# Data 603 – Big Data Platforms

UMBC

Lecture 1

# Instructor Contact Info

- Contact info:
  - enkeboll@umbc.edu

# Course Goal

- To introduce methods, technologies, and computing platforms for performing data analysis **at scale**
- Prepare the class for the Apache Spark developer certification from Databricks (https://academy.databricks.com/exam/databricks-certified-associate-developer)

# Topics Covered

- Topics include:
  - Theory and techniques for data acquisition, cleansing, aggregation, management of large heterogeneous data collections, processing, information and knowledge extraction.
  - Practical hands-on experience using Apache Spark on Databricks (cloud) platform.

# Techniques Covered

- Class will introduce:
  - MapReduce, streaming, and external memory algorithms and their implementations using the Apache Spark ecosystem.

# Course Prerequisites

- Students will gain practical experience in analyzing large existing databases.
- **Prerequisite:**
  - Enrollment in the Data Science program
  - Programming experience
  - Other students may be admitted with instructor permission
- **Corequisite:**
  - DATA 601: Introduction to Data Science

# Class Introduction

- Introduce yourself:
  - Name
  - UMBC affiliation
  - Big Data Experience
  - Goals of the class

# Class Logistics

- Every Thursday during Spring 2022
- Start at 7:00 PM and end at 9:40 PM
- Every class we will try to maintain the following agenda:
  - 20 Minutes – Homework Quick Review / Questions
  - 60 Minutes – Lecture Presentation, labs
  - 15 Minutes – Break
  - 45 Minutes – Lecture Presentation, labs
  - 10 Minutes – Review / Prepare for next lecture / Quiz

# Course Topics & Syllabus

- Introduction & Foundation of Data Sciences
- Platforms overview
- Tools in Data Science
  - Hadoop & HDFS
  - Apache Spark (RDD, DataFrame, Dataset, SQL)
  - Apache Hive
  - Distributed DB (Hbase/Accumulo, Cassandra)
  - Machine Learning
  - Cloud Computing

# Course Grading

| Course work | Grade distribution |
|---|---|
| **Attendance/class participation/presentations** | 10% |
| **Homework & Assignments** | 25% |
| **Quizzes** | 10% |
| **Technical Research Paper** | 20% |
| **Final Project** | 35% |

# Course Grading

| Letter Grade | Score (Percent Grade) |
|---|---|
| A | 90% - 100% |
| B | 80% - 89% |
| C | 70% - 79% |
| D | 60%-69% |
| F | <60% |

# Quick Notes About Grading

- Graduate students are expected to participate in class discussions
  - Extra points in some cases!
- For quizzes and exams, there will be no make ups
- Post due homework will receive immediate 50% deduction
  - Usually homework are due Wednesday at 11:59 PM
  - Once class start on Thursday, undelivered homework will get a 0

# Optional Text Books

- Chambers, Bill, and Matei Zaharia. *Spark: The Definitive Guide: Big Data Processing Made Simple.* O'Reilly Media, 2018.
- Damji, Jules S., et al. *Learning Spark: Lightning-Fast Data Analytics 2nd Edition* O'Reilly Media, 2020. **NOTE: Free download at: https://databricks.com/p/ebook/learning-spark-from-oreilly, use your UMBC email address.**

# Norms

- Respect everyone
- Communication is key
- Ask lots of questions
- Mistakes are good

# Why Data Science

- High demand for Data Scientists
- Growing job market
- Essential skills for IT professionals, business professionals, managements, statistician, and others

# What is Data Science

- A data scientist is a popular field (somehow new too) that encompass knowledge from the following fields:
  - – Data architecture
  - – Data analysis
  - – Data development
  - – … and others

# As a Data Scientist!

To qualify as a Data Scientist, you have to have experience in these four quadrants:

1. Database Management, including traditional SQL and Querying
2. Predictive Analytics, including modeling and Machine Learning
3. Big Data, for unstructured data analysis, mining,  and trends
4. Data Visualization and presentation

# Important Activities

- 14 remaining weeks of classes
- One big data project
- One technical paper

# Class Schedule

- Refer to the syllabus.

# The Class Project

- Start thinking about your project today
- Formulate the idea then draft a problem statement, and be ready to defend it
- Sample project topics will be shared in the next lecture
- Your responsibility is to enhance on the presented project topic and implement something new
- Project Logistics:
  - Must use Apache Spark
  - Data must have minimum of 1000 records

# Project Schedule

| Week # | Activity | Expected Outcome |
|---|---|---|
| 1 | Present the project assignment to students | Start thinking about big data project |
| 5 | Project idea ready | Prepare a slide deck for presenting the project idea |
| 10 | Present project progress report | Every student will prepare and submit a project progress status report |
| 14<br>15 | Project presentations | Prepare a slide deck for presenting your project to students |
| 15 | Project report | Final slide deck and 1-page summary due |

# Project Presentation & Defense

- In your project proposal defense:
  - Clearly illustrate the idea
  - Present the expected outcome
- In your final project delivery:
  - Show your implementation
  - Present results in graph formats
  - Show your contributions

# The Technical Paper

- Start working on the technical (research) paper today
- The paper should cover an innovative topic in Big Data
  - Copied or regurgitated papers will not be accepted!!
- Sample Topics:
  - Cognitive Computing & Big Data
  - Machine Learning & Big Data
  - Cybersecurity & Big Data
  - Cloud Computing & Big Data

# Technical Paper Defense & Presentation

- Individual papers, no teaming up!
- In your proposal defense, you should demonstrate the following:
  - Authenticity of the paper
  - Innovation and new ideas
  - Quality of the work
- In your final delivery:
  - Ensure solid technical writing
  - Presentation is of good quality
  - Organized presentation so other students can benefit from

# Technical Paper Schedule

| Week # | Activity | Expected Outcome |
|---|---|---|
| 1 | Present the technical research paper assignment to students | Start thinking about proposals for the paper |
| 4 | Technical paper proposal ready for defense | Every student will submit his paper proposal |
| 9 | Present near complete paper and share progress | Every student will prepare and submit a paper progress report |
| 13 | Deliver Final paper | Final paper deliver (due 11:59PM) |

# Important: Class Benefits

- What will you get from this class?
  - Big data foundation skills
  - Expertise on Apache Spark
  - Intelligently talk about big data platforms
  - Pass a big data skills interview
  - Prepare you to build expertise and skills in your job
  - Eventually (and through job training and hands-on expertise), become an expert

# Homework

- Sign up for Databricks Community Edition
  - Databricks: https://community.cloud.databricks.com/
- Download Docker
- Fork the class repository in Github and add me as a collaborator (enkeboll)

# Homework

- Assignment: write a script that calculates the number of **unique** words in Tolstoy's *War and Peace*
- Requirements:
  - Can be written in any modern, open programming language (Python, Node, R, Java, Bash; no Matlab, SAS, SPSS, etc)

# Homework

- Requirements
  - If running it requires anything more than executing the file, include a readme with instructions. They should not involve downloading third party libraries.
  - **In a branch**, add this file to your new github repo with the name **homework/hw01-tolstoy.[py|js|sh|etc]**

# Homework

- Requirements
  - Open a pull request and **tag me as the reviewer**.
  - Finally, you will take the link to your pull request and submit it in Blackboard to the open assignment (forthcoming).

# Questions