Disrupting Big Data with Apache Spark in the Cloud

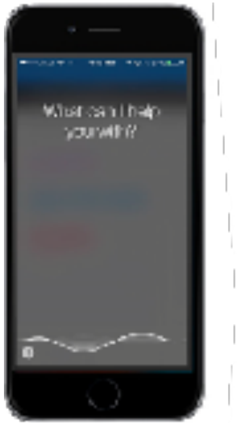
Ali Ghodsi

databricks™

The Dawn of Advanced Analytics

Self-driving cars

SIRI/assistants

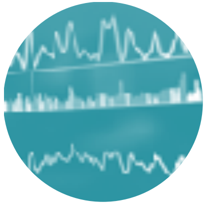


Watson

Not just sci-fi, important applications for businesses

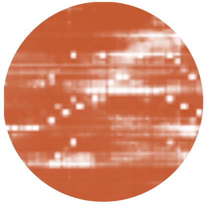
Analytics Transforming Industries

Predictive analytics



Predict Product Revenue
Customer Assessment
Targeted Advertising

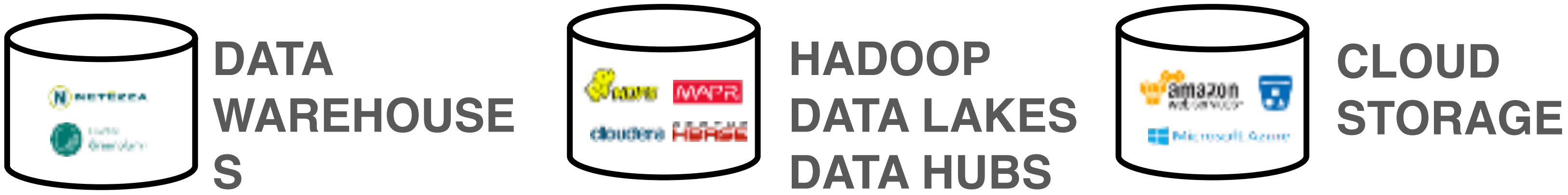
Anomaly Detection



Fraud Detection
Risk Assessment
Equipment Failure

Data-Driven Real-time Analytics Applications

Today’s Data Reality



DATA
WAREHOUSE
S

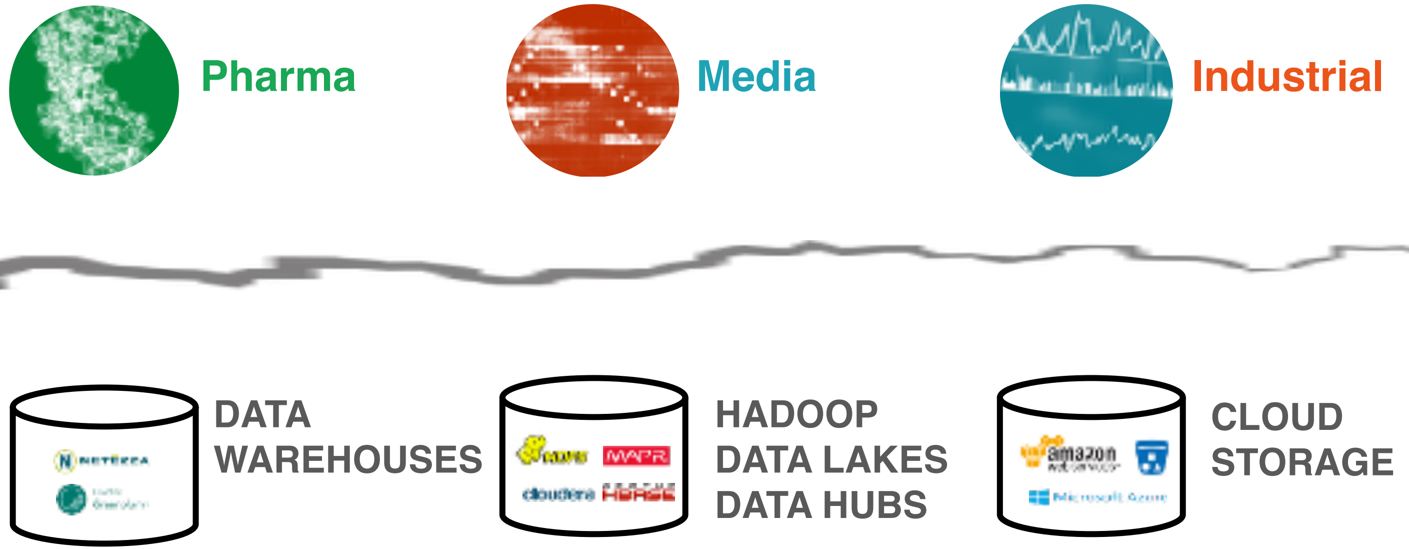
HADOOP
DATA LAKES
DATA HUBS

CLOUD
STORAGE

Siloed, Fast-Growing Size, Cost

The Analytics Gap

Real-time Data-Driven Analytics Applications



Pharma

Media

Industrial

DATA WAREHOUSES

HADOOP DATA LAKES DATA HUBS

CLOUD STORAGE

Siloed, Fast-Growing Size, Cost

Why is there a gap?

Real-time Data-Driven Analytics Applications

1 **Manage Data infrastructure**

- Create, tune, monitor **compute clusters**.
- **Securely** access silos of **disparate data sources**.
- Enforce **proper data governance**.

2 **Empower teams to be productive**

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

3 **Establish Production-Ready Applications**

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

Siloed, Fast-Growing Size, Cost

Databricks Cloud-Hosted Platform

1 **Just-in-Time Data Platform**

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

Agile

2 **Integrated Workspace**

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

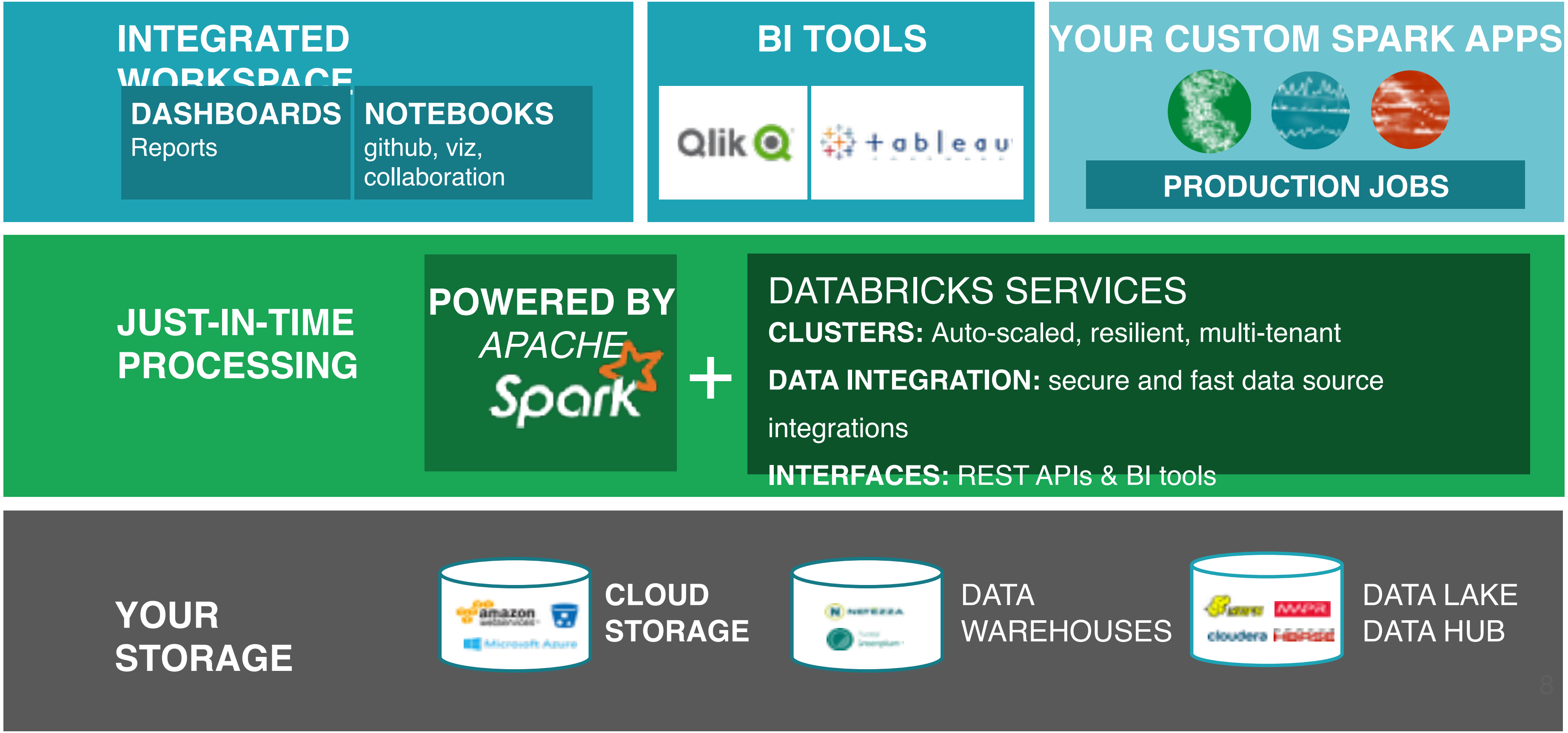
Democratize Big Data

3 **Automated Apache Spark Management**

- Workflow scheduler for ML, streaming, SQL, ETL
- High availability, fault-tolerant, performance-optimized

Production-Ready

Databricks Just-in-Time Data Platform



INTEGRATED WORKSPACE

DASHBOARDS
Reports

NOTEBOOKS
github, viz, collaboration

BI TOOLS

YOUR CUSTOM SPARK APPS

PRODUCTION JOBS

JUST-IN-TIME PROCESSING

POWERED BY
APACHE
Spark

+

DATABRICKS SERVICES

CLUSTERS: Auto-scaled, resilient, multi-tenant

DATA INTEGRATION: secure and fast data source integrations

INTERFACES: REST APIs & BI tools

YOUR STORAGE

CLOUD STORAGE

DATA WAREHOUSES

DATA LAKE
DATA HUB

The Challenge of Securing Analytics

End-to-end security a challenge for enterprises

Securing file management

Secure table management

Secure cluster management

Secure job workflows

Secure dashboards, report, notebook management

Today there are piecemeal solutions, but no comprehensive solution

Databricks Enterprise Security (DBES)

Holistic end-to-end security for Data Analytics

Files

Tables

Clusters

Workflows

Notebooks, Dashboards, Reports

DBES provides

- **Role-based access control**
- **Auditing and governance**
- **Integrated identity-management**
- **Encryption on-disk and on-the-wire**

The First End-to-End Security Solution for Apache Spark

Enterprise use-cases

Capital One	Preventing credit card fraud
DNV·GL	Predict energy demand based on massive weather data
Riot GAMES	Predict player churn, predicting network outages
ELSEVIER	Natural language processing to extract author graph
NBC UNIVERSAL	Generating tailored programs based on big data

Thank you.

databricks™

Try Apache Spark with Databricks

Try latest version of Apache Spark and preview of Spark 2.0

http://databricks.com/try