23/01/24 **Lecture 1: Introduction to MapReduce**

**Introduction**

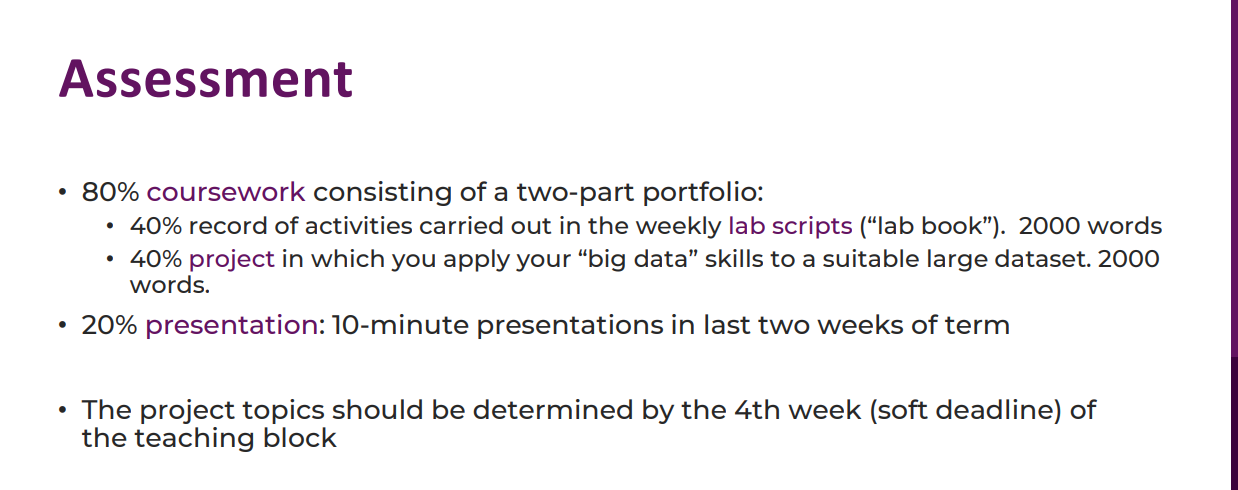
* Data becomes very large and becomes hard to store and process
* Solution: Big data technology

Big data technology cover in this module

* Hadoop
* Spark
* Tensorflow

Why new techs (old techs: SQL and etc)

* Scalability
* Parallel processing
* Distributed resources

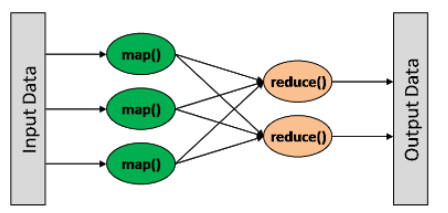


Deadline: late May

**Hadoop and MapReduce**

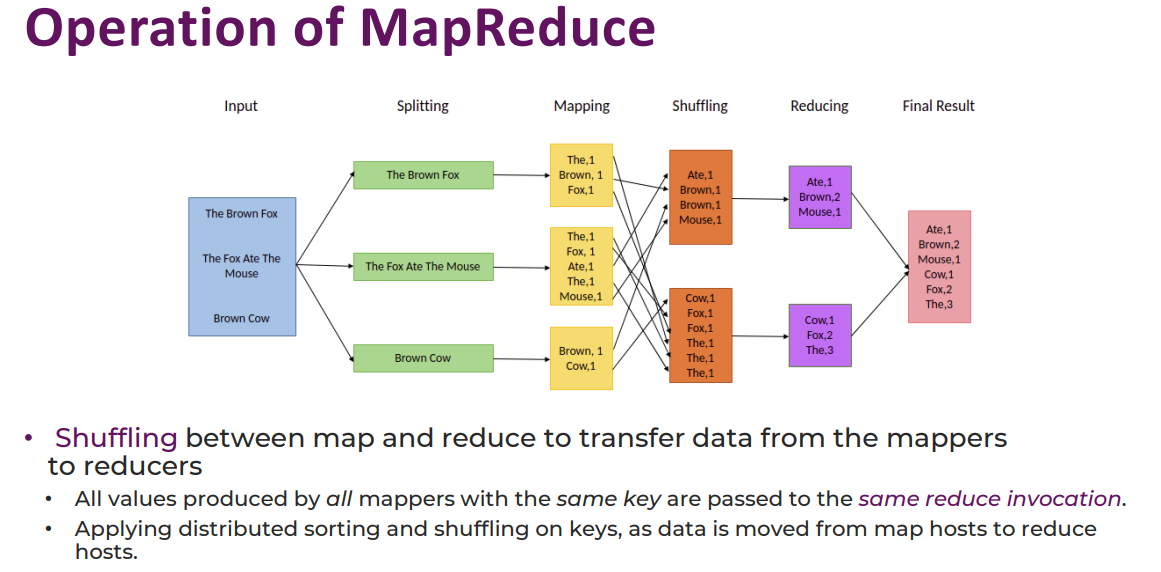
MapReduce: Programming framework for parallel processing

* Parallel programming is typically harder than SQL
* Map & Reduce function is a simple framework
* Scalable



* Map() process each input line by line
* Result in Key, value pairs (Python dictionary)
* Reduce() accumulates pairs with the same key

Example word count:



* As a programmer, we only need to write the Map() & Reduce() function

Pros & Cons

* Simple implementation
* Scalability
* Fault tolerance
* Not optimal for iterative processing (eg. machine learning). It’s slow.

Apache Hadoop

* Software to implement MapReduce
* Mediumware
* One machine will be a masternode to manage the whole cluster of nodes

Components (Architecture)

* Storage layer (HDFS: Hadoop File System)
* Resource Manager (YARN)
* Computational Engine (Map & Reduce)

