Lecture 0: Course Introduction

Print to PDF

Contents

Teaching squad

- This course has been redeveloped in 2024 to a great extent to accommodate the latest trends.

Office Hours

- Check MDS calendar

Todays Agenda

- Course Overview
- How does data appear on a website
- Data management in a big data environment
  - What is big data?
  - Which tool to use?
  - How big is big data?
  - Explore big data locally
  - What are things to consider when it is big data?

Course Overview

My Goals for the Course

- Understand various dimensions of Big Data and identify solutions based on the nature of your data.
- Learning various approaches to store and process big data.
- Getting comfortable with using AWS cloud solutions.
- Setup infrastructure in AWS for scale-UP and scale-OUT solutions and have an understanding of important services.
- Apply your learning's from previous courses to act as an independent individual in a corporate environment.

Prerequestiles

- Python, R, and SQL programming.
- Understanding of workflows and basic commands to interact with Linux based systems.

Course plan

Date	Topic	Assessments/Milestone Due due
Lecture 1	Getting Data, Big data, Processing Big data locally	
Lecture 2	Big data file formats	Milestone 1 - Check MDS calendar
Lecture 3	Getting ready for the cloud, security before everything	
Lecture 4	Important cloud components - storage (S3), Computing (EC2)	Milestone 2 - Check MDS calendar
Lecture 5	Review of MapReduce, Hadoop, EMR	
Lecture 6	Distributed computing in the cloud	Milestone 3- Check MDS calendar
Lecture 7	Options for Deployment	
Lecture 8	Class Conclusions/ Special Topics	Milestone 4- Check MDS calendar

General course announcements/logistics

- There are 2 `.yml` files in student repo. `525.yml` is the one you need for milestone 1 and the non-optional parts of the lectures. You can use `525_dev.yml` if you want to try out optional stuff in the lecture notes.
- Start working on milestones early so that you can ask questions earlier during the week. You might not get a response if you plan to post questions over the weekend.
- Refresher questions will be updated based on the previous class.

- Details regarding "Thought/Discussion" during class will be updated with more information (if needed) after the class.
- Milestones will be released on Monday before 8 AM.
- Grading will be done 1 week after the due date. After release, you can find solutions in the `solution` folder in the student repo.
- For any regrade request, reach out to TA who graded.
- You will receive an email welcoming you to AWS before week 2, and you will be given a $100 credit for AWS.

Course Model

Individual Milestones

Important

This is a project-based course, but everyone is expected to work on it individually, just like how it is with any lab-based course. You are allowed to discuss with your peers like you would in a lab-based course. There are some parts of milestones where you need to work with a partner, which will be mentioned in the milestone.

Various roles you will be taking in this course and milestone

- Data Engineer
- Devops Engineer / Analytical Engineer
- Infrastructure Engineer
- Data Scientist

Toolsets we will be using

- Data format - CSV, Parquet
- Cloud Vendor - AWS
- Cloud services - EC2, EMR, S3, RDS, EBS, Sagemaker
- Analysis - Python, R, SQL
- Notebook - Jupyter, JupyterHub(entrprise solution), Sagemaker notebooks

Milestone 1 (25 %)

Get the data from the web using API, process it, and convert it to an efficient file format.

- Experience the big data, and understand how your data life gets complicated when it's big data.
- Apply techniques to handle this situation locally.
- Play around with some big data file formats and reason how it can help.
- Understand the process of serialization and deserialization and why it matters.

Milestone 2 (25 %)

Move the data to the cloud set up the infrastructure in the cloud for development.

- Introduce you to AWS.
- Get experience in S3, EC2.
- Configuring the security.
- Set up a collaborative environment (JupyterHub) in the cloud.

Milestone 3 (25 %)

Setup your distributed infrastructure in the cloud and run the same ML model using Spark.

- Setup and configure EMR for Hadoop/spark cluster.
- Use of spark.

Milestone 4 (25 %)

Deploy your ML model in the cloud so that other consumers can use it.

- AWS scripts
- Apply your learnings on flask and deployment.

Things to keep in mind:

- Unlike your previous courses where the standards/basics include using pandas, NumPy, dplyr, databases, and scikit-learn, things are evolving fast as we get into this big data zone. It is a must-have skill to adapt to any package/infrastructure/file-format that comes our way, so starting with the documentation is essential. In this course, you will have many opportunities to develop this skill (if you don't have it already). I will be using some experimental elements and discussing them, as I think they are gaining traction in the big data zone, but things can change at any time.