Spark in the Wild:

150+

An In-Depth Analysis of 50+ Production Deployments

Arsalan Tavakoli-Shiraji

June 16, 2015

databricks™

Databricks

Founded by the creators of Spark in 2013

Largest contributor to Spark

End-to-end hosted cloud platform

Questions we’re hoping to answer

1. Why did they choose Spark?
2. How did they use Spark?
3. What were the challenges?

Study set: 150+ production deployments

Company size

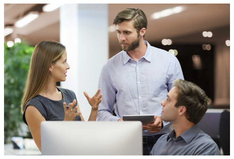
< 10 employees

Fortune 50

Industry

Advertising & Marketing

Energy & Utilities



Enterprise Technology

Financial & Insurance

Healthcare & Pharma

Media & Entertainment

Retail and Consumer

Telecom

Questions we're hoping to answer

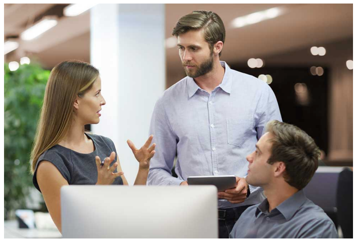
1. **Why did they choose Spark?**
2. **How did they use Spark?**
3. **What were the challenges?**

databricks

From a business perspective…

Productivity & Time to Value

F100 Media


F100 Technology

Consumer electronics

New Product Enablement

myfitnesspal

OpenTable™

picwell®
choose smarter

Depends on previous Hadoop usage

Existing Hadoop user (60%)

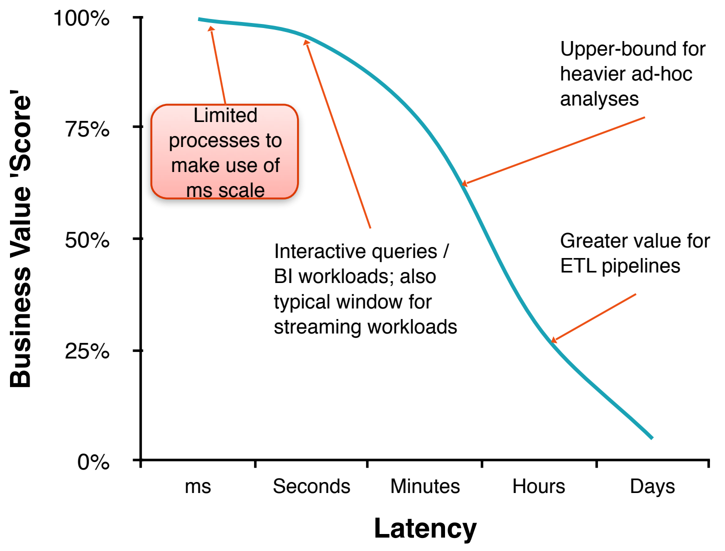
1. Efficiency of ETL / data pipeline — RADIUS
2. Ease and speed of ad-hoc exploration — CONVIVA
3. Combining multiple analytics capabilities — Large education company

Never used Hadoop (40%)

1. Production ML and data science at scale — Digital Health company
2. Analysis across multiple data sources — Financial services company
3. Moving beyond SQL-based analyses — Large gaming company

Extreme scale and latency not the norm…

“Real-time” is relative



Limited processes to make use of ms scale

Interactive queries / BI workloads; also typical window for streaming workloads

Upper-bound for heavier ad-hoc analyses

Greater value for ETL pipelines

Big Data != Massive Clusters

A Cluster size often driven by storage vs. processing needs

Large financial services firm

B Significant performance inefficiency in user code

Large ad-tech company

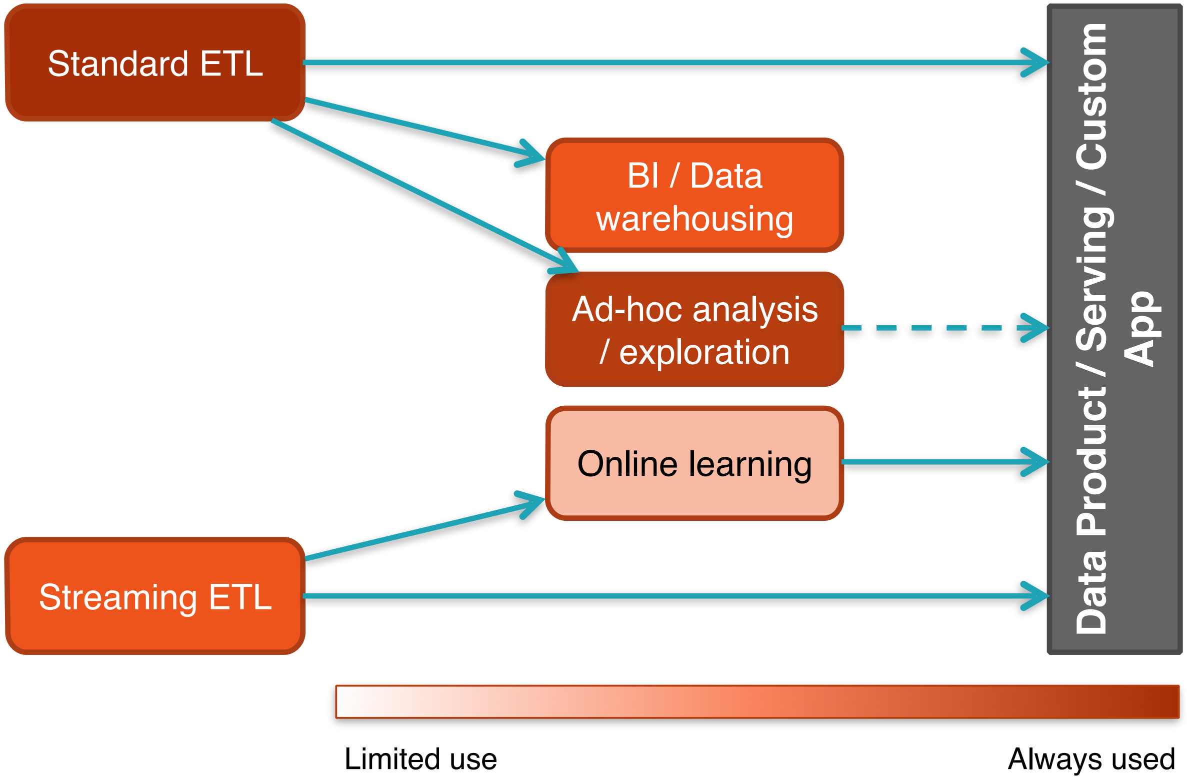
C Separate clusters for use cases becoming more common

Large gaming company

Questions we’re hoping to answer

1. **Why did they choose Spark?**
2. **How did they use Spark?**
3. **What were the challenges?**

Every use case leverages Spark for ETL



databricks

Nearly 100% of deployments use SQL

Many organizations have **data analysts** that are most comfortable with SQL

Used fairly often for **ETL pipelines** – often in conjunction with custom UDFs

SchemaRDD kick-started usage; **DataFrames** have accelerated this

Data becoming more distributed

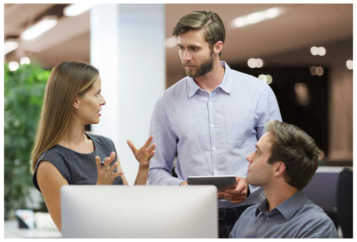
>1/3rd

Used multiple data sources

Financial Company

60%+

Used a non-HDFS data source



Software Company

1. **In many cases, data 'unification' taking place at processing layer**
2. **As such, seeing compute and storage become decoupled**

Questions we’re hoping to answer

1. Why did they choose Spark?
2. How did they use Spark?
3. **What were the challenges?**

Easier than alternatives, but still not easy

Configuration and tuning still difficult

Often lots of room for **performance optimizations** but these require Spark expertise

Debugging distributed systems is still a fundamentally hard problem

Spark has and continues to make significant strides here

Difficulty sizing environment

F100 Manufacturer Example

A TB's of existing data in AWS S3; significant growth expected

B Initial use by 7-10 data scientists; likely grow to 50+ within 12 months

C Workload: data exploration, machine learning, and streaming analytics

Cluster size?

Some high-level thoughts

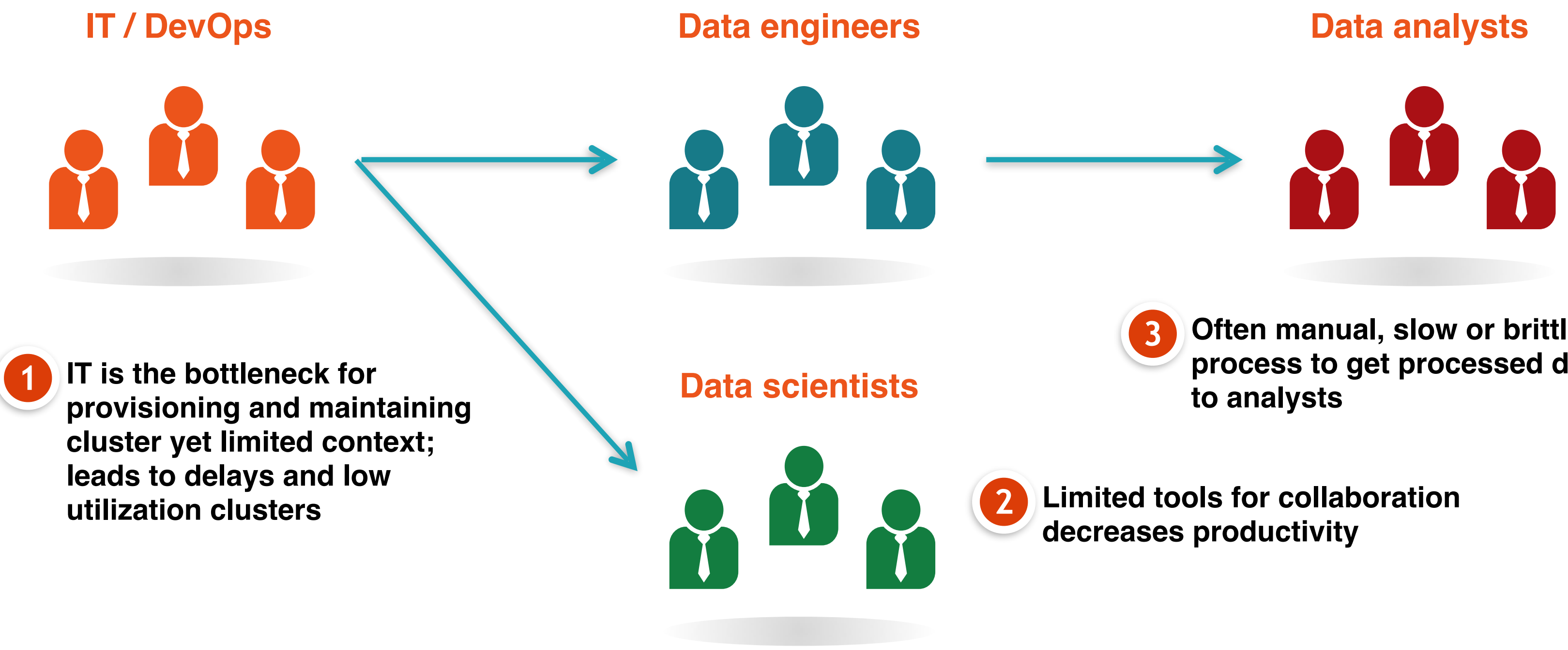
1 Spark doesn't require data to be cached in memory so data size is poor sizing metric

2 Lots of inefficiencies in user code; bigger impact than cluster size

3 Performance and cluster size typically highly correlated (to a point)

4 Best solution is often separate clusters per use case – ideally with dynamic sizing

Many dependencies; collaboration hard



IT / DevOps

Data engineers

Data analysts

Data scientists

1 IT is the bottleneck for provisioning and maintaining cluster yet limited context; leads to delays and low utilization clusters

2 Limited tools for collaboration decreases productivity

3 Often manual, slow or brittle process to get processed data to analysts

Significant push for 'data democratization' and self-service

Enterprise security model evolving

Large Technology Company

Ⓐ Significant amounts of data spread across S3 and Redshift

Ⓑ 500 projected users with a variety of permissions; column level role-based access control needed

Ⓒ Leveraging just Spark as their processing engine

Best way to secure?

Some high-level thoughts

1. Leveraging storage-level security mechanisms difficult with heterogeneous storage sources
2. Many organizations have existing security mechanisms they want to integrate with
3. Solution likely needed at compute or application level

Key takeaways from today

Spark is being used in production across a broad range of verticals and enterprises today

Data – importing, transforming, exploring, and making it readily accessible – is at the **core of Spark adoption**

Traditional approaches for Hadoop deployments may **not be the most applicable** for Spark

Thank you.

Sign-up for a Databricks trial at www.databricks.com

databricks™