Big Data Course

High Level Design

& Course Details

Course Brief

Topic	Big Data Course
Course Schedule	▪ Face-to-face 4 hours daily course: 3 months (total 240 hours) - Lecture & Exercise: 2 months (160 hours) - Capstone Project: 1 month (80 hours)
Learning Environment	▪ Classroom - Projector, Microphone, Laser pointer, White board ▪ PC - Minimum: Windows10 64 Bit + CPU i5 + RAM 16GB - Recommended: Windows10 64 Bit + CPUi7-8700 3.19 GHz + RAM 32GB
Learning Objectives	Learn the entry level skills need to be a data engineer by understanding a set of processes for Big Data architecture, ingestion, storage, analysis, and visualization. Cultivate the practical skills through a series of hands-on exercises. Finally, combine the acquired skills by completing a capstone project. At the end of the course, students will be expected to: ▪ Understand the background and trends of Big data technology. Figure out how the Big data ecosystem is applied and operated in real use cases ▪ Ingest the data from the source and store it in the distributed storage system ▪ Analyze complex data elements and systems, data flow, dependencies, and relationships using analysis methods ▪ Design data models as well as process them against new datasets on distributed computing environments ▪ Understand trends, outliers, and patterns of data with visualization tools
Course Prerequisites	▪ Required - Basic understanding of program in at least one of these languages (e.g. Java, Python, Scala) ▪ Recommendation - Previous experience in Cloud Environment - Basic knowledge of SQL and DBMS
Audience & Characteristics	▪ Target - Youth (age 18~24), interested in pursuing a career in Big Data Engineering, who need the appropriate education for the career - Official Target for SIC program will follow the given characteristic. However, actual participants will mainly consist of undergraduate students in STEM major and a few graduate students within the given age range because secondary school graduates will barely meet prerequisites given above ▪ Characteristics - Educational background: successfully completed high school level STEM courses or higher education - Level for understanding: possess basic knowledge in programming. - Expectations: expects to obtain necessary knowledge and skills for entry-level job placement in Big Data field

▶ **Lecture and Exercise** (2 months, 160hours)

Course Contents	Duration
Chapter 1. Introduction to Big Data	**4H (total)**
- Unit 1. Big Data Overview and Background	2.5H
- Unit 2. Current Trends in Big Data	1H
- Quiz	0.5H
Chapter 2. Fundamentals of Big Data	**9H (total)**
- Unit 1. Big Data Processing	2H
- Unit 2. Hadoop Core & Eco system overview	2H
- Unit 3. Hadoop Architecture for Big Data	4H
- Quiz	1H
Chapter 3. Big Data Ingestion	**26H (total)**
- Unit 1. Data Migration from EDW	3H
- Unit 2. Streaming Real-Time Data Sources	3H
- Unit 3. Creating Data Pipelines with Apache NiFi	14H
- Unit 4. Apache Kafka	4H
- Quiz	2H
Chapter 4. Big Data Storage	**24H (total)**
- Unit 1. Data Storage Alternatives	8H
- Unit 2. NoSQL	14H
- Quiz	2H
Chapter 5. Big Data Analytics	**32H (total)**
- Unit 1. Introduction to SQL	8H
- Unit 2. Basic Analytics	8H
- Unit 3. Advanced Analytics	8H
- Unit 4. Streaming Data Analysis Architecture	4H
- Quiz	4H
Chapter 6. Big Data Processing with Apache Spark	**44H (total)**
- Unit 1. Unstructured Data Processing	15H
- Unit 2. Structured Data Processing	13H
- Unit 3. Processing Streaming Data	7H
- Unit 4. Apache Spark Applications	5H

- Quiz	4H
Chapter 7. Big Data Modeling and AI	**14H (total)**
- Unit 1. Machine Learning Models	9H
- Unit 2. Apache Spark Machine Learning Library (MLlib)	4H
- Quiz	1H
Chapter 8. Data Visualization	**4H (total)**
- Unit 1. Open Source Visualization Tools	2H
- Unit 2. Commercial Visualization Tool	1H
- Quiz	1H
Chapter 9. Security and Access Control	**3H (total)**
- Unit 1. Secure Access to Cluster Services	1H
- Unit 2. Secure Access to Cluster Data	1H
- Quiz	1H
Total	**160H**

▶ **Capstone Project** (1 month, 80hours)

Course Contents	Duration
Capstone Project Overview	**1H (total)**
- Student Guide Explanation	0.5H
- Document Explanation	0.5H
Chapter 10. Starting a Big Data Project	**3.5H (total)**
- Project Preparation	0.5H
- Big Data Architecture Design	3.5H
Chapter 11. Big Data Capstone Project Tutorial	**2.5H (total)**
- Designing a data ingest pipeline.	1.5H
- Fundamentals of exploring and transforming data for ETL	1H
- Creating and presenting data analysis reports	0.5H
※ During the capstone project, student's project activities take more time than lecture itself. Please expect up to 80 hours to complete the whole project	**73H**
Total	**80H**

Course Details

Chapter	Details	Duration
1	**Chapter 1. Introduction to Big Data**	**4H (Total)**
	Objective: Understand the role of data in the human evolution. Gain insight to the major changes in how data is generated and how savvy businesses are taking advantage of it to gain valuable insight. Understand how existing technology and platforms failed in the face of changing data landscape and what new requirements were incorporated into Hadoop to meet these challenges.	
	Unit 0. Course Roadmap	0.5H
	Unit 1. Big Data Overview and Background • Background of the emergence of big data • Big data definition • The difference between traditional data analysis and big data analysis • Key elements and characteristics of big data • The evolution of processing and organizing data • Changing hardware landscape - availability and cost • Cost-effective scalable and fault-tolerant architecture	2H
	Unit 2: Current Trends in Big Data • Role of public cloud based services • XOps - hype or real? • Edge Computing to Artificial Intelligence • Citizen Data Scientists and Citizen Data Engineers	1H
	Quiz (Written)	0.5H
2	**Chapter 2. Fundamentals of Big Data**	**9H (Total)**
	Objective: Gain insight to the objectives and goals of Big Data processing through examining some use-case scenarios. Understand the fundamentals of a distributed parallel cluster architecture and how enterprises deploy Big Data platforms, either on-premise or through the use of public cloud services. Understand how on-premise platforms, such as Hadoop, provide storage and computing for large scale data. Gain an overview to the data pipeline as data travels from its source to its final destination and the transformations that occurs within. Review Hadoop ecosystem tools and how Apache Spark has changed the landscape with the advent of in-memory processing.	
	Unit 1. Big Data Processing • Big Data use case for industry • Public Cloud Introduction (Amazon AWS, Microsoft Azure) • Instructor Demo	2H (Demo 0.5)

	Unit 2. Hadoop Core & Eco system overview • Apache Hadoop & Spark platform overview • Data ingest / storage / processing / BI introduction	2H
	Unit 3. Hadoop Architecture for Big Data • HDFS Storage • YARN resource manager and compute architecture • Exercise - Setting up our VirtualBox environment • Exercise - HDFS and Running applications through Yarn	4H (Exercise 2H)
	Quiz (Written & Hands-on)	1H
3	**Chapter 3. Big Data Ingestion**	**26H (Total)**
	Objective: Acquire detailed knowledge in the tools available in the Hadoop Ecosystem for data ingestion. Learn how to migrate data from EDW data silos. Create data pipelines for capturing streaming real-time data. Learn how NiFi allows user to create data pipelines that include data transformations using a browser based GUI. Finally, understand why Kafka has become such an integral tool in a big data ingestion architecture.	
	Unit 1. Data Migration from EDW • Apache Sqoop • Exercise - Sqoop	3H (Exercise 1H)
	Unit 2. Streaming Real-Time Data Sources • Apache Flume • Micro-batch based tools (Apache Spark Streaming) • Event-driven tools (Apache Flink) • Exercise - Flume	3H (Exercise 1H)
	Unit 3. Creating Data Pipelines with Apache NiFi • Processors, Connections, Dataflows and Process Groups • Exercise - NiFi	14H (Exercise 8H)
	Unit 4. Apache Kafka • Apache Kafka Fundamentals • Producers and Consumers • Exercise - Kafka on console	4H (Exercise 1H)
	Quiz (Written + Hands-on)	2H
4	**Chapter 4. Big Data Storage**	**24H (Total)**
	Objective: Understand the various options available for Big Data Storage and their pros and cons. Understand how data storage for analysis has evolved overtime from flat-file based to database to 3NF relational databases. Gain insight to why relational databases do not fit to processing massive amount of data and how NoSQL has evolved to meet these new challenges.	
	Unit 1. Data Storage Alternatives	8H

	• On-premise Storage - HDFS, HDFS EC, Kudu • Public Cloud Storage - S3, Glacier, BLOB • Exercise - AWS S3 and Glacier	(Exercise 4H)
	Unit 2. NoSQL • NoSQL Basics • Apache HBase • Apache Cassandra • MongoDB • Exercise - HBase and Cassandra	14H (Exercise 6H)
	Quiz (Written & Hands-on)	2H
5	**Chapter 5. Big Data Analytics**	**32H (Total)**
	Objective: Understand how analyzing massive amount of data differ from typical relational databases. Understand how new ecosystem tools based on Map Reduce evolved to meet these challenges. Learn to use SQL based queries and scripts to create, modify and analyze big data tables. Gain insight to how data can be organized to increase performance of the operations. Understand the use cases of currently available big data analytic tools from latency sensitive interactive ad-hoc queries to working with petabytes of data on thruput enhanced tools. Learn how to work with batch data and streaming data to analyze and compare historical data with current streaming data.	
	Unit 1. Introduction to SQL • Creating tables using DDL • Modify table properties with DML • Query data with SQL • Exercise - SQL Basics	8H (Exercise 4H)
	Unit 2. Basic Analytics • Data preprocessing and Basic data analysis with Apache Pig • Basic Querying with Apache Hive and Impala • Exercise - HiveQL basics	8H (Exercise 4H)
	Unit 3. Advanced Analytics • Hive and Impala Data Management for Data storage and Performance • Complex Data and Relational Data Analysis with Hive and Impala • Exercise - HiveQL advanced and Impala basics	8H (Exercise 4H)
	Unit 4. Streaming Data Analysis Architecture • Lambda Architecture • Kappa Architecture • Exercise - Experimenting with Lambda Architecture	4H (Exercise 2H)
	Quiz (Written & Hands-on)	4H
6	**Chapter 6. Big Data Processing with Apache Spark**	**44H**

	Objective: Understand why Apache Spark has become one of the most popular tools in big data. Learn to work with the various APIs in Spark and the different use cases for each. Use Apache Spark Core API to analyze and transform unstructured and semi-structured data. Use Dataframe API to work with structured data and incorporate the catalyst optimizer for maximum performance. Use DStream API for unstructured/semi-structured streaming data and Structured Streaming API for structured streaming data. Understand the Spark architecture and learn the basic steps for application optimization.	**(Total)**
	Unit 1. Unstructured Data Processing • Introduction to Apache Spark • Python Basics • Data Transformation with Core API • Working with Pair RDDs • Exercise - Spark Core API	15H (Exercise 7H)
	Unit 2. Structured Data Processing • Introduction to Spark SQL • Spark SQL Operations • Interoperation RDDs and DataFrames • Exercise - Spark Dataframe API	13H (Exercise 7H)
	Unit 3. Processing Streaming Data • Introduction to Spark Streaming • Working with Unstructured Streaming Data • Working with Structured Streaming Data • Exercise - Spark DStream and Structured streaming API	7H (Exercise 3H)
	Unit 4. Apache Spark Applications • Spark Application • Distributed Processing • Persistence • Exercise - Apache Spark applications	5H (Exercise 1H)
	Quiz (Written & Hands-on)	4H
7	**Chapter 7. Big Data Modeling and AI**	**14H (Total)**
	Objective: Understand how to use tools learned in previous chapters to preprocess and transform datasets in preparation for machine learning. Learn how to create Apache Spark ML/MLlib Transformers, Estimators and Pipelines and utilize standardized machine learning algorithms provided within the API to create data models. Gain insight to how classification is performed using regression based algorithms. Understand the basic algorithms behind collaborative filtering and use them to create models.	
	Unit 1. Machine Learning Models	9H

	• Machine Learning Basic • Machine Learning on Public Cloud • Apache Spark ML • Exercise - Machine Learning on Public Cloud	(Exercise 2H)
	Unit 2. Apache Spark Machine Learning Library (MLlib) • Creating Spark MLlib Pipelines • Classification and Regression • Clustering and Collaborative Filtering • Exercise - Spark ML Models	4H (Exercise 2H)
	Quiz (Written & Hands-on)	1H
8	**Chapter 8. Data Visualization**	**4H (Total)**
	Objective: Understand the importance of good data visualization in clearly communicating the result of a data analysis. Learn to use Notebook based tools to combine code, presentation and visualizations. Gain insight to current popular data visualization tools.	
	Unit 1. Open Source Visualization Tools • Data Visualization Basics • Introduction to Apache Hue and Jupyter • Exercise – Simple BI using Hue	2H (Exercise 1H)
	Unit 2. Commercial Visualization Tool • Introduction to Power BI • Exercise – Power BI	1H (Exercise 0.5H)
	Quiz (Written & Hands-on)	1H
9	**Chapter 9. Security and Access Control**	**3H (Total)**
	Objective: Understand the basics of security including the concepts of authentication, authorization, and encryption. Learn to use Kerberos to control authentication and Apache Ranger to control access to objects. Understand how to use Apache Atlas with Ranger to create tags, associate objects with tags and control low-level access with them.	
	Unit 1. Secure Access to Cluster Services • Kerberize a Hadoop Custer	1H
	Unit 2. Secure Access to Cluster Data • ACL with Apache Ranger and Atlas	1H
	Quiz (Written & Hands-on)	1H