CSC 735 – Data Analytics

INTRODUCTION

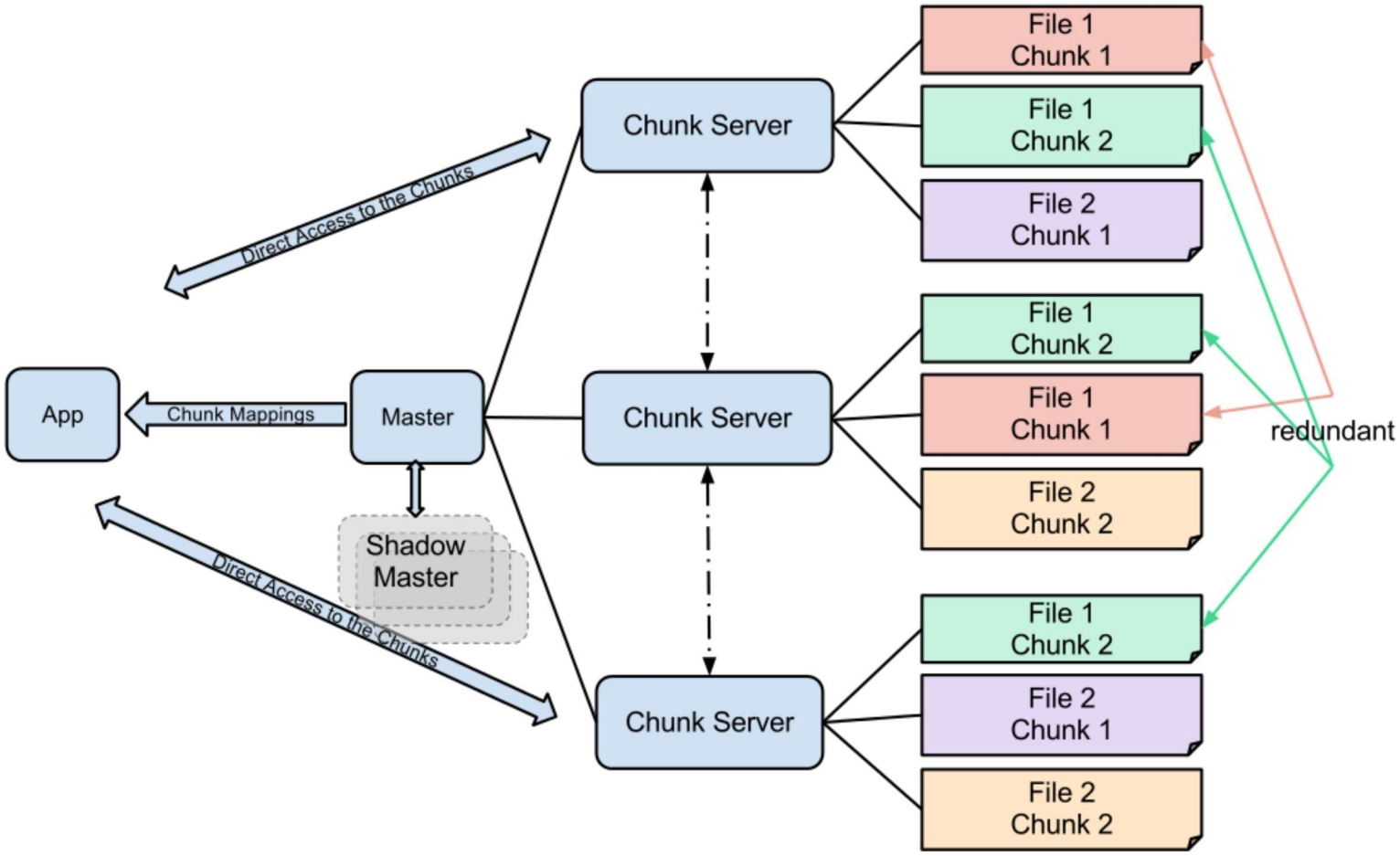
Brief History of Big Data

In early 2000s, search engine providers faced the challenge of Internet Scale Problems

Google and Yahoo! worked on possible solutions

In 2003, Google released a whitepaper titled "The Google File System" (GFS)

Google File System (GFS)



Brief History of Big Data (cont.)

In 2004, Google released another whitepaper, titled "MapReduce: Simplified Data Processing on Large Clusters."

These white papers inspired **Doug Cutting** and **Mike Cafarella** to develop **Hadoop**

What is Hadoop?

It is an Apache open source big data platforms

Processes large datasets across a cluster of commodity computers

It is written in Java

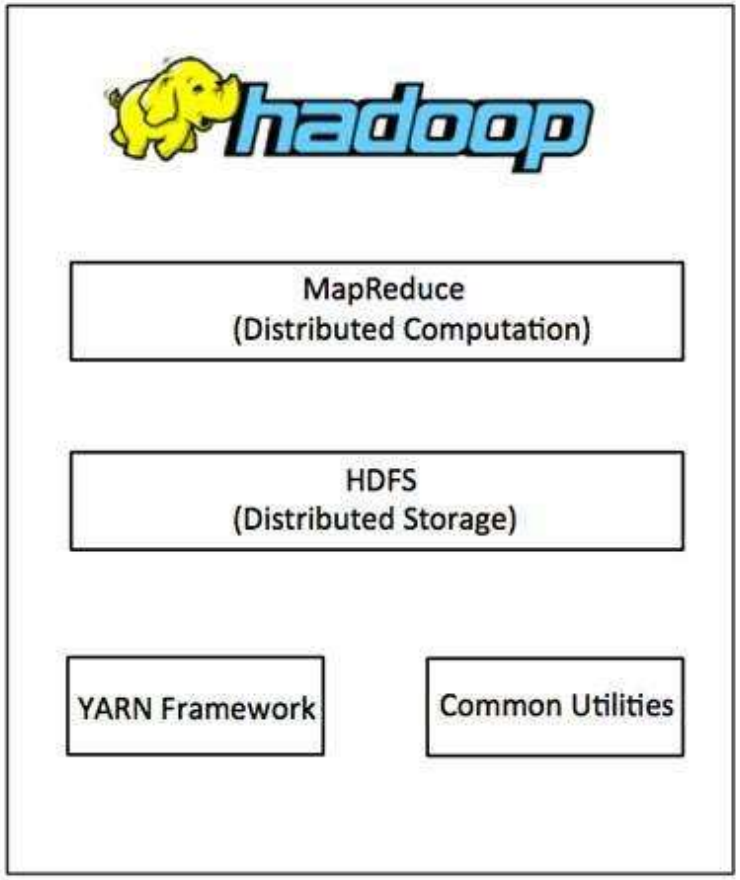
Scalable

Fault-tolerant

Hadoop Architecture

1. Hadoop Common: Java files and libraries necessary to start Hadoop and for supporting the other Hadoop modules
2. Hadoop Distributed File System (HDFS): distributed storage system with ideas similar to GFS
3. Hadoop YARN: A framework for job scheduling and cluster resource management
4. Hadoop MapReduce: A framework for parallel processing of large data sets.

Hadoop Architecture



Big Data

Big data is essential for many organizations

Big data is growing exponentially

Units of Storage

Unit	Equivalent
1 **kilobyte** (KB)	1,024 **bytes**
1 **megabyte** (MB)	1,048,576 **bytes**
1 **gigabyte** (GB)	1,073,741,824 **bytes**
1 **terabyte** (TB)	1,099,511,627,776 **bytes**
1 **petabytes** (PB)	1,024 TB
1 **exabyte** (EB)	1,024 PB
1 **zettabyte** (ZB)	1,024 EB

Rate of Data Growth

in 2017:

2.5 quintillion bytes of data created each day

90% of the data was generate in the last two years

Google processes more than 40,000 searches EVERY second (3.5 billion searches per day)!

77% of searches are conducted on Google

About 5 billion searches a day

Rate of Data Growth (cont.)

Every minute (2017):

- Facebook users post 510,000 comments
- 456,000 tweets on Twitter
- 46,740 photos on Instagram
- Users watch 4,146,600 YouTube videos
- 527,760 photos shared on Snapchat

Definition of Big Data

People define it in different ways

- one definition relates to the volume of data
- another definition relates to the richness of data
- another definition is “too big” by traditional standards

Characteristics of Big Data

Three Vs of big data

- Volume
- Velocity
- Variety

Characteristics of Big Data

Three Vs of big data

- Volume
- Velocity
- Variety

4th V

- Veracity

Characteristics of Big Data

Three Vs of big data

- Volume
- Velocity
- Variety

4th V

- Veracity

5th V

- Value

What is Apache Spark?

Unified computing engine and set of libraries for parallel data processing on computer clusters

Open source engine for big data

It support languages: Scala, Python, Java, and R

Has APIs for multiple analytics tasks such as: SQL, streaming, machine learning

It can run on a laptop and on a cluster of thousands of computers

Brief History

Research project at UC Berkeley AMPlab in 2009

Brief History

Research project at UC Berkeley AMPlab in 2009

Motivation

Brief History

Research project at UC Berkeley AMPlab in 2009

Motivation

Initially batch applications

Brief History

Research project at UC Berkeley AMPlab in 2009

Motivation

Initially batch applications

Then allowed interactive analysis and SQL queries

Brief History

Research project at UC Berkeley AMPlab in 2009

Motivation

Initially batch applications

Then allowed interactive analysis and SQL queries

More APIs added over time: MLlib, Streaming, GraphX

Brief History

Research project at UC Berkeley AMPlab in 2009

Motivation

Initially batch applications

Then allowed interactive analysis and SQL queries

More APIs added over time: MLlib, Streaming, GraphX

In 2013, project contributed as open-source vender-independent to Apache Software Foundation

Brief History

Research project at UC Berkeley AMPlab in 2009

Motivation

Initially batch applications

Then allowed interactive analysis and SQL queries

More APIs added over time: MLlib, Streaming, GraphX

In 2013, project contributed as open-source vender-independent to Apache Software Foundation

Databricks

Brief History

Research project at UC Berkeley AMPlab in 2009

Motivation

Initially batch applications

Then allowed interactive analysis and SQL queries

More APIs added over time: MLlib, Streaming, GraphX

In 2013, project contributed as open-source vender-independent to Apache Software Foundation

Databricks

Spark 1.0 in 2014 and 2.0 in 2016

Hadoop MapReduce vs Spark

Spark is faster

- Hadoop stores data on disk
- Spark keeps as much data in memory as possible

Hadoop MapReduce vs Spark

Spark is faster

- Hadoop stores data on disk
- Spark keeps as much data in memory as possible

Spark provides much more functionality

- Hadoop only uses Map & Reduce
- Spark uses most functional programming

Hadoop MapReduce vs Spark

Spark is faster

- Hadoop stores data on disk
- Spark keeps as much data in memory as possible

Spark provides much more functionality

- Hadoop only uses Map & Reduce
- Spark uses most functional programming

Spark

- Easier to use
- REPL & interactive environment

Why Scala?

A language such as R or MATLAB does not scale with Scala, it's easier to scale your problem to large datasets

Distributed computations in Spark are simple to write in Scala

Running Spark

You can download and install Spark on your computer

- All you need is **java** installed on your system PATH

- YouTube Video- Installing Apache Spark and Scala on Windows

Databrick's Community Edition: free **cloud** environment for learning Spark

- Create an account

Launching Spark's Interactive Consoles

Launching the Python console

- From Spark's home directory, run

```
.\bin\pyspark
```

Launching the Scala console

- From Spark's home directory, run

```
.\bin\spark-shell
```

Using Databricks Community Edition

Hassle free environment for using Spark

Has all the data used by our book

Provides a notebook experience for using Spark

Basic overview

Book's GitHub page