DS-610 - Big Data Analytics

Course Description

Big Data (Structured, semi-structured, & unstructured) refers to large datasets that are challenging to store, search, share, visualize, and analyze.

Gathering and analyzing these large data sets are quickly becoming a key basis of competition. This course explores several key technologies used in acquiring, organizing, storing, and analyzing big data.

Topics covered include

- Hadoop
- unstructured data concepts (key-value)
- Map Reduce technology
- Pig and Hive
- NoSQL storage solutions like HBase, Cassandra, and Oracle NoSQL and analytics for big data.

A part of the course is devoted to public Cloud as a resource for big data analytics. The objective of the course is for students to gain the ability to employ the latest tools, technologies and techniques required to analyze, debug, iterate and optimize the analysis to infer actionable insights from Big Data.

Prerequisites

This course requires following courses to be taken before this course.

- Introduction to Data Science (DS-510)
- Data Analysis and Decision Modeling (DS-520)
- Big Data & Data Management (DS-530)

Learning Outcomes

This course aims to teach you the following concepts.

- Understand the challenges comes with big datasets.
- Learn and use new algorithms to analyze large datasets
- Gain experience in tackling complex level big data problem
- Learn the big data ecosystem and be able to understand it's tools and methods.

Textbooks & Tools

Book we will follow for Spark. You can find it here.

```
Title: PySpark Recipes
Author(s): Raju Kumar Mishra
Release date: 2018
Publisher(s): Apress
ISBN: 978-1-4842-3141-8
```

Course Outline

Below are the distribution of content and assignments for the semester.

Week	Tasks	Description
Week 01	Assignment0	Python Warmup
Week 02	Assignment1	Introduction to parallelism
Week 03	FP Kickoff	Introduction to clustering systems, MongoDB, and Cassandra, Cloud Computing
Week 04	FP Review	Map-Reduce Algorithm
Week 05	Assignment2, FP Review	Spark and Hadoop Ecosystem. Understanding Spark architecture.
Week 06	FP Review	Spark API Introduction (RDDs, Accumulators, Broadcast variables, Dataframe)
Week 07	Midterm	Midterm (Everything upto Spark), FP Start
Week 08	FP Review	Spark Examples 1 (RDD and Dataframe), FP Session 1
Week 09	Assignment3, FP Review	Spark Examples 2 (Machine Learning), FP Session 2
Week 10	Final	Spark Examples 3, FP Session 3
Week 11	FB Presentations	Final Project Presentations, What's Next

- FB stands for `Final Project`.

Grade Determination

The items that student is responsible for this class is the following. Students will be graded according to this chart.

Item	Percentage
Assignments	20 %
Projects	30 %
Midterm	20 %

Item	Percentage
Final	20 %
Participation	10 %
Bonus (Maybe)	10 %

Partication means attending to discussions in Blackboard, class sessions, etc.

Above list is subject to change.

Grade Scale

Based on the Grade Determination, a final letter grade will be assigned to student based on the following scale.

Min	Max	Grade
94	100	A
87	92	A-
83	86	B+
80	82	B
77	79	B-
73	76	C+
60	72	C
0	59	F

Attendence Policy

Due to COVID-19, there is no restriction on attending the lecture, however, attending the classes will benefit on your learning curve a lot more instead of watching lectures online. As said, here are the guides:

- We record the classes and you can always revisit or watch previous classes. Links will be shared.
- When attending to class, please keep your **camera on** and your **microphone off**.

Academic Honesty and Student Conduct

Students need to submit **only their own original work** (e.g. Code, PPT, figures, visualizations...). Student need to familiarize themselves with the academic rules of the University. In case a student is found guilty under the act of plagiarism, his/her test or assignment will be graded zero.

If plagiarism occurs twice, student will receive 'F' grade with immediate effect. It is expected that everybody turns off/mute all devices that emit sounds and noises that may interrupt the class (e.g. mobile phones, pagers, watch alarms). If an occasion arises, in which a student may need to leave the class to receive a phone call (important call), use rest room or get drinking water, he or she should silently walk out without disturbing rest of the class.

Working on assignments and project work, that belongs to another course is **STRICTLY NOT ALLOWED**.

Copying your colleagues code is **STRICTLY NOT ALLOWED**. All submissions will be submitted to autodetection plagiarism tool that our university provides!

Agreement between You and the Instructor

I certify that in this course, my contribution and assignments will be my own work, based on my personal study and/or research and that I am acknowledging all material and sources will be used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication.

I also certify that that in this course, the assignments I will be submitting will not previously been submitted for assessment in any other unit, except where specific permission has been granted from all unit coordinators involved, or at any other time in this unit, and that I have not copied in part or whole or otherwise plagiarized the work of other students and/or persons.

Student		**Instructor**	
Student's Name	.....	Instructor's Name	.....
Signature	.....	Signature	.....

References

Useful links regarding university.

- Check out more information about the prerequisites on curriculum of the SPU's data science.
- Additional information about this class can be found on SPU's website, which requires student's login.
- Blackboard Classroom DS-610-PHYB-21SPTR: Big Data Analytics.