A Crash Course in Apache Hadoop

Event Outline

1. What is Hadoop
2. Current data challenges
3. Hadoop Solutions
4. Architecture
5. Workshop

Who & When

- Origin from Google papers
- Originally developed at Yahoo!
  - Doug Cutting, Michael Cafarella
- Project officially began around 2005.
- Named after a toy elephant

Why

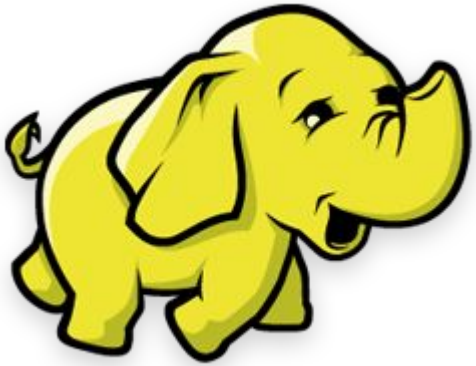
- More ways to collect data
- Too much data
  - CERN Laboratory
  - Google/Yahoo/Facebook
- Forget about processing when you can’t even store it

Analogy

- Imagine you needed to transport 2,000,000kg of raw material
- How would you do it? (Let's assume that horsepower is proportional to the mass that each vehicle can carry)

  - Ferrari 458 will run - $243,000 ~560 Horsepower
  - Bugatti Veyron - $2,310,688 ~1200 horsepower
  - Brand new Ford F-150 will cost $30,000 ~325 horsepower
  - Dodge Caravan - $20,000 ~280 horsepower

What



- Provide a way to reliably *access* and *process* large volumes of data
- Designed to scale across many, many machines.

The ASF

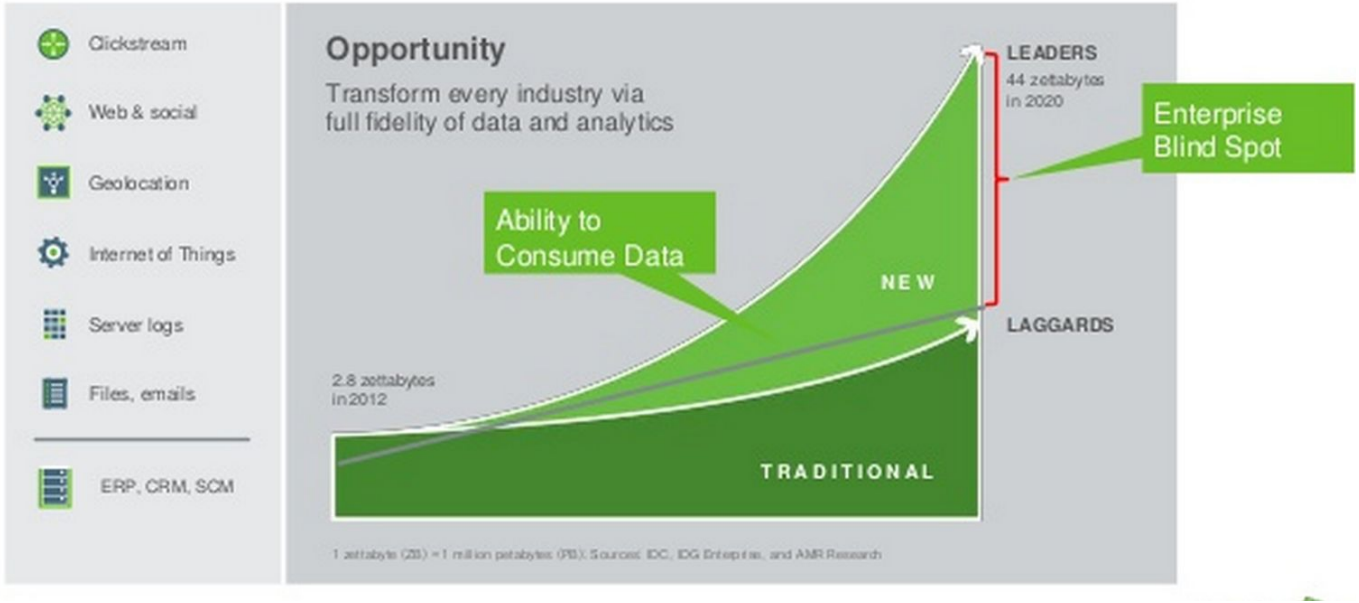
- Apache Open Source
  - OpenOffice
  - HTTP Server
  - Subversion
  - Tomcat Webserver
  - Commons
  - Maven
  - *Hadoop*
- Anyone can view the source code!
  - Build/edit/modify on your own machine

The Apache Software Foundation

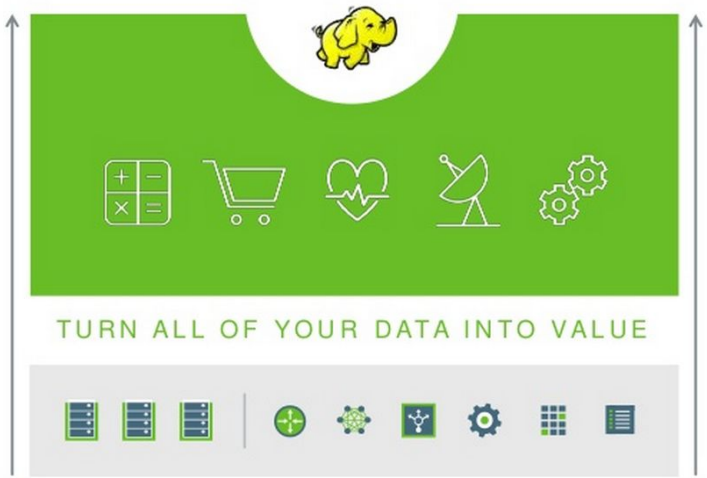
Community-led development since 1999.

Why data?

Opportunities and analytic insights for businesses

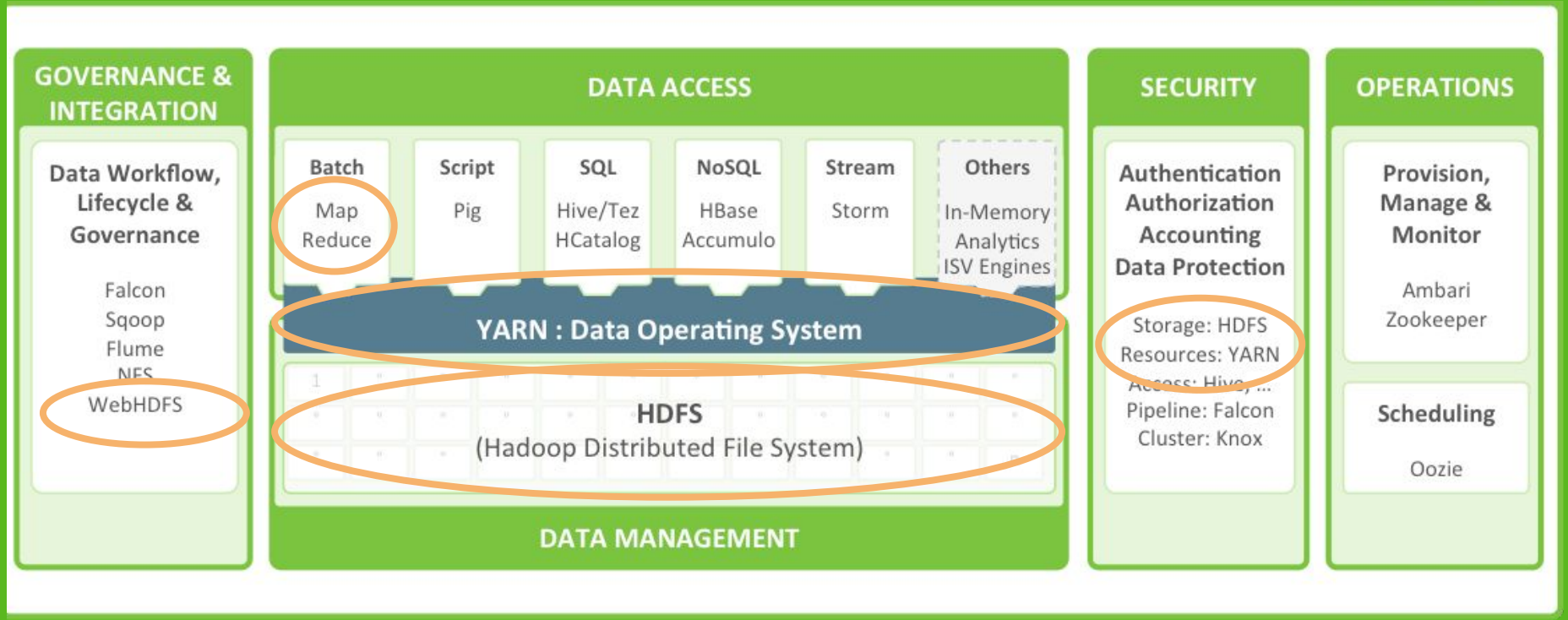


Why Hadoop?



- *Benefits of the Hadoop Architecture*
  - Consolidates Data
  - Integrates with many existing platforms
  - Scalable and Affordable
  - Real-Time Insights

The Hadoop Ecosystem



Hadoop Architecture

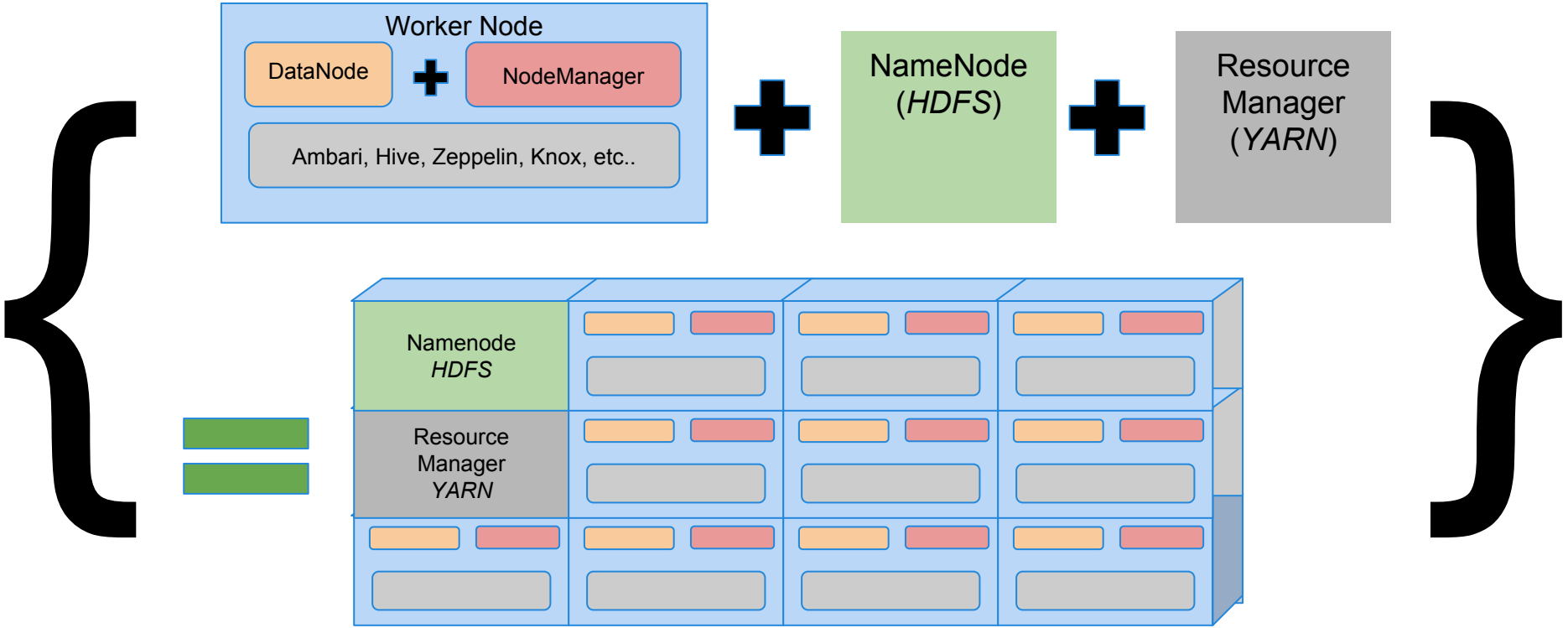
Notable Hadoop Projects

- Apache Kafka → Data Streaming
- Apache HBase → Big Data Management
- Apache Hive → Read and Query from HDFS
- Apache ZooKeeper → HA management
- Apache Spark → Processing Engine
- Apache Ambari → Cluster management

At the Core of Hadoop

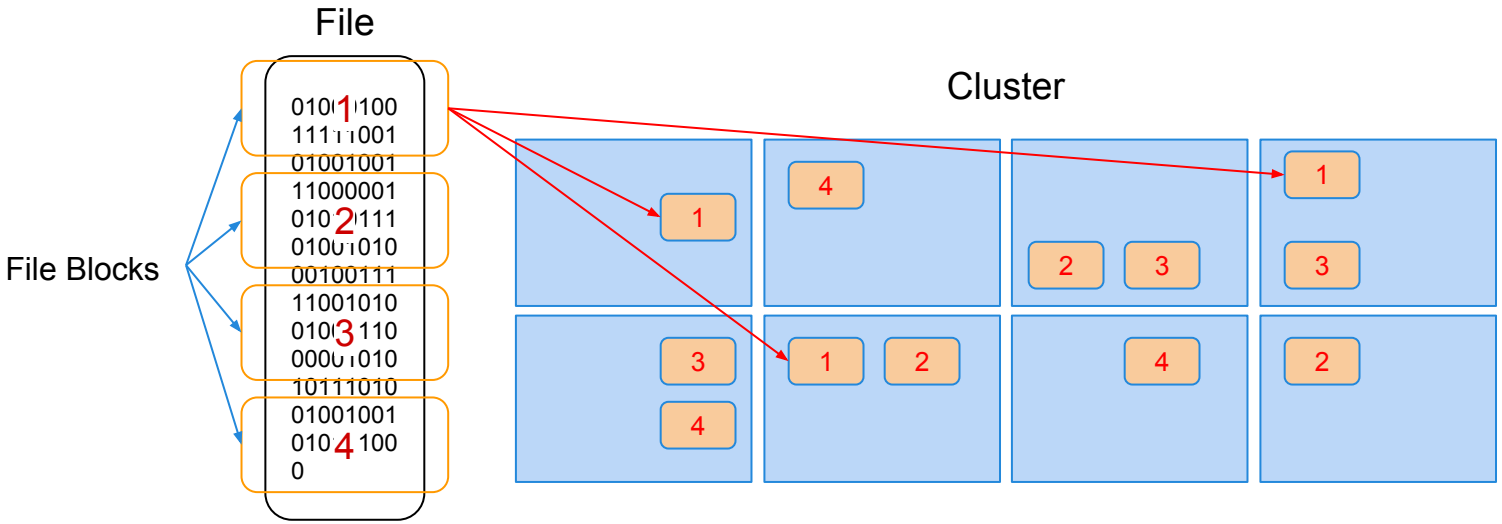
- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce (Processing Engine)
- Hadoop Common (Core Hadoop Libraries)
- Hadoop YARN (Yet Another Resource Manager)
    - CPU/Storage/Memory management (parallel jobs)

Cluster Architecture



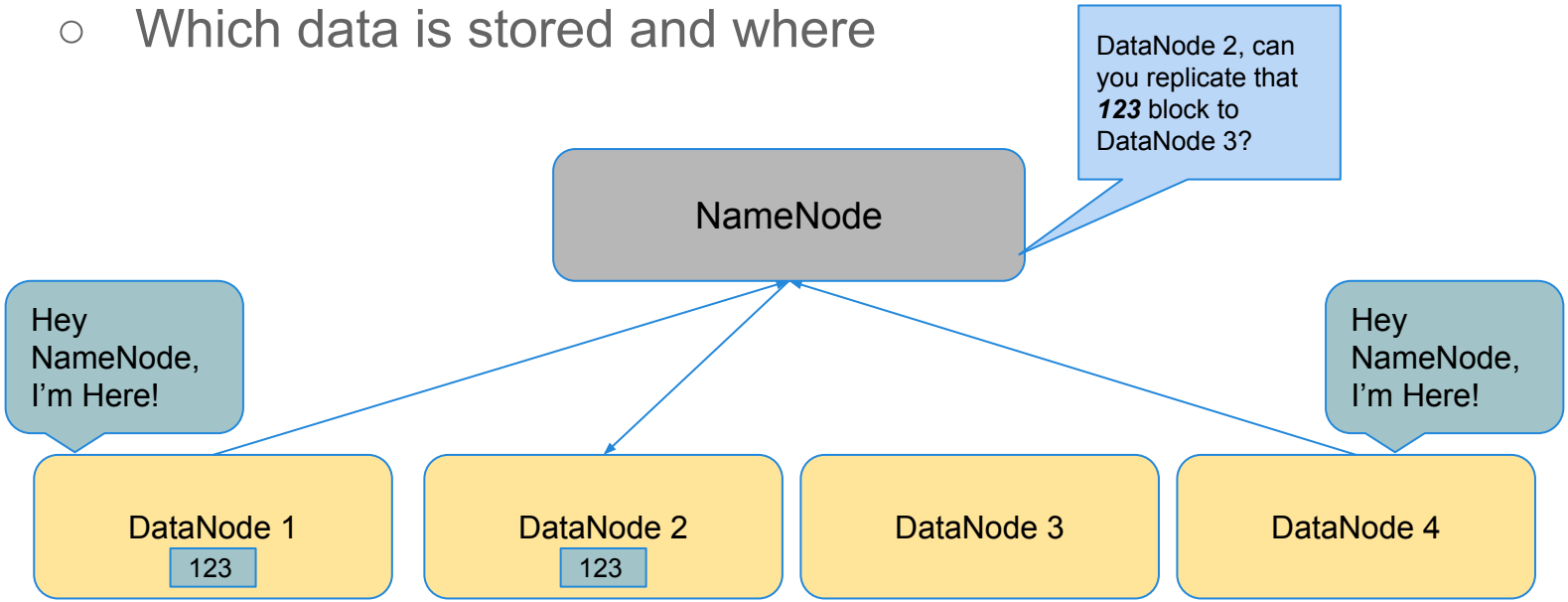
HDFS Architecture

- Fault-tolerant distributed storage
  - Split file into logical blocks
  - Store multiple copies of each block

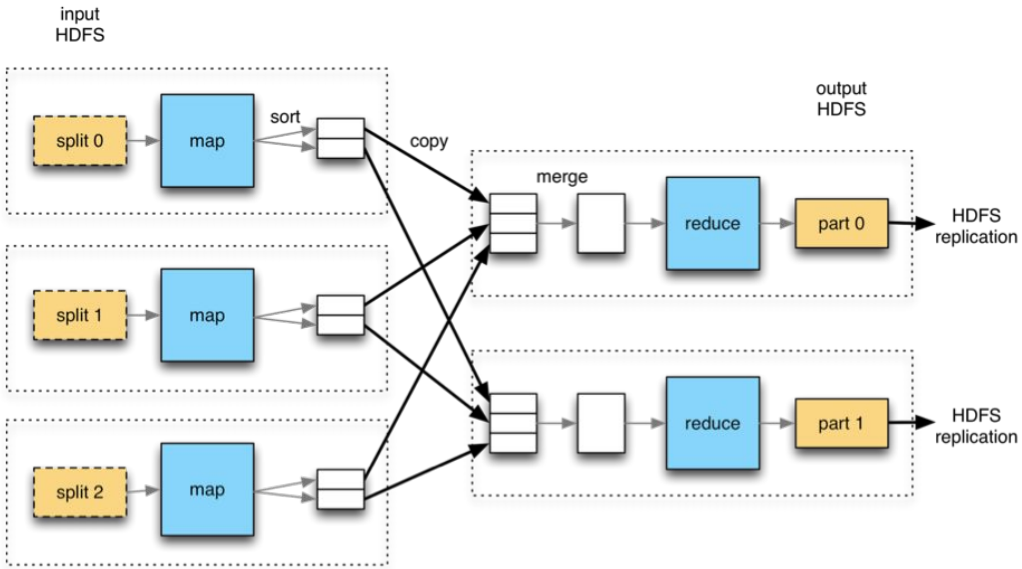


HDFS - Namenode and Heartbeats

- Namenode communicates through Heartbeats
  - Keep track of all the data notes
  - Which data is stored and where



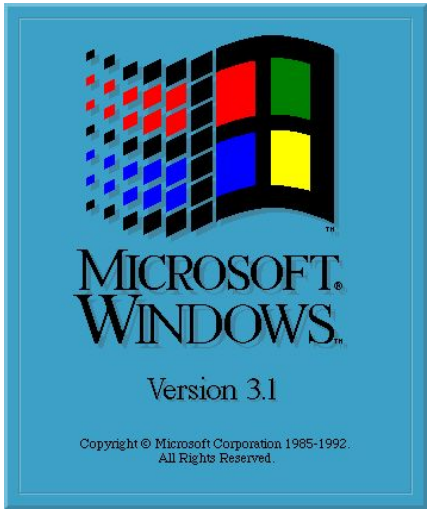
MapReduce in Hadoop



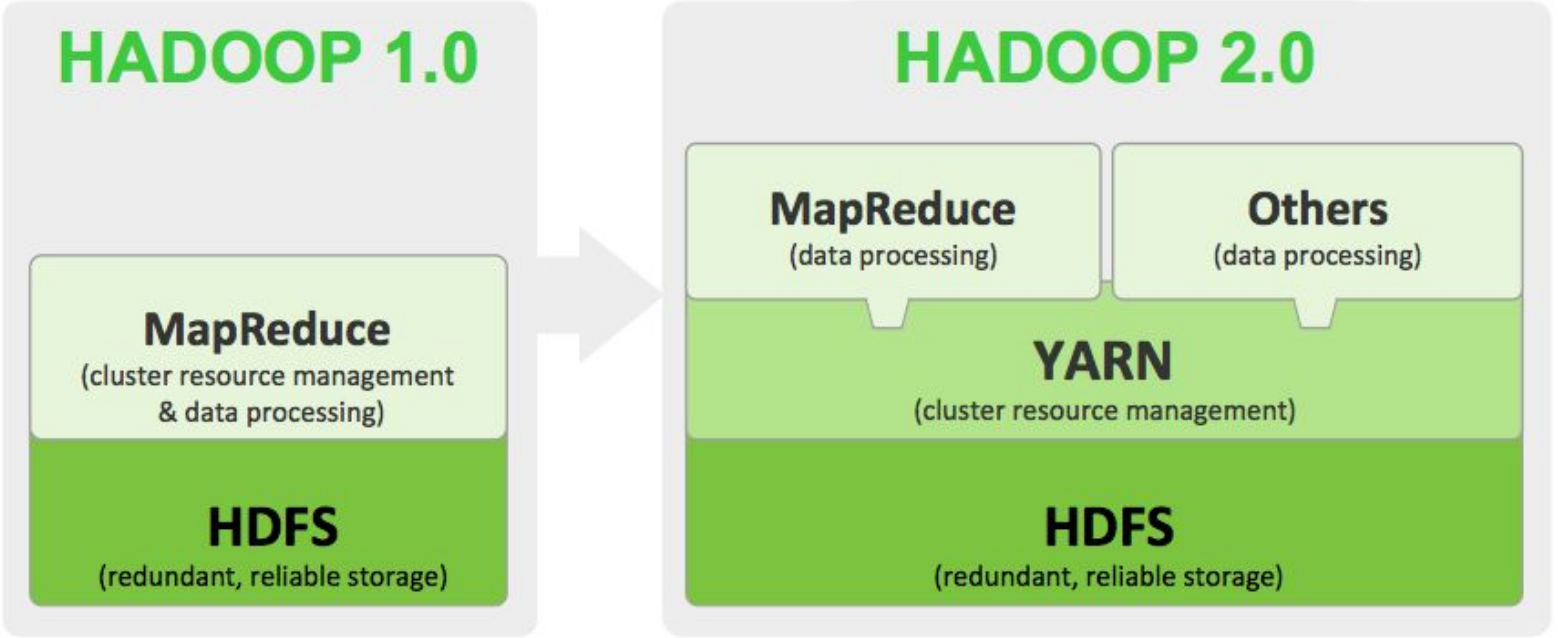
- Shuffle and Sort
  - Break a problem into sub-problems
- Batch Processing

What do iOS 4 and Windows 3.1 have in common?

iOS 4



Multi-Use vs. Batch



Workshop