SPARK DEVELOPMENT BOOTCAMP

September 2015 Strata + Hadoop World NYC

databricks
Databricks:

- 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](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
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


Work by Databricks engineers: Reynold Xin, Parviz Deyhim, Xiangrui Meng, Ali Ghodsi, Matei Zaharia
Startup Crunches 100 Terabytes of Data in a Record 23 Minutes

BY KUNT FINLEY 10.13.14 | 2:36 PM | PERMALINK

Databricks demolishes big data benchmark to prove Spark is fast on disk, too

by Derrick Harris  Oct. 10, 2014 - 149 PM PST
Distributions:
- CDH
- HDP
- MapR
- DSE

- SQL
- Streaming
- Tachyon
- GraphX
- MLib
- MLlib
- BlinkDB

Apps

RDBMS
Hadoop Input Format
mongoDB
Neo4j
HBase
Apache

Distributions:
- CDH
- HDP
- MapR
- DSE
Hadoop vs Spark

- **Hadoop**
  - Map Reduce
  - HDFS

- **Spark**
  - SQL
  - MLlib
  - Streaming

- **Mesos**
  - YARN
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

Languages

- Scala 73%
- Java 9%
- Python 10%
- 11 Other 8%

...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/