CS 6350 Big Data Analytics and Management
(Graduate Level)
Spring 2019

People:
Instructor: Dr. Latifur Khan
Office: ECSS (ES) 3.228
Phone: (972) 883 4137
E-mail: lkhan@utdallas.edu
Office Hours: Wednesday: 1.00 p.m. to 2.00 p.m. or via Email

Course Info:
CS 6350.001
24102 Big Data Management and Analytics
Monday & Wednesday
11:30am - 12:45pm
ECSW 1.315

Teaching Assistant (TA):
TBD
Office Hour: TBD
Office Location: TBD

Course Summary

Popular relational database systems like IBM DB2, Microsoft SQL Server, Oracle, and Sybase are struggling to handle massive scale of data introduced by the Web, Social network and cyber physical systems/Internet of Things (IOT) devices. Now-a-days, companies have to deal with extremely large datasets. For example, on one hand, Facebook handles 15 TeraBytes of data each day into their 2.5 PetaByte Hadoop-powered data warehouse and on the other hand, eBay maintains a 6.5 PetaByte data warehouse. To handle emerging data at massive scale, "big data analytics” and “big data management” areas are emerging. Many traditional assumptions are not working, instead, new query and programming interfaces are required, and new computing models are emerging.
The course will focus on data mining and machine learning algorithms for analyzing very large amounts of data or Big data. Map Reduce and NoSQL system will be used as tools/standards for creating parallel algorithms that can process very large amounts of data.

The course material will be drawn from textbooks as well as recent research literature. The following topics will be covered this year: Hadoop, Mapreduce, NoSQL systems (Spark, Cassandra, Pig, BigTable), Large scale supervised machine learning, Data streams, Clustering, advanced machine learning and Applications including recommendation systems, Web and security.

Requirements

Two pop up quiz, two Exams, a couple of assignments (preferably 4) and a group project. Your course grade will be based on 10% of the quiz, 30% of the exams (Exam I 10% & Exam II 10%), 48% of assignments, and 12% of the project. The project includes implementation and demonstration.

<table>
<thead>
<tr>
<th>Class Learning Outcomes</th>
<th>Number of Students</th>
<th>Material Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability to understand of conceptual, logical and physical organization of big data</td>
<td></td>
<td>HW 1, 2, 3, 4 EXAMs</td>
</tr>
<tr>
<td>Ability to understand of large data processing using Map-Reduce</td>
<td></td>
<td>HW 1, 2, 3, 4 EXAMs</td>
</tr>
<tr>
<td>Ability to understand of NoSQL models, theory and practices</td>
<td></td>
<td>HW 1.2 3, 4, EXAMs</td>
</tr>
<tr>
<td>Ability to understand of data modeling, indexing, query processing for big data</td>
<td></td>
<td>HW 1, 2, 3, 4, EXAMs</td>
</tr>
<tr>
<td>Ability to understand of recommendation methods for big data</td>
<td></td>
<td>HW 4, EXAMs</td>
</tr>
<tr>
<td>Ability to understand of unsupervised learning for big data</td>
<td></td>
<td>HW 4, EXAMs</td>
</tr>
<tr>
<td>Ability to Understand of supervised learning for big data</td>
<td></td>
<td>HW 4, EXAMs</td>
</tr>
<tr>
<td>Ability to communicate and work on team software project</td>
<td></td>
<td>Project</td>
</tr>
</tbody>
</table>

Perquisite:

Database Management Systems, JAVA (intermediate/advanced), Linux OS, Scala (Preferable)
Course Materials

The following textbook can be used this semester to augment the material presented in lectures:

  

- B3: Anand Rajaraman and Jeff Ullman, Mining of Massive Datasets, Cambridge Press. [Mandatory]


Papers Related to Big Data Analytics and Management


- P6: Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints, IEEE Transactions on Knowledge & Data Engineering (TKDE), 2011, IEEE Computer Society, June 2011, Vol. 23, No. 6, Page 859-874.
- P12: Ahsanul Haque*, Latifur Khan, Michael Baron and Charu Aggarwal. Efficient Semi-Supervised Adaptive Classification and Novel Class Detection over Data Stream, 32nd IEEE International Conference on Data Engineering, May 16-20, 2016 · Helsinki, Finland

Recommendation System:

  
  http://doi.ieeecomputersociety.org/10.1109/MC.2009.263

Software

Mahout: http://mahout.apache.org/
Piglatin: http://pig.apache.org/docs/r0.7.0/tutorial.html
Hadoop: http://hadoop.apache.org/
Cassandra: http://cassandra.apache.org/
Kafka: https://cwiki.apache.org/confluence/display/KAFKA/Kafka+papers+and+presentations
Storm: http://storm-project.net/
Spark https://spark.apache.org/
Flink http://flink.apache.org/
## Lectures

<table>
<thead>
<tr>
<th>Topic</th>
<th>Chapters/Papers</th>
<th>Homework/Lecture Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop+ Mapreduce</td>
<td>Chapter 1, 2, 3 [B1], [B3]. Paper: P1, P2, P3, P4</td>
<td>[Hadoop with Mapreduce]</td>
</tr>
<tr>
<td>Big Data Management: Spark, Spark SQL, Hive, NoSQL Pig Latin, Cassandra, BigTable, HBase</td>
<td>Paper: P5, P7, P8, P9, P10</td>
<td></td>
</tr>
<tr>
<td>Stream Data Management: Spark Stream</td>
<td>Paper: P11</td>
<td></td>
</tr>
<tr>
<td>Clustering Analysis</td>
<td>Chapter 8 [B2]</td>
<td>[Lecture Note in Clustering]</td>
</tr>
<tr>
<td></td>
<td>Chapter 9 [B2]</td>
<td></td>
</tr>
<tr>
<td>Classification, Prediction, Stream Mining</td>
<td>Chapter 5 [B2]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chapter 4 [B3]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paper: P6, P12</td>
<td></td>
</tr>
<tr>
<td>Recommendation System</td>
<td>Chapter 9 [B3], P13</td>
<td></td>
</tr>
<tr>
<td>Advanced Analytics</td>
<td>Papers TBD</td>
<td></td>
</tr>
<tr>
<td>Exam I, II</td>
<td>TBD</td>
<td></td>
</tr>
<tr>
<td>Project Demonstration</td>
<td>TBD (May)</td>
<td></td>
</tr>
</tbody>
</table>
Assignment (Tentative)

Assignments will be based on:

1. Map Reduce Programming—Basic & Hands on HDFS setup & Advanced Map Reduce Programming

2. Text Data Processing and Analytics using Spark

3. Big Stream Text Data Management Using Spark SQL, Hive, Data Frame

4. Problem Solving Questions/Exercise Problems from Books; and Large Scale Recommendation Using Spark, Spark and Mlib

Projects (Sample—Expanded further)

- Text Mining with LDA
- Geo Location Mining
- Smart Phone Apps Fingerprinting
- Smart Phone Data Analytics
- NLP Processing for big Web Data
- Cyber Data Analysis
- Secure Data Analytics wit Trusted Execution Environment