Data vs Information

- **Data** :

  Simply fact or figure

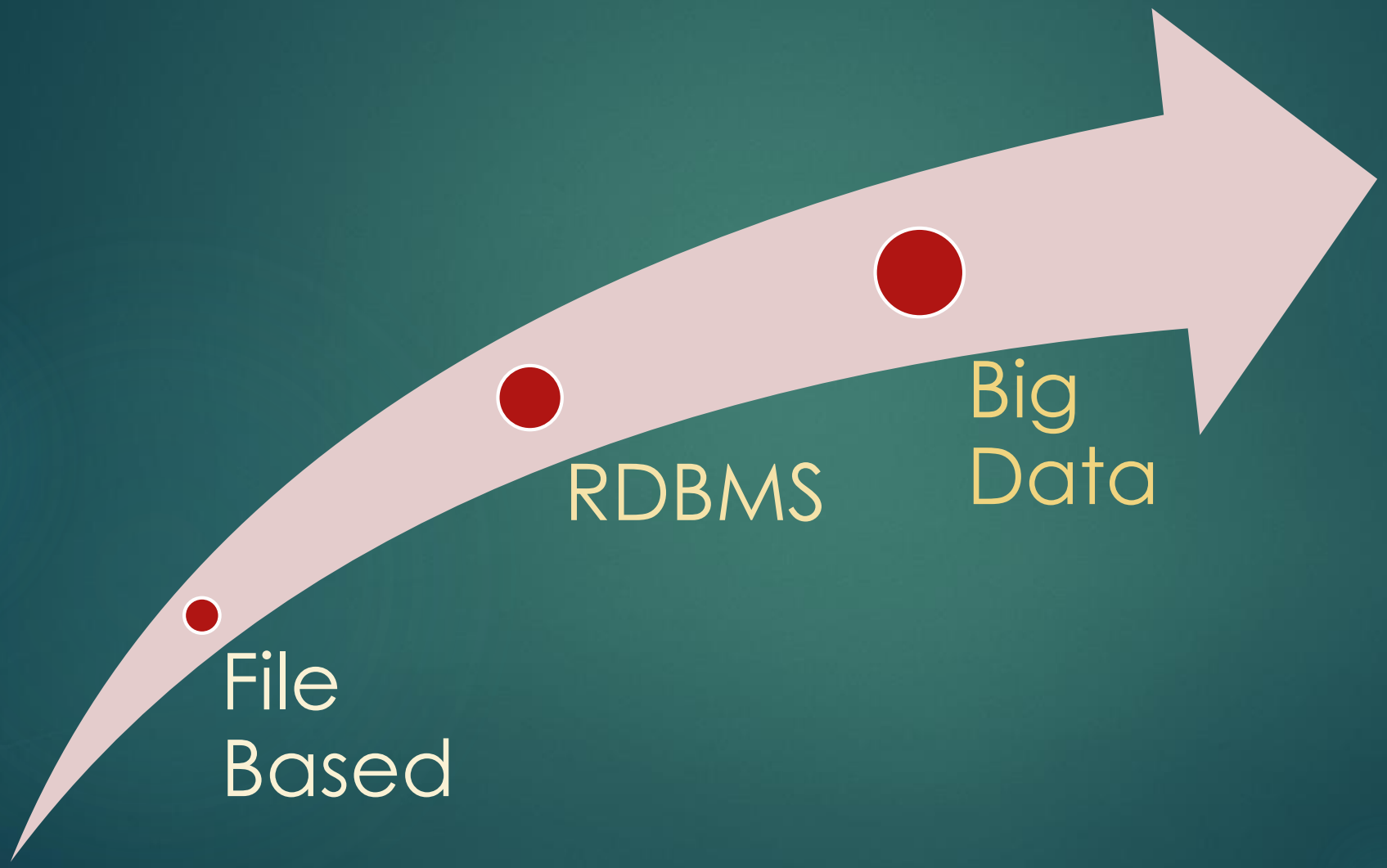
  For example: a number 15

- **Information**:

  Context + data

  For example: 15 degree centigrade is the temperature of Montreal on 26th May 2018 at 09:35 AM.

Evolution in Data management



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?

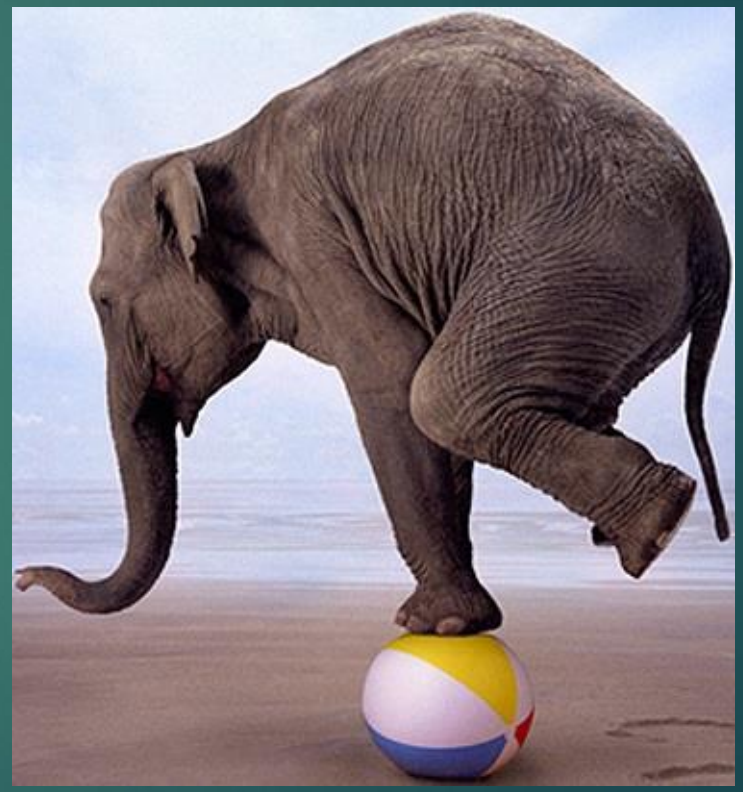
Ref: Hadoop definitive guide 4th edition, Oreilly publications

Characteristics of Big Data

- Volume
- Velocity
- Variety
- Value

Characteristic: Volume

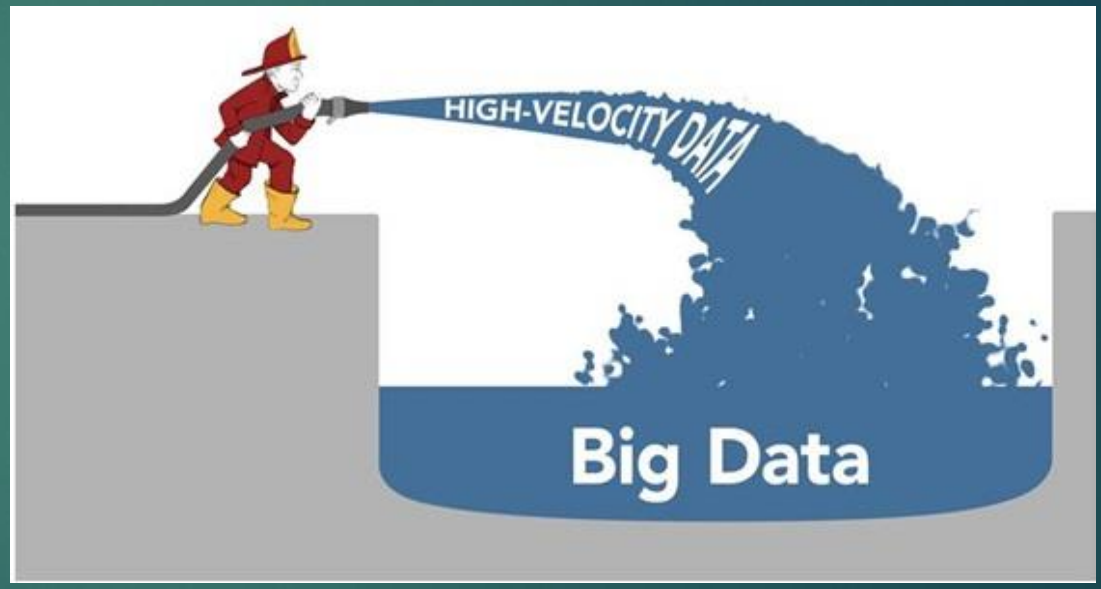
- Any guess how much amount of data we are producing within this room?
- Connected smart cars will generate 25GB data per hour



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

Course Outline

- Topics:

  https://github.com/shyam-kantesariya/big_data_course/blob/master/lecture1/Topics.pdf

- Email: kantesariyashyam@gmail.com

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

Compete to Innovate

- Go to www.innovatank.com
- Register as Innovator free trial
- Subscribe for NASA web log analysis challenge
- Make your team of two members
- Choose any programming language of your choice
- Submit your proposed solution
- Time Limit: *two hours*

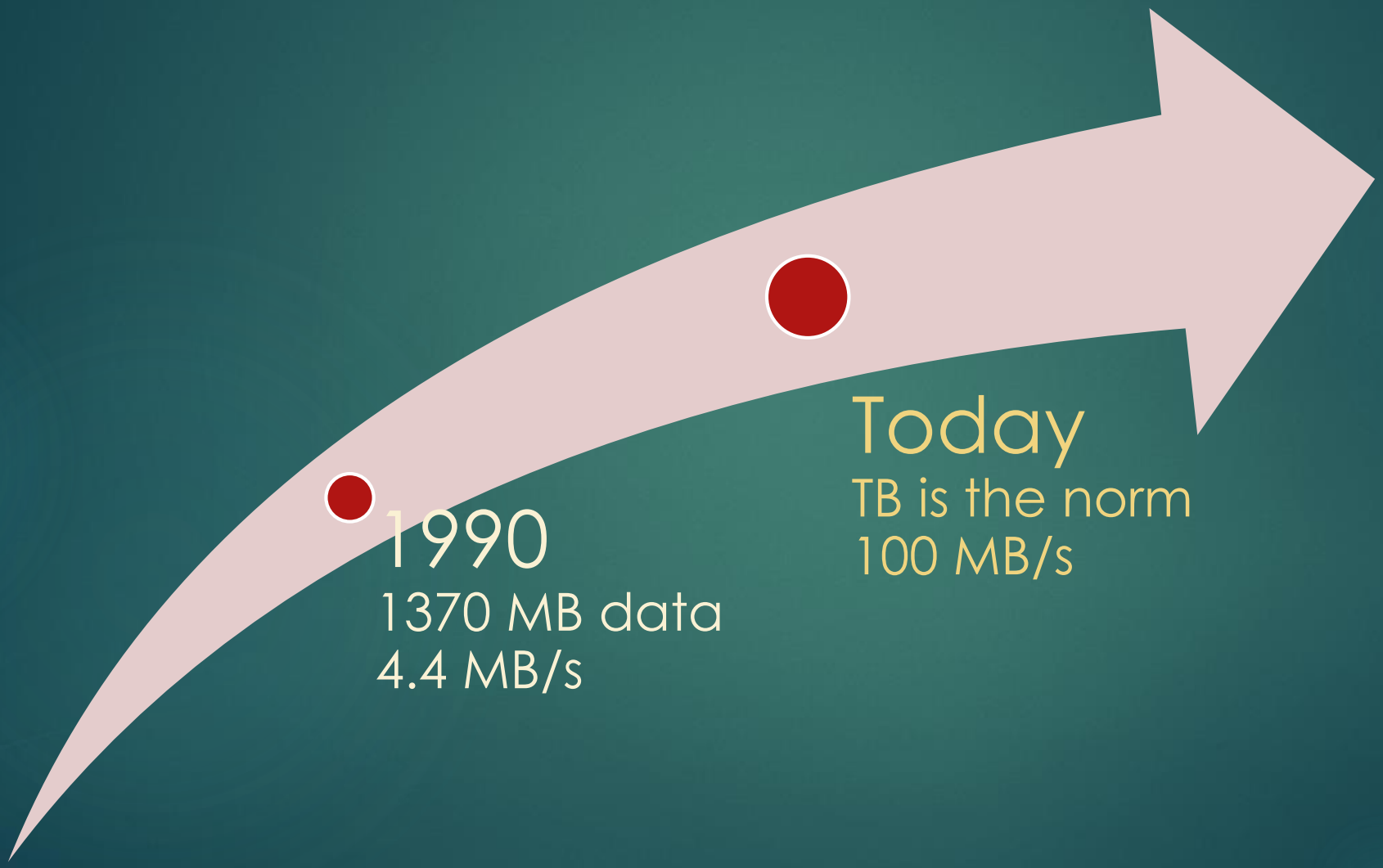
What is next?

- 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



Grid computing

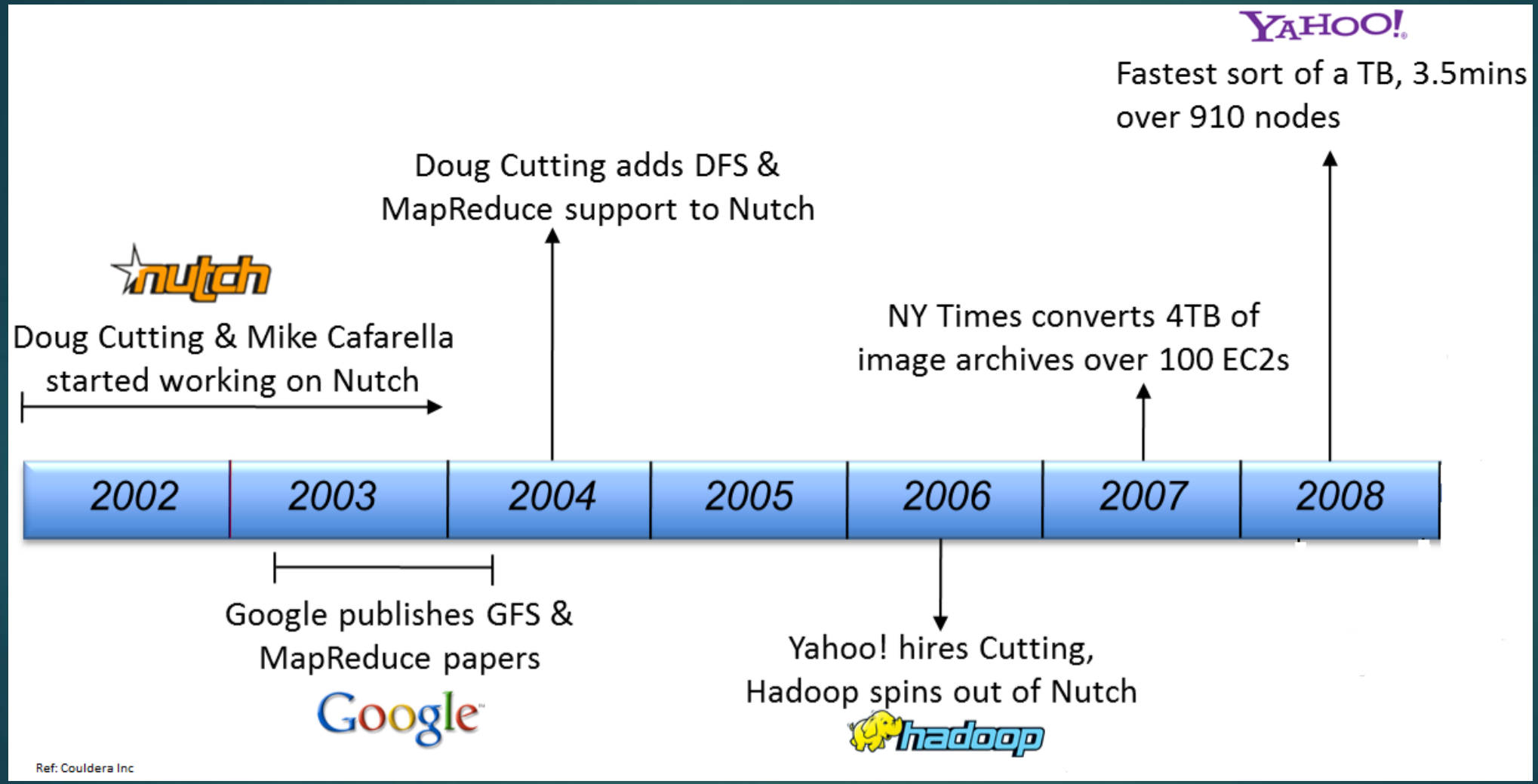
- Based on Message Passing Interface (MPI)
- Uses shared filesystem
- Programmer has to think at task level as opposed to data level
- Missing abstraction of fault tolerance

Volunteer computing

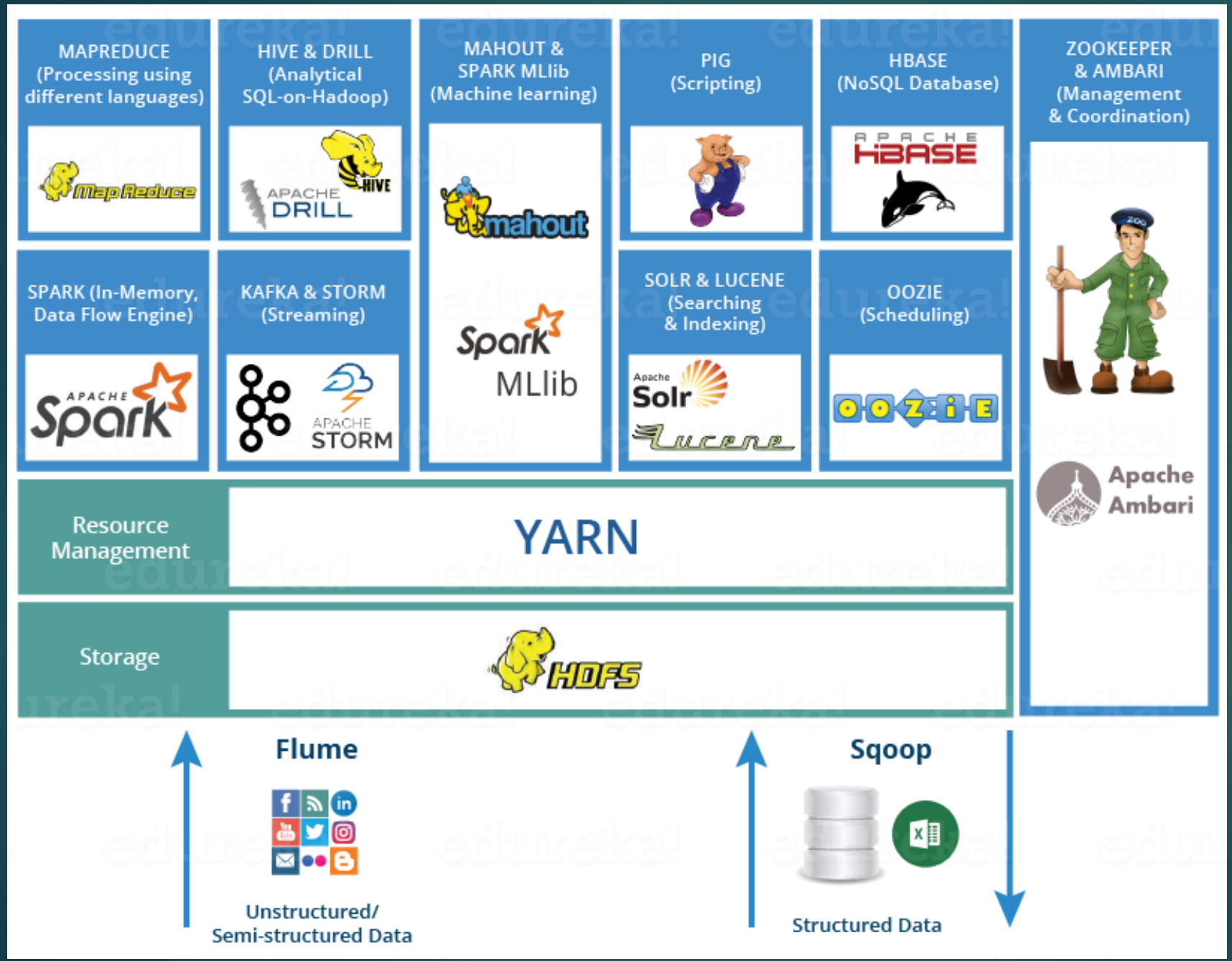
- System is highly compute intensive
- Small amount of data on remote machine
- Low bandwidth
- Based on Internet

Distributed Computing

History of Hadoop



Major Vendors



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

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

Resources

- Intellij Idea

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

- Git bash

https://git-scm.com/downloads

- Unix
  - http://www.ee.surrey.ac.uk/Teaching/Unix/