



DATABRICKS

Spark Summit

June 2014

Apache Spark and Databricks

Adoption

All major Hadoop distributions include Spark



Beyond Hadoop



Partnerships

Partner with Spark distributors to provide great experience to **every** Spark user

Partners

The Cloudera logo, featuring the word "cloudera" in a bold, blue, sans-serif font with a registered trademark symbol.The DataStax logo, featuring the word "DATASTAX" in a bold, black, sans-serif font, followed by a circular icon composed of several blue dots of varying sizes.The MapR logo, featuring the word "MAPR" in a bold, white, sans-serif font inside a red rectangular box.The SAP logo, featuring the word "SAP" in a bold, white, sans-serif font inside a blue rectangular box with a white diagonal line.

Certification

Build a strong application ecosystem

Spark Apps



App Cert



Spark API

Spark Distros



Distros Cert



Certification

Free certification process

Scripts for certifying Spark distributions

- Developed by community
- Open-source

Anyone will be able to certify any Spark distribution

Training

We've been teaching Spark since 2012

- 400+ people this year through Databricks

Just launched a new training program

- Already hold workshops in 5 cities

300+ people signed up for training on Wednesday

Solve Big Data Challenges

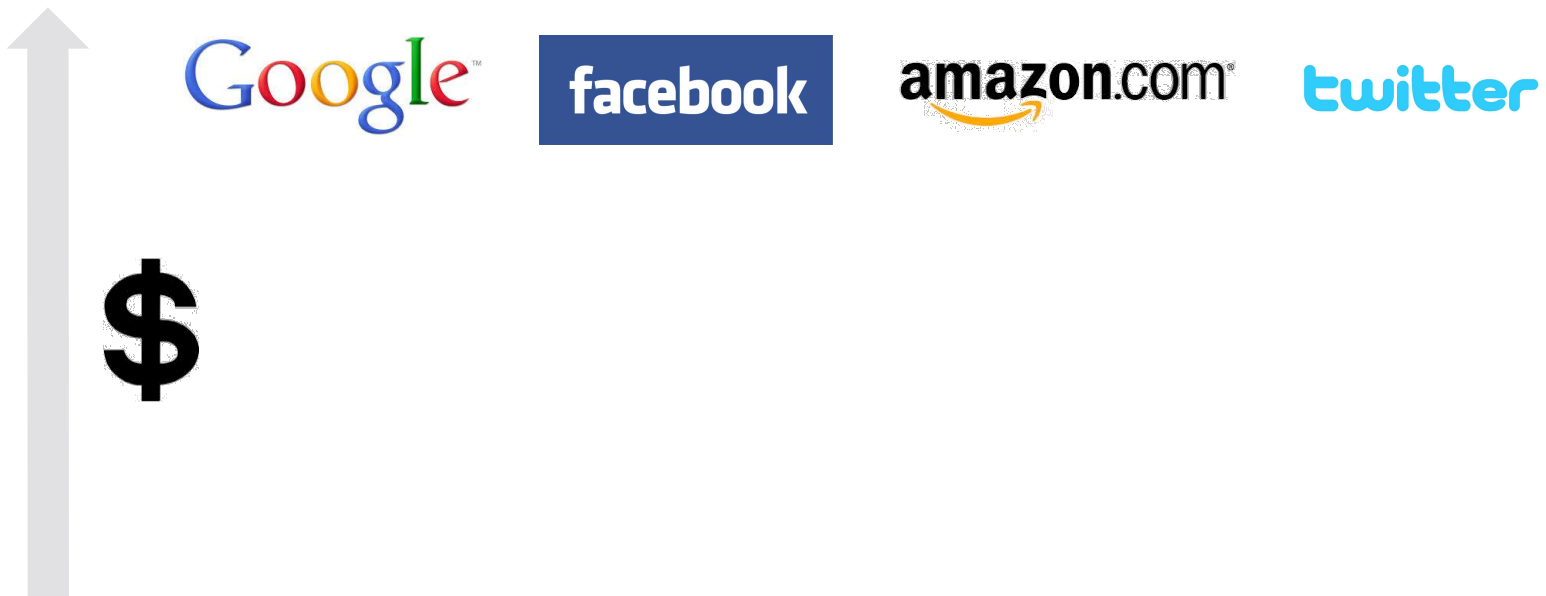
Big Promise

Great successes using Big Data



Big Promise

Great successes using Big Data



Every organization collects data

Your company here!

Big Challenge

Great successes using Big Data



Google™

facebook

amazon.com®

twitter

?

Google, Facebook spend billions \$ to develop, implement, and run data analysis tools and products

Every organization collects data

Your company here!

Typical Story

Your company starts a Big Data initiative

You are tasked to...



1) Build a Hadoop cluster
(IT)



Clusters hard to set up
and manage

2) Build a data pipeline
(engineers, data scientists)



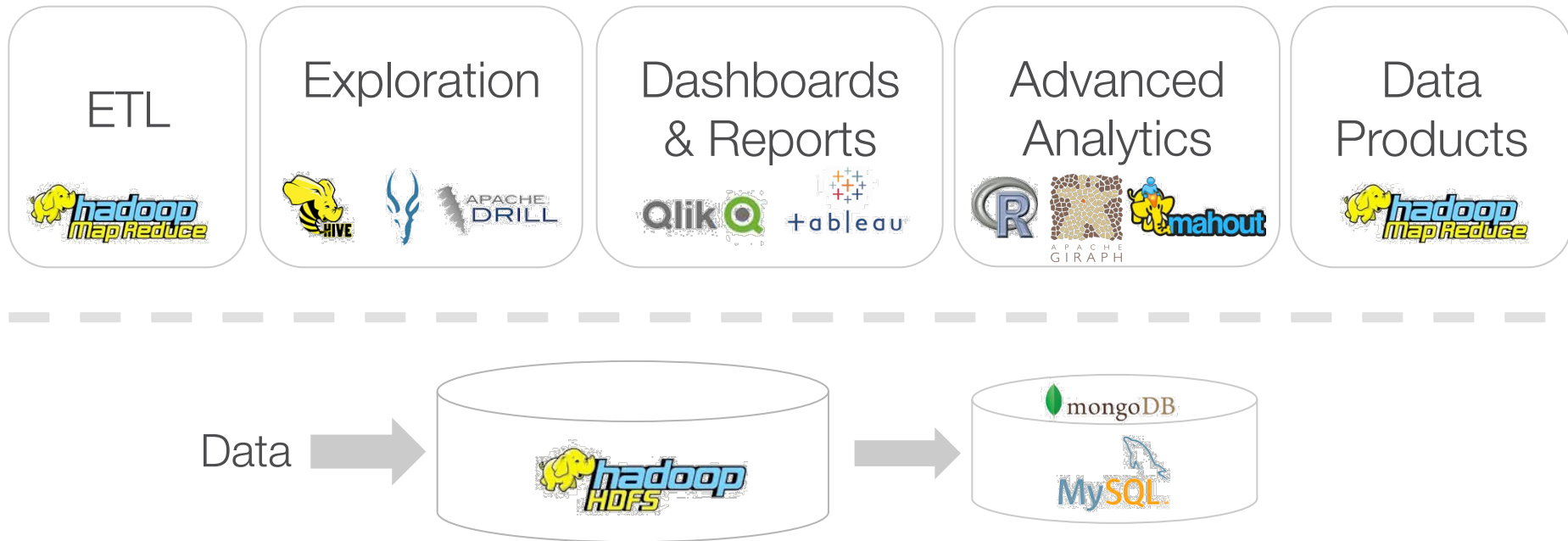
Need to integrate a zoo
of tools

3) Get insights &
build data products
(engineers, data scientists, analysts)



Tools are hard to use

Typical Data Pipeline



Integrate disparate, clunky tools
Hard to navigate data, develop and deploy apps

Vision

Make big data easy

From Challenges to Solutions

Challenges	Solutions
Clusters hard to set up and manage	Hosted platform
Need to integrate a zoo of tools	Apache Spark
Tools are hard to use	Interactive Workspace

Databricks Cloud

Databricks Workspace



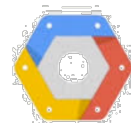
Databricks Platform

Databricks Platform

Databricks Workspace



Databricks Platform



Google Cloud Platform

Databricks Platform

Clusters							
+ Add Cluster							
Name	Memory	State	Nodes	Notebooks	Dashboard		
▶ Test12	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossei...	Current	restart	
▶ TD's Cluster	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, Test	Activate	restart	
▶ AWS direct	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossei...	Current	restart	
▶ Matei Dev	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, Test	Activate	restart	
▶ AndyK	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossei...	Current	restart	
▶ App Store	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, Test	Activate	restart	
▶ Shopping Cart	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossei...	Current	restart	
▶ IonData	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, Test	Activate	restart	

Start clusters in seconds

Zero-cost management

Dynamically scale up & down

Apache Spark

Databricks Workspace



Databricks Platform

Unifies

- Streaming
- SQL
- Machine learning
- Graphs

Single system,
single API

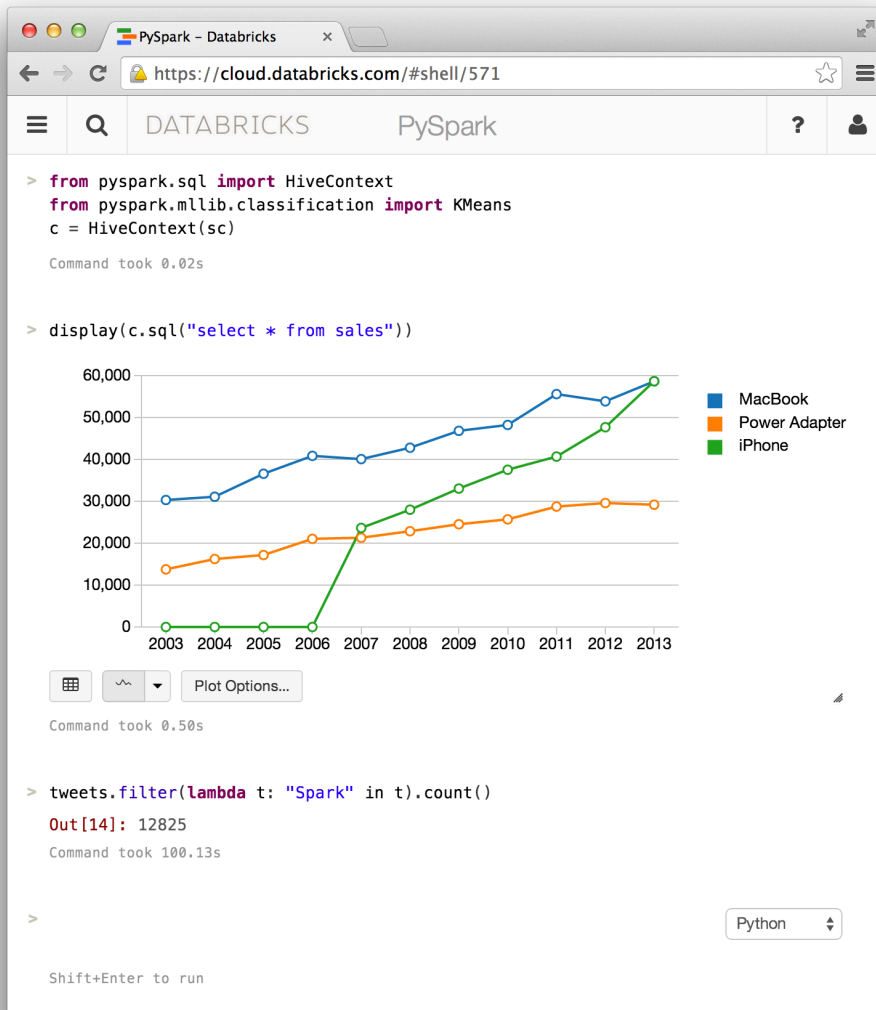
Databricks Workspace

Databricks Workspace



Databricks Platform

Notebooks

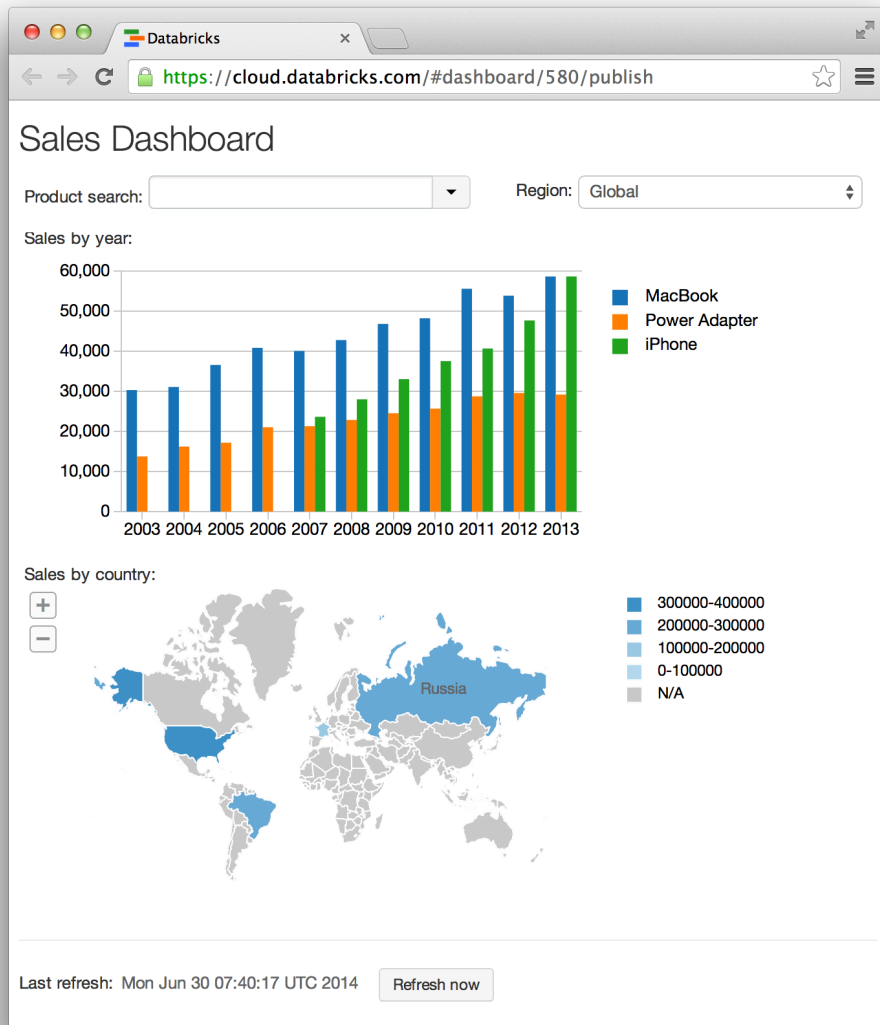


Support Python, SQL, Scala

Interactive commands & plots

On-line collaboration

Dashboards



WYSIWYG builder

Interactive plots

One-click publishing

Job Launcher

Elastic Jobs New Job		
Job Name ▼	Recent Runs	Active
Market Basket Analysis	05/26/2014 06/02/2014 06/09/2014 ...	Today at 8:57 PM Cluster: Default Cluster Actions + databricks/analysis/transform.jar : Triggers + Every Week: Monday
		Minimize Remove
Sales Dashboard ETL	Today at 5:00 PM Today at 6:00 PM Today at 7:00 PM Today at 8:00 PM Today at 9:00 PM Today at 10:00 PM Today at 11:00 PM ...	Tomorrow at 12:00 AM Cluster: Default Cluster
		Edit Remove
Fraud Model Training	06/09/2014 Last Tuesday at 1:00 AM Last Wednesday at 1:00 AM Last Thursday at 1:00 AM Last Friday at 1:00 AM Last Saturday at 1:00 AM Yesterday at 1:00 AM ...	Actions + databricks/ml/training.jar : Triggers + Daily: 1am
		Minimize Remove

Run arbitrary Spark jobs, programmatically

Dramatically Simplify Data Pipeline

ETL

Exploration

Advanced Analytics

Dashboards & Reports

Data Products

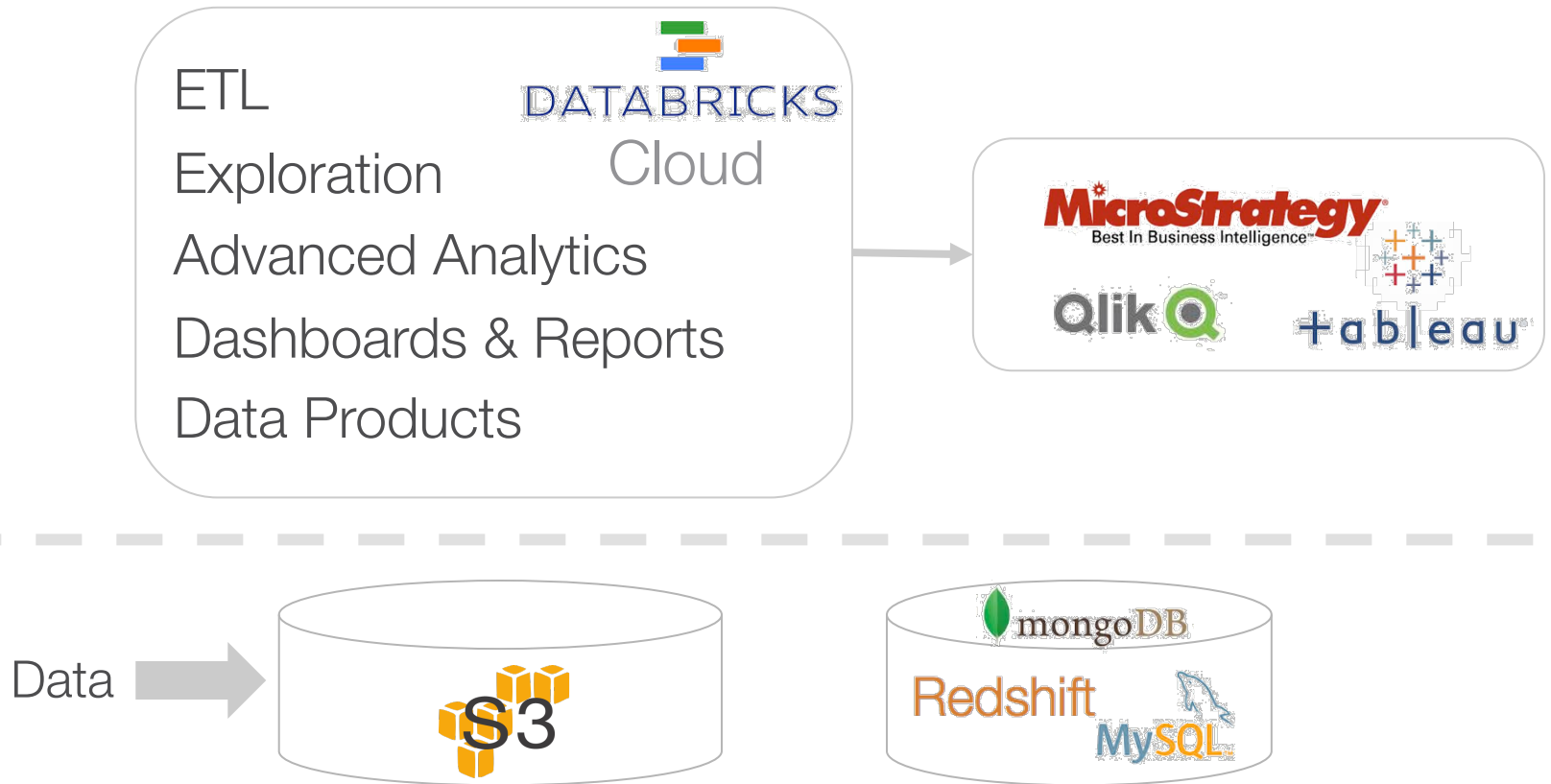

DATABRICKS

Cloud

Data



Dramatically Simplify Data Pipeline



Free users to focus on
finding answers & building products

Demo

Availability

Started closed beta program earlier this year

Limited availability soon

- Gradually ramping up
- Sign up on databricks.com!

3rd Party Apps

Databricks
Workspace



Databricks Platform

3rd Party Apps

Databricks
Workspace

Apps



Databricks Platform

Databricks Cloud and Spark

Databricks Cloud runs 100% Apache Spark

- **No lock in:** any Databricks Cloud app runs on any certified Spark distribution

Databricks Cloud accelerates Spark adoption

- Provide easiest way to learn and use Apache Spark

Databricks Cloud

Dramatically simplify

- analyzing big data
- building data products

Databricks Workspace



Databricks Platform

Fuel growth of Spark ecosystem

Make big data easy



DATABRICKS

Thank You!