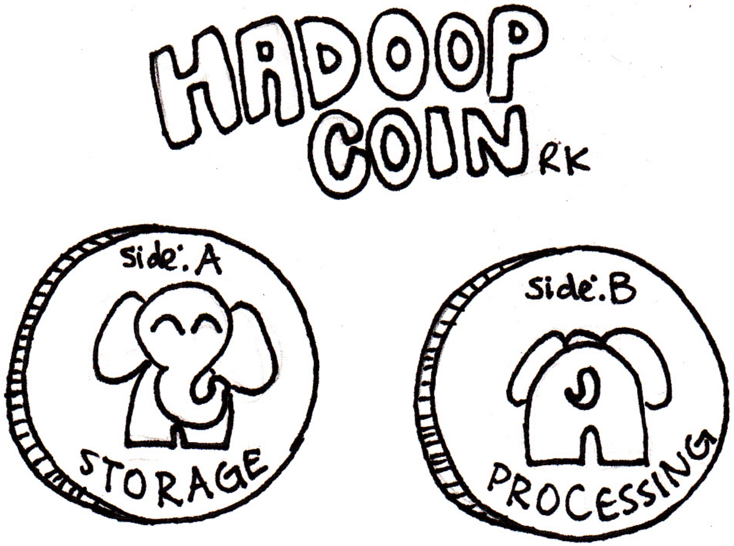
How Hadoop Works

School of Information Studies
Syracuse University

Two Goals of Hadoop

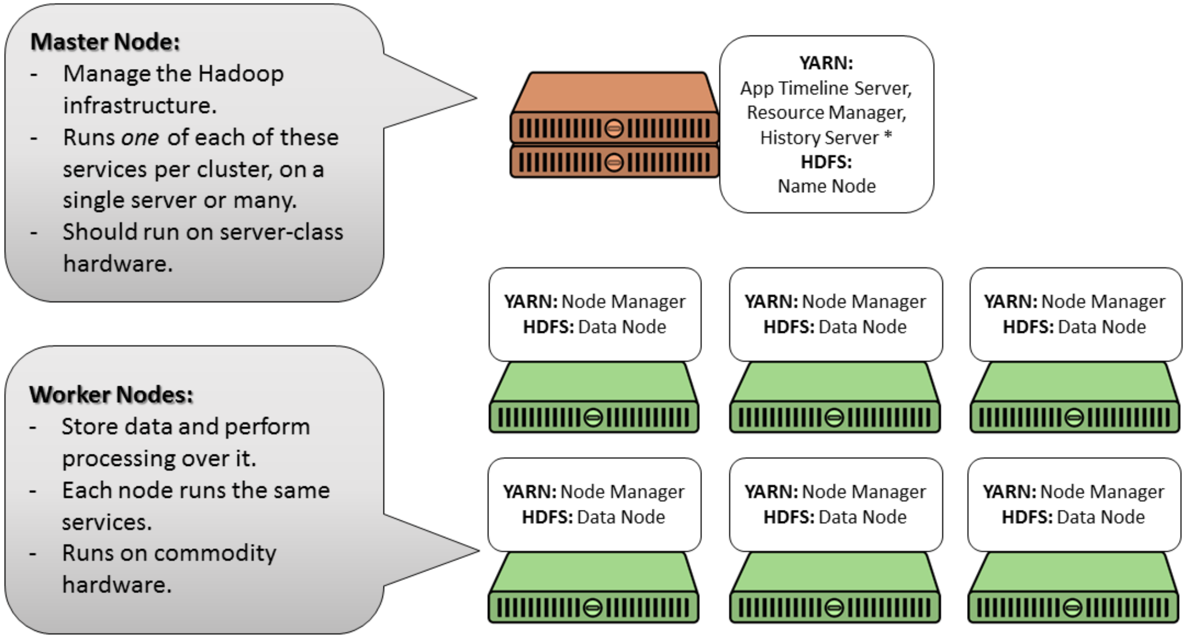
1. Distribute the data.
   HDFS Does this
2. Move processing to the data.
   MapReduce /
   YARN Does this



Hadoop Clusters

3 Node Types:

1. Master Nodes
2. Worker Nodes
3. Client Nodes



Master Node:
- Manage the Hadoop infrastructure.
- Runs *one* of each of these services per cluster, on a single server or many.
- Should run on server-class hardware.

YARN:
App Timeline Server,
Resource Manager,
History Server *
HDFS:
Name Node

Worker Nodes:
- Store data and perform processing over it.
- Each node runs the same services.
- Runs on commodity hardware.

YARN: Node Manager
HDFS: Data Node

YARN: Node Manager
HDFS: Data Node

YARN: Node Manager
HDFS: Data Node

YARN: Node Manager
HDFS: Data Node

YARN: Node Manager
HDFS: Data Node

YARN: Node Manager
HDFS: Data Node

* Map Reduce 2 service on YARN

How Does Hadoop Differ from Relational?

Relational	Hadoop
Schema On Write	Schema On Read
Fast Reads	Fast Writes
Highly Structured Data	Loosely Structured Data
Declarative Data Processing (SQL)	Declarative & Procedural Data Processing
Good For: ACID Transactions, Business Data	Good For: Logs, Data Streams, Unstructured, Data Discovery

What *Exactly* is “Schema on Read?” Again?

Traditional RDBMS

You cannot write data without a table.

Cannot insert data unless data fits into table’s single design.

Large up front design costs.

- Conceptual Models
- Table Design

“Schema on Write”

Hadoop’s HDFS

You write the data “as-is”, to HDFS.

Schema applied when data is read – multiple designs.

Very little up-front design costs

- Just Write to disk
- Apply schema when you need it

“Schema on Read”