Developing a Remote Sensing and Cloud Computing Curriculum for the Association of Computer/Information Sciences and Engineering Departments at Minority Institutions (ADMI)

Jazette Johnson
Computer Science
Spelman College
Atlanta, Georgia
jjohn133@scmail.spelman.edu

Jimil Perkins
Computer Science
Norfolk State University
Norfolk, Virginia
j.m.perkins@spartans.nsu.edu

Jerome Mitchell (Mentor)
Computer Science
Indiana University
Bloomington, Indiana
jeromitc@indiana.edu

Abstract - **In the past decade, online learning initiatives have become increasingly comprehensive and have allowed students to be unburdened from learning complex subjects in a traditional teach-learn environment. Universities have recognized the need to adapt new teaching-learning approaches for meeting students' diverse inadequacies. Cloud computing, which offers a scalable and flexible approach to storing, processing, and analyzing big data, has benefited from a variety of science applications except for remote sensing. The research explored the potential for a cloud computing and remote sensing curriculum through the use of video resources and hands-on assessments. This research discusses a curriculum for coupling two diverse research areas, cloud computing and remote sensing. The solution acquired information about cloud computing and remote sensing in order to develop five 15-20 minute self-contained modules. Understanding the challenges recognized by minority serving institutions in adapting from a teaching-learning environment to an online environment was also explored.**

Keywords- Cloud Computing, storage area networks, virtual private networks, computer networks, Google, web services, Digital video broadcasting, remote sensing, curriculum development, educational programs

I. INTRODUCTION

In fall 2006 19.6% of students were learning online and in fall of 2011 that percentage was increased to 32.0%. Online education is increasing [1]. According to previous research, findings show that in 2012 56.4% of students taking online courses had the same learning outcome compared to students taking courses face to face with professors. With the increasing of technology, online education will continue to increase [1]. Online education has the ability to teach subjects that some professors cannot teach at historically black colleges and universities (HBCU). Some HBCU's have limit to no resources to teach cutting edge research options like cloud computing. If the institutions have resources to teach subjects like cloud computing, the downfall is not having the teachers who can communicate the subject effectively. Producing cutting edge online courses for Massive Open Online Courses (MOOC) allows students at ADMI institutions to experience cutting edge education in virtual time and train the next generation of engineers.

II. RELATED WORK

This project used two highly known online educational services as a reference for the course, MIT OpenCourseWare (OCW) and Stanford Online. Massachusetts Institution of Technology (MIT) has developed online courses on a cloud called MIT OpenCourseWare (OCW). OCW allows materials that are being taught in MIT's classrooms available on the Internet for no charge. MIT OpenCourseWare

currently has 2,150 courses and about 125 million people who have visited the site [2]. Stanford Online is an educational tool for people to experience Stanford University's high quality education by unleashing innovation in online learning.

III. Remote Sensing and Cloud Computing

A. Remote Sensing

Remote Sensing is the art and science of obtaining information about an object without being in direct physical contact with the object [3].

B. Cloud Computing

Cloud Computing is the next big thing in computing and Internet evolution because it is allowing you to have access to the services wherever and whenever you need it. Cloud computing includes three main services, they include: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Infrastructure as a Service supplies resources to data centers that hold large pools of information. The customer can use the Internet or dedicated virtual private networks. Examples of IaaS include networks, virtual machines, servers, and storage. Software as a Service is a delivery method for software that provides access to software and its functions remotely as a Web-based service [4]. Examples of Software as a Service include email, virtual desktop, and games. Platform as a Service allows IT to develop, test, deploy, host and also update from a single streamline environment [5]. Examples include execution runtimes, development tools, and webservers. With the use of these services school can freely access, edit and develop with for teaching, learning etc. The services also allow businesses to have full control of the business documents because they the have ability to process, store, and analyze all documents.

C. How can they be related?

Remote sensing creates a plethora of data, which creates difficulty with storing, processing, and analyzing the big data. The Center for Remote Sensing of Ice Sheets (CReSIS) collects a plethora of big data, which makes it difficult to process all of the data. A possible solution is to find an easy way to store, analyze, and process the big data within a cloud. The cloud's scalable and flexible environment produces a simple way for storing, processing, and analyzing. One of the biggest problems with remote sensing is processing, taking the raw remote sensing data and putting them to images. Using infrastructure as a service a developer can use MapReduce to process the data. MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. Many companies such as Google, Netflix, and Facebook use MapReduce to make their websites user friendly.

IV. Methodology

Research was conducted to become famailar with topics such as Big Data, Parallel and Distributed Computing, Designing Parallel Programs, Cloud Computing and MapReduce. Then modules were created to give a indroduction to these topics.

A. Big Data

1) Module 1: What is Big Data?

Big Data is large and complex data sets that exceeds and organization's ability to handle, store, analyze and process.

2) Module 2: Three V's of Big Data

Volume, Variety and Velocity of Information are the three key factors of Big Data.

3) Module 3: Volume of Information

There has been an increase in data volume. Previously, the increase in data caused a storage issue. As time progressed storage cost decreased. Issues arose like determining relevance amongst large volumes of data and finding the value from the data.

4) Module 4: Variety of Information

There is an abundance of different formats that represent Big Data. All these different types must be analyzed and processed.

5) Module 5: Velocity of Information

There is an ongoing tug-of-war with the data production versus data processing speeds. Problems occur when data cannot be processed fast enough to meet industry demands.

6) Module 6: Other Factors of Big Data

Veracity is the truth behind Big Data. Data flow can become inconsistent with peaks, especially with any social media involved. Complexity occurs because large volumes of data are coming from a variety of different sources.

7) Module 7: Uses

Big Data ultimately is used to increase knowledge. First data is extracted, and then analyzed.

8) Module 8: Challenges

It is hard to find value within huge masses of data and data can often get too large or too varied to deal with effectively.

9) Module 9: Solutions

There are two solutions. All the data can be analyzed or the relevance of the data is determined upfront and only the relevant data is analyzed.

10) Module 10: Technological Advancements

There is now cheap abundant storage and server processing capacity. Due to Moore's Law, there are now faster processors. Also, new methods like parallel processing and Map-Reduce have helped the Big Data era flourish.

B. Distributed and Parallel Computing

1) Module 1: Distributed Computing

Distributed Computing is a computing concept that refers to multiple computer systems working on a single problem. The system must be networked in order to communicate.

2) Module 2: Benefits

There are many benefits to using distributed computing some include scalability, redundancy, low cost, easy access, and the simplicity.

3) Module 3: Client/Server Model

The client server model explains the basic architecture of distributed computing systems. Functions separated into two parts client and servers. Clients are programs that use services other programs provide while servers are programs that provide the services

4) Module 4: Parallel Computing

Parallel Computing is a division of distributed computing which refers to multiple processors working on a single problem. Problem is broken into parts that can be solved concurrently. The parts are then broken down even further into instructions that can be carried out simultaneously.

5) Module 5: Uses

Parallel Computing is used to model many challenging problems in science and engineering and has commercial and industry applications to develop faster computers.

6) Module 6: Why use parallel?

Parallel computing saves time and money. Parallel computing helps to solve larger problems, provides concurrency, and it surpasses the limits of serial computing.

7) Module 7: Flynn's Classical Taxonomy

Flynn's Classical Taxonomy is a method that classifies multi-processor computer architectures based on their dimension and state.

8) Module 8: Single Instruction Single Data Stream

These are serial computers with one instruction stream and one data stream. These are the oldest most common computer.

9) Module 9: Single Instruction Multiple Data Stream

These are parallel computers with a single instruction stream and multiple data streams.

10) Module 10: Multiple Instruction Single Data Stream

These are parallel computers with multiple instruction streams and a single data streams. Only a few of these computers actually exist.

11) Module 11: Multiple Instruction Multiple Data Stream

These are parallel computers with multiple instruction and data streams. They are the most common parallel computer and most modern supercomputers. Most Multiple Instruction Multiple Data units have Single Instruction Multiple Data execution sub-units

C. Designing Parallel Programs

1) Module 1: Auto vs. Manual

Designing Parallel Programs is usually a very manual process however there are conversion tools that help convert serial programs to parallel programs. Some examples are pre-processors and parallelizing compiler. Compiler works in two ways, fully automatic or programmer directed.

2) Module 2: Understanding the Problem

In order to design a parallel program the problem must be understood. The user must determine the parallelizability of the program. Also the user should identify bottlenecks, inhibitors and dependencies.

3) Module 3: Partitioning

Program must be broken down into chunks and this is called partitioning or decomposition. There are two types of decomposition: Domain and Functional. With domain decomposition data linked with the problem is decomposed. On the other hand, functional decomposition bases focus on the performed computation rather than data manipulation.

4) Module 4: Communication

Communication is dependent on the problem at hand yet most parallel applications require the share of data that can only be done through communication

5) Module 5: Factors of Communication

There are many factors that come into play when dealing with communication. Cost of communication, latency vs. bandwidth, synchronous vs. asynchronous, and scope of communication are just few factors to consider.

6) Module 6: Synchronization

Synchronization is a concept that deals with multiple processes joining up at a certain point in goal to further receive instructions or reach agreement point. Synchronization has a great impact on the overall performance of the program.

7) Module 7: Data Dependencies

Data Dependencies are important because they are the primary inhibitors of parallel programs. Dependence occurs when the order of statement execution affect the result of the program while data dependence occurs when more than one task share the same location for storage.

8) Module 8: Load Balance

Load Balance refers to distributing equal amounts of work so that all tasks are kept busy all the time. It is importance for performance, so all the work should be partitioned equally.

9) Module 9: Granularity

Granularity is the qualitative measure of the ratio of computation to communication. There are two types of parallelism, fine-grain and coarse-grain.

10) Module 10: Input / Output Negative Aspects

Input and Output operations are inhibitors to parallelism. They have troubles with availability, overwriting, dependability on file server, and crashes.

11) Module 11: Input / Output Tips

There are parallel file systems that can be investigated to help with input output operations. The overall goal is to reduce input output operations.

D. Cloud Computing

1) Module 1: Why is Cloud Computing Important?

Cloud computing is the next big thing in computing and Internet evolution because it is allowing users to have access to the services wherever and whenever they need it.

2) Module 2: What is Cloud Computing?

Cloud computing allows applications, infrastructures and business to be available to you as a service.

3) Module 3: Infrastructure as a Service (IaaS)

IaaS supplies their resources to data centers that hold large pools of information. With IaaS customers can use the Internet or dedicated virtual private networks. To start up the infrastructure the cloud users install operating- system images and their application software on the cloud infrastructure. IaaS are the networks, Virtual machines, servers, and storage. Example infrastructures as a service include MapReduce, Google Compute Engine, and HP Cloud.

4) Module 4: Platform as a Service (PaaS)

Platform as a Service is a development platform where the development tool is hosted in a cloud and accessed through a browser. This allows the user to develop and run their software solutions on a cloud platform. PaaS are the execution runtime, development tools, and webservers.

Examples of Platform as a Service include Google App Engine and Windows Azure.

5) *Module 5: Software as a Service (SaaS)*
Software as a Service is provided by could providers who manage the infrastructures and platforms that run the application, users are able to access application software and databases. SaaS allows you to access your files the same way you will access them on your home computer, but via your browser. SaaS are the email, virtual desktop, and games. One Example of SaaS is Google Docs; there is many more application that serves as Software as a service.

E. MapReduce

1) *Module 1: What is MapReduce?*
MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster.

2) *Module 2: Who is using MapReduce?*
Companies such as Google, Yahoo, Facebook, Apple, EBay, IBM, Amazon, Oracle and Cisco etc. use MapReduce to process the large amounts of data the receive.

3) *Module 3: MapReduce Structure*

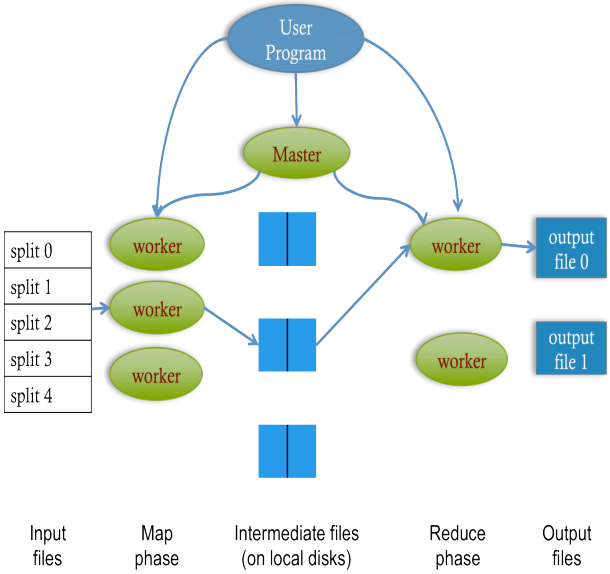

Figure 1: MapReduce Structure, the details of the MapReduce phases can be view in Modules 4, 5, & 6.

4) *Module 4: Map*
The map phase breaks up large portions of work into smaller ones and then takes action on each portion.

5) *Module 5: Sort*
The sort takes the data and sorts the information into groups determined any way determined within your program.

6) *Module 6: Reduce*
The reduce phase is the data collation or processing phase. Reduce combines the many results from the map step into a single output

V. Conclusion

We have created five modules to be used as educational pre-requisites to a Cloud Computing and Remote Sensing Curriculum. Some of the topics covered were Big Data, Parallel and Distributed Computing, Designing Parallel Programs, Cloud Computing and MapReduce. Modules were designed to appeal to multiple learning styles. Visual and auditory enhancements were included in our modules in order to incorporate multiple learning styles. These modules will be placed on the Massive Open Online Course provided by Google when the page is completed.

VI. Future Work

In further production of the remote sensing and cloud-computing curriculum the modules will be expanded and user experience studies will be conducted with students who attend ADMI institutions (i.e Spelman College, Norfolk State University, Elizabeth City State University). The user study will allow the student to take the course online; feedback will be gathered to see how the curriculum is effective. A Message Passing Interface (MPI) and Hadoop virtual appliance will be developed so students can apply theoretical concepts gained from the curriculum.

VII. References

[1] Allen, E., & Seaman, J. (2013). Changing course: Ten years of tracking online education in the united states. (pp. 24-41).

[2] Abelson, H. (2007). The creation of opencourseware at mit.Journal of Science Education and Technology

[3] Jensen, 2004. Introductory.Digital Image Processing. 3rd edition.

[4] Sriram , I., & Khajeh-Hosseini, A. (n.d.). Research agenda in cloud technologies.

[5] Waggner, S. (2010). Cloud computing: Managing data in the cloud. UC Berkeley
iNews