DATABRICKS

Spark's Role in the Big Data Ecosystem

Matei Zaharia

An Exciting Year for Spark

Very fast community growth

1.0 release in May

7+ distributors, 20+ apps

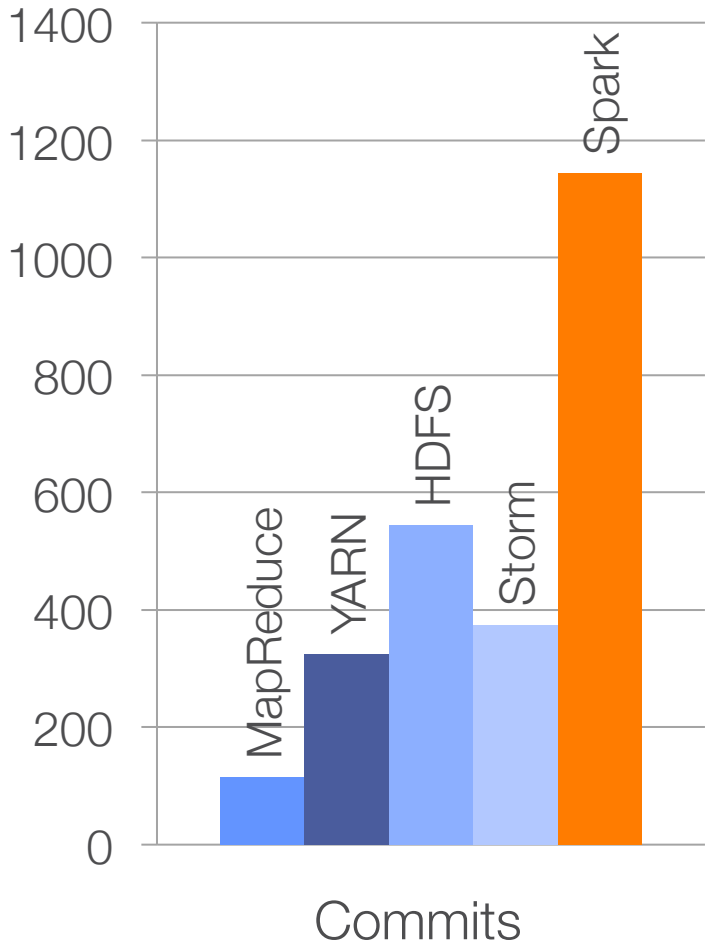
Project Activity

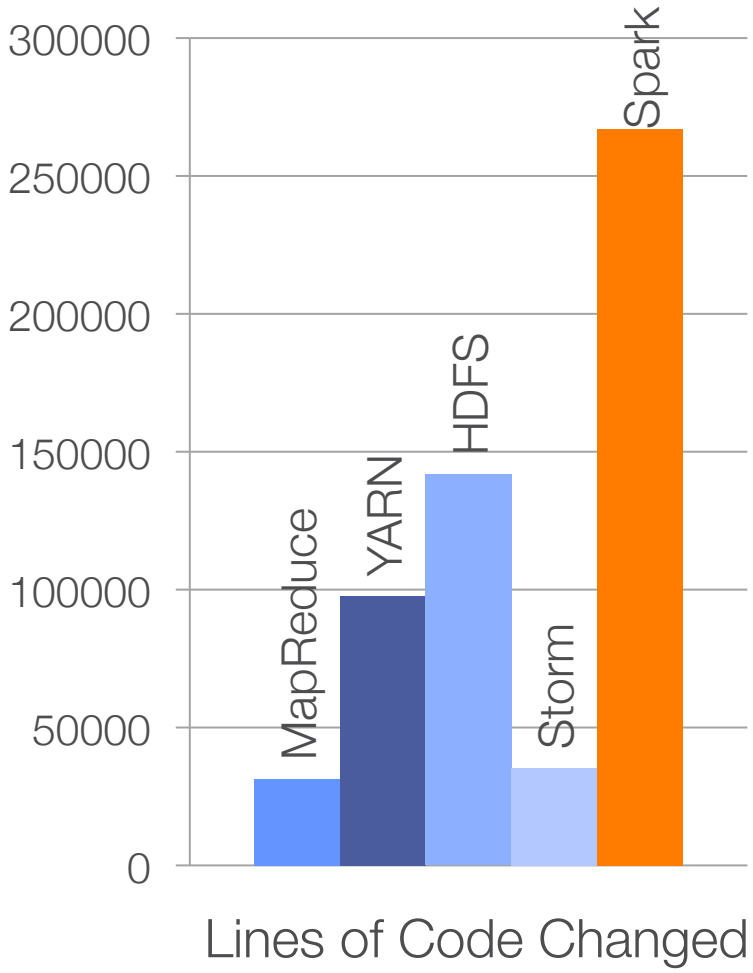
	June 2013
total contributors	68
companies contributing	17
total lines of code	63,000

Project Activity

	June 2013	June 2014
total contributors	68	255
companies contributing	17	50
total lines of code	63,000	175,000

Compared to Other Projects

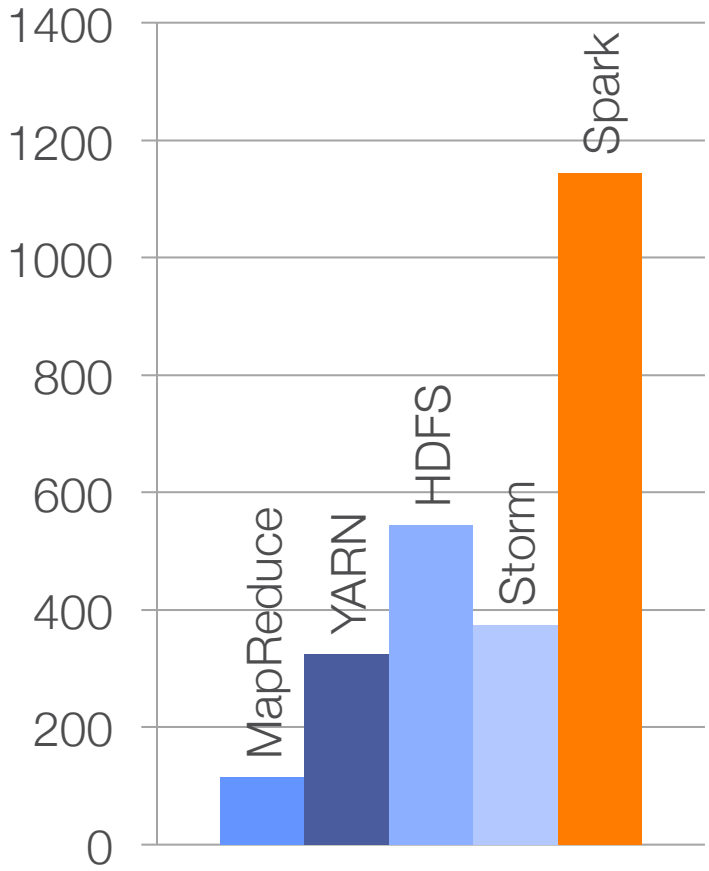


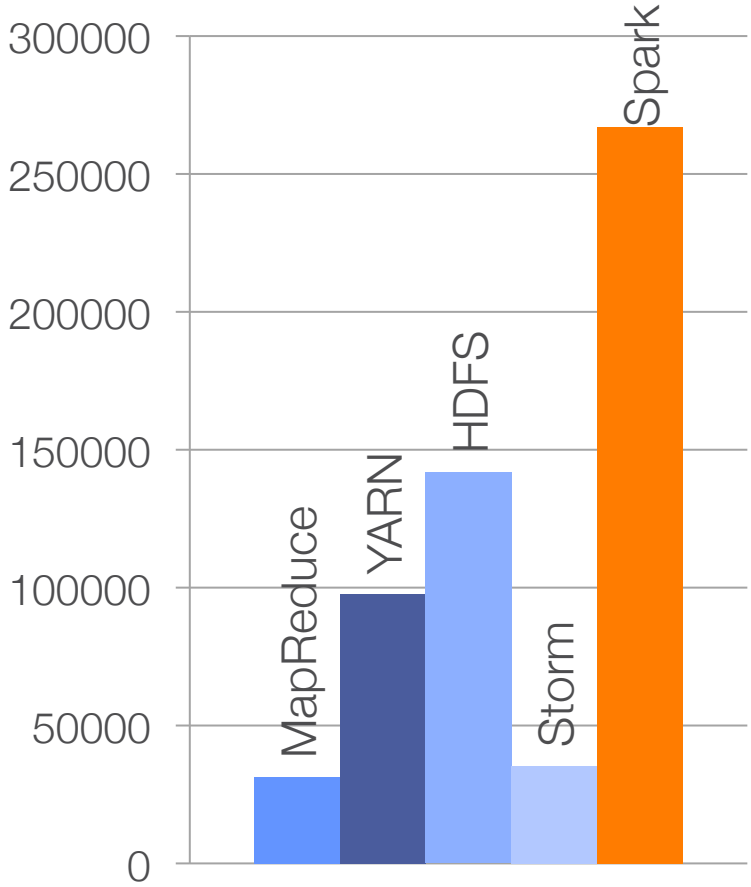


Activity in past 6 months

DATABRICKS

Compared to Other Projects





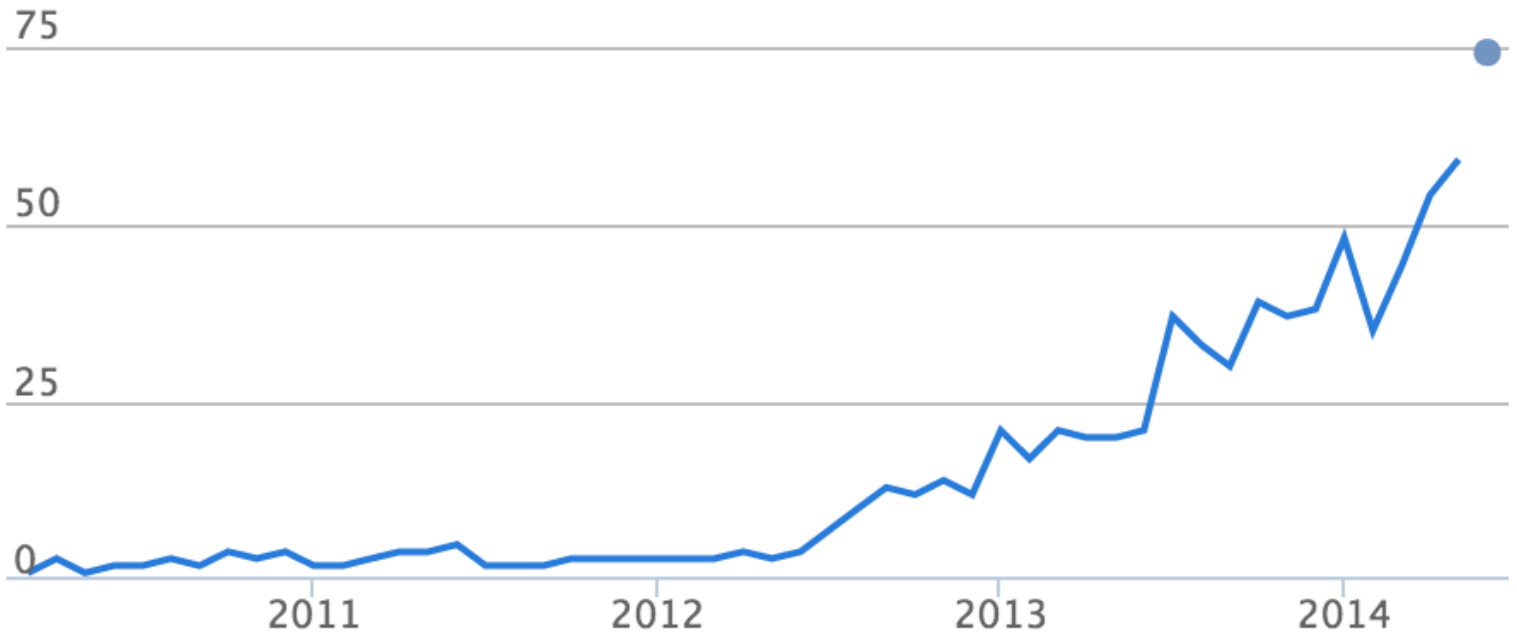
Spark is now the most active project in the Hadoop ecosystem

Compared to Other Projects

Spark is one of top 3 most active projects at Apache

More active than “general” data processing projects like NumPy, matplotlib, SciKit-Learn

Continuing Growth



Contributors per month to Spark

source: ohloh.net

Major new additions

Last Summit

Last Summit we said we'd focus on two things:

- Standard libraries
- Enterprise features

New libraries: Spark SQL, MLlib (machine learning), GraphX (graph processing)

Enterprise features: security, monitoring, HA

Spark SQL

Enables loading & querying structured data in Spark

From Hive:

```
c = HiveContext(sc)
rows = c.sql("select text, year from hivetable")
rows.filter(lambda r: r.year > 2013).collect()
```

From JSON:

```
c.jsonFile("tweets.json").registerAsTable("tweets")
c.sql("select text, user.name from tweets")
```

tweets.json

```
{"text": "hi",
 "user": {
   "name": "matei",
   "id": 123
}}
```

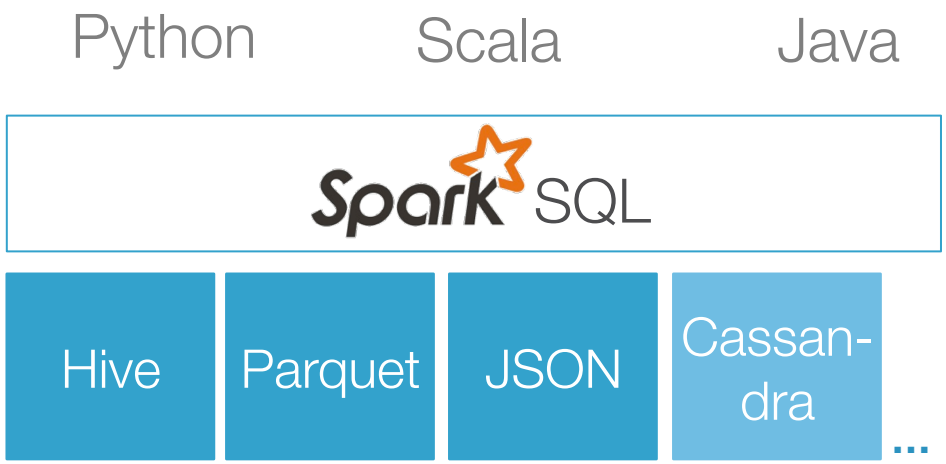
DATABRICKS

Spark SQL

Integrates closely with Spark's language APIs

```
c.registerFunction("hasSpark", lambda text: "Spark" in text)
c.sql("select * from tweets where hasSpark(text)")
```

Uniform interface for data access



44 contributors in past year

DATABRICKS

Machine Learning Library (MLlib)

Standard library of machine learning algorithms

Now includes 15+ algorithms

- New in 1.0: decision trees, SVD, PCA, L-BFGS
- In development: non-negative matrix factorization, LDA, Lanczos, multiclass trees, ADMM

```
points = context.sql(“select latitude, longitude from tweets”)
model = KMeans.train(points, 10)
```

40 contributors in past year

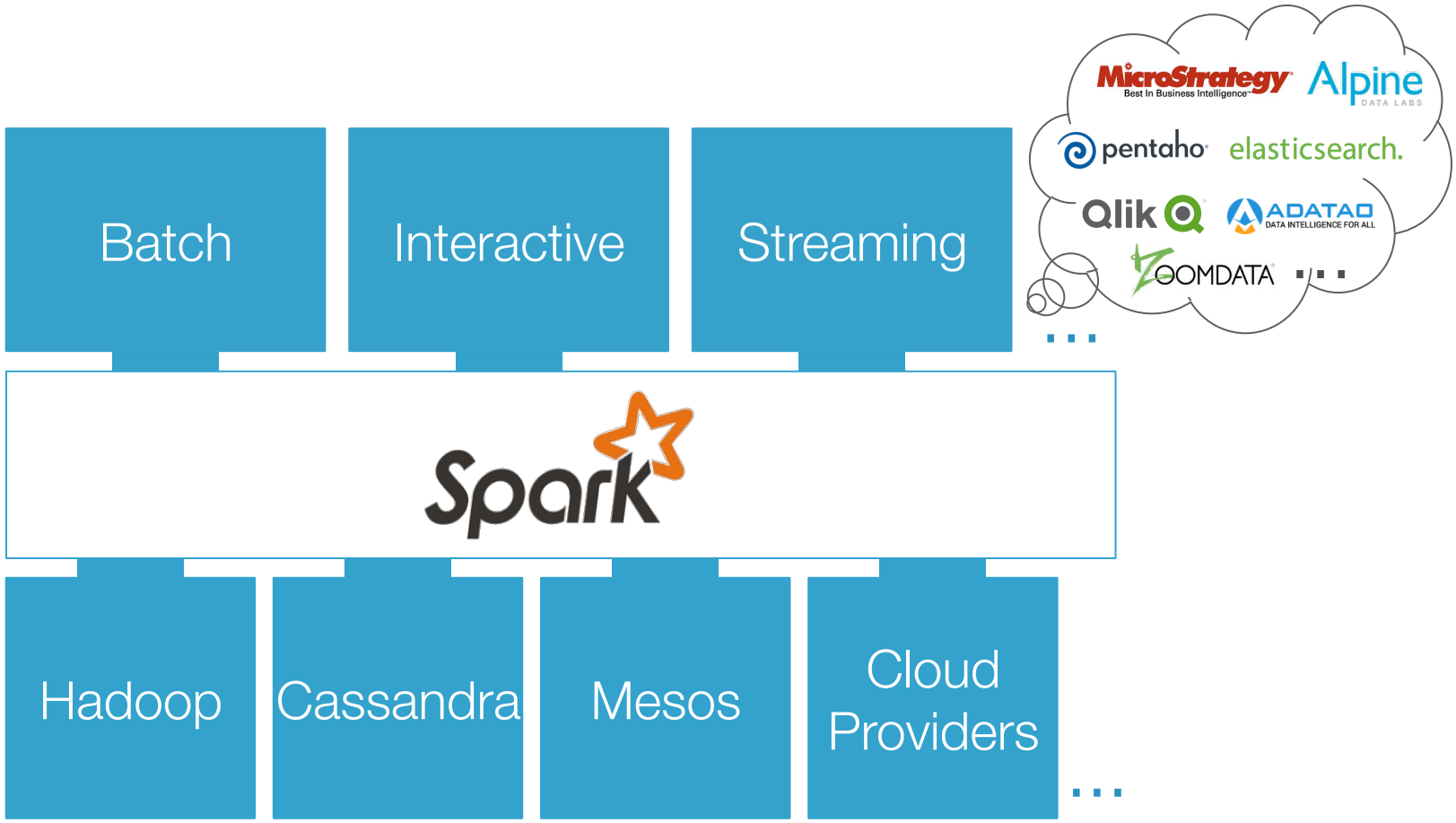
Java 8 API

Enables concise programming in Java similar to Scala and Python

```
JavaRDD<String> lines = sc.textFile("data.txt");
JavaRDD<Integer> lineLengths = lines.map(s -> s.length());
int totalLength = lineLengths.reduce((a, b) -> a + b);
```

What is our vision for Spark?

1. Unified Platform for Big Data Apps



Uniform API for *diverse workloads* over *diverse storage systems and runtimes*

DATABRICKS

Why a Platform Matters

Good for developers: one system to learn

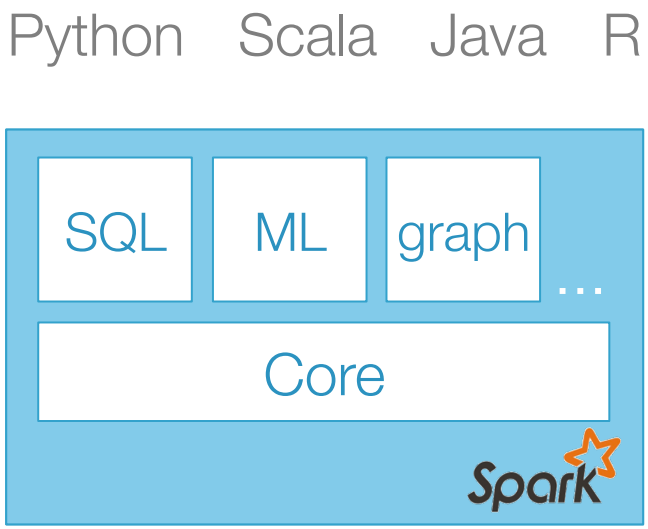
Good for users: take apps anywhere

Good for distributors: more applications

2. Standard Library for Big Data

Big data apps lack libraries of common algorithms

Spark's generality + support for multiple languages make it suitable to offer this



Much of future activity will be in these libraries

Databricks & Spark

At Databricks, we are working to keep Spark 100% open source and compatible across vendors

All our work on Spark is at Apache

Check out project-specific talks to see what's next!

DATABRICKS

Thank You and Enjoy Spark Summit!