Spring 2020 Blueprint

Data Mining With MySQL, NoSQL, Hadoop, Spark, And Hive

COURSE DESCRIPTION

A deep dive into the principles and techniques of data mining and driving business insight using noSQL, Hadoop, Apache Spark, and Apache Hive. Topics include querying datasets using SQL using the Dataiku Studio. What is more, topics will include machine learning using PySpark, querying data using Hive (HQL), and creating business value based on data stored in Hadoop (Spark environment).
Dataiku Studio will be used to practice SQL, queries on databases in PostgreSQL and MongoDB.

Students will also learn how use Python packages to connect to a Hadoop database. Students will be using Dataiku to run Pyspark queries in Hadoop and perform data mining in the Dataiku environment.

COURSE TOPICS

This course will cover many data science topics including machine learning, predictive modeling, and driving business insight from quantitative frameworks. All these topics will be related to the Hadoop ecosystem that uses cluster computing to accommodate large datasets ("Big Data"). What is more, the course will give students the ability to learn new Python language frameworks that help when working with Big Data. Besides additional Python skills, students will learn the basics of SQL languages and NoSQL such as MongoDB.

LEARNING OUTCOMES

CLO1: Understanding core concepts behind Hadoop and cluster computing for Big Data.
CLO2: Data mining using noSQL and HiveQL
CLO3: Use Dataiku to access a Hadoop cluster and perform data mining using Pyspark
CLO4: Performing machine learning models with Spark.

Required textbooks

- Required: "Data Analytics with Hadoop: An Introduction for Data Scientists" By Benjamin Bengfort, Jenny Kim, 2016, ISBN-10: 9781491913703

Assessment Descriptions

Assessment 1 Description (required):

This will be an online quiz to test the knowledge gained in classes 1, 2, 3,and 4. The quiz will have 3-4 questions. 15 minutes per question. The exam will be performed in DATAKU using one of the PySpark notebooks.

Assessment 2 Description (required):
Students will have to solve a given business case.
IBM-HR-Attrition Business case. This is a proprietary business case, designed for the purpose of this class. It is using publicly available data.
This business case will be discussed in the 2nd class.

Assessment 3
Prior to the first class, students will have to complete a course on Data Camp covering basic SQL.

Assessment 4 Description: Optional (Class Participation):
Total of 2 participation assessments completed in class in a form of a quiz.

Assignment 1 [A1]
To be copied from the previous Data Mining class

Assignment 2 [A2]
To be copied from the previous Data Mining class

Assignment 3 [A3]
Marked as complete or incomplete – Has to be completed on time (prior to class1) in order to be marked complete

Assignment 4 [A4]
To be copied from the previous Data Mining class

Course Outline

Class 1 (4 hours)

Topic(s) Covered	**INTRO / SETUP + mySQL** -Standard database structure overview, relational databases -> basic SQL -How does Hadoop fit in a data process (aka. "The pipline") -Hadoop architecture – understanding how the nodes interact **SQL syntax types:** -basic -analytical -conditional **Use Dataiku to query a PostgreSQL:**
Guest speaker (if applicable)	
Required Reading & Preparation	

Class 2 (4 hours)

Topic(s) Covered	**noSQL:** -noSQL -> comparing with mySQL -driving business insight using noSQL -document based data containers -using the mongoDB database -using the mongoDB Atlas interface to query the documents -the noSQL syntax in MongoDB - the "$find:" query -the "$aggregate:" query -other useful queries that are comparable with mySQL **Use Dataiku to query MongoDB using document based syntax:** **A4 assignment**
Guest speaker (if applicable)	
Required Reading & Preparation	"Data Analytics with Hadoop: An Introduction for Data Scientists" By Benjamin Bengfort. Chapters 1,2,3

Class 3 (4 hours)

Topic(s) Covered	**Connecting SQL and NoSQL** **Using DATAIKU as a platform to combine the SQL and NoSQL (MongoDB) databases.** -connecting and visualizing data from SQL Server -connecting and visualizing data from MongoDB **Hadoop 1 (Hive)** **Deploying a Hadoop cluster** -Understanding the Hadoop deployment process – master and slave nodes - using HiveView to analyze performance of the Hadoop cluster -using Jupyther to write PySpark programs and execute in Hadoop **Basic data cleaning, massaging, manipulation in Hive and PySpark** -using SQL in the Jupyther Notebook PySpark environment %%sql -writing python programs to clean the data and create summary tables describing the data
Guest speaker (if applicable)	
Required Reading & Preparation	"Data Analytics with Hadoop: An Introduction for Data Scientists" By Benjamin Bengfort. Chapters 4,5

Class 4 (4 hours)

Topic(s) Covered	**Hadoop 2 (PySpark)** **Creating a data mining pipeline using PySpark in a DATAIKU notebook:** -leveraging the work from class 3 -data ingestion in Hive + data cleaning and massaging+ plotting -building a simple linear regression model to find causation **Logistic regression:** -understanding the difference between logistic regression and linear regression -using logistic regression to drive business decisions -modeling probability of business success **Decision trees:** -building a decision tree in PySpark -comparing a decision tree results with a logistic regression. **A4 assignment**
Guest speaker (if applicable)	
Required Reading & Preparation	"Data Analytics with Hadoop: An Introduction for Data Scientists" By Benjamin Bengfort. Chapters 6,7

Class 5 (4 hours)

Topic(s) Covered	**MACHINE LEARNING with PYSPARK** **Mid-term quiz.** **Analytics and other machine learning concepts using PYSPARK in DATAIKU:** -Advance machine learning using Spark -Neural Networks -Ensembled models -Managing a Hive node using HQL -transferring data from a Hive node to the Spark environment. **Group presentations**
Guest speaker (if applicable)	
Required Reading & Preparation	"Data Analytics with Hadoop: An Introduction for Data Scientists" By Benjamin Bengfort. Chapters 8,9,10

Note: At least one class session must focus on an ESR (Ethics, Sustainability, and Responsibility) topic. Please indicate which class session(s) will address ESR and what topic(s) you will cover.

Class session(s) related to ESR	**Class 5: SPARK MACHINE LEARNING**
ESR topic(s) covered	How can Spark help make ethical decisions when running machine learning classification frameworks.