Harnessing the Power of Spark with Databricks Cloud

Ion Stoica

March 18, 2015

Accelerating Spark Adoption

Certification

Applications (35+)

Distributions (11+)

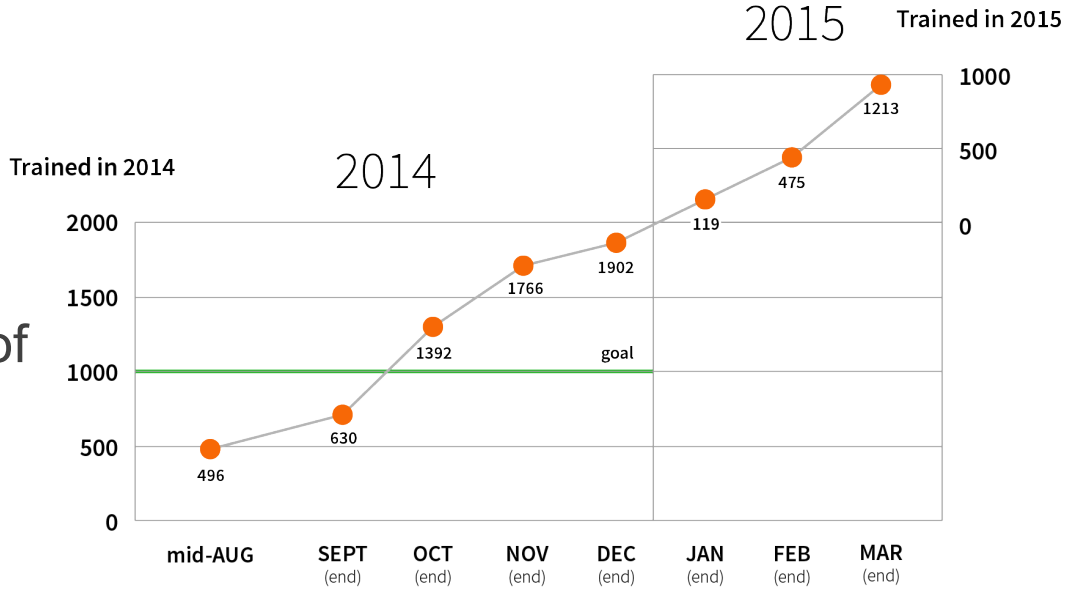
Training

Spark training since 2011

~2000 people trained in 2014

1200+ people trained by end of March, 2015

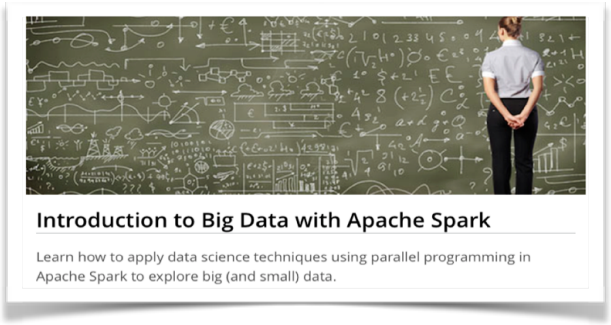
- 500+ people trained at this Spark Summit alone!



MOOCs

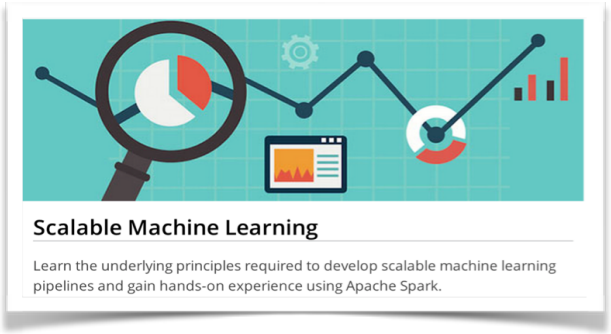
"Intro to Big Data with Apache Spark"

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



"Scalable Machine Learning"

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



Making Big Data Simple

Databricks Cloud

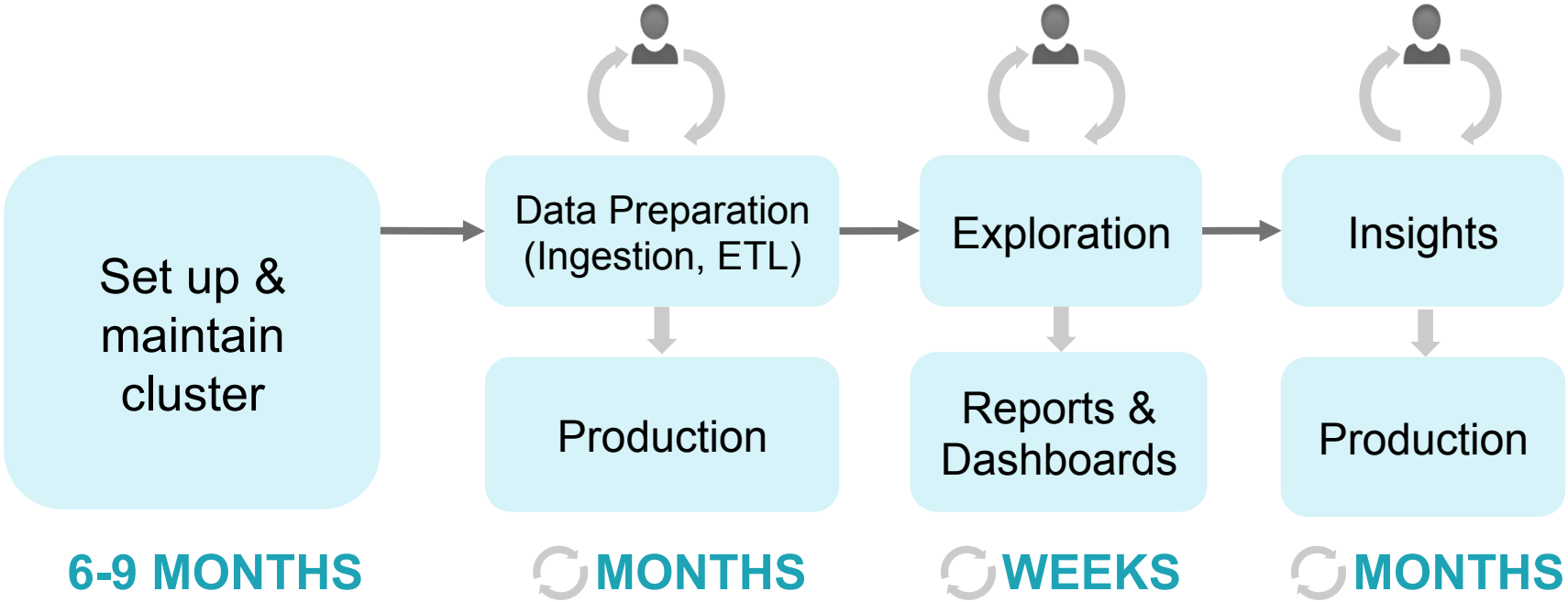
July, 2014: Unveiled Databricks Cloud

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?

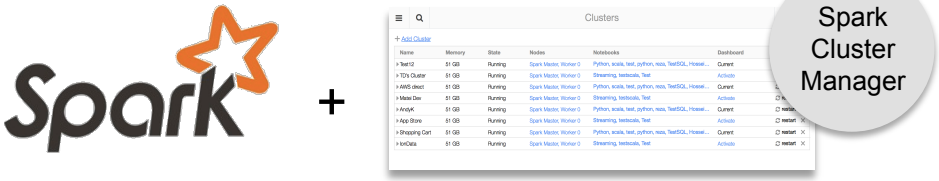
Accelerate time-to-results from months to days

- Zero management
- Real-time
- Unified platform

Open platform

Databricks Cloud

Workspace



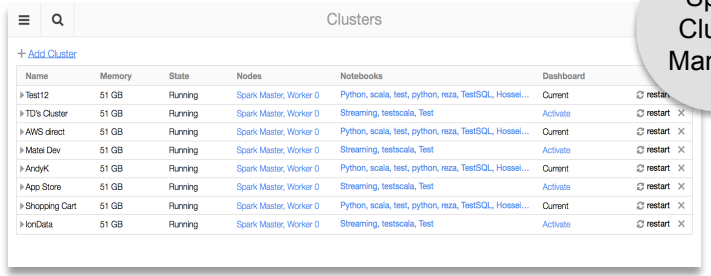
Spark + Cluster Manager

Cloud Infrastructure

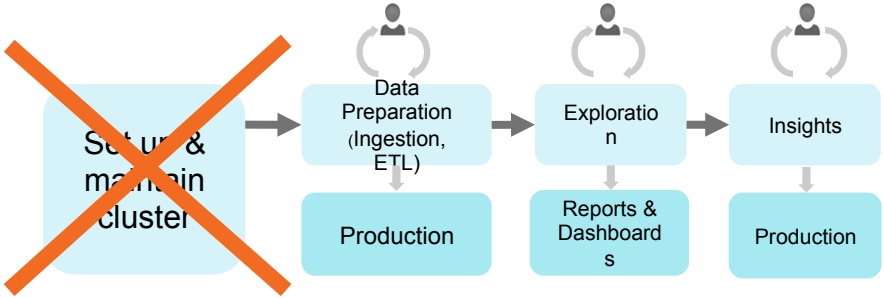
Zero Management

Zero Management

Spark Cluster Manager



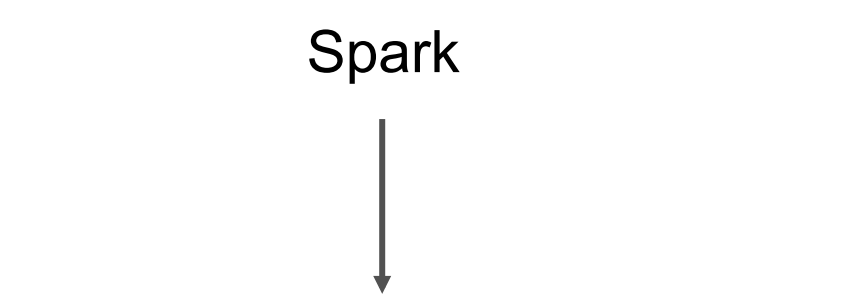
No need to set up clusters



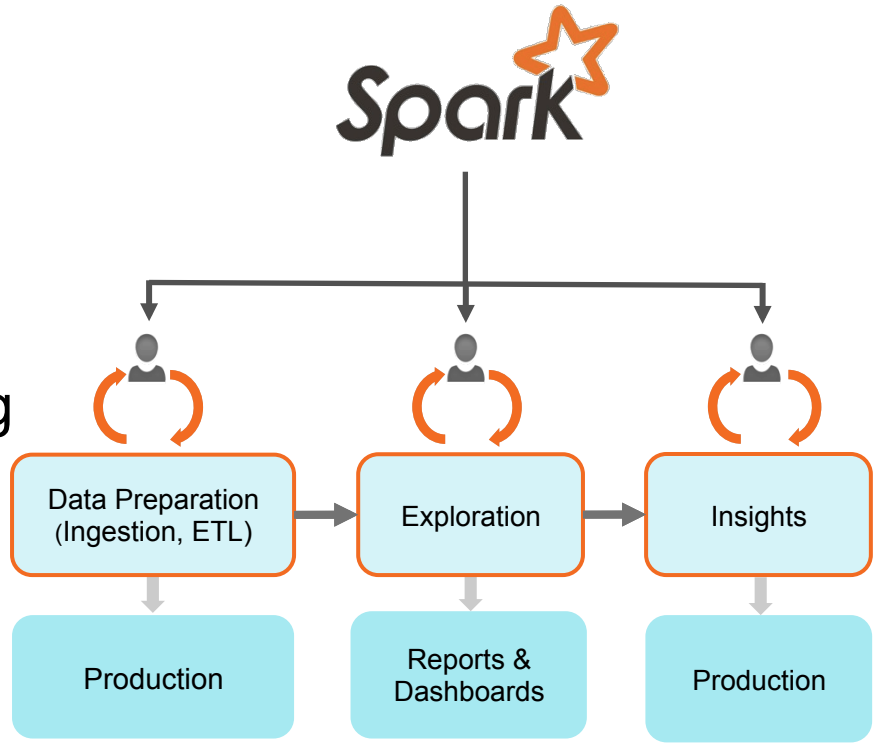
Real Time

Real-Time

Spark

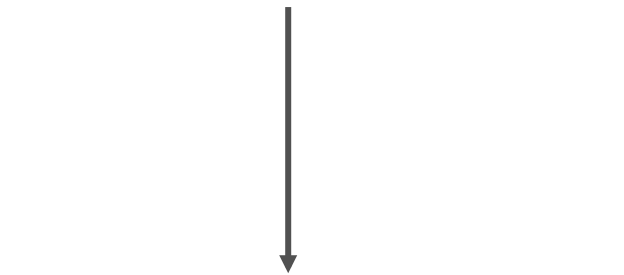


Interactive Queries & Streaming

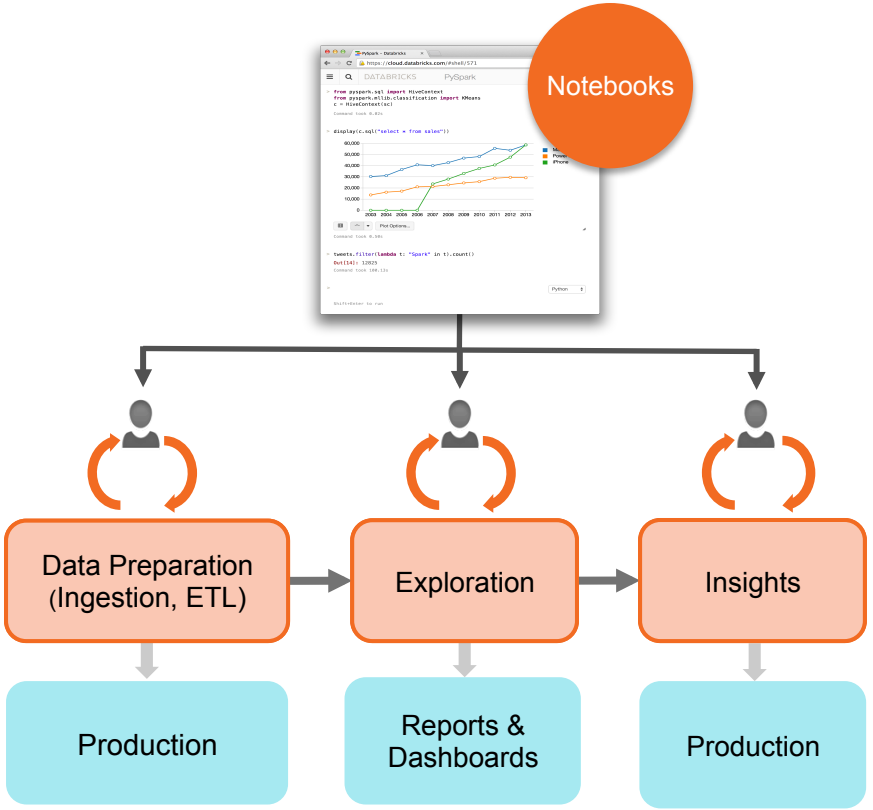


Real-Time

Notebooks



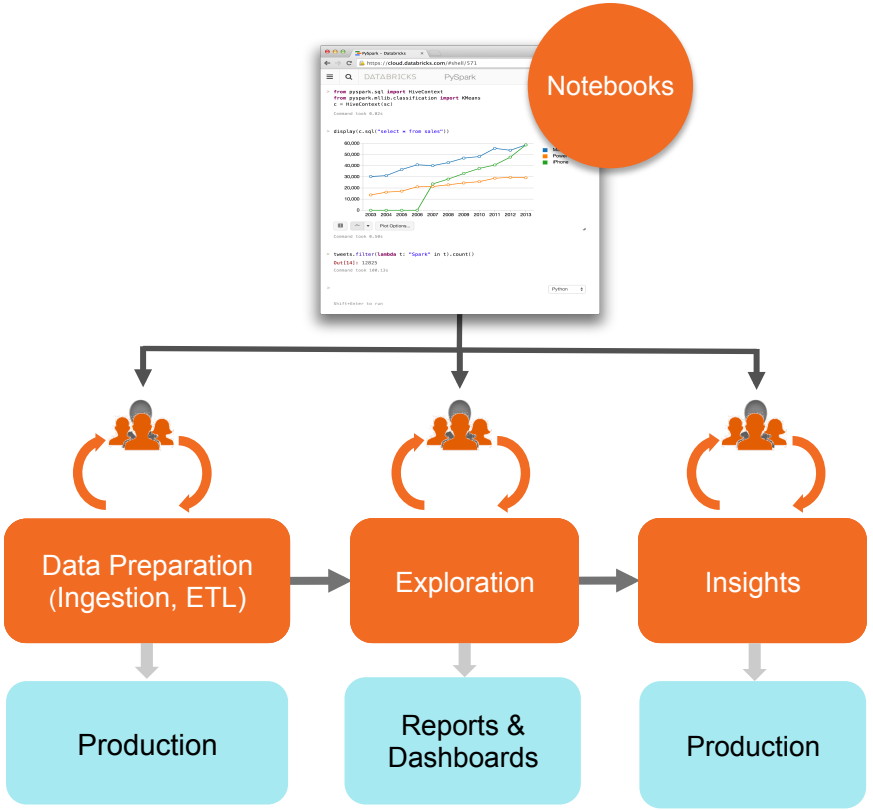
Interactive Visualization



Real-Time

Notebooks

Real-Time Collaboration

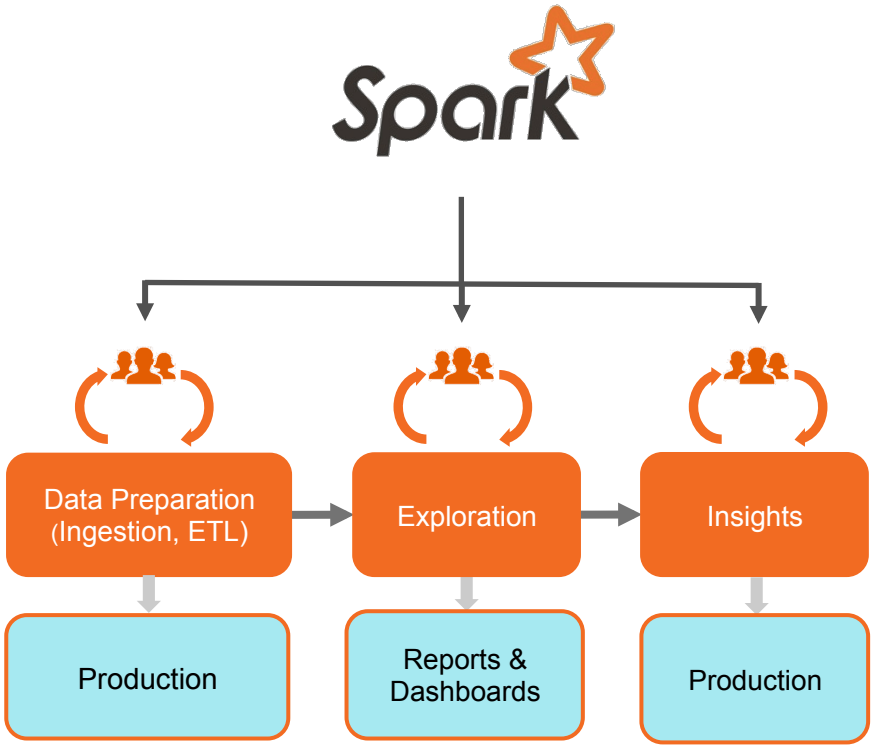


Unified Platform

Unified Platform

Spark

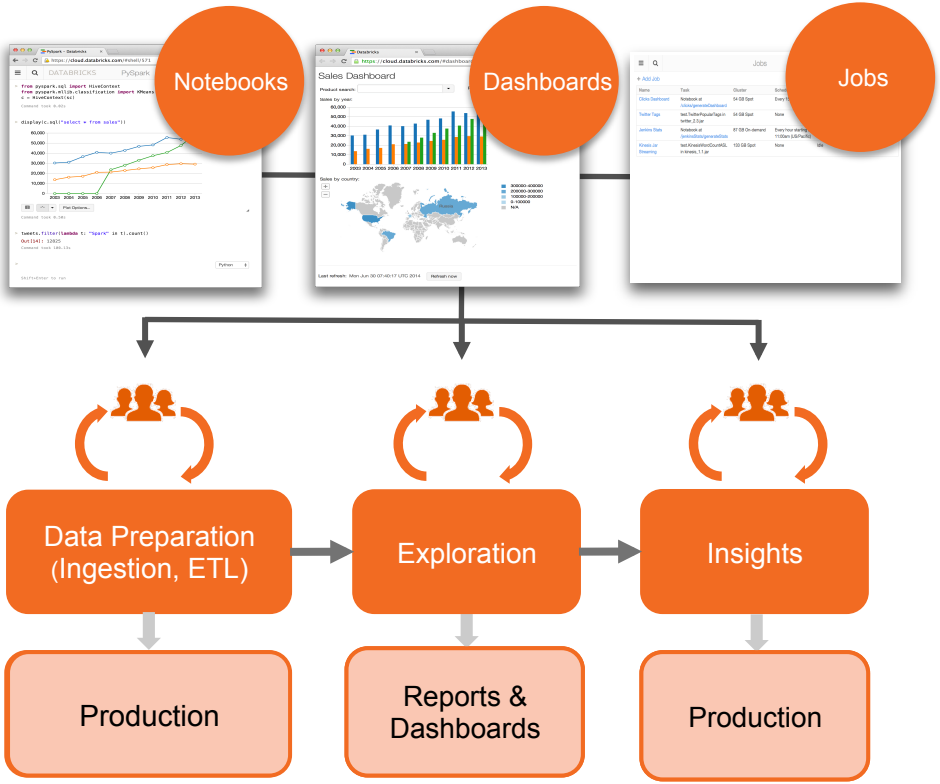
One API, One Engine
Supporting All
Workloads



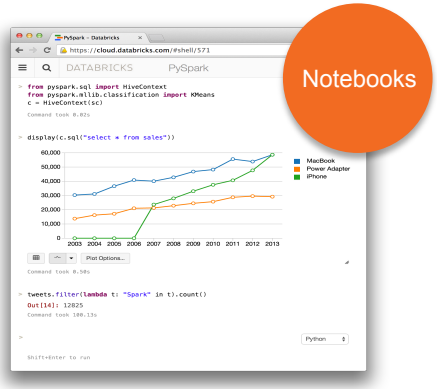
Unified Platform

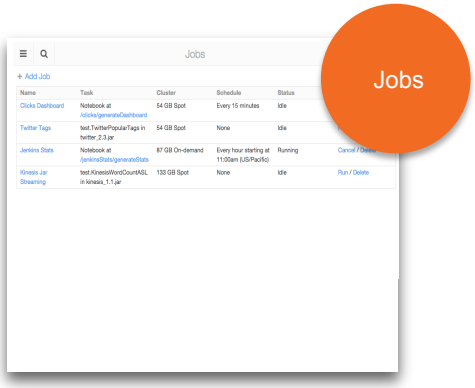
Notebooks,
Dashboards,
Jobs

One Set of Tools



Unified Platform





Use notebooks to interactively develop

- ETL
- Data analysis
- ML Models
- …

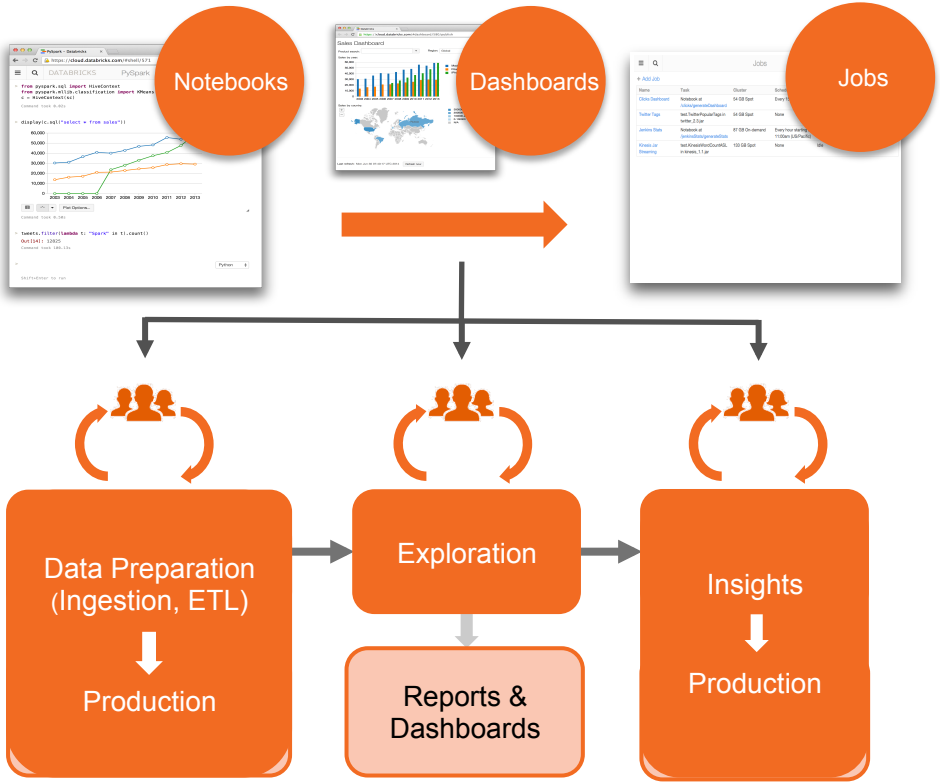
Run notebooks as jobs!

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

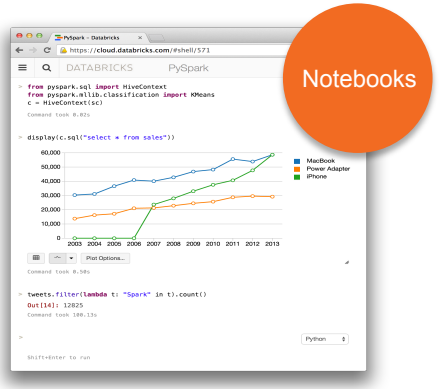
Unified Platform

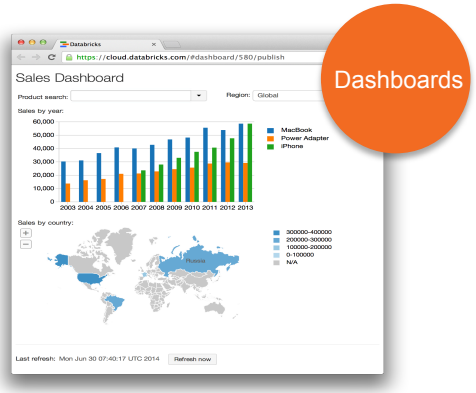
Run Notebooks as Jobs

No Code to Rewrite



Unified Platform





Use notebooks to compute and plot

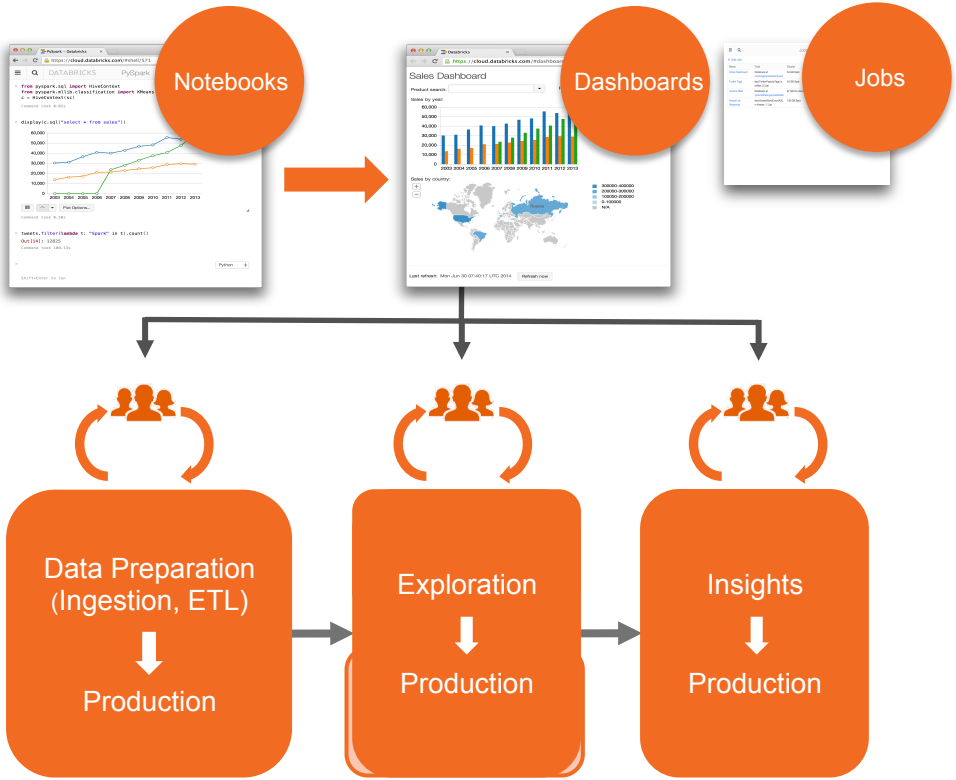
- KPIs
- Funnels
- …

Drag and drop notebook plots to instantly create dashboards.

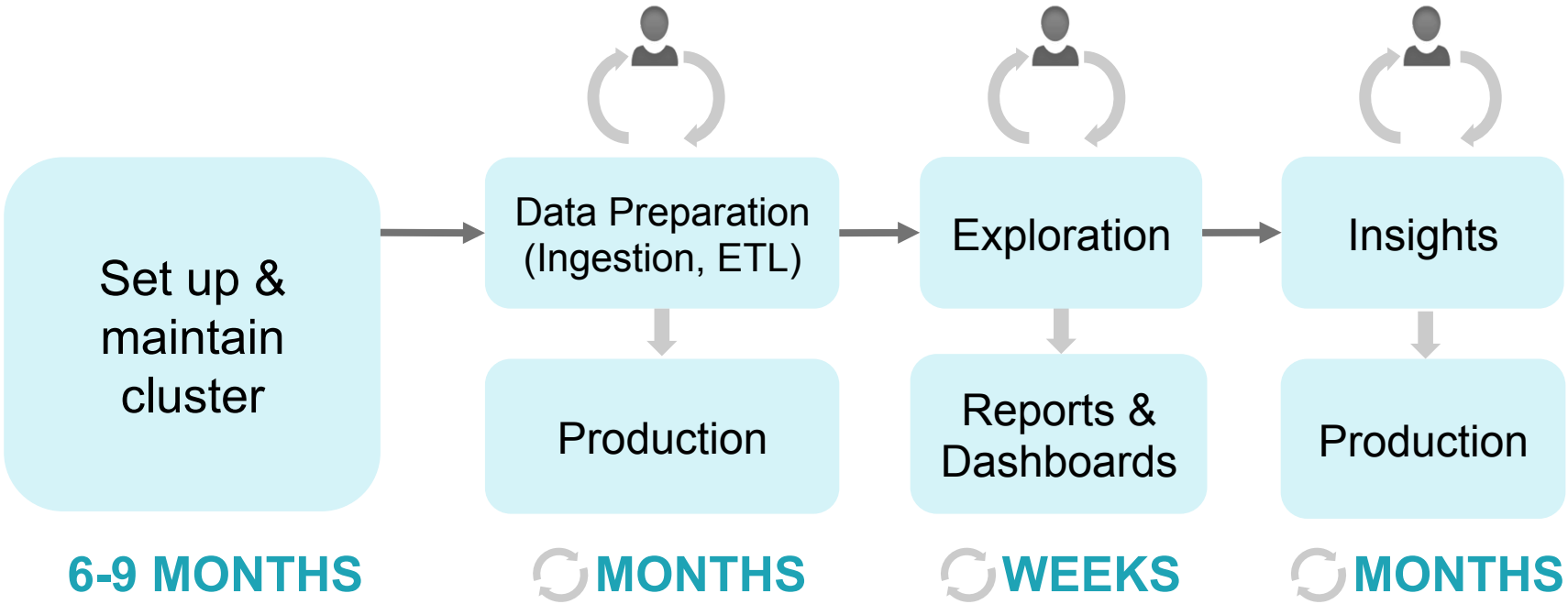
Unified Platform

Notebooks as Dashboards

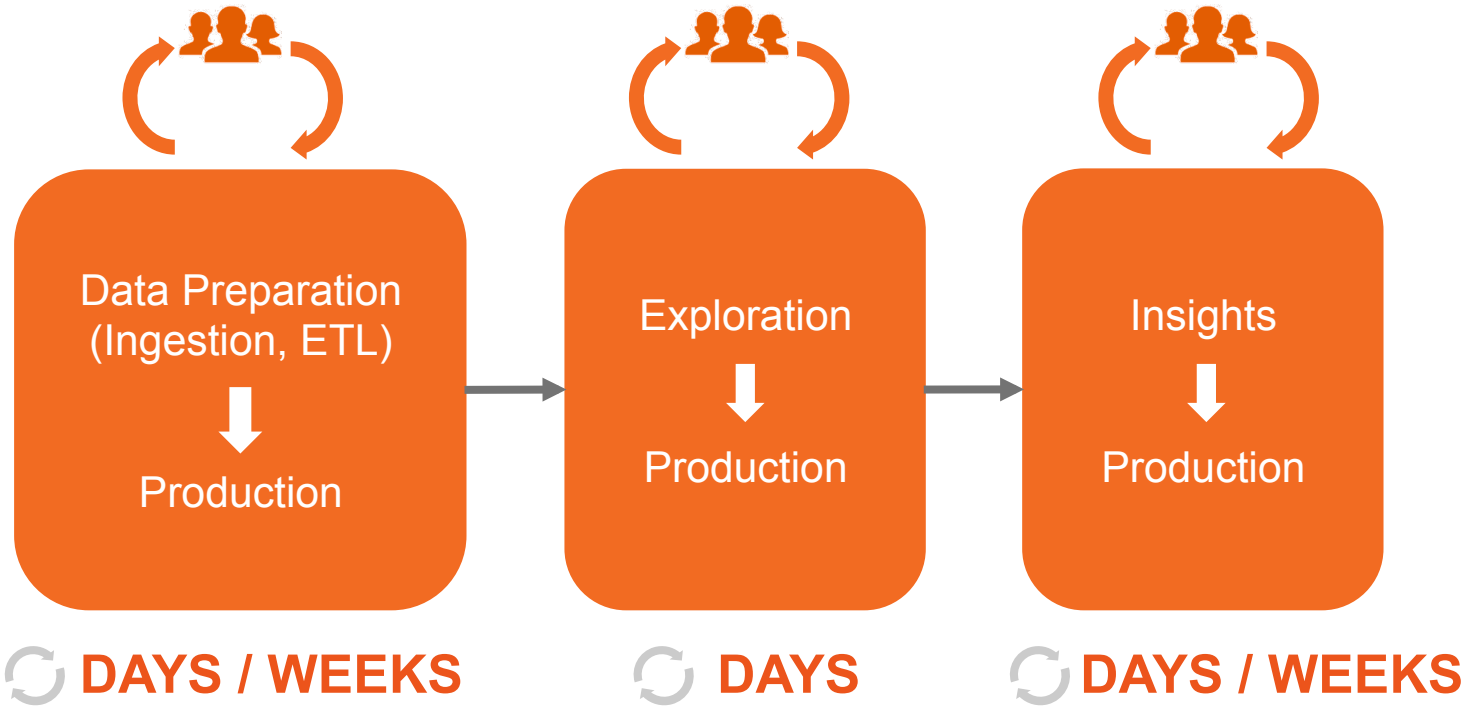
Easily Go From
Exploration
to Production



From Months


Set up & maintain cluster
Data Preparation (Ingestion, ETL)
Production
Exploration
Reports & Dashboards
Insights
Production
6-9 MONTHS
MONTHS
WEEKS
MONTHS

From Months to Days



Open Platform

Open Platform

Data Sources

S3 · cassandra · HDFS · Kinesis · Redshift · MySQL · kafka · ...

External Packages

- JARs
- Libraries
- ...

Databricks Cloud

Notebooks · Dashboards · Jobs

Spark + Spark Cluster Manager

amazon web services

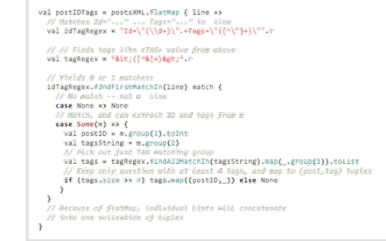
ODBC driver

BI Tools

+ableau · Qlik · MicroStrategy Best In Business Intelligence · ...

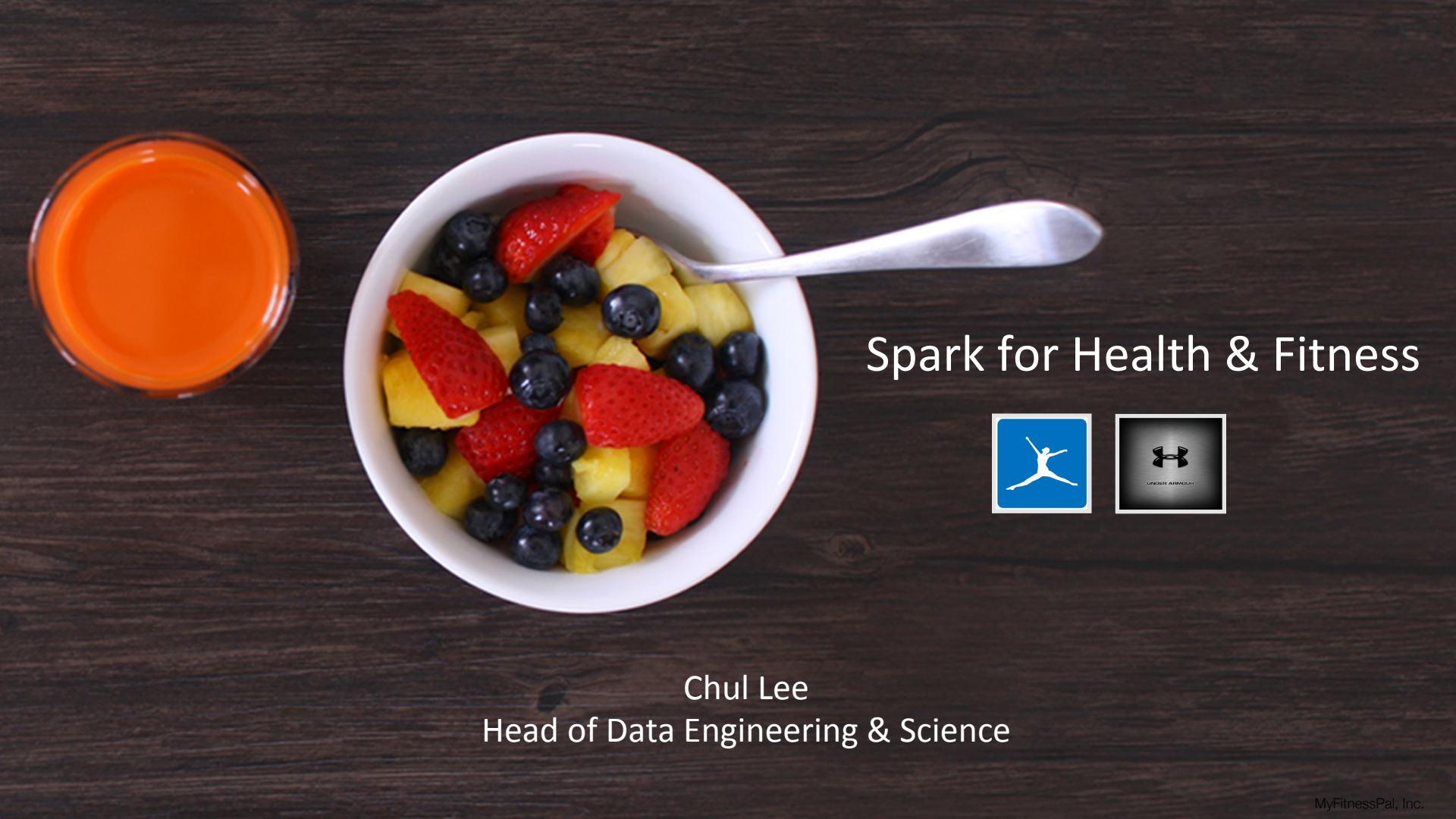
No Lock-In

Run Code
Certified Spark Distribution



databricks


myfitnesspal
UNDER ARMOUR



Spark for Health & Fitness

Chul Lee
Head of Data Engineering & Science
MyFitnessPal, Inc.

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)

myfitnesspal + UNDER ARMOUR

databricks

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)

databricks + Spark

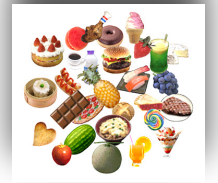
MyFitnessPal's food DB (other related data) is the richest and largest in industry

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

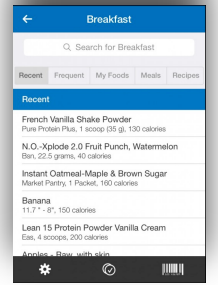
DataBricks helps to reduce "time to value" allowing to focus on data product innovation and customer understanding

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

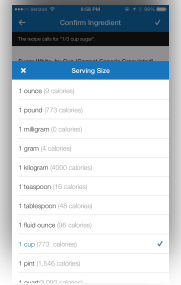
Past



Food Data Cleaning



Search



Suggested Serving Sizes

And more....

Future

Ad-targetting/RecSys

Deep-Dive into Customer Understanding

Large-Scale ETL

And more...

databricks

MyFitnessPal, Inc.


AUTOMATIC™

Open Platform: 3rd Party Apps


ZOOMDATA®


uncharted™
formerly Oculus Info Inc


tresata

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!