

# Harnessing the Power of Spark with Databricks Cloud

Ion Stoica

March 18, 2015



# Accelerating Spark Adoption

Applications  
(35+)



Distributions  
(11+)



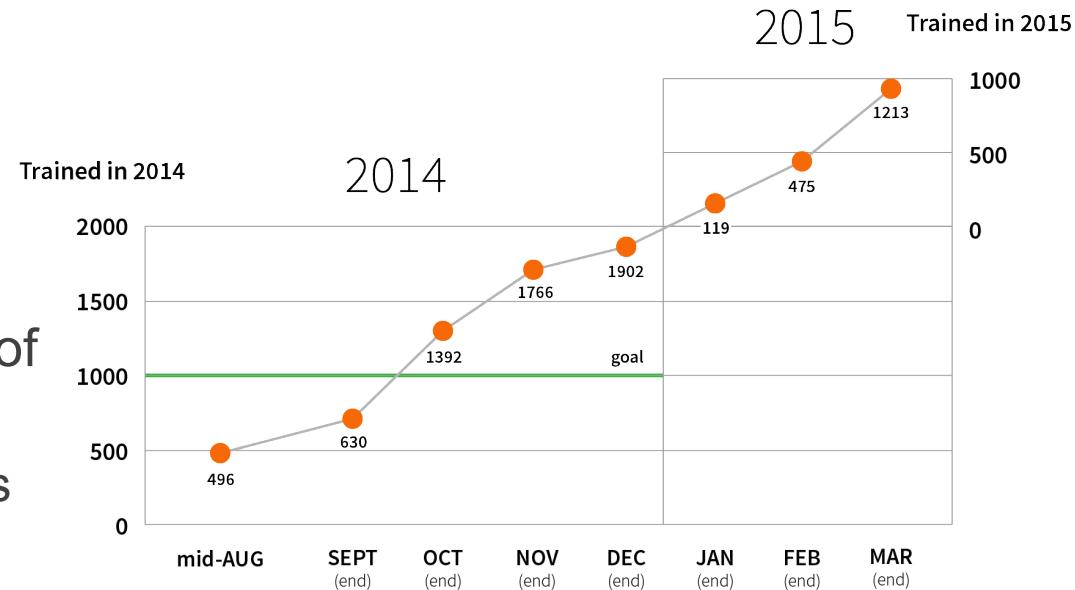
# Training

Spark training since 2011

~2000 people trained in 2014

1200+ people trained by end of March, 2015

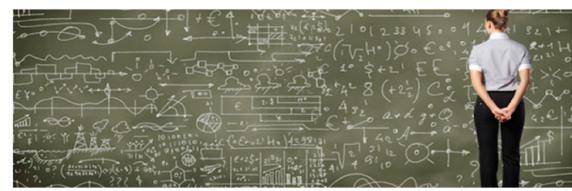
– 500+ people trained at this Spark Summit alone!



# MOOCs

## *“Intro to Big Data with Apache Spark”*

- Anthony Joseph, UC Berkeley
- **30,000+** already registered



### **Introduction to Big Data with Apache Spark**

Learn how to apply data science techniques using parallel programming in Apache Spark to explore big (and small) data.

## *“Scalable Machine Learning”*

- Ameet Talwalkar, UCLA
- **16,000+** already registered



### **Scalable Machine Learning**

Learn the underlying principles required to develop scalable machine learning pipelines and gain hands-on experience using Apache Spark.

# Making Big Data Simple

# Databricks Cloud

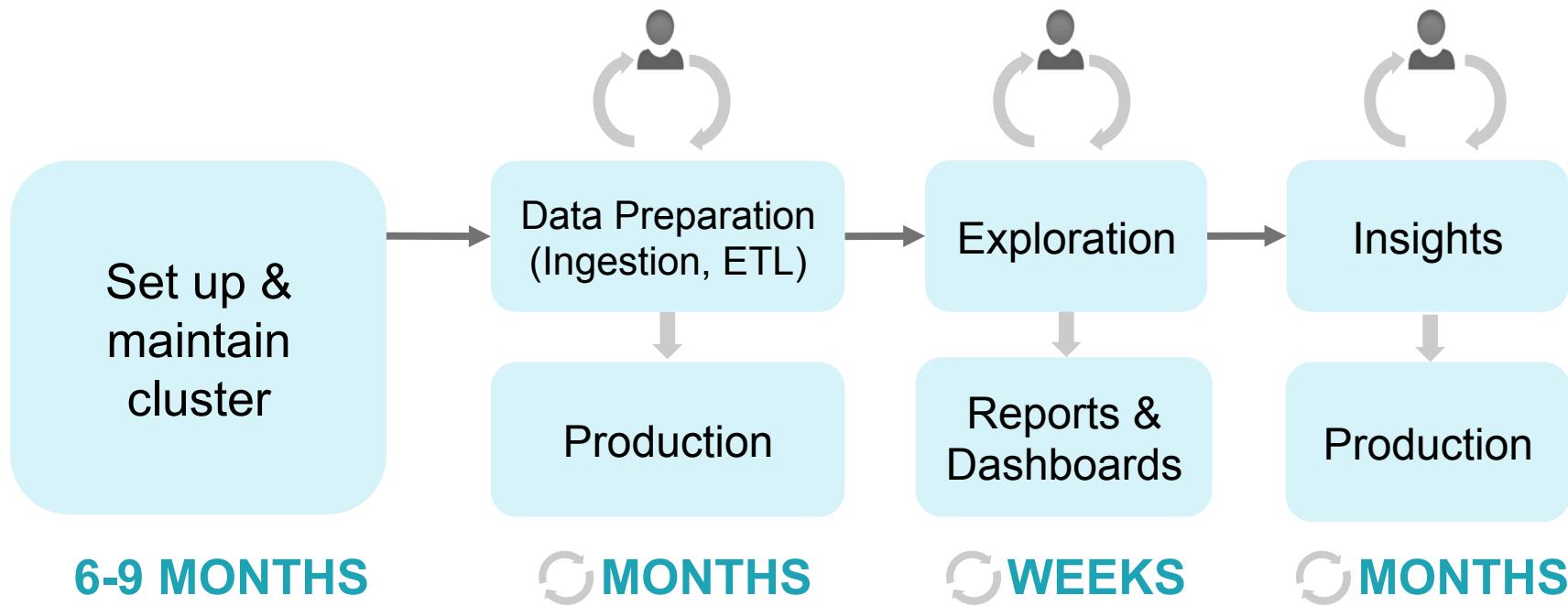
**July, 2014:** Unveiled Databricks Cloud

Over **3,500+** have registered to use Databricks Cloud

**November, 2014:** Limited availability

**100+** companies have been using Databricks Cloud

# Big Data Projects are Hard



# Why Databricks Cloud?

Accelerate time-to-results from **months** to **days**

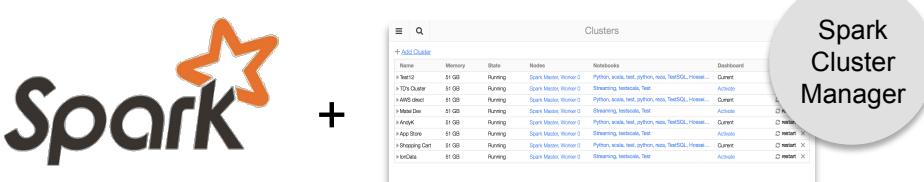
- Zero management
- Real-time
- Unified platform

Open platform

# Databricks Cloud



# Workspace



## Spark + Cluster Manager



# Cloud Infrastructure

# Zero Management

# Zero Management

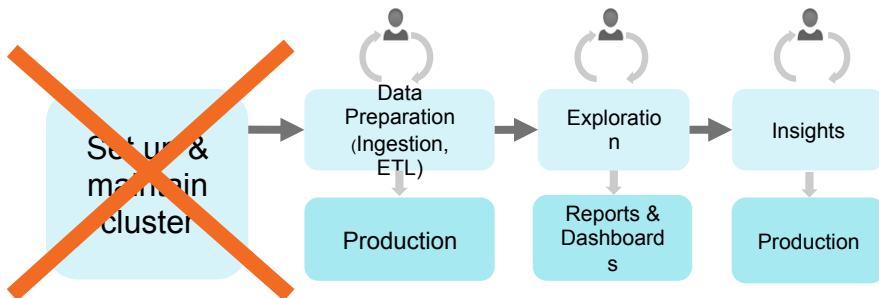
## Spark Cluster Manager



No need to set up clusters

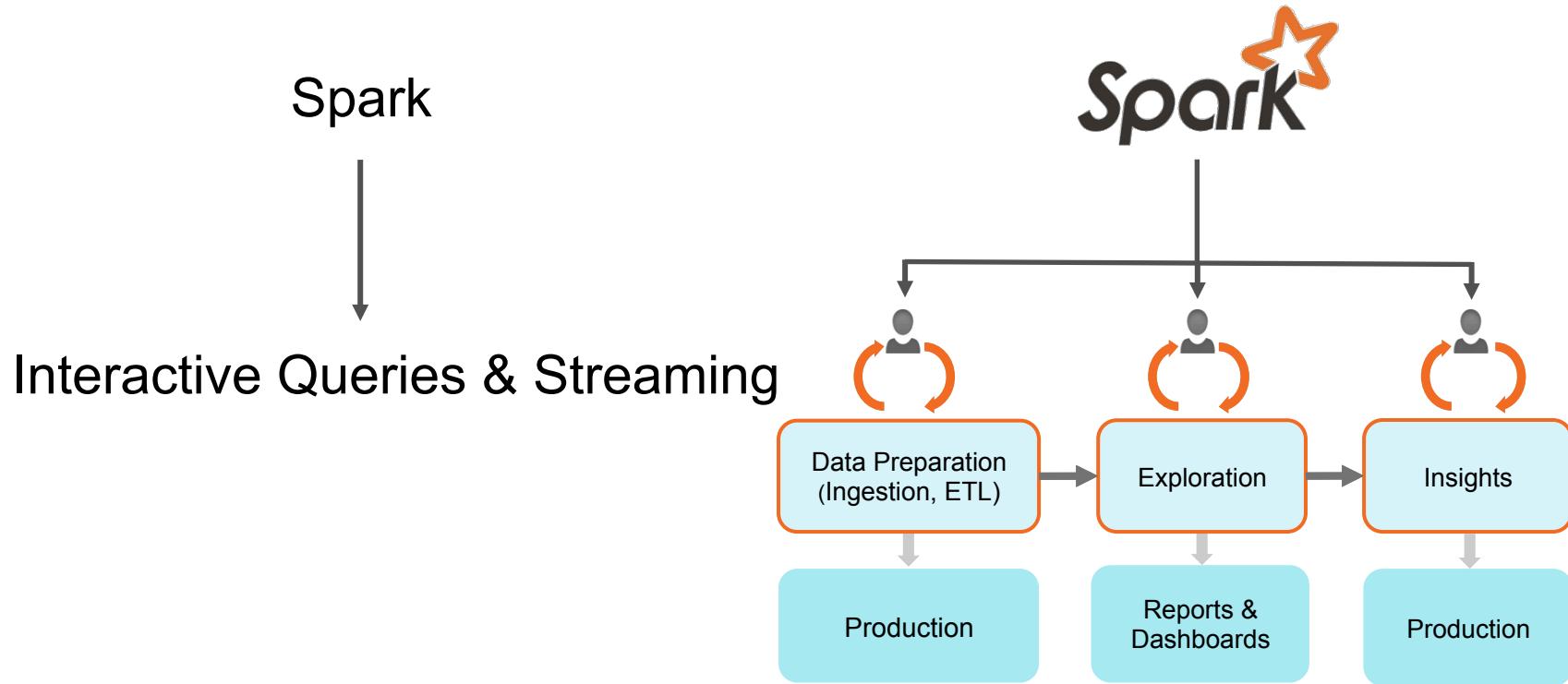
Clusters						
+ Add Cluster						
Name	Memory	State	Nodes	Notebooks	Dashboard	
» Test12	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossel...	Current	<input type="button"/> restart <input type="button"/> X
» TDS Cluster	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, Test	Activate	<input type="button"/> restart <input type="button"/> X
» AWS direct	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossel...	Current	<input type="button"/> restart <input type="button"/> X
» Mata Dev	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, test	Activate	<input type="button"/> restart <input type="button"/> X
» AndyK	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossel...	Current	<input type="button"/> restart <input type="button"/> X
» App Store	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, Test	Activate	<input type="button"/> restart <input type="button"/> X
» Shopping Cart	51 GB	Running	Spark Master, Worker 0	Python, scala, test, python, reza, TestSQL, Hossel...	Current	<input type="button"/> restart <input type="button"/> X
» IonData	51 GB	Running	Spark Master, Worker 0	Streaming, testscala, Test	Activate	<input type="button"/> restart <input type="button"/> X

Spark  
Cluster  
Manager



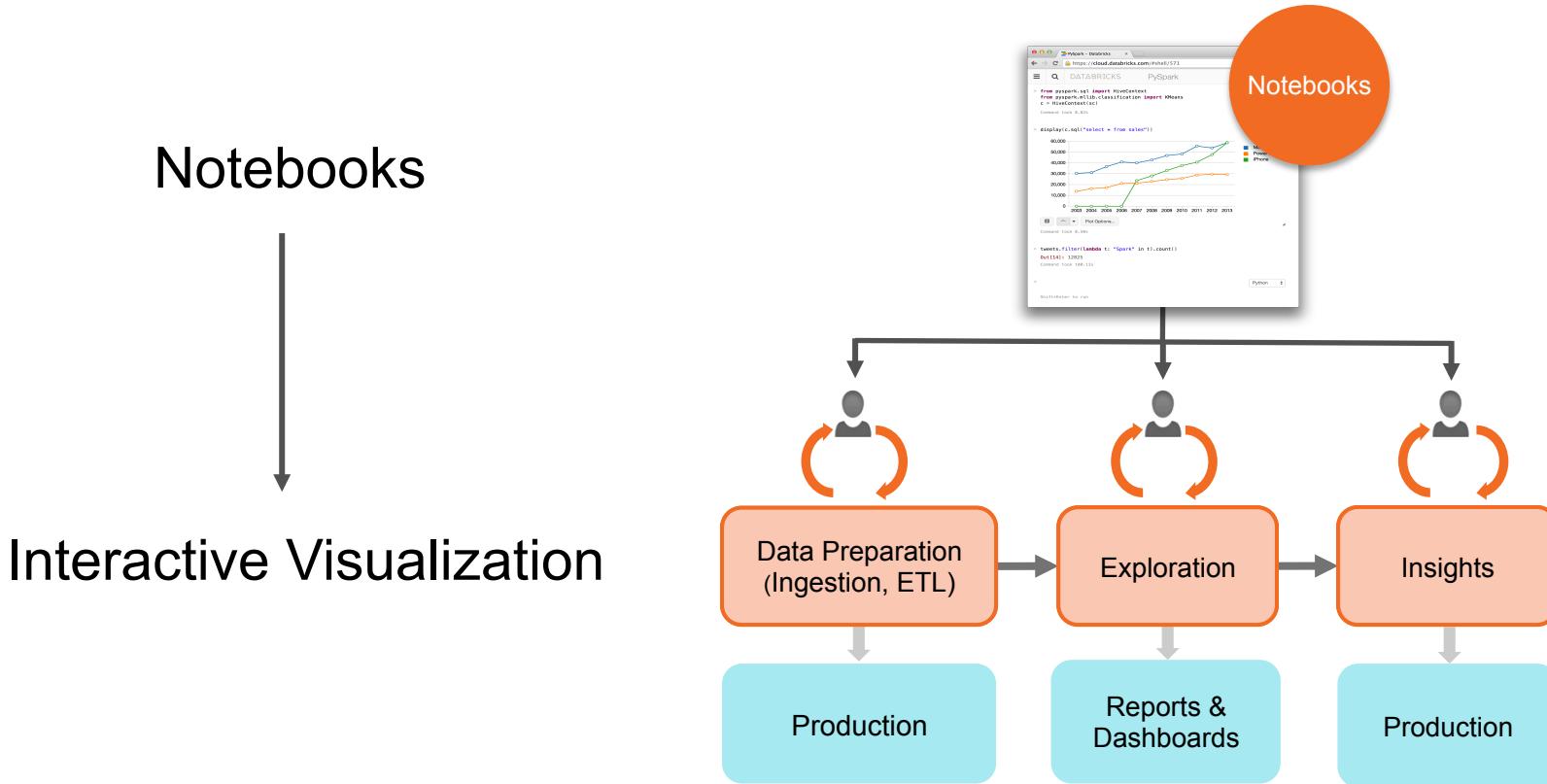
# Real Time

# Real-Time



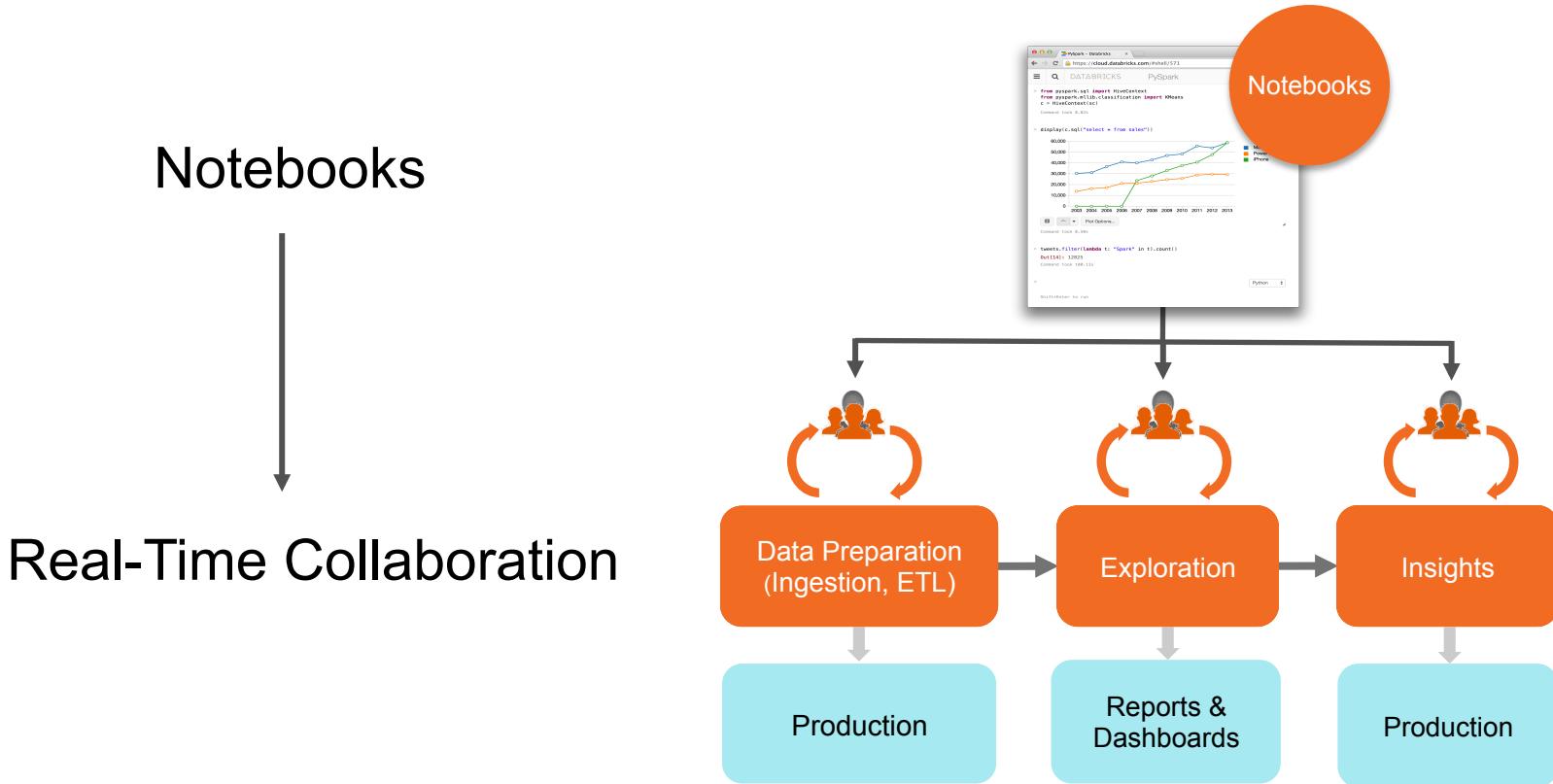
# Real-Time

## Notebooks



# Real-Time

## Notebooks

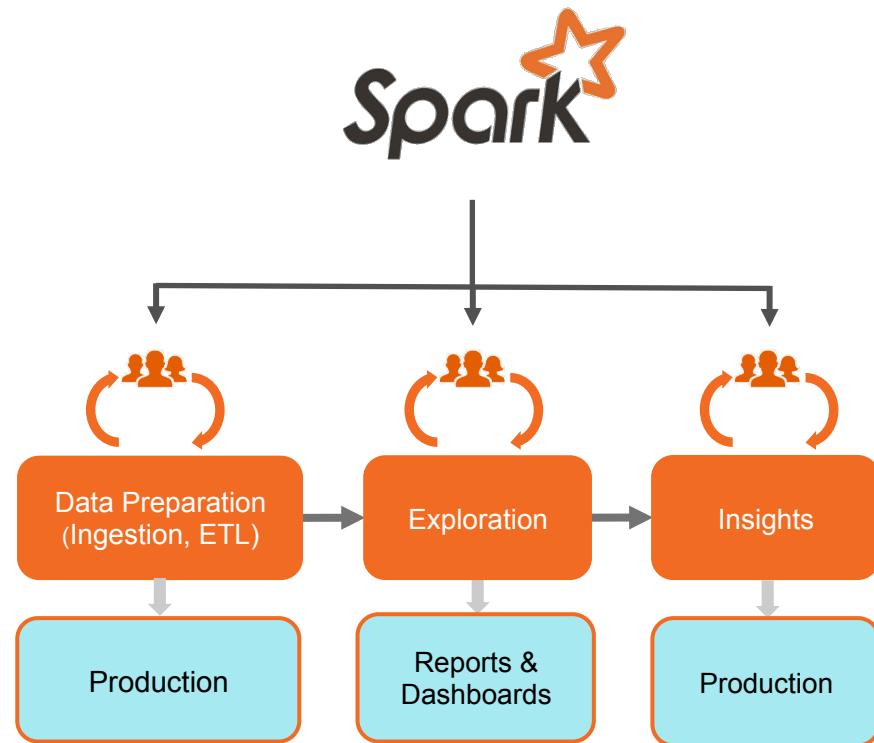


# Unified Platform

# Unified Platform

Spark

One API, One Engine  
Supporting All  
Workloads

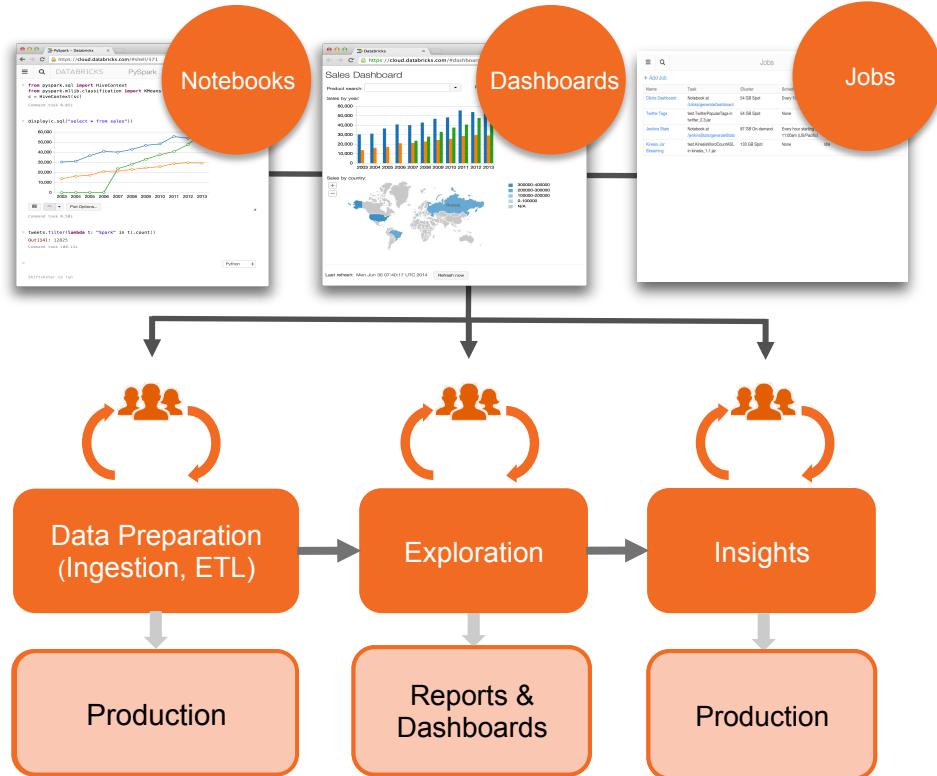


# Unified Platform

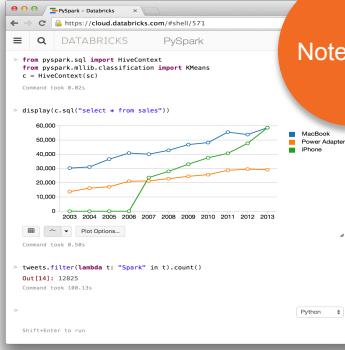
Notebooks,  
Dashboards,  
Jobs



One Set of Tools

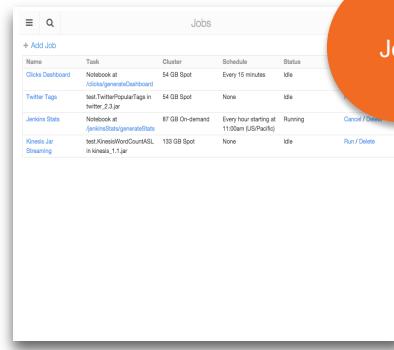


# Unified Platform



A screenshot of a Databricks PySpark notebook. The interface includes a top navigation bar with 'PySpark' and 'Databricks'. Below the bar, there's a code editor with PySpark code and a line chart. The chart displays data for 'MacBook', 'Power Adapter', and 'iPhone' from 2003 to 2013. The chart shows a significant increase in sales for all three categories over the period.

Notebooks



A screenshot of the Databricks Jobs dashboard. It shows a table with columns for Name, Task, Cluster, Schedule, and Status. There are five entries listed:

Name	Task	Cluster	Schedule	Status
Clicks Dashboard	Notebook at /clicks/generalDashboard	54 GB Spot	Every 15 minutes	Idle
Twitter Tags	test.TwitterPopularTags in twitter_2.3.jar	54 GB Spot	None	Idle
Jenkins Shuts	Notebook at /jenkins/generalJenkins	87 GB On-demand	Every hour starting at 11:00am (US/Pacific)	Running
Kinesis Jar Streaming	test.KinesisWordCountSQL in kinesis_1.1.jar	130 GB Spot	None	Idle

Jobs

Use notebooks to interactively develop

- ETL
- Data analysis
- ML Models
- ...

Run notebooks as jobs!

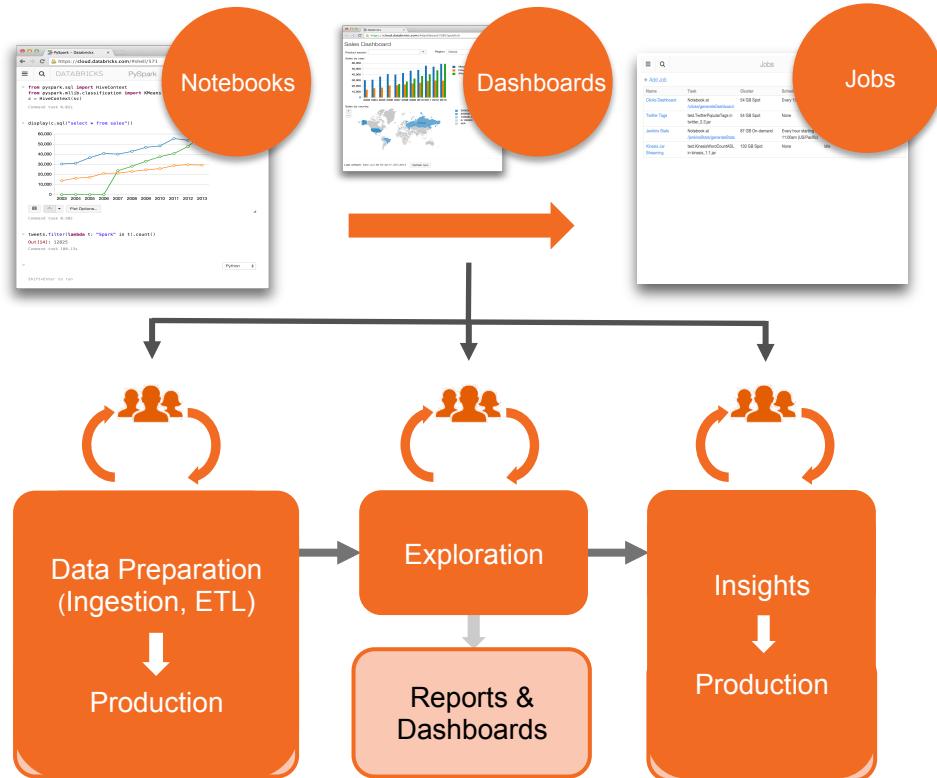
- Can take input arguments
- No need to re-engineer

# Unified Platform

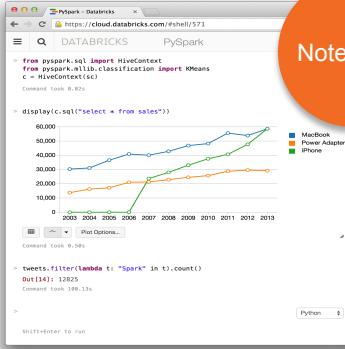
Run Notebooks as Jobs



No Code to Rewrite



# Unified Platform



A screenshot of a PySpark notebook in a Databricks interface. The notebook contains Python code for importing PySpark, creating a Hive context, performing K-Means clustering, and displaying a line plot of sales data from 2003 to 2013 for MacBook, Power Adapter, and iPhone. The plot shows a general upward trend for all three categories.

Notebooks



Dashboards

Use notebooks to compute and plot

- KPIs
- Funnels
- ...

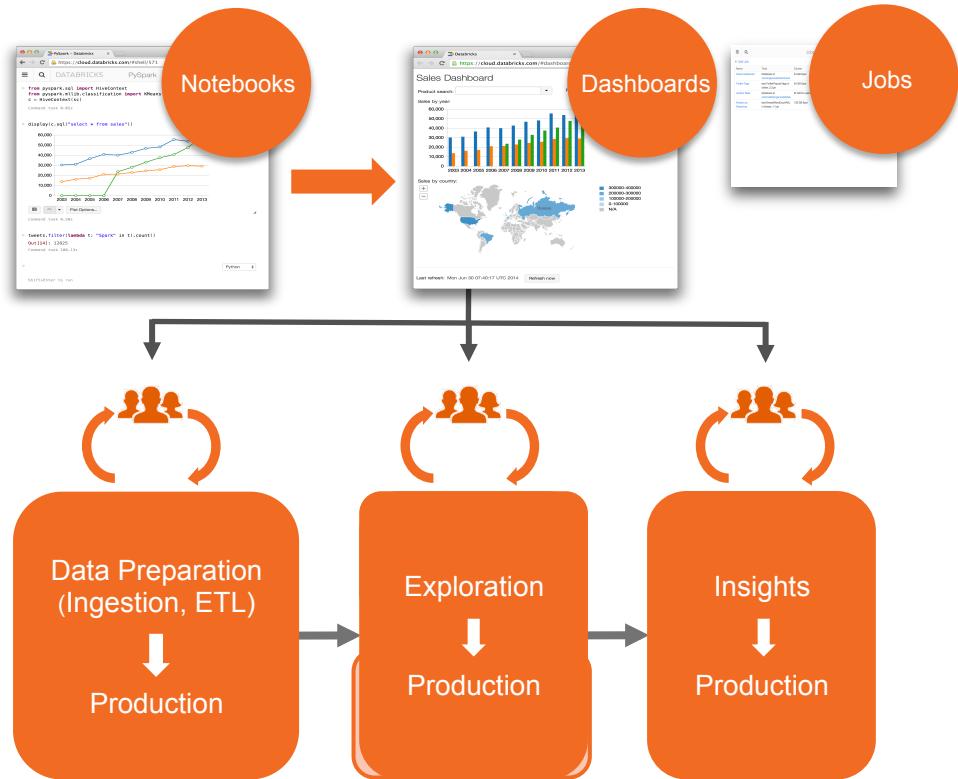
Drag and drop notebook plots to instantly create dashboards.

# Unified Platform

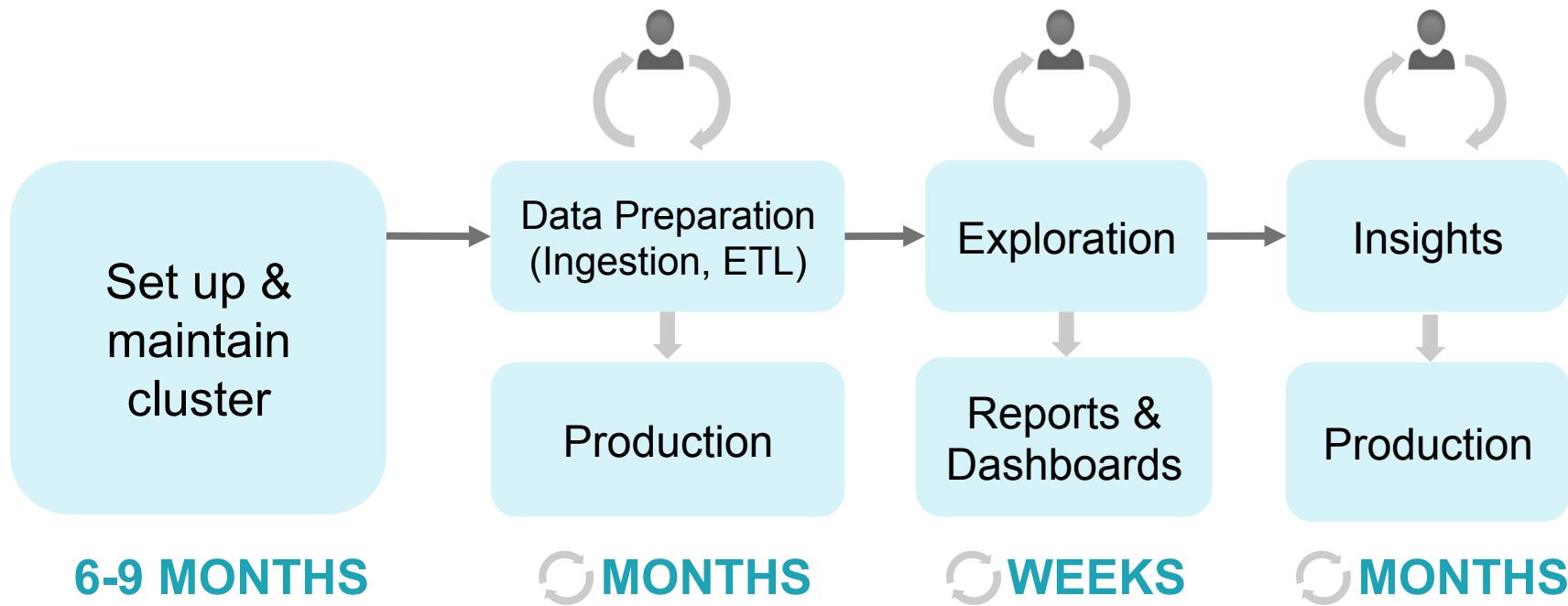
Notebooks as Dashboards



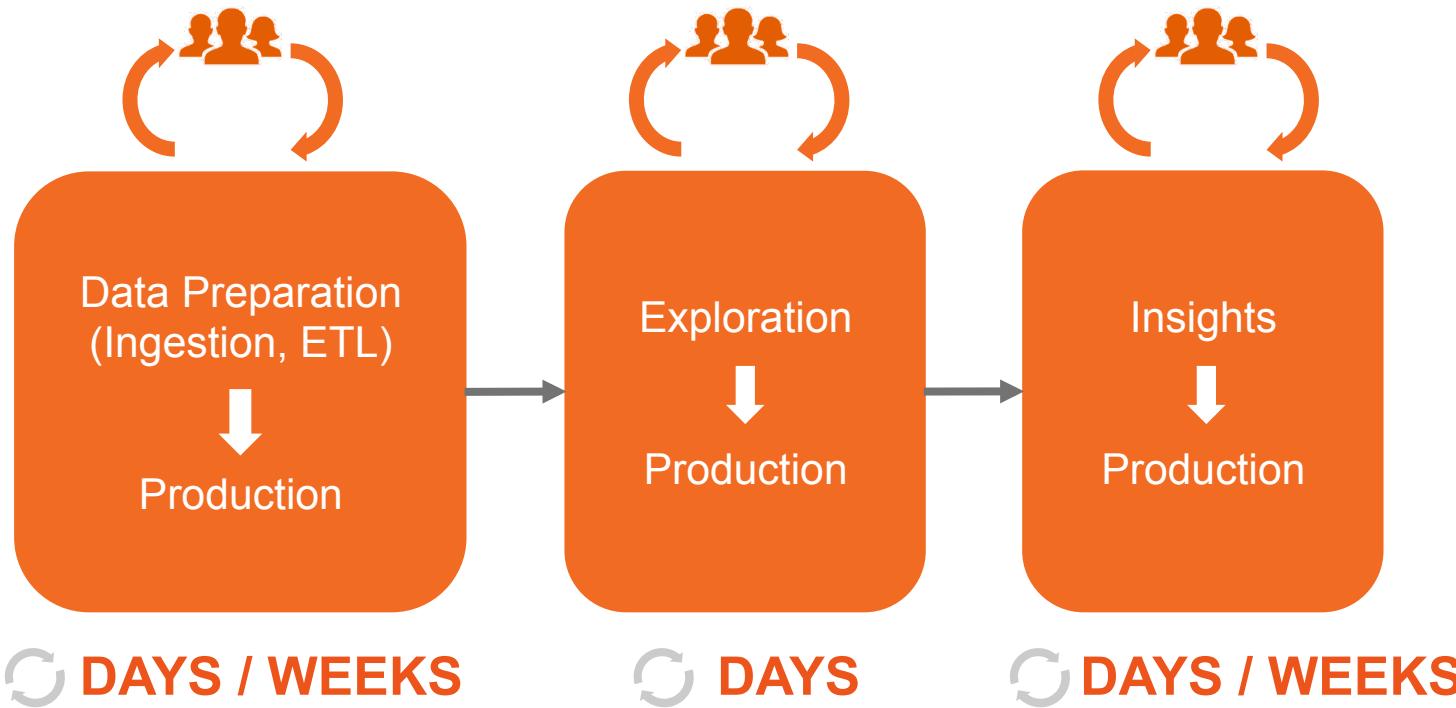
Easily Go From  
Exploration  
to Production



# From Months

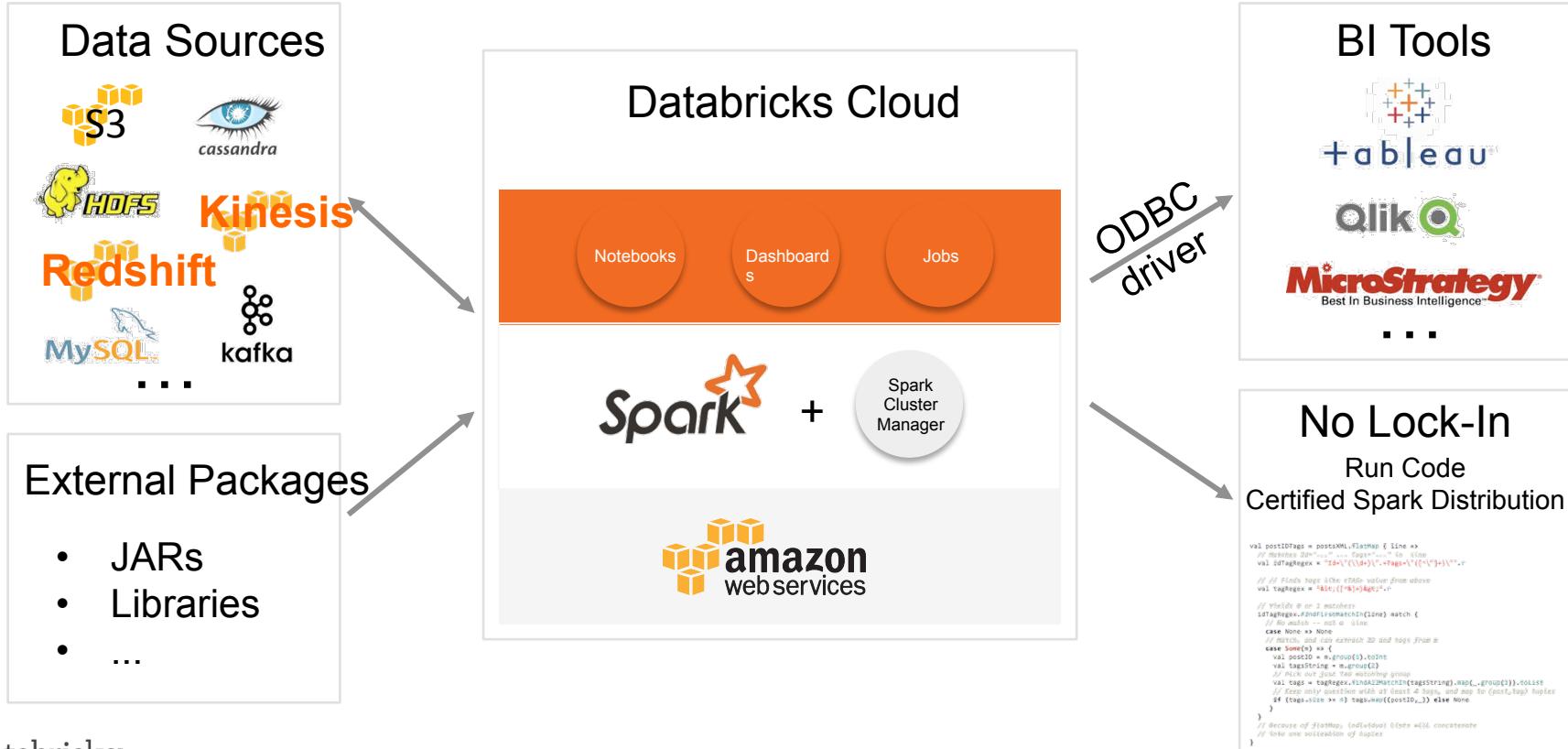


# From Months to Days



# Open Platform

# Open Platform







# Spark for Health & Fitness



Chul Lee  
Head of Data Engineering & Science

# What is MyFitnessPal?



## Simple & Effective Health/Fitness Tracking Tool

#1 health & fitness app for iOS & Android  
over 1 million 5 star ratings in the App Store



## Big Engaged Community

80+ million registered users



## Massive DB of foods

Over 5 million food items  
Over 14.5 billion logged foods  
Over 36 million recipes

*(plus Massive DB of exercise data)*



## Success Factors of Data Product Innovation



Big Data (Foods, Recipes, Diets, etc)



Solid & Highly Scalable Data Infrastructure



Product Fit



Large-Scale Algorithms (ML, NLP, etc)



*MyFitnessPal's* food DB (other related data) is the richest and largest in industry

**DataBricks** provides a flexible and scalable data infrastructure for the rapid and solid development of data products

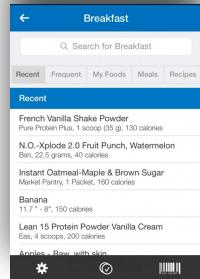
**DataBricks** helps to reduce “time to value” allowing to focus on data product innovation and customer understanding

**Spark** provides an easy access to large scale ML and data mining algorithms (i.e. MLlib)

# Past



## Food Data Cleaning



## Search



## Suggested Serving Sizes

And more....

# Future



## Ad-targetting/RecSys



## Deep-Dive into Customer Understanding



## Large-Scale ETL

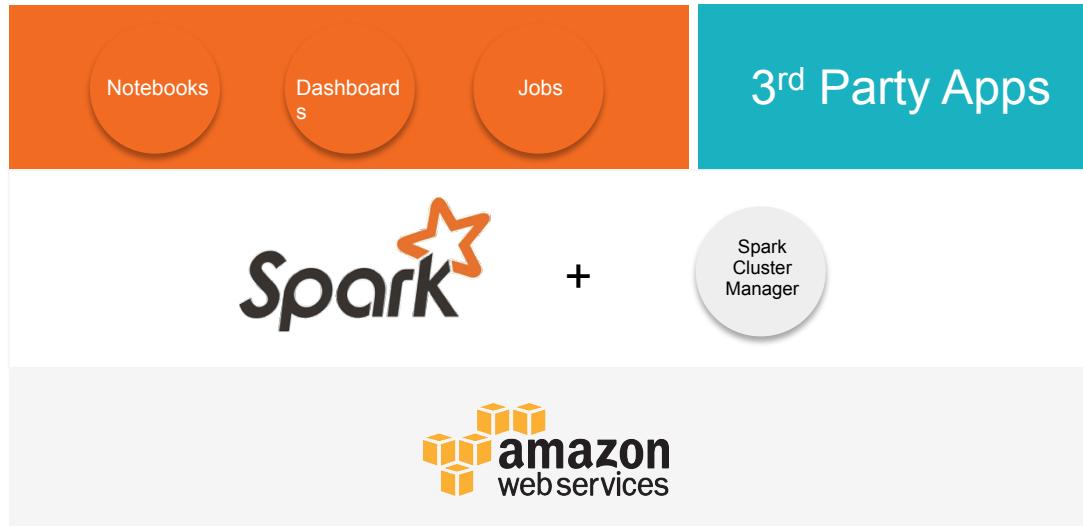
And more...



AUTOMATIC™

# Open Platform: 3<sup>rd</sup> Party Apps

## Databricks Cloud





**uncharted**<sup>TM</sup>  
formerly **Oculus Info Inc**



treasta

# Databricks Cloud

Dramatically accelerate time-to-results for big data

Open platform, no lock-in

Everyone here will receive access to  
Databricks Cloud within next week!