Fundamental Concepts: Big Data

Thuan L Nguyen, PhD

Introduction to Big Data & Hadoop Framework

1. Big Data: What Is It?
2. Big Data: The Big V's and Its Management
3. Apache Hadoop Framework: Overview
4. Apache Hadoop Ecosystem: Major Components
5. Apache Hadoop Ecosystem: Why We Use It?
6. Apache Hadoop Ecosystem: The Focus

Big Data: What is It?

Big Data: What is It?

- Big data is a term to describe large volumes of data that can be both **structured** and **unstructured**.
- These enormous volumes of data overwhelm the digital world every second.
- However, it is not the amount of data that is important.
- It is what we can do with the data that matters: Big data analytics can provide insights that lead to better decisions and strategic moves.

Big Data: Where does It Come From?

Source: ADEC Group (http://www.adec-innovations.com/)

Big Data: Where does It Come From?

- Ubiquitous availability of internet connections using broadband networks
- Advent of sensors
- Advanced mobile technologies
- High performance computing devices
- …

All leads to a situation in which firms have been overwhelmed by the staggering amount of data collected via digital interactions with consumers.

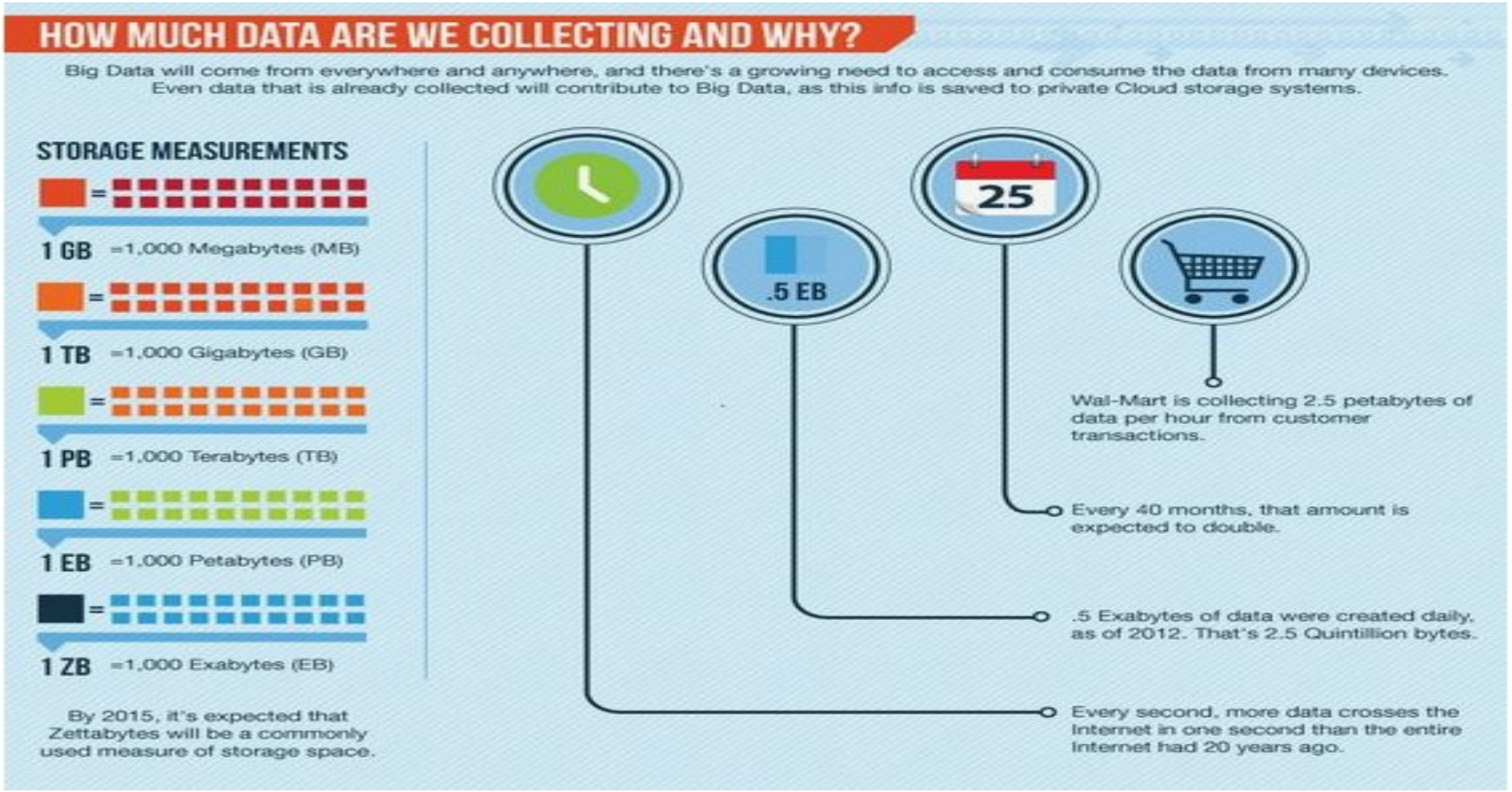
Big Data: How Big is It?

- It is staggering!
- According to IBM: 90% of data existing today has been created just for last two years.

- IBM also predicted by 2020:
  - Six billion people will have cell phones.
  - 2.5 quintillion bytes of data will be generated every day.

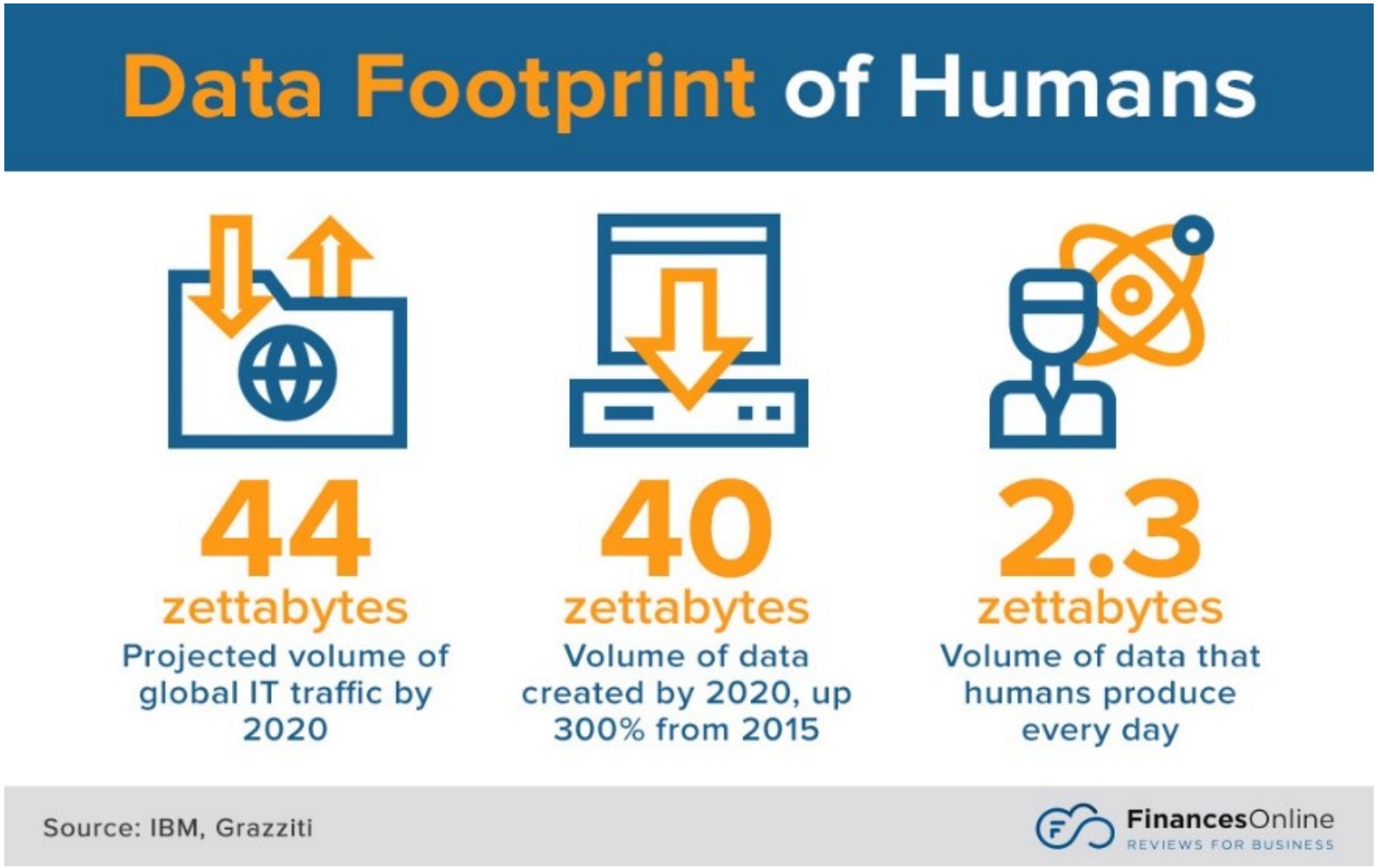
- A quintillion = 1 000 000 000 000 000 000 bytes ~ 1 Zetabytes

Big Data: How Big is It?



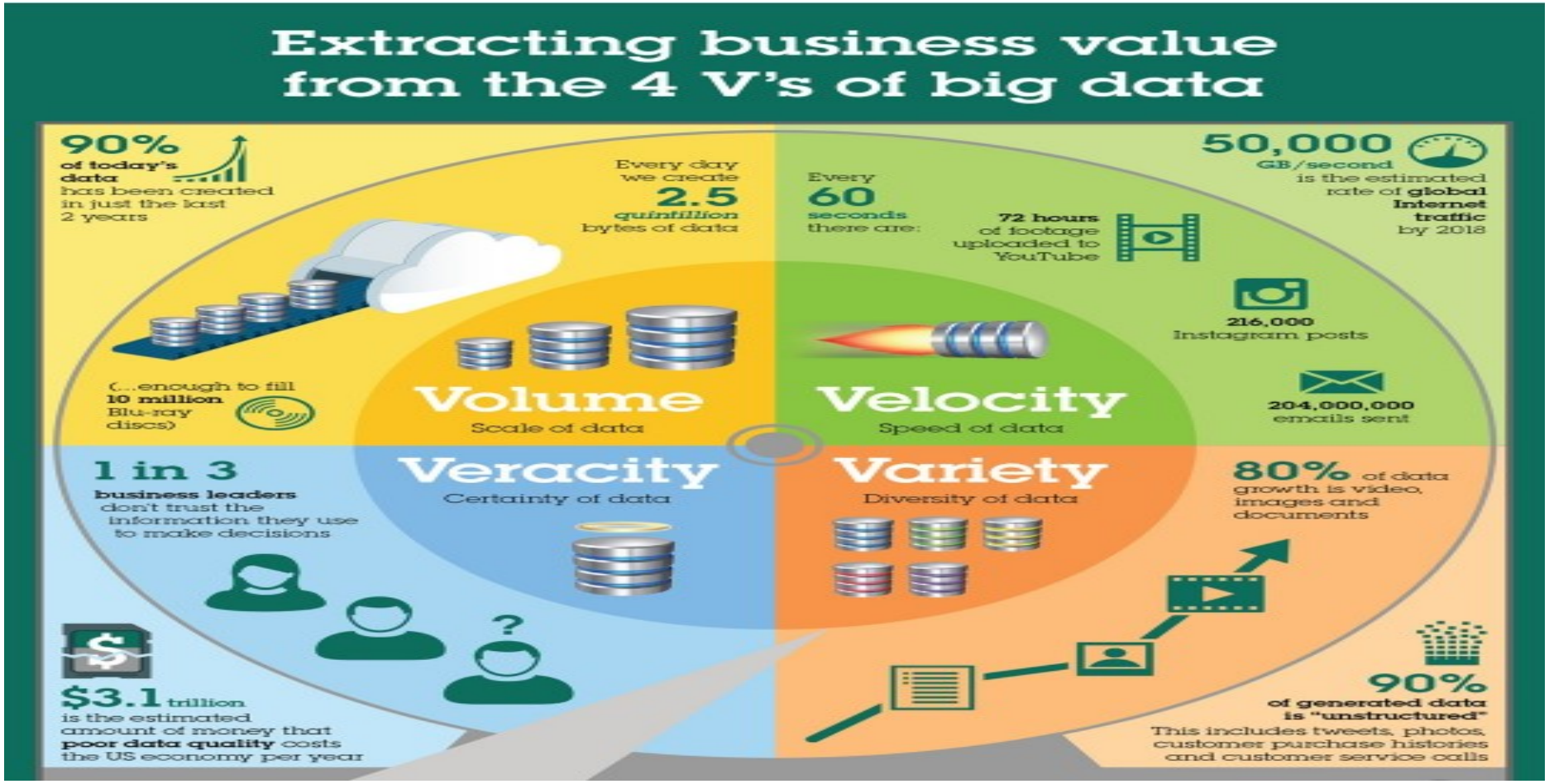
Source: MasterInit.org

Big Data: How Big is It?



Source: IBM, Grazzilti, FinancesOnline.com

Big Data: The Five Big V's



Source: IBM Cloud

Big Data: The Five Big V's

Big data has five elements. The first four are:

- **Volume**: It is enormous. The amount of data created in a unit of time – either second, minute, hour, or day – that is huge in comparison with traditional data sources.
- **Velocity**: It is very fast. The 2nd V refers to how fast new data is generated and transmitted.
- **Variety**: It is not uniformed. The 3rd V indicates different formats of data that are generated in big volume: emails, photos, videos, documents, sensor signals, to name a few.
- **Veracity**: The 4th V refers to the fact that when data comes in staggering volumes, with extremely high speed, and in all kinds of formats, the quality and accuracy of the data may be not at the high levels.

Big Data: The Five Big V's

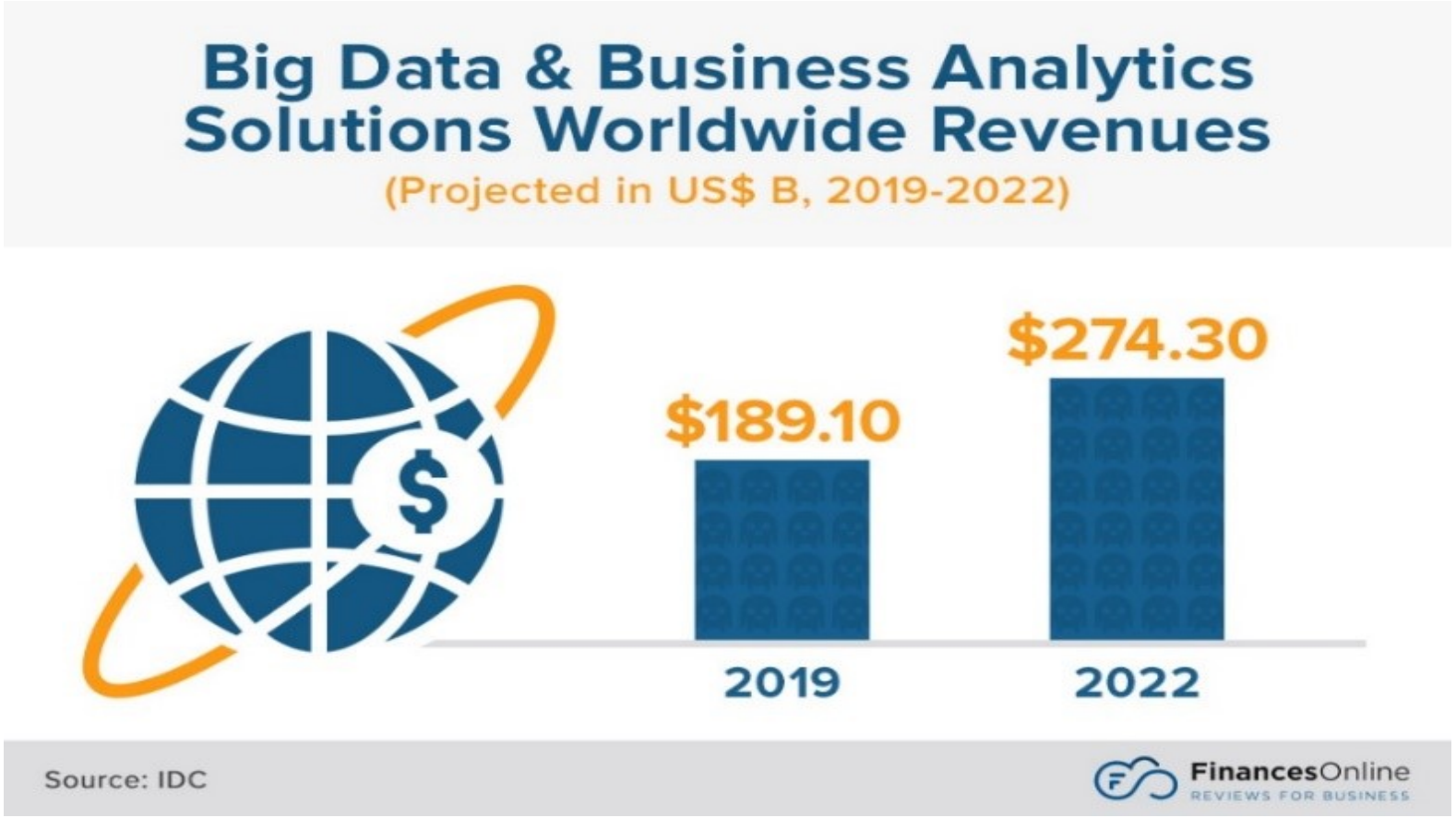
However, no matter how big or how fast the big data is, it is not BIG DEAL at all if big data does not bring out any business value.

Then, the 5th V comes out: **VALUE**!

Source: IBM Cloud

Big Data: The Five Big V's

BIG DATA: The 5th V comes out: **VALUE**!



Source: IDC, FinancesOnline.com

Big Data: How to Manage It?

So, **big data** is

- Very **big** (of course!): in enormous volumes
- Very **fast**: coming and being generated in extremely high speed
- Available in **all kinds of formats**: both structured and unstructured
- Originally **without high levels of quality**

Big Data: How to Manage It?

Can we use the well-known relational database systems like Oracle, Microsoft SQL Server, IBM DB2, MySQL, etc., to efficiently manage big data?

No!

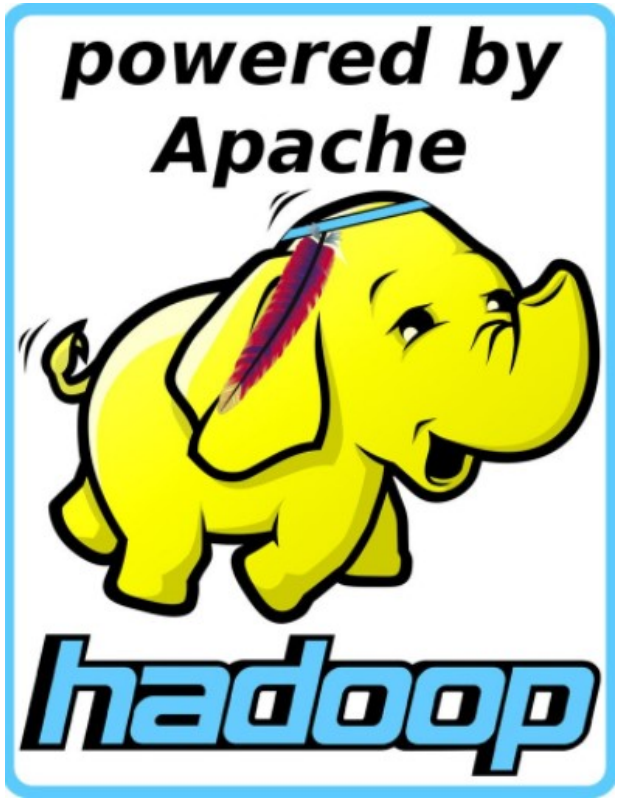
Because big data is very different from the data that can be found in these relational database systems, as seen in the previous slide.

In other words, big data is **NOT** the traditional data that is big.

We need a new type of data management system.

Big Data: How to Manage It?

The Apache Hadoop Framework



Apache Hadoop Framework: A Bit of History

- 2002: Doug Cutting & Mike Cafarella built Nutch to crawl the web. Nutch is a primitive distributed system that could run on several computers at most.
- 2003: Google released the white paper on Google File System that was also a distributed system but much more efficient and powerful.
- 2004: Google released the white paper on MapReduce that is a parallel processing system.
- Based on these white papers, Doug Cutting & Mike Cafarella built a new distributed system that is the foundation for an open source distributed system: **Apache Hadoop Framework**.
- And the history was made!

Apache Hadoop Framework: Overview

What is the Apache Hadoop?

Apache Hadoop is an open source framework:

- Being capable of processing large amounts of heterogeneous data sets in a distributed fashion across clusters of commodity computers and hardware
- Using a simplified programming model
- Providing a reliable shared storage and analysis system.

Apache Hadoop Ecosystem: Core Components

What are the major components of the Apache Hadoop ecosystem?

- Hadoop Distributed File System (HDFS)
- YARN (Hadoop 2)
- MapReduce
- Hive
- Pig
- Hbase
- Spark