Spark in the Hadoop Ecosystem

Eric Baldeschwieler (a.k.a. Eric14)
twitter: @jeric14

Who is Eric14

- A Hadoop ecosystem cheerleader & Tech Advisor
- Previously
  - CTO/CEO of Hortonworks
  - VP Hadoop Engineering @ Yahoo!

Spark “on the radar”

- 2008 - Yahoo! Hadoop team collaboration w Berkeley Amp/Rad lab begins
- 2009 - Spark example built for Nexus -> Mesos
- 2011 - “Spark is 2 years ahead of anything at Google”
  - Conviva seeing good results w Spark
- 2012 - Yahoo! working with Spark / Shark
- Today - Many success stories
  - Early commercial support

Spark Today

Spark updates Hadoop

- Hardware had advanced since Hadoop started:
  - Very large RAMs, Faster networks (10Gb+)
  - Bandwidth to disk not keeping up
- MapReduce awkward for key big data workloads:
  - Low latency dispatch (E.G. quick queries)
  - Iterative algorithms (E.G. ML, Graph…)
  - Streaming data ingest

Spark, “lingua franca?”

- Support for many development techniques
  - SQL, Streaming, Graph & in memory, MapReduce
  - Write “UDFs” once and use in all contexts
- Small, simple & elegant API
  - Easy to learn and use; expressive & extensible
  - Retains advantages of MapReduce (fault tolerance…)

Spark often better

- Today you will hear many success stories from teams who have converted Hadoop based workloads to Spark and seen:
  - Huge speedups
  - Big cost savings
- But there do exist cases were Hadoop is superior…
  - Proven to work at the largest scales
  - Mature & widely commercially supported
  - Much larger ecosystem of solutions and tools

Spark complements Hadoop

- Hadoop 2.x being widely deployed this year
  - Spark now just another kind of Yarn Job
- Spark leverages Hadoop ecosystem
  - HDFS, HCatalog, Data Input/OutputFormats
  - Huge investment in data collection & tooling

The Future

Spark **and** Hadoop

- Hadoop vs. BDAS is not the story!
  - Yes spark is useful separately from Hadoop and may flourish independently…
  - But with Yarn, Spark can be brought to the data in every Hadoop 2 cluster in the world…
  - Complimenting the investment already made by enterprise.

Spark **the** “lingua franca”

Data scientists & Developers need an open standard for sharing their Algorithms & functions, an “R” for big data.

- Spark best current candidate:
  - Open Source - Apache Foundation
  - Expressive (MR, iteration, Graphs, SQL, streaming)
  - Easily extended & embedded (DSLs, Java, Python…)

“The future is already here.
It’s just not very evenly distributed.”

– William Gibson