**ITEC452**

**Final Exam Study Guide**

**Lecture 8 Cloud Computing**

* What does cloud computing do? (slide 4)
* What is a cloud? (slide 6)
* Cloud Models
	+ Public Cloud
	+ Private Cloud
	+ Community Cloud
	+ Hybrid Cloud
* Cloud Service Models
* What is each of the following? What are benefits of each of the following?
	+ Software as a Service (SaaS)
	+ Platform at a Service (PaaS)
	+ Infrastructure as a Service (IaaS)
	+ More models: XaaS
		- Network as a Service (NaaS)
		- Database as a Service (DaaS)
		- Business as a Service (BaaS)
* Cloud Benefits
	+ High Efficiency, Reliability, Flexibility
	+ Applications as Utilities over Internet
	+ Manipulate and Configure Apps Online
	+ Cost Effective
	+ No Software Required
	+ Online Development and Deployment tools
	+ On-demand Self Service
	+ Resources Available on Network
* Cloud Computing Characteristics
	+ Essential Characteristics
		- On Demand Self-Service
	+ Common Characteristics
		- Broad Networks Access
		- Rapid Elasticity
		- Resource Pooling
		- Measured Services
		- Massive Scale
		- Resilient Computing
		- Homogeneity
		- Geographic Distribution
		- Virtualization
		- Service Orientation
		- Low Cost Software

**Lecture 9 Cloud Services Model: Big Data and Hadoop**

* Big Data
	+ Big Data’s 4V Big Challenges
		- Volume – Data Size
			* 40 Zettabytes (1021) of data is predicted to be created by 2020
			* 2.5 Quintillionbytes (1018) of data are created every day
			* 6 Billion (109) people have mobile phones
			* 100 Terabytes (1012) of data (at least) is stored by most U.S. companies
			* 966 Petabytes (1015) was the approximate storage size of the American manufacturing industry in 2009
		- Variety – Data Formats
			* 150 Exabytes (1018) was the estimated size of data for health care throughout the world in 2011
			* More than 4 Billion (109) hours each month are used in watching YouTube
			* 30 Billon contents are exchanged every month on Facebook
			* 200 Million monthly active users exchange 400 Million tweets every day
		- Velocity – Data Streaming Speeds
			* 1 Terabytes (1012) of trade information is exchanged during every trading session at the New York Stock Exchange
			* 100 sensors (approximately) are installed in modern cars to monitor fuel level, tire pressure, etc.
			* 18.9 Billion network connections are predicted to exist by 2016
		- Veracity – Data Trustworthiness
			* 1 out of 3 business leaders have experienced trust issues with their data when trying to make a business decision
			* $3.1 Trillion (1012) a year is estimated to be wasted in the U.S. economy due to poor data quality
* Hadoop
	+ Demand: Data Storage, Access, and Analysis
		- Hard drive storage capacity has tremendously increased
		- But the data read and write speeds to and from the hard drives have not significantly improved yet
		- Simultaneous parallel read and write of data with multiple hard disks requires advanced technology
	+ Hadoop is a Reliable Shared Storage and Analysis System
	+ Hadoop = HDFS + MapReduce + α
		- HDFS (Hadoop Distributed FileSystem) provides Data Storage
		- MapReduce provides Data Analysis
			* MapReduce = (Map Function) + (Reduce Function)
	+ HDFS
		- DFS (Distributed FileSystem) is designed for storage management of a network of computers
		- HDFS is optimized to store huge files with streaming data access patterns
		- HDFS is designed to run on clusters of general computers
		- HDFS was designed to be optimal in performance for a WORM (Write Once, Read Many times) pattern, which is a very efficient data processing pattern
		- HDFS was designed considering the time to read the whole dataset to be more important than the time required to read the first record
	+ MapReduce
		- MapReduce is a program that abstracts the analysis problem from stored data
		- MapReduce transforms the analysis problem into a computation process that uses a set of keys and values
		- MapReduce Architecture
			* MapReduce was designed for tasks that consume several minutes or hours on a set of dedicated trusted computers connected with a broadband high-speed network managed by a single master data center
		- MapReduce Characteristics
			* MapReduce uses a somewhat brute-force data analysis approach
			* The entire dataset (or a big part of the dataset) is processed for every query

🡺 *Batch* Query Processor model

* + - * MapReduce enables the ability to run an ad hoc query against the whole dataset within a scalable time
			* Many distributed systems combine data from multiple sources (which is very difficult), but MapReduce does this in a very effective and efficient way

**Lecture 10 Cloud Services Model: MapReduce and HDFS**

* Hadoop uses **HDFS** to move the **MapReduce** computation to several distributed computing machines that will process a part of the divided data assigned
* MapReduce
* Need to know how does it work?
	+ Jobs
		- Map Task
		- Reduce Task
	+ Node types for Job Execution
		- Jobtracker
		- Tasktracker
	+ Data Flow
		- Split
	+ [MapReduce paper by Google](file:///H%3A%5Cpublic_html%5Cclasses%5Citec452_fall2016%5CClassNotes%5CMapReduce1.pdf)
		- Needs to be able to explain:
			* the execution overview (Section 3.1)
			* how it reacts at a worker failure (Section 3.3)
* HDFS
	+ [Hadoop Distributed File System by Yahoo](file:///H%3A%5Cpublic_html%5Cclasses%5Citec452_fall2016%5CClassNotes%5CHadoopDistributedFileSystem.pdf)
		- Hadoop project components (Section 1; Table 1)
		- Architecture
			* Name Node: What is Name Node? How does it work? (Section II.A)
			* Data Nodes: What is Data Node? How does it work? (Section II.B)
			* Image and Journal: What are these? How do they work? (Section II.D)
		- File I/O Operations and Replica Management
			* How the block placement works? (Section III.B)
			* How the replication management works? (Section III.C)