Sungkyunkwan University (SKKU) International Summer Semester (ISS) 2023

INTRODUCTION TO BIG DATA ANALYSIS

Prof. Jongwook Woo, California State University Los Angeles

SHORT COURSE DESCRIPTION

Students develop practical knowledge of Data Engineering, Data Analysis, and Visualization in Big Data platform. The students learn the Big Data architecture to store, analyze, and visualize large-scale data using Hive QL in Hadoop with hands-on examples with cloud computing systems. The students understand the following from the lecture and practical lab:

- Understand the genesis of Big Data Systems
- Understand practical knowledge of Big Data Analysis using Hadoop, Hive and QL
- Provide the student with a detailed understanding of effective behavioral and technical techniques in Cloud Computing on Big Data
- Demonstrate knowledge of Big Data in industry and its Architecture
- Learn data analysis, modeling and visualization in Big Data systems

READING MATERIALS

1. Instructional materials (Lecture and Lab) from the instructor
2. Hadoop: The Definitive Guide by Tom White
3. https://hadoop.apache.org/
4. https://www.cloudera.com/tutorials.html

COURSE REQUIREMENTS AND GRADING

Students are expected to attend every class session. Since Cloud computing and Big Data concepts are presented during class time, class attendance is essential for successful completion of assignments and tests. As a large part of the course involves work on cloud computing, it is essential that you utilize the time in class for discussion and exercises on the computer. If attendance is not possible for one of the class meetings, please contact the instructor beforehand. Students are expected to use the equipment of the computer labs on campus if you do not have a personal computer nor internet.

SKKU regulations require students to attend at least 80% of all classes. All ISS classes are pass/fail based on the student academic achievement evaluated by grades on a scale of 100 points (grade of 60 or above is Pass).

Grading Policy:

- Class Activities (Pop quizzes, Attendance, Participation in Class): 10%
- Lab Assignments: 30%
- Midterm Exam 25%
- Final Exam 35%
- Total 100%

COURSE SCHEDULE

– WEEK I –

Monday (26 June)

DAILY TOPIC & CONTENTS	**COURSE MATERIAL & ASSIGNMENTS**	**REFERENCE**
Course Overview	Syllabus	
Ch 1 An Introduction to Big Data and Cloud Computing Systems (p1 - p17) Lab 1: set up Linux CLI and Connect to Big Data Server	Introduction To Big Data Introduction to Hadoop Data Analysis with Big Data	Reading Instructor's material about the systems of Big Data and Cloud Computing

Tuesday (27 June)

Ch 1 (Continued, p18 – p45) Ch 2 Big Data Cluster (p1 – p11) a. Introduction to Hadoop b. Motivation for Hadoop	Understanding HDFS Understanding Hadoop Clusters Understanding YARN Architecture Understanding MapReduce	Reading Instructor's material about Hadoop
Lab 2: set up cloud computing accounts such as Oracle Big Data Compute Edition and Practice Linux/HDFS Shell Commands		

Wednesday (28 June)

Ch 3 HDFS and Hive (p1 - p37)	Understanding Hive Understanding Hive Architecture Learn Hive QL	Reading Instructor's material about HDFS, MR, Hive
Lab 3 Part 1: HVAC Sensor Data Analysis in Hive		

Thursday (29 June)

Ch 4 Hive Detail (p1 - p29)	Complex Data Type Operators External Table Insert Data	Reading Instructor's material about MR, Cluster, Ecosystems, Hive

Lab 3 Part 2: HVAC Sensor Data Analysis in Hive	Cloud Architecture and Code in detail	

– WEEK II –

Monday (3 July)

Ch 5 Sqoop and Join in Hive (p1 – p 28)	Sqoop Inner Join Outer Join Union	Reading Instructor's material about Join in Hive
Lab 4 Part 1: IoT Sensor Log Data Analysis using Hive in Oracle Big Data	Code in detail about RegEx expression	

Tuesday (4 July)

Ch 6 Text Processing in Hive (p1 – p24)	Functions Text Processing String Functions Table Generating Functions	Reading Instructor's material about Hive Text Analysis
Lab 4 Part 2: IoT Sensor Log Data Analysis using Hive in Oracle Big Data		

Wednesday (5 July)

Ch 7 Text Processing with NGram (p1 – p16)	Ngram Function Context Ngram Function	Reading Instructor's material about NGram and Functions of Hive
Lab 5 Part 1: Sentiment Analysis with N-Grams Text Processing		

Thursday (6 July)

Midterm Exam	Questions for Lectures and Labs learned in Lecture 1 through Ch6/Lecture 6	

– WEEK III –

Monday (10 July)

Ch 8 Advanced Text Processing in Hive (p1 – p11)	Text Processing Regular Expression RegEx Function	Reading Instructor's material about RegEx

	RegEx SerDe	
Lab 5 Part 2: Sentiment Analysis with N-Grams Text Processing		

Tuesday (11 July)

Ch 8 Advanced Text Processing in Hive (Cont., p12 – p20) Ch 9 Cast, Time, Alias (p1 – p20)	Type Conversion Date Format Alias View	Reading Instructor's material about Data Type, Format, Alias, View
Lab 6 Part 1: NGram Sentiment Text analysis of Twitter social media data		

Wednesday (12 July)

Ch 10 Table for Json and Extended (p1 – p26)	Json file Some Regex and Case Extended Describe	Reading Instructor's material about Data File and Extended
Lab 6 Part 2: NGram Sentiment Text analysis of Twitter social media data		

Thursday (13 July)

Ch 10 Table for Json and Extended (p26 – p31) Ch 11 Alter Table, Hive CLI and Other Join (p1 – p22) Ch 12 File Type and Data Type Delimiter (p1 – p5)	File Type Data Type Delimiter Hive SerDe	Reading Instructor's material about Hive SerDe, File & Data Types
Lab 7 Part 1: IoT data of TruckEvent		

– WEEK IV–

Monday (17 July)

Ch 12 File Type and Data Type Delimiter (p6 – p16) Ch 13 Hadoop Cluster for Computing	Alter Table Hive CLI Semi Join, Cross	Reading Instructor's material about Hive CLI, Cross, Semi-Join

(p1 – p15)		
Lab 7 Part 2: IoT data of TruckEvent		

Tuesday (18 July)

Ch 13 Hadoop Cluster for Computing (p16 – p47)	Understanding HDFS	Reading Instructor's material about Big Data HDFS and Cluster
Lab 8: Sentiment Analysis using Big Data and Tableau		

Wednesday (19 July)

Final Exam	Questions for Lectures and Labs learned in Lecture 8 through Lecture 14	