Spark, Hadoop, and Friends

(and the Zeppelin Notebook)

Douglas Eadline

Jan 4, 2017

NJIT

Presenter

Douglas Eadline
deadline@basement-supercomputing.com

@thedeadline

- HPC/Hadoop Consultant/Writer
- http://www.basement-supercomputing.com

Today's Topics

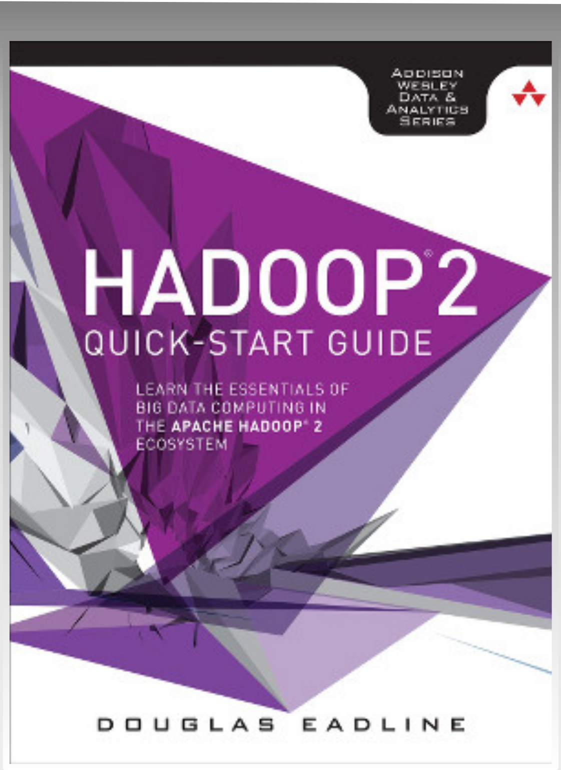
1. Hadoop Overview and Review
2. Spark (and Hadoop)
3. A Little Bit about Data Science
4. Getting Data into Hadoop (hands-on)

   (lunch)

5. Spark and The Zeppelin Notebook
6. Examples with Zeppelin (hands-on)

Hadoop Overview and Review

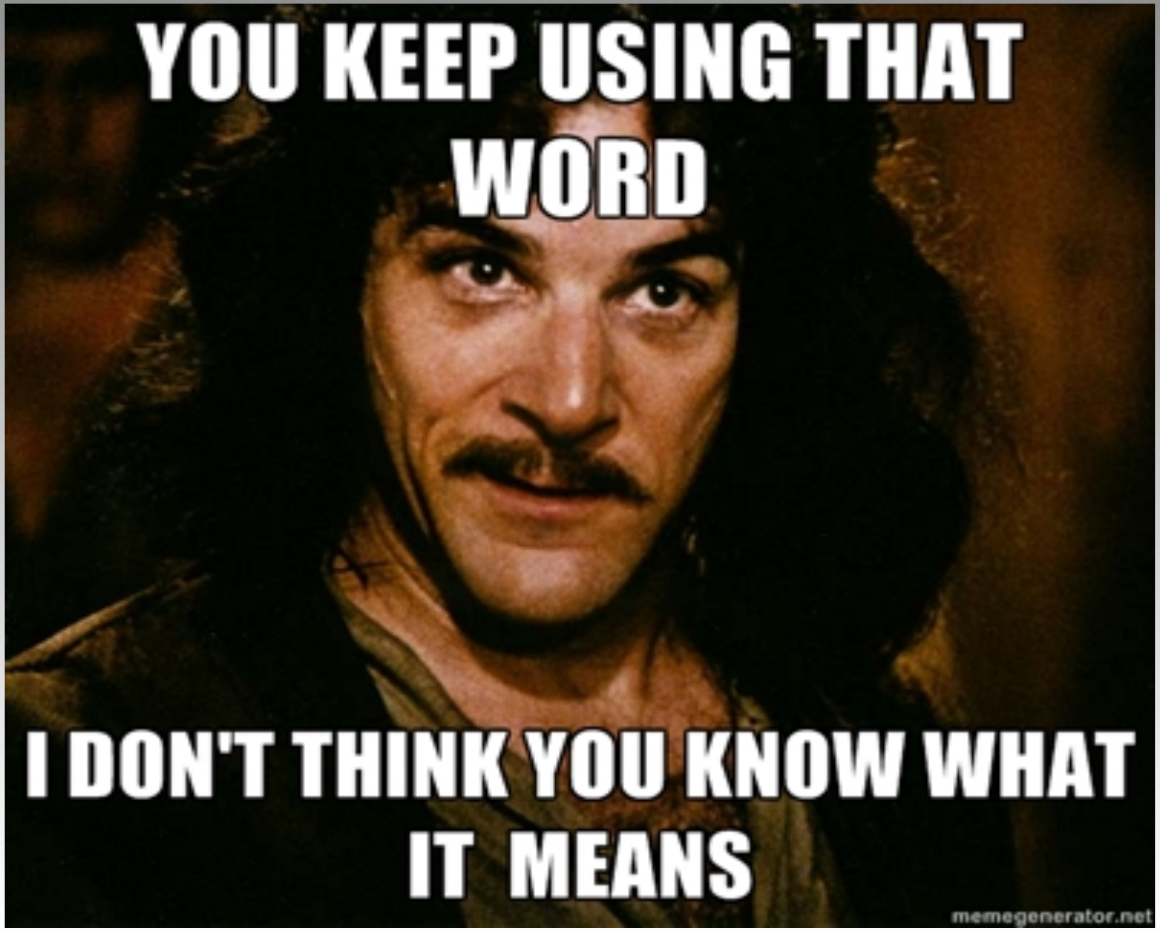
- Covered Hadoop in more detail last year (you should have the book)
- Today: Quick review in case you missed it
- Consider *Hadoop 2 Quick Start Guide*


Addison Wesley Data & Analytics Series
HADOOP 2
QUICK-START GUIDE
LEARN THE ESSENTIALS OF BIG DATA COMPUTING IN THE APACHE HADOOP 2 ECOSYSTEM
DOUGLAS EADLINE

http://www.clustermonkey.net/Hadoop2-Quick-Start-Guide

BASEMENT SUPERCOMPUTING

Big Data & Hadoop Are The Next Big Thing!



BASEMENT SUPERCOMPUTING

Data Volume – Let's Call It a Lake

BASEMENT SUPERCOMPUTING

Data Velocity – Let’s Call It a Waterfall

BASEMENT SUPERCOMPUTING

Data Variation – Let’s Call it Recreation

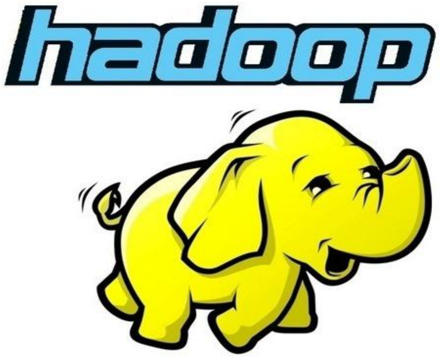
What is Big Data?
(besides a marketing buzzword)

- The Internet created this situation
- Three V’s (Volume, Velocity, Variability)
- Data that are large (lake), growing fast (waterfall), unstructured (recreation) - not all may apply.
- May not fit in a "relational model and method"
- Can Include: video, audio, photos, system/web logs, click trails, IoT, text messages/email/tweets, documents, books, research data, stock transactions, customer data, public records, human genome, and many others.

Why All the Fuss about Hadoop?

- We hear about Hadoop scale
- We hear about Hadoop storage systems
- We hear about Hadoop parallel processing – something called "MapReduce"

At Yahoo, 100,000 CPUs in more than 40,000 computers running Hadoop, 455 petabytes of storage (10E15 Bytes)



Is BIG DATA Really That big

Two analytics clusters at Yahoo and Microsoft, median job input sizes are under 14GB and 90% of jobs on a Facebook cluster have input sizes under 100 GB. (“Nobody ever got fired for using Hadoop on a cluster,” HotCDP 2012)

BASEMENT SUPERCOMPUTING

What Is Really Going on Here?

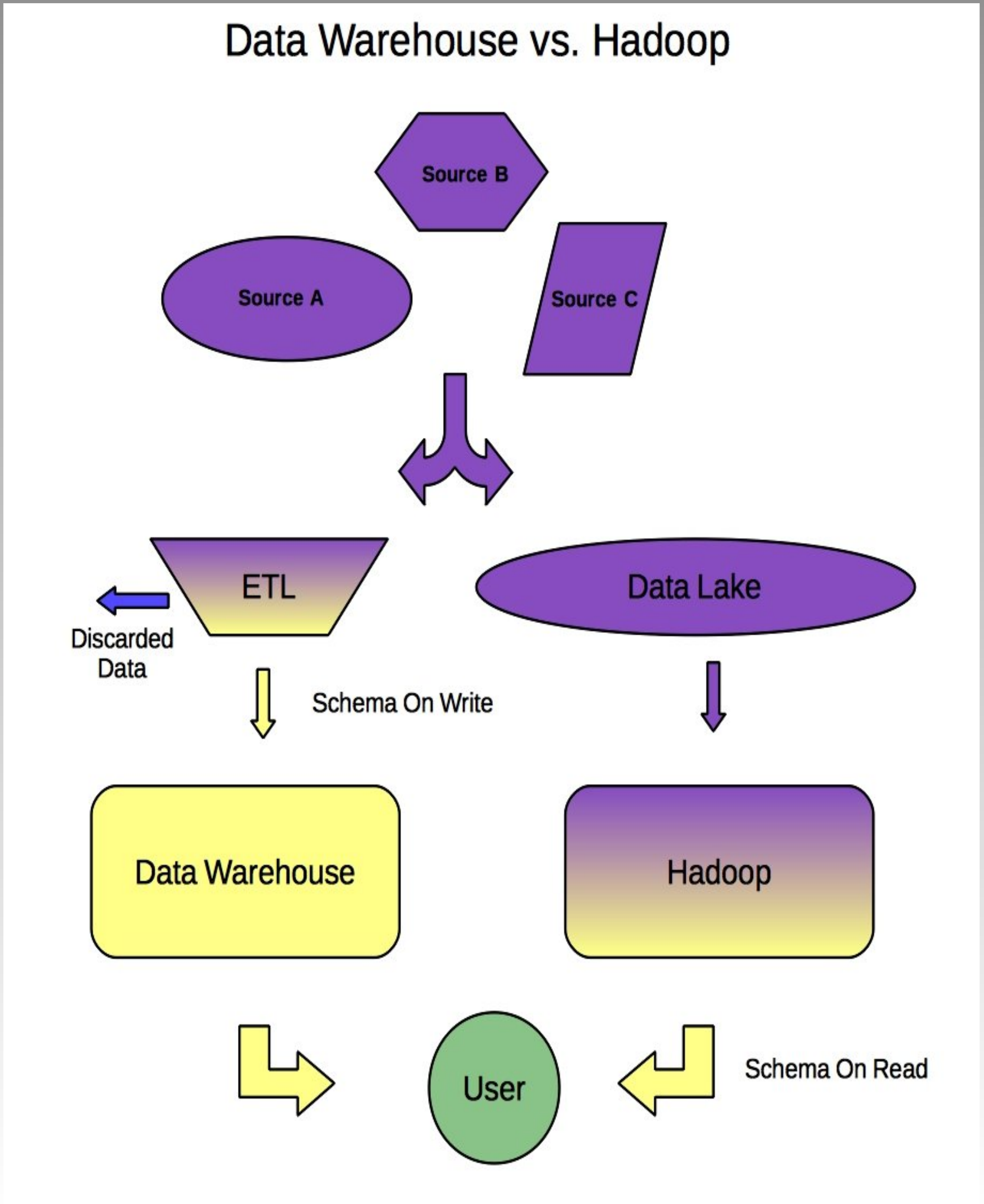
- Hadoop is not necessarily about scale
  (although it helps to have a scalable solution that can handle large capacities and large capabilities)
- Big Data is not necessarily about volume
  (although volume and velocity are increasing, so is hardware)

However, The "Hadoop Data Lake" represents a fundamentally new way to manage data.

Hadoop Data Lake

Data Warehouse applies "schema on write" and has an Extract, Transform, and Load (ETL) step.

Hadoop applies "schema on read" and the ETL step is part of processing. All input data are placed in the lake in raw form.



Defining Hadoop (Version 2)

A PLATFORM for data lake analysis that supports software tools, libraries, and methodologies

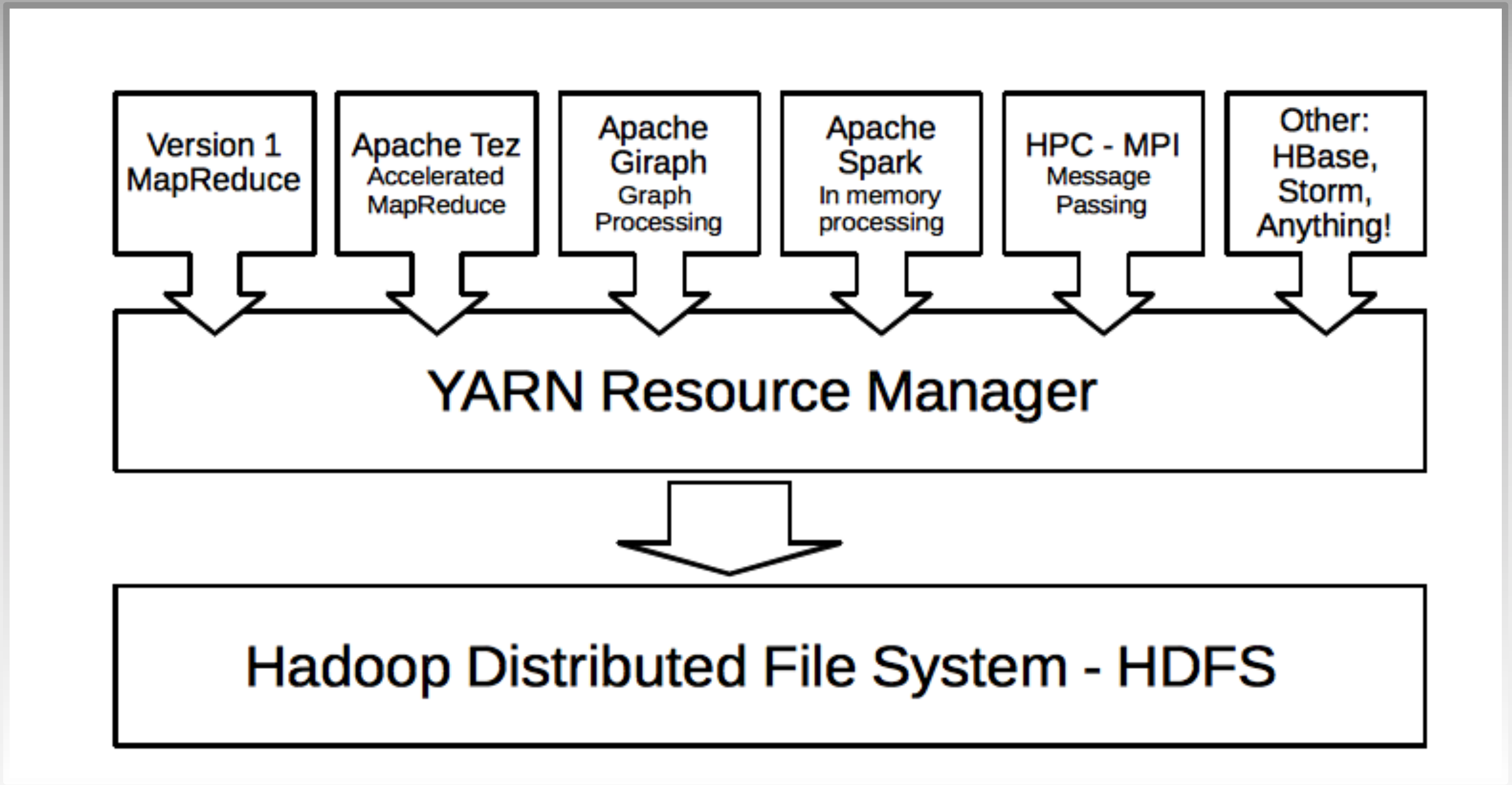
- Core and most tools are open source (Apache License)
- Large unstructured data sets (Petabytes)
- Written in Java, but not exclusively a Java platform
- Primarily GNU/Linux, Windows versions available
- Scalable from single server to thousands of machines
- Runs on commodity hardware and the cloud
- Application level fault tolerance possible
- Multiple processing models and libraries (>MapReduce)

BASEMENT SUPERCOMPUTING

Hadoop Core Components

- **HDFS – Hadoop Distributed File System.** Designed to be a fault tolerant streaming file system using multiple “DataNodes.”
- **YARN – Yet Another Resource Negotiator.** Master scheduler and resource allocator for the entire Hadoop cluster using worker nodes with “NodeManagers.”
- **MapReduce – YARN Application.** Provides MapReduce processing (Compatible with version 1 of Hadoop).
- **Cluster servers (nodes) are usually both DataNodes and NodeManagers (move processing to data)**

Hadoop V2 The Platform



Version 1 MapReduce

Apache Tez Accelerated MapReduce

Apache Giraph Graph Processing

Apache Spark In memory processing

HPC - MPI Message Passing

Other: HBase, Storm, Anything!

YARN Resource Manager

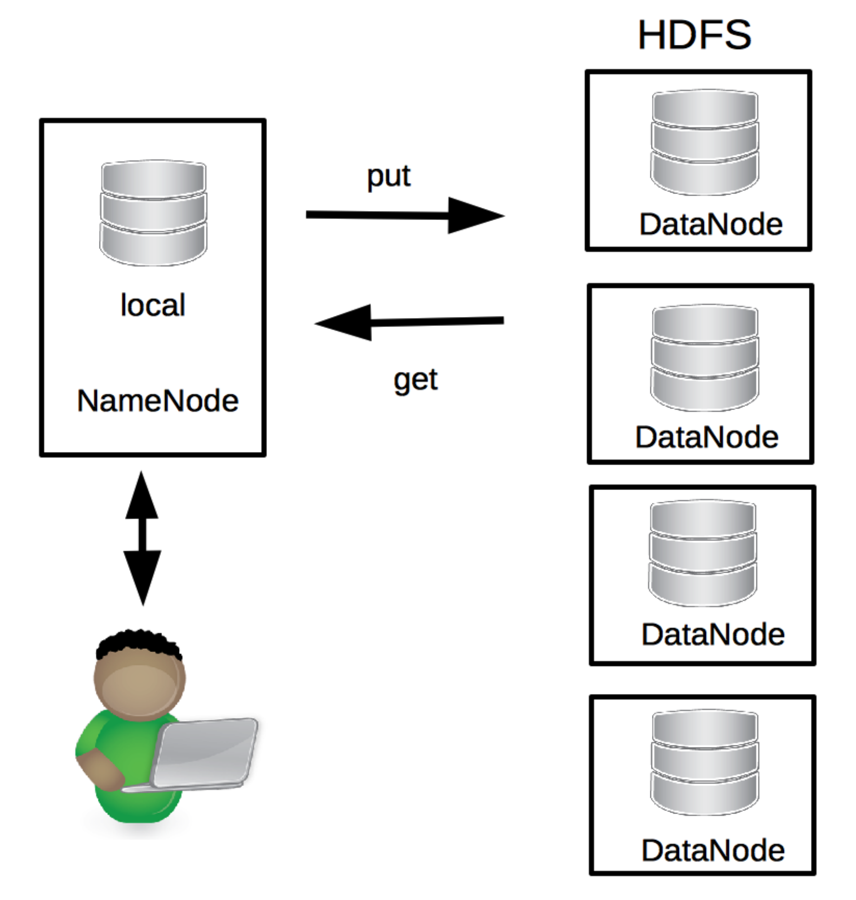
Hadoop Distributed File System - HDFS

Hadoop Distributed File System (HDFS)

- Master/Slave model
- NameNode - metadata server or "data traffic cop" (kept in memory for performance)
- Single Namespace - managed by the NameNode
- DataNodes - where the data live (data are sliced and placed across the DataNodes allowing concurrent access)
- Secondary NameNode - checkpoints NameNode metadata to disk, but not failover node
- Built with "simple servers"
- Features: High Availability (multiple NameServers), Federation (spread namespace across NameServers), Snapshots (instant read-only point-in-time copies), NFSv3 Access, federation

How the User “Sees” HDFS

- HDFS is a separate file system from the host machine
- Data must be moved to (put) and from (get) HDFS (some tools do this automatically)
- Hadoop processing happens in HDFS



A Few Hadoop Use Cases

- **NextBio** is using Hadoop MapReduce and HBase to process multi-terabyte data sets of human genome data. Processing wasn't feasible using traditional databases like MySQL.
- **US Xpress**, one of the largest trucking companies in the US, is using Hadoop to store sensor data (100s of data points and geo data) from thousands of trucks. The intelligence they mine out of this, saves them $6 million year in fuel cost alone. Hadoop allows storing enormous amount of sensor data and querying/joining this data with other data sets.
- A leading **retail bank** is using Hadoop to validate data accuracy and quality to comply with federal regulations like Dodd-Frank. The platform allows analyzing trillions of records which currently result in approximately one terabyte per month of reports. Other solutions were not able to provide a complete picture of all the data.