ADSA

Welcome to Intro to Hadoop

Presented by: Jerry Chen
Chris Kim
Jackson Davis

What is the biggest problem today?

- We are being flooded with data everyday.
- Ninety percent of all the data collected by humans have been created in the last two years!
- The problem: We have all this data, but traditional tools are poorly equipped to deal with the volume and complexity of this much data.

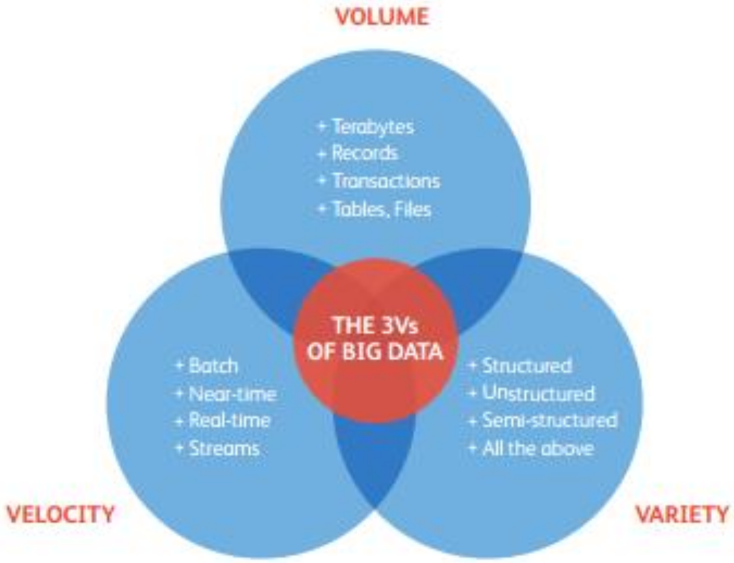
INTERNET SECURITY SHARE SYSTEMS
MANAGE CLOUD RISK STATISTIC
HADOOP
DATA BUSINESS MANAGE SHARE RISK SECURITY SHARE HADOOP
OPEN SOURCE
ANALYTICS SHARE
SETS STORE BUSINESS INTERNET ANALYSIS
BIG DATA
SECURITY STORAGE INTERNET HADOOP DATA
SHARE MANAGE VOLUME HADOOP ENTERPRISE
TREND
CLOUD BUSINESS ANALYTICS SYSTEMS
SOCIAL MEDIA SHARE
HADOOP LEARNING COMPUTING ANALYTICS
ENTERPRISE PROJECTS SERVER ENTERPRISE
BIG DATA
CLOUD MANAGE
SYSTEMS ENTERPRISE
COMPLEX SERVER VOLUME LARGE DATA
SOLVED SOFTWARE OPEN SOURCE BIG DATA
SETS DATA VOLUME SOLUTIONS DATABASE RISK TOOLS TECHNOLOGIES
SOLVED SETS BIG DATA
SOCIAL MEDIA SOFTWARE
SYSTEMS COMPUTING
TREND
HADOOP CLOUD
LEARNING MARKETING TOOLS RISK SYSTEMS SOFTWARE COMPLEX
SYSTEMS COMPLEX BIG SERVER
DATA SECURITY OPEN SOURCE
BIG CLOUD DATA TOOLS BUSINESS
RISK NETWORKS
DATABASE
SOCIAL SOLVED
HADOOP

What is data? The “3 V’s”

1. **Volume** - Large amounts of data ranging from a many terabytes to a few petabytes.

2. **Variety** - Data that is organized in multiple structures.

3. **Velocity** - The speed at which data can be processed. The larger your data set, bigger your velocity challenge.



Relatable Analogy

- Let’s say you’re given two assignments every 5 days
- It takes you to complete one assignment every 5 days
- You continuously receive two more assignments every 5 days
- Eventually there’s so much work that you can’t complete them all in time

This is the Enterprise Approach...

How do we solve this problem?

With data continuously entering the computer, there’s two main approaches:

Enterprise Approach - Big Data is processed by a powerful computer. Eventually, the computer gets to a limit where the amount of data gets too large. Then the work starts piling up.

Hadoop’s Approach - Big Data is broken into multiple pieces. The cluster of computers compute each computation then combines them into a single result.

Back to the analogy...

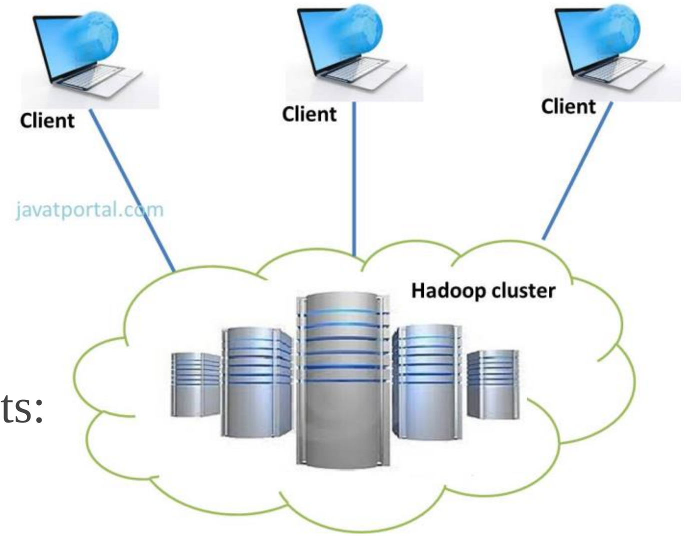
What is Hadoop Exactly?

Hadoop is a linux based set of tools.

Hadoop works on a distributed model with numerous low cost computers.

Each low cost computer, called slaves, has two components: a task tracker and a data node.

The task tracker processes the piece of data given to this particular node.



So, why Hadoop?

By dividing the load, the massive amount of inputted data can be parallel processed by each slave. This significantly speeds up processing time.

Hadoop brings the computation to the data.

Hadoop is highly scalable. It can go up from one computer to thousands and the processing speed is linear to the number of computers.

HDFS (Hadoop Distributed File System)

HDFS is what allows Hadoop to store large amounts of data.

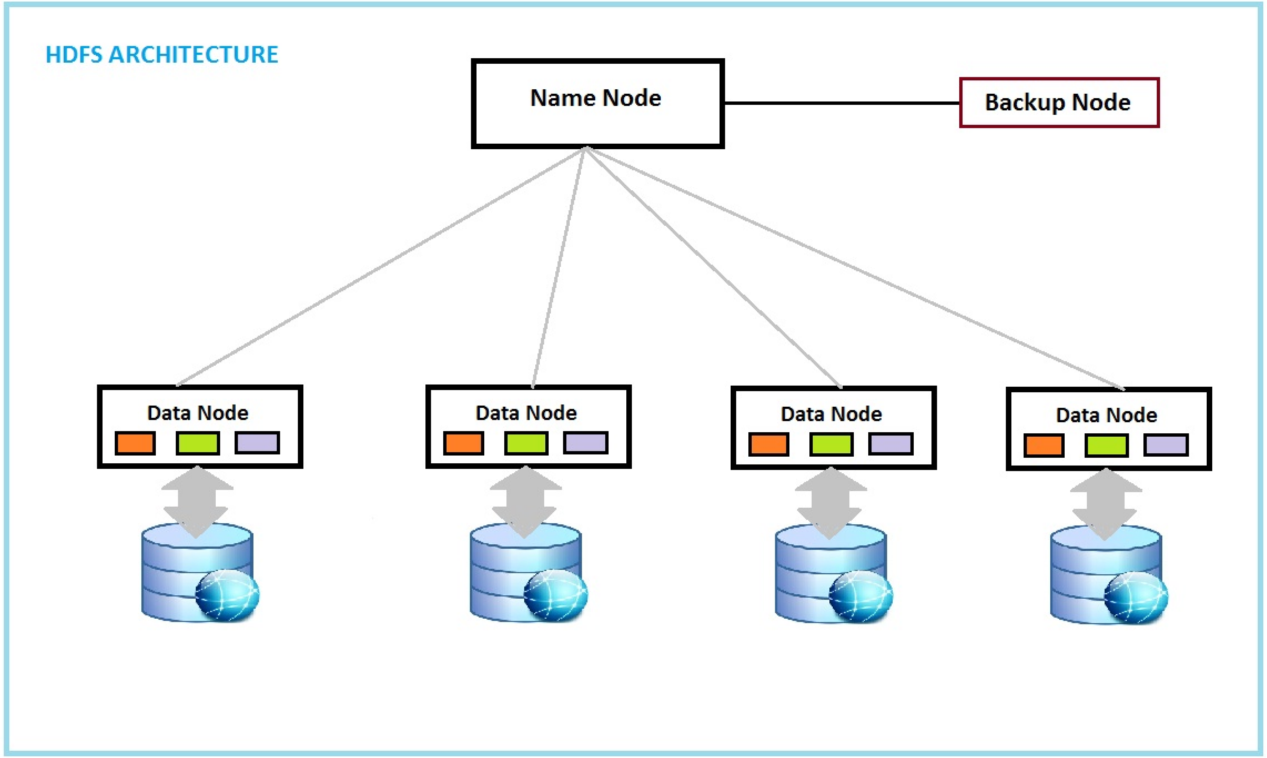
It contains the following

NameNode - This runs on the “master node”. It also tracks and controls storage of the clusters.

DataNode - Runs on the “slave nodes”. It replicates the data files three times and stores it across the cluster.

Client Machine - In charge of uploading data, submitting Mapreduce jobs, and viewing processed data

Part of a Bigger Picture



Hadoop Map Reduce (YARN)

Map Phase: Input data is copied and split into a large number of fragments with map tasks. Taken care of by Hadoop’s ecosystem.

Distributing: Map tasks are distributed across cluster of computer units.

Processing: Each unit uses map tasks to process data and create immediate key-value pairs.

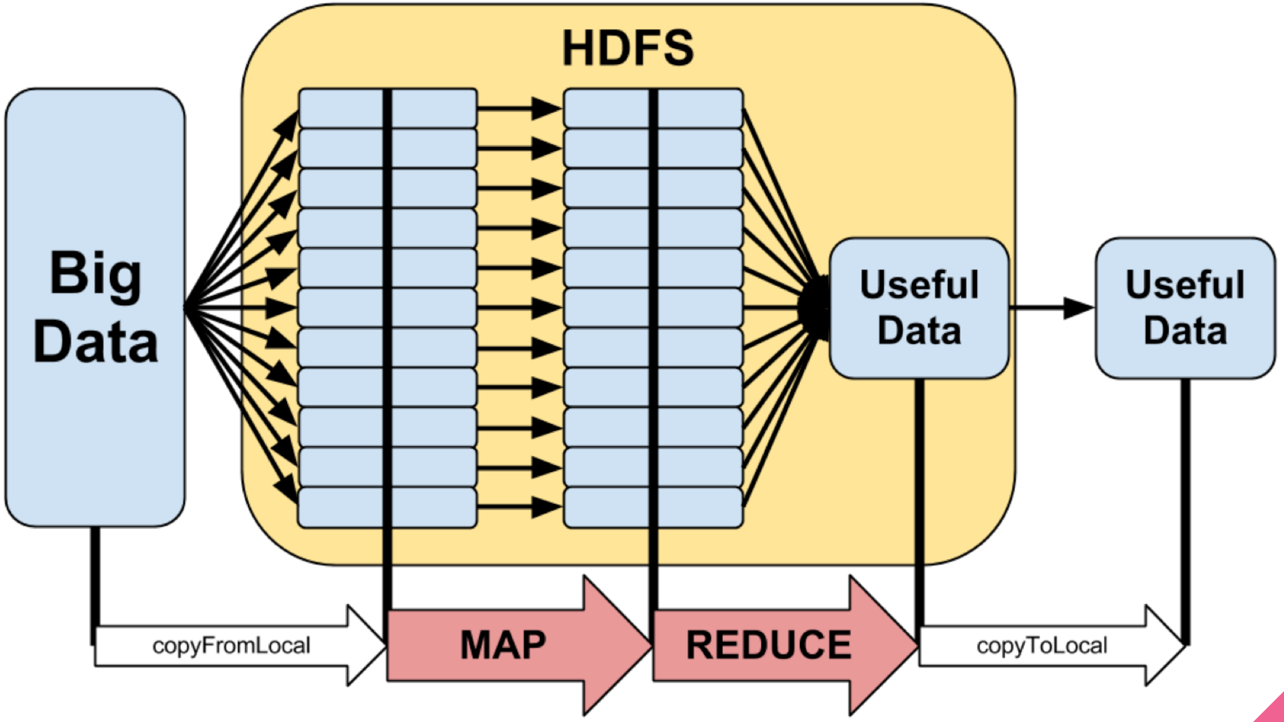
Sorting: The pairs are sorted by keys and partitioned into fragments

Reduce Phase: Each unit reduces the data fragments and creates an output key-value pair

Distributing: Reduce tasks are distributed across the cluster.

Processing: The reduce tasks write the outputs to HDFS when finished

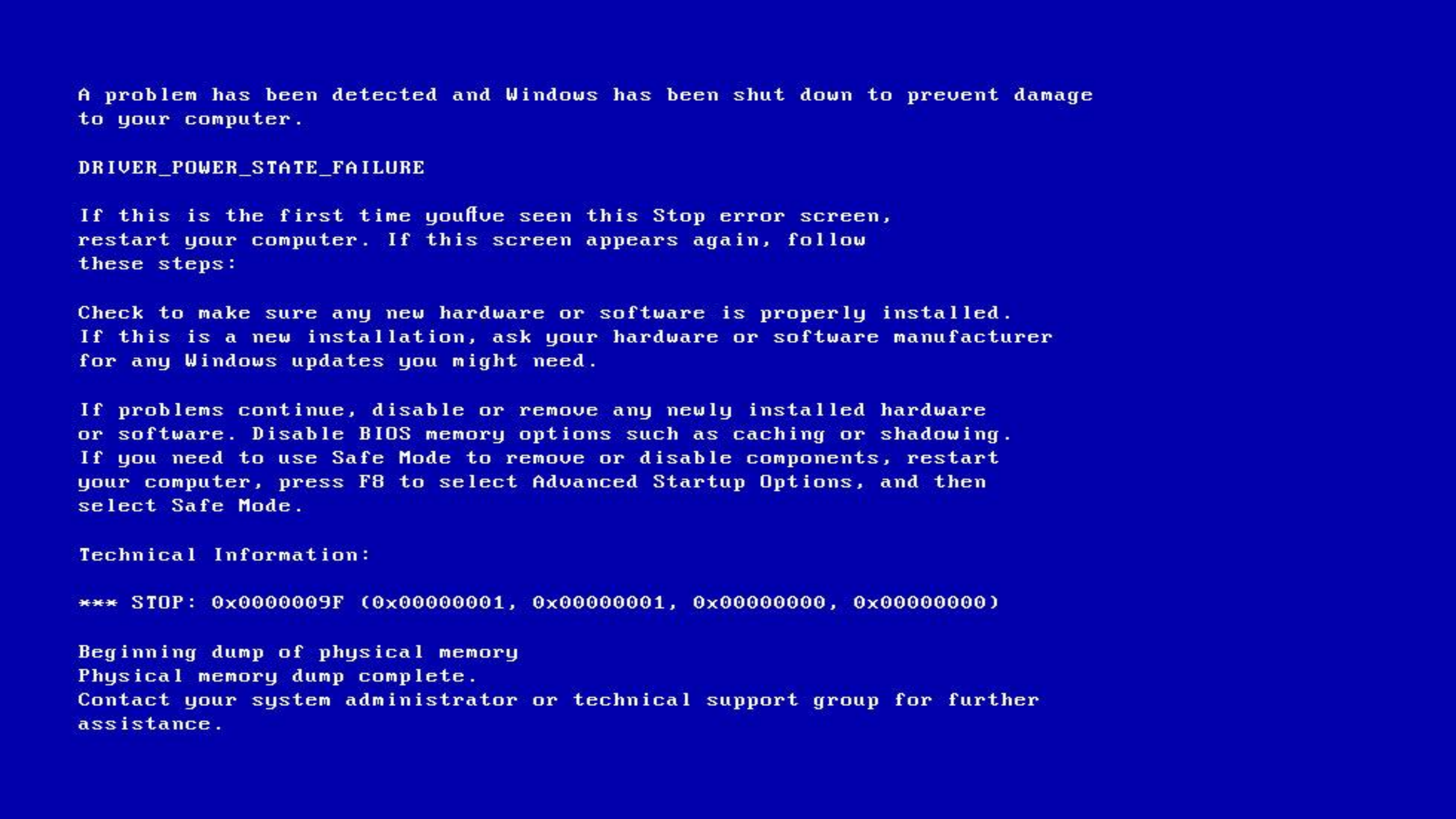
Visualizing the Process



Different Flavors of Hadoop

1. Hbase - Designed to operate on top of the Hadoop distributed file system (HDFS) for scalability, fault tolerance, and high availability.
    a. Written in Java
    b. Open source
2. Hive - data warehousing application in Hadoop.
    a. Written in HQL (similar to SQL)
    b. Created by facebook, but is now open source
3. Pig - large-scale data processing system
    a. Written in Pig Latin (similar to Perl)
    b. Created by Yahoo!, but is now open source

Common Idea: Provide high level languages that can be used to for large data processing



```
A problem has been detected and Windows has been shut down to prevent damage
to your computer.

DRIVER_POWER_STATE_FAILURE

If this is the first time youﬂve seen this Stop error screen,
restart your computer. If this screen appears again, follow
these steps:

Check to make sure any new hardware or software is properly installed.
If this is a new installation, ask your hardware or software manufacturer
for any Windows updates you might need.

If problems continue, disable or remove any newly installed hardware
or software. Disable BIOS memory options such as caching or shadowing.
If you need to use Safe Mode to remove or disable components, restart
your computer, press F8 to select Advanced Startup Options, and then
select Safe Mode.

Technical Information:

*** STOP: 0x0000009F (0x00000001, 0x00000001, 0x00000000, 0x00000000)

Beginning dump of physical memory
Physical memory dump complete.
Contact your system administrator or technical support group for further
assistance.
```

Fail Safe Management

To counteract different failures…

Slave Name Node: Hadoop keeps multiple copies of each file and scattered amongst other nodes. When the node is fixed, Hadoop copies the data from the other nodes.

Slave Task Tracker: Master computer tracks the slave's failure and contact other computers to complete the job. All data is backed up in the master's name node

Master Computer: If the master computer fails, there's a second computer called the backup master.

The system must continue against all odds.

Realistic Application for Hadoop

Logs - commonly referred data about what webpages people have visited and in which order they have visited them.

Logs are tedious to analyze because of its volume.

With Hadoop, you can quickly analyze all this data.

Example, when you’re on amazon

- Relevant items - what items are typically bought with other items
- If you put something into your cart and you decide you do not want to purchase - final page visited, list in shopping cart, state of transaction

All of this data can be useful when it is collected.

Realistic Application for Hadoop cont.

Credit card fraud

- What you usually buy
- Where you usually buy
- How much you usually spend

This data can be used to determine if a purchase on your card was fraud

Applications

Risk Management: Determine which patients have the highest risk of dying at a hospital, for priority treatment.

Company Image Analysis: Mine social media conversations to analyze what people think about your company.

Where is Hadoop used?

Companies that heavily used it:

Yahoo - Security, they had a presentation a couple weeks back how they use Hadoop for advertisement click fraud or real by detecting uncommon click patterns.

Facebook - Users login data, and various posts across the domain.

Amazon - Sorting merchandise and maintaining purchases from customers

How to Get Started

Cluster Setup Download

http://hortonworks.com/products/sandbox/

VirtualMachine Download (v. 16.04.1 LTS)

https://www.ubuntu.com/download/desktop

After installation, click on new

Set any name, pick type: linux, version: ubuntu (64bit)

Content/Resources/Reference

https://www.youtube.com/watch?v=xWgdny19yQ4, https://www.youtube.com/watch?v=Pq3OyQO-l3E, https://www.youtube.com/watch?v=DLutRT6K2rM (Introductory Concept Videos)

http://www.glennklockwood.com/data-intensive/hadoop/overview.html (Overview Map Reduce & Hadoop)

https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html (Hadoop Tutorial)

http://www.plottingsuccess.com/hadoop-101-important-terms-explained-0314/

Hadoop For Dummies (book)