Introduction Big Data

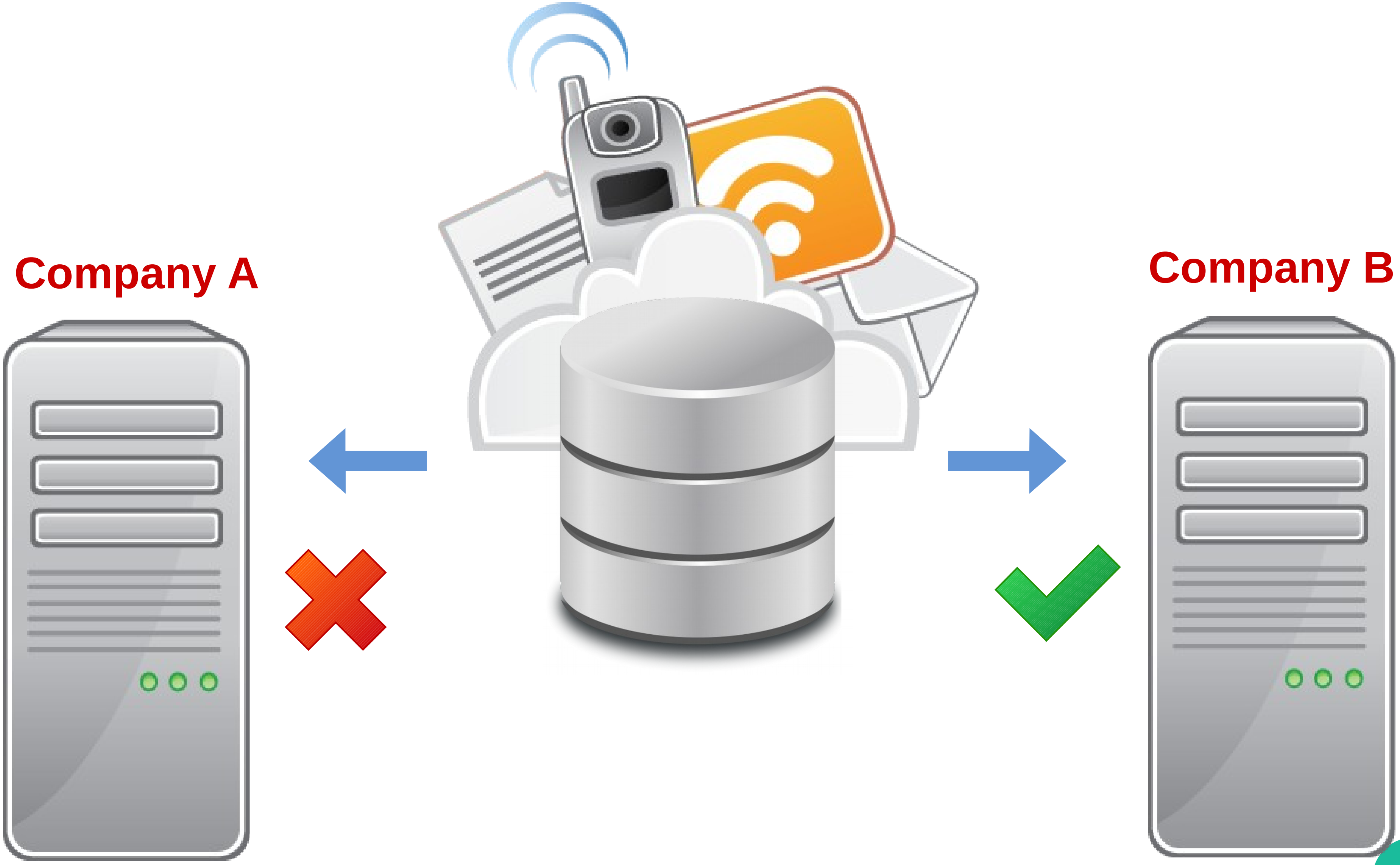
Agenda

- What is Big Data?
- How Big is our Data Universe?
- Sources of Big Data?
- Four V's of Big Data
- Data Storage & Analysis of Big Data
- Hadoop as Solution

What is Big Data?

- **Large Volume of Data both Structured & Unstructured**

Big Data



Company A

Company B

How Big is Our Data Universe?

- Every Day, we create 2.5 quintillion bytes of Data.
- 90% of the data in the world today has been created int the last two years alone.
- The production of data is expanding by 4300% increase in annual data generation by 2020.

How Big is Our Data Universe?

- **IDC estimate “digital universe” at 4.4 zettabytes in 2013 and is forecasting a tenfold growth by 2020 to 44 zettabytes.**

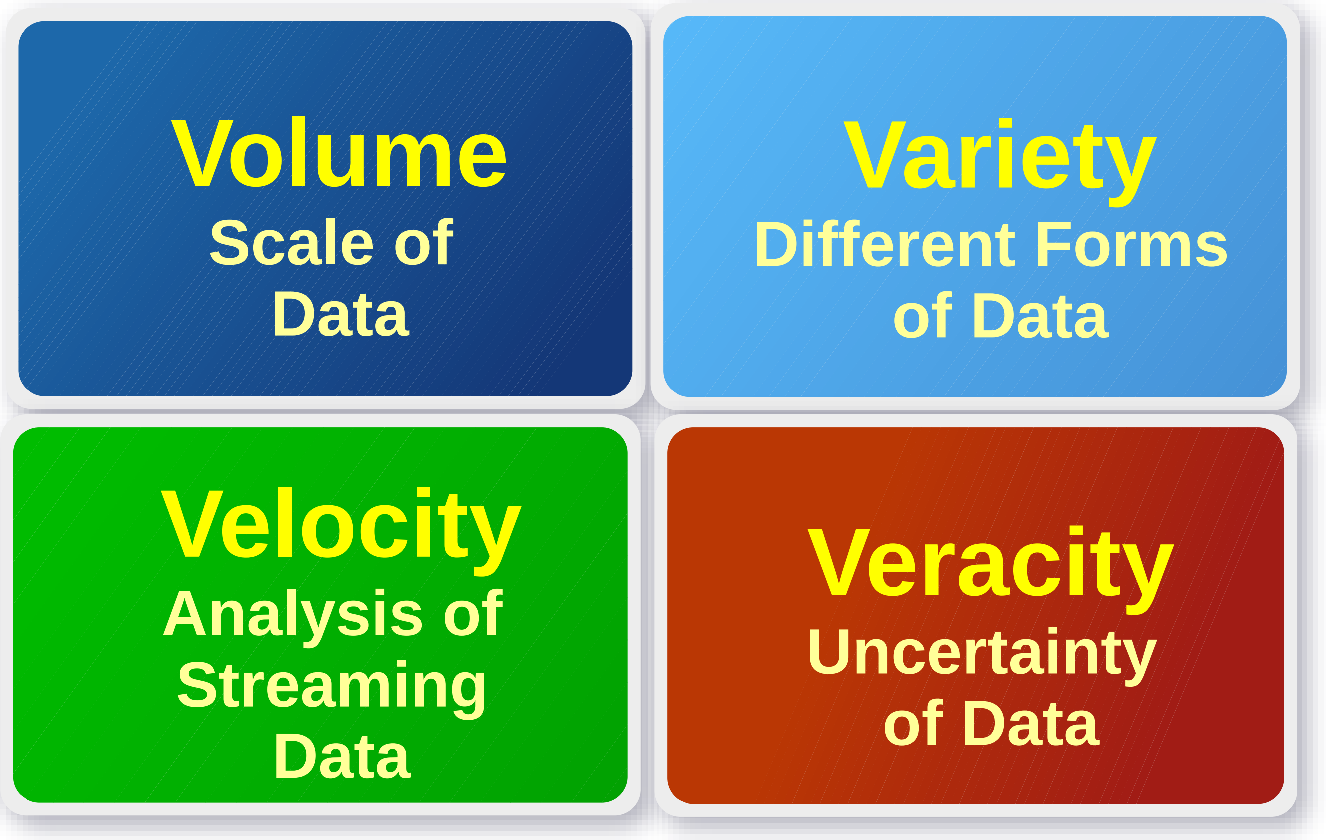
What is Zettabyte?

1 Bit	Binary Digit
8 Bits	1 Byte
1024 Bytes	1 Kilobyte
1024 Kilobytes	1 Megabyte
1024 Megabytes	1 Gigabyte
1024 Gigabytes	1 Terabyte
1024 Terabytes	1 Petabyte
1024 Petabytes	1 Exabyte
1024 Exabytes	1 Zettabyte

Big Data Sources

- The **New York Stock** Exchange generates about **4−5 terabytes** of data per day.
- **Facebook** hosts more than 240 billion photos, growing at **7 petabytes per month.**
- **Ancestry.com,** the genealogy site, stores around **10 petabytes** of data.
- The Internet Archive stores around **18.5 petabytes** of data.
- The **Large Hadron Collider** near Geneva, Switzerland, produces about **30 petabytes of data per year.**

Four V's of Big Data



Volume
Scale of Data

Variety
Different Forms of Data

Velocity
Analysis of Streaming Data

Veracity
Uncertainty of Data

Big Data Problem

WHAT WE KNOW...
THE REST...
BIG DATA

Data Storage & Analysis of Big Data

- **Storage capacity has grown exponentially but read speed has not kept up**
  - **1990:**
    - **Store 1,400 MB**
    - **Transfer speed of 4.5MB/s**
    - **Read the entire drive in ~ 5 minutes**
  - **2010:**
    - **Store 1 TB**
    - **Transfer speed of 100MB/s**
    - **Read the entire drive in ~ 3 hours**

Data Storage & Analysis of Big Data

- **Why not reading from multiple disks at once to reduce time?**
  - Imagine 100 drives, each holding one hundredth of the data. Working in parallel
  - Read entire data in under two minutes.
- **Yes, Hadoop does that.**

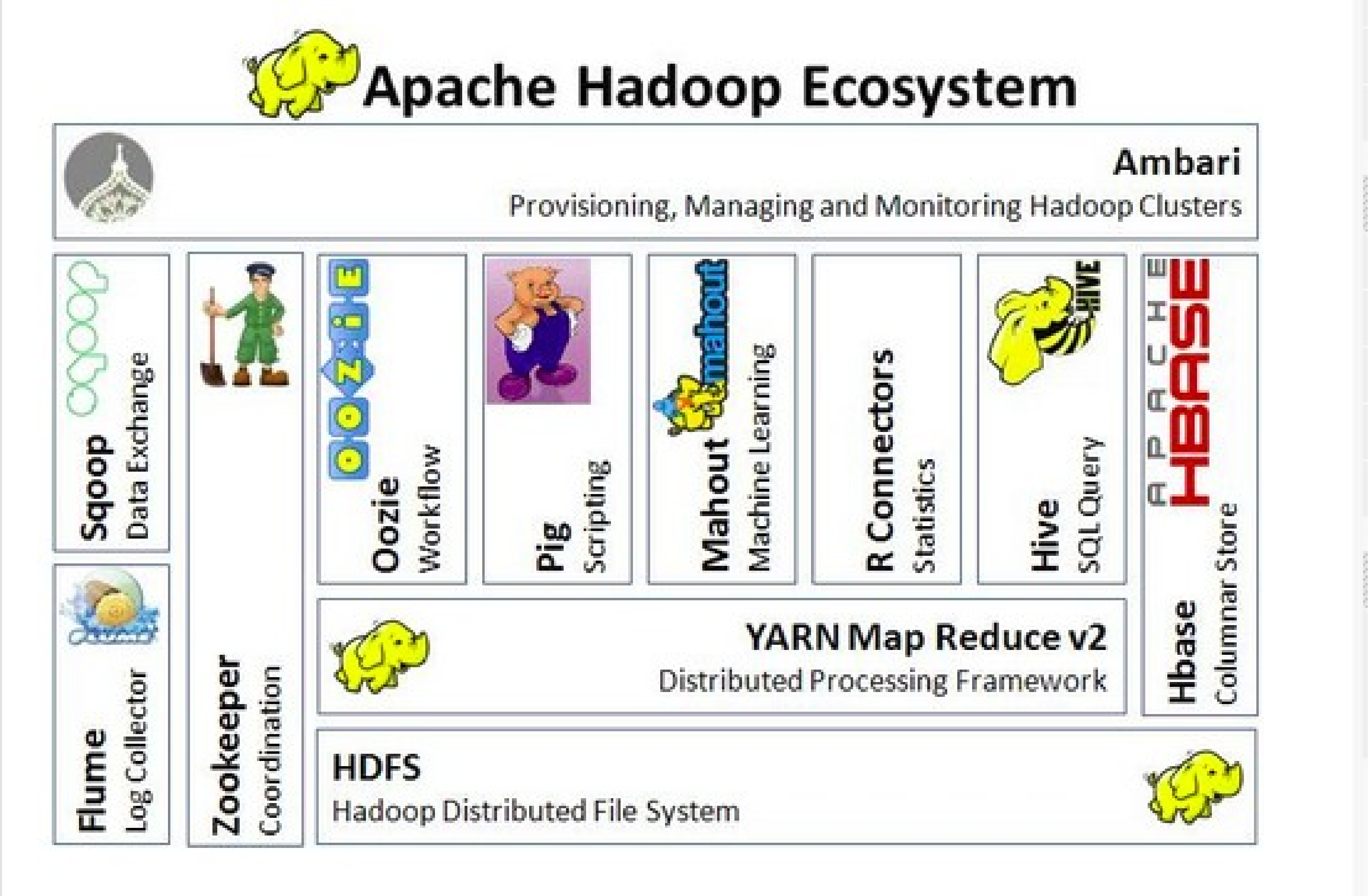
Problems with Reading from Multiple Disks Together

- **Hardware failure**
  - **HDFS solves this.**
- **Combining data from Multiple Disks after reads**
  - **Map Reduce Solves this.**

Hadoop as Affordable Solution

- Hadoop provides: a reliable, scalable platform for storage and analysis.
- It runs on commodity hardware.
- It is open source.

Hadoop Eco System



Resources

- **Hadoop: The Definitive Guide**
  - Tom White (Author)
  - O'Reilly Media; 4th Edition.