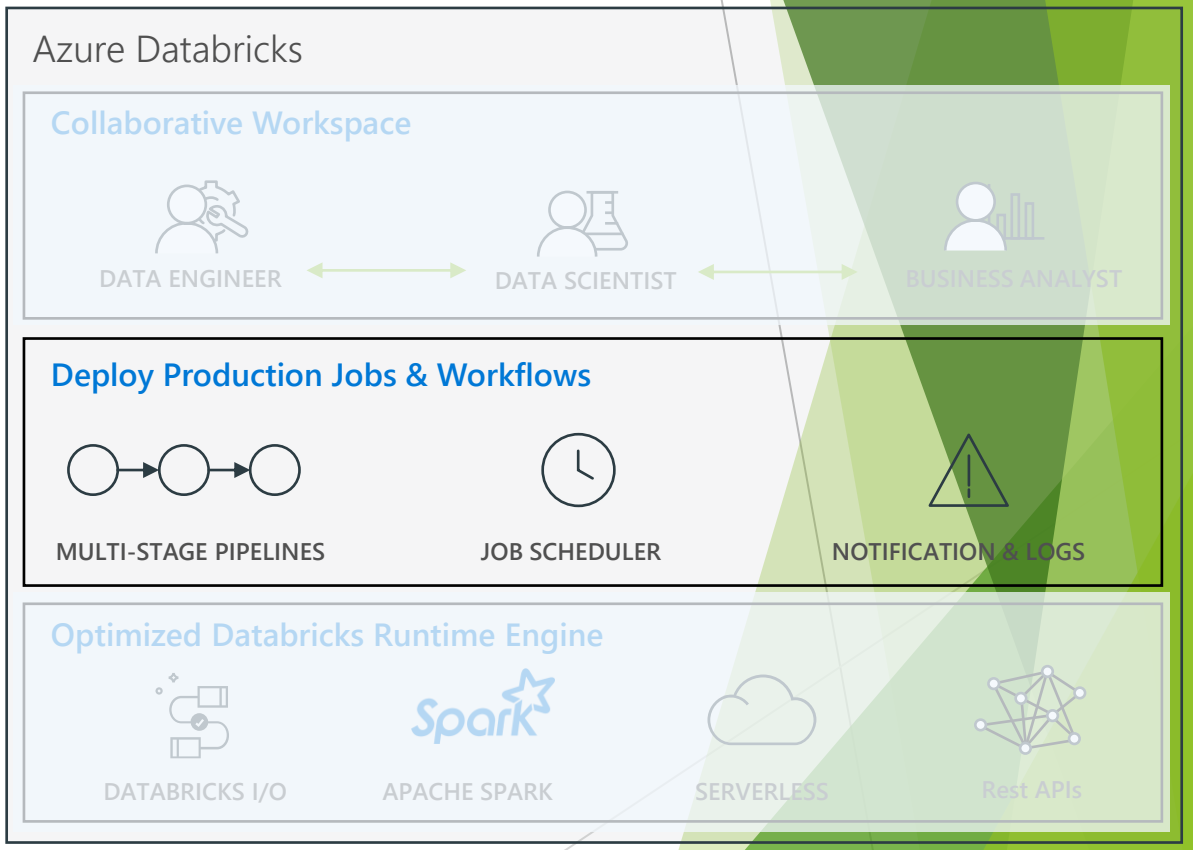
RULEAZĂ NOTEBOOK-URI CA JOB-URI

Plasează notebook-urile sau JAR-ul in job-uri Spark reziliente de o maniera extrem de simpla

NOTIFICATICARI SI LOG-URI

SE INTEGREAZA NATIV CU AZURE SERVICES

Integrare cu: Azure SQL Data Warehouse, Cosmos DB, Azure Data Lake Store, Azure Blob Storage, si Azure Event Hub



DRE(Databricks Runtime Engine) optimizat

PERFORMANTE I/O OPTIMIZATE

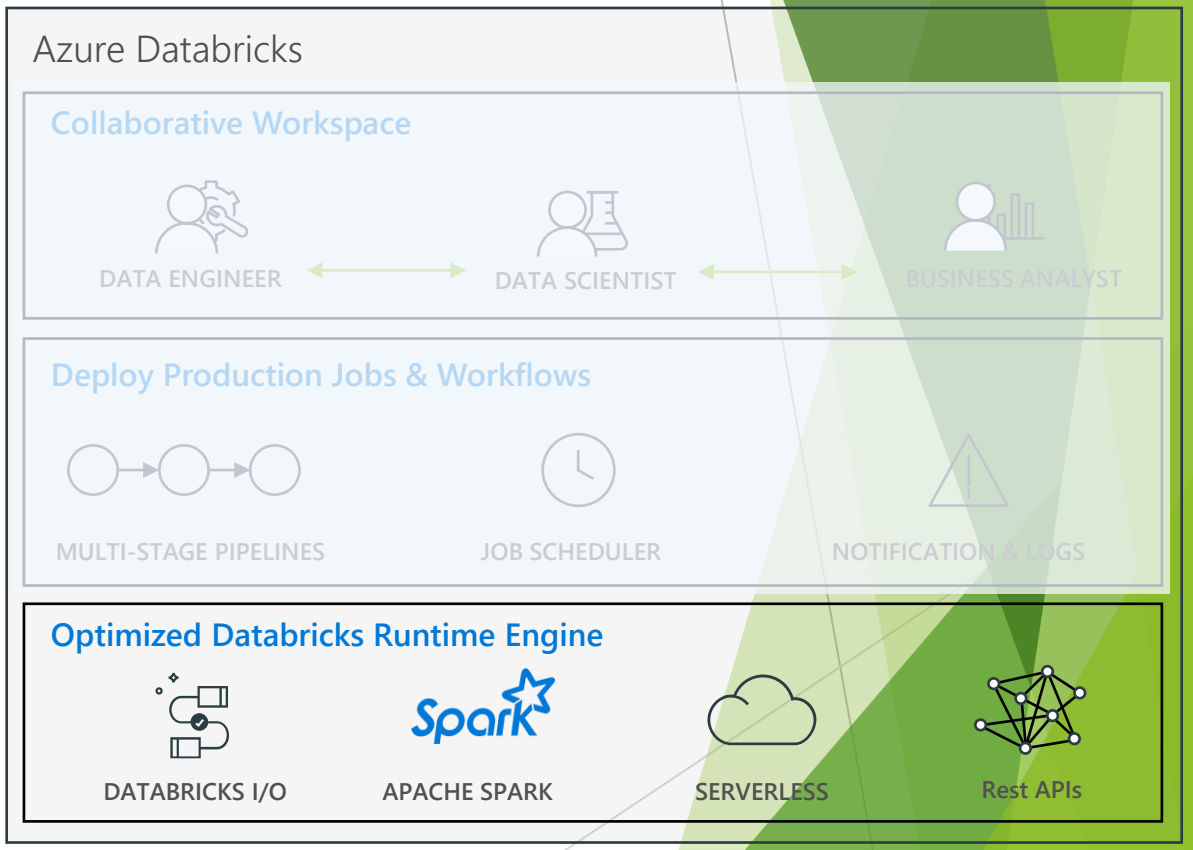
Databricks I/O (DBIO) optimizat trece viteza proceselor la un nivel superior odată cu portarea Spark in cloud

PLATFORMĂ FULL-MANAGED ÎN AZURE

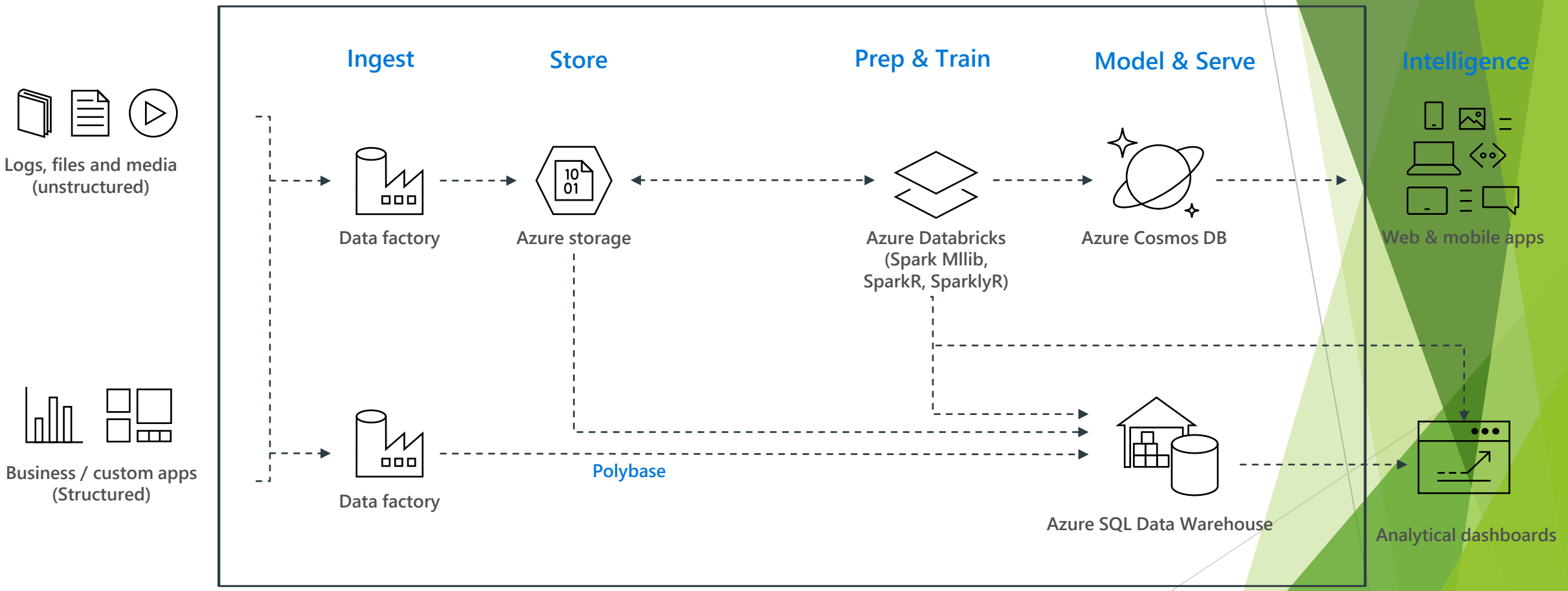
Serviciu complet gestionat ce produce reducerea complexității soluțiilor Big Data și ML într-un context SERVERLESS elastic proiectat a reduce complexitatea operațională.

OPEREAZĂ CU SCALARE MASIVĂ

Fără limite globale



Analitice pentru Big Data



DATABRICKS ACCESS CONTROL

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

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

Databricks Access Control

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

DEMO