

Big Data Programming

1. Instructor

Hae Joon Lee (philip.lee@kiwiplus.io)

2. Course Summary

- These days, data analysis and engineering are gradually becoming popular in the computer field. For these, basic knowledge of big data analytical methodologies has to be proceeded.
- A course covers the analytical methodologies, which is fundamental to big data engineers or scientists. You will gain an understanding of what insights big data could offer through hands-on experience with representative tools.
- This course would be for the third or fourth year students.

3. Qualifications

- Java Basic Programming Skill
- (not required) Python Basic Skill

4. Textbooks

- **Main textbook 1** : Tom White, Hadoop: The Definitive Guide, 4th edition, O'Reilly.
- **Main textbook 2** : Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia, Learning Spark: Lightning-Fast Data Analytics, O'Reilly, 2015.
- **Auxiliary textbook 3** : Donald Miner & Adam Shook, MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems

5. Grading Policy

- Individual Project : MapReduce Algorithm Implementation
- Group Project : Paper Presentation
- Midterm and Final-term

6. Tentative Schedule

Week	Main Contents	Reference
1	Introduction to Big Data	Tutorials
2	Overview of Hadoop, MapReduce	Chap.1,2,3 of textbook 1
3	Hadoop Basics	Chap.1,2,3 of textbook 1
4	Hadoop Basics & Yarn	Chap.2,3,4 of textbook 1
5	Hadoop I/O	Chap.4,5 of textbook 1
6	MapReduce Applications	Textbook 1 & Tutorial
7	MapReduce Workflow & Algorithms	Textbook 1 & Tutorial
8	Midterm	

9	MapReduce Workflow & Algorithms	Textbook 1 & Tutorial
10	Introduction Spark	Chap.1,2 of textbook 2
11	Spark Basics	Chap.3,4 of textbook 2
12	Advanced Spark 1	Chap.5,6 of textbook 2
13	Advanced Spark 2	Chap.7,8 of textbook 2
14	Spark Configuration and Spark SQL	Chap.9,10 of textbook 2
15	Flink, Spark Competitor & Zeppelin	Tutorials
16	Finalterm	

Related Papers

week 1

- Piranha : Optimizing Short Jobs in Hadoop, Elmeleegy K
- Robert H Bonczek, Clyde W Holsapple, and Andrew B Whinston. Foundations of decision support systems. Academic Press, 2014.
- Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D Ernst. Haloop: efficient iterative data processing on large clusters. Proceedings of the VLDB Endowment,

week 2

- An Experimental Comparison of Pregel-like, Systems G Han M Daudjee K Ammar KOzsu M Wang X Jin T
- Twister : A Runtime for Iterative MapReduce, Ekanayake J Li H Zhang B Gunarathne TBae S Qiu J Fox G
- Piranha : Optimizing Short Jobs in Hadoop, Elmeleegy K
- Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D Ernst. Haloop: efficient iterative data processing on large clusters. Proceedings of the VLDB Endowment,

week 3 - 6

- The Hadoop Distributed File System, Shvachko K Kuang H Radia S Chansler
- MapReduce : Simplified Data Processing on Large Clusters, Dean J Ghemawat S
- Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. Communications of the aCM, 51(1):107–113, 2008.
- Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB

week 7 - 9

- MapReduce Online, Condie T Conway N Alvaro P Hellerstein JEIemeleegy K Sears R
- PACMan: Coordinated memory caching for parallel jobs, Ananthanarayanan G Ghodsi A Wang A
- Jaliya Ekanayake, Hui Li, Bingjing Zhang, Thilina Gunarathne, Seung- Hee Bae, Judy Qiu, and Geoffrey Fox. Twister: a runtime for iterative mapreduce. In Proceedings of the 19th aCM International Symposium

week 10

- Hive-a petabyte scale data warehouse using hadoop

- Hive: a warehousing solution over a map-reduce framework

week 11

- Twister : A Runtime for Iterative MapReduce, Ekanayake J Li H Zhang B Gunarathne TBae S Qiu J Fox G
- MapReduce Online, Condie T Conway N Alvaro P Hellerstein JElmeleegy K Sears R
- Resilient Distributed Datasets : A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Zaharia M Chowdhury M Das T Dave A Ma JMccauley M
- Flink Forward conference in Berlin. Flink vs spark slideshare. [http://www.slideshare.net/sbaltagi/flink-vs-spark? related=2](http://www.slideshare.net/sbaltagi/flink-vs-spark?related=2).

week 12

- Resilient Distributed Datasets : A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Zaharia M Chowdhury M Das T Dave A Ma JMccauley M Franklin M
- Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. HotCloud, 10:10–10, 2010.

week 13

- Peter Bailis. Large-scale dataflow engines, spark-sql. [http://www. redbook.io/all-chapters.html](http://www.redbook.io/all-chapters.html)
- avrilia Floratou, Umar Farooq Minhas, and Fatma O'zcan. Sql-on- hadoop: Full circle back to shared-nothing database architectures. Proceedings of the VLDB Endowment, 7(12):1295–1306, 2014.

week 14

- An Experimental Comparison of Pregel-like, Systems G Han M Daudjee K Ammar KOzsu M Wang X Jin T
- distributed machine learning and graph processing with sparse matrices. In Proceedings of the 8th ACM European Conference on Computer Systems

week 15

- Streaming Data Analysis using Apache Cassandra and Zeppelin
- Analysis of Hadoop performance and unstructured data using Zeppelin