Course Introduction

Big Data Analytics

Professor Anis Koubaa

SE 446
Alfaisal University
https://github.com/aniskoubaa/big_data_course

Spring 2026

Outline

1. Welcome & Course Overview
2. Course Logistics
3. Tools & Platforms
4. Semester Project
5. GitHub Workflow
6. Summary & Next Steps

Welcome to SE 446!

Course Info

- **Credits**: 3 hours
- **Duration**: 13 weeks
- **Schedule**: 2 sessions/week

Instructor

- Prof. Anis Koubaa
- `akoubaa@alfaisal.edu`
- Office: SG-10

What is Big Data?

Definition

Data that is too large, fast, or complex for traditional tools to process.

Scale Examples:

- Facebook: 4 PB/day
- YouTube: 500 hrs video/min
- Twitter: 500M tweets/day
- Google: 20 PB processed/day

Why It Matters:

- Better business decisions
- Scientific discoveries
- Real-time insights
- Competitive advantage

Course Learning Outcomes (CLOs)

By the end of this course, you will be able to:

1. **Knowledge & Understanding**
   - Explain essential concepts, challenges, and approaches in Big Data
2. **Skills**
   - Implement scalable data processing pipelines for batch, streaming, and distributed real-time workflows on distributed platforms
3. **Perform Data Analysis**
   - Perform data analysis on large datasets and interpret results to support evidence-based decision making in real-world contexts
4. **Values, Autonomy, & Responsibility**
   - Demonstrate ethical, responsible, and collaborative practices when working with data, including respect for privacy, security, and teamwork principles

Weekly Schedule

Week	Topic	Milestone	Assessment
1	Course Introduction	–	–
2	Big Data + HDFS	–	–
3-4	Data Formats + MapReduce	M1	–
5-6	Hive + M2	M2	Midterm 1
7-8	Apache Spark	M3	–
9-10	Kafka + Streaming + M4	M4	Midterm 2
11-12	Project Completion	M5	Quiz 1, 2
13	Final Review	–	–

Grading Breakdown

Component	**Weight**
Midterm 1	20%
Midterm 2	20%
Final Exam	30%
Quizzes (2)	10%
Project Work	20%
Total	**100%**

Project Work (20%)

5 Milestones (4% each)

What counts:

- Github commits
- Regular submissions
- Milestone Quality

Attendance Policy

Important

Attendance is **mandatory** and missing classes would affect your grades.

- Each class has an in-class Moodle quiz (last 15 min)
- Missing an in-class submission will affect your grade
- Medical/official excuses within 48 hours

Tools We'll Use

Tool	Purpose	Weeks
Google Colab	Python, Pandas, PySpark basics	1-4
Databricks	Spark, Hive, Streaming	5-10
VS Code	Local development (optional)	All
GitHub	Code collaboration	All
Moodle	In-class quizzes	All

No Installation Required!

Everything runs in the **cloud**. You only need a web browser.

Google Colab & Databricks

Google Colab

- Free Jupyter notebooks in the cloud
- Python + libraries pre-installed
- Easy sharing via Google Drive
- GPU access when needed

URL: `colab.google.com`

Databricks Community

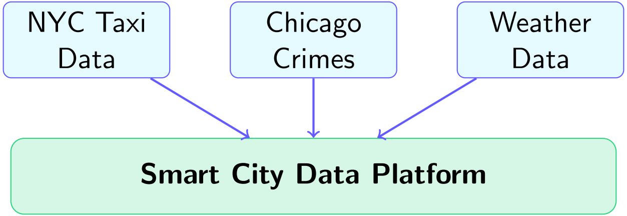
- Free cloud Big Data platform
- Apache Spark pre-configured
- Industry-standard tool
- Notebooks + cluster management

URL: `databricks.com/try`

Project: Smart City Data Platform

Concept

Build a data analytics platform using real urban datasets.



5 Milestones

M	Topic	Week Due	Weight
M1	Data Loading (HDFS concepts)	4	4%
M2	MapReduce Processing	6	4%
M3	Hive Analytics	8	4%
M4	Spark Analysis	10	4%
M5	Streaming Pipeline	12	4%
	Total		**20%**

Per Milestone

GitHub Commits and **Related Assessment** will be counted towards project grade

Datasets We'll Use

Dataset	Size	Description
NYC Yellow Taxi	∼50 MB	Trip records, fares, locations
Chicago Crimes	∼30 MB	Crime types, dates, locations
NYC Weather	∼5 MB	Daily temperature, precipitation
Air Quality Index	∼3 MB	Daily AQI by city

Good News!

All datasets are pre-hosted. No downloading required.

Course Repository

GitHub Repository

https://github.com/aniskoubaa/big_data_course

- All course materials (slides, notebooks, data)
- Weekly updates
- Milestone templates
- Clone it to get started!

Clone Command

```
git clone https://github.com/aniskoubaa/big_data_course.git
```

Team Repository Structure

Repository Organization:

Each team gets **ONE shared repository**

```
se446-team-01/
  milestone_1/
    student_ahmed/
    student_fatima/
  milestone_2/
    student_ahmed/
    student_fatima/
  ...
```

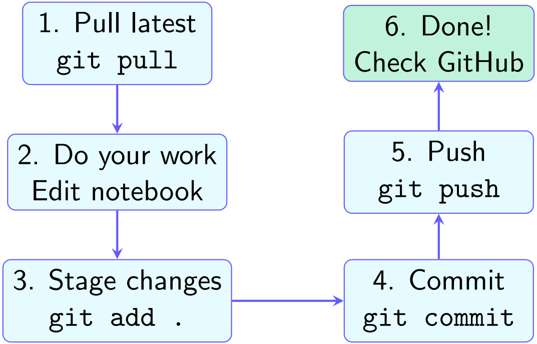
Important Rules:

- Each student works in their **own folder**
- Individual commits are tracked separately
- Work on your assigned tasks only
- Quality matters more than quantity

Note

Your individual contributions will be evaluated based on your folder's commits

Git Workflow (Simplified)



1. Pull latest
`git pull`

2. Do your work
Edit notebook

3. Stage changes
`git add .`

4. Commit
`git commit`

5. Push
`git push`

6. Done!
Check GitHub

Commit Message Standards

Format

```
<MILESTONE>:  <Short description>
```

Good Examples:

- `M1:  Loaded NYC taxi data and checked schema`
- `M2:  Implemented mapper for crime type count`
- `M3:  Added HiveQL query for average fare`

Bad Examples:

- `update` ← Too vague
- `asdfasdf` ← Meaningless

Summary

1. **Course**: Learn Big Data processing with Hadoop, Spark, Kafka
2. **Grading**: Exams (70%) + Quizzes (10%) + Project (20%)
3. **Tools**: Colab, Databricks, VS Code, GitHub, Moodle
4. **Project**: 5 milestones with real urban datasets
5. **GitHub**: Your commits are tracked and analyzed

Course Repository

`https://github.com/aniskoubaa/big_data_course`

Action Items for This Week

1. **Create accounts** (if you don't have):
   - GitHub: `github.com`
   - Google (for Colab): `google.com`
2. **Clone the course repository**
3. **Watch pre-class video** for Week 2:
   - "What is Big Data?" - Simplilearn (∼15 min)

Next Week Preview

Week 2: Introduction to Big Data & HDFS

- The 5 V's of Big Data
- HDFS Architecture
- File Formats (CSV, JSON, Parquet)
- First hands-on notebook!

Get ready to dive into Big Data!

Questions?

Let's set up your accounts!

Prof. Anis Koubaa
akoubaa@alfaisal.edu

https://github.com/aniskoubaa/big_data_course