Democratizing AI with Apache Spark

Ali Ghodsi
Co-Founder and CEO

databricks™

AI is changing the world

Self-driving cars

SIRI/assistants

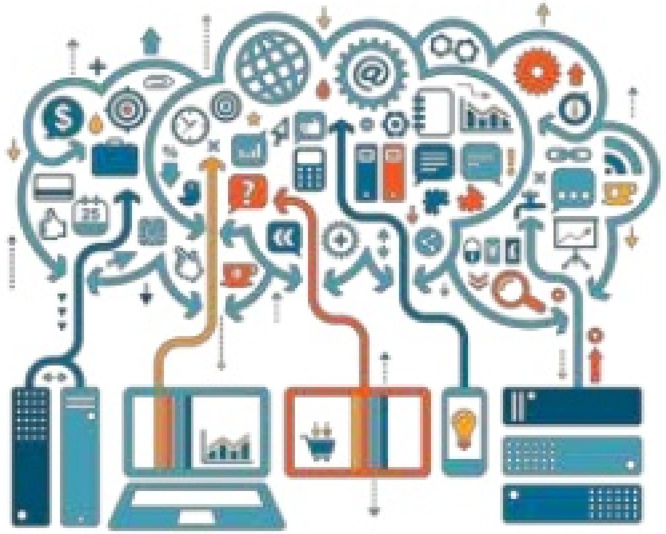
What can I help you with?

AlphaGo

Why now?

Data is the catalyst

More data



Clickstreams

Sensor data (IoT)

Video

Speech

Handwriting

…

Better training, tuning, validation

AI hasn't been democratized

The hardest part of AI isn’t AI

“Hidden Technical Debt in Machine Learning Systems “, Google NIPS

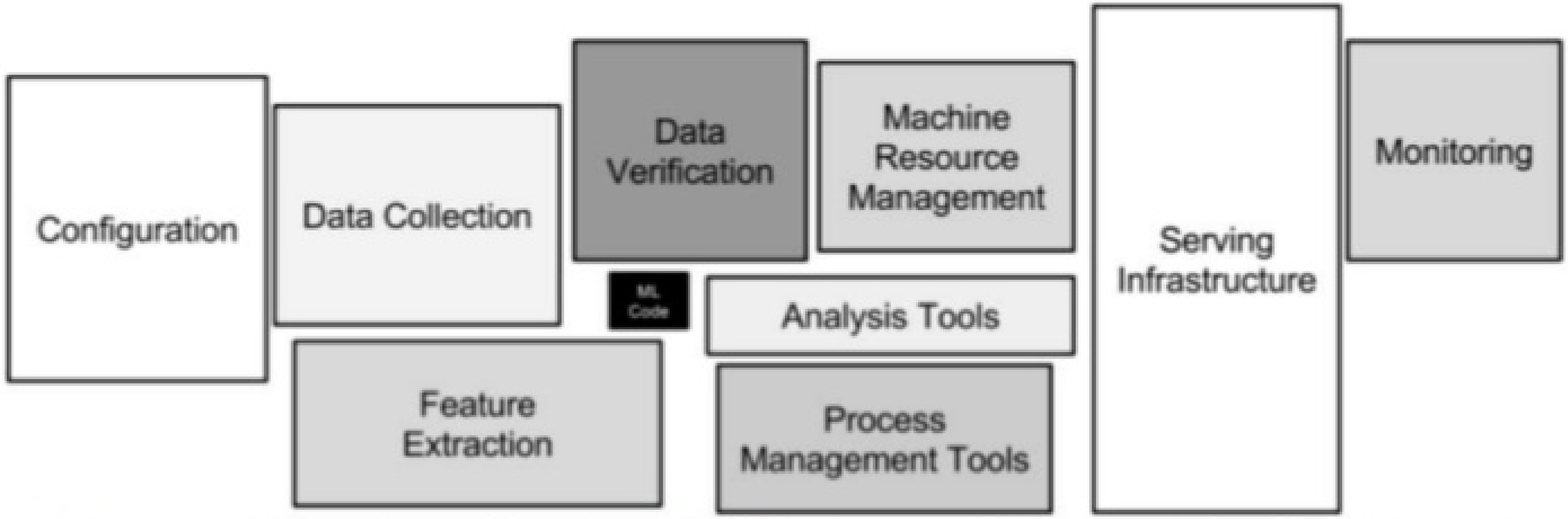


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

How do we democratize AI?

APACHE Spark + AI

"Hidden Technical Debt in Machine Learning Systems ", Google

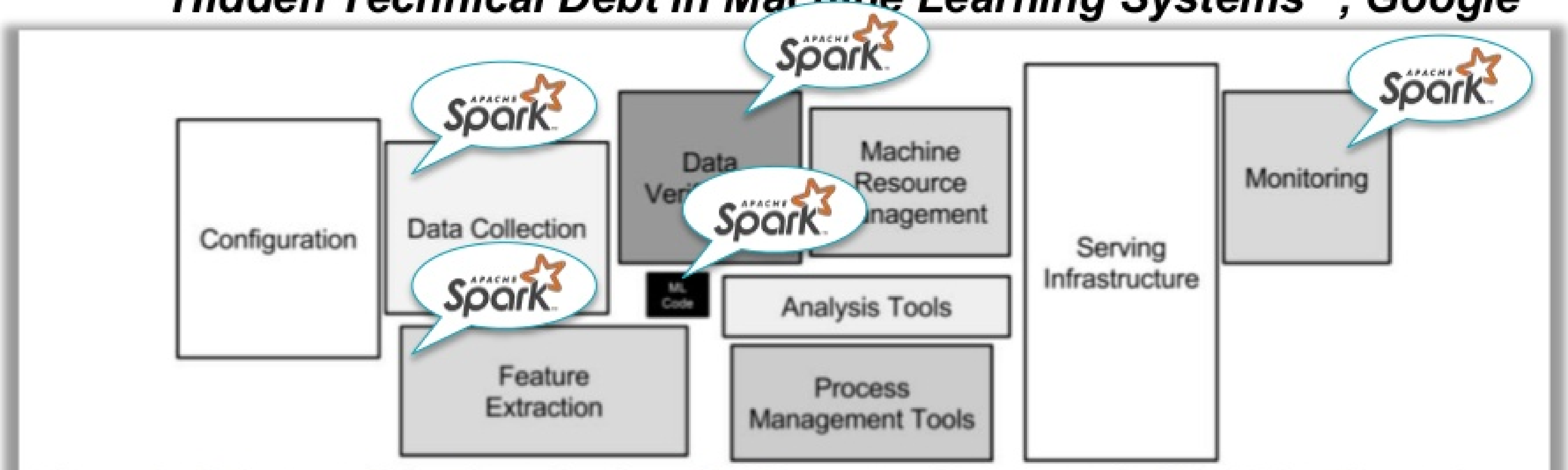


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

FLEXIBLE

FAST

BIG DATA

Some gaps remain

1 **Manage Data infrastructure**

- Create, configure, monitor resilient **big data clusters**.
- **Securely** access silos of **disparate data sources**.
- Enforce **proper data governance**.

2 **Empower teams to be productive**

- **Interactively explore** data and prototype ideas.
- Securely share big data clusters among analysts.
- Debug, troubleshoot, version-control big data applications.

3 **Establish Production-Ready Applications**

- Setup **robust ML data pipelines** for ETL/ELT.
- **Productionize real-time** applications with HA, FT.
- Build, serve, maintain advanced machine learning models.

Databricks: Closing the gap

1 Just-in-Time Data Platform

- Separate compute & storage
- Integrate existing data stores
- Efficient cache on first access

Agile + Low TCO

2 Integrated Workspace

- Interactive notebooks, dashboards, reports
- Real-time exploration, machine learning, graph use cases

Accelerate Time to Value

3 Automated Spark Management

- Workflow scheduler for ML, streaming, SQL, ETL
- Performance-optimized, high availability, fault-tolerant

Performance

Enterprise AI use-cases

CapitalOne	Predict credit score, credit limit, anomalies
DNV·GL	Predict energy demand based on massive weather data
ELSEVIER	Natural language processing to extract author graph
RIOT GAMES	Predict player churn, predicting network outages
SIEMENS	Predict machine equipment failure

New Frontier of AI: Deep Learning

Detect cancer

Improve cancer detection

Understand speech

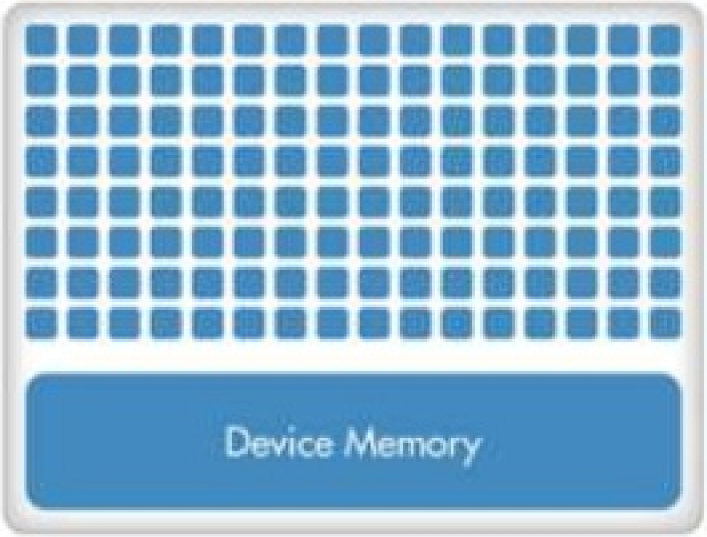
Recognize Mandarin and English

Infer location

Identify landmarks in photos

Faster and easier deep learning with Databricks

GPUs



Massive parallelism

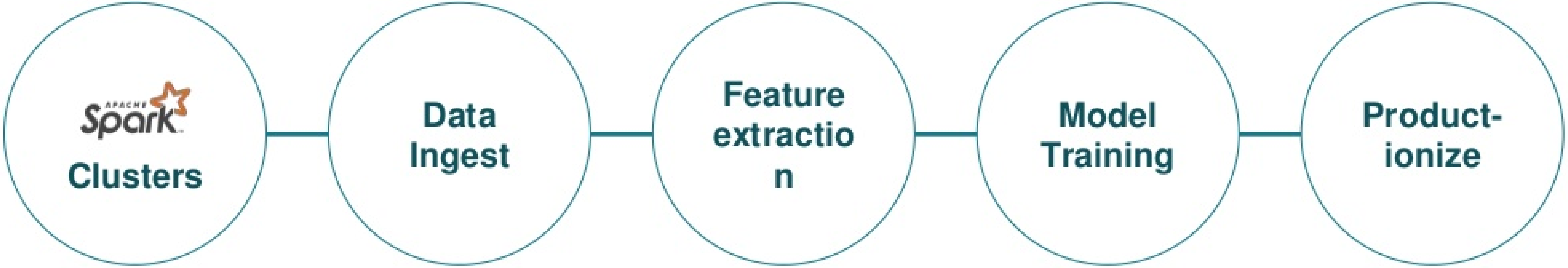
+

TensorFlow on Spark

- TensorFlow: The most popular deep learning framework.
- TensorFrames: Makes TensorFlow computations faster and easier to program on Spark.

TensorFrames and GPUs support out-of-the-box

Deep Learning on Databricks



Automated management

Just-in-time data platform

Interactive exploration

TensorFrames

+

GPUs

Jobs & Workflows

Thank you.

databricks™

Deep Learning references

- Image recognition (Geo ID):
  - https://www.technologyreview.com/s/600889/google-unveils-neural-network-with-superhuman-ability-to-determine-the-location-of-almost/
- Cancer screening:
  - http://www.popsci.com/how-deep-learning-technology-could-be-next-step-in-cancer-detection
  - https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/
- Speech translation:
  - https://www.technologyreview.com/s/544651/baidus-deep-learning-system-rivals-people-at-speech-recognition/

Analytics Transforming Industries

PHARM
A

gsk

Predicting Diabetes
in Rural Counties

Next-Gen Product R&D

MEDIA

NBC UNIVERSAL

Generating programs
based on Nielsen ratings

Predictive Analytics

INDUSTRIA
L

SIEMENS

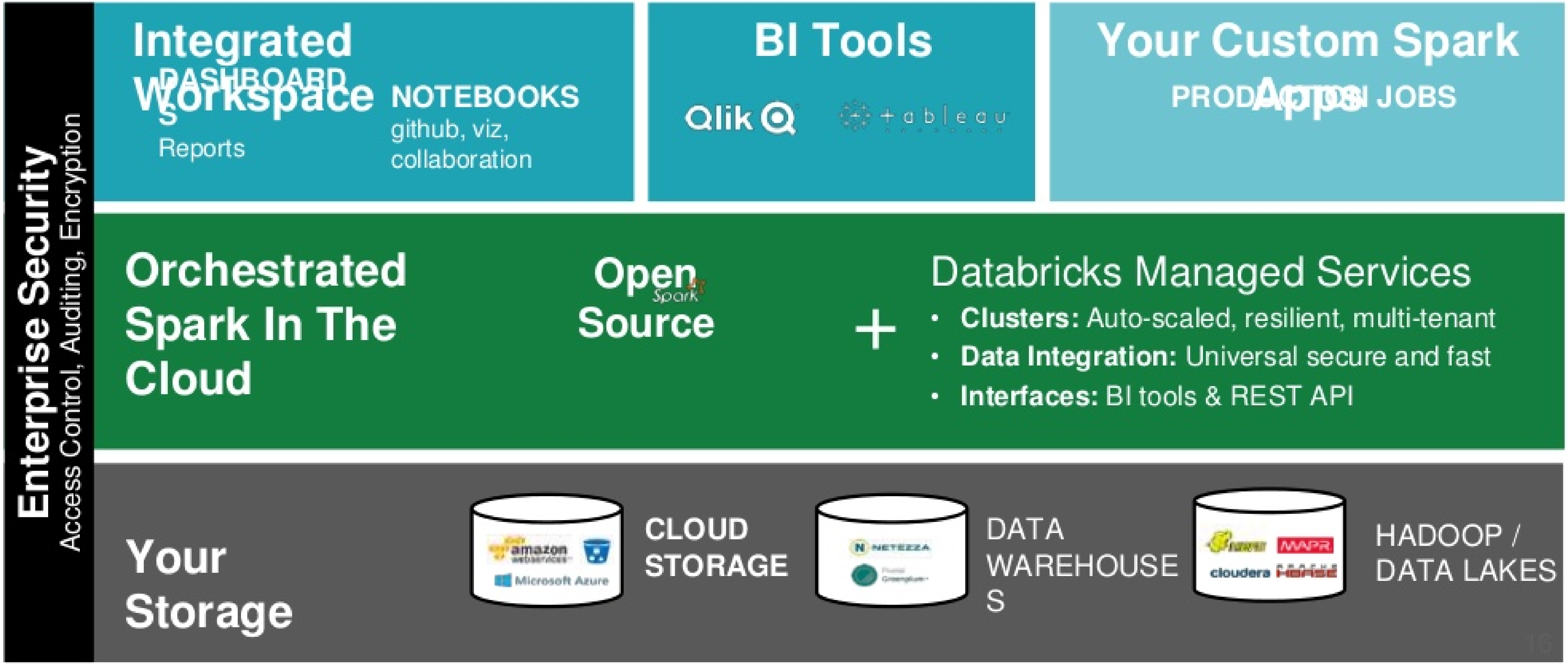
Real-time detection
of failing wind-turbines

Anomaly Detection

Real-time Data-Driven Analytics Applications

Databricks Just-in-Time Data Platform

Powered by Apache Spark



Enterprise Security

Access Control, Auditing, Encryption

Integrated Workspace

DASHBOARDS

Reports

NOTEBOOKS

github, viz, collaboration

BI Tools

Your Custom Spark Apps

PRODUCTION JOBS

Orchestrated Spark In The Cloud

Open Source

+

Databricks Managed Services

- **Clusters:** Auto-scaled, resilient, multi-tenant
- **Data Integration:** Universal secure and fast
- **Interfaces:** BI tools & REST API

Your Storage

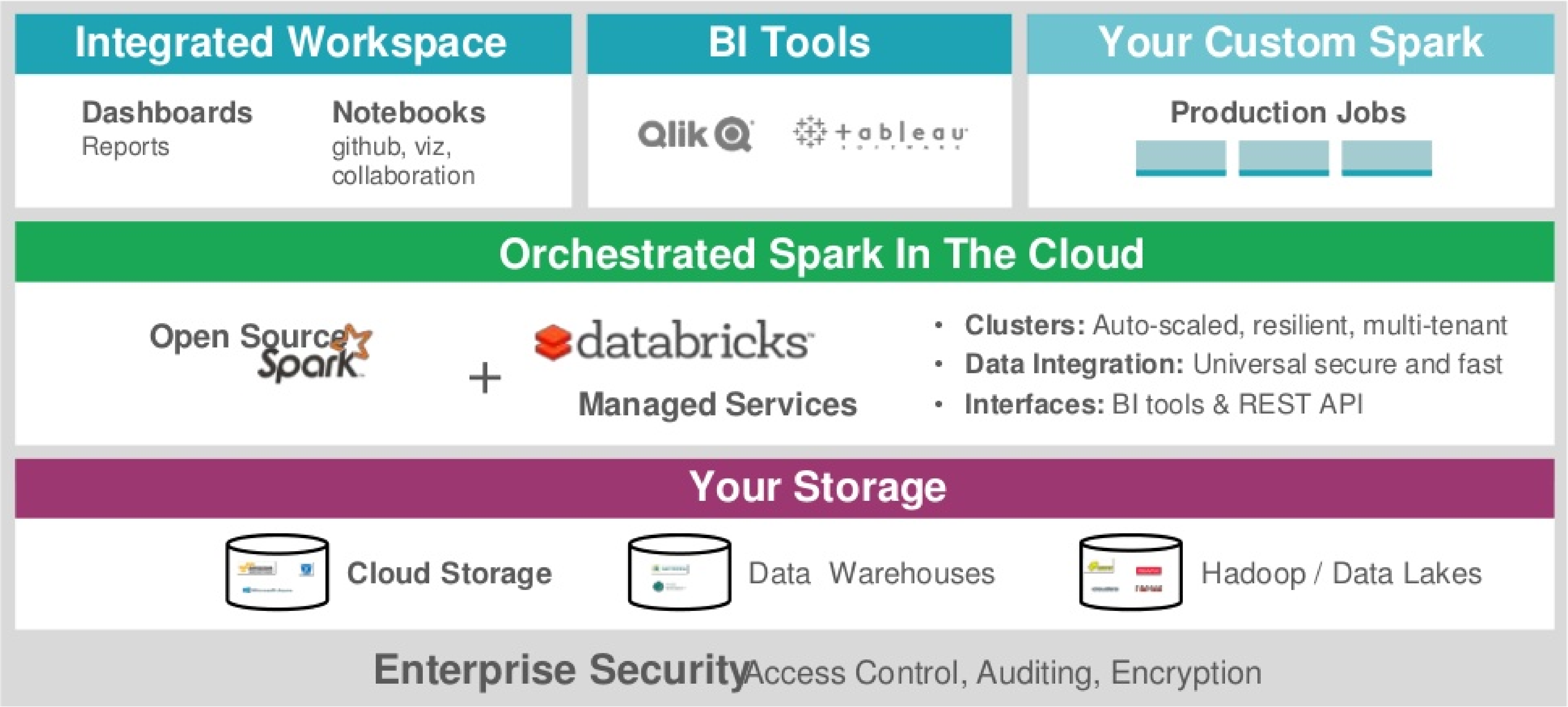
CLOUD STORAGE

DATA WAREHOUSES

HADOOP / DATA LAKES

Databricks Just-in-Time Data Platform

Powered by Apache Spark



Integrated Workspace

Dashboards
Reports

Notebooks
github, viz, collaboration

BI Tools

Qlik

tableau

Your Custom Spark

Production Jobs

Orchestrated Spark In The Cloud

Open Source Spark + databricks **Managed Services**

- **Clusters:** Auto-scaled, resilient, multi-tenant
- **Data Integration:** Universal secure and fast
- **Interfaces:** BI tools & REST API

Your Storage

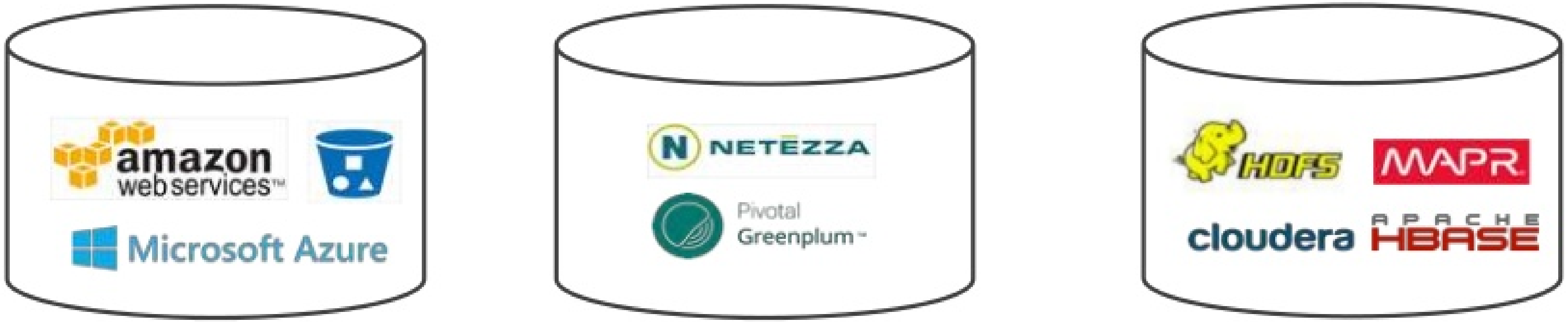
Cloud Storage

Data Warehouses

Hadoop / Data Lakes

Enterprise Security Access Control, Auditing, Encryption

Today’s Data Reality



Cloud Storage

Data Warehouses

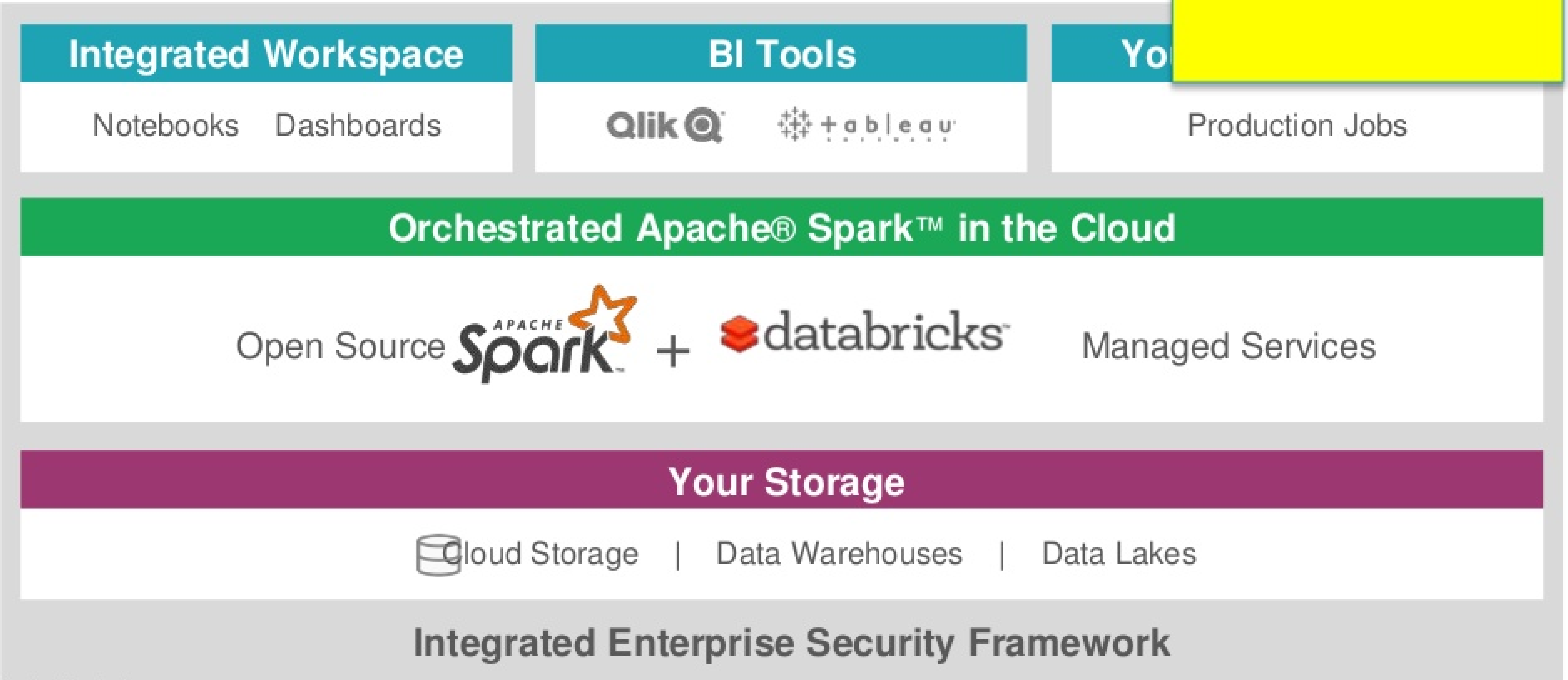
Hadoop / Data Lakes

Siloed, Unstructured, Fast-Growing Data

Databricks Just-in-Time Data

Powered by Apache® Spark™

OPTIONAL



Integrated Workspace

Notebooks Dashboards

BI Tools

Qlik tableau

Yo

Production Jobs

Orchestrated Apache® Spark™ in the Cloud

Open Source Apache Spark + databricks Managed Services

Your Storage

Cloud Storage | Data Warehouses | Data Lakes

Integrated Enterprise Security Framework