## Schedule of Course Activities: Session 26

## *[Cloud 519: Introduction to Cloud Computing Online-Based]*

## *[Instructor: John C. Chan]*

|  |  |
| --- | --- |
| **Overview of Session** |  |
| We will answer the following questions: | 1. Introduction to Hadoop 2. Major features. 3. … |

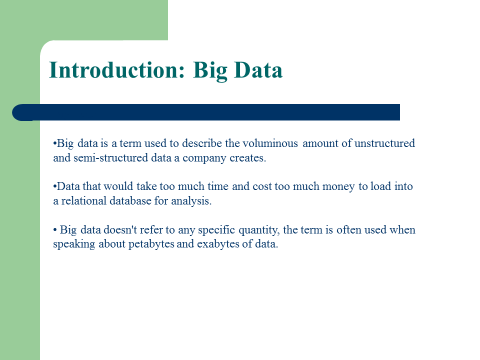
# **Apache Hadoop**

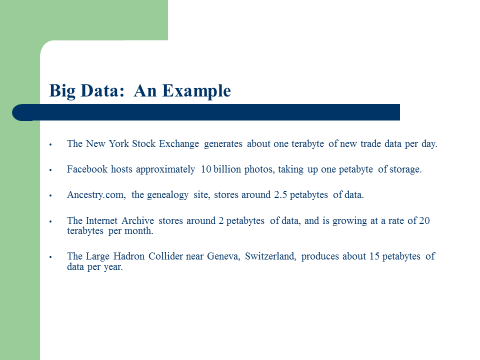
**Apache Hadoop** is an open-source software framework written in Java for **distributed storage** and **distributed processing** of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

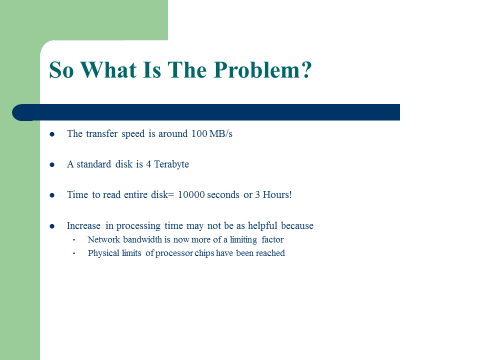
The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop MapReduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality[—nodes manipulating the data that they have on hand—to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking.

The base Apache Hadoop framework is composed of the following modules:

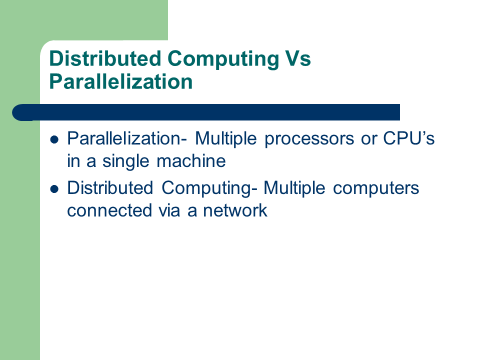
* *Hadoop Common* – contains libraries and utilities needed by other Hadoop modules;
* *Hadoop Distributed File System (HDFS)* – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;
* *Hadoop YARN* – a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications; and
* *Hadoop MapReduce* – a programming model for large scale data processing.

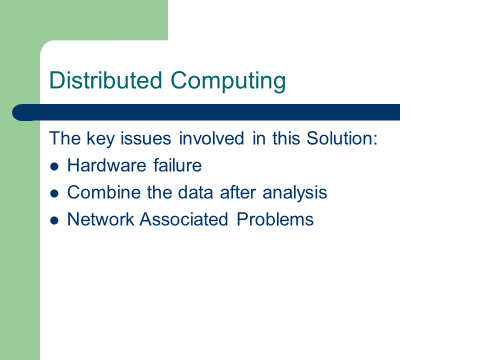
****

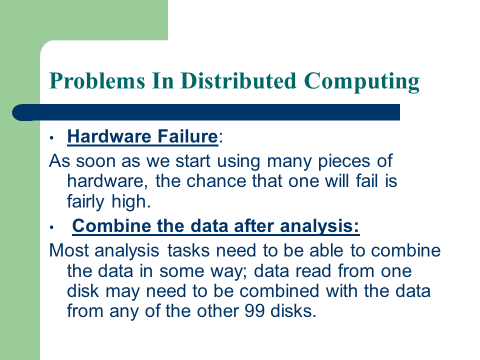
****

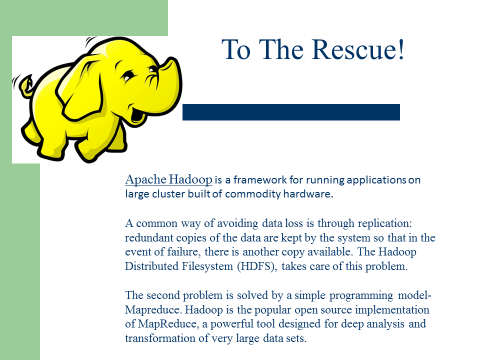
****

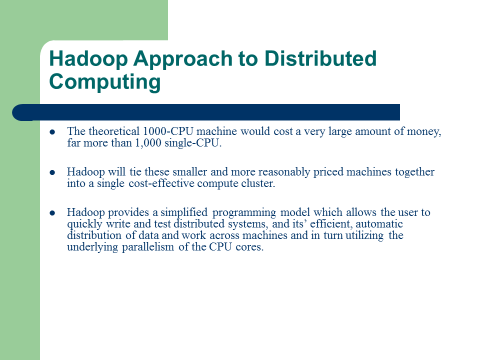
****

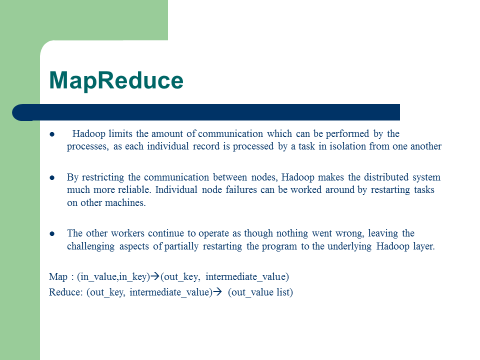
****

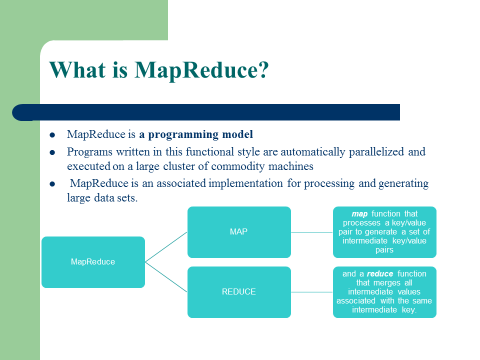
****

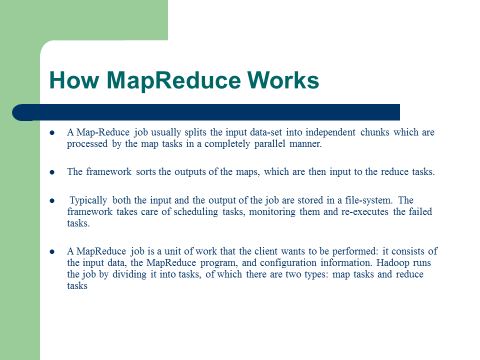
****

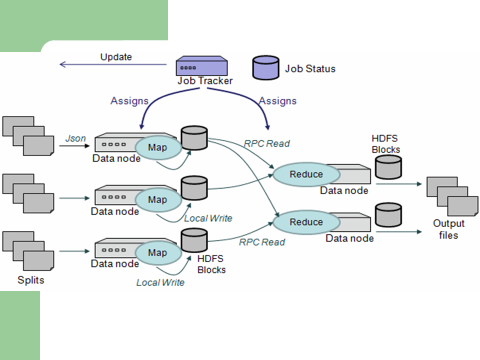
****

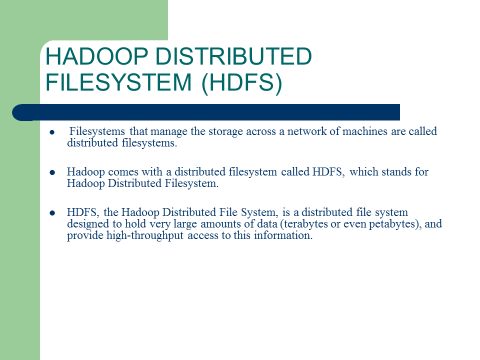
****

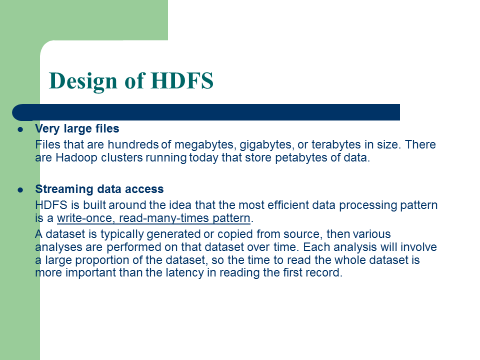


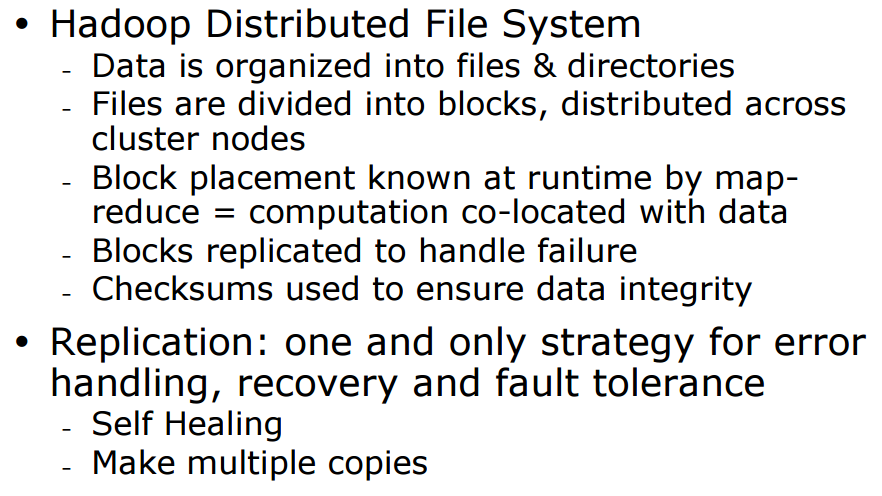






****

****



**Now, we have the basic ideas of Big Data and Hadoop, let’s watch this video to make common sense out of it.**

[**https://www.youtube.com/watch?v=FHVuRxJpiwI**](https://www.youtube.com/watch?v=FHVuRxJpiwI)

Key Take-Away:

* We have the data collected, but are scattered all over the places physically.
* Locate the data file, then send your job to it. (Job Tracker).
* Data duplication (case a node fail). MapReduce facilitate inter-node data exchange, and computation needs.
* Difference to centralized relational database. Hadoop is written once, read many times. BoSQL.
* Hadoop eco-system. Alternative tools to SQL.

Class Assignment: What is Hadoop most relevant to?

1. Big Data, Big organizations, Big companies, where data are located in many geographically area, and are constantly updated.
2. Small company, where data is largely located in the same facility.

End-of-Class Module.

Questions? Please email to me, or post it on Blackboard.

Thank you.