Institute of Business Administration

CSE 452: Big Data Analytics

(Course Outline and Syllabus)

IBA SMCS

School of Mathematics and Computer Science

Fall 2025

Institute of Business Administration
School of Mathematics and Computer Science

CSE 452: Big Data Analytics

“Without big data, you are blind and deaf and in the middle of a freeway.” — Geoffrey Moore

“Big data isn’t about bits, it’s about talent.” — Douglas Merrill

“Where there is data smoke, there is business fire.” — Thomas Redman

“Big data is at the foundation of all of the megatrends that are happening today, from social to mobile to the cloud to gaming.” — Chris Lynch

Artificial intelligence is useless without data. And data is useless without artificial intelligence – Andrew Ng

Machine learning and AI are fueled by data. The better and bigger the data, the more powerful the AI becomes – Fei-Fei Li (Stanford Professor)

Course Logistics

Instructor: Dr. Tariq Mahmood

Course schedule: Tuesday, Thursday MTC25 1130am – 1245pm
Credits: 3 (3 + 0)

Pre-requisites: CS341 Database Systems

Course Description:

This course targets a 360-degree learning related to big data and its analysis. The primary motivation for doing big data analysis is big data itself. So, in the presence of big data, the data storage and analysis methods change. These methods are the core scope of this course, with a focus on NoSQL storage and querying, and big data analysis through interactive methods. The course goes one step further in using the well-known Docker platform for containerization and orchestration. Moreover, data engineering aspects are also included pertaining to building data pipelines for big data.

Course Objectives:

- Acquire understanding of big data and its differences from small data with respect to ingestion, storage, management and analysis
- Acquire understanding and skillset of big data storage technologies under the NoSQL umbrella

- Acquire knowledge of cloud-based big data storage (AWS, Azure and Google Cloud)
- Acquire theoretical and practical knowledge of Docker platform (containerization in general)
- Acquire a theoretical and practical perspective of data engineering with respect to building data engineering pipelines
- Get a strong understanding of the relationship between the AI revolution and big data analysis

Program Learning Outcomes:

- PLO-2: Knowledge for Solving Computing Problems
- PLO-3: Problem Analysis
- PLO-4: Design/Development of Solutions
- PLO-5: Modern Tool Usage

Course Learning Outcomes:

- CLO1: Knowledge and skills for big data storage based on NoSQL
- CLO2: Knowledge and skills for containerization-based BDA
- CLO3: Knowledge of big data management, software architectures, governance and security
- CLO4: Knowledge and skills related to data engineering

PLO to CLO Mapping:

	PLO-2	PLO-3	PLO-4	PLO-5
CLO-1	✓	✓		✓
CLO-2	✓	✓		✓
CLO-3			✓	
CLO-4	✓	✓	✓	✓

Format and Procedures:

The LMS site will be used to share the syllabus, give out assignments, and to share other course resources. The University's standard policies on attendance, inclusivity, office hours, and academic integrity apply in this course. These are described in later sections below.

Course Textbooks:

- Big Data: Principles and Best Practices of Scalable Realtime Data Systems
- Mining Massive Datasets, Rajaraman, Leskovec, Ullman, Stanford University

(http://www.mmds.org/)

- The Guide to Big Data Analytics, Data Meer

Reference Textbooks:

- Hadoop: The Definitive Guide (OReilly) + Reference Manual (Online)
- Apache Hive Essentials + Hive-Hadoop: The Definitive Guide (https://cwiki.apache.org/confluence/display/Hive/Books+about+Hive)
- MongoDB: The Definitive Guide (OReilly) + Reference Manual (Online)
- Redis in Action + Reference Manual (Online)
- Cassandra: The Definitive Guide (OReilly) + Reference Manual (Online)
- Building Knowledge Graphs: A Practitioner's Guide
- Designing Data-Intensive Applications by Martin Kleppmann
- Fundamentals of Data Engineering by Joe Reis
- Big Data Fundamentals: Concepts, Drivers & Techniques

Required Tools:

List of Tools:

Data Wrangling: Python

Docker, Apache Hadoop, Apache Hive, Redis, Apache Casandra, Neo4j, MongoDB, Airflow, Luigi
BI Tools: PowerBI, Tableau, Custom

Setup Links:

Docker Desktop: https://docs.docker.com/desktop/install/windows-install/
VM's on HPC will be made available.

Grading Procedures: (tentative)

- Quiz ever alternate week (10%): Total 5 quizzes worth 2% each
- Mid-Term: 20%
- Final Exam: 30%
- Assignments: 15% (5 assignments of 3% each)
- Project: 25% (Video and Analytical Report)
- *Team Formation:* You may form teams of 2 members for all your assignments, class activities and projects. You must decide your team at the start of the semester, and it shall continue to the end. Absence of any member during any activity will result in loss of marks for that member.

Attendance Policy

IBA attendance policy applies.

Academic Integrity

Each student in this course is expected to abide by the IBA Code of Conduct. Scholastic dishonesty shall be considered a serious violation of these rules and regulations and is subject to strict disciplinary action as prescribed by IBA regulations and policies. Scholastic dishonesty includes, but is not limited to, cheating on exams, plagiarism on assignments, and collusion. Kindly refer to https://examination.iba.edu.pk/CheatingPlagiarism.php for more details.

- PLAGIARISM: Plagiarism is the act of taking the work created by another person or entity and presenting it as ones own for the purpose of personal gain or of obtaining academic credit. Plagiarism includes the submission of or incorporation of the work of others without acknowledging its provenance or giving due credit according to established academic practices. This includes the submission of material that has been appropriated, bought, received as a gift, downloaded, or obtained by any other means. Students must not, unless they have been granted permission from all faculty members concerned, submit the same assignment or project for academic credit for different courses.
- CHEATING: The term cheating shall refer to the use of or obtaining of unauthorized information in order to obtain personal benefit or academic credit.
- COLLUSION: Collusion is the act of providing unauthorized assistance to one or more person or of not taking the appropriate precautions against doing so. Any student violating academic integrity a second time in this course will receive a failing grade for the course, and additional disciplinary sanctions may be administered.
- SHARING CREDENTIALS: It has been observed that some students share their credentials (log in id's and passwords) of LMS, portal, email, etc., with other students. These credentials are private and confidential and not to be shared with anyone. Any violation will be considered as aiding in plagiarism/collusion/cheating and appropriate action might be taken against such students.

Office hours

Monday and Wednesday 430 pm to 515 pm
If you need to speak to the instructor besides the designated office hours, you may book an appointment via email.

Late Submission Policy:

All assignments and graded class activities must be timely submitted via LMS. For any assignment, late submission up to one day late will be accepted with a 10% late penalty of the maximum score. Beyond that, no late submissions will be acceptable.

Missed assessments policy:

There will be no makeup for any missed assessments including assignments, project, quizzes and exams. In case of any medical emergency, proof must be submitted for any consideration.

Weekly breakdown of classes (tentative)

Week 1: Introduction to Big Data, History, Evolution, Concepts, Definitions, Techniques, Big Data Landscapes, BDA

Week 2-3: Fundamentals of Distributed Computing and Parallel Processing, Containerization and orchestration, Virtualization, Hypervisor, Hypervisor types, Docker architecture and theory in detail, Hands-on with Docker, Experimental testbed for BDA experiments, Hands-on with Linux commands, Hands on with Linux Shell Scripting

Week 4-5: Document databases: MongoDB, History, Need, Core theoretical aspects in detail (architecture, querying language, design, distributed databases), Practical on Docker, Applications, Usage, Limitations

Week 6-9: Apache Hadoop, HDFS, MapReduce and Apache Spark, Apache Kafka, Apache Hive, History, Need, Theoretical aspects in detail, Practical on Docker, Applications, Usage, Limitations

Week 10: Columnar Warehouses. HBase History, Need, Core theoretical aspects in detail (architecture, querying language, design, distributed databases), Practical on Docker, Applications, Usage, Limitations

Week 11: Key-value databases: Redis, History, Need, Core theoretical aspects in detail (architecture, querying language, design, distributed databases), Practical on Docker, Applications, Usage, Limitations

Week 12: Big Data software architectures, AI-based big data analysis techniques, tools and trends

Week 13-14: Data Engineering: definition, history, theory, current practices, building a data pipeline (ingestion, storage, analysis etc.), best practices and rules for making a data pipeline, relationship to DevOps, building an end-to-end CI/CD data pipeline using Apache Airflow and related tools

Week 15: Big Data and AI