연세대학교 YONSEI UNIVERSITY

IEE3593-02 INTRODUCTION TO BIG DATA ANALYSIS

COURSE TITLE	INTRODUCTION TO BIG DATA ANALYSIS	**COURSE CODE-SECTION**	IEE3593-02
CREDIT	3	**CLASS PERIOD**	P2(11:00~12:40)
OFFICE		**OFFICE HOURS**	
INSTRUCTOR	JONG WOOK WOO	**EMAIL**	jwoo5@exchange.calstatela.edu

[COURSE INFORMATION]

COURSE DESCRIPTION & BRIEF INTRODUCTION OF THE COURSE	The course is to develop (1) Practical knowledge of Big Data, (2) How to process and store Big Data using scalable distributed systems in cloud computing, and (3) How to analyze data with MapReduce and Hive. The course includes hands-on examples and a project to implement data engineering, data analysis and visualization using Big Data platform in Amazon AWS Cloud Computing.					
COURSE GOALS	1. Understand the genesis of Big Data Systems					
	2. Understand practical knowledge of Big Data Analysis using Hive, Sqoop, Linux Shell					
	3. Provide the student with a detailed understanding of effective behavioral and technical techniques in Cloud Computing on Big Data					
	4. Demonstrate knowledge of Big Data in industry and its Architecture. Learn data analysis, modeling and visualization in Big Data systems					
COURSE METHODS (100% TOTAL)	LECTURE	PRACTICE TRAINING	PRESENTATION	DEBATE	TEAM PROJECT	
	50	50				
GRADING POLICY (100% TOTAL)	MIDTERM	FINAL EXAM	QUIZ	INDIVIDUAL ASSIGNMENT	TEAM ASSIGNMENT	ATTENDANCE	OTHERS
	20	30		25	15	10	
PREREQUISITE	a. Mastery over Microsoft Windows and its File Management (Windows Explorer, Mac Finder) facilities b. Basic knowledge of any programming language (SQL, Python, Java) c. Basic knowledge of BI tools					
COURSE REQUIREMENTS	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 at Yonsei University if you do not have a personal computer nor internet.					

Course Syllabus
2025 YONSEI INTERNATIONAL SUMMER SCHOOL
6-WEEK PROGRAM

연세대학교 YONSEI UNIVERSITY

TEXTS & REFERENCES	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

[WEEKLY SCHEDULE]

WEEK	DAILY TOPIC & CONTENTS	COURSE MATERIAL & ASSIGNMENTS	REFERENCE
WEEK1 (June 30 to July 3, 2025)	Course Overview 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 Ch 2 Big Data Cluster (p1 – p11) Lab 2: set up cloud computing accounts Ch 3 HDFS and Hive (p1 – p13) Lab 3 Part 1: HVAC Sensor Data Analysis in Hive	Introduction To Big Data Introduction to Hadoop Data Analysis with Big Data Understanding HDFS Understanding Hadoop Clusters Understanding YARN Architecture Understanding MapReduce Understanding Hive Understanding Hive Architecture Learn Hive QL	Reading Instructor's material about the systems of Big Data and Cloud Computing Reading Instructor's material about Hadoop Reading Instructor's material about HDFS, MR, Hive
WEEK 2 (July 7 to July 10, 2025)	Ch 3 HDFS and Hive (p13 – p40) Lab 3 Part 2: HVAC Sensor Data Analysis in Hive Ch 4 Hive Detail (p1 - p29) Lab 3 Part 3: HVAC Sensor Data Analysis in Hive Ch 5 Sqoop and Join in Hive (p1 – p 28) Lab 4 Part 1: IoT Sensor Log Data Analysis Ch 6 Text Processing in Hive (p1 – p15) Lab 4 Part 2: IoT Sensor Log Data Analysis	Learn Hive QL Complex Data Type Operators External Table Insert Data Sqoop Inner Join Outer Join Union Functions Text Processing	Reading Instructor's material about HDFS, MR, Hive Reading Instructor's material about MR, Cluster, Ecosystems, Hive Reading Instructor's material about Join in Hive Reading Instructor's material about Hive Text Analysis
WEEK3 (July 14 to July 17, 2025)	Ch 6 Text Processing in Hive (p16 – p24) Lab 4 Part 3: IoT Sensor Log Data Analysis	String Functions Table Generating Functions	Reading Instructor's material about Hive Text Analysis

	Ch 7 Text Processing with NGram (p1 – p16) Lab 5 Part 1: Sentiment Analysis with N-Grams Text Processing Lab 5 Part 2: Sentiment Analysis with N-Grams Text Processing Ch 8 Advanced Text Processing in Hive (p1 – p20) Lab 5 Part 3: Sentiment Analysis with N-Grams Text Processing	Ngram Function Context Ngram Function Text Processing Regular Expression RegEx Function RegEx SerDe	Reading Instructor's material about NGram and Functions of Hive Reading Instructor's material about RegEx
WEEK4 (July 21 to July 23, 2025)	Midterm Exam Ch 9 Cast, Time, Alias (p1 – p20) Lab 6 Part 1: NGram Sentiment Text analysis of Twitter social media data Lab 6 Part 2: NGram Sentiment Text analysis of Twitter social media data Ch 10 Table for Json and Extended (p1 – p17) Lab 6 Part 3: NGram Sentiment Text analysis of Twitter social media data	Type Conversion Date Format Alias View Json file Some Regex and Case Extended Describe	Reading Instructor's material about Data Type, Format, Alias, View Reading Instructor's material about Data File and Extended
WEEK5 (July 28 to July 31, 2025)	Ch 11 Alter Table, Hive CLI and Other Join (p1 – p22) Lab 6 Part 4: NGram Sentiment Text analysis of Twitter social media data Ch 12 File Type and Data Type Delimiter (p1 – p26) Lab 7 Part 1: IoT data of TruckEvent Ch 13 Hadoop Cluster for Computing (p1 – p47) Lab 7 Part 2: IoT data of	File Type Data Type Delimiter Hive SerDe Alter Table Hive CLI Semi Join, Cross	Reading Instructor's material about Hive SerDe, File & Data Types Reading Instructor's material about Hive CLI, Cross, Semi-Join

	TruckEvent Lab: Final Project Preparation I Lab: Final Project Preparation II Lab: Final Project Preparation III	 Lab: Final Project Preparation I 1. Grouping: (3 ~ 4 members of a team) - choose team role: What tasks each member to do (Group score and individual score depending on the project contribution) 2. Data Engineering: Data Collection, BigDAI center: https://www.calstatela.edu/centers/hipic/related-site Lab: Final Project Preparation II 1. Data Engineering - Upload Data set to HDFS - Hive QL Table Creation in a group DB - Share the group DB 2. Data Engineering: - Query the data to confirm - Start Data Analysis Lab: Final Project Preparation III 1. Data Analysis - Build Hive QL - Group By 2. Data Visualization: - Excel, Power BI, Google Spreasheet, etc	
WEEK6 (August 4 to August 6, 2025)	Final Exam Lab: Final Project Slide Ready Final Project Presentation	 Lab: Final Project Presentation Slide Ready	