Apache Spark in 2015 and Beyond

Reynold Xin (@rxin)

ApacheCon, Apr 16, 2015

databricks™

Who am I?

Reynold Xin

Spark PMC member

Databricks co-founder & architect

UC Berkeley AMPLab PhD (on leave)

About Databricks

Founded by creators of Spark and is the largest contributor to Spark

Databricks Cloud (in limited availability)

- Fully managed Spark clusters (one-click launch)
- Interactive workspace
- Production data pipelines
- 3rd party applications

Show of Hands!

How familiar are you with Spark?

A. Heard of it, but haven't used it before.
B. Kicked the tires with some basics.
C. Worked or working on a POC.
D. Worked or working on a production deployment.

Alpha / Pre-alpha

BlinkDB
Approximate SQL

Spark SQL

Spark Streaming
Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark R
R on Spark

Spark Core Engine

Fast and general engine for distributed data processing

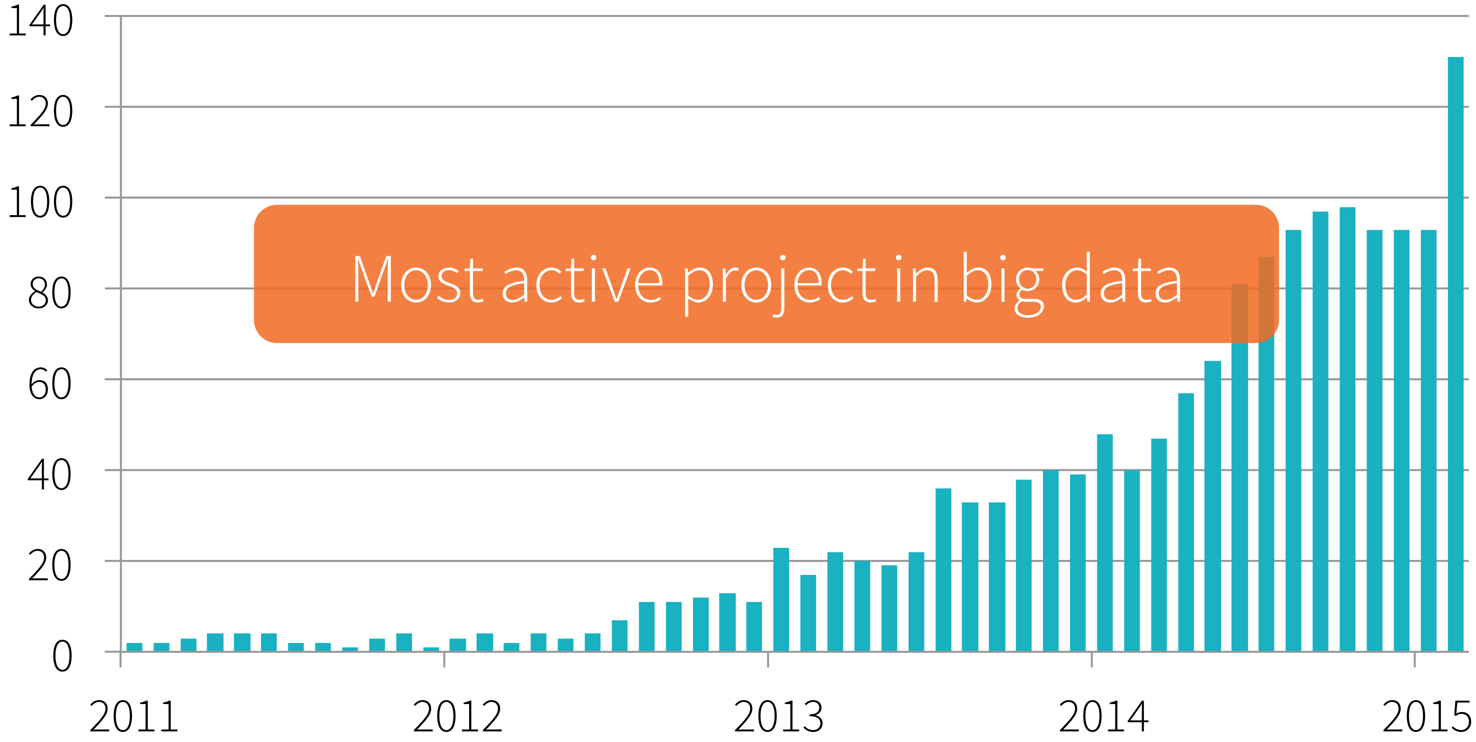
2014: an Amazing Year for Spark

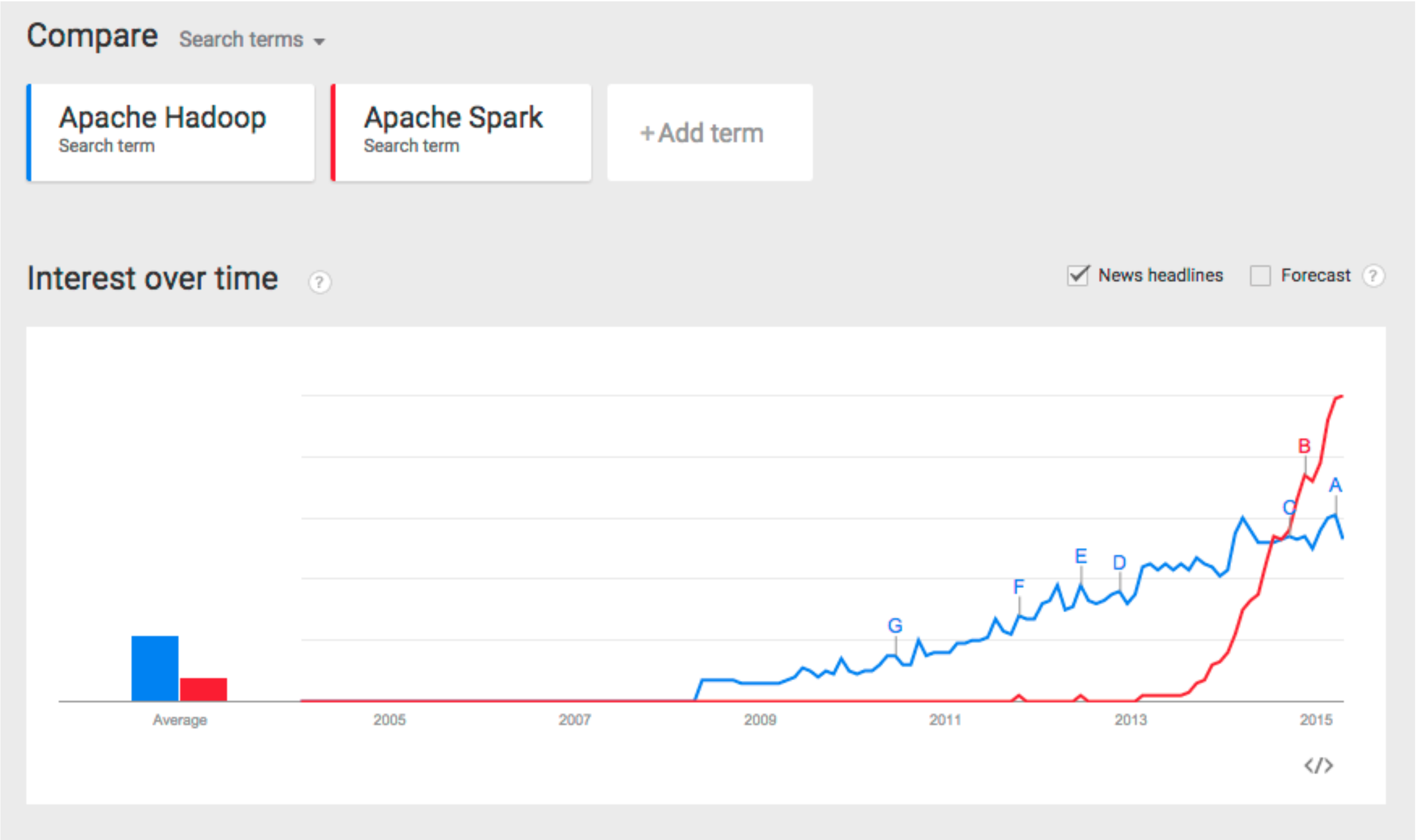
Total contributors: 150 → 500

Lines of code: 190K → 370K

500+ active production deployments

Contributors per Month to Spark





Note： not a scientific comparison.

databricks

2014 Focus

Richer libraries

Core engine: stability, performance, and enterprise readiness

Libraries in 2014

Spark SQL

Java 8 closure support

Python streaming

…

GraphX

Random forests

Streaming MLlib

Core Engine in 2014

Revamped “shuffle” and network transport

– Higher throughput and lower memory usage

Much better memory estimation and management

– harder to get OutOfMemoryError

Enterprise Readiness

– YARN integration, security, …

On-Disk Sort Record:
Time to sort 100TB

2013 Record: Hadoop

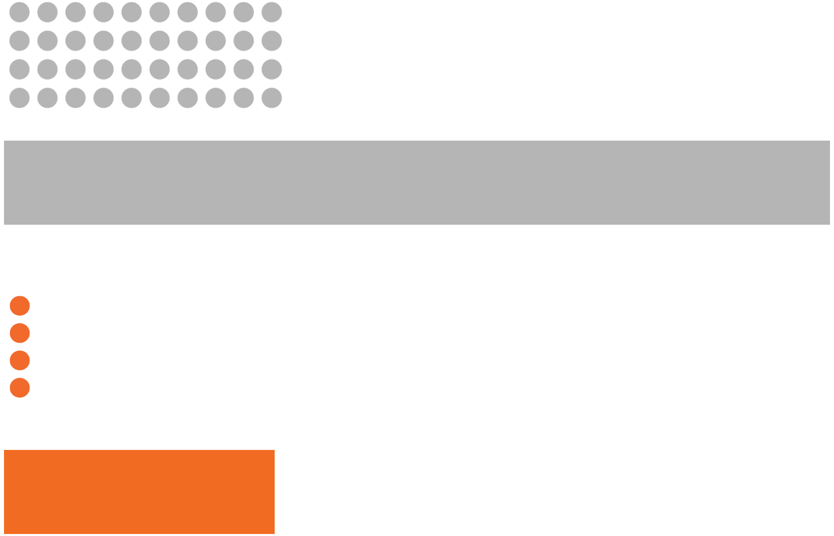
2100 machines

72 minutes

2014 Record: Spark

207 machines

23 minutes



Source: Daytona GraySort benchmark, sortbenchmark.org

Continuing in 2015

Performance & robustness

Security

– Option to encrypt all communication (control & data planes)

Usability

– Visualization and debugging tools

Continuing in 2015

MLlib

- Many new algorithms

GraphX

- Java API

Streaming

- Stronger integration with Kafka, etc

SQL

- Performance improvements, robustness

New Directions in 2015

Data Science

Making it easier for wider class of users

Platform Interfaces

Scaling the ecosystem

From MapReduce to Spark

```java
public static class WordCountMapClass extends MapReduceBase
  implements Mapper<LongWritable, Text, Text, IntWritable> {

  private final static IntWritable one = new IntWritable(1);
  private Text word = new Text();

  public void map(LongWritable key, Text value,
                  OutputCollector<Text, IntWritable> output,
                  Reporter reporter) throws IOException {
    String line = value.toString();
    StringTokenizer itr = new StringTokenizer(line);
    while (itr.hasMoreTokens()) {
      word.set(itr.nextToken());
      output.collect(word, one);
    }
  }
}

public static class WorkdCountReduce extends MapReduceBase
  implements Reducer<Text, IntWritable, Text, IntWritable> {

  public void reduce(Text key, Iterator<IntWritable> values,
                     OutputCollector<Text, IntWritable> output,
                     Reporter reporter) throws IOException {
    int sum = 0;
    while (values.hasNext()) {
      sum += values.next().get();
    }
    output.collect(key, new IntWritable(sum));
  }
}
```

```scala
val file = spark.textFile("hdfs://...")
val counts = file.flatMap(line => line.split(" "))
                 .map(word => (word, 1))
                 .reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://...")
```

databricks

Beyond Hadoop Users

Early adopters

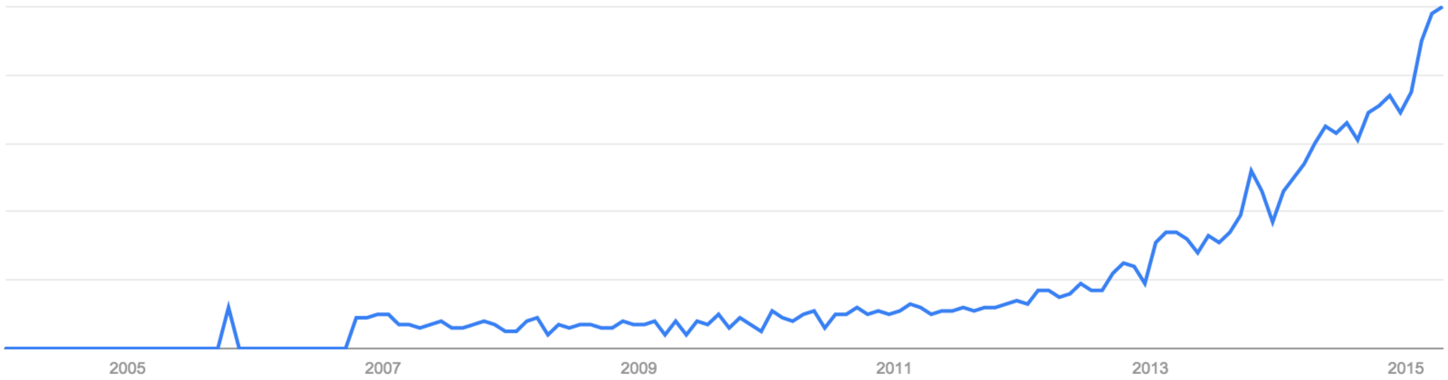
Users

Understands
MapReduce
& functional APIs

Data Scientists
Statisticians
R users …
PyData

Data Frames

De facto data processing abstraction for data science
(R and Python)



Google Trends for “dataframe”

databricks

Spark DataFrames

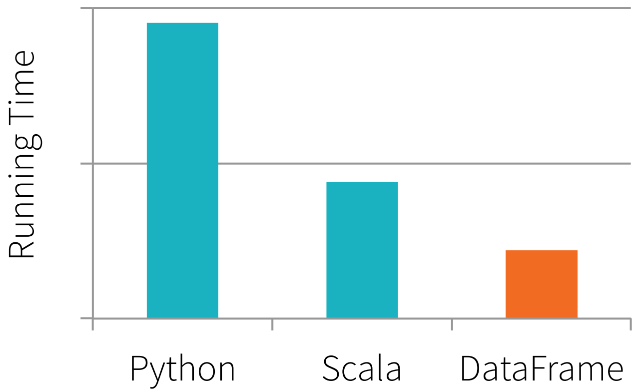
Similar API to data frames in R and Pandas

Automatically optimized via Spark SQL

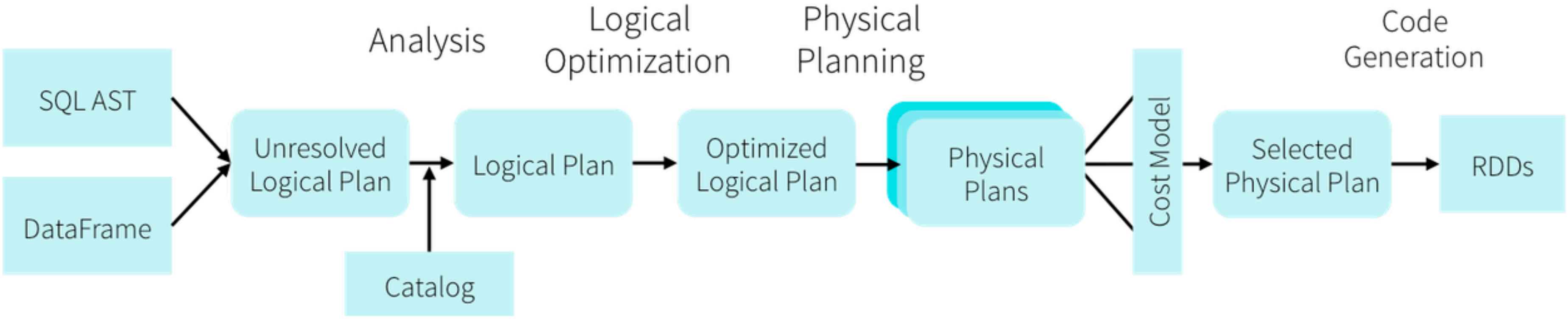
Out in Spark 1.3

```
df = jsonFile(“tweets.json”)

df[df[“user”] == “matei”]
  .groupBy(“date”)
  .sum(“retweets”)
```



Convergence of SQL and DataFrames



DataFrames and SQL share the same optimization/execution pipeline

Maximize code reuse & share optimization efforts

PySpark RDD vs DataFrame

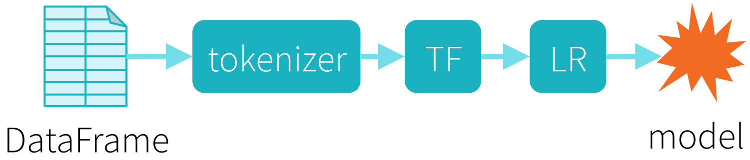
```
pdata.map(lambda x: (x.dept, [x.age, 1])) \
  .reduceByKey(lambda x, y: [x[0] + y[0], x[1] + y[1]]) \
  .map(lambda x: [x[0], x[1][0] / x[1][1]]) \
  .collect()
```

```
data.groupBy("dept").avg("age")
```

Machine Learning Pipelines

High-level API inspired by SciKit-Learn

Featurization, evaluation, parameter search




```
tokenizer = Tokenizer()
tf = HashingTF(numFeatures=1000)
lr = LogisticRegression()

pipe = Pipeline([tokenizer, tf, lr])
model = pipe.fit(df)
```

R Interface (SparkR)

Spark 1.4 (June)

Exposes DataFrames, and ML library in R

```
df = jsonFile("tweets.json")

summarize(
  group_by(
    df[df$user == "matei",],
    "date"),
  sum("retweets"))
```

Data Science at Scale

Higher level interfaces in Scala, Java, Python, R

Drastically easier to program Big Data

- With APIs similar to single-node tools

New Directions in 2015

Data Science

Making it easier for
wider class of users

Platform Interfaces

Scaling the ecosystem

Spark Mailing Lists (Mar 2015)

user:	2,577	dev:	456
issues:	4,473	reviews:	9,993

apache / spark
mirrored from git://git.apache.org/spark.git

Watch 648 Un

Pull requests | Labels | Milestones

Filters is:open is:pr

Clear current search query, filters, and sorts

290 Open ✔ 5,246 Closed Author Labels Milestones As

[SPARK-6867][MLlib] Add dropout regularization to logistic regression.
#5539 opened an hour ago by rakeshchalasani

[SPARK-6198][SQL] Support "select current_database()"
#5538 opened 2 hours ago by DoingDone9

[SPARK-6955][NETWORK]Do not let Yarn Shuffle Server retry its server port.
#5537 opened 3 hours ago by SaintBacchus

[SPARK-6954] [YARN] Dynamic allocation: numExecutorsPending in ExecutorAllocationManager should become negative
#5536 opened 4 hours ago by piaozhexiu

[SPARK-6952] Handle long args when detecting PID reuse
#5535 opened 5 hours ago by punya

[SPARK-6893][ML] default pipeline parameter handling in python

Feature Requests and Patches

Can I add this new API zipWithIndexWithContext to RDD?

I want support for HBase/Cassandra/Accumulo/…

I want this algorithm for machine learning…

Platformization

Standardize interfaces for integration points (so implementations can run in many versions to come)

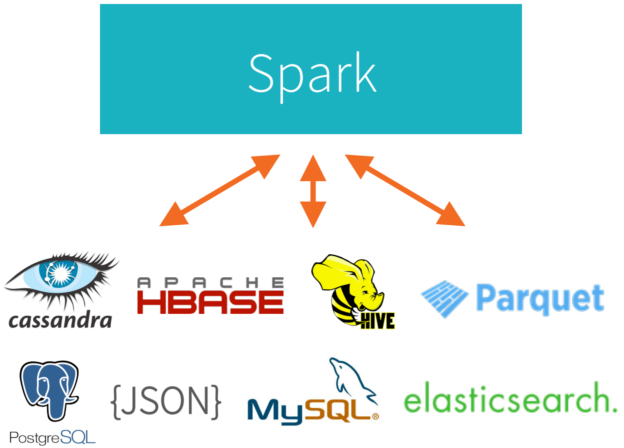
Make it as easy as possible to consume these implementations that are not in Apache Spark repo

External Data Sources

Platform API to plug smart data sources into Spark

Returns DataFrames usable in Spark apps or SQL

Pushes logic into sources

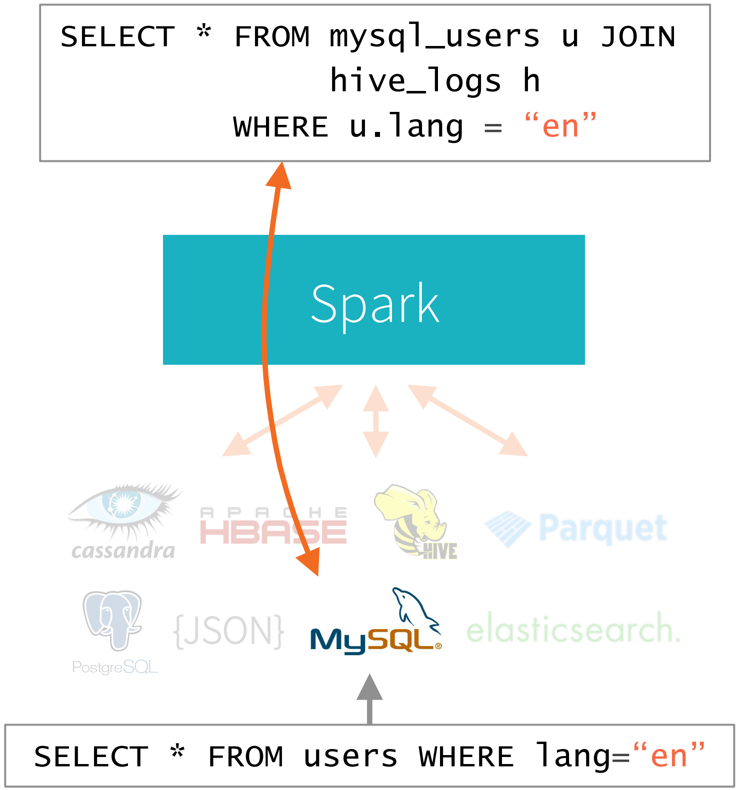


External Data Sources

Platform API to plug smart data sources into Spark

Returns DataFrames usable in Spark apps or SQL

Pushes logic into sources



databricks

```scala
/**
 * ::DeveloperApi::
 * A BaseRelation that can eliminate unneeded columns and filter using selected
 * predicates before producing an RDD containing all matching tuples as Row objects.
 *
 * The actual filter should be the conjunction of all `filters`,
 * i.e. they should be "and" together.
 *
 * The pushed down filters are currently purely an optimization as they will all be evaluated
 * again.  This means it is safe to use them with methods that produce false positives such
 * as filtering partitions based on a bloom filter.
 */
@DeveloperApi
trait PrunedFilteredScan {
  def buildScan(requiredColumns: Array[String], filters: Array[Filter]): RDD[Row]
}
```

Existing Data Sources

built-in

Parquet

JDBC

{ JSON }

PostgreSQL

HIVE

MySQL

HDFS

amazon web services S3

H2

external

AVRO

CSV

dBase

APACHE HBASE

elasticsearch.

cassandra

amazon web services
Amazon Redshift

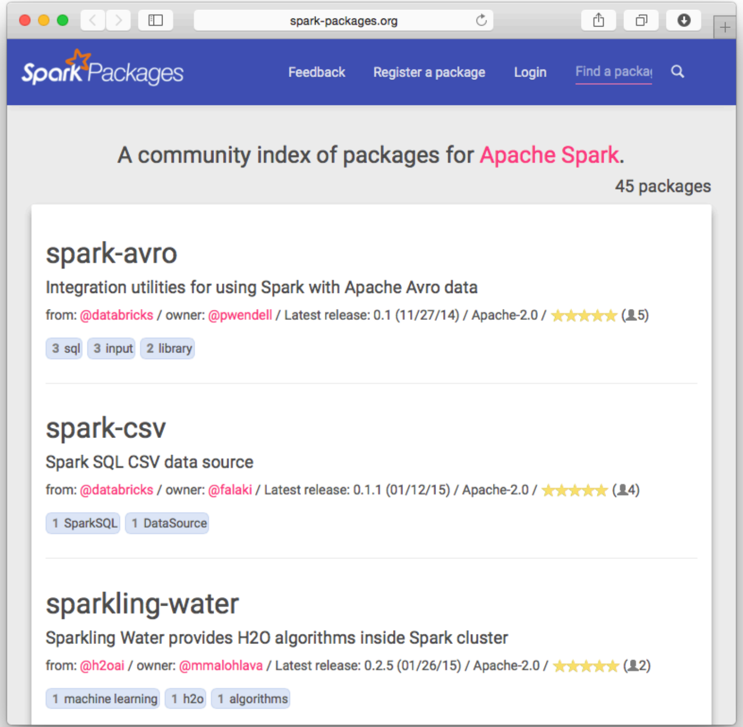
and more …

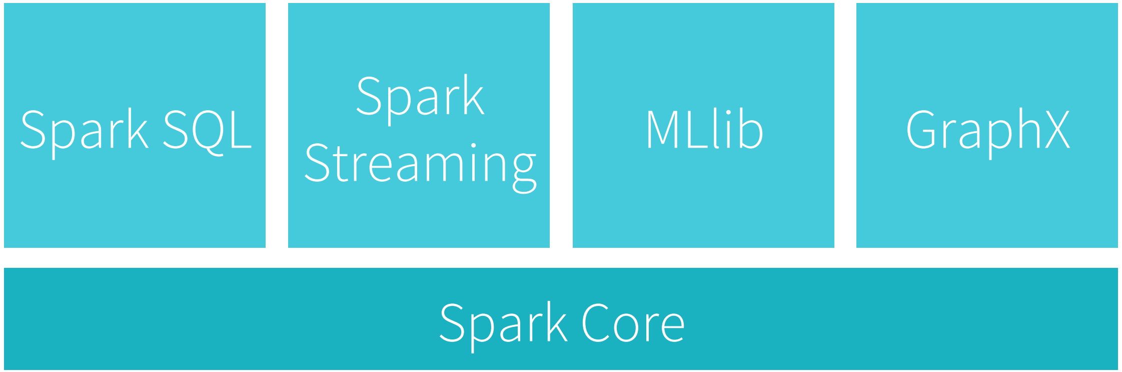
Spark Packages

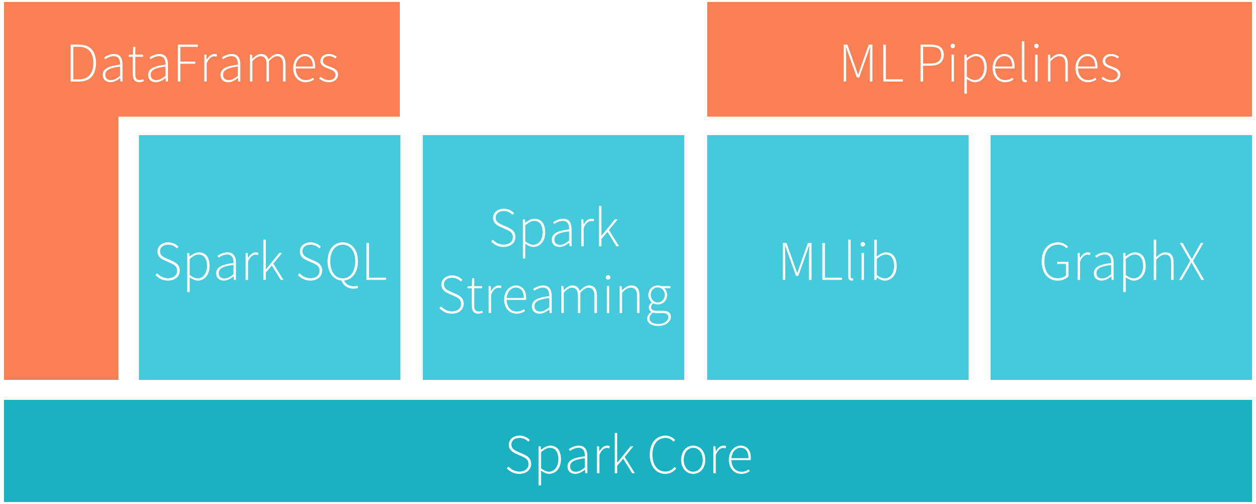
Community index of third party packages

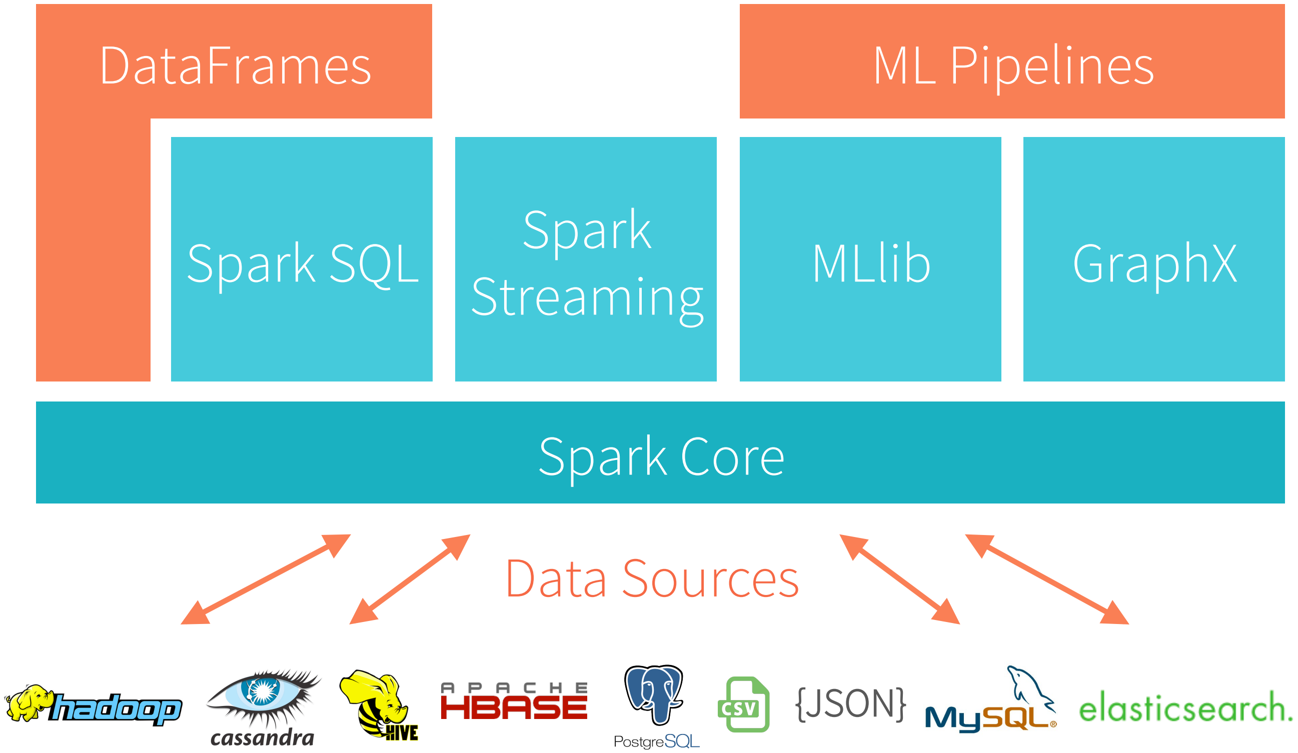
bin/spark-shell --packages databricks/spark-csv:0.2

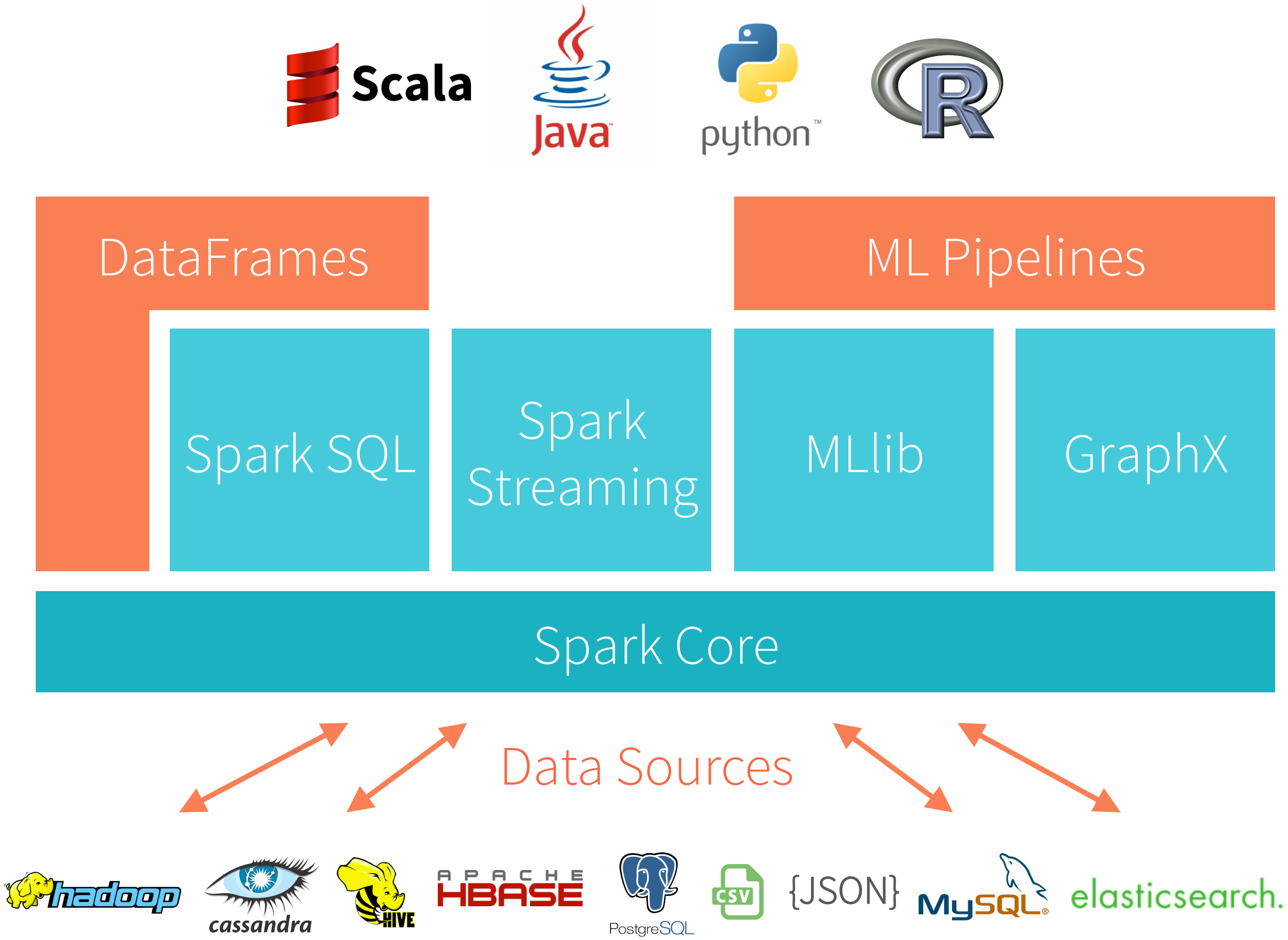
spark-packages.org




Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core


DataFrames
ML Pipelines
Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core


DataFrames
ML Pipelines
Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core
Data Sources
hadoop
cassandra
HIVE
APACHE HBASE
PostgreSQL
CSV
{JSON}
MySQL
elasticsearch.



Scala
Java
python
R
DataFrames
ML Pipelines
Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core
Data Sources
hadoop
cassandra
HIVE
APACHE HBASE
PostgreSQL
CSV
{JSON}
MySQL
elasticsearch.

Goal: unified engine across data sources, workloads and environments

Spark Summit 2015

SAN FRANCISCO | JUNE 15-17 2015

REGISTER NOW >

REGISTER | AGENDA | SPARK TRAINING | VENUE | JOB BOARD

DAY 1 - SPARK SUMMIT 2015

TIME	06/15/2015		
07:00 AM - 09:00 AM	Registration		
09:00 AM - 12:00 PM	Keynotes		
12:00 PM - 12:00 PM	Lunch		
01:00 PM - 01:30 PM	TRACK A - Developer Beyond SQL: Spark SQL Abstractions For The Common Spark Job Michael Armbrust (Databricks) 30 min	TRACK B - Data Science Large-scale Lasso and Elastic-Net Regularized Generalized Linear Models DB Tsai (Alpine Data Labs) 30 min	TRACK C - Applications Use of Spark MLlib for Predicting the Offlining of Digital Media Christopher Burdorf (NBC Universal) 30 min
	TRACK A - Developer Spark-on-YARN: The	TRACK B - Data Science Hybrid Community Detection for Web-scale	TRACK C - Applications Spark and Spark Streaming at Netflix

SPONSORS

PLATINUM

databricks

intel

IBM

GOLD

ZOOMDATA

platfora

SILVER

IMPETUS

sharethrough


sexualization protest
SPARK
action
resistance
knowledge
SPARK Movement
Blog
The SPARKteam
Staff
Research Blog
SPARKit!
Join SPARK
For Educators
SPARK

Enjoy Spark Forum at ApacheCon!