# Course Outline

- http://www.johnabbott.qc.ca/continuing-education/specialized-it/emploi-quebec/management-and-treatment-of-big-data/

- Email: kantesariyashyam@gmail.com

- LinkedIn: https://www.linkedin.com/in/kantesariyashyam/

# Data vs Information

- **Data** :

  Simply fact or figure

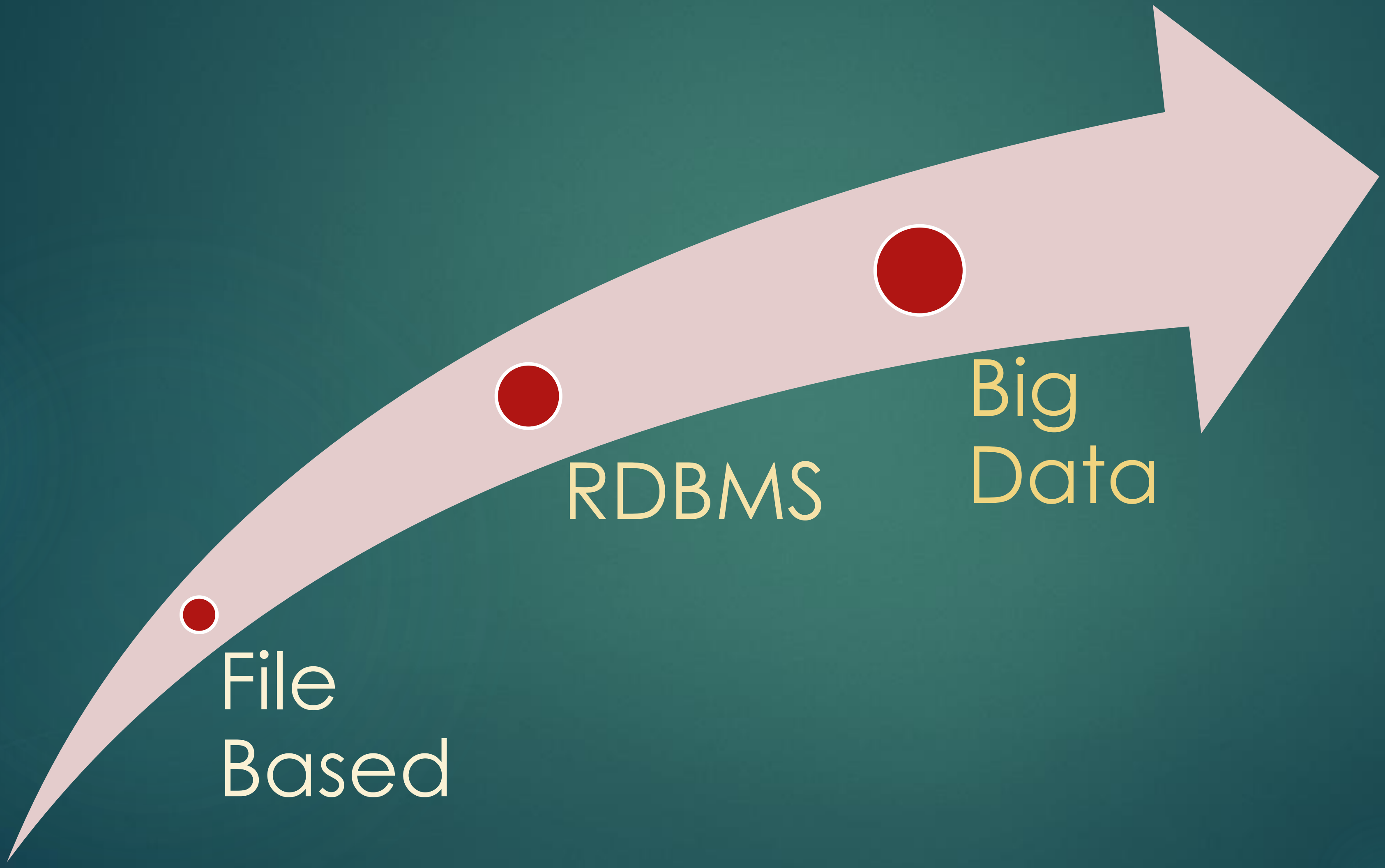
  For example: a number -10

- **Information**:

  Context + data

  For example: -10 degree centigrade is the temperature of Montreal on 27th Jan 2018 at 10:35 AM.

# Evolution in Data management



File Based

RDBMS

Big Data

# What's Big Data?

- International Data Corporation (IDC) has measured data footprint in 2013: 4.4 zettabytes
- 1 zettabyte = 1 billion terabytes
- Forecast is to have 44 zettabytes by 2020
- Where does this data come from?

Ref: Hadoop definitive guide 4th edition, Oreilly publications

# Characteristics of Big Data

- Volume
- Velocity
- Variety
- Value

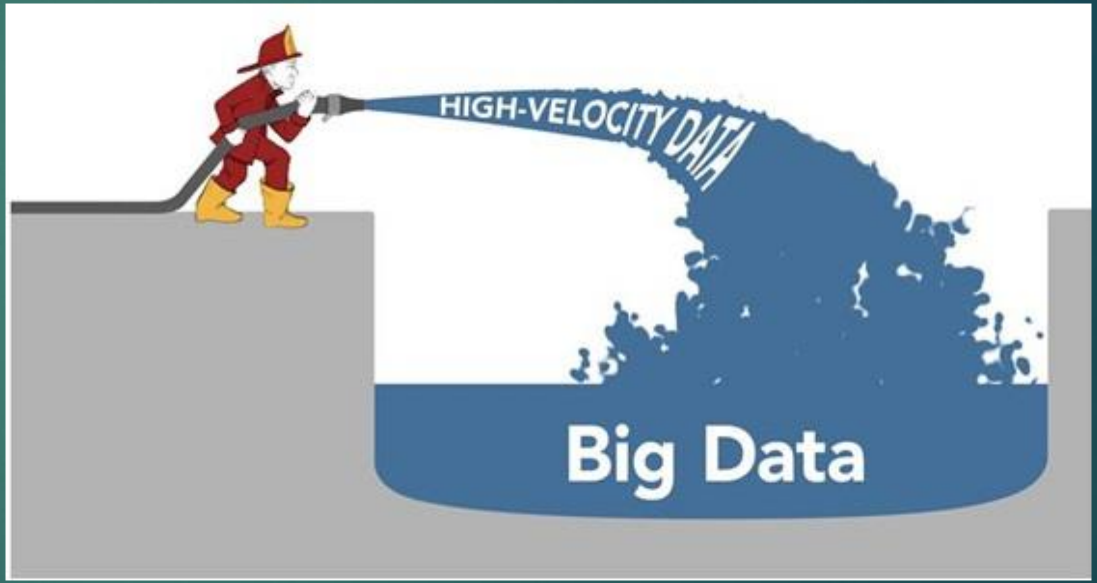
# Characteristic: Volume

- Any guess how much amount of data we are producing within this room?
- Connected smart cars will generate 25GB data per hour

Ref: https://qz.com/344466/connected-cars-will-send-25-gigabytes-of-data-to-the-cloud-every-hour/

# Characteristic: Velocity

- What happens in an internet second
  - 54,907 Google searches
  - 7,252 tweets
  - 125,406 YouTube videos
  - 2,501,018 emails sent



Ref: http://www.dailymail.co.uk/sciencetech/article-3662925/What-happens-internet-second-54-907-Google-searches-7-252-tweets-125-406-YouTube-video-views-2-501-018-emails-sent.html#ixzz4sNJmz06e

# Characteristic: Variety

- Structured
- Semi structured
- Unstructured
- XML
- Json
- Web logs
- Sensor data

# Characteristic: Value

# Applications

- Finance
- Pharma
- Retail
- Manufacturing
- Insurance
- Travel industry

# Environment set up

- Intellij Idea

https://www.jetbrains.com/idea/download/#section=windows

- Git bash

https://git-scm.com/downloads

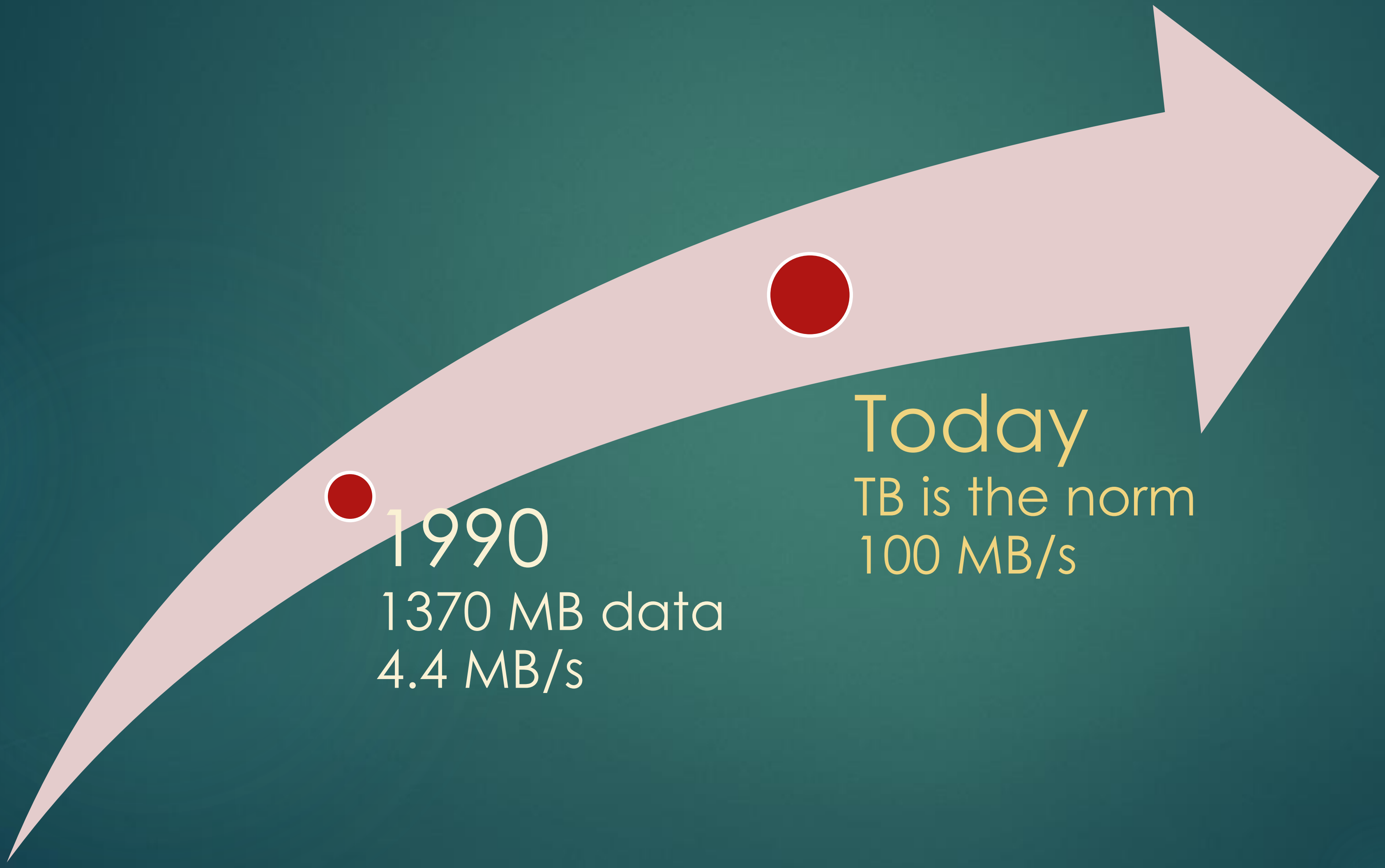
# What is next?

- The good news is “We have big data to analyze”
- But the challenge is “How to store and process it”

# What’s the solution

- Build a bigger system with increased computing power
- “In pioneer days they used oxen for heavy pulling, and when one ox couldn’t budge a log, they didn’t try to grow a larger ox. We shouldn’t be trying for bigger computers, but for more systems of computers” – Grace Hopper

# Storage Technology



1990

1370 MB data

4.4 MB/s

Today

TB is the norm

100 MB/s

# Grid computing

- Based on Message Passing Interface (MPI)
- Uses shared filesystem
- Programmer has to think at task level as opposed to data level
- Missing abstraction of fault tolerance

# MPI program example

- **Serial**

```
#include <stdio.h>
int main(){
    printf (“Hello World!");
    return 0;
}
```

- **Parallel**

```
#include <stdio.h>
#include <mpi.h>
int main(){
    MPI_Init (NULL, NULL);
    printf (“Hello World!");
    MPI_Finalize();
    return 0;
}
```

# Volunteer computing

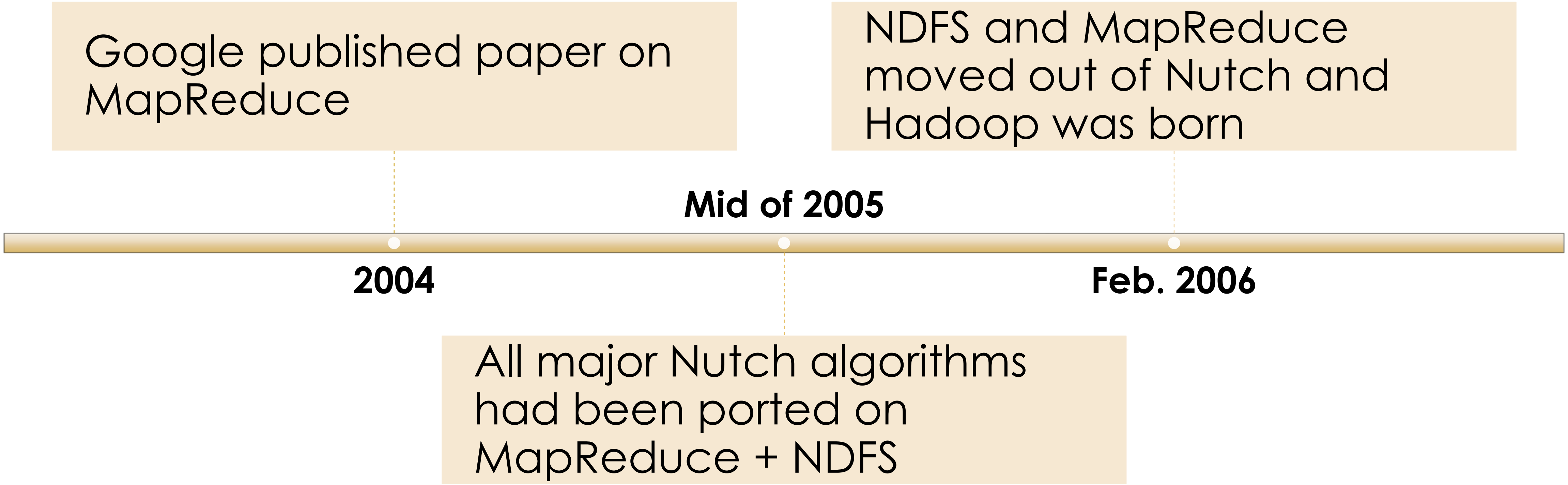
- System is highly compute intensive
- Small amount of data on remote machine
- Low bandwidth
- Based on Internet

# Distributed Computing

# History of Hadoop: Origin

- Origin: Apache Nutch - Open source web search engine
- Cost: 0.5 million $ hardware and 30,000$ running cost to support one billion page index
- Nutch started in 2002 and was ready to crawl and search quickly
- Challenge: Scale to billions of web pages

# History of Hadoop: Hadoop born



Google published paper on MapReduce

2004

Mid of 2005

All major Nutch algorithms had been ported on MapReduce + NDFS

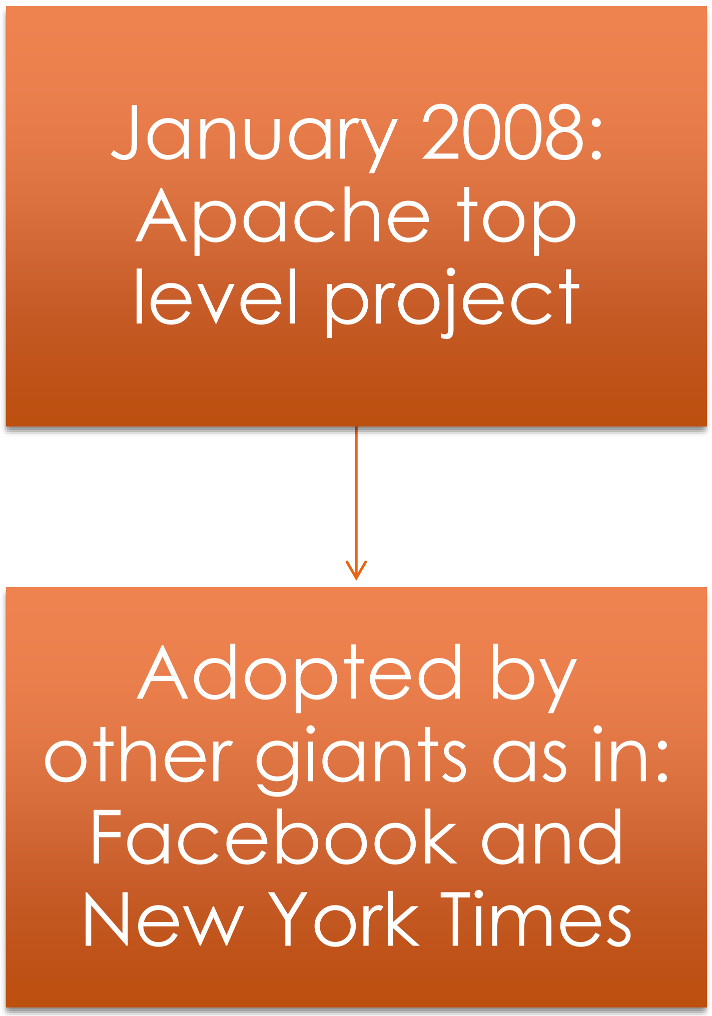
Feb. 2006

NDFS and MapReduce moved out of Nutch and Hadoop was born

# History of Hadoop: Hadoop at Yahoo

- Dreadnaught: System to build WebMap
- Started new project in C++ based on GFS and MapReduce
- January 2006: Daug Cutting joined Yahoo!
- Set-up 200 node cluster to accelerate Hadoop project

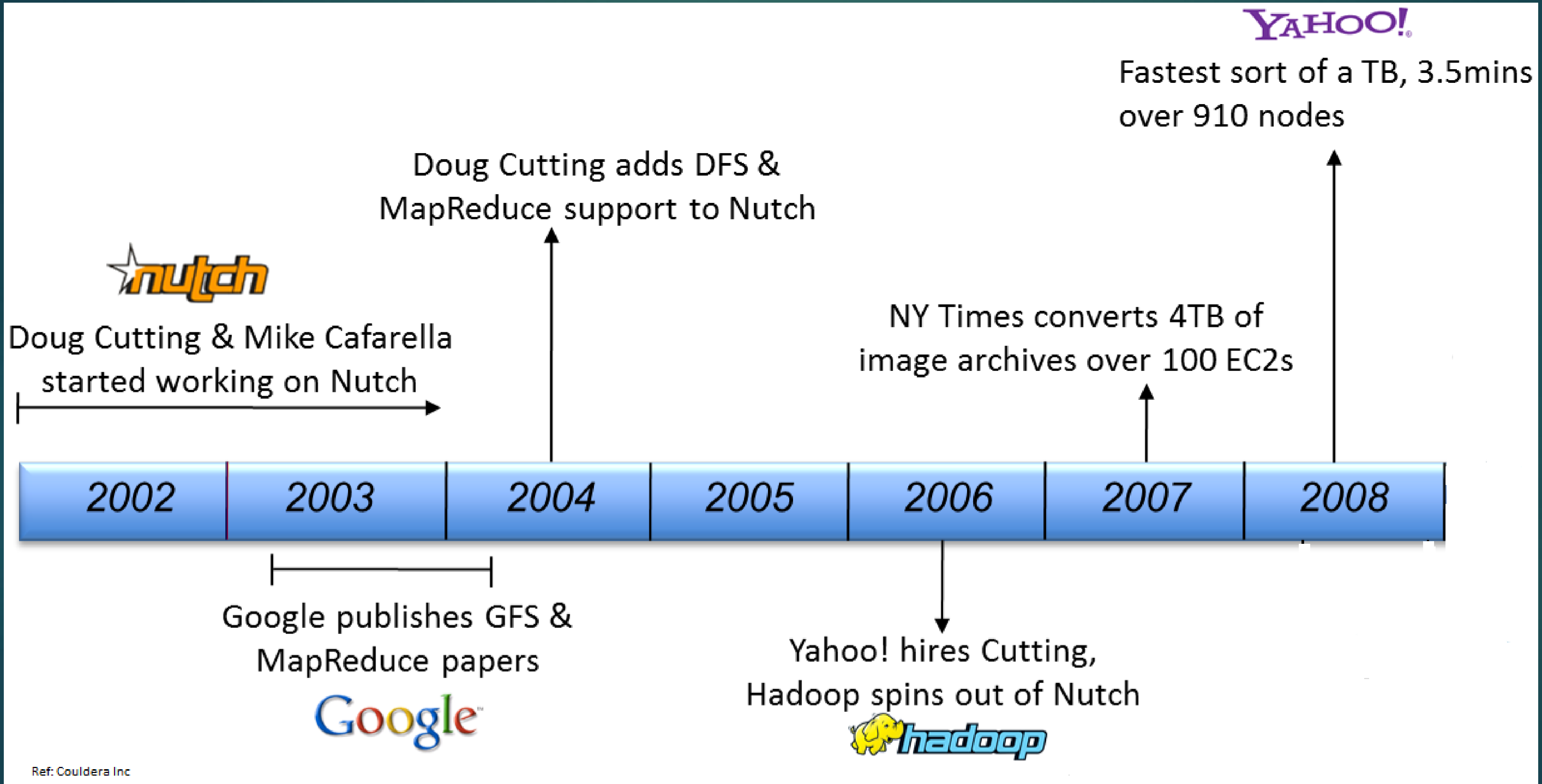
# History of Hadoop: Apache



January 2008: Apache top level project

Adopted by other giants as in: Facebook and New York Times

# History of Hadoop



Doug Cutting & Mike Cafarella started working on Nutch

Google publishes GFS & MapReduce papers

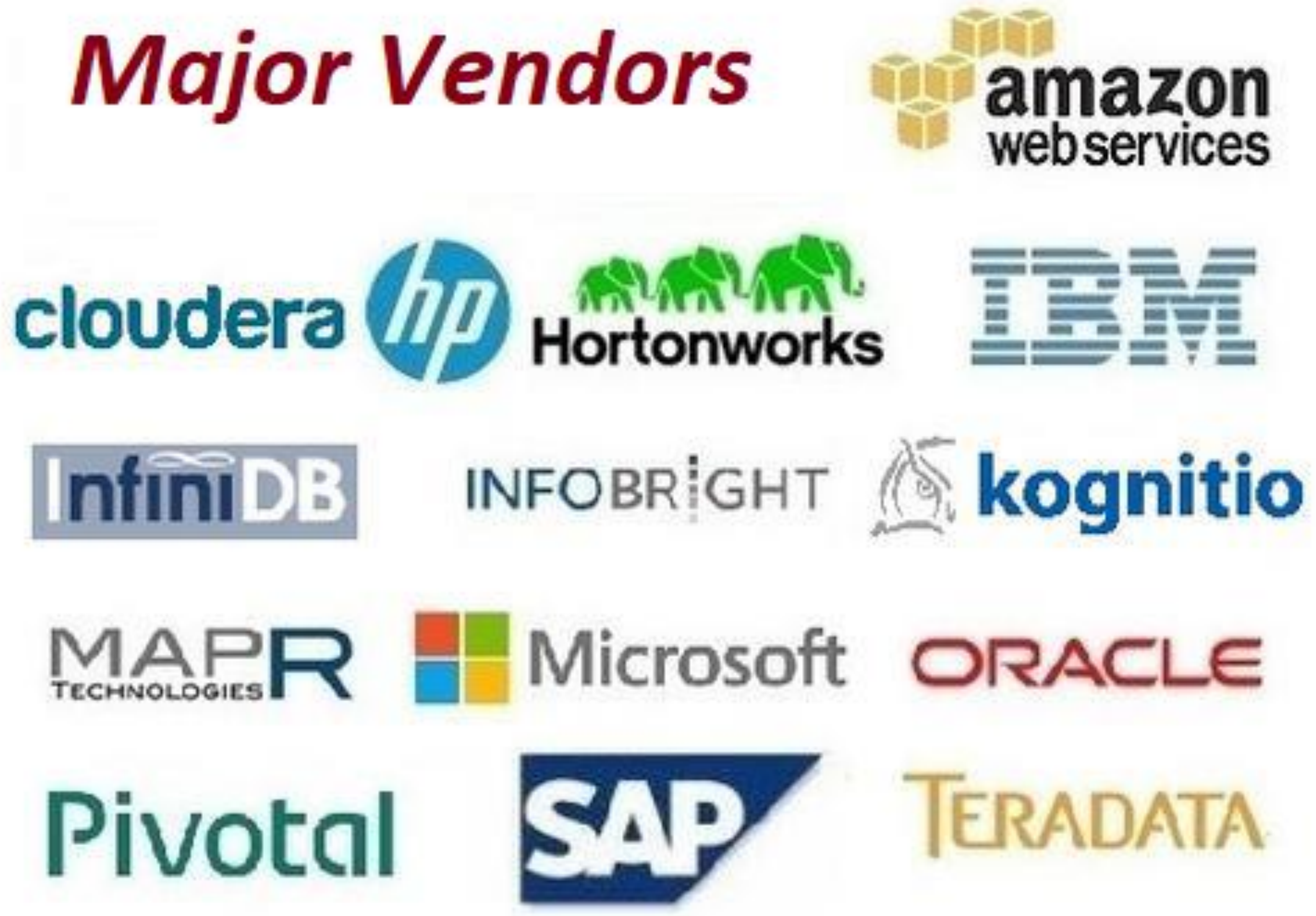
Doug Cutting adds DFS & MapReduce support to Nutch

Yahoo! hires Cutting, Hadoop spins out of Nutch

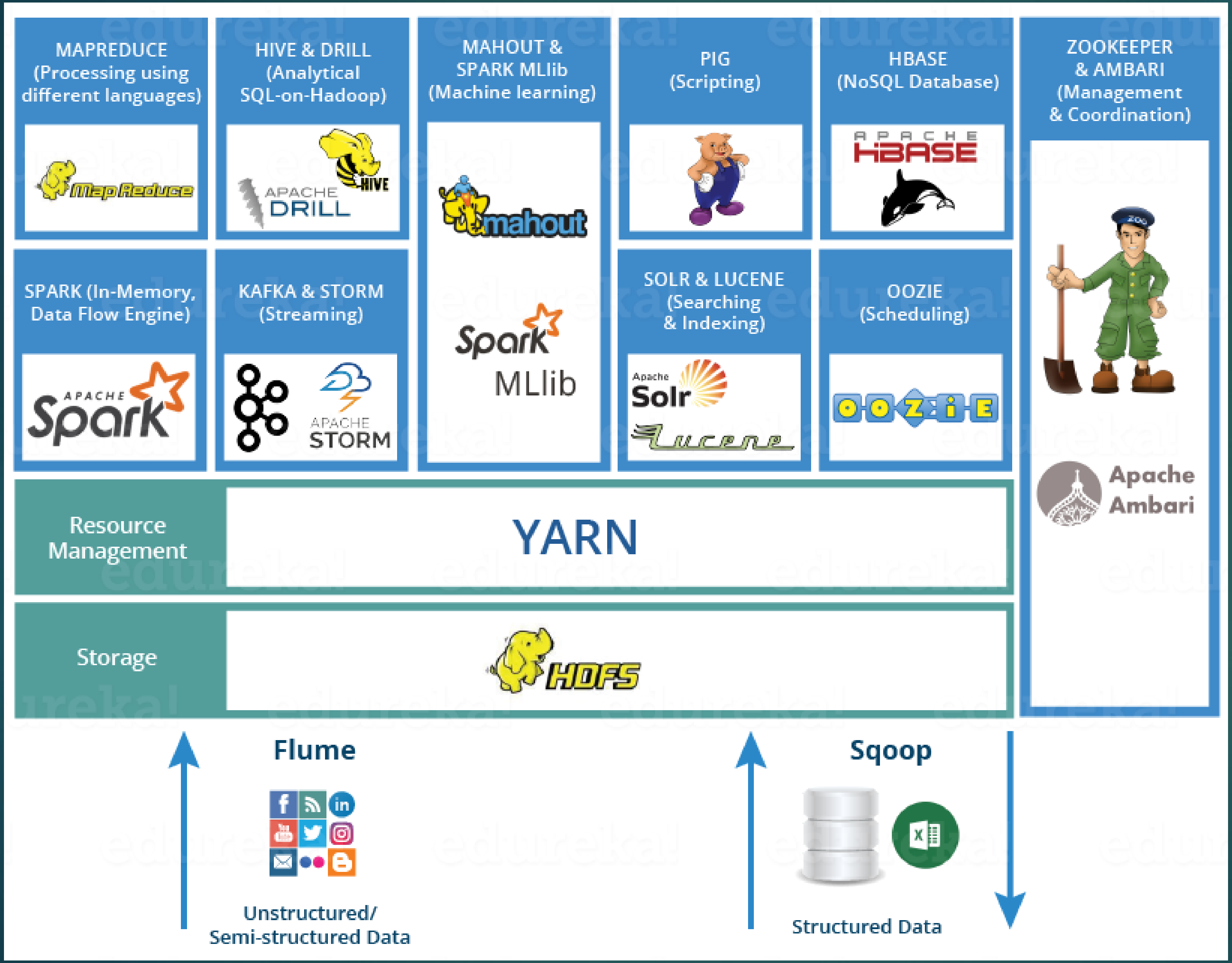
NY Times converts 4TB of image archives over 100 EC2s

Fastest sort of a TB, 3.5mins over 910 nodes

Ref: Couldera Inc



# Major Vendors


MAPREDUCE (Processing using different languages)
Map Reduce
HIVE & DRILL (Analytical SQL-on-Hadoop)
APACHE DRILL
HIVE
MAHOUT & SPARK MLlib (Machine learning)
mahout
Spark MLlib
PIG (Scripting)
HBASE (NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI (Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM (Streaming)
APACHE STORM
SOLR & LUCENE (Searching & Indexing)
Apache Solr
Lucene
OOZIE (Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data
edureka!

https://www.edureka.co/blog/hadoop-ecosystem

# Activity

CALCULATE THE SUM

# Major Components

- HDFS
- Namenode
- Data node
- Job Tracker
- Task Tracker

# RDBMS vs Hadoop

| **Attribute** | **RDBMS** | **Hadoop** |
|---|---|---|
| Data Size | Gigabytes | Petabytes |
| Access | Interactive & Batch | Batch |
| Updates | Multiple Read/Write | Write once, Read multiple times |
| Transaction | ACID | None |
| Structure | Schema-on-write | Schema-on-read |
| Integrity | High | Low |
| Scaling | Nonlinear | Linear |

# Exercise

- Java:
  - Class & Object
  - Method
  - Inheritance
- Unix
  - http://www.ee.surrey.ac.uk/Teaching/Unix/