SMU
School of
Computing and
Information Systems

Course Code:	**IS459**
Course Name:	**Big Data Architecture**
When was the course design document last verified by the Course Manager:	**Jan 2024**

NOTE: The information given in this document is for reference only; the updates given during the class sessions and/or eLearn will supersede the information given in this document.

IMPORTANT NOTICE

Please note that the below information is for Term 2, AY2024-25 and earlier. For AY2025-26 and onwards, there will be a new public course catalogue. Please watch out for email updates from RO/in BOSS.

1. Synopsis

Big data has become an essential part of our digital world in the last decade. Governments, e-commerce websites, even short video platforms, are now relying on big data technologies to gain business insight and design their strategies. Different from the traditional data analytics industry, the infrastructure of big data is mostly built on top of cheap commodity PCs and open source software. This trend has successfully lowered the deployment as well as operational cost of such big data platforms. On the other hand, it raises new challenges to data engineers and scientists when building their own system with such infrastructure.

This course aims to bridge the gap between the big data practice and the skills of undergraduate students in data infrastructure. It will provide an overview of big data infrastructure, which enables the student to build their own system based on the data characteristics and processing demands. It will get the students involved in the whole process of big data system building, i.e., the design, planning and implementation. It will also cover additional tools and techniques crucial to the success of big data, including data visualization, and monitoring.

Students are expected to learn skills of 1) software development infrastructure in mainstream data-driven IT products; 2) data crawling and extraction frameworks for web-based data collection; 3) data storage architecture for highly variant data in big data systems; 4) massive data processing frameworks; 5) real-time and streaming data processing frameworks; 6) data visualization tools for big data analytics; 7) orchestration of the open source systems with high-level data interface.

2. Prerequisites/Co-requisites

Pre-Requisite: IS112 Data Management OR IS105 Business Data Management OR ANLY104/IS217 Analytics Foundations OR MGMT108 Intro to Business Analytics
(Please check Course Catalogue in BOSS for updated information!)

3. Course Areas

Business Options
Econ Major Rel/Econ Options
Business-Oriented Electives
ITSD Major
Social Sciences/PLE Major-related
IS (Business Analytics Track)
IS (Software Development Track)
IS (Digitalisation & Cloud Solutions Track)
SMU-X

(Please check Course Catalogue in BOSS for updated information!)

4. Course Objectives

Upon completion of the course, students will learn:

- ❖ System development infrastructure
  - ✓ Source code management with Git and GitHub
  - ✓ Task Management with ClickUp
  - ✓ Free document system, Confluence
- ❖ Data crawling and extraction framework
  - ✓ Extract information from web pages with Scrapy
- ❖ Data storage system
  - ✓ Choose the right storage architecture based on data characteristics
  - ✓ Use data store and document store for various types of data
- ❖ Massive data processing frameworks
  - ✓ Install and deploy Hadoop and Spark
  - ✓ Program big data processing logics with Hadoop and Spark
- ❖ Data interface between modules
  - ✓ Process JSon files
  - ✓ Adopts GraphQL as the data interface
- ❖ Other open source big data tools
  - ✓ Visualise data results with D3.js
  - ✓ Monitors online data with Prometheus

5. Competencies

1. Basic engineering skills as a data engineer
2. Big data project scoping
3. Architecture design and implementation

6. Teaching Staff

Faculty:

- Sachin GUPTA, sachingupta@smu.edu.sg

7. Course Assessments

Assessment Categories	Type	Weightage (%)
Class Participation	Individual	10
Assignments	Individual	20

Project	Group	40
Quizzes	Individual	20
Topic Research & Sharing	Group	10
Total		**100**

Note: Weightage may vary depending on the year the course is offered and the faculty/instructor teaching the course.

8. Course Assessments Details

Class Participation (10%)

- Contribution to online discussion in Slack discussions
- In-class discussion, presentation, quizzes

Assignment (20%)

- Individual assignments

Quizzes (20%)

- There will be TWO (2) quizzes.
- It will focus on the key concepts taught throughout the term *plus* some topics from assigned videos/readings.

Topic Research & Sharing (10%)

- Each group (of 5-6 students) will research a Big Data topic and share with classmates. Each presentation will be 10-15 minutes and delivered during lesson time.
- Topic groups will be *assigned* by the teaching team.

Project (40%)

The project is intended to complement the class materials by getting students to investigate selected topics in greater depth or breadth. The project should be done in teams of 4-6 students. Students must **form their own groups**.

9. Lesson Plan

Week	Topic	Remarks
1	Course Introduction & tool setup	
2	Data Ingestion - Part 1	
3	Data Ingestion - Part 2	*Register project group in e-Learn*
4	Data Transformation - Big Data Tools	
5	Big Data Storage Data Pipeline Orchestration	*Assignment 1 Released*
6	Big Data on Cloud - Part 1	*Quiz 1*
7	Special Topic	
8	*Recess (No Lesson)*	
9	Project proposal presentation	*Project Proposal* *Assignment 1 Due*
10	Big Data on Cloud - Part 2	*Assignment 2 Released*
11	Big Data on Cloud - Part 3	
12	Guest speaker	*Quiz 2*
13	Final project presentation	*Final Presentation* *Final Project Submission*
14	Study Week	*Assignment 2 Due*

Big Data Tools/Technologies
Git, Scrapy, Hadoop, Spark, Flink, Celery, D3.js, GraphQL, Prometheus, Databricks tools, Neo4j tools, AWS Big Data services, Microsoft Azure, and others.

10. Resources

- Supplementary reading/watching materials will be provided 1-2 weeks prior to each lesson.

11. University Policies

Academic Integrity

All acts of academic dishonesty (including, but not limited to, plagiarism, cheating, fabrication, facilitation of acts of academic dishonesty by others, unauthorized possession of exam questions, or tampering with the academic work of other students) are serious offences.

All work (whether oral or written) submitted for purposes of assessment must be the student's own work. Penalties for violation of the policy range from zero marks for the component assessment to expulsion, depending on the nature of the offense.

When in doubt, students should consult the instructors of the course. Details on the SMU Code of Academic Integrity may be accessed at https://smu.sharepoint.com/sites/oasis/SitePages/DOS-WKLSWC/UCSC.aspx.

Copyright Notice

Please note that all course materials are meant for personal use only, namely, for the purposes of teaching, studying and research. You are strictly not permitted to make copies of or print additional copies or distribute such copies of the course materials or any parts thereof, for commercial gain or exchange.

For the full copyright notice, please visit: https://smu.sg/Copyright-notice or *OASIS -> CAMPUS LIFE & EXCHANGE -> CONDUCT & DISCIPLINE -> UNIVERSITY COUNCIL OF STUDENT DISCIPLINE*

Accessibility

SMU strives to make learning experiences accessible for all. If you anticipate or experience physical or academic barriers due to disability, please let me know immediately. You are also welcome to contact the university's student accessibility support team if you have questions or concerns about academic provisions: accessibility@smu.edu.sg. Please be aware that the accessible tables in our seminar room should remain available for students who require them.

Digital Readiness for Teaching and Learning (DRTL)

As part of emergency preparedness, instructors may conduct lessons online via the Zoom platform during the term, to prepare students for online learning. During an actual emergency, students will be notified to access the Zoom platform for their online lessons. The class schedule will mirror the current face-to-face class timetable unless otherwise stated.