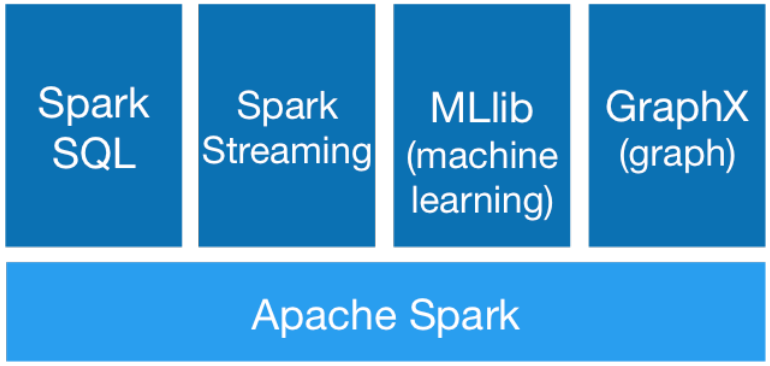
Apache Spark™

February 2015

Grigor Vladimir
vladimir.grigor@kiosked.com

What is Apache Spark™?

- Spark is a fast and general engine for large-scale data processing
- High-level APIs in Java, Scala and Python
- Higher-level tools:
  - Spark SQL for SQL and structured data processing
  - MLlib for machine learning
  - GraphX for graph processing
  - Spark Streaming.

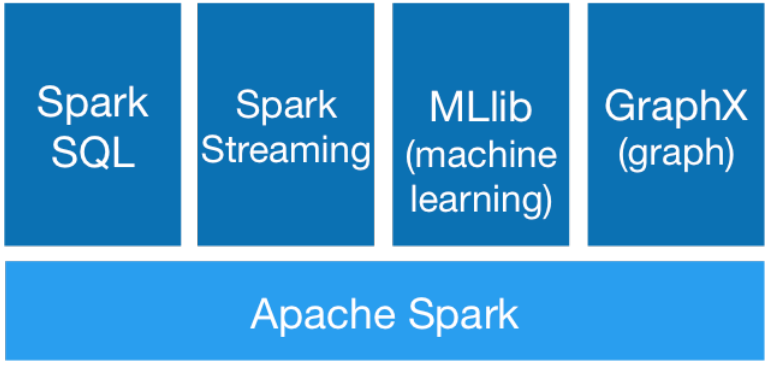


Who uses Apache Spark™?

https://cwiki.apache.org/confluence/display/SPARK/Powered+By+Spark

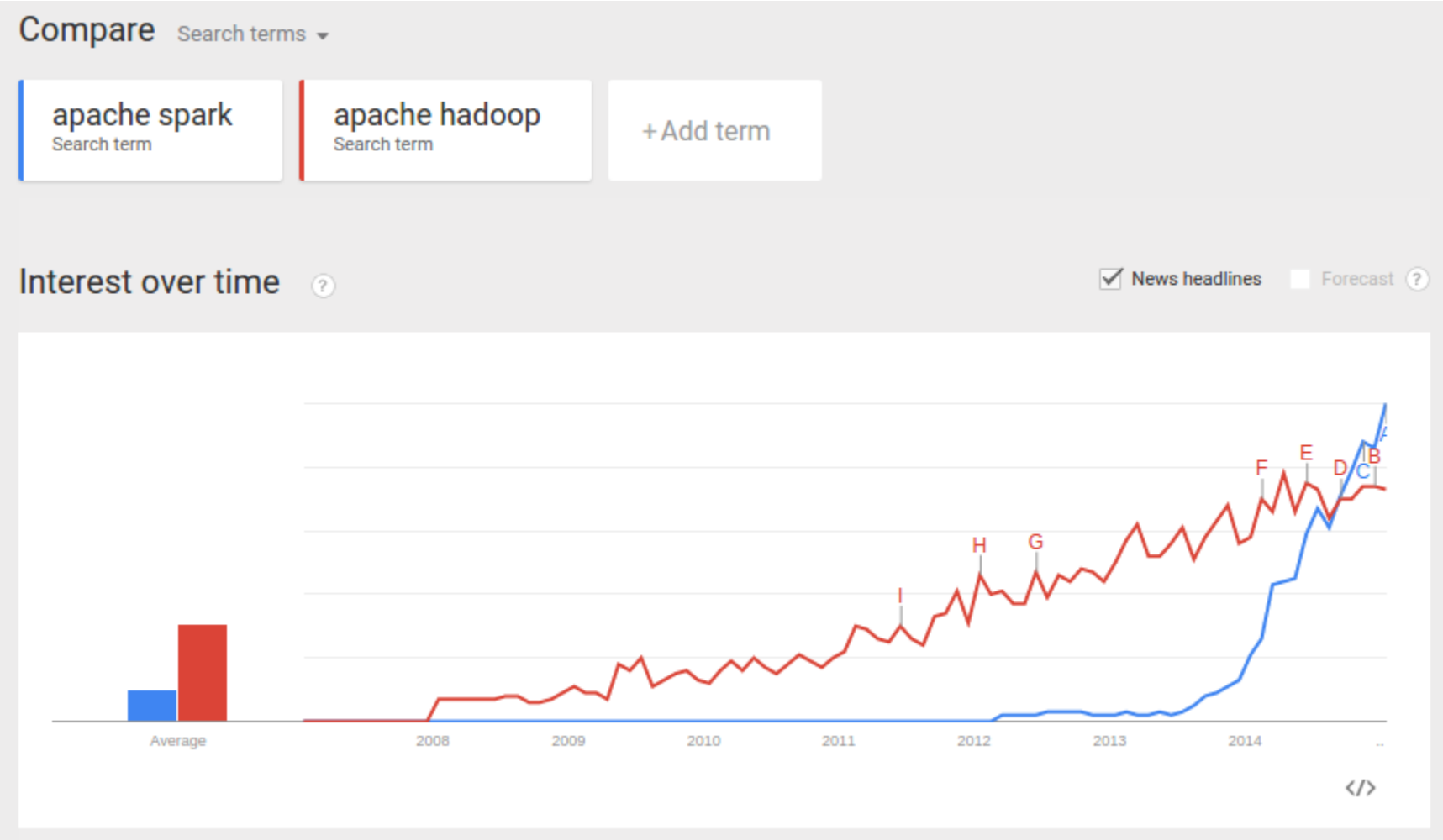
What Apache Spark can be used for?

- Exploratory big data analysis
- Real-time analytics
- Batch processing - EMR jobs



WARNING:
TECHNICAL
STUFF AHEAD

Google Trends

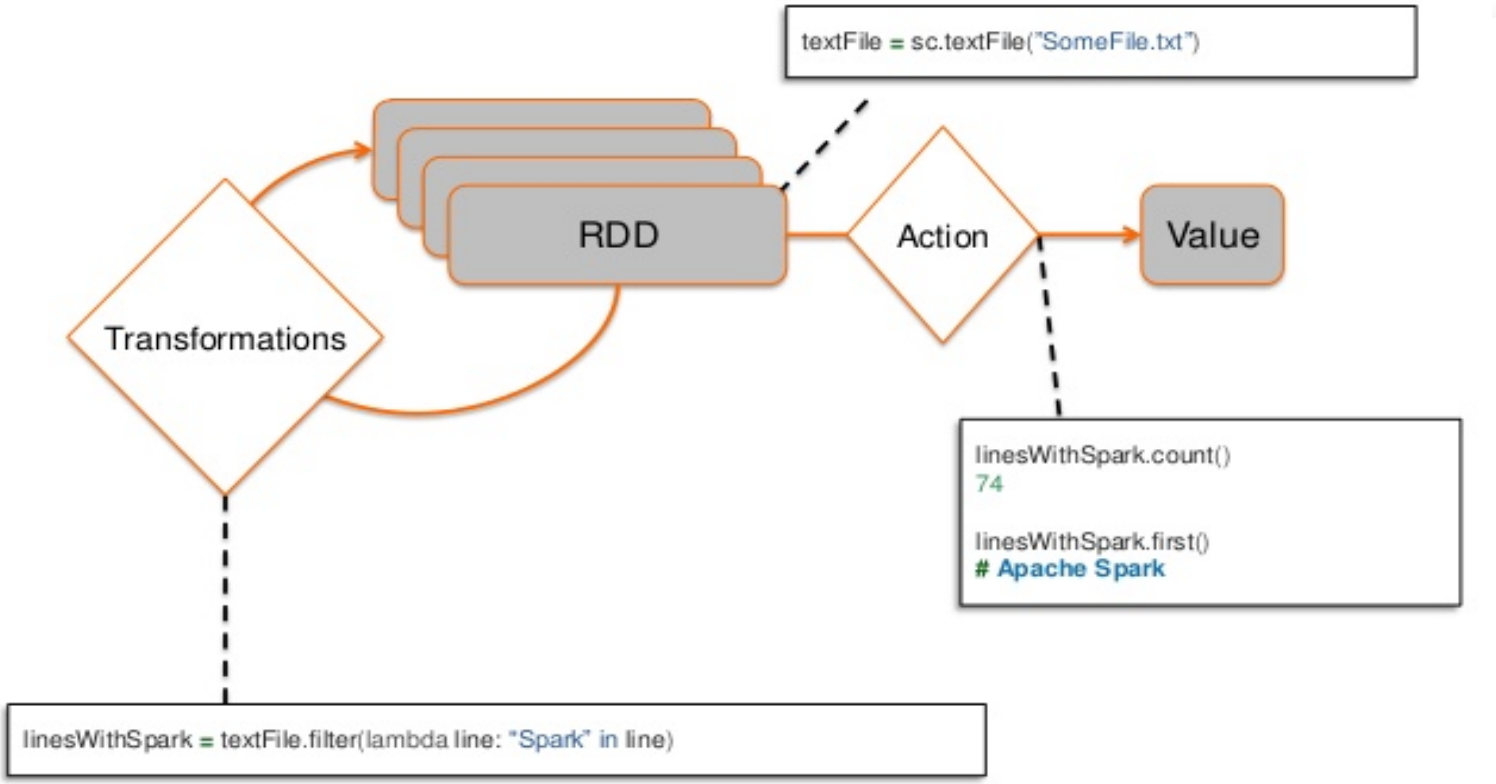


http://www.google.com/trends/explore?hl=en-US#q=apache%20spark%2C%20apache%20hadoop&date=1%2F2007%2097m&cmpt=q&tz=

Demo

TECHNICAL STUFF WARNING

How it works: Resilient Distributed Dataset (RDD)



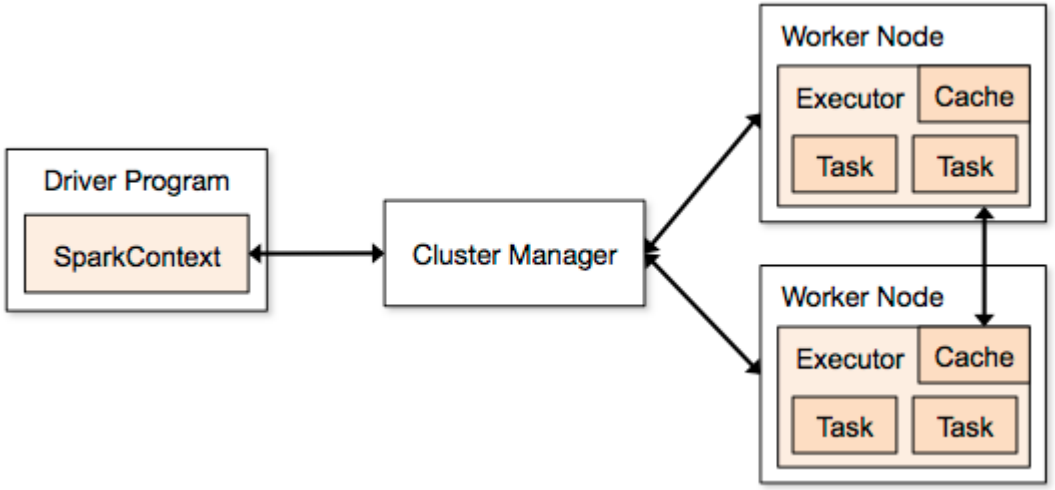
databricks™

http://www.slideshare.net/tsliwowicz/reversim2014

TECHNICAL STUFF WARNING

How it works: cluster mode

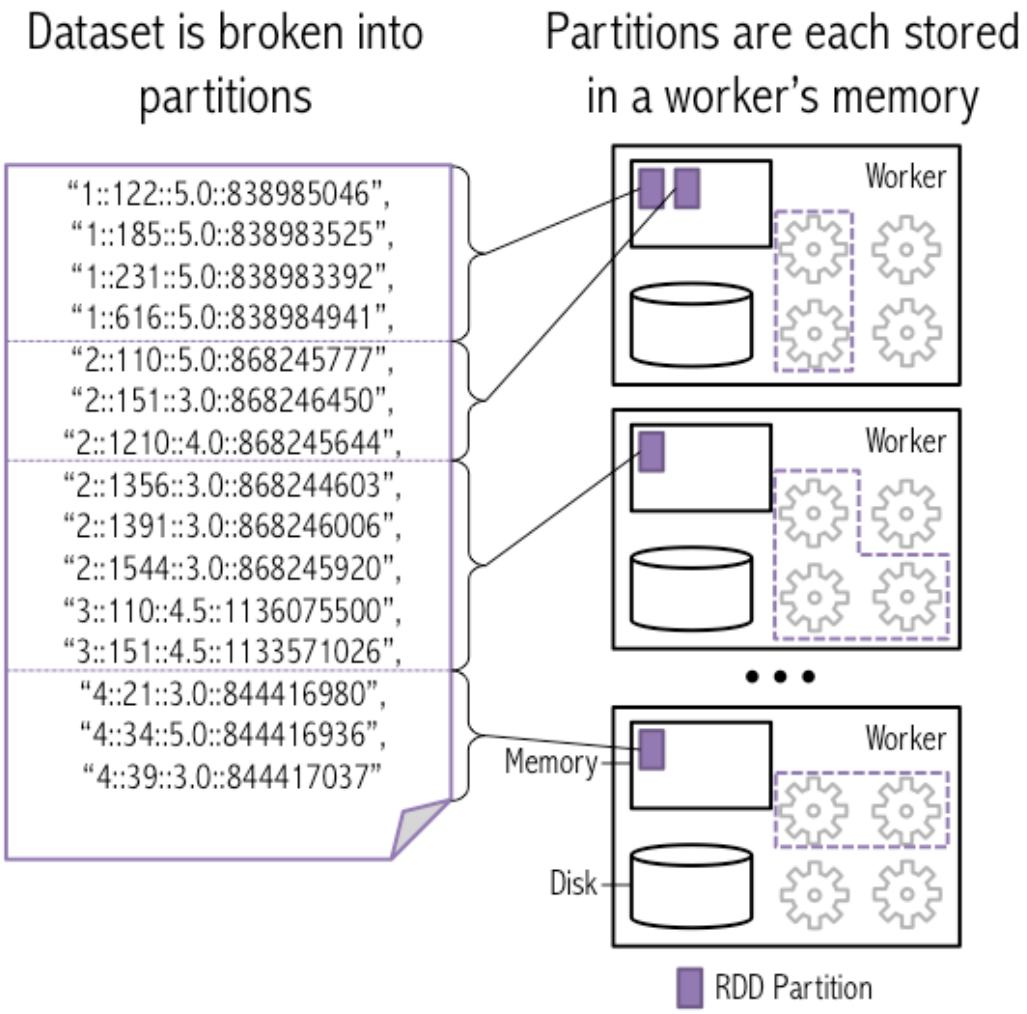
- Application is coordinated by SparkContext run within Driver
- Cluster Manager allocate resources - Executors
- SparkContext sends application code to Executors
- SparkContext sends tasks to Executors



http://spark.apache.org/docs/latest/cluster-overview.html

TECHNICAL STUFF WARNING

How it works: RDD partitions



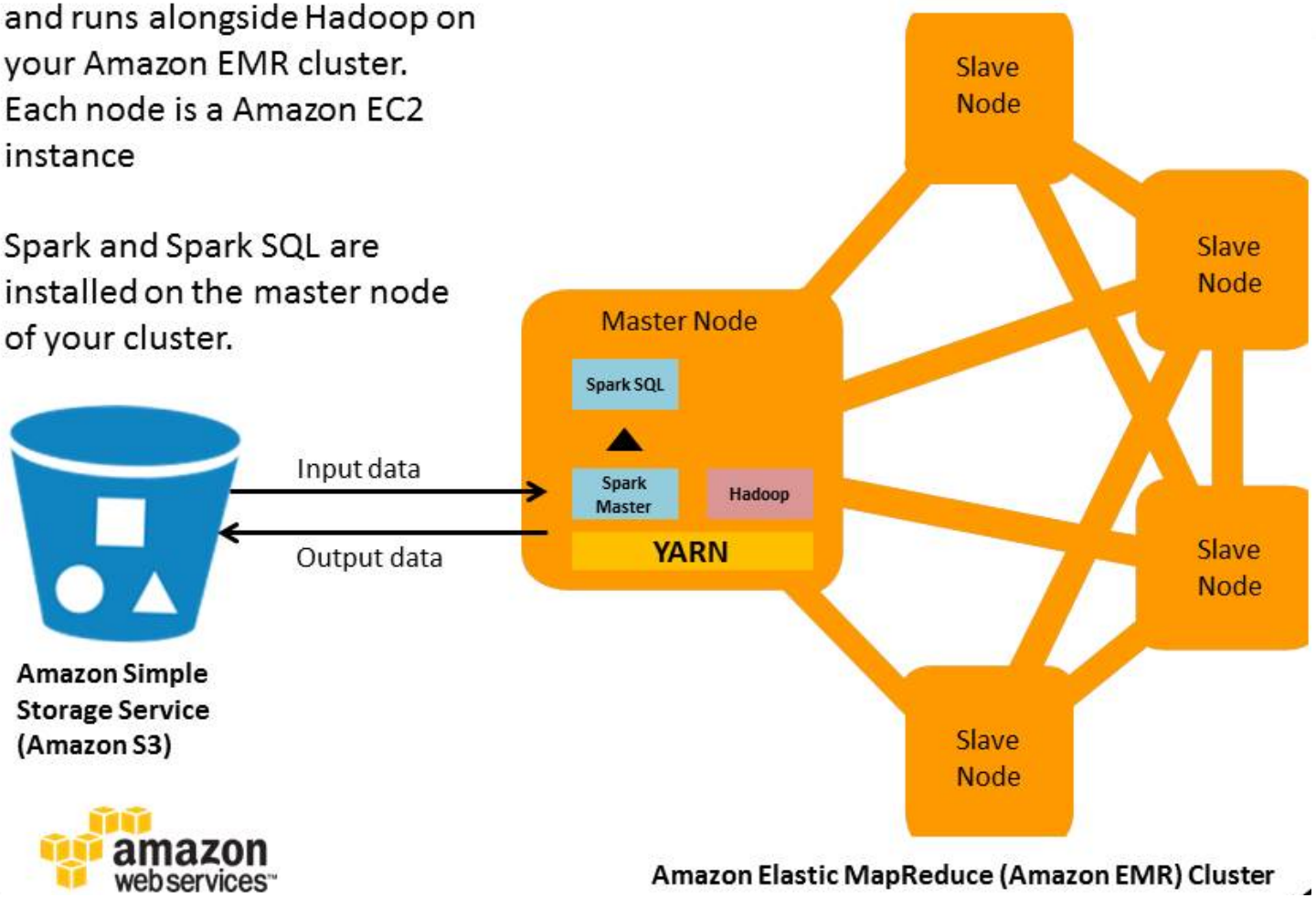
http://nbviewer.ipython.org/github/amplab/datascience-sp14/blob/master/lab8/lab8-withsolutions.ipynb

TECHNICAL STUFF WARNING

How it works: Yarn on EMR

Spark is managed by YARN, and runs alongside Hadoop on your Amazon EMR cluster. Each node is a Amazon EC2 instance

Spark and Spark SQL are installed on the master node of your cluster.



https://aws.amazon.com/articles/4926593393724923

TECHNICAL STUFF WARNING

Pitfalls

- “cluster mode is currently not supported for standalone clusters, Mesos clusters, or python applications”
  - http://spark.apache.org/docs/latest/submitting-applications.html
- Pickle serialization issues with Python code
- No clear deployment
- JVM
- Myriad of configuration possibilities

Application	User program built on Spark. Consists of a *driver program* and *executors* on the cluster.
Application jar	A jar containing the user's Spark application.
Driver program	The process running the main() function of the application and creating the SparkContext
Cluster manager	An external service for acquiring resources on the cluster (e.g. standalone manager, Mesos, YARN)
Deploy mode	Distinguishes where the driver process runs. In "cluster" mode, the framework launches the driver inside of the cluster. In "client" mode, the submitter launches the driver outside of the cluster.
Worker node	Any node that can run application code in the cluster
Executor	A process launched for an application on a worker node, that runs tasks and keeps data in memory or disk storage across them. Each application has its own executors.
Task	A unit of work that will be sent to one executor
Job	A parallel computation consisting of multiple tasks that gets spawned in response to action (e.g. `save`, `collect`)
Stage	Each job gets divided into smaller sets of tasks called *stages* that depend on each other

http://spark.apache.org/docs/latest/cluster-overview.html

TECHNICAL STUFF WARNING

Spark WordCount vs 50+ lines of Java MR

Scala:

```
val f = sc.textFile("README.md")
val wc = f.flatMap(l => l.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
wc.saveAsTextFile("wc_out")
```

Python:

```
from operator import add
f = sc.textFile("README.md")
wc = f.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).reduceByKey(add)
wc.saveAsTextFile("wc_out")
```

http://training.databricks.com/workshop/itas_workshop.pdf

… is hiring!

Vladimir Grigor
vladimir.grigor@kiosked.com

Facebook.com/kiosked
Twitter @Kiosked