+

Spark – A Quick Primer

Sujee Maniyam

sujee@elephantscale.com

Who Invited This Guy? Hi, I am Sujee Maniyam ☺

- 15 years+ software development experience
- Consulting & Training in Big Data
- Author
  - "Hadoop illuminated" open source book
  - "HBase Design Patterns" coming soon
- Open Source contributor (including Hadoop) http://github.com/sujee
- Founder / Organizer of '**Big Data Guru**' meetup http://www.meetup.com/BigDataGurus/
- http://sujee.net/

Spark

- Fast & Expressive Cluster computing engine
- Compatible with Hadoop
- Came out of Berkeley AMP Lab
- Now Apache project
- Version 1.1 just released (Sep 2014)

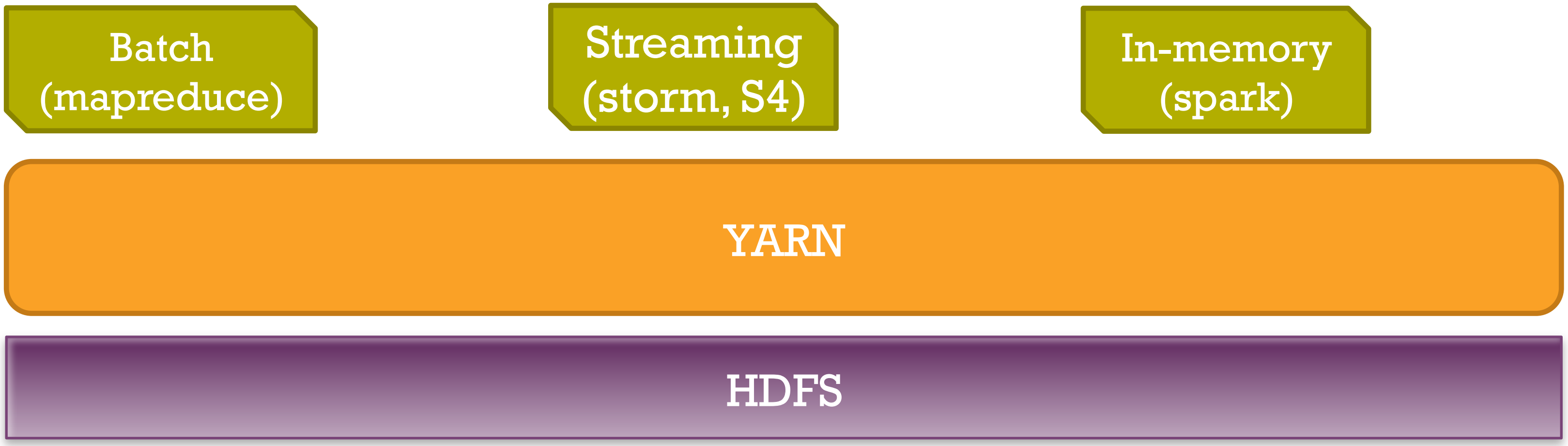
Comparison With Hadoop

Hadoop	Spark
MapReduce framework	Generalized computation
Usually data on disk (HDFS)	On disk / in memory
Not ideal for iterative work	Great at Iterative workloads (machine learning ..etc)
Batch process	- Upto 10x faster for data on disk - Upto 100x faster for data in memory
	Compact code Java, Python, Scala supported
	Shell for ad-hoc exploration

Is Spark Replacing Hadoop?

- Right now, Spark runs on Hadoop / YARN
  - Complimentary
- Can be see as generic MapReduce
- If data fits in memory (few hundred gigs),
  - Spark can really excel
- Future ???

Hadoop + Yarn : Universal OS for Cluster Computing



Batch (mapreduce)

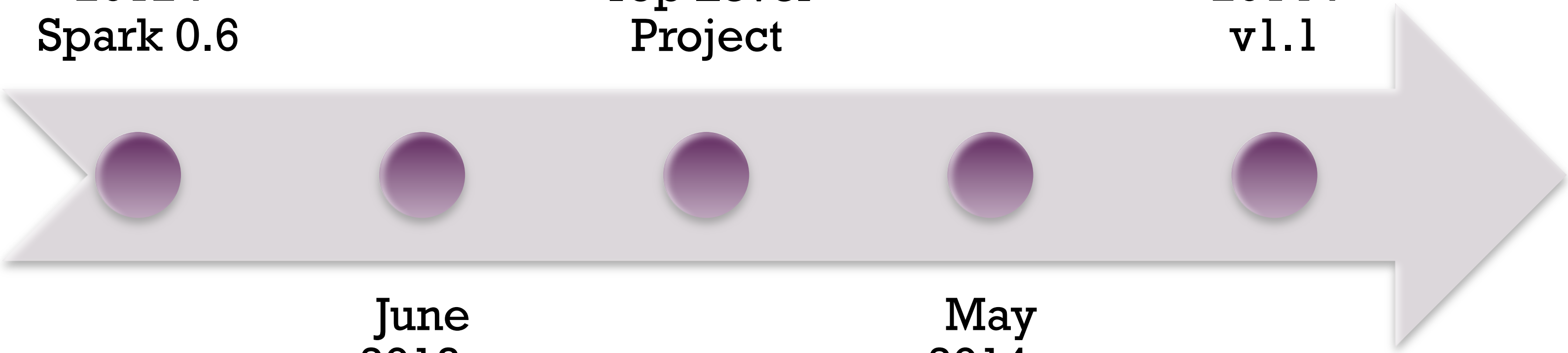
Streaming (storm, S4)

In-memory (spark)

YARN

HDFS

Bit of History



Sep 2012 : Spark 0.6

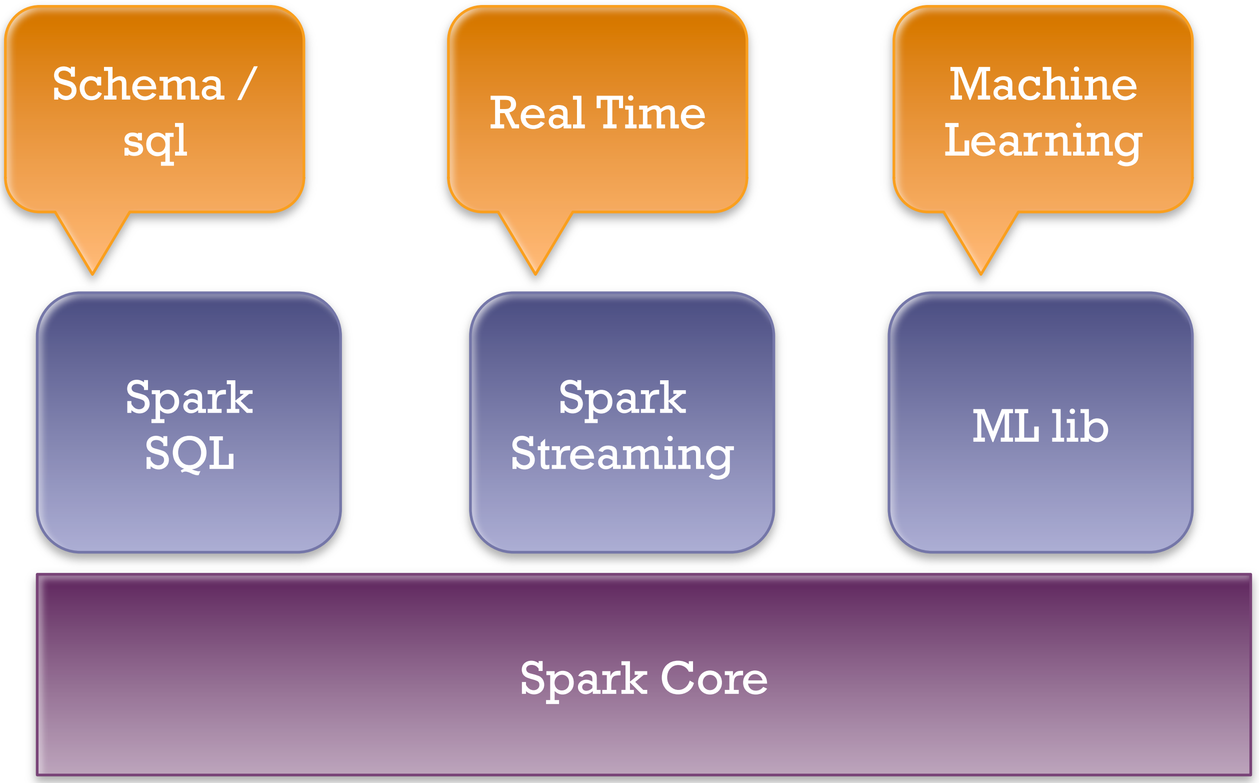
June 2013 : Apache incubator project

Feb 2014 : Apache Top Level Project

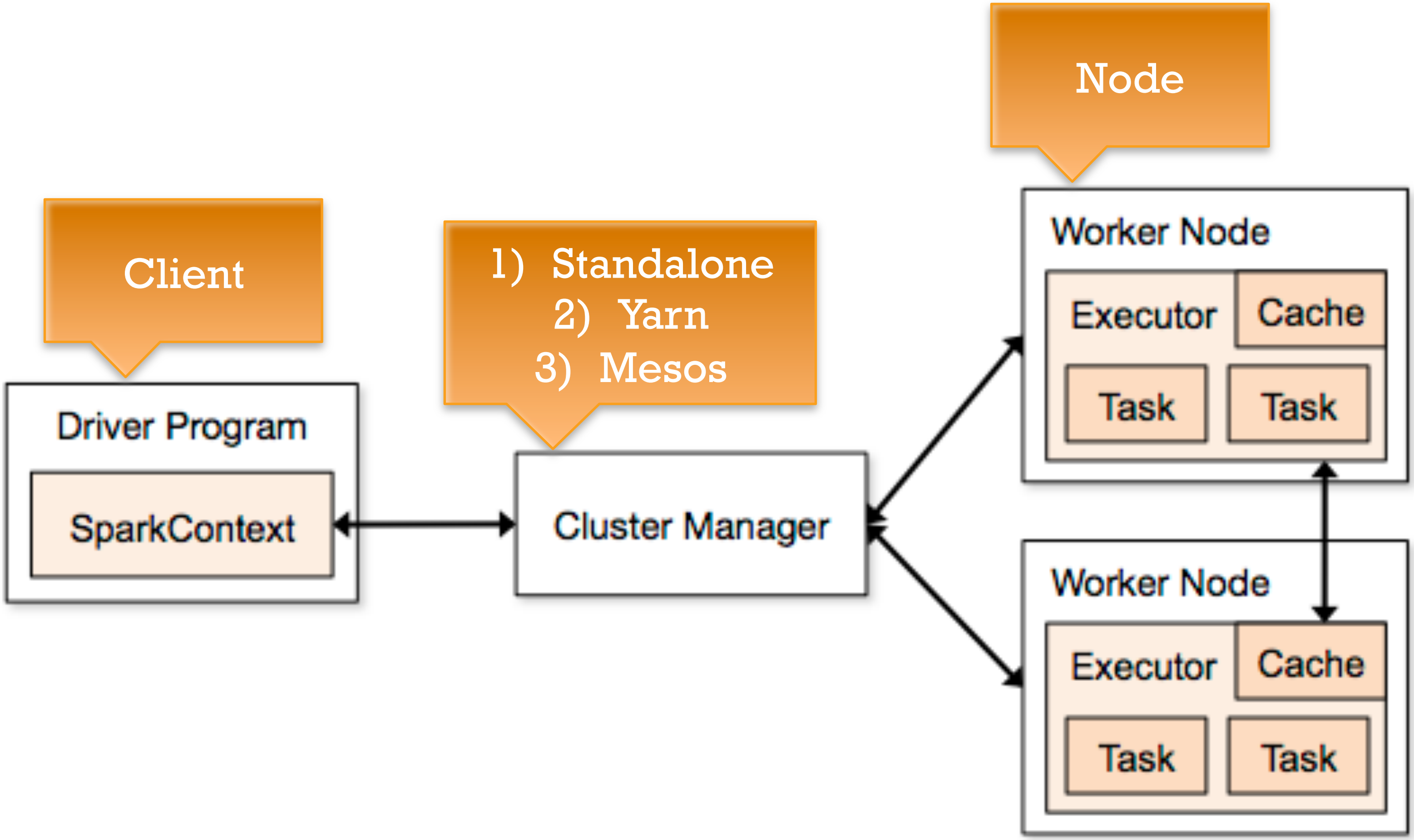
May 2014 : v1.0

Sep 2014 : v1.1

Spark Eco-System



Spark Architecture



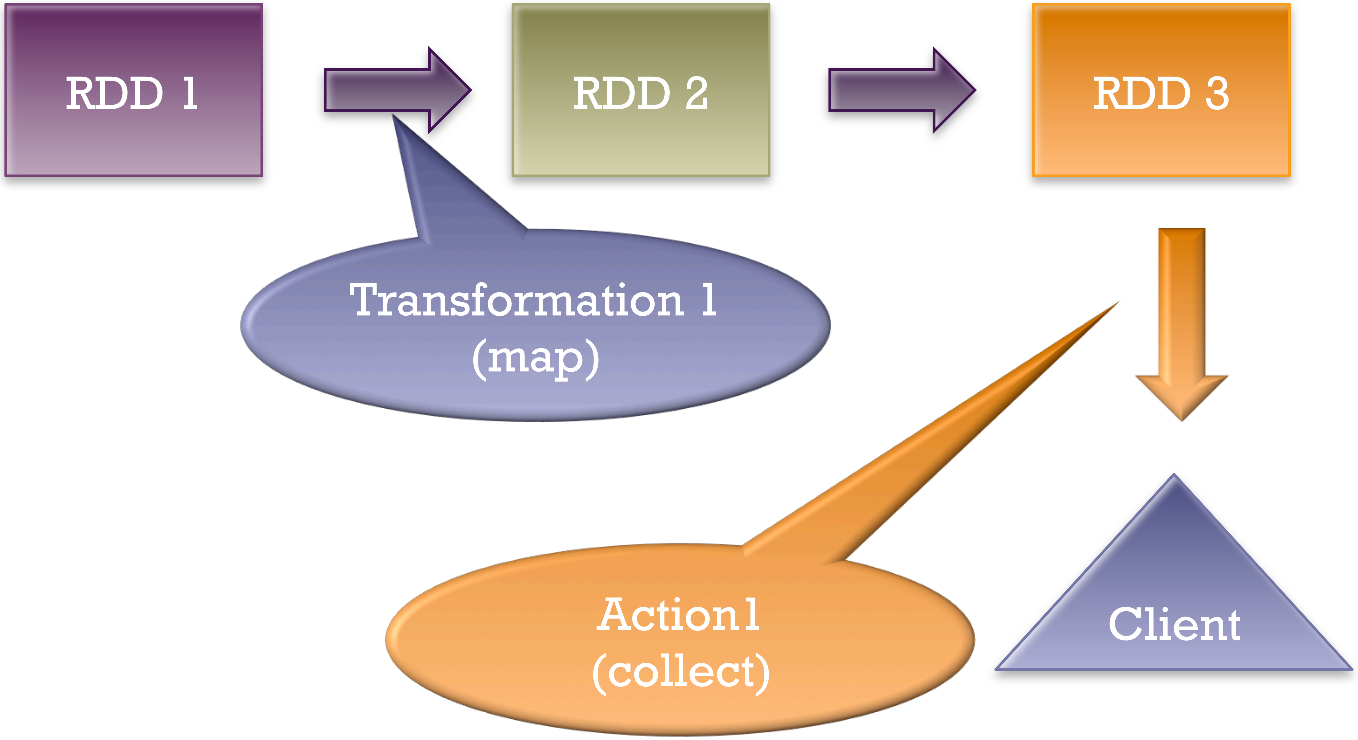
Spark Architecture

- Multiple ‘applications’ can run at the same time
- Each application gets its own ‘executor’
  - Isolated (runs in different JVMs)
  - Also means data can not be shared across applications
- Cluster Managers:
  - multiple cluster managers are supported
  - 1) Standalone : simple to setup
  - 2) YARN : on top of Hadoop
  - 3) Mesos : General cluster manager (AMP lab)

Data Model

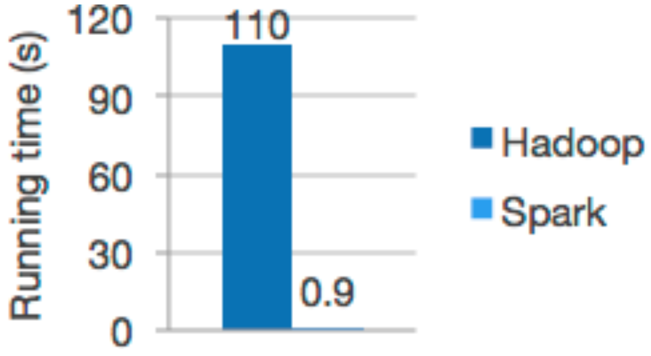
- Resilient Distributed Dataset (RDD)
- Can live in
  - Memory (best case scenario)
  - Or on disk (FS, HDFS, S3 …etc)
- Operations on RDDs
  - 1) Transformations
    - Create a new RDD from existing ones (e.g. Map)
  - 2) Actions
    - E.g. Returns the results to clients (e.g. Reduce)
- Transformations are **lazy**.. Actions force transformations

Transformations / Actions

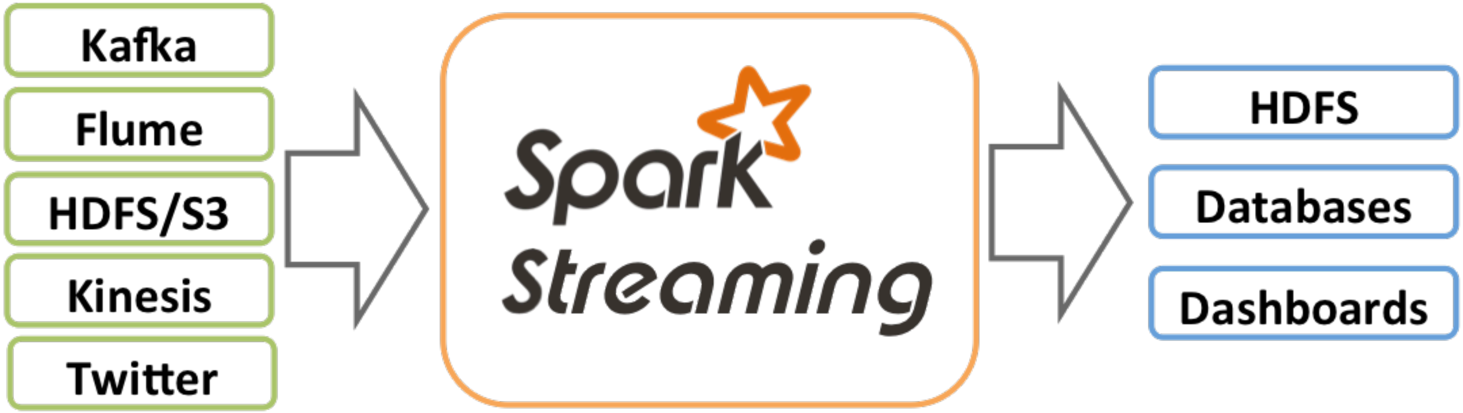


Caching of RDDs

- RDDs can be loaded from disk and computed
  - Hadoop mapreduce model
- Also RDDs can be cached in memory
- Subsequent operations are much faster
- In memory RDDs are great for iterative workloads
  - Machine learning algorithms



Spark Streaming



Kafka

Flume

HDFS/S3

Kinesis

Twitter

Spark Streaming

HDFS

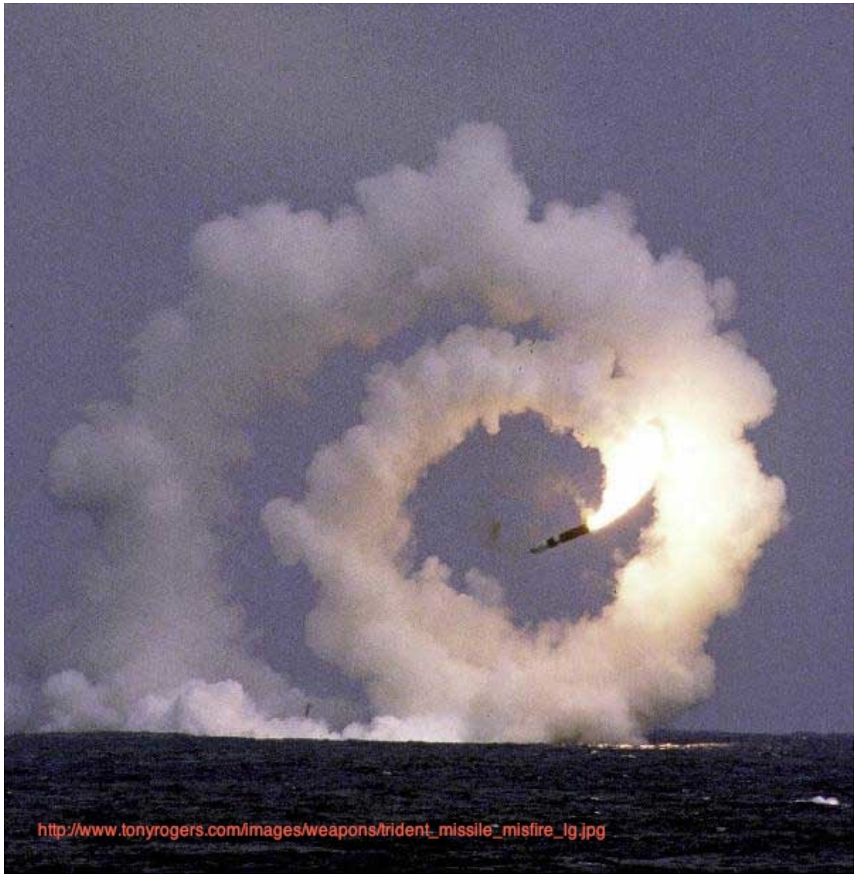
Databases

Dashboards

Machine Learning (ML Lib)

- Out of the box ML capabilities !
- Lots of common algorithms are supported
- Classification / Regressions
  - Linear models (linear R, logistic regression, SVM)
  - Decision trees
- Collaborative filtering (recommendations)
- K-Means clustering
- …
- More to come

DEMO

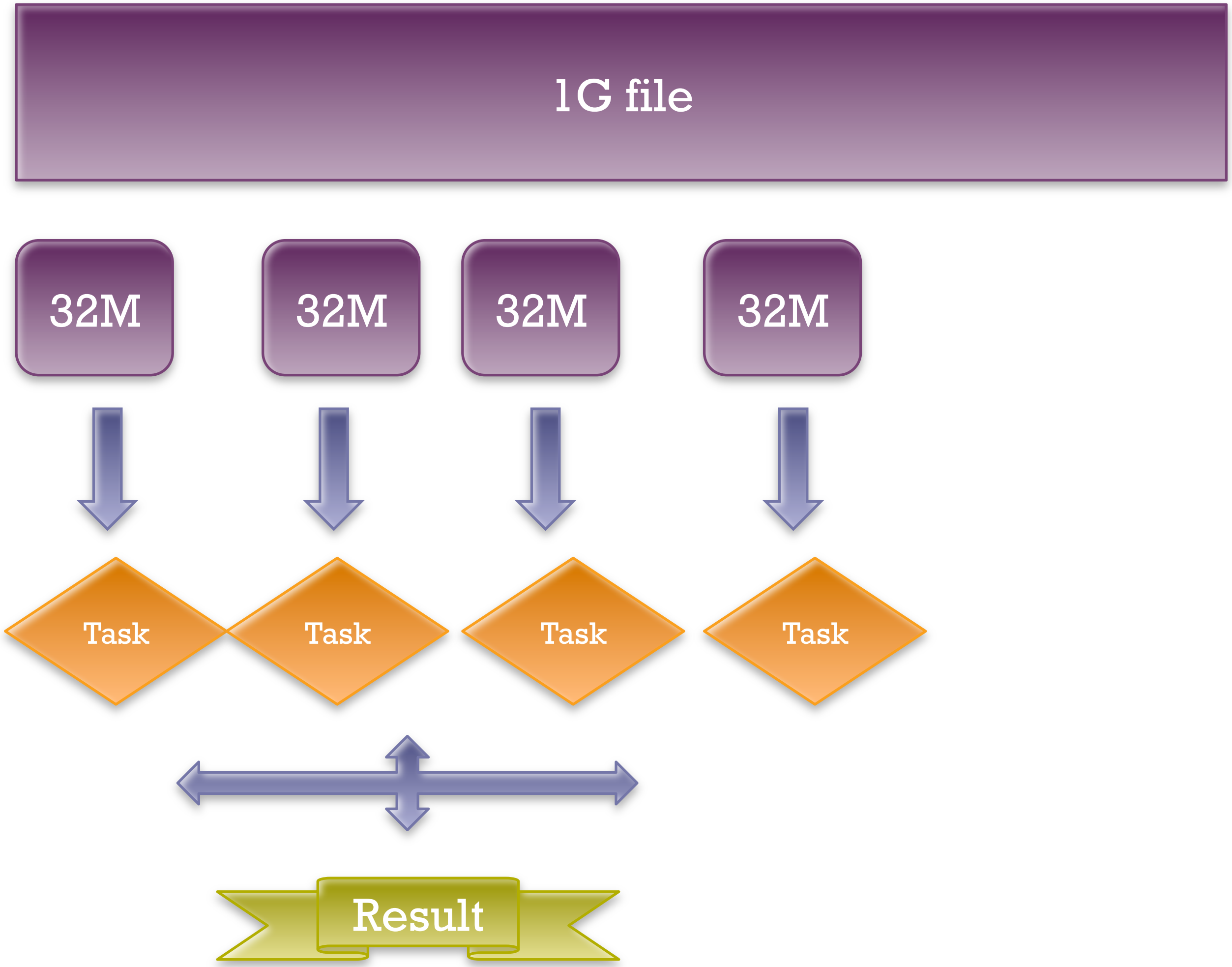


http://www.tonyrogers.com/images/weapons/trident_missile_misfire_lg.jpg

Demo 1 : Quick Start on Single Node

- Run Spark
- Spark Shell
- Load file and count
- Mapreduce

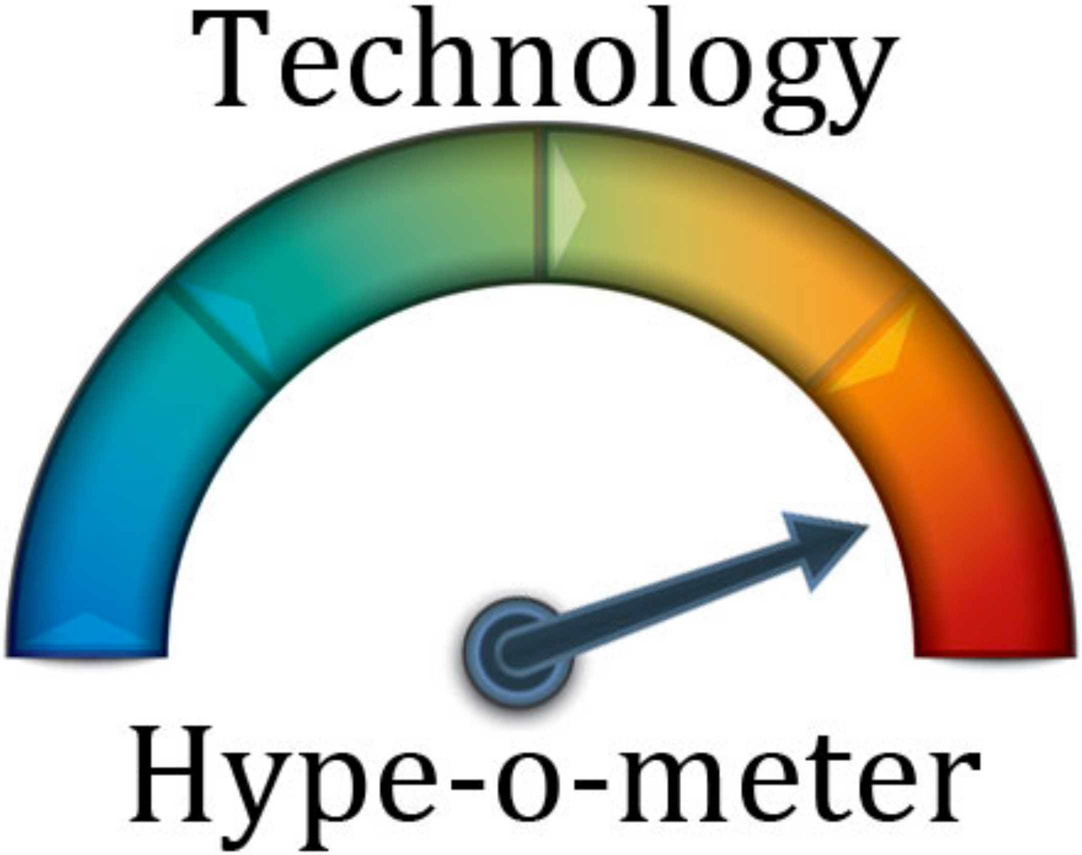
Demo Explained



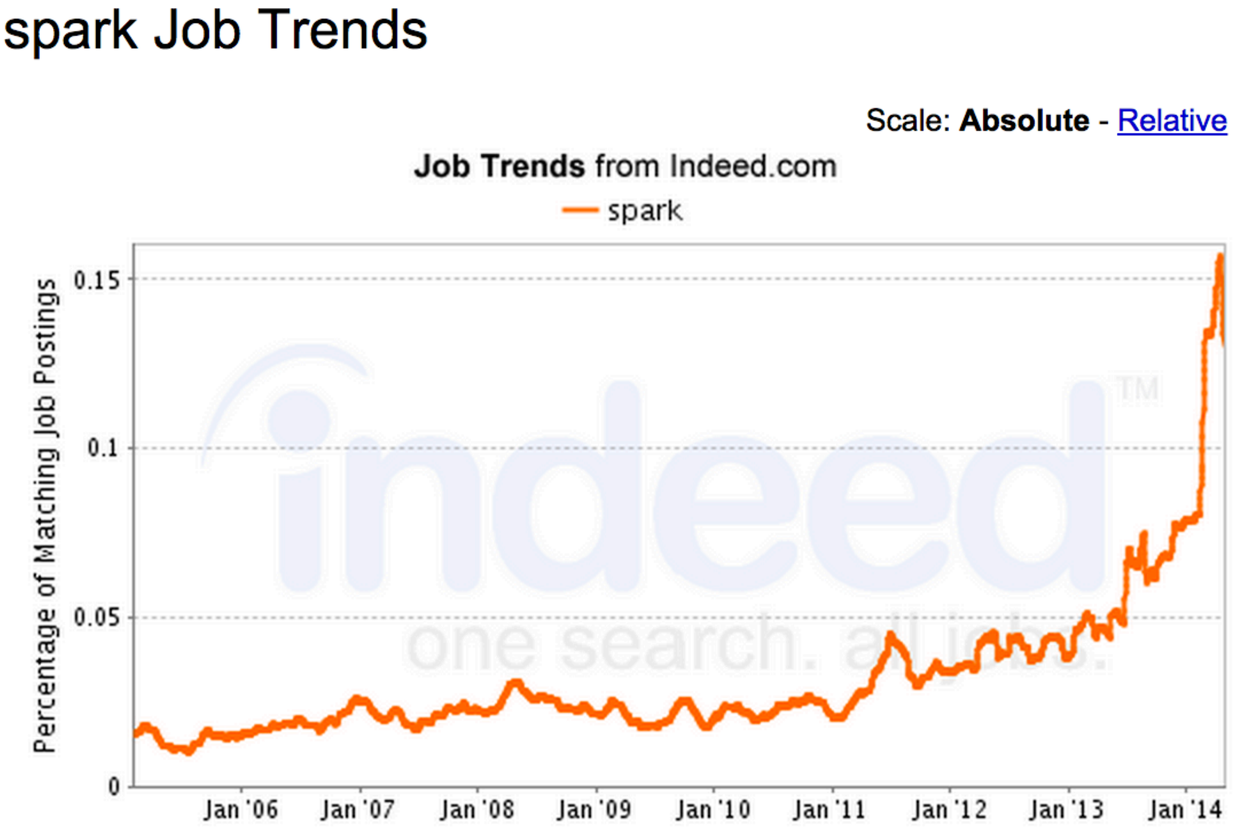
Demo 2 : Distributed Spark

- On Amazon

Hypo-meter ☺



Spark Job Trends



spark Job Trends

Scale: **Absolute** - Relative

Learning More

- http://spark.apache.org
- AMP Camp training
  http://ampcamp.berkeley.edu/big-data-mini-course/
- Spark summit (videos)

Re-cap

- Spark is easier to get started
- Tremendous interest in community
- Plays nice with Hadoop
- Could be the ‘next MapReduce’

Thanks !

Sujee Maniyam

sujee@elephantscale.com

http://elephantscale.com

Expert consulting & training in Big Data

Lab Time !

- Download Spark from http://spark.apache.org
- I have it on USB drive too
- You do need JDK 7

- GitHub : https://github.com/sujee/svcodecamp-2014

Credits

- http://spark.apache.org/
- http://www.strategictechplanning.com