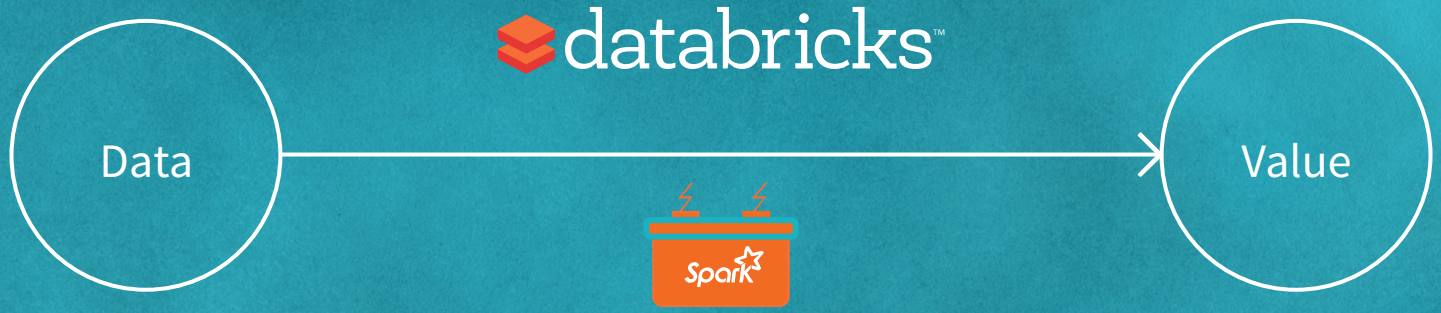
Making big data simple with Databricks

We are Databricks, the company behind Spark

Founded by the creators of Apache Spark in 2013

75%

Share of Spark code contributed by Databricks in 2014



Created **Databricks** on top of Spark to **make big data simple.**

WORKING WITH BIG DATA IS DIFFICULT

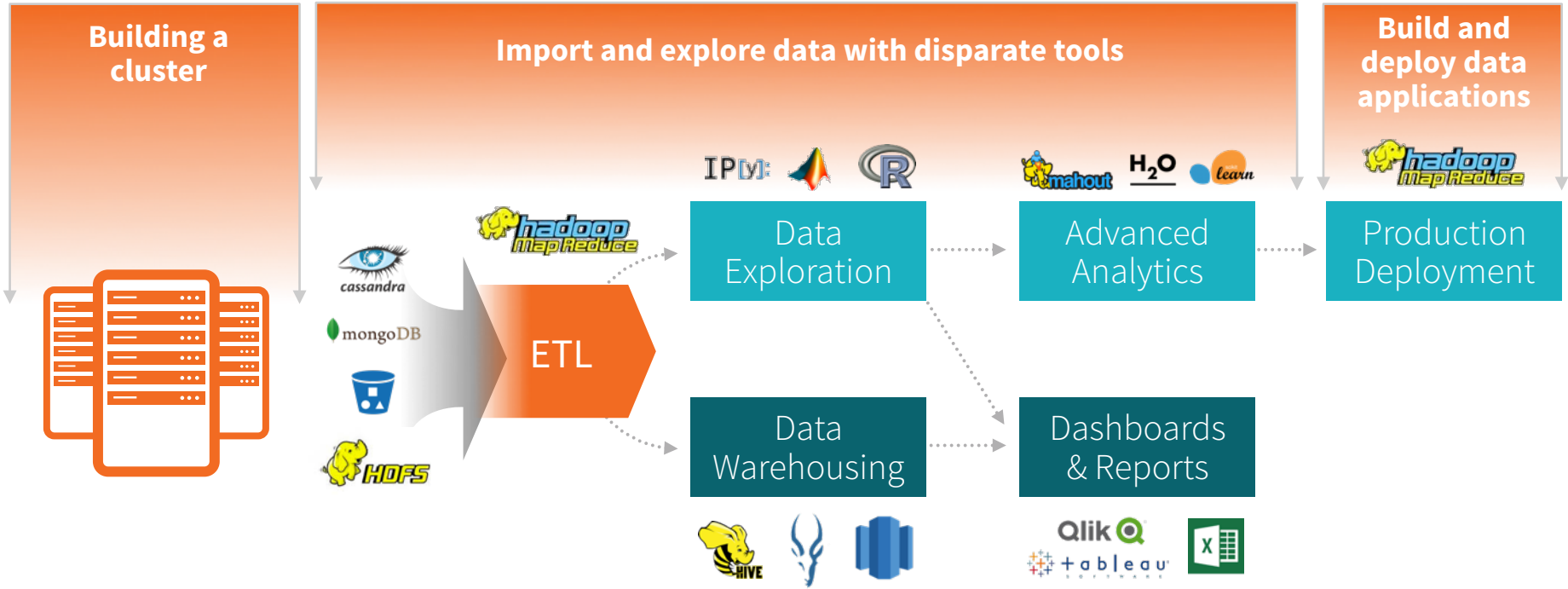
"Through 2017, 60% of big-data projects will fail to go beyond piloting and experimentation and will be abandoned."

GARTNER

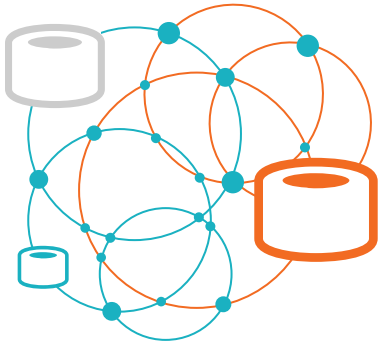
PROBLEM

Building infrastructure and data pipelines is complex

Your difficult journey to finding value in data



3 main causes of this problem:

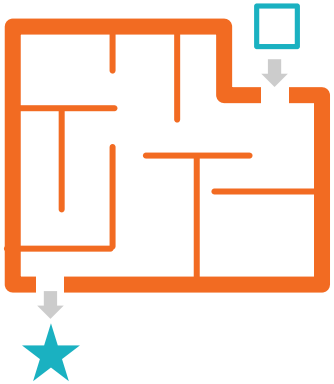


Infrastructure is complex to build and maintain

- Expensive upfront investment
- Months to build
- Dedicated DevOps to operate

Tools are slow, clunky, and disparate

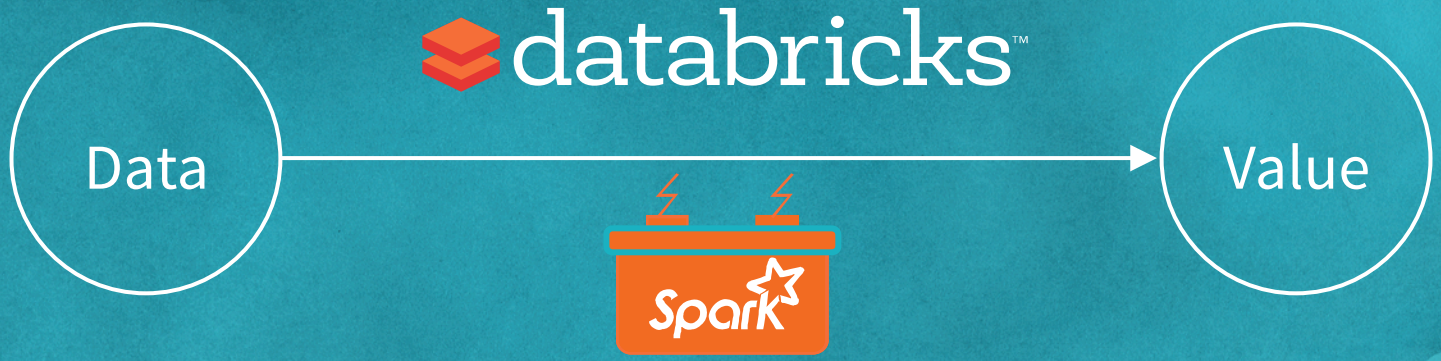
- Not user-friendly
- Long time to compute answers
- Lots of integration required



Re-engineering of prototypes for deployment

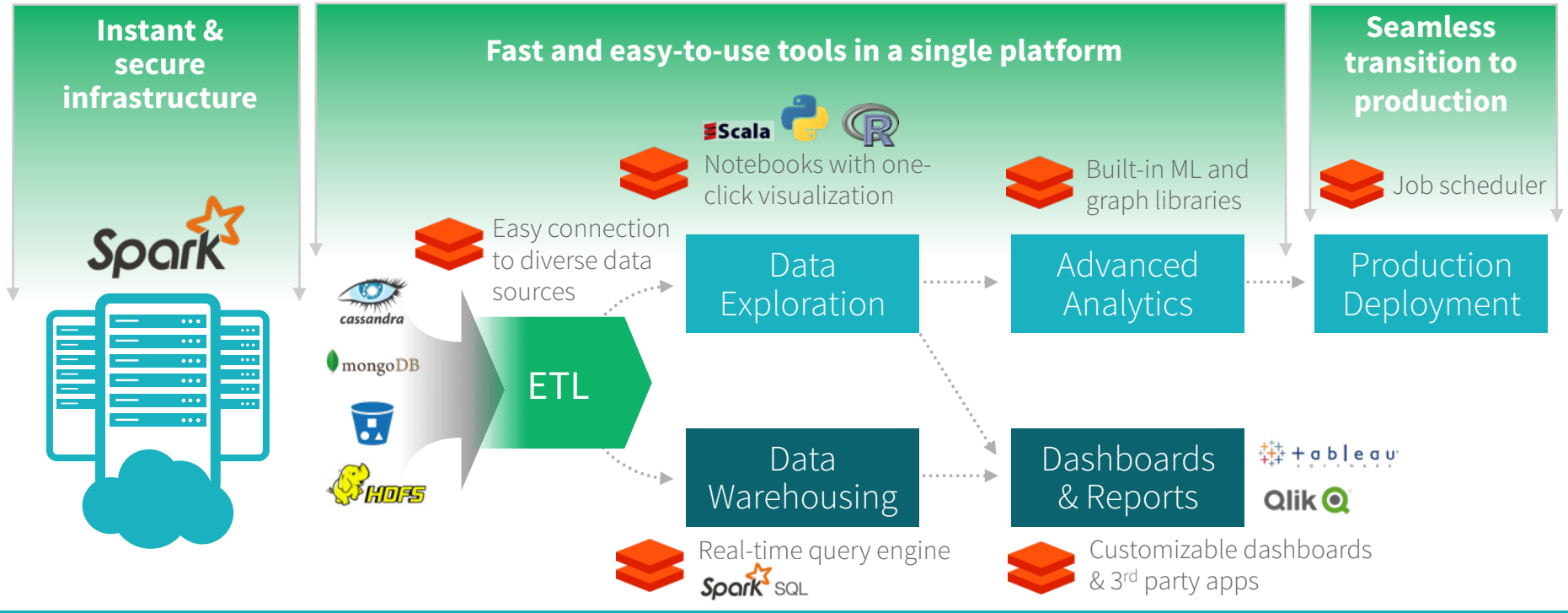
- Duplicated effort
- Complexity to achieve production quality

SOLUTION



Build **Databricks** on top of Spark to **make big data simple**

A complete solution, from ingest to production



Short time to value

Lower costs

Four components of Databricks

Make Big Data simple

Managed Spark clusters

- Easily provision clusters
- Harness the power of Spark
- Import data seamlessly

Production pipeline scheduler

- Schedule production workflows
- Implement complete pipelines
- Monitor progress and results

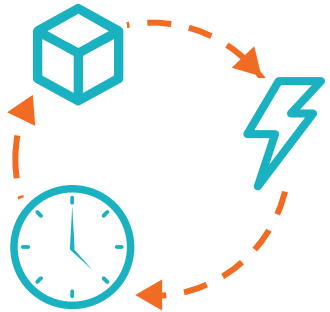
Interactive workspace with notebooks

- Explore data and develop code in Java, Python, Scala, or SQL
- Collaborate with the entire team
- Point and click visualization
- Publish customized dashboards

3rd party applications

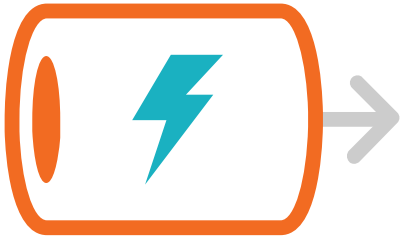
- Connect powerful BI tools
- Leverage a growing ecosystem of applications

Databricks benefits



Higher productivity

- Maintenance-free infrastructure
- Real time processing
- Easy to use tools



Faster deployment of data pipelines

- Zero management Spark clusters
- Instant transition from prototype to production

Data democratization

- One shared repository
- Seamless collaboration
- Easy to build sophisticated dashboards and notebooks

A few examples of Databricks in action

Prepare data

- Import data using APIs or connectors
- Cleanse mal-formed data
- Aggregate data to create a data warehouse

Perform analytics

- Explore large data sets in real-time
- Find hidden patterns with regression analysis
- Publish customized dashboards

Build data products

- Rapid prototyping
- Implement advanced analytics algorithms
- Create and monitor robust production pipelines

CUSTOMER CASE STUDIES

Customer testimonials

"Without Databricks and the real-time insights from Spark, we wouldn't be able to maintain our database at the pace needed for our customers"

Darian Shirazi, CEO, Radius Intelligence

"We condensed the 6 months we had planned for the initial prototype to production process to just about a couple of weeks with Databricks."

Rob Ferguson, Director of Engineering, Automatic Labs

"Databricks is used by over a third of our staff; After implementation, the amount of analysis performed has increased sixfold, meaning more questions are being asked, more hypotheses tested."

Jaka Jančar, CTO, Celtra

Radius Intelligence

Gathering customer insights for marketers

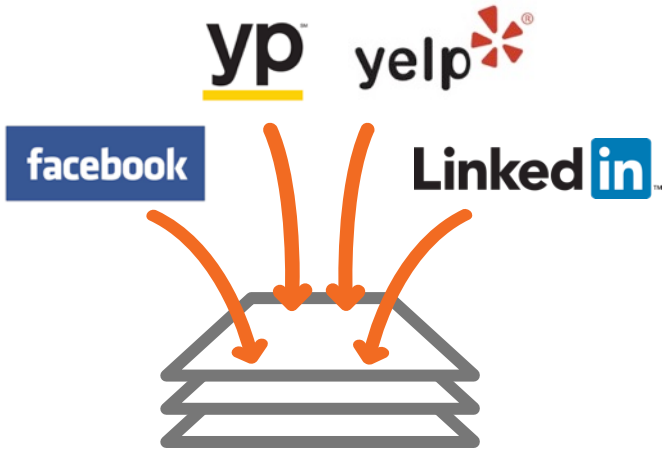
RADIUS®

CHALLENGE: Complex data integration

- 25 million businesses
- Over 100 billion points of data

RESULT: Speed up the data pipeline

- Entire data set processed in hours instead of days
- Deploy weekly updates to customers instead of monthly



BENEFIT:
Higher productivity

Automatic Labs

IoT for drivers – making car sensor data useful

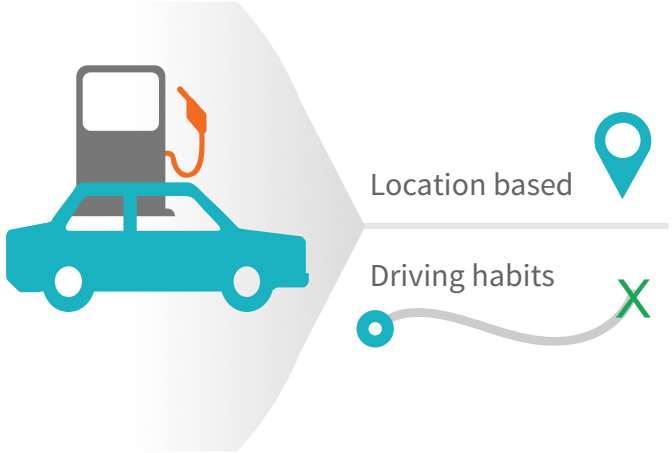
AUTOMATIC

CHALLENGE: Product idea validation

- Ingest billions of data points
- Rapidly test hypothesis
- Iterate on ideas in real-time

RESULT: Shorter time from idea to product

3 weeks with Databricks vs.
2 months with previous solution



BENEFIT:
Faster Deployment of Data Pipelines

databricks

Celtra

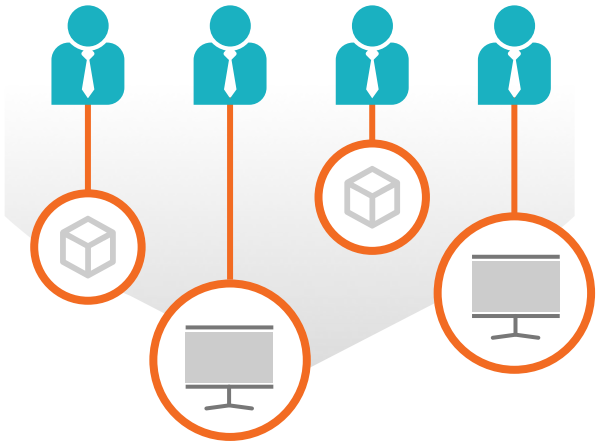
Building and serving rich digital ads across platforms

CHALLENGE: Analytics specialist bottleneck

- Billions of data points, operational data of the entire company
- Huge backlog of analytics projects

RESULT: Enable self-service for non-specialists

- Grew the number of analysts by 4x
- Increased analytics project completed by 6x in four months



BENEFIT:
Data Democratization

Sharethrough

Intelligent ad placement

CHALLENGE: Slow performance, costly DevOps

- Terabyte-scale clickstream data, long delays in new feature prototyping
- Two full-time engineers to maintain infrastructure

RESULT: Faster answers, zero management

- Prototyped new feature in record time
- Reduced system downtime with faster root-cause analysis
- Dramatically easier to maintain than Hive

BENEFIT: Higher Productivity

Yesware

Delivering tools and analytics for sales teams

CHALLENGE: Infrastructure pain, slow & clunky tools

- 6 months to setup Pig, very problematic pipeline
- Too slow to extend reporting history beyond 1 month
- Need to develop machine-learning algorithms

RESULT: Instant infrastructure, full suite of capabilities

- 3 weeks to setup robust Spark pipeline w/Databricks
- Double data processed, in fraction of time
- Built-in machine learning libraries

BENEFIT:
Faster Deployment of Data Pipelines

A few of our customers

AUTOMATIC

celtra

getexp

Mediative

myfitnesspal

OpenTable

picwell
choose smarter

RADIUS

sellpoints

sharethrough

trueffect

Yesware

WHAT'S NEW?

What's new with Databricks

- Databricks is now generally available (announced on June 15th, 2015)
- Upcoming features during second half of 2015:
  - **R-language notebooks:** Analyze large-scale data sets using R in the Databricks environment.
  - **Access control and private notebooks:** Manage permissions to view and execute code at an individual level.
  - **Version control:** Track changes to source code in the Databricks platform.
  - **Spark streaming support:** Enabling a fault-tolerant real-time processing

What's new with Spark

- The general availability of Spark 1.4 was announced on June 10th 2015
- Spark 1.4 is largest Spark release: more than 220 contributors and 1,200 commits.
- Key new features introduced in Spark 1.4:
  - New R language API (SparkR)
  - Expansion of Spark's Dataframe API's: window functions, statistical and mathematical functions, support for missing data.
  - API to build complete machine learning pipelines.
  - UI visualizations for debugging and monitoring programs: interactive event timeline for jobs, DAG visualization, visual monitoring for Spark Streaming.

Data science made easy with Apache Spark

From ingest to production

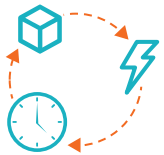
Spark

- ☑ Unified
- ☑ Fast at any scale
- ☑ Flexible
- ☑ No lock-in

+

databricks™

- ☑ Zero management
- ☑ Real-time and collaborative
- ☑ Instant to production
- ☑ Open and extensible



Higher productivity

Faster deployment of data pipelines

Data democratization

Databricks is available today

Contact sales@databricks.com

Or sign up for a trial at
https://databricks.com/registration

Thank you