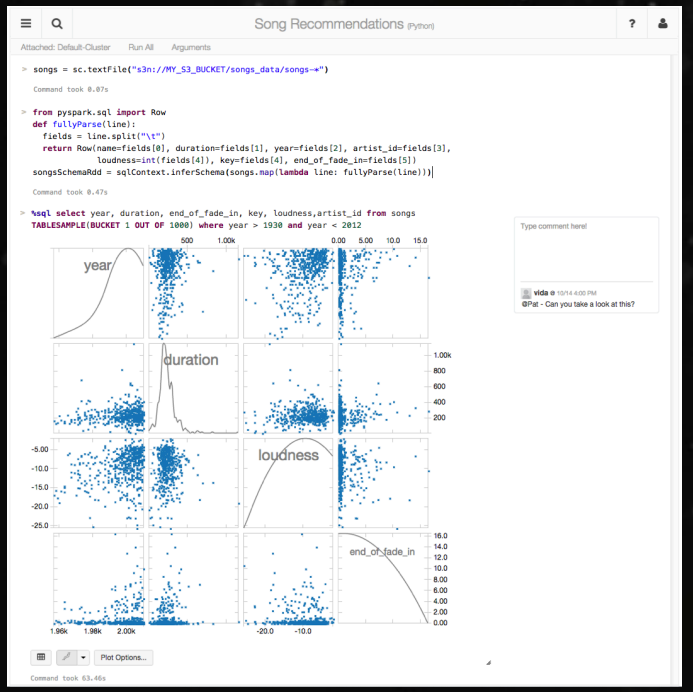
SPARK DEVELOPMENT BOOTCAMP

September 2015 Strata + Hadoop World NYC

making big data simple

- Founded in late 2013
- by the creators of Apache Spark
- Original team from UC Berkeley AMPLab
- Raised $47 Million in 2 rounds
- ~60 employees
- We're hiring! (http://databricks.workable.com)
- Level 2/3 support partnerships with
  - Hortonworks
  - MapR
  - DataStax



Databricks Cloud:
"A unified platform for building Big Data pipelines – from ETL to Exploration and Dashboards, to Advanced Analytics and Data Products."

Databricks has contributed more than **75%** of the code added to Spark in the past year

AGENDA

Day 1

- Welcome
- Login + Introductions
- Just enough Scala for Spark 
- DataFrames 
- RDD Fundamentals
- RDD lab 

Day 2

- Review
- Stages
- Accumulators & Broadcast Variables 
- Spark Runtime Architecture (w/ YARN)
- Memory & Persistence
- Transformations & Actions 
- Spark UI

Day 3

- Review
- Spark Streaming 
- MLlib
- GraphX 
- Config, tunings, settings
- Putting it all together 
- Certification Prep.

INSTRUCTOR: LAURENT WEICHBERGER

Homepage: www.ompoint.com
LinkedIn: www.linkedin.com/profile/view?id=2218412
@meherfalcon

- 25+ years experience building & maintaining software systems
- Hadoop (at Hortonworks, and Cloudera), Cassandra (at DataStax), Couchbase (at Couchbase), Java & C# (at JP Morgan Chase, Pacific Controls, etc.) now Scala and much more …
- Spark instructor for Databricks

100TB Daytona Sort Competition 2014

databricks

	Hadoop MR Record	Spark Record	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400 physical	6592 virtualized	6080 virtualized
Cluster disk throughput	3150 GB/s (est.)	618 GB/s	570 GB/s
Sort Benchmark Daytona Rules	Yes	Yes	No
Network	dedicated data center, 10Gbps	virtualized (EC2) 10Gbps network	virtualized (EC2) 10Gbps network
Sort rate	**1.42 TB/min**	**4.27 TB/min**	**4.27 TB/min**
Sort rate/node	**0.67 GB/min**	**20.7 GB/min**	**22.5 GB/min**

Spark sorted the same data **3X faster** using **10X fewer machines** than Hadoop MapReduce in 2013.

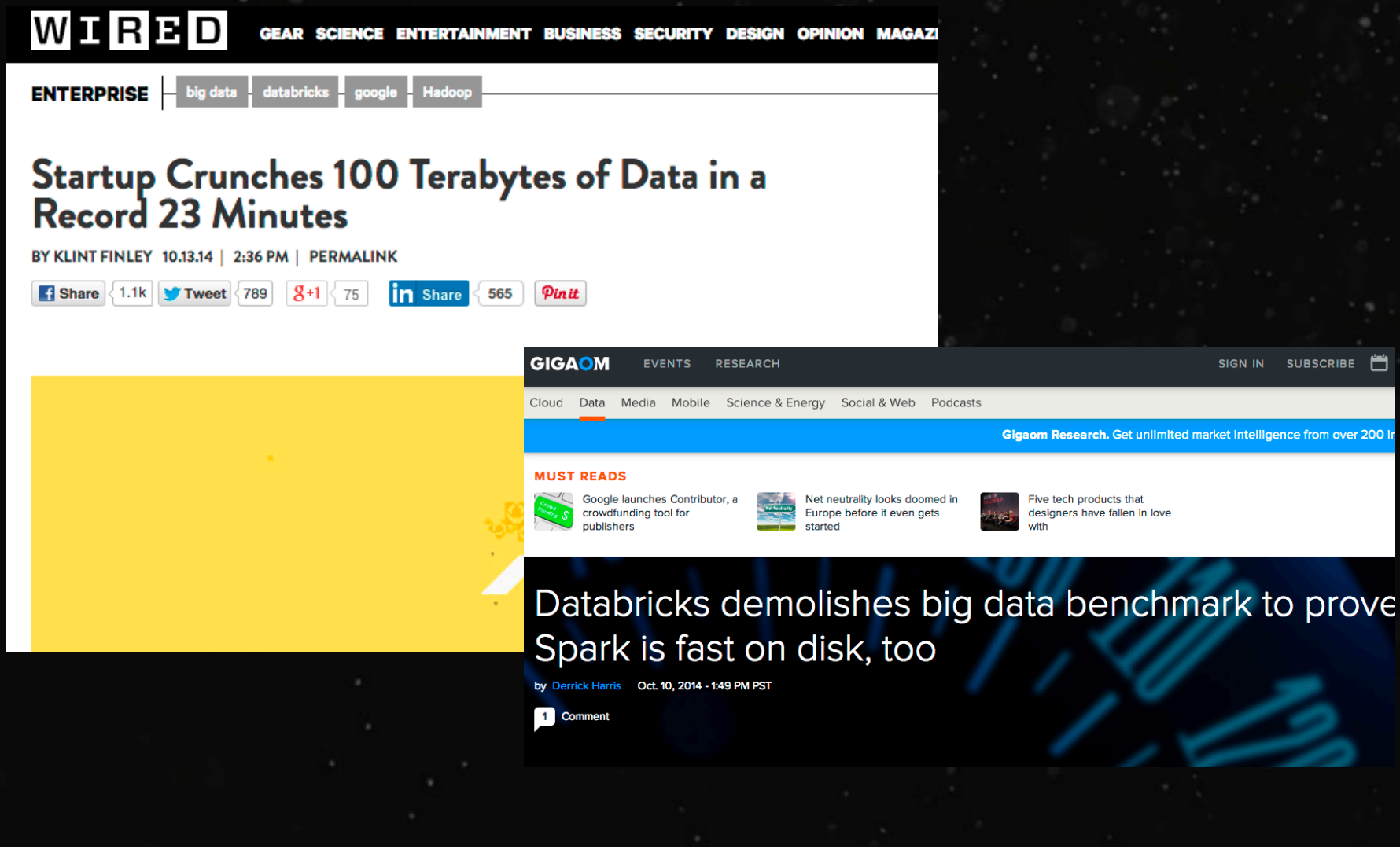
All the sorting took place on disk (HDFS) without using Spark's in-memory cache!

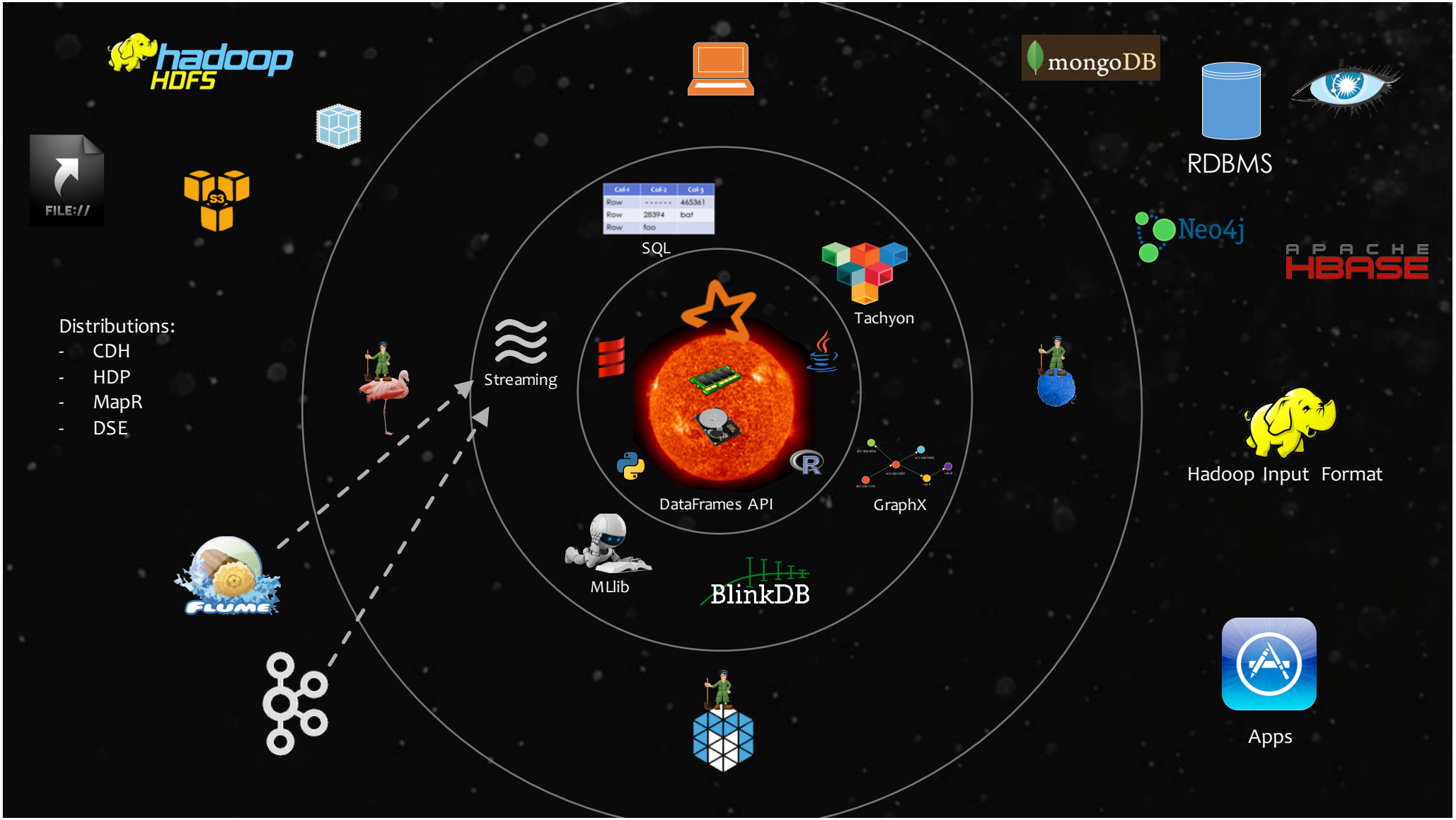
More info:

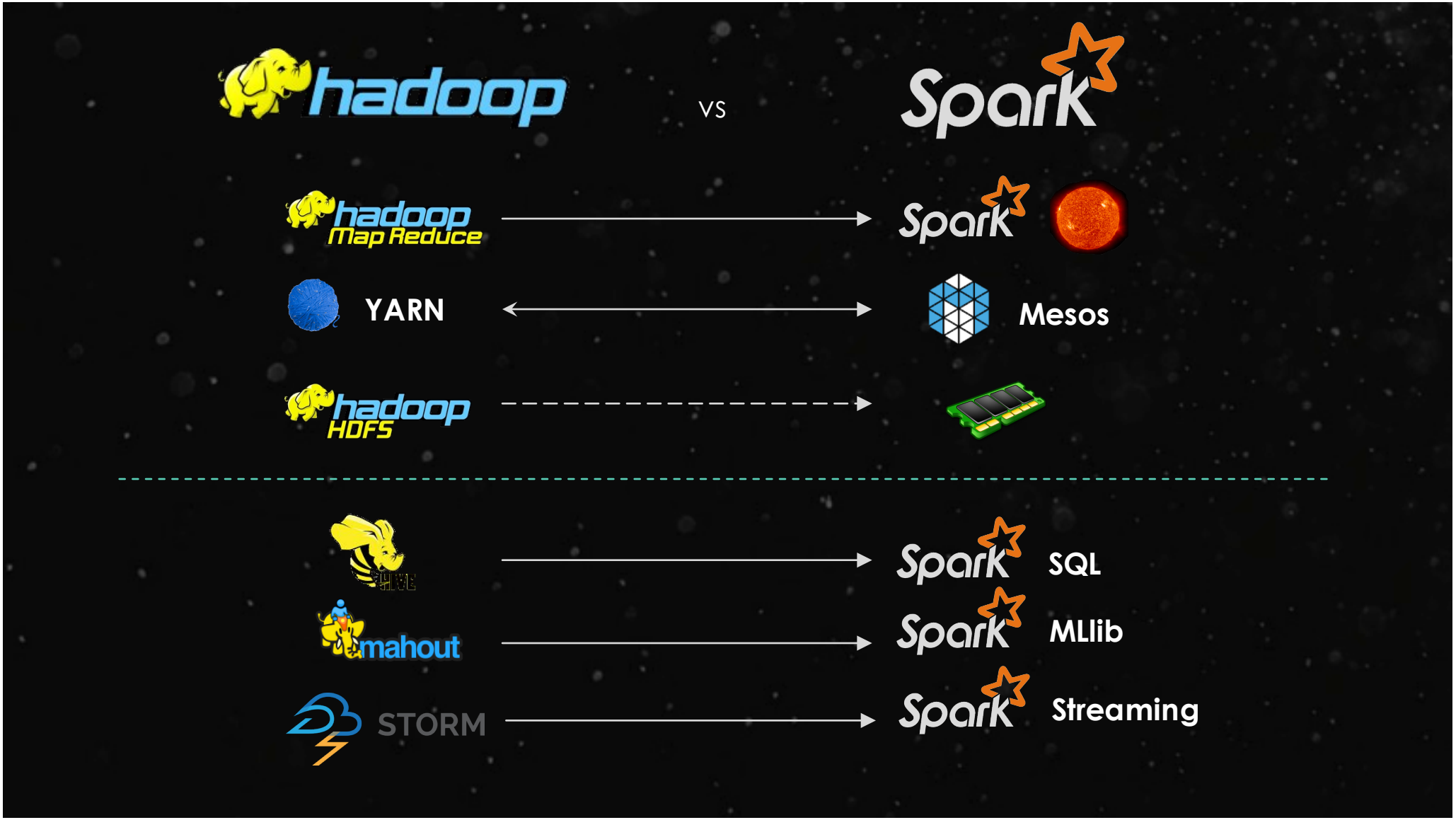
http://sortbenchmark.org

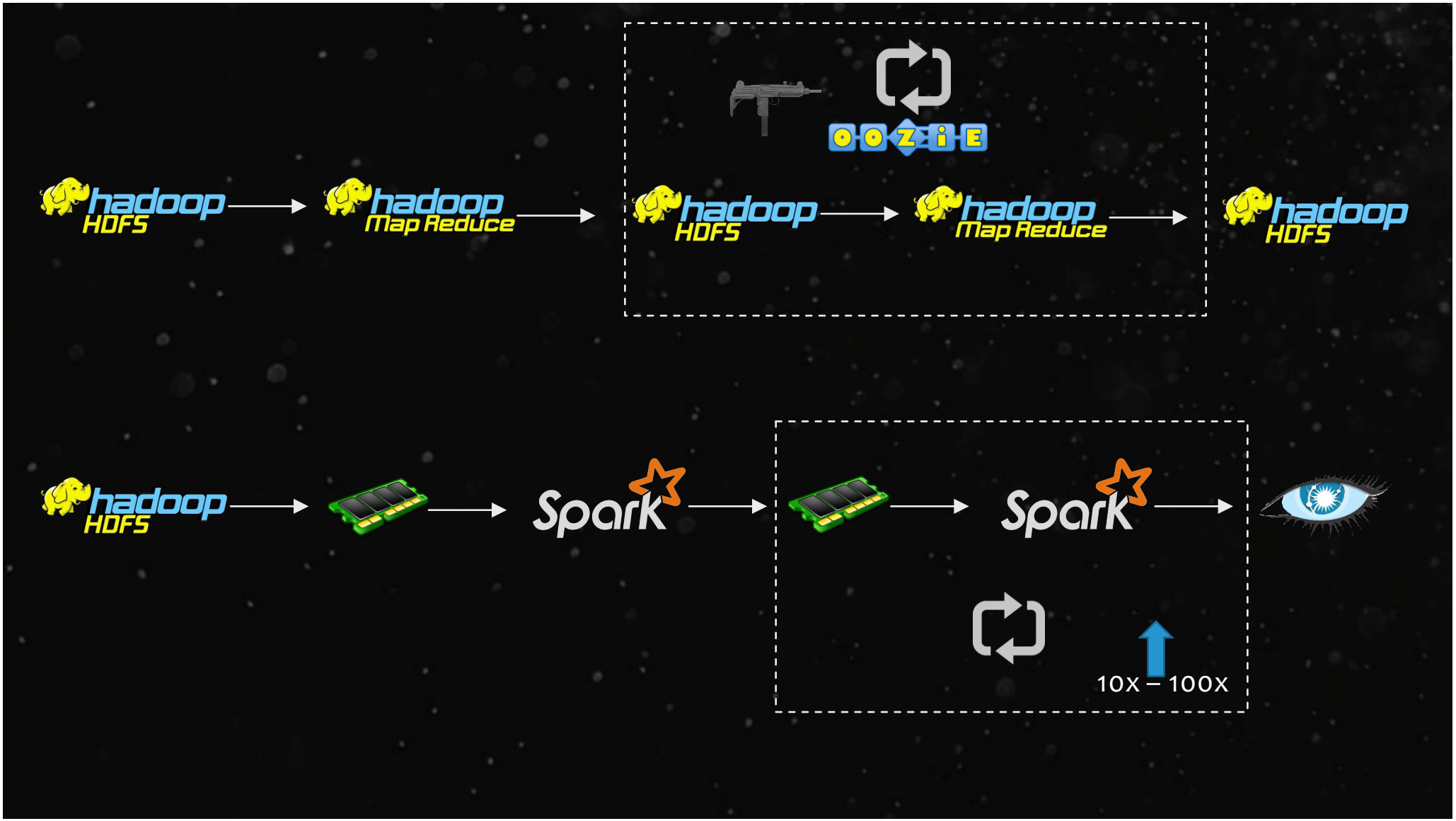
http://databricks.com/blog/2014/11/05/spark-officially-sets-a-new-record-in-large-scale-sorting.html

Work by Databricks engineers: Reynold Xin, Parviz Deyhim, Xiangrui Meng, Ali Ghodsi, Matei Zaharia

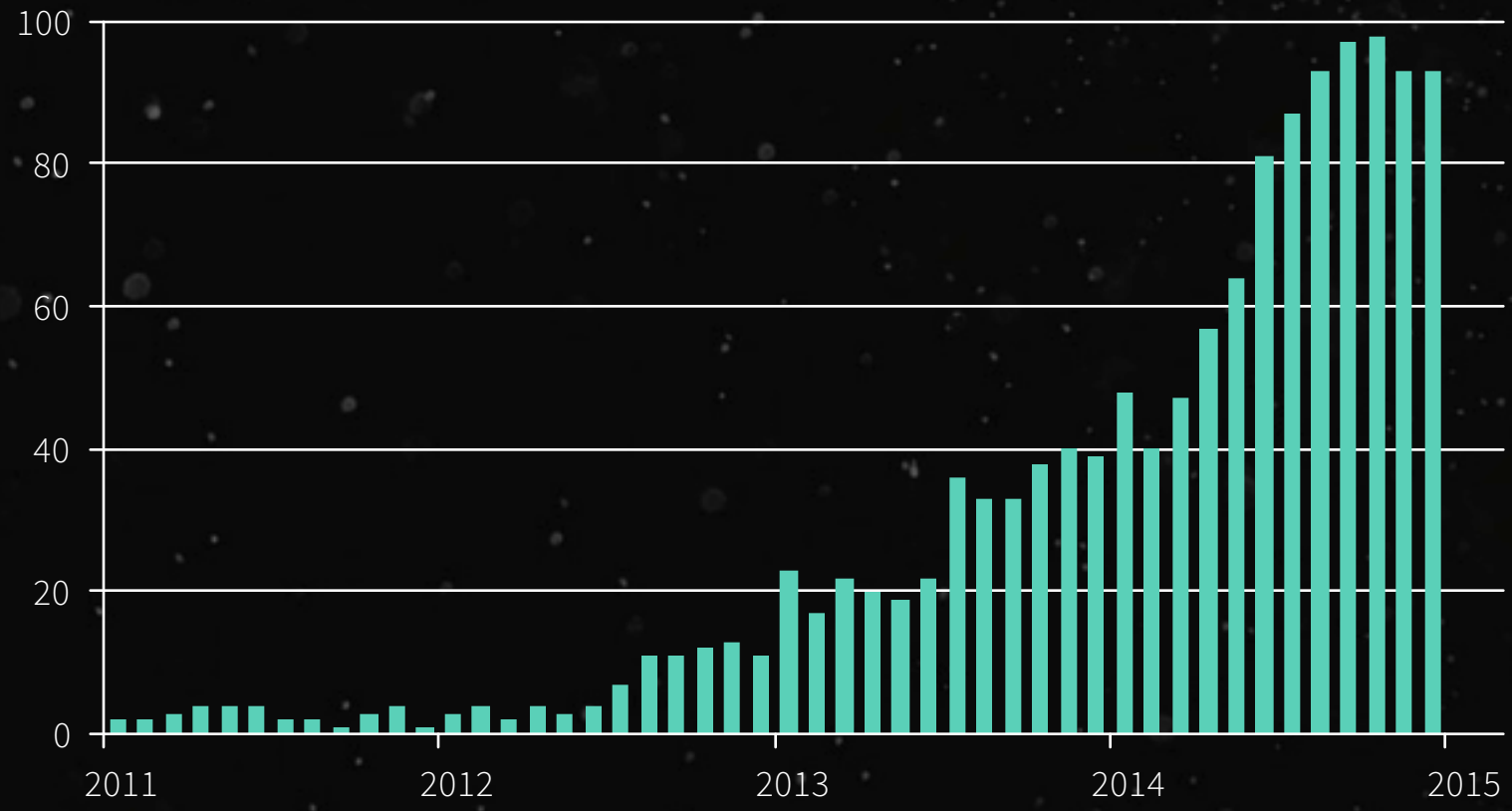

WIRED
GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION MAGAZ
ENTERPRISE
big data
databricks
google
Hadoop
Startup Crunches 100 Terabytes of Data in a Record 23 Minutes
BY KLINT FINLEY 10.13.14 | 2:36 PM | PERMALINK
Share 1.1k
Tweet 789
+1 75
Share 565
Pin it
GIGAOM
EVENTS
RESEARCH
SIGN IN
SUBSCRIBE
Cloud
Data
Media
Mobile
Science & Energy
Social & Web
Podcasts
Gigaom Research. Get unlimited market intelligence from over 200 ir
MUST READS
Google launches Contributor, a crowdfunding tool for publishers
Net neutrality looks doomed in Europe before it even gets started
Five tech products that designers have fallen in love with
Databricks demolishes big data benchmark to prove Spark is fast on disk, too
by Derrick Harris Oct. 10, 2014 - 1:49 PM PST
1 Comment


hadoop
HDFS
FILE://
S3
mongoDB
RDBMS
Neo4j
APACHE
HBASE
Distributions:
- CDH
- HDP
- MapR
- DSE
SQL
Tachyon
Streaming
DataFrames API
GraphX
MLlib
BlinkDB
FLUME
Hadoop Input Format
Apps


hadoop
vs
Spark
hadoop Map Reduce
Spark
YARN
Mesos
hadoop HDFS
SQL
mahout
MLlib
STORM
Streaming


hadoop
HDFS
hadoop
Map Reduce
OOZIE
hadoop
HDFS
hadoop
Map Reduce
hadoop
HDFS
hadoop
HDFS
Spark
Spark
10x – 100x

Contributors per month to Spark



COMMUNITY STATS

- Most active Apache project in contributors per month
- Most active open source project in a functional language

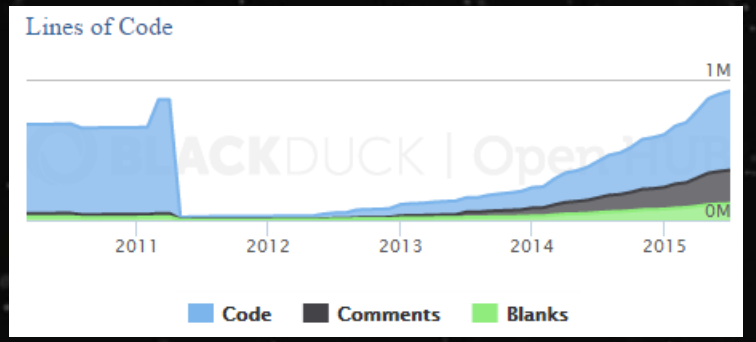
In a Nutshell, Apache Spark...

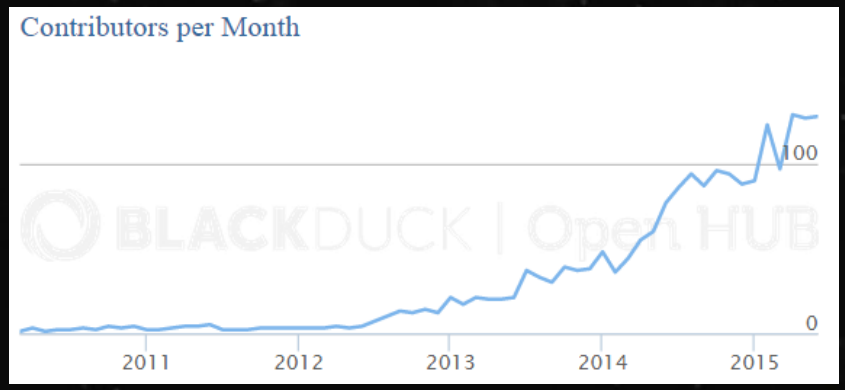
... has had 23,706 commits made by 797 contributors
representing 563,770 lines of code

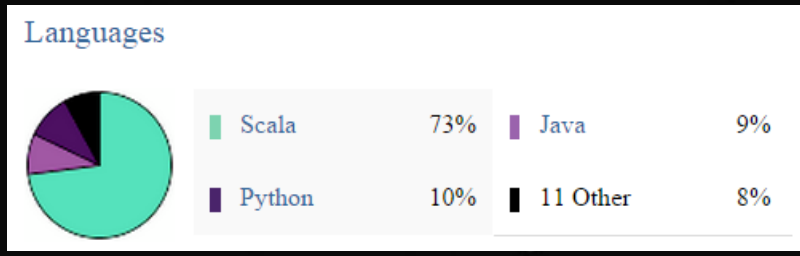
... is mostly written in Scala
with a well-commented source code

... has a well established, mature codebase
maintained by a very large development team
with stable Y-O-Y commits

... took an estimated 154 years of effort (COCOMO model)
starting with its first commit in ~~March, 2010~~ **Aug 2009**
ending with its most recent commit 1 day ago







...in June 2013

Source: openhub.net

INTRODUCTIONS

1. Your Name?
2. Last programming language you used?
3. When you wrote that code?

4. Go ahead and login:
https://classeast04.cloud.databricks.com/