Introduction To Big Data Analytics INSY 8413

Instructor:

- Eric Maniraguha | eric.maniraguha@auca.ac.rw | LinkedIn Profile

6h00 pm – 8h50 pm

- Monday A-G104
- Tuesday E-G108
- Wednesday A-G104
- Thursday E-G108

June 2025

Course Introduction: Introduction to Big Data Analytics

In today's world, data is everywhere—and growing fast. Traditional tools can't keep up with the **scale, speed, and complexity** of modern data. That's where **Big Data Analytics** comes in.

This course gives you the **essential knowledge and hands-on skills** to work with large datasets, understand key concepts, and use tools that professionals rely on. It's designed for beginners, so no prior experience with data or coding is required.

What You'll Learn:
- **What Big Data is** and why it matters (the 5 V's: Volume, Velocity, Variety, Veracity, Value)
- **Basics of Python programming**—no coding background needed
- How to **clean, analyze, and visualize large datasets**
- An overview of **popular Big Data tools** like Hadoop and Spark
- How to use **data to uncover insights and support decisions**

Tools We'll Use:
- **Google Colab**: Run Python code in the cloud—easy, fast, no installation needed.
- **Kaggle**: Practice with real datasets and compete in fun challenges.
- **GitHub:** Collaborate, manage code, and track your progress like a real data scientist

Why Take This Course?
By the end of the course, you'll be able to:
- Write and run basic Python code for data analysis
- Work with real-world datasets using industry tools
- Share and manage your work using GitHub
- Build a strong foundation for advanced data science and analytics courses

Whether you're interested in tech, business, health, or research—**Big Data skills will set you apart**.

Course Outline

Lecture 1: Introduction to Big Data
- What is Big Data?
- The 5 V's: Volume, Velocity, Variety, Veracity, Value
- The importance of Big Data in today's world
- Real-world applications across sectors

Module 2: Big Data Ecosystem Overview
- Hadoop ecosystem (HDFS, MapReduce, YARN)
- Apache Spark fundamentals
- NoSQL databases (MongoDB, Cassandra, HBase)
- Cloud platforms (AWS, Azure, Google Cloud)

Lecture 3: Introduction to Git and GitHub
- Version control concepts
- Creating and managing repositories
- Pushing notebooks and scripts to GitHub

Lecture 4: Introduction to Machine Learning
- What is Machine Learning?
- Types: Supervised vs. Unsupervised learning
- Examples of classification and regression
- Basics of scikit-learn and ML pipeline
- Hands-on: Simple ML demo (e.g., Linear Regression or K-Means)

Lecture 5: Basics of Python for Data Analytics
- Installing Python and Jupyter Notebook - Anaconda
- Introduction to Python syntax and structure
- Variables, data types, conditionals, loops
- Basic input/output and functions

Lecture 6: Working with Data in Python
- Lists, dictionaries, and data structures
- Introduction to Pandas and NumPy
- Reading data from CSV and Excel
- Data cleaning basics
- Using Google Colab & Kaggle Platform for Python coding
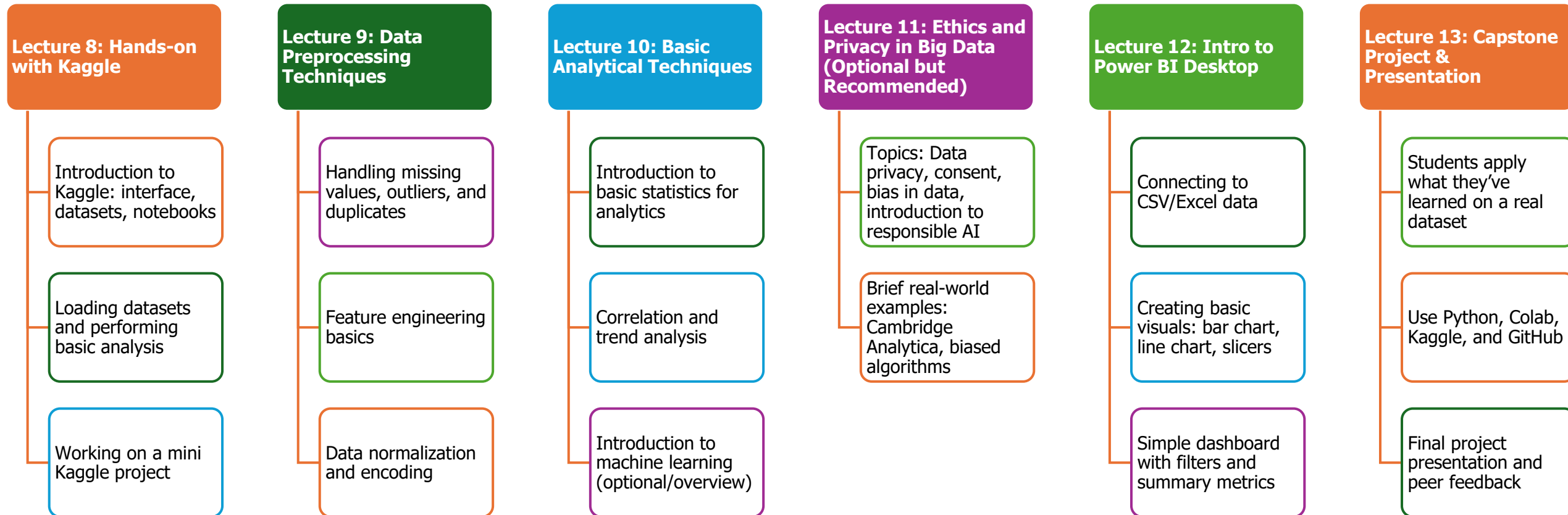
Lecture 7: Data Exploration and Visualization
- Summary statistics and data profiling
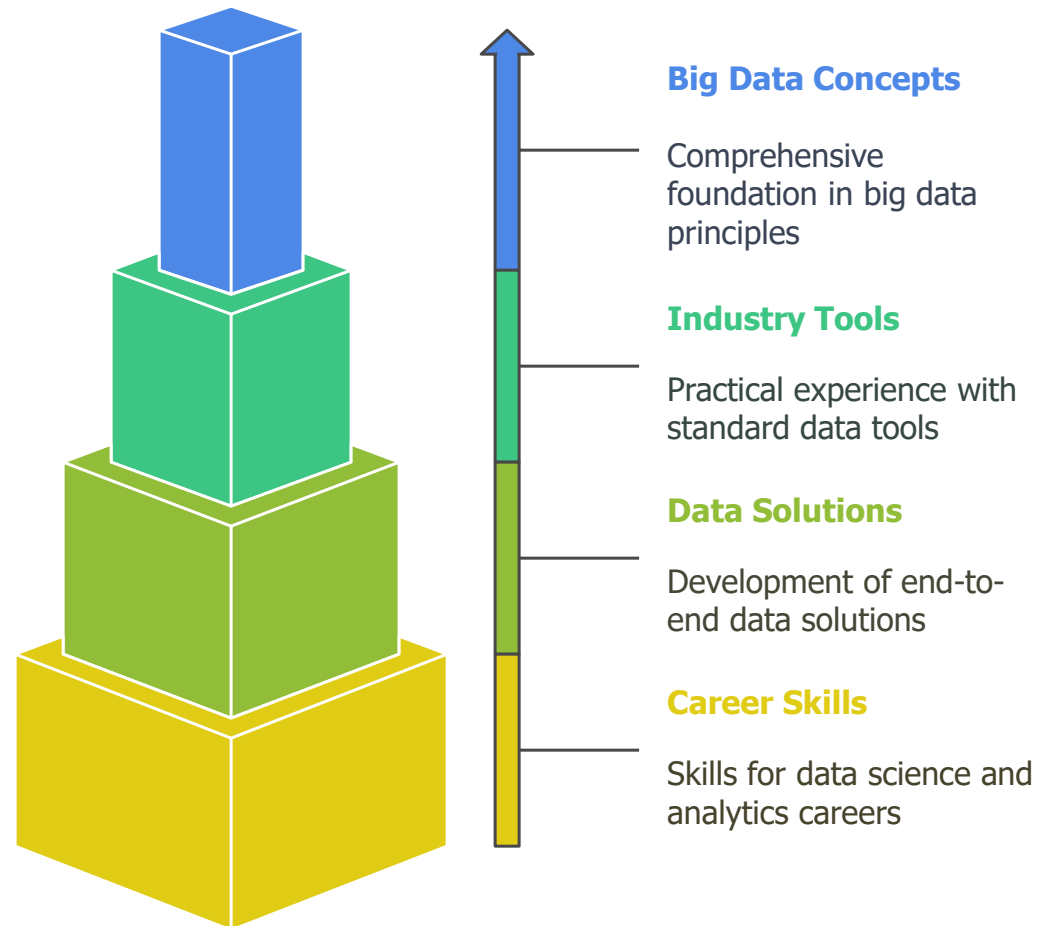- Basic EDA (exploratory data analysis) using pandas
- Visualizing data and interpretation
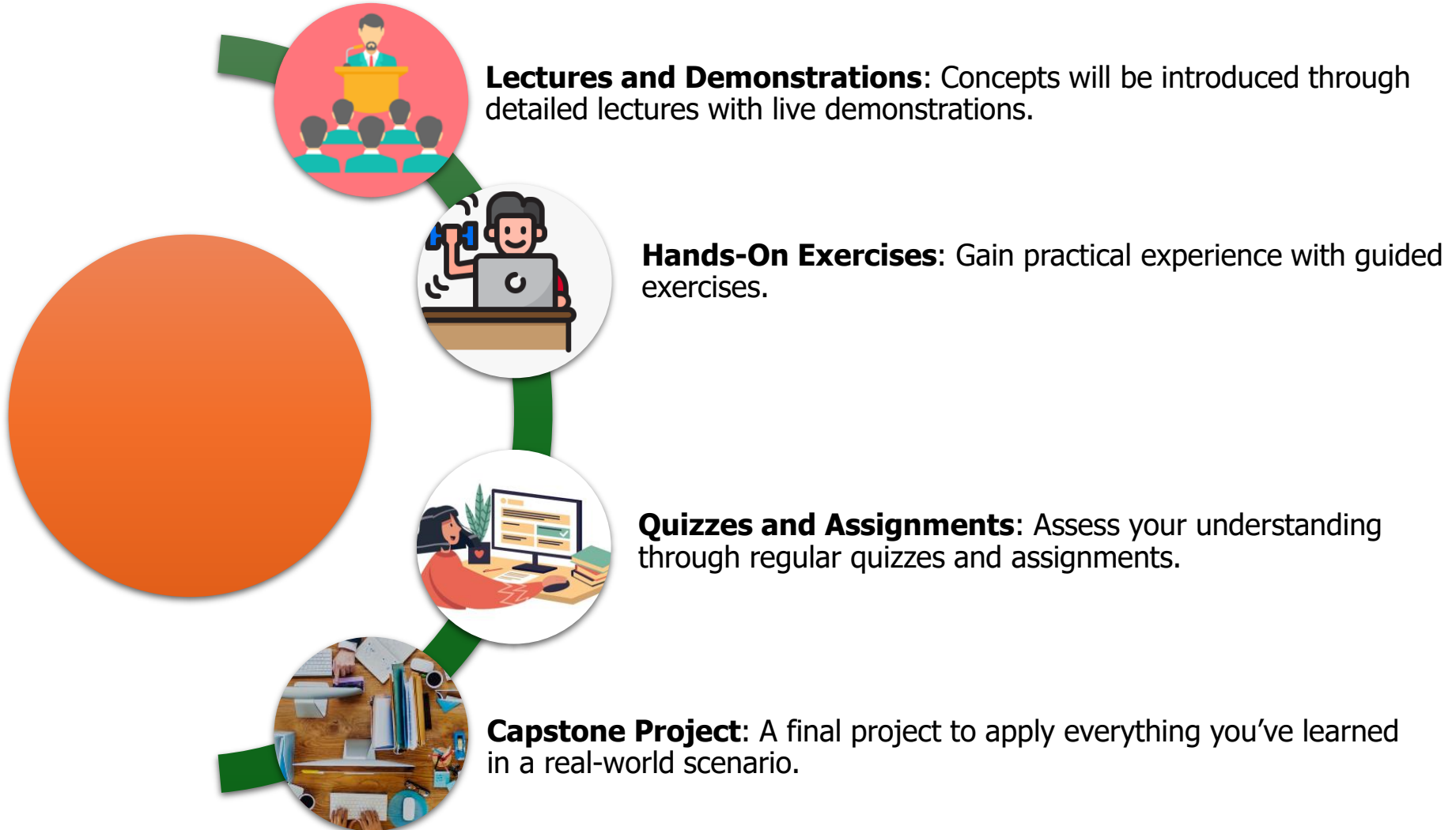- Identifying patterns and trends

Course Outline

Lecture 8: Hands-on with Kaggle

- Introduction to Kaggle: interface, datasets, notebooks
- Loading datasets and performing basic analysis
- Working on a mini Kaggle project

Lecture 9: Data Preprocessing Techniques

- Handling missing values, outliers, and duplicates
- Feature engineering basics
- Data normalization and encoding

Lecture 10: Basic Analytical Techniques

- Introduction to basic statistics for analytics
- Correlation and trend analysis
- Introduction to machine learning (optional/overview)

Lecture 11: Ethics and Privacy in Big Data (Optional but Recommended)

- Topics: Data privacy, consent, bias in data, introduction to responsible AI
- Brief real-world examples: Cambridge Analytica, biased algorithms

Lecture 12: Intro to Power BI Desktop

- Connecting to CSV/Excel data
- Creating basic visuals: bar chart, line chart, slicers
- Simple dashboard with filters and summary metrics

Lecture 13: Capstone Project & Presentation

- Students apply what they've learned on a real dataset
- Use Python, Colab, Kaggle, and GitHub
- Final project presentation and peer feedback

Big Data Mastery Pyramid

Big Data Concepts

Comprehensive foundation in big data principles

Industry Tools

Practical experience with standard data tools

Data Solutions

Development of end-to-end data solutions

Career Skills

Skills for data science and analytics careers

Teaching Methodology

SQL*Plus

Lectures and Demonstrations: Concepts will be introduced through detailed lectures with live demonstrations.

Hands-On Exercises: Gain practical experience with guided exercises.

Quizzes and Assignments: Assess your understanding through regular quizzes and assignments.

Capstone Project: A final project to apply everything you've learned in a real-world scenario.

Who Should Attend?

This course is designed for **undergraduate students** from any discipline who are curious about data and its power to solve real-world problems. It is ideal for:

- **Beginners with no prior coding experience** who want to build foundational skills in Python and data analytics.
- **Students in IT, business, health, economics, or engineering** looking to understand how data is transforming their fields.
- **Aspiring data scientists or analysts** who want to explore Big Data tools and platforms like Kaggle, Google Colab, and GitHub.
- **Anyone interested in working with real-world datasets** and learning how to generate insights using modern tools.

Whether you're aiming for a career in **tech, research, entrepreneurship, or public service**, this course will give you valuable skills to work with data confidently.

Prerequisites

No prior experience in programming or data analytics is required, although having a basic understanding of fundamental concepts can be helpful.

This course is designed for beginners, and all foundational topics will be introduced from scratch. However, the following will be beneficial:

- Basic computer literacy (e.g., using web browsers, saving files)
- Curiosity and a willingness to learn new tools and technologies
- Comfort with basic mathematics and logical thinking

All coding will be done using **Google Colab or Jupyter Notebook or Anaconda**, so no software installation is necessary. Step-by-step guidance will also be provided for using Kaggle and GitHub.

Define Big Data and its key characteristics (Volume, Velocity, Variety, Veracity, and Value) and explain its impact across different industries.

Clean and preprocess large datasets, including handling missing values, outliers, and applying transformations.

Write and execute basic Python code for data manipulation, analysis, and visualization using tools like Pandas, NumPy, Matplotlib, and Seaborn.

Key Learning Outcomes (By the end of this course, you will be able to):

Access and analyze real-world datasets through the Kaggle platform, and participate in hands-on data challenges.

Use Google Colab to write and run Python scripts in the cloud without requiring any local installation.

Visualize data insights effectively to support interpretation and decision-making.

Build a solid foundation for more advanced courses in data science, analytics, or machine learning.

Understand the purpose and architecture of Big Data technologies like Hadoop and Spark at a conceptual level.

Key Learning Outcomes (By the end of this course, you will be able to):

Conduct a complete data analysis project, from data collection to presenting insights using modern platforms and tools.

Apply version control using Git and GitHub to manage and share coding projects in a collaborative environment.

9

Unlocking Career Paths with Big Data: Roles, Skills, and Future Prospects

As data continues to shape decisions across industries, Big Data Analytics has become one of the most in-demand skillsets globally. This course sets the stage for students to explore a wide range of data-driven career paths. By learning to work with real datasets, cloud platforms, and essential tools, you gain valuable hands-on experience that can lead to impactful roles.

In-Demand Career Roles:
- **Data Analyst** – Collect, clean, and analyze data to guide business decisions.
- **Business Intelligence (BI) Analyst** – Design dashboards and reports to deliver insights.
- **Data Scientist (Entry-level)** – Apply programming, statistics, and modeling to extract insights.
- **Data Engineer (Foundation level)** – Prepare and manage data pipelines for analytics.
- **AI/ML Assistant** – Support machine learning projects with data preparation and analysis.
- **Digital Research Assistant** – Analyze data in academic or applied research contexts.

Key Skills You'll Build:
- Programming with **Python**
- Using platforms like **Kaggle**, **Google Colab**, and **GitHub**
- Data wrangling and **preprocessing**
- **Data visualization** and storytelling
- Understanding **Big Data tools** like Hadoop and Spark (conceptually)
- Collaborating using **version control systems**

Future Prospects:
- Progress into advanced courses in **Data Science**, **Machine Learning**, or **AI**
- Gain a competitive edge in internships, research, or graduate study
- Enter job markets across industries like healthcare, finance, tech, logistics, public policy, and more

Big Data is not just a technical skill—it's a new way of seeing the world. Whether you're interested in solving real-world problems, building smarter businesses, or contributing to data-driven research, this course is your starting point.

Further Learning

Take your next step toward becoming a data-driven professional.
Completing *Introduction to Big Data Analytics* gives you more than just foundational knowledge—it sets you on a clear path to deeper expertise and industry-recognized credentials. Whether you want to specialize in analytics, enter the job market, or pursue research, the learning doesn't stop here.

Recommended Next Steps:
Advanced Data Analytics & Machine Learning
- Deepen your understanding of data modeling, clustering, and prediction using Python.
- Start building ML models with Scikit-learn, TensorFlow, or PyTorch.

Big Data Ecosystems & Technologies
- Learn to manage and process massive datasets using **Apache Spark**, **Hadoop**, **Kafka**, and **Hive**.
- Explore distributed computing and real-time data streaming.

Data Engineering Foundations
- Build skills in data pipeline creation, ETL processes, cloud storage, and APIs.
- Learn tools like **Airflow**, **SQL**, and **Google BigQuery**.

Business Intelligence & Visualization
- Turn complex data into clear insights using **Tableau**, **Power BI**, or **Looker**.
- Master dashboard design and interactive reporting.

Cloud-Based Data Solutions
- Explore platforms like **Google Cloud**, **AWS**, and **Microsoft Azure** for scalable data services.
- Learn about cloud-based ML, data lakes, and data warehousing.

Top Certifications to Boost Your Career

These globally recognized certifications can help validate your skills and improve your job readiness:

- **Google Data Analytics Professional Certificate** *(Coursera)*
- **IBM Data Analyst Professional Certificate** *(Coursera)*
- **Microsoft Certified: Data Analyst Associate** *(Power BI)*
- **AWS Certified Data Analytics – Specialty**
- **Google Cloud Certified – Professional Data Engineer**
- **Cloudera Certified Associate (CCA): Data Analyst / Spark Developer**

💡 *Many of these certifications offer beginner-friendly paths and include hands-on labs using tools you'll already know from this course (like Python, Colab, and GitHub).*

Why Continue Learning?

- **Stay competitive** in a fast-changing job market
- **Specialize** in a domain that matches your career goals
- **Qualify for internships, research roles, or tech jobs** with data skills that are in high demand
- **Strengthen your portfolio** and online presence with real project work and certifications

Academic Calendar – Semester III (2024/2025)

Month	Date	Major Event
June 2025	8th – 13th	Orientation and Registration
	15th	Beginning of Summer Semester
July 2025	1st	**Public Holiday** – Independence Day
	4th	**Public Holiday** – Liberation Day
	6th – 11th	Mid-Semester Exams
August 2025	1st	End of Summer Semester Classes
	3rd – 8th	Final Exams
	15th	End of Summer Semester
	17th – 21st	AUCA Staff Weekend Retreat

Grading and Assessment Breakdown – Introduction to Big Data Analytics

Assessment Component	Weight (%)	Details
Quiz	10%	Short tests to assess understanding of early course concepts.
Class Work (CW)	10%	In-class participation, exercises, and activities to enhance engagement.
Assignments (Ass)	10%	Practical take-home tasks applying learned concepts.
Mid-Semester Exam	30%	Formal evaluation to assess progress at mid-semester.
Continuous Assessment Tests (CATS)	60%	Aggregate of quizzes, assignments, and mid-semester exam performance.
Final Exam (Project Capstone)	40%	Comprehensive final project or exam to evaluate overall course mastery.
Total Marks	100%	Combined total of all assessment components.
Final Score (SCORE)	20 Points	Final grade converted and recorded out of 20 points.

Mastering Big Data Analytics Design and Development: Final Capstone Project

Final Exam – 40%

▪ **Format**:
- o Capstone project (25%)
- o Theoretical exam (15%)

▪ **Project Scope**:
- o Covers all key tasks learned throughout the course
- o Includes GitHub code repository and PowerPoint insights presentation

▪ **Implementation Phases**:
- o Executed step-by-step in structured phases
- o Allows for continuous progress tracking and evaluation at each stage

Why This Approach

This project-based and phased approach is used because it:

1. **Reflects Industry Practice**: Real-world big data projects are executed in **phases**, from problem definition to deployment and presentation. This format mimics that.
2. **Builds Higher-Order Thinking**: Instead of just testing memory or isolated skills, it **integrates critical thinking, problem-solving, and communication**.
3. **Encourages Ownership**: Students take responsibility for the full data analytics cycle—from idea to delivery.
4. **Promotes Agile Thinking**: Breaking the project into phases encourages **iterative development**, feedback incorporation, and incremental learning.

PRACTICUM

Mastering Big Data Analytics Design and Development: Final Capstone Project

Key Deliverables Explained

1. Problem Statement Design
- A strong, well-scoped problem defines project direction.
- Evaluates ability to translate **real-world problems into analytical questions**.
- You must clarify: Who is affected? Why is it important? What data is needed?

2. Solution Development
- Demonstrates your ability to **design an appropriate pipeline** (ETL, data wrangling, modeling, etc.).
- Incorporates tools such as Spark, Hadoop, NoSQL, or ML frameworks.
- Emphasizes **model testing, optimization, and interpretation**.

3. Code Submission on GitHub
- GitHub use ensures **version control, collaboration readiness, and professional practice**.
- Code organization and commit messages show your **development discipline**.
- Makes the project **transparent, reproducible, and presentable** to future employers or evaluators.

4. Documentation Report
- A comprehensive written report validates your ability to:
  - Explain technical workflows.
  - **Reflect on decision-making**, like why a particular model or tool was used.
  - Document real project constraints (e.g., missing data, performance trade-offs).
  - Share lessons learned and justify your choices.

5. PPT Presentation
- Trains you to **communicate insights to non-technical stakeholders**.
- Focuses on data storytelling with:
  - Clear visuals (charts, dashboards).
  - Code snippets to highlight important logic.
  - Impact assessment and possible future extensions.
- Tests your ability to **summarize a complex project concisely and persuasively**.

PRACTICUM

Overall Assessment Goals

Skill Area	What's Evaluated
Technical Proficiency	Coding, modeling, data handling
Analytical Thinking	Framing questions, interpreting results
Tool Mastery	Use of platforms like Kaggle, Google Colab, etc.
Communication	Reports, GitHub documentation, presentations
Professionalism	Project management, version control, clarity

PRACTICUM

Course Policies: Attendance, Integrity, and Respect

Guideline	Details
1. Attendance Required	- Active participation is essential for success.- A **minimum of 75% attendance** is required to be eligible for the final exam.- Absences must be minimized and justified with valid reasons.
2. Academic Integrity	- Submit **original work only**—plagiarism, cheating, or dishonest conduct will not be tolerated.- Violations are subject to **disciplinary action**, including potential course failure per university policy.
3. Respect Assignment Deadlines	- Timely submission is crucial for academic progress.**Late Submission Policy:** ▪ Within 2 weeks: Graded at **50%** of the original score. ▪ After 2 weeks: Not accepted unless pre-approved with a valid reason.

The Power of Faithfulness: Biblical Verses and Reflections

The Power of Faithfulness: Biblical Verses and Reflections
Faithfulness is a powerful virtue that strengthens our character, builds trust, and deepens our relationship with God and others. Below are some key biblical verses on faithfulness, along with reflections to inspire and encourage you.

2. Proverbs 3:3-4 (NIV)
"Let love and faithfulness never leave you; bind them around your neck, write them on the tablet of your heart. Then you will win favor and a good name in the sight of God and man."
Reflection:
Faithfulness is a mark of integrity and loyalty that brings honor both from God and people. It's a lifelong commitment that shapes our reputation.

3. 1 Corinthians 4:2 (NIV)
"Now it is required that those who have been given a trust must prove faithful."
Reflection:
Faithfulness means being trustworthy and dependable in all God has entrusted to us — our talents, relationships, and responsibilities.

4. Psalm 33:4 (NIV)
"For the word of the Lord is right and true; he is faithful in all he does."
Reflection:
God's actions and promises are always reliable. We can trust Him completely because His faithfulness never fails.

5. Galatians 5:22-23 (NIV)
"But the fruit of the Spirit is love, joy, peace, patience, kindness, goodness, faithfulness, gentleness and self-control."
Reflection:
Faithfulness is a fruit of the Holy Spirit. When we live guided by the Spirit, faithfulness naturally grows within us as part of our character.

Closing Thought:
Faithfulness is a journey and a choice we make daily. By embracing God's faithfulness and reflecting it in our lives, we build lasting relationships, deepen our spiritual walk, and fulfill our purpose with integrity and love.

Thank you!

Stay Connected!