Introduction to Spark

Matei Zaharia

Databricks Intern Event, August 2015

What is Apache Spark?

Fast and general computing engine for clusters

Makes it easy and fast to process large datasets

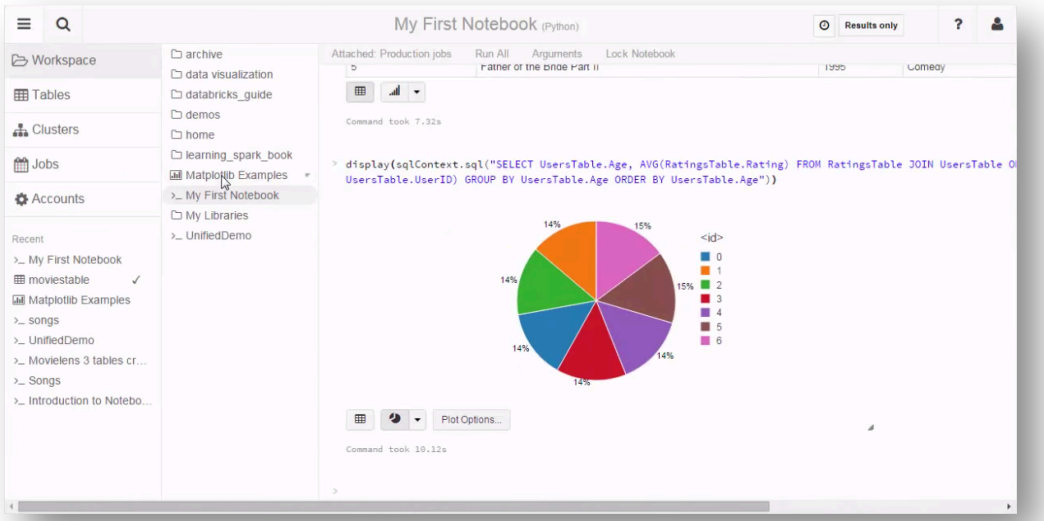
- APIs in Java, Scala, Python, R
- Libraries for SQL, streaming, machine learning, …
- 100x faster than Hadoop MapReduce for some apps

About Databricks

Founded by creators of Spark in 2013

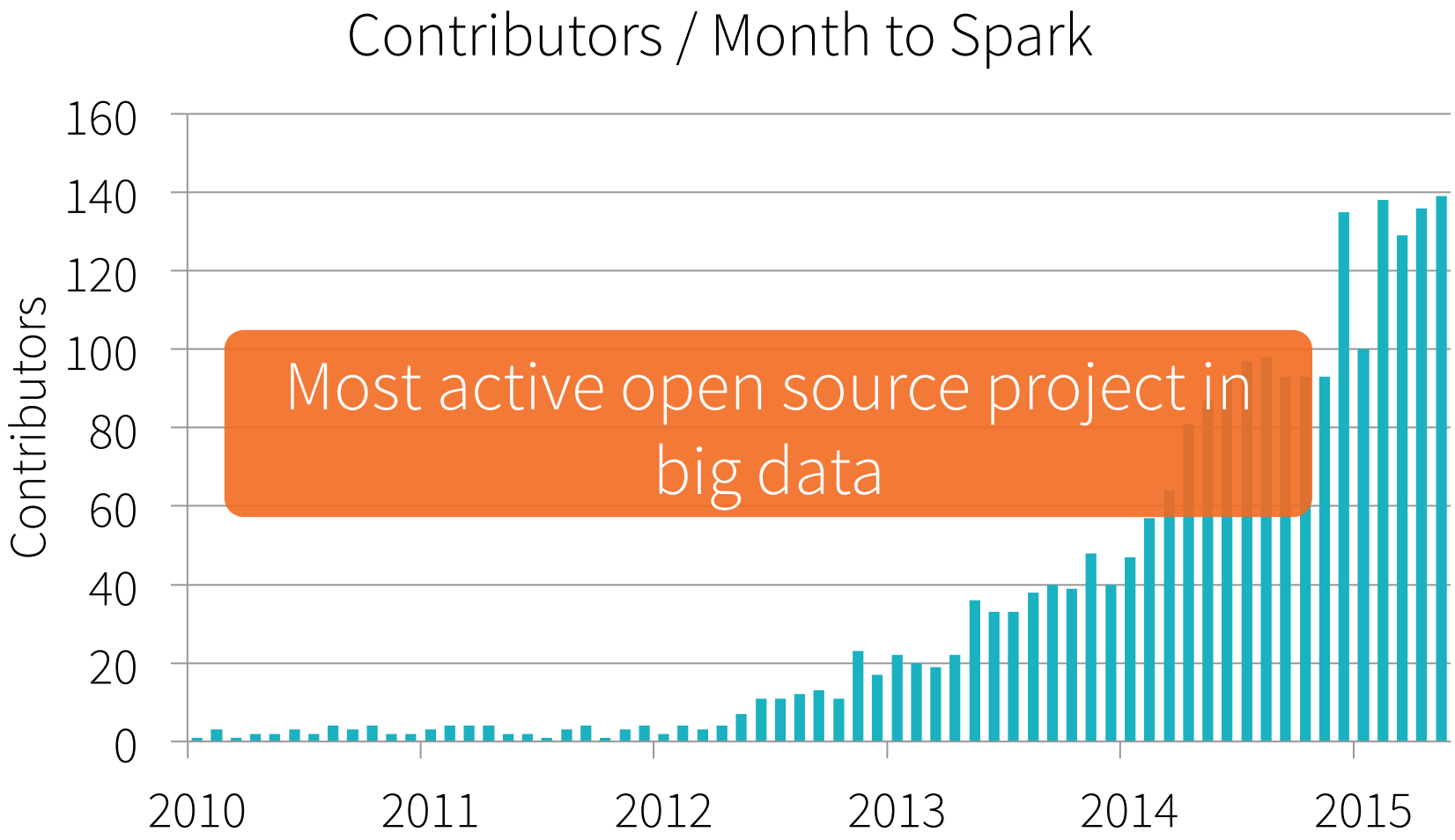
Offers a hosted cloud service built on Spark

- Interactive workspace with notebooks, dashboards, jobs



databricks

Community Growth



Spark Programming Model

Write programs in terms of transformations on distributed datasets

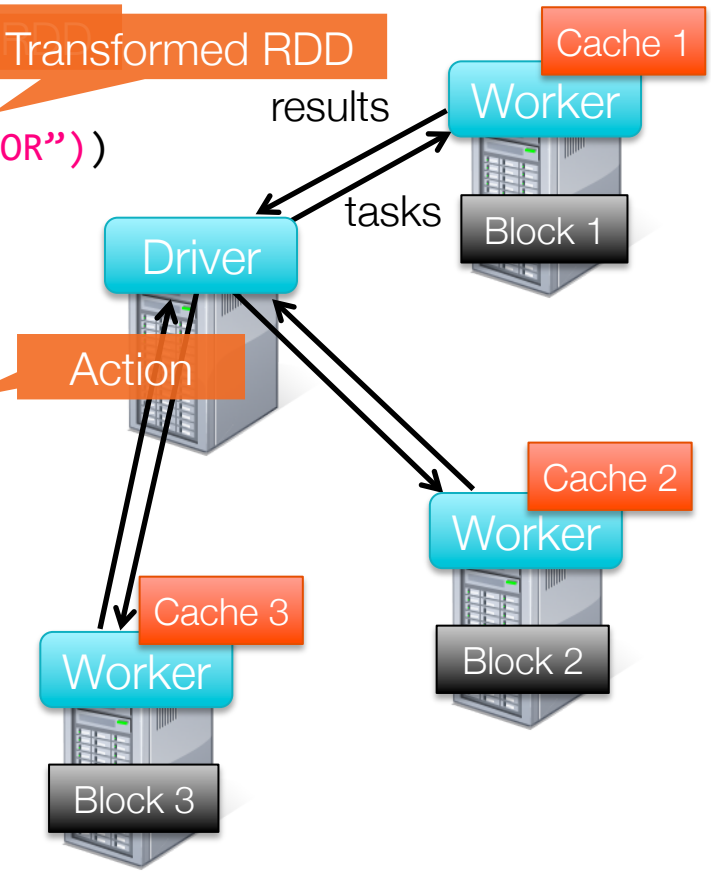
Resilient Distributed Datasets (RDDs)

- Collections of objects stored in memory or disk across a cluster
- Built via parallel transformations (map, filter, …)
- Automatically rebuilt on failure

Example: Log Mining

Load error messages from a log into memory, then interactively search for various patterns

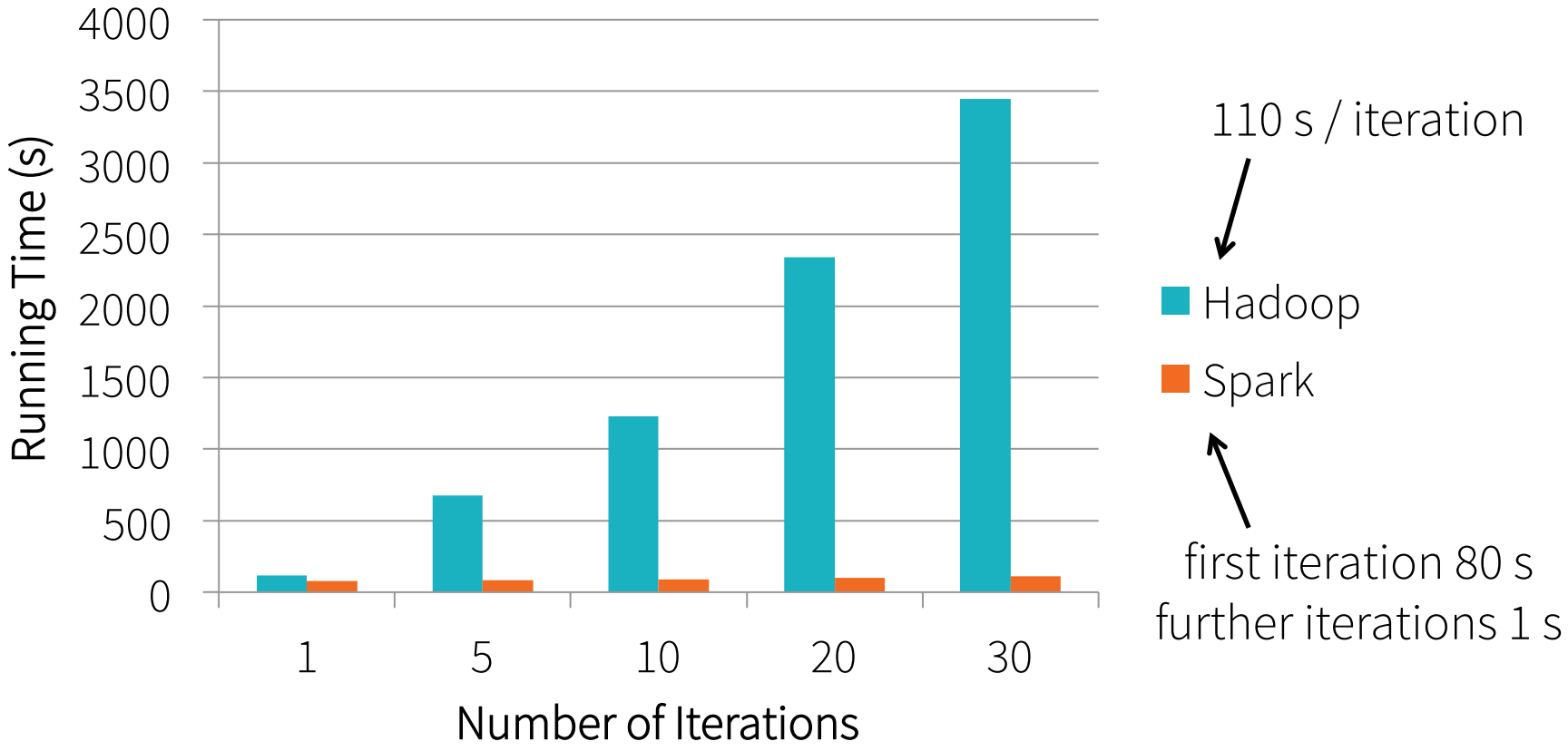
```
lines = spark.textFile(“hdfs://...”)
errors = lines.filter(lambda s: s.startswith(“ERROR”))
messages = errors.map(lambda s: s.split(‘\t’)[2])
messages.cache()

messages.filter(lambda s: “MySQL” in s).count()
messages.filter(lambda s: “Redis” in s).count()
. . .
```



Result: full-text search of Wikipedia in 0.5 sec (vs 20s for on-disk data)

Example: Logistic Regression

Iterative algorithm used in machine learning



On-Disk Performance

Time to sort 100TB

2013 Record: Hadoop

2100 machines

72 minutes

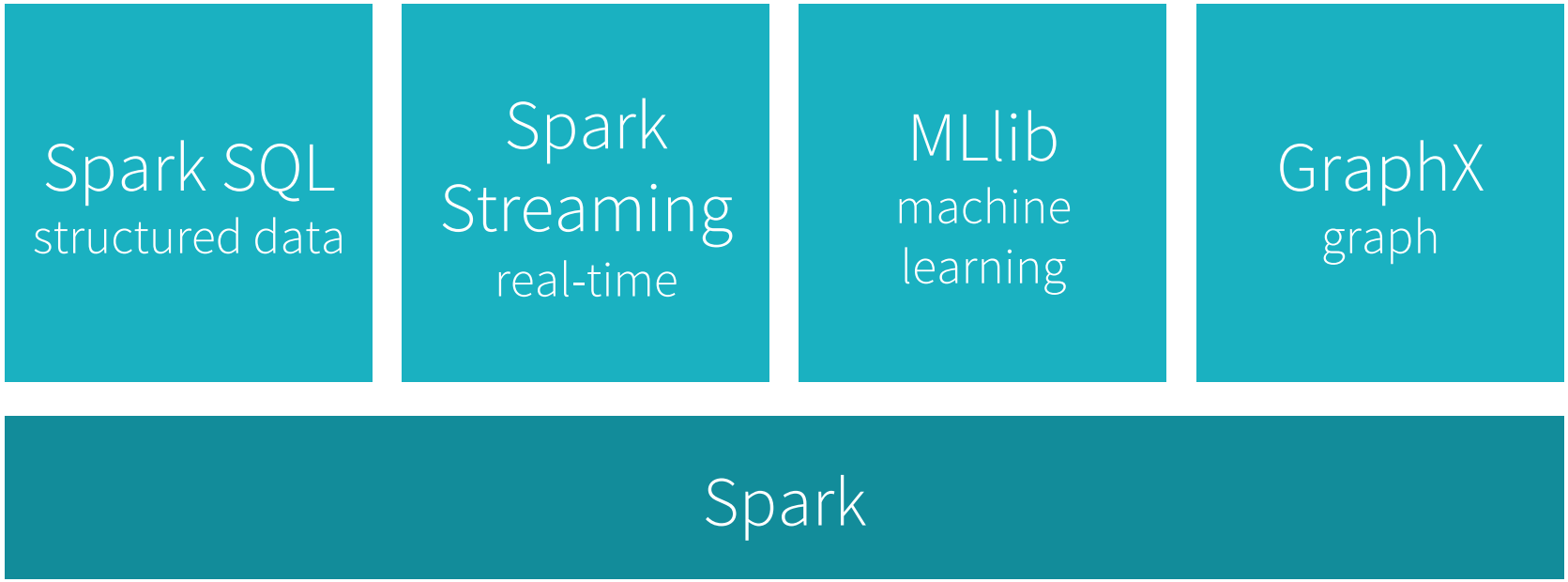
2014 Record: Spark

207 machines

23 minutes

Source: Daytona GraySort benchmark, sortbenchmark.org

Higher-Level Libraries



Higher-Level Libraries

```
// Load data using SQL
points = ctx.sql(“select latitude, longitude from tweets”)

// Train a machine learning model
model = KMeans.train(points, 10)

// Apply it to a stream
sc.twitterStream(...)
  .map(lambda t: (model.predict(t.location), 1))
  .reduceByWindow(“5s”, lambda a, b: a + b)
```

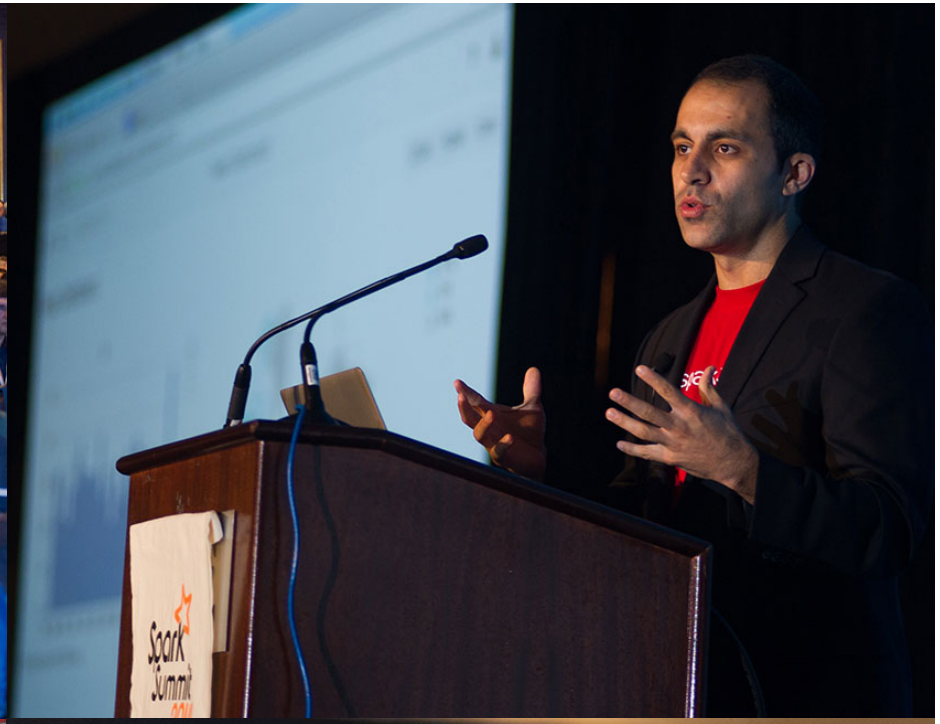
Demo

Spark Community

Over 1000 production users, clusters up to 8000 nodes

Many talks online at spark-summit.org


Spark Summit 2014


Spark Summit


GridGain
amazon web services
Spark Summit 2014

Ongoing Work

Speeding up Spark through code generation and binary processing (Project Tungsten)

R interface to Spark (SparkR)

Real-time machine learning library

Frontend and backend work in Databricks (visualization, collaboration, auto-scaling, …)

Thank you.

We're hiring!