Course Description

IT Information Technology 471 Big Data on Cloud Systems (3:3:0)

This course focuses on cloud-based Big Data solutions, methods, and tools. Students learn to create and use big data analytical environments using cloud data services and tools. Also covered is employing best practices for designing big data cloud solutions for performance, security, and cost effectiveness.

Prerequisites

The prerequisites for this course are:

- IT-109 (Intro to Computer Programming - Python),
- IT-214 (Database Fundamentals),
- STAT-250 (Introductory Statistics) or STAT-344,
- IT-442 (Cloud Infrastructure).

A grade of "C" or better must be achieved in all prerequisite courses before a student is qualified to take this course. The prerequisite courses must be completed prior to, not concurrently with, this course.

Course Outcomes

Upon successful completion of this course, students will demonstrate:

- How to host, manage, and analyze large cloud-based datasets
- How to identify and use cloud services to support big data analysis projects
  ○ Describe the advantages and challenges of hosting big data projects on cloud systems
  ○ Explore and compare selected commercial big data solutions
- How to optimize cloud-based data analytics for scalability, performance, and security

Supported Student Outcomes at the Program Level

- SO1: Analyze a complex big data computing problem and apply principles of distributed/cloud computing and other relevant disciplines to identify solutions.
- SO2: Design, implement, and evaluate a big data cloud-based solution to meet a given set of computing requirements in the context of the program’s discipline.
- SO3: Identify and analyze user needs and take them into account in the selection, creation, integration, evaluation, and administration of cloud-based systems.
- SO4: Recognize professional responsibilities and make informed judgments in big data analytics computing practice based on legal and ethical principles
# Major Topics and Schedule

The course covers the hardware and software architecture components required to conduct big data analytics in cloud computing environments, including:

- Cloud Computing Infrastructure Review
  - Distributed Storage and Processing
- Big Data Analytics Concepts, Methods, and Software
  - Apache Hadoop and Spark, MapReduce, Hive, Kafka, Streaming data analysis
  - Use cases: Artificial Intelligence, Machine Learning, Predictive Analytics
  - Data analytics programming tools (Python, R, SQL, NoSQL)
- AWS tools and services for Big Data
- Data classification and management, security, privacy, ethics
- Commercial cloud products supporting big data analytics
- Case studies of both successful and unsuccessful big data projects

## Week 1: Review of General Cloud concepts

- Cloud Computing Definitions
- Cloud Computing General Reference Architecture

Assigned Readings:
- CCSA*: Chapter 1: Introduction to Cloud Computing
- NIST: SP 800-145: The NIST Definition of Cloud Computing
- NIST: SP 500-292: Cloud Computing Reference Architecture

## Week 2: Foundational AWS Cloud Services

- Distributed Storage and Processing
- Elastic Compute Cloud (EC2)
- Simple Storage Service (S3)

Assigned Readings:
- CCSA: Chapter 2: Virtual Machines and Compute Services
- CCSA: Chapter 5.1: Cloud Storage

## Week 3: Big Data Analytics Concepts

- Big Data Definitions
- Big Data General Reference Architecture

Assigned Readings:
- CCSA: Chapter 8: Big Data Analytics in the Cloud
- NIST: SP 1500-1: Definitions
- NIST: SP 1500-6: Reference Architecture
<table>
<thead>
<tr>
<th>4-5</th>
<th><strong>Basic Data Analytics Programming Tools</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Python, R, SQL</td>
</tr>
</tbody>
</table>

**Assigned Readings:**

• Practical Statistics for Data Scientists 2nd Ed.

<table>
<thead>
<tr>
<th>6-7</th>
<th><strong>Big Data Analytics Methods</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Apache Hadoop and Spark</td>
</tr>
<tr>
<td></td>
<td>• MapReduce, Hive, Kafka</td>
</tr>
<tr>
<td></td>
<td>• Amazon EMR</td>
</tr>
</tbody>
</table>

**Assigned Readings:**

• CCSA: Chapter 10: Batch Analytics

<table>
<thead>
<tr>
<th>8</th>
<th><strong>Data classification and management</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Data Security &amp; Privacy</td>
</tr>
</tbody>
</table>

**Assigned Readings:**

• CCSA: Chapter 13: Cloud Security
• NIST: SP 1500-4: Security and Privacy

<table>
<thead>
<tr>
<th>9-10</th>
<th><strong>Case studies: successful &amp; unsuccessful big data projects</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Use cases: Artificial Intelligence, Machine Learning, Predictive Analytics</td>
</tr>
</tbody>
</table>

**Assigned Readings:**

• NIST: SP 1500-3: Use Cases and Requirements

<table>
<thead>
<tr>
<th>11</th>
<th><strong>Commercial cloud products supporting big data analytics</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Microsoft Azure, Google Cloud, Oracle Cloud, IBM Cloud, …</td>
</tr>
</tbody>
</table>

**Assigned Readings:**

• [https://cloud.google.com/](https://cloud.google.com/)
• [https://www.ibm.com/cloud](https://www.ibm.com/cloud)
• [https://azure.microsoft.com/en-us/](https://azure.microsoft.com/en-us/)
• [https://www.oracle.com/cloud/](https://www.oracle.com/cloud/)

<table>
<thead>
<tr>
<th>12</th>
<th><strong>Big Data Ethics</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Examine and discuss key ethical and legal issues for entry level professionals in data science and analytics</td>
</tr>
</tbody>
</table>

**Assigned Readings:**

• Selected reading from Cyberethics: Morality and Law in Cyberspace 7th Ed., R.A. Spinello, 2021 Jones & Bartlett Learning
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Individual/Group work &amp; consultations on projects</td>
</tr>
<tr>
<td>14</td>
<td>Project Presentations</td>
</tr>
<tr>
<td>15</td>
<td>Final Projects &amp; Course Wrap-up</td>
</tr>
</tbody>
</table>

**Textbooks and References**

Required Textbook:

- AWS Certified Data Analytics Study Guide

Recommended References:

- Cloud Computing Solutions Architect (CCSA)
- Practical Statistics for Data Scientists 2nd Ed.
- Becoming an AWS Certified Data Analytics
  - [https://towardsdatascience.com/becoming-an-aws-certified-data-analytics-new-april-2020-4a3ef0d9f23a](https://towardsdatascience.com/becoming-an-aws-certified-data-analytics-new-april-2020-4a3ef0d9f23a)

NIST References:

- NIST Cloud and Big Data Definitions, Reference Architectures, Use Cases
  - [https://bigdatawg.nist.gov/V3_output_docs.php](https://bigdatawg.nist.gov/V3_output_docs.php)
  - [https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r2.pdf](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r2.pdf)
  - [https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6r2.pdf](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6r2.pdf)

AWS References:

- Big Data Analytics Options on AWS
  - [https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf)
- Building Big Data Solutions with Amazon EMR and Amazon Redshift
  - [https://www.youtube.com/watch?v=1AHGcLwvinI&feature=youtu.be](https://www.youtube.com/watch?v=1AHGcLwvinI&feature=youtu.be)
- Analyze Big Data with Hadoop
  - [https://aws.amazon.com/getting-started/hands-on/analyze-big-data/](https://aws.amazon.com/getting-started/hands-on/analyze-big-data/)
- Amazon EMR
  - [https://aws.amazon.com/emr/](https://aws.amazon.com/emr/)

**Student Computer Technical Requirements:**

Students must have access to a [Mason-recommended](#) laptop computer (including a *webcam*, and *Internet access*), and an [AWS Classroom](#) account.

Hardware: You will need access to a Windows or Macintosh computer with at least 2 GB of RAM and to a *fast, reliable broadband Internet connection* (e.g., cable, DSL). For optimum visibility of course material, the recommended computer monitor and laptop screen size is 13-inches or larger. You will need computer speakers or
headphones to listen to recorded content. A headset microphone is recommended for live audio sessions using course tools like Blackboard Collaborate.

**Grading**

Grades will be earned in accordance with the GMU Grading System for undergraduate students. See http://catalog.gmu.edu/policies/academic/ under Grading for more information.

*Letter grades will be assigned according to the following scale:*

<table>
<thead>
<tr>
<th>Numeric Score</th>
<th>Letter Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>97 – 100</td>
<td>A+</td>
</tr>
<tr>
<td>93 – 96</td>
<td>A</td>
</tr>
<tr>
<td>90 – 92</td>
<td>A–</td>
</tr>
<tr>
<td>87 – 89</td>
<td>B+</td>
</tr>
<tr>
<td>83 – 86</td>
<td>B</td>
</tr>
<tr>
<td>80 – 82</td>
<td>B–</td>
</tr>
<tr>
<td>77 – 79</td>
<td>C+</td>
</tr>
<tr>
<td>73 – 76</td>
<td>C</td>
</tr>
<tr>
<td>70 – 72</td>
<td>C– Passing*</td>
</tr>
<tr>
<td>60 – 69</td>
<td>D Passing*</td>
</tr>
<tr>
<td>0 – 59</td>
<td>F Failing</td>
</tr>
</tbody>
</table>

* Grades of "C–" and "D" are considered passing grades for undergraduate courses. However, a minimum grade of "C" is required in the BSIT program for any course that is a prerequisite for one or more other courses.

*Final grades will be determined based on the following components:*

<table>
<thead>
<tr>
<th>Assignments</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercises</td>
<td>20%</td>
</tr>
<tr>
<td>Lab Projects (3)</td>
<td>50%</td>
</tr>
<tr>
<td>Final Project</td>
<td>20%</td>
</tr>
<tr>
<td>Discussion Board / Quizzes</td>
<td>10%</td>
</tr>
</tbody>
</table>

*These components are outlined in the following sections.*

**Assigned Exercises:**

Throughout the term students will be assigned exercises in order to practice concepts learned. Upon completion, students will be required to submit completed exercise to Blackboard in order to receive credit. Late submissions will be penalized.

**Lab Projects:**

Students are required to individually design, build, publish, and submit individual projects in accordance with the requirements. For more information about the projects, check the Project folder on Blackboard.

**Final Project:**

There is a final project which will be used to evaluate your mastery of terms and concepts along with the successful application of those terms and concepts. You should prepare for the final project by understanding the material presented in the course and by successfully completing the assigned exercises.
Discussion Board:

Students will be required to stay engaged via the online discussion board. The instructor will post weekly graded discussion questions or other topics to the discussion board. Students are expected to check the discussion board frequently (at least 1-2 times per week), read the latest postings, and reply to classmates and instructor postings.

Attendance Policy

Students are expected to attend each online class session, to complete any required preparatory work and to participate actively in presentations, demos, discussions and exercises.

Students are expected to make prior arrangements with Instructor if they know in advance that they will miss any online session or assignment and to consult with the Instructor as soon as possible. Contact the instructor in advance if you have a serious problem that prevents you from meeting course requirements.

Communications

Registered students will be given access to a section of the Blackboard Learning System (http://mymasonportal.gmu.edu/) for this course. Blackboard will used as the primary mechanism to disseminate course information, including announcements, lecture slides, homework and other assignments, and scores for homework and exams.

Communication with the Instructor on issues relating to the individual student must be conducted using GMU email, via telephone, or in person - not in the public forums on Blackboard. Email is the preferred method. Federal privacy law and GMU policy require that any communication with a student related in any way to a student's status be conducted using secure GMU systems – if you use email to communicate with the Instructor you MUST send messages from your GMU email account.

Privacy

Instructors respect and protect the privacy of information related to individual students. As described above, issues relating to an individual student will discussed via email, telephone or in person. Instructors will not discuss issues relating to an individual student with other students (or anyone without a need to know) without prior permission of the student.

Assessable work will be returned to individual students directly by the Instructor (or by a faculty or staff member or a Teaching Assistant designated by the Instructor, or via another secure method). Under no circumstances will a student's graded work be returned to another student. Faculty and staff will take care to protect the privacy of each student's scores and grades.

Disability Accommodations

Disability Services (DS) at George Mason University is committed to providing equitable access to learning opportunities for all students by upholding the laws that ensure equal treatment of people with disabilities. If you are seeking accommodations for this class, please first visit http://ds.gmu.edu/ for detailed information about the Disability Services registration process. Then please discuss your approved accommodations with me. Disability Services is located in Student Union Building I (SUB I), Suite 2500. Email: ods@gmu.edu | Phone: (703) 993-2474

Honor Code

All members of the Mason community are expected to uphold the principles of scholarly ethics. Similarly, graduating students are bound by the ethical requirements of the professional communities they join. The ethics requirements for some of the communities relevant to IT graduates are available via the following links:
- ACM Code of Ethics and Professional Conduct
  - https://www.acm.org/code-of-ethics
- IEEE Code of Ethics
  - https://www.ieee.org/about/corporate/governance/p7-8.html
- Ethical Guidelines for Statistical Practice

On admission to Mason, students agree to comply with the requirements of the Honor Code (https://oai.gmu.edu/mason-honor-code/) at George Mason University. Student members of the George Mason University community pledge not to cheat, plagiarize, steal, and/or lie in matters related to academic work. The Honor Code will be strictly enforced in this course.

Any use of the words or ideas of another person(s), without explicit attribution that clearly identifies the material used and its source in an appropriate manner, is plagiarism and will not be tolerated. The Instructor reserves the right to use manual and/or automated means (including such services as SafeAssign to detect plagiarism in any work submitted by students for this course, and to direct Teaching Assistants and/or other faculty and/or staff members to do likewise in support of this course.