APACHE SPARK FOR DATA SCIENTISTS

Session 1: Introduction to Apache Spark

SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Recommended Prerequisite Knowledge in Python Language

- Primitive Data Types in Python
- Flow Control in Python (for, while loops, if...else etc.)
- Standard Functions & Lambda Expressions
- Python Collections and optionally Classes
- Pandas, Numpy, Scipy
- String Operations & Regular Expressions
- Data Visualization in Python (matplotlib, bokeh, plotly etc.)
- Basic Functional Methods in Python (map, filter, reduce, itertools etc.)
- Scikit-learn (good to know it resembles syntactically to Spark MLlib)

https://www.codecademy.com/learn/learn-python

https://lectures.quantecon.org/py/

https://runestone.academy/runestone/static/pythonds/index.html

Session 1

I. Course Overview

II. Spark: Background & Position in Big Data Analytics

III. Core Concepts & Challenges of Distributed Computing

IV. Overview of Spark Components

V. Conceptual Introduction to Spark Application

VI. Appendix

Apache Spark Overview

Apache Spark:

- is a fast and general purpose big data processing engine.
- is an open source project incubated by Apache Software Foundation

https://spark.apache.org/

https://github.com/apache/spark

- is an unified engine with built-in modules for SQL, streaming, machine learning, graph processing & third-party packages.

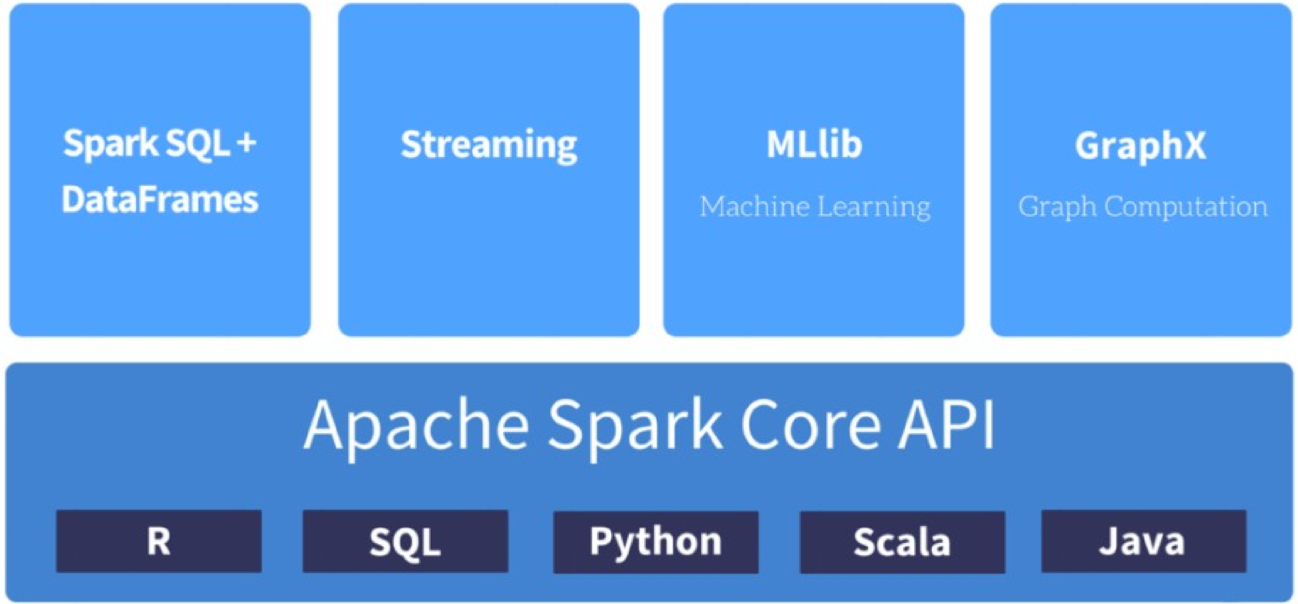


Figure Ref: http://syntelli.com/data-capability/apache-spark/

Common Use Cases of Apache Spark for Big Data

I. Extract-Transform-Load (ETL) operations

II. Predictive analytics and machine learning

III. Data access operations, such as interactive SQL queries and visualizations

IV. Text mining and processing

V. Real-time / Near real-time event processing

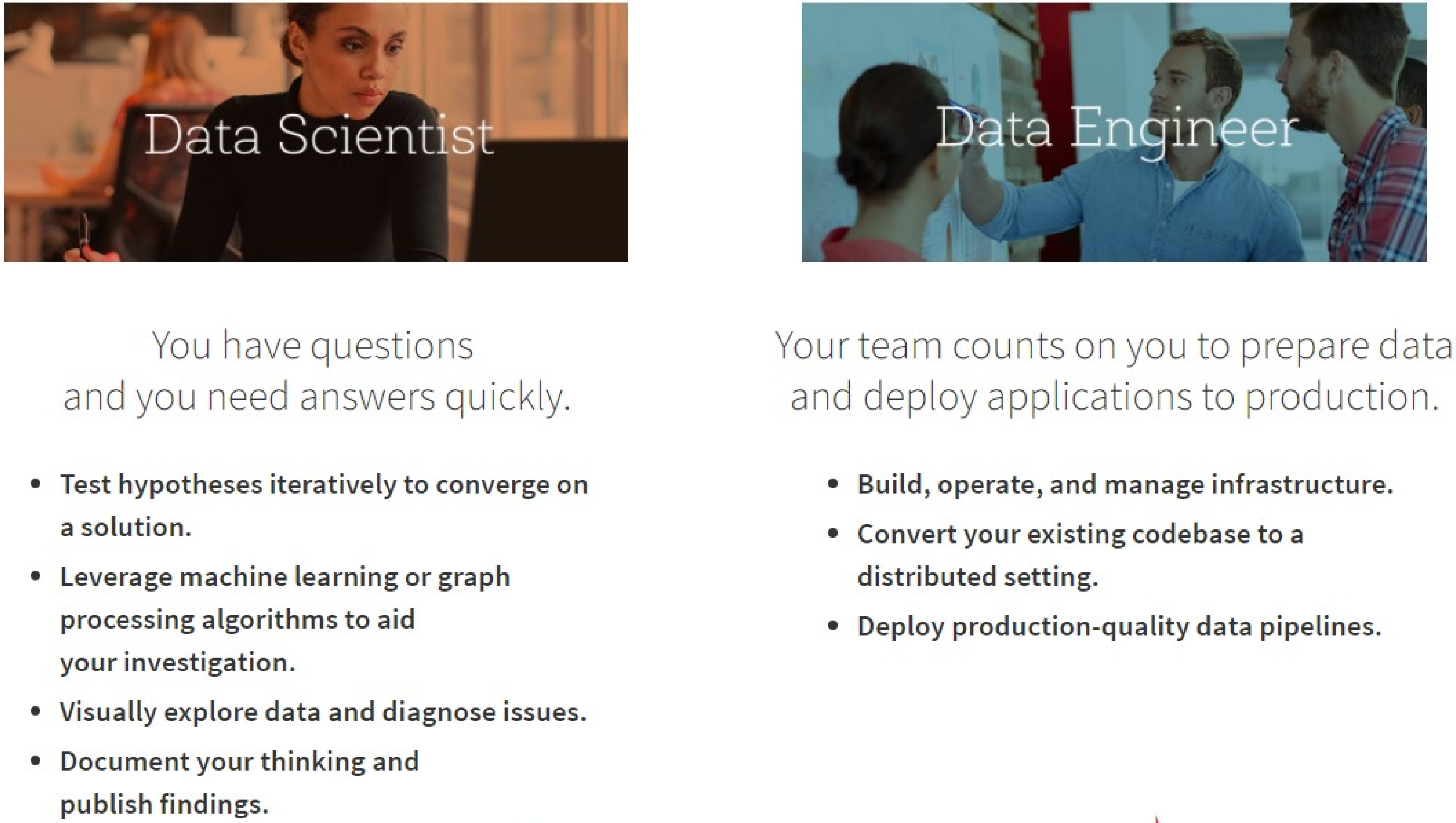
VI. Graph analysis applications

VII. Pattern recognition & Deep Learning

VIII. Recommendation engines

IX. and so on…

Apache Spark for Data Scientists & Data Engineers



Data Scientist

You have questions
and you need answers quickly.

- Test hypotheses iteratively to converge on a solution.
- Leverage machine learning or graph processing algorithms to aid your investigation.
- Visually explore data and diagnose issues.
- Document your thinking and publish findings.

Data Engineer

Your team counts on you to prepare data
and deploy applications to production.

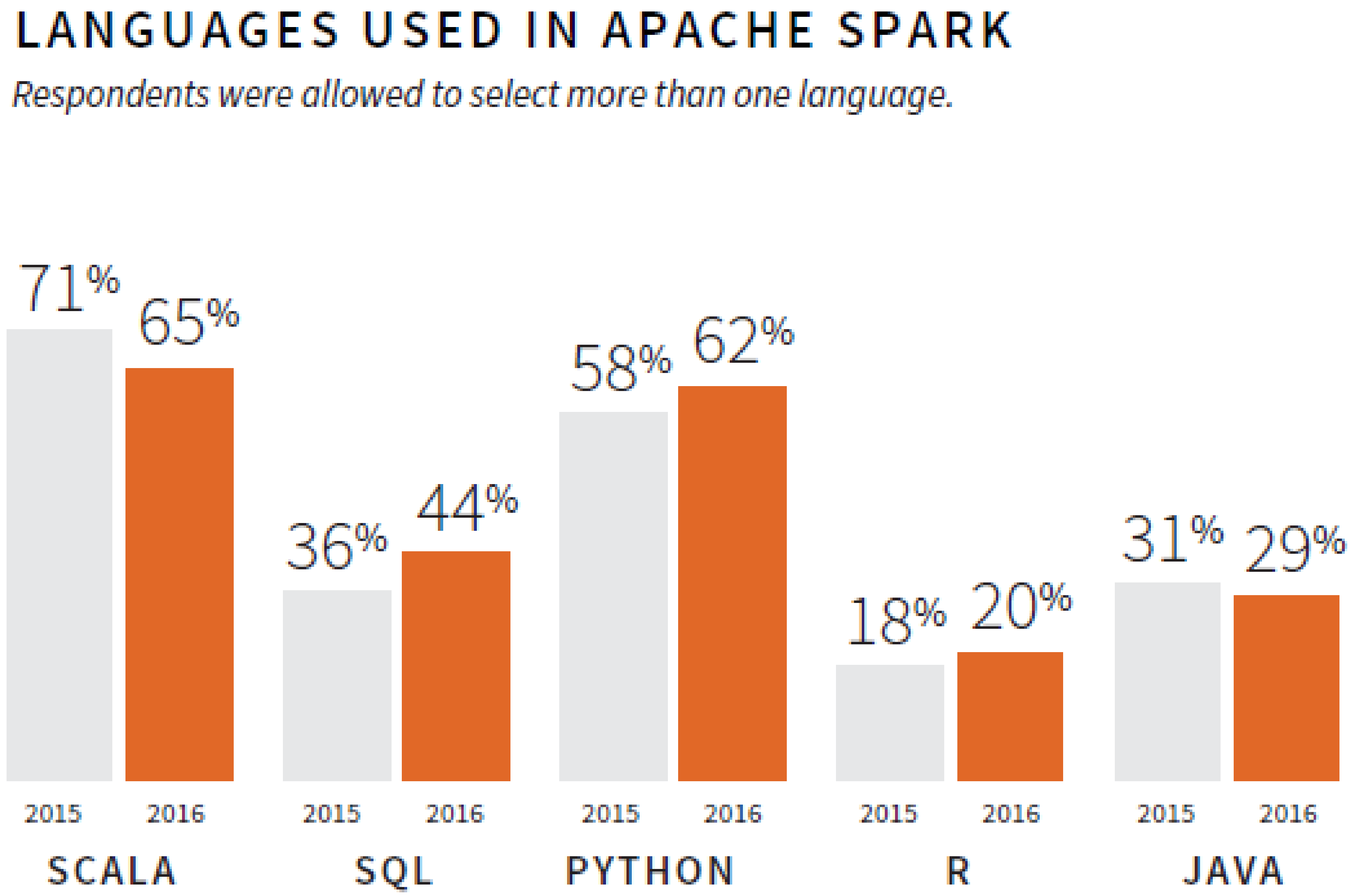
- Build, operate, and manage infrastructure.
- Convert your existing codebase to a distributed setting.
- Deploy production-quality data pipelines.

python™ R Spark SQL

Scala Java python™

Ref: https://databricks.com/solutions/by-role

Spark Programming Languages

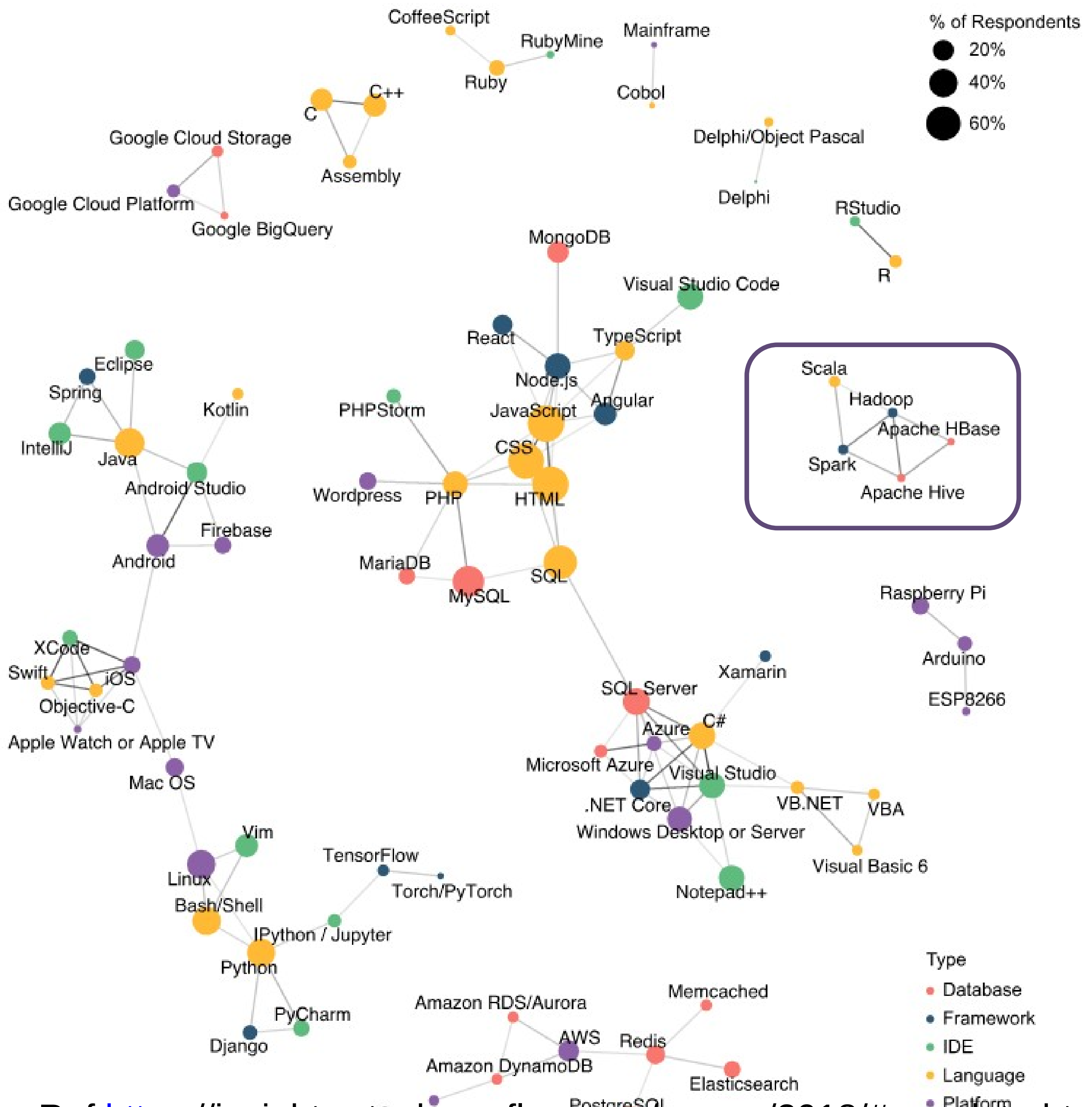


LANGUAGES USED IN APACHE SPARK

Respondents were allowed to select more than one language.

March 2017 Ref:
https://www.slideshare.net/JohanPicard/a-short-introduction-to-spark-and-its-benefits

How Technologies Are Connected



Ref:https://insights.stackoverflow.com/survey/2018/#correlated-technologies

Python vs. Scala for Spark Development (1/2)

python™

Python is an interpreted scripting language and Pyspark API is a set of wrappers on Scala-Spark methods/functions, therefore there is always additional computational overhead.

It involves many widely used data science libraries like numpy, pandas, scipy, pyarrow, matplotlib, tensorflow, pixiedust, numba, spacy etc. which can be mixed-in with Pyspark.

Scala

Apache Spark is built with Scala, therefore it is:

✓ 10x faster than Pyspark

✓ New features are always initially available in Scala API

✓ More robust & extensive integration with other ecosyste tools like Kafka, Akka, Hadoop etc.

Scala & Java has less commonly used data science libraries like breeze, vegas, standford NLP, deeplearning4j etc.

Python vs. Scala for Spark Development (2/2)

python™

It is dynamically typed, which limits type-safe Spark features and hardens debugging.

Suits better to interactive data exploration, model training & evaluation and prototyping projects.

Limited streaming features more relevant to prototyping rather than production applications & there are limitations of Pyspark Connectors to other Big Data Tools.

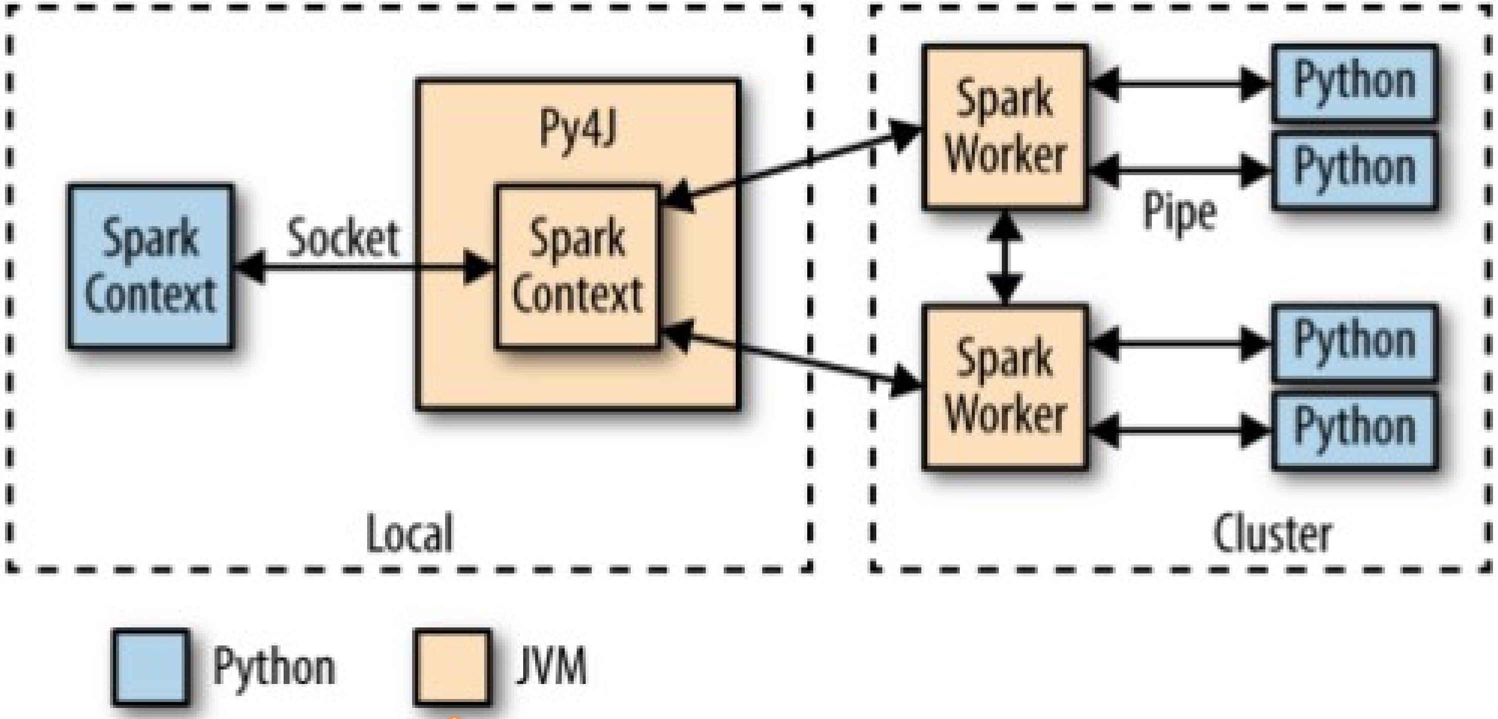
Scala

It is a statically typed & functional, complex functional & type-safe implementations are feasible and majority of bug can be caught in compile time.

Suits better for production (live system) implementations & potentially a more robust option in any real-time analytics

Advanced streaming & data product development capabilities, more variety of Scala-Spark Connectors to other Big Data Tools which are also developed in Scala or Java.

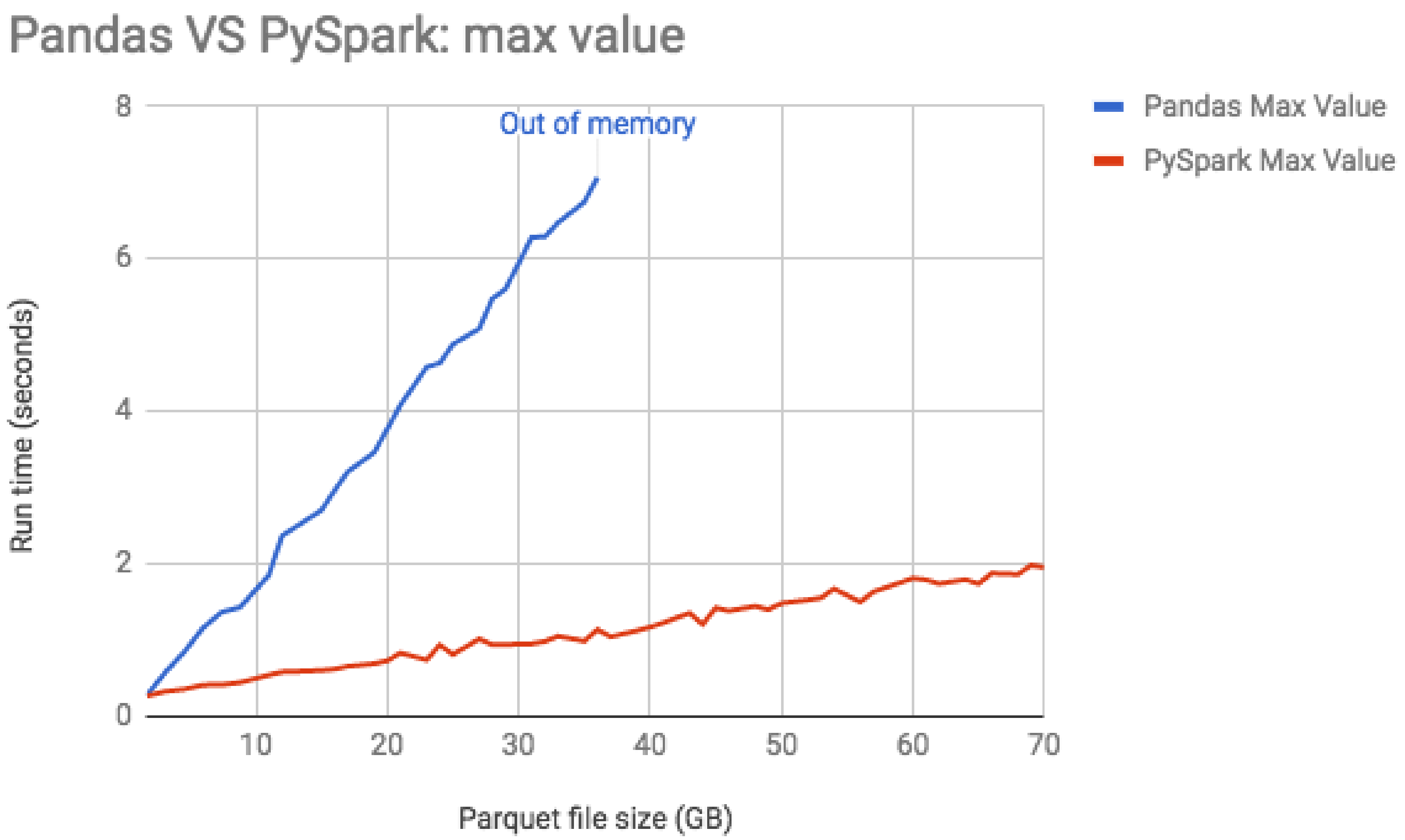
About Pyspark (Python API for Spark)



Spark is internally developed in Scala. Scala Apps run on JVM (Java Virtual Machine) Threads similar to Java. List of JVM Languages

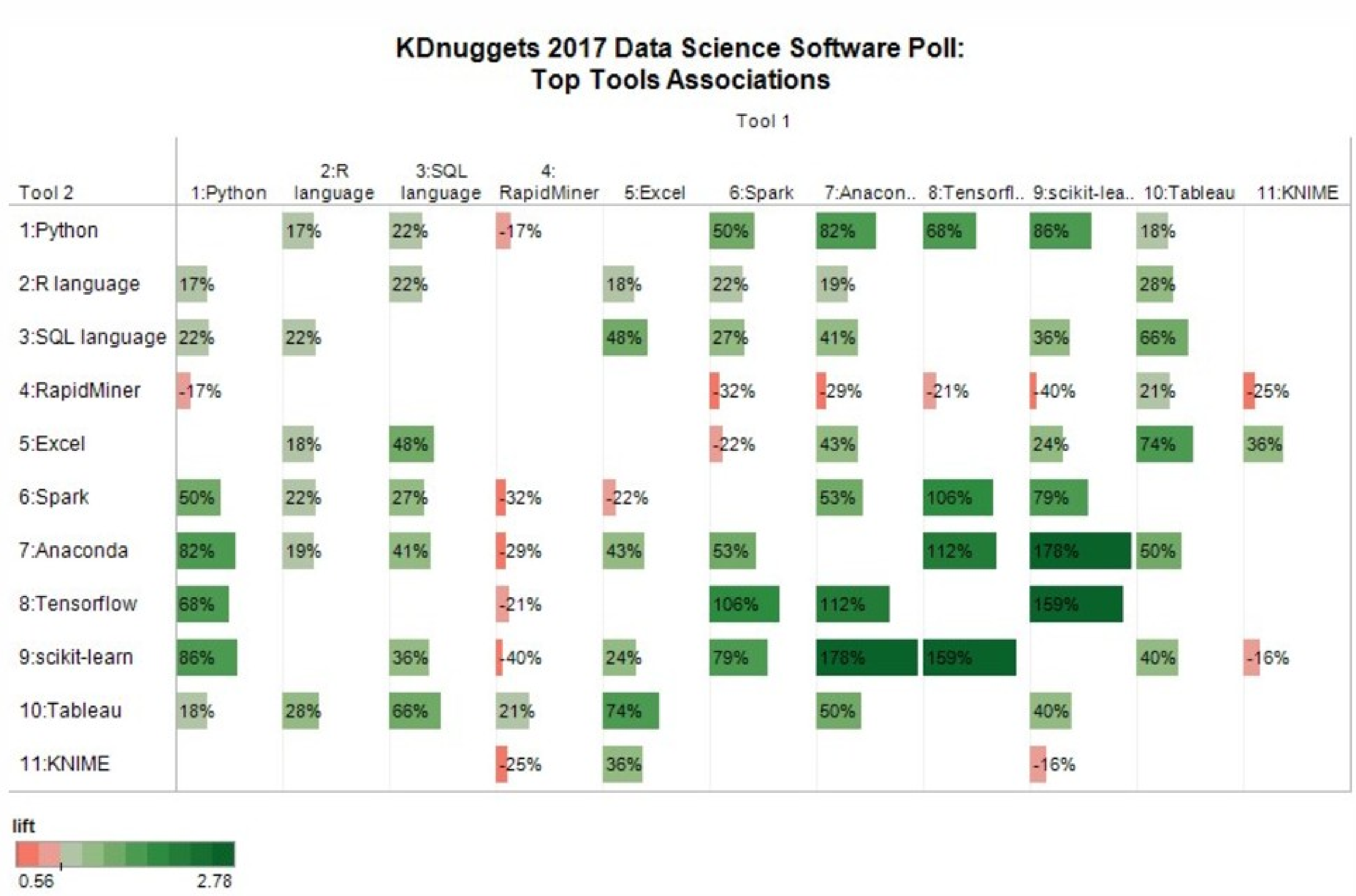
IMPORTANT: Pyspark API Documentation: https://spark.apache.org/docs/latest/api/python/index.htmln

Benchmanking Pandas Vs Pyspark Dataframes on a Single Node Machine



Ref: https://databricks.com/blog/2018/05/03/benchmarking-apache-spark-on-a-single-node-ma

Top Data Science Tools Associations (2017 KDnuggets Survey)

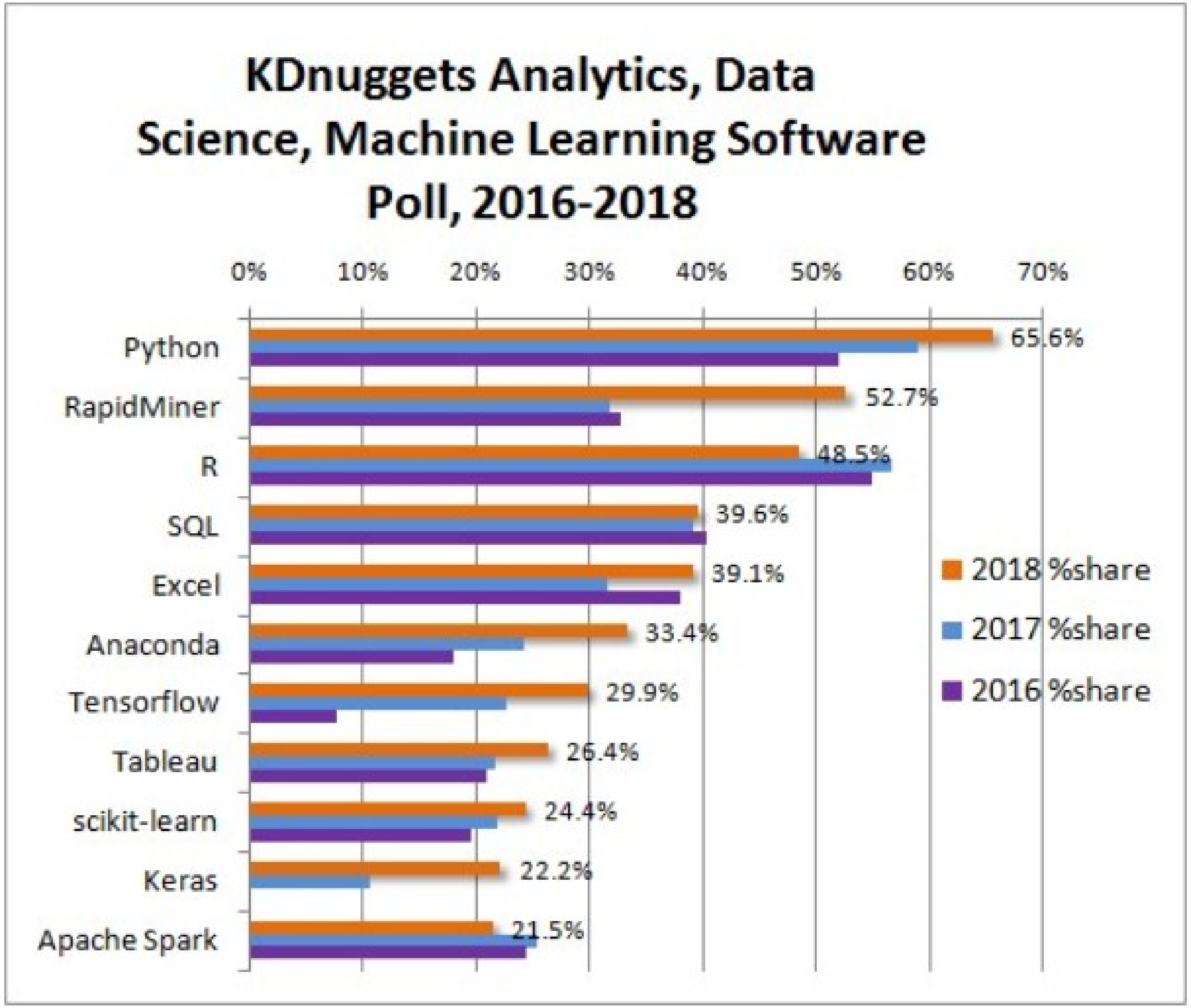


KDnuggets 2017 Data Science Software Poll:
Top Tools Associations

	Tool 1										
Tool 2	1:Python	2:R language	3:SQL language	4: RapidMiner	5:Excel	6:Spark	7:Anacon..	8:Tensorfl..	9:scikit-lea..	10:Tableau	11:KNIME
1:Python		17%	22%	-17%		50%	82%	68%	86%	18%	
2:R language	17%		22%		18%	22%	19%			28%	
3:SQL language	22%	22%			48%	27%	41%		36%	66%	
4:RapidMiner	-17%					-32%	-29%	-21%	-40%	21%	-25%
5:Excel		18%	48%			-22%	43%		24%	74%	36%
6:Spark	50%	22%	27%	-32%	-22%		53%	106%	79%		
7:Anaconda	82%	19%	41%	-29%	43%	53%		112%	178%	50%	
8:Tensorflow	68%			-21%		106%	112%		159%		
9:scikit-learn	86%		36%	-40%	24%	79%	178%	159%		40%	-16%
10:Tableau	18%	28%	66%	21%	74%		50%		40%		
11:KNIME				-25%	36%				-16%		

Ref: http://www.kdnuggets.com/2017/06/ecosystem-data-science-machine-learning-software.html

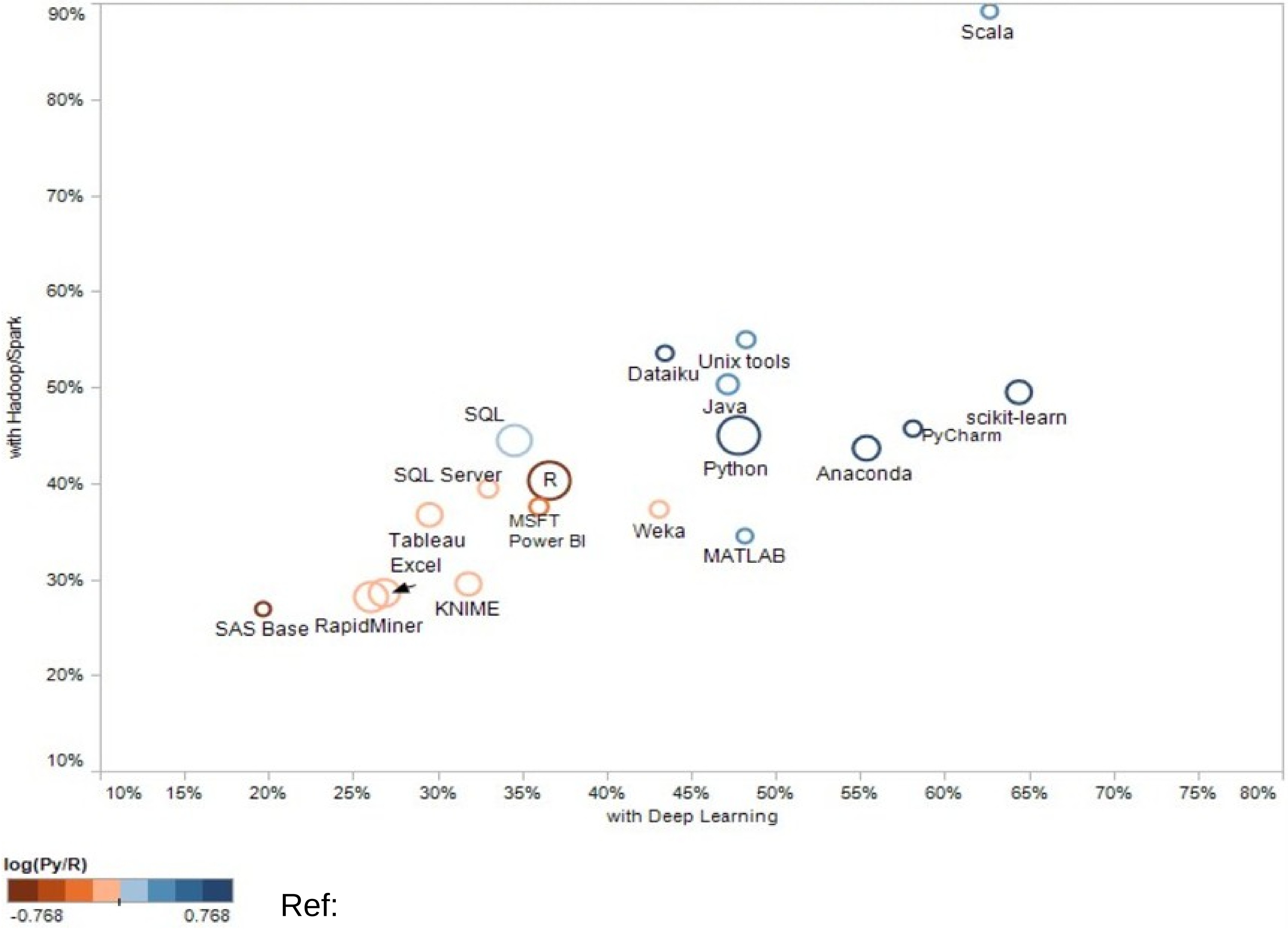
Top Data Science Tools Associations (2017 KDnuggets Survey)



Bar length is Bias_Py_R as defined above, bar height is the popularity of the tool.

Ref: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

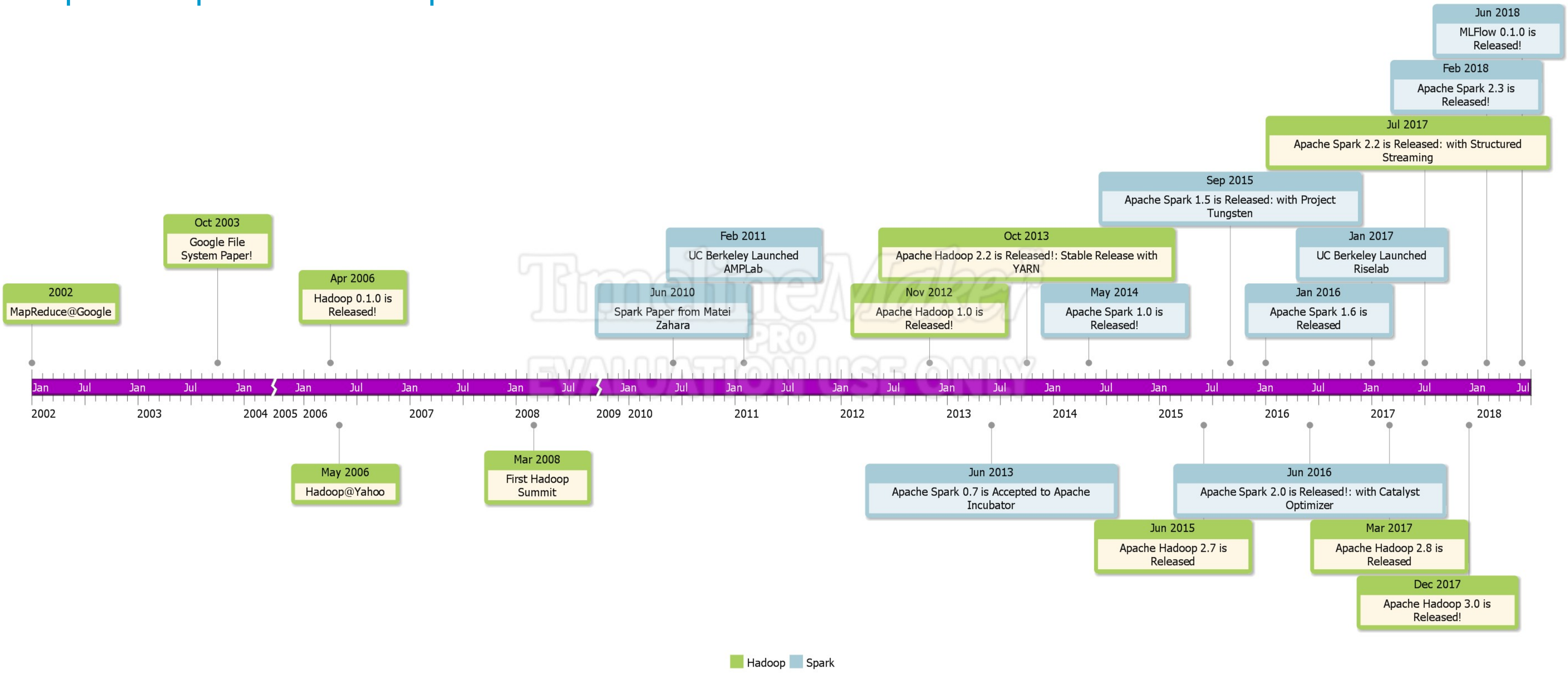
Deep Learning vs Spark/Hadoop affinity for top Data Science Tools(2017 KDnuggets)



Circle size corresponds to tool share of use, and color to Python (blue) vs R (Orange) bias.

Ref: http://www.kdnuggets.com/2017/06/ecosystem-data-science-machine-learning-software.html/2

Apache Spark & Hadoop Timeline



2002 MapReduce@Google

Oct 2003 Google File System Paper!

Apr 2006 Hadoop 0.1.0 is Released!

May 2006 Hadoop@Yahoo

Mar 2008 First Hadoop Summit

Jun 2010 Spark Paper from Matei Zahara

Feb 2011 UC Berkeley Launched AMPLab

Nov 2012 Apache Hadoop 1.0 is Released!

Jun 2013 Apache Spark 0.7 is Accepted to Apache Incubator

Oct 2013 Apache Hadoop 2.2 is Released!: Stable Release with YARN

May 2014 Apache Spark 1.0 is Released!

Jun 2015 Apache Hadoop 2.7 is Released

Sep 2015 Apache Spark 1.5 is Released: with Project Tungsten

Jan 2016 Apache Spark 1.6 is Released

Jun 2016 Apache Spark 2.0 is Released!: with Catalyst Optimizer

Jan 2017 UC Berkeley Launched Riselab

Mar 2017 Apache Hadoop 2.8 is Released

Jul 2017 Apache Spark 2.2 is Released: with Structured Streaming

Dec 2017 Apache Hadoop 3.0 is Released!

Feb 2018 Apache Spark 2.3 is Released!

Jun 2018 MLFlow 0.1.0 is Released!

Jan Jul Jan Jul Jan Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
🟩 Hadoop 🟦 Spark

Created with Timeline Maker Pro v4. Produced on Aug 11 2018.