Getting Started with HDFS

Understanding HDFS

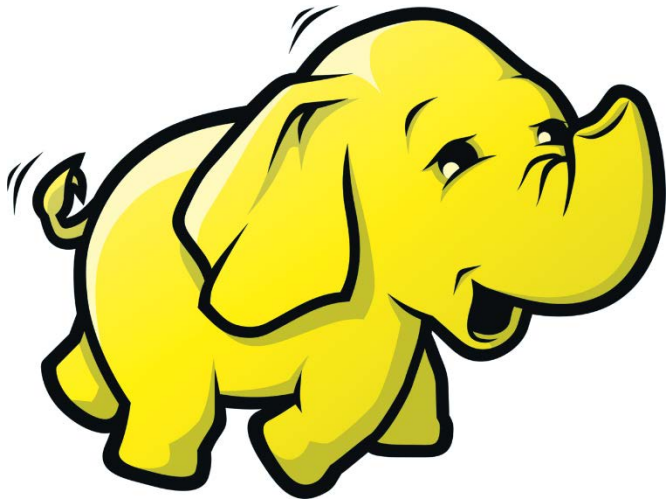
Thomas M. Henson

@henson_tm | www.thomashenson.com

HDFS

Hadoop Distributed File System

Overview



Explain HDFS architecture

Discuss YARN

Learn about fault tolerance

Setup sandbox in Azure

Good to Know

- Basic Linux command line knowledge
- Basic programming knowledge
- Hadoop development environment

HDFS

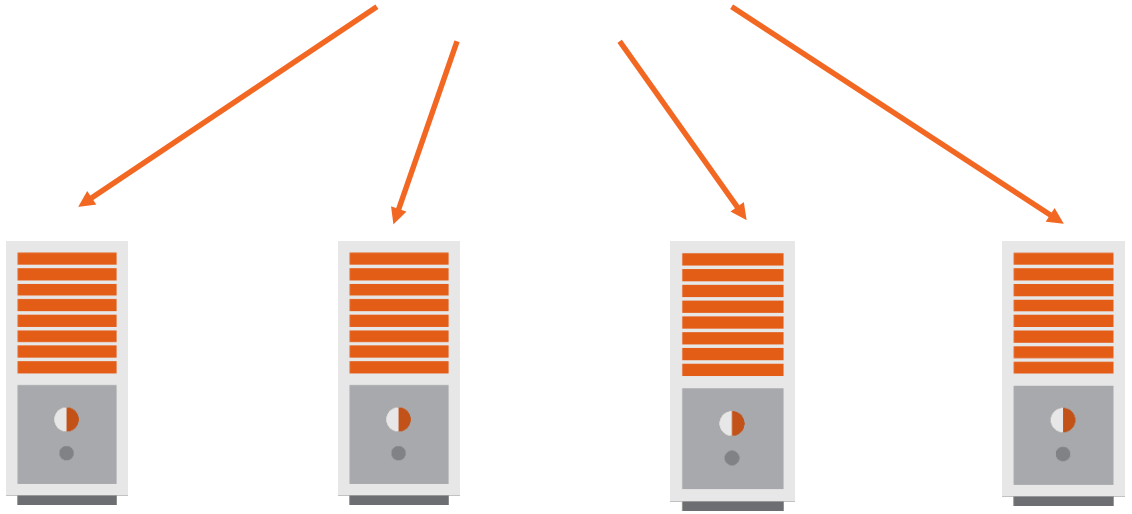
Distributed file system | Batch processing | Written in Java

HDFS consists of a NameNode
and DataNode

Namenode

Remembers where the data is stored in the cluster

Datanode



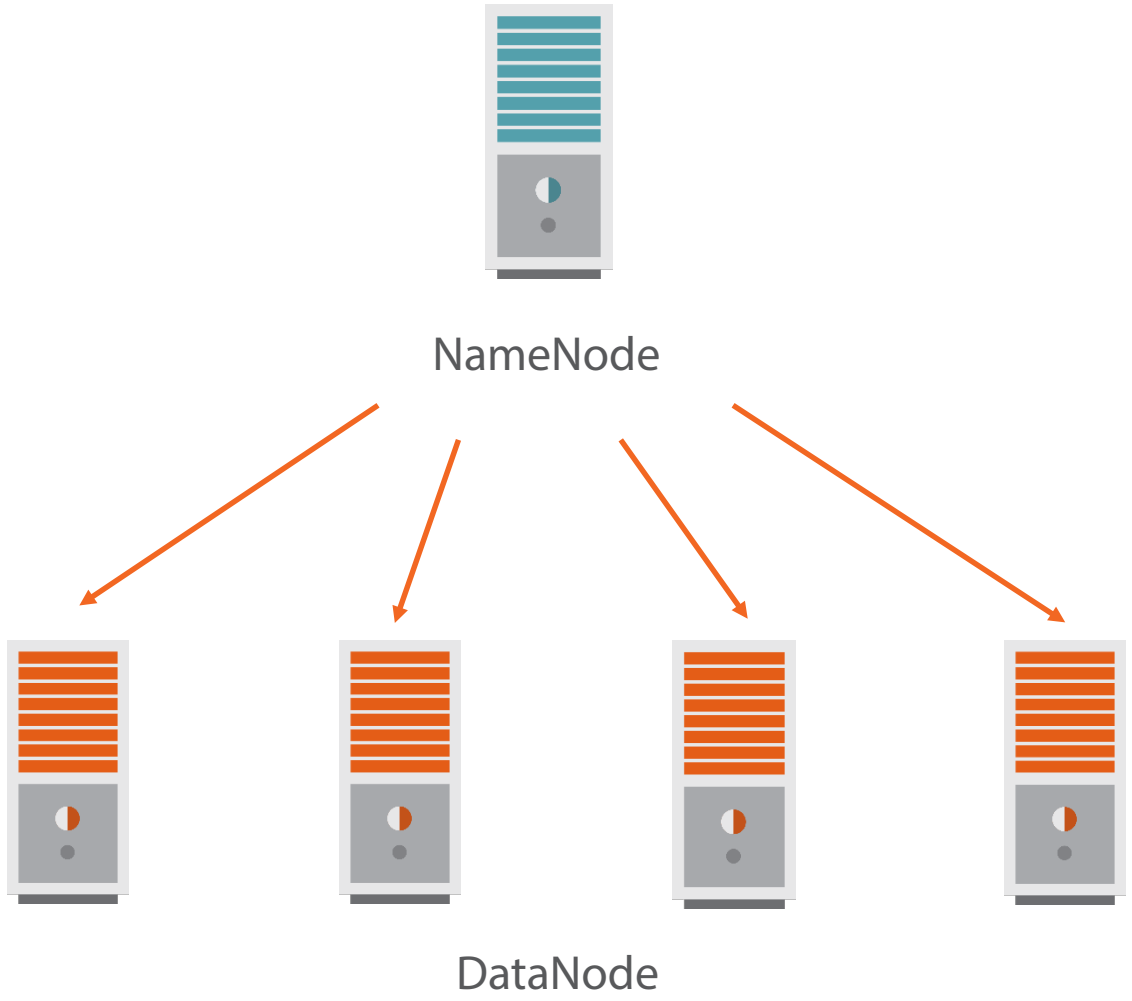
Stores actual data

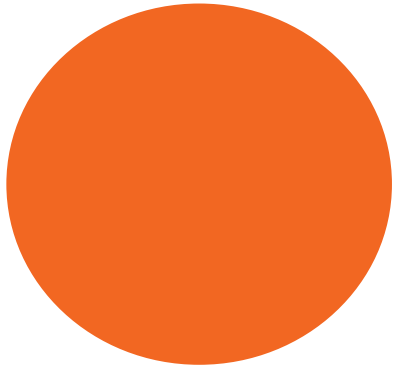
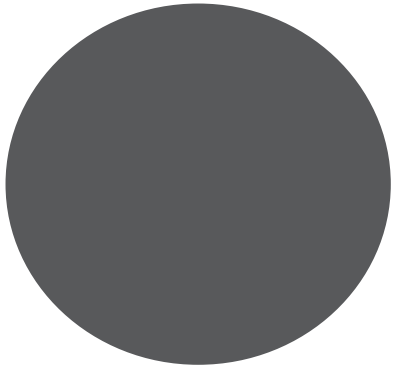
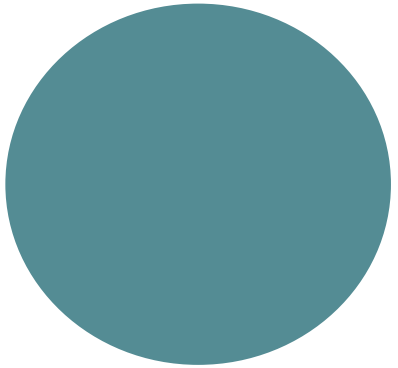
HDFS was built under the premise that hardware **will fail**

Fault Tolerance

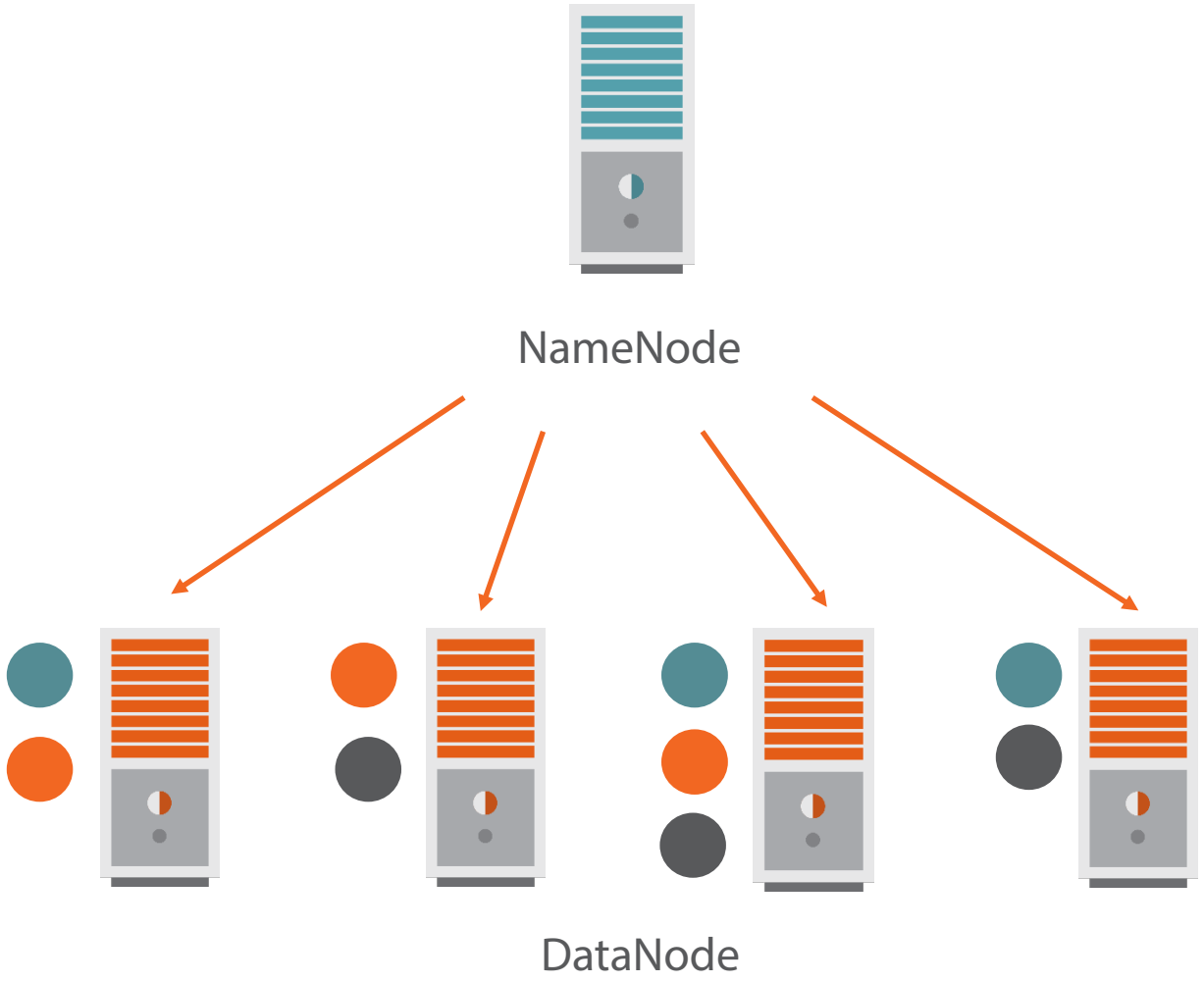
Ensures that when hardware fails, users will still have their data available. Fault tolerance is achieved through storing multiple copies throughout cluster.

Fault Tolerance



Fault Tolerance



Comparing Versions

HDFS 1.0

- NameNode single point of failure
- YARN
- Scalability and performance suffer with larger clusters

HDFS 2.0

- NameNode high availability
- YARN
- Scalability and performance do well in larger clusters

YARN

Split the role of the of the resource manager into global resource manager and application master

YARN

Yet Another Resource Negotiator

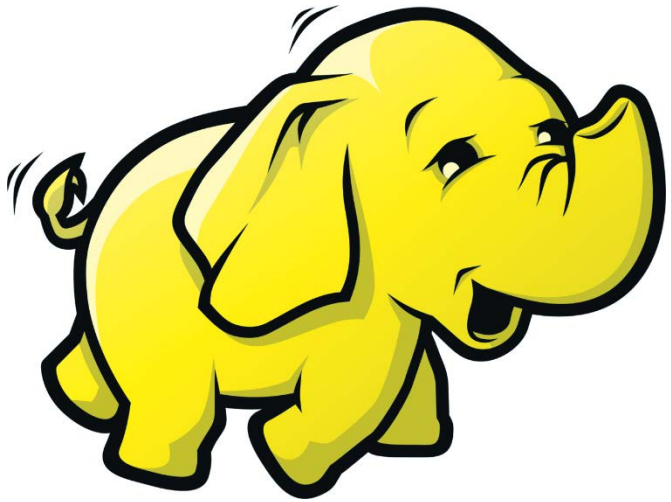
Resource management

MapReduce v2

Hadoop Development Enviroment

Pig Latin: Getting Started

Summary



Know how HDFS architecture works

Understanding of fault tolerance

Compared HDFS 1.0 & 2.0

Development environment ready