Apache Spark & Big Data Course

ITC - Demi Ben-Ari

About Me

Demi Ben-Ari, Co-Founder & VP R&D **@ Panorays**

- BS'c Computer Science – Academic College Tel-Aviv Yaffo
- **Co-Founder @**
  - **"Big Things" Big Data Community**
  - **Cloud Google Developer Group**

In the Past:

- Sr. Data Engineer @ Windward
- Software Team Leader & Senior Java Software Engineer
- Missile defense and Alert System - **"Ofek" – IAF**

Course Agenda - Syllabus

1) Introduction to Big Data - Ecosystem.

2) Install Spark & Deployments (About spark's infrastructure too).

3) Spark Core

4) Spark Development Process

5) Running your own Spark cluster on local VM's

Course Agenda - Syllabus

6) Spark High Level API's - Introduction (SparkSQL, SparkStreaming, Spark Notebooks).

7) Spark real world use case.

8) Monitoring Big Data Systems

8) Spark Best Practices.

9) Open Talk, Questions, other things that you'd like to ask me.

Course Agenda - Syllabus

- Around 3 hours per half day (~ 09:00 - 12:30).
  - Really depends on the amount of questions.
- Every ~45-50 minutes we'll do a short break.
- Feel free to ask question during the lectures.
  - If I don't know the answer I promise to find an answer later. (Or I'll just bluntly lie and smile :))
- Every student will know about the infrastructure of Apache Spark and about the development process of a real world application.

Course Agenda - Syllabus

- Please, I urge you! Interrupt me during the lectures, Ask Questions!
  - Together, we’ll learn how to be better Israeli People :)
- Be aware,
  If you don’t stop me...I’m just gonna keep talking!

Spark Introduction

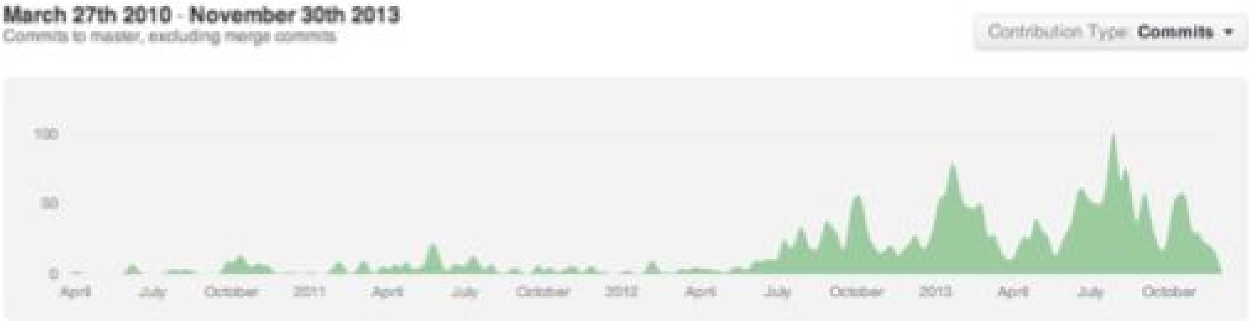
What is spark?

- □Apache Spark is a general-purpose, cluster computing framework
- Spark does computation In Memory & on Disk
- Apache Spark has low level and high level APIs

Spark Philosophy

- □Make life easy and productive for data scientists
- Well documented, expressive API’s
- Powerful domain specific libraries
- Easy integration with storage systems… and caching to avoid data movement
- Predictable releases, stable API’s
- Stable release each 3 months

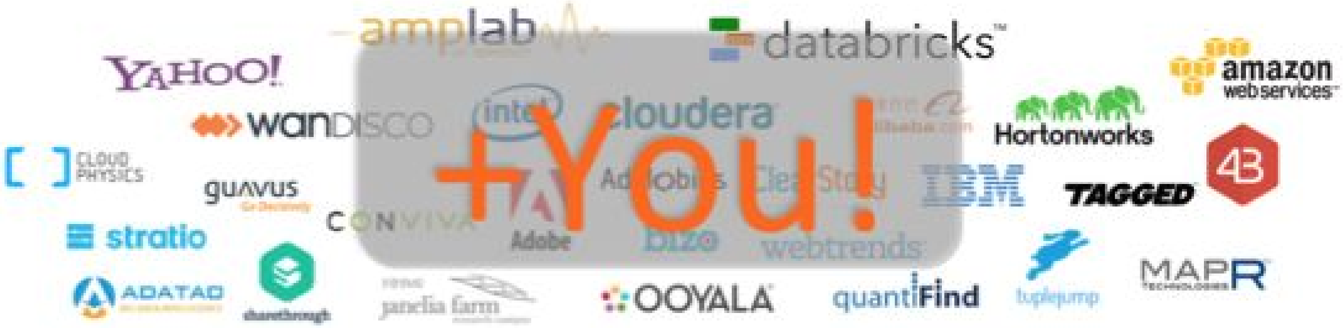
The Spark Community



March 27th 2010 - November 30th 2013

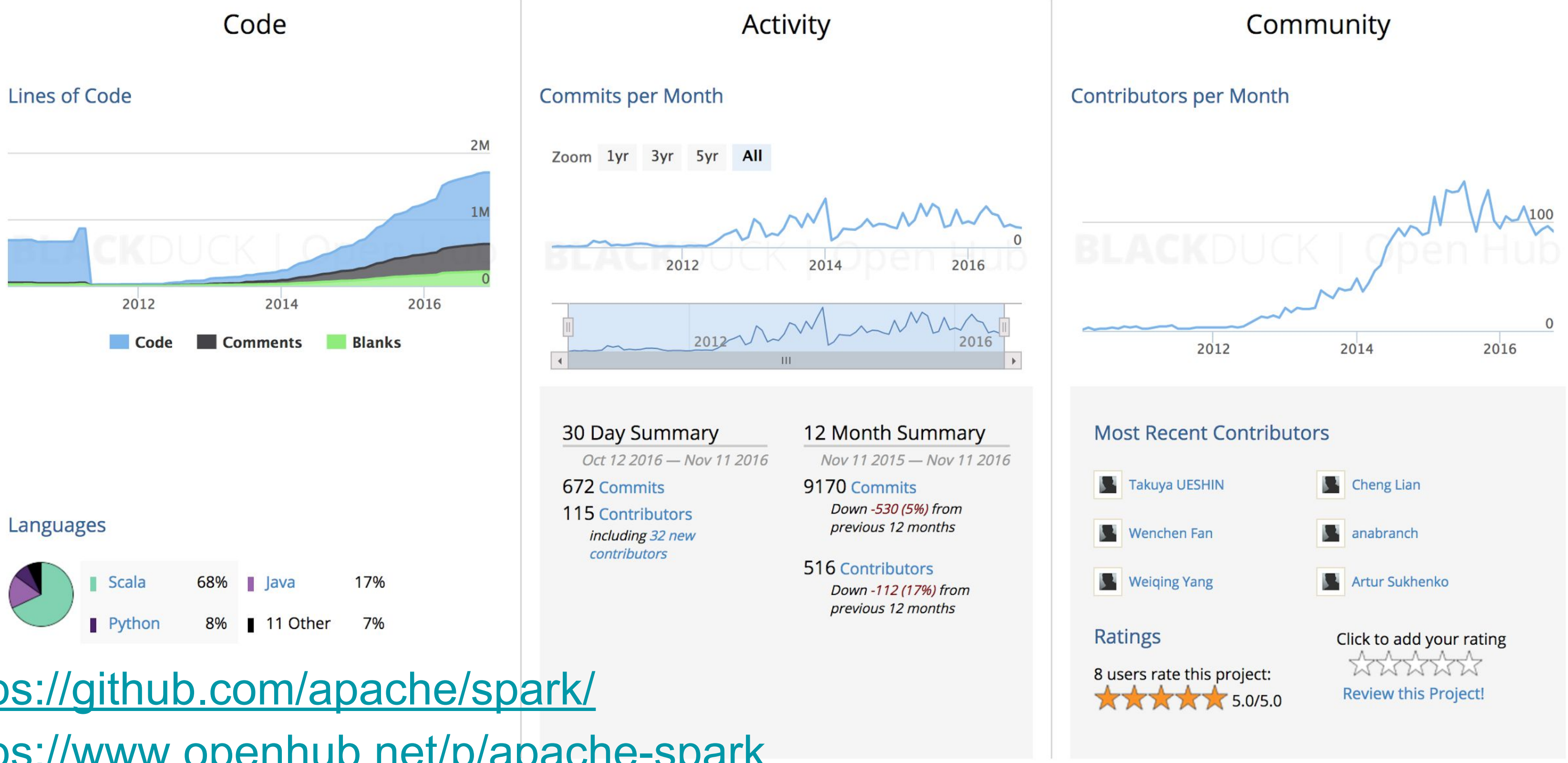
Commits to master, excluding merge commits

Contribution Type: Commits



+You!

Spark Contributors



Code

Lines of Code

Languages

Scala	68%	Java	17%
Python	8%	11 Other	7%

Activity

Commits per Month

30 Day Summary

Oct 12 2016 — Nov 11 2016

672 Commits

115 Contributors

including 32 new contributors

12 Month Summary

Nov 11 2015 — Nov 11 2016

9170 Commits

Down -530 (5%) from previous 12 months

516 Contributors

Down -112 (17%) from previous 12 months

Community

Contributors per Month

Most Recent Contributors

Takuya UESHIN	Cheng Lian
Wenchen Fan	anabranch
Weiqing Yang	Artur Sukhenko

Ratings

8 users rate this project:

★★★★★ 5.0/5.0

Click to add your rating

☆☆☆☆☆

Review this Project!

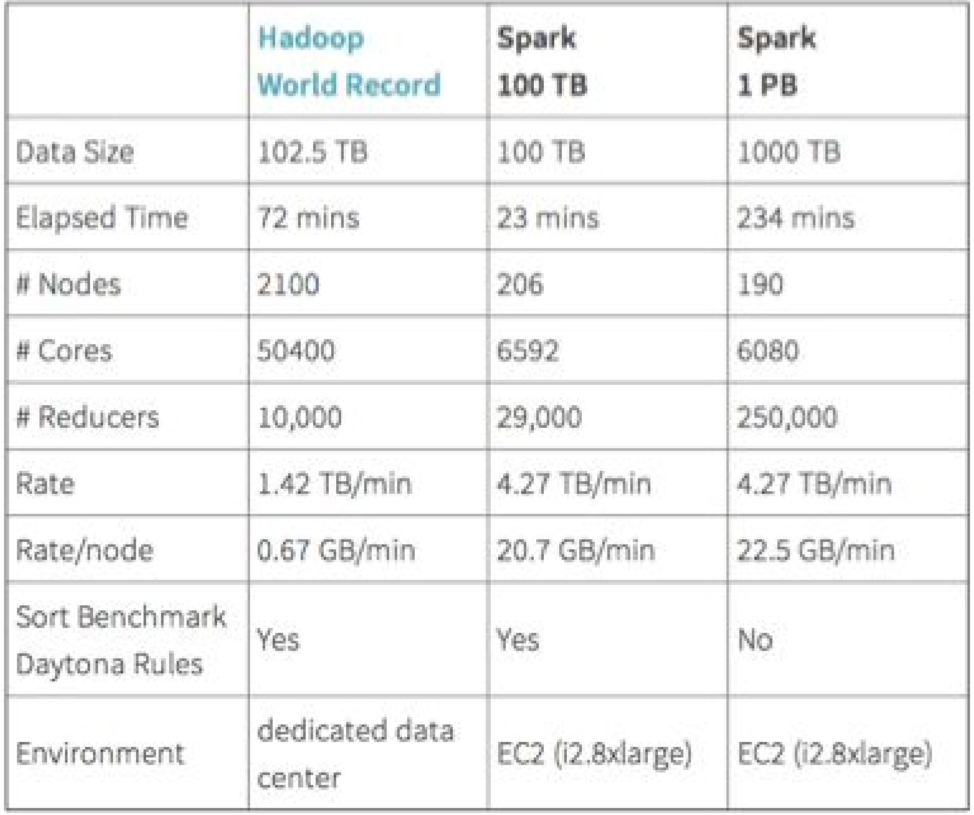
https://github.com/apache/spark/

https://www.openhub.net/p/apache-spark

About Spark Project

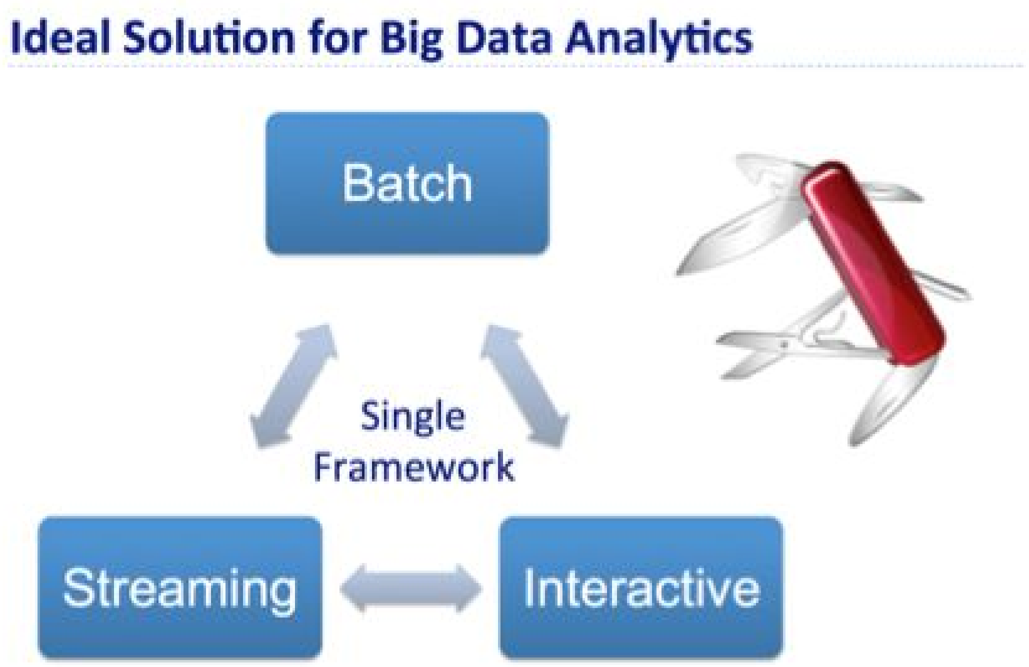
- □Spark was founded at UC Berkeley and the main contributor is “**Databricks**”.
- □Interactive shell Spark in Scala and Python (spark-shell, pyspark)
- Currently stable in version 2.1

Spark Petabyte Sort



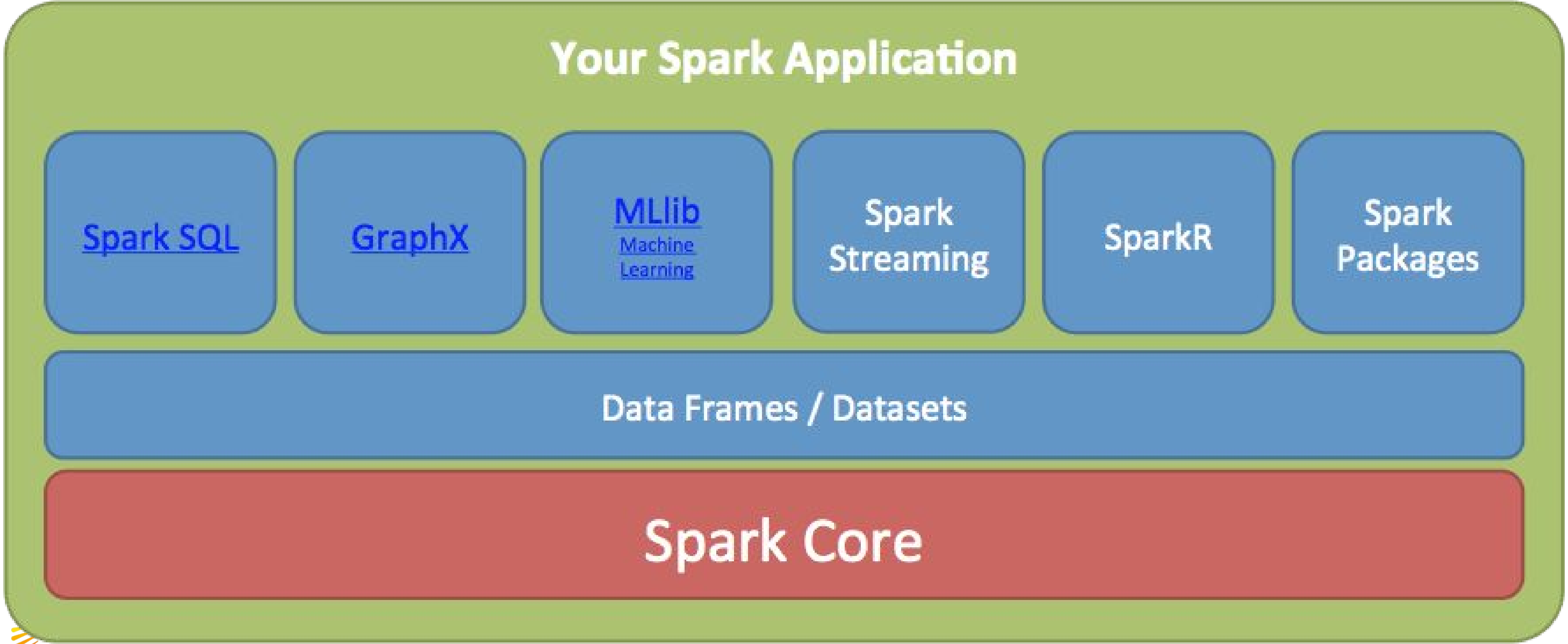
	Hadoop World Record	Spark 100 TB	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400	6592	6080
# Reducers	10,000	29,000	250,000
Rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min
Sort Benchmark Daytona Rules	Yes	Yes	No
Environment	dedicated data center	EC2 (i2.8xlarge)	EC2 (i2.8xlarge)

United Tools Platform

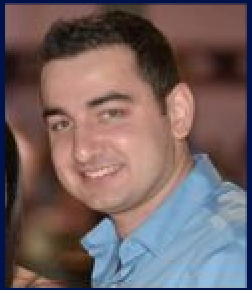


Ideal Solution for Big Data Analytics

Unified Tools Platform



Panorays



- LinkedIn
- Twitter: @demibenari
- Blog: http://progexc.blogspot.com/
- **demi.benari@gmail.com**

- “Big Things” Community

□Meetup, YouTube, Facebook, Twitter

- GDG Cloud

:BIG THINGS

Thank You

Panorays

Good Luck!

Mind the
Attack Surface