

# ISIT312/ISIT912 Big Data Management

# Cluster Computing

Dr Guoxin Su and Dr Janusz R. Getta

School of Computing and Information Technology -  
University of Wollongong

# Cluster Computing

## Outline

[Computer Cluster](#)

[Big Data](#)

[Traditional Data Architectures](#)

[Meet Hadoop !](#)

[Big Data on Database Clusters](#)

[Big Data on Kubernetes](#)

# Computer Cluster

What is a **computer cluster** ?

A **computer cluster** is a collection of computers (also called as **nodes**) connected through high speed local are network that work together to simulate a single much more powerful computer system

Each node in a **computer cluster** is controlled by its own operating system

Each node in a **computer cluster** performs a different version of the same task

A difference between **computer cluster** and **computer grid** is such that the nodes in a computer grid perform different tasks

An architecture of **computer cluster** ranges from a simple two-node system connecting two personal computers to a supercomputer with a cluster architecture

# Computer Cluster

**Computer clusters** are used to speed up computing through **shared nothing (sharding)** partitioning of data and paralellization of data procesing on the nodes of a cluster

**Computer clusters** provide high availability through automatic replacement of a failed node with a **replica node**

Advantages of **computer clusters**: faster processing speed, larger storage capacity, better data integrity, greater reliability and wider availability of resources

A **Linux cluster** is a collection of connected computers that can be viewed and managed as a single system

A sample **computer cluster**: 54 regular compute nodes (with two 32-Core Intel 8358 processors, 1.6TB of local NVME storage and 512GB of memory each) and 5 GPU nodes with two 24-Core AMD EPYC 7413 processors, eight A100 GPU cards, 960GB of local storage and 512GB of memory each

# Computer Cluster

What is a **cluster computing** ?

**Cluster computing** is the process of sharing the computation tasks among multiple computers included in a **computer cluster**

Advantages of **cluster computing**: cost efficiency, processing speed, expandability, high availability of resources

At the moment **cluster computing** is an attractive paradigm for processing large scale science, engineering and commercial applications

**Cluster computing** requires the specialized algorithms like load balancing, resource sharing and resource scheduling for optimization of data processing

**Cluster computing** is an attractive alternative to data processing on **large parallel supercomputers**

The simplest configuration of nodes for **cluster computing** consists of a **master node** and **slave nodes**

# Cluster Computing

## Outline

[Computer Cluster](#)

[Big Data](#)

[Traditional Data Architectures](#)

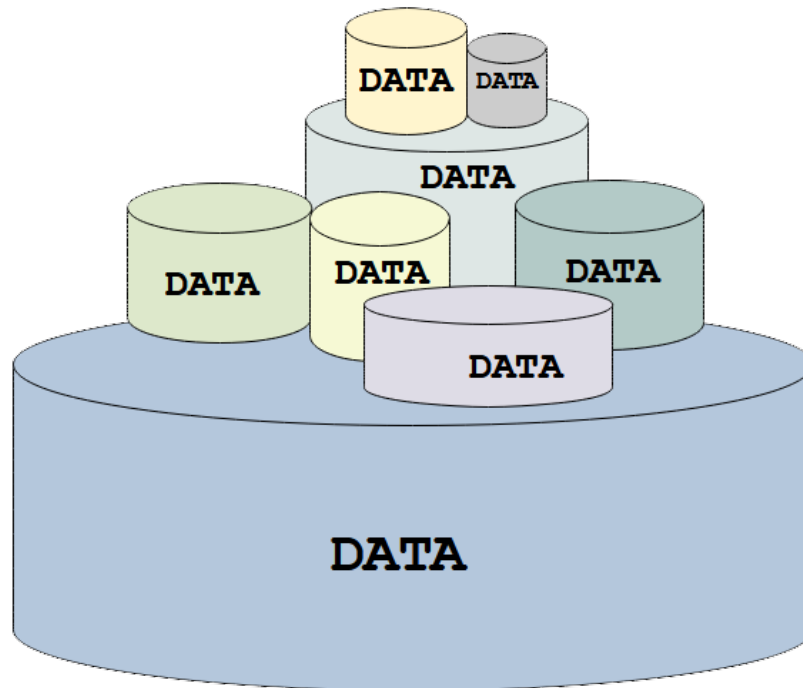
[Meet Hadoop !](#)

[Big Data on Database Clusters](#)

[Big Data on Kubernetes](#)

# Big Data

What does **Big Data** mean and how big is **Big Data** ?

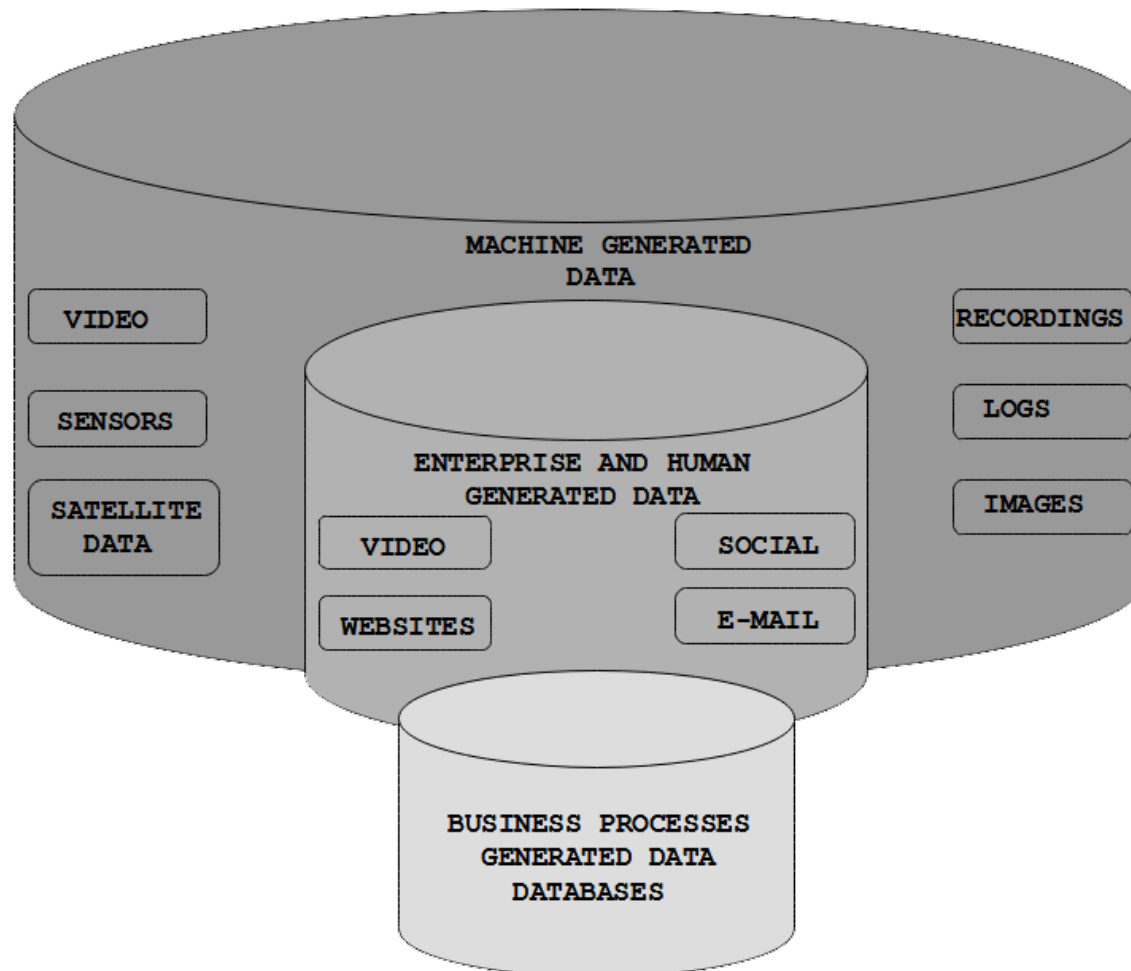


**Big Data** is so big that it cannot be stored on the persistent storage devices attached to a single computer system

**Big Data** may also mean **an infinite amount of data**

# Big Data

What are the sources of **Big Data** ?



[TOP](#)



# Big Data

**Big Data** is characterized by so called **3V features**:

- **Volume**: e.g., billions of rows ? millions of columns
- **Variety**: Complexity of data types and structures
- **Velocity**: Speed of new data creation and growth

Additional **Vs**:

- **Veracity**: Ability to represent and process uncertain and imprecise data
- **Value**: Data is the driving force of the next-generate business
- **Viability**: Benefits we can potentially have from data analysis

There are many, many other **Vs**, the largest number of **Vs** I found on Web was **42** !

- **Vagueness**: The meaning of found data is often very unclear, regardless of how much data is available
- **Validity**: Rigor in analysis is essential for valid predictions where data is the driving force of the next-generate business
- **Vane**: Data science can aid decision making by pointing in the correct direction
- ... and many, many others. ;)

# Big Data

## Examples of Big Data:

- Clickstream data
- Call centre data
- E-mail and instant-messaging
- Sensor data
- Unstructured data
- Geographic data
- Satellite data
- Image data
- Temporal data
- and more ...

# Cluster Computing

## Outline

[Computer Cluster](#)

[Big Data](#)

[Traditional Data Architectures](#)

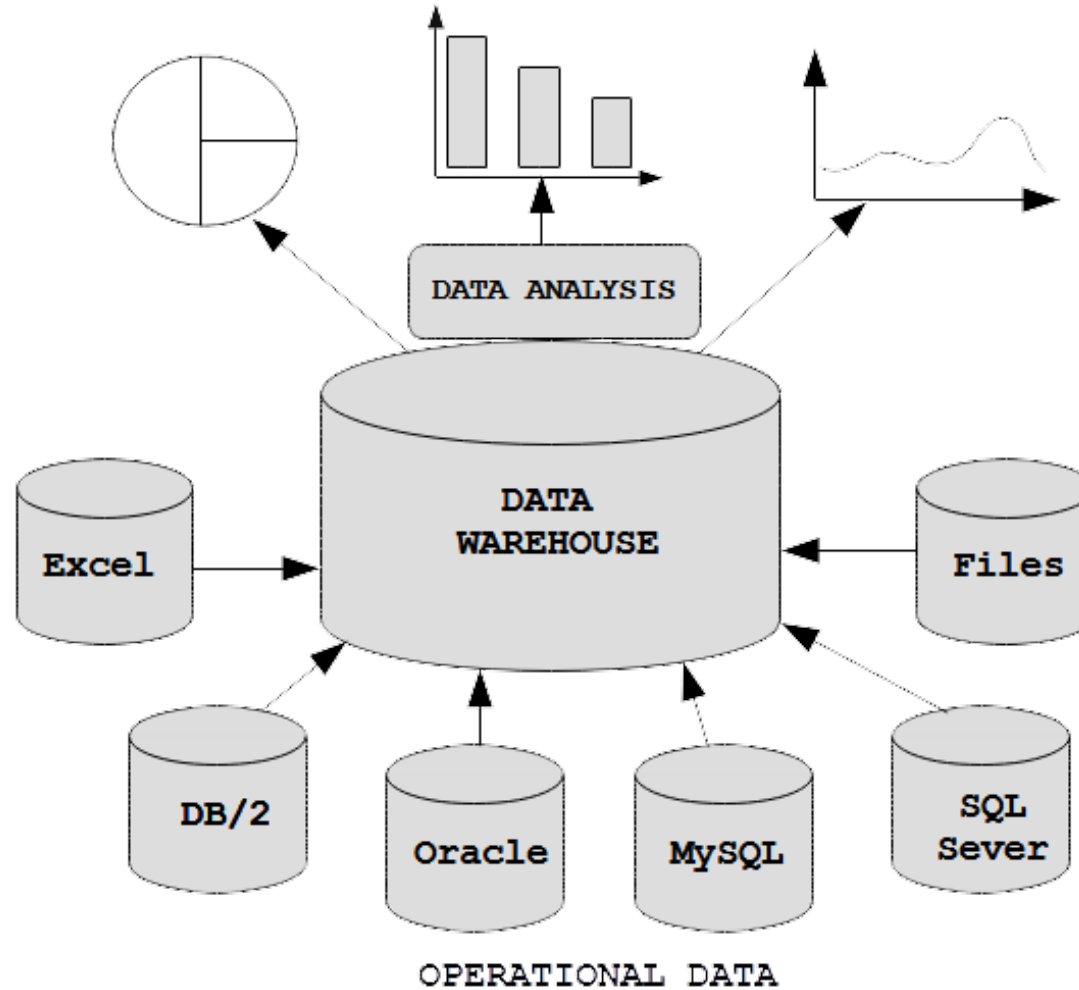
[Meet Hadoop !](#)

[Big Data on Database Clusters](#)

[Big Data on Kubernetes](#)

# Traditional Data Architectures

Data warehousing technologies



[TOP](#)

# Traditional Data Architectures

The strength of **traditional data architectures**:

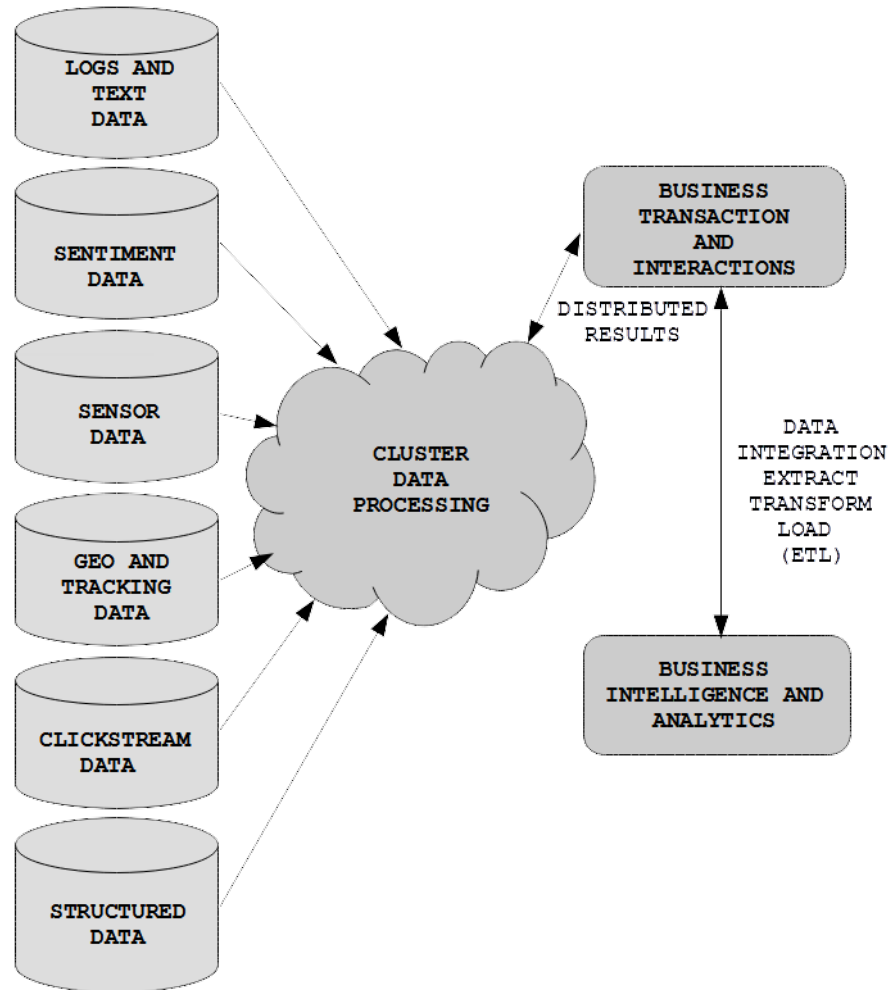
- Centralised governance of data repositories
- Light-fast inquiries performed regularly in daily business
- Optimisation for OLTP and OLAP
- Security and access control
- Fault-Tolerance and backup

The challenges for **traditional data architectures**:

- New types of data such as unstructured data and semi-structured data
- Increasingly large amounts of data flowing into organisations
- New computational paradigms use non-traditional NoSQL databases to rapidly mine and analyse very large data sets
- Increasing cost of storing and analysing the large amounts of data
- Increasing use of data analytics, which requires significant storage and processing capabilities

# Traditional Data Architectures

A sample **Data Lake** architecture



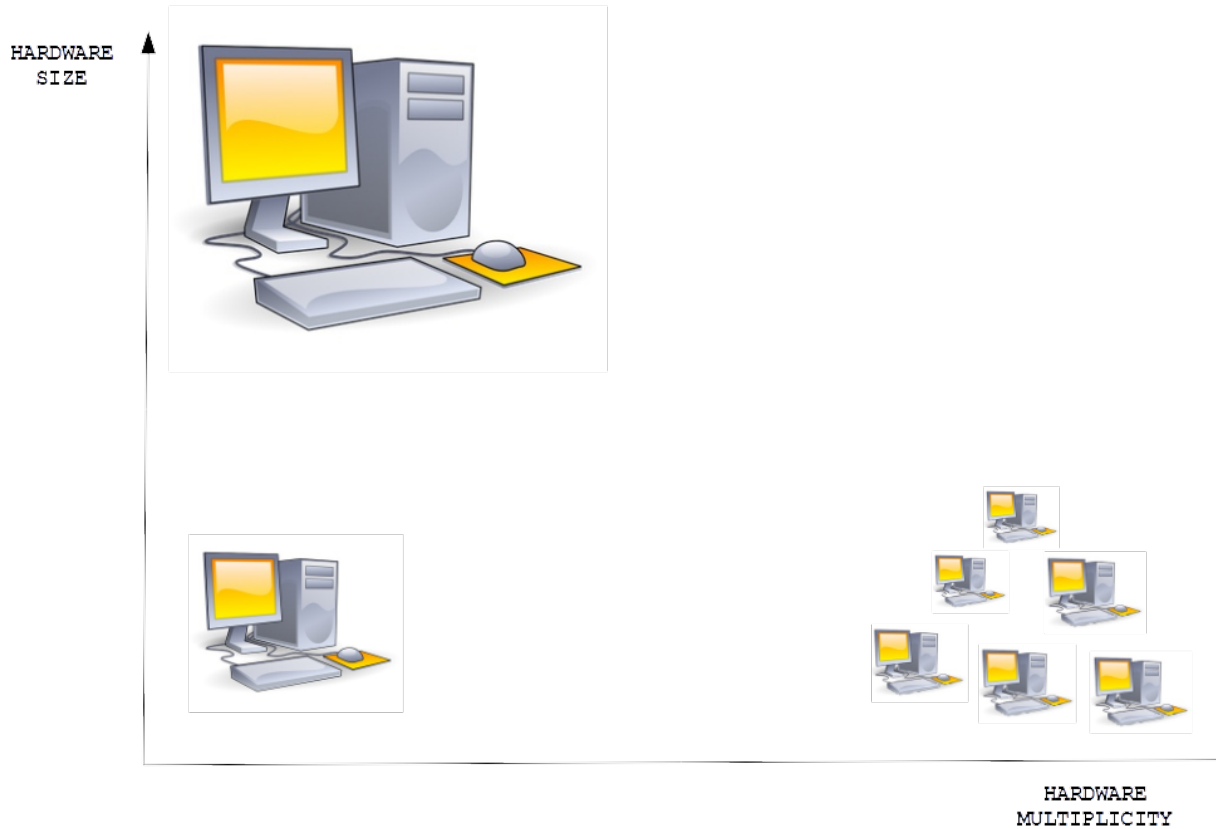
[TOP](#)

ISIT312/ISIT912 Big Data Management, Spring 2023

14/28

# Traditional Data Architectures

Hardware for **Big Data** has two scalability dimensions



# Cluster Computing

## Outline

[Computer Cluster](#)

[Big Data](#)

[Traditional Data Architectures](#)

[Meet Hadoop !](#)

[Big Data on Database Clusters](#)

[Big Data on Kubernetes](#)



# Meet Hadoop !

**Hadoop**, in terms of its developers, is a project that develops open-source software for reliable, scalable, distributed computing

## Features of Hadoop

- Capability to handle large data sets, e.g. simple scalability and coordination
- File size range from gigabytes to terabytes
- Can store millions of those files
- High fault tolerance
- Supports data replication
- Supports streaming access to data
- Supports batch processing
- Support interactive, iterative and stream processing
- Implements a data consistency model of **write-once-read-many** access model
- Run on commodity hardware, not high-performance computers
- Inexpensive
- It can be deployed on premises or in the cloud

# Meet Hadoop !

Core components of **Hadoop**

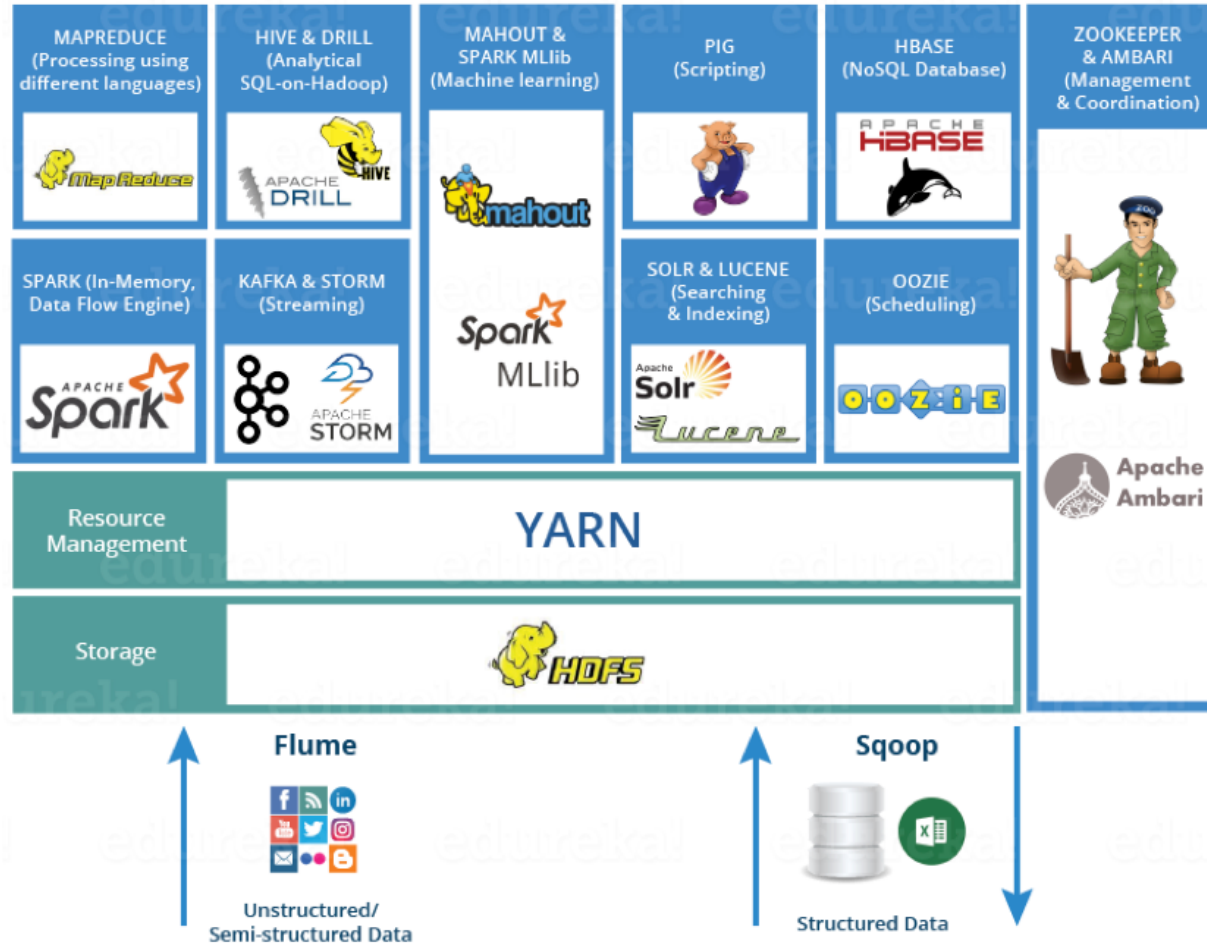
Different data-processing frameworks  
(e.g., MapReduce)

YARN: An Operating System for Hadoop  
(Hadoop Cluster Resource Management)

HDFS  
(Hadoop Distributed File System)

# Hadoop Ecosystem

## Hadoop ecosystem



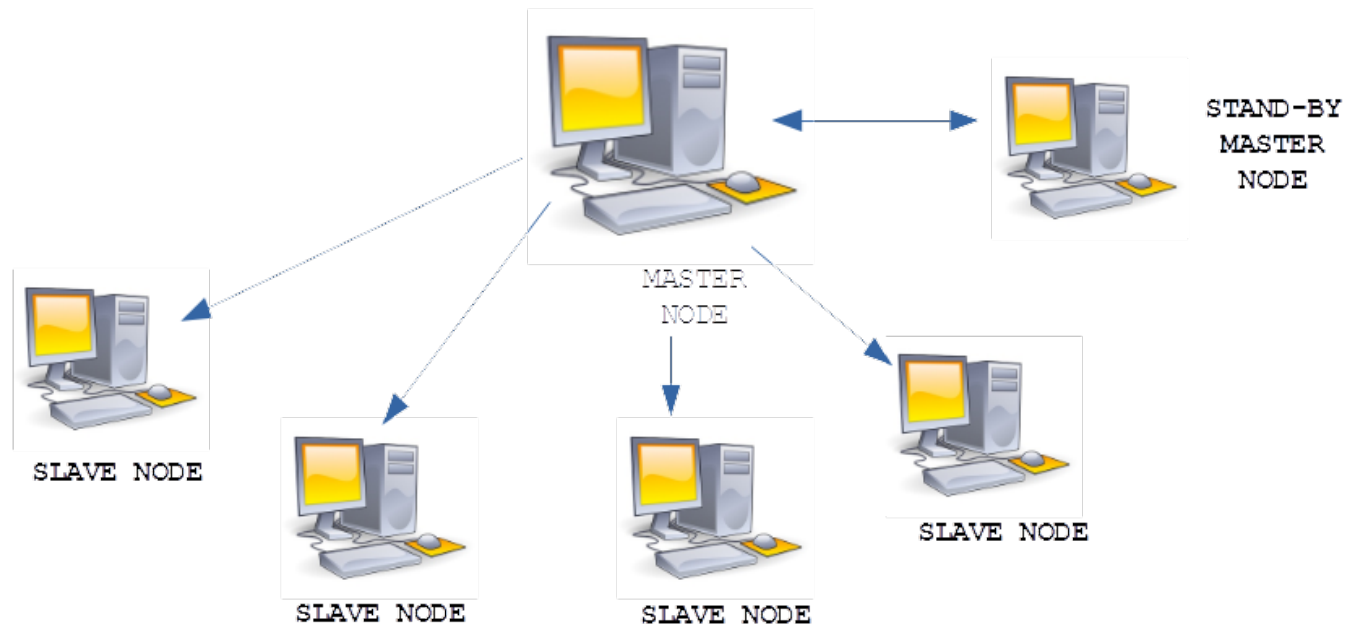
# Commercial Hadoop Landscape

Commercial **Hadoop** landscape



# Meet Hadoop !

## Master-slave architecture of Hadoop clusters



# Meet Hadoop !

**Hadoop clusters** can support up to 10,000 server and receives near-to-linear scalability in computing power

A typical **Hadoop cluster** consists of:

- A set of **master nodes** (servers) where the daemons supporting key Hadoop frame-works run
- A set of **worker nodes** that host the storage (HDFS) and computing (YARN) work
- One or more **edge servers**, which are used for accessing the Hadoop cluster to launch applications
- One or more **relational databases** such as MySQL for storing the metadata repositories
- **Dedicated servers** for special frameworks such as Kafka

# Meet Hadoop !

**Hadoop** also support the **pseudo-distributed mode**

- All HDFS and YARN daemons running on a single node.
- Highly simulate the full cluster
- Easy for beginner's practice
- Easy for testing and debug

Our lab setting is the **pseudo-distributed mode**

- The single node is a Ubuntu 14.04 Virtual Machine (VM)

# Cluster Computing

## Outline

[Computer Cluster](#)

[Big Data](#)

[Traditional Data Architectures](#)

[Meet Hadoop !](#)

[Big Data on Database Clusters](#)

[Big Data on Kubernetes](#)



# Big Data on Database Clusters

A **database cluster** is a collection of databases that is managed by a single instance of a running database server

A very large database in a **database cluster** is partitioned over a number of smaller databases each located on a separate node of a computer cluster

**Database clustering** requires replication and sharding

**Database clustering** improve performance, availability, and scalability

The classes of database system that allow for **database clustering**:

- **NoSQL** systems: MongoDB, RavenDB, Cassandra, Amazon Aurora, ...
- **NewSQL** systems: ClustrixDB, NuoDB, CockroachDB, Pivotal GemFire XD, Alibase, MemSQL, VoltDB, ...
- **Improved OldSQL** systems: Oracle RAC, SQL Server (Windows server Failover Cluster), DB2 Cluster, PostgreSQL, MySQL Cluster, ...

# Cluster Computing

## Outline

[Computer Cluster](#)

[Big Data](#)

[Traditional Data Architectures](#)

[Meet Hadoop !](#)

[Big Data on Database Clusters](#)

[Big Data on Kubernetes](#)

# Big Data on Kubernetes

**Kubernetes** (K8) is a container or microservice platform that orchestrates computing, networking, and storage infrastructure workloads

in a plain language **Kubernetes** is an **orchestration platform** to manage any **containerized application**

A **Kubernetes** cluster consists of a single **master node** and potentially multiple corresponding **worker nodes**

The benefits of **Kubernetes**:

- horizontal scaling,
- automated rollouts and rollbacks,
- service discovery and load balancing,
- storage orchestration,
- self healing,
- batch execution,
- automatic binpacking

# References

White T., Hadoop The Definitive Guide: Storage and analysis at Internet scale, O'Reilly, 2015 (Available through UOW library)

Vohra D., Practical Hadoop ecosystem: a definitive guide to Hadoop-related frameworks and tools, Apress, 2016 (Available through UOW library)

Aven J., Hadoop in 24 Hours, SAMS Teach Yourself, SAMS 2017

Alapati S. R., Expert Hadoop Administration: Managing, Tuning, and Securing Spark, YARN and HDFS, Addison-Wesley 2017