The seminar deals with recent developments in database systems and big data systems. The focus will be on Graph databases, Document Stores, Systems for Large Scale Distributed Processing, Properties of Distributed Systems (theory), High Availability Distributed Processing and Integrated Cluster Computing Framework.
LECTURE TOPICS (the CS library holds the books)


In all these 4 presentations you need to demo in python, Java or Scala.

1. Chapters 1-3 (Introduction, Installation, RDDs)
2. Chapters 4-5 (Key-Value pairs, Loading/Storing data)
3. Chapter 6 (Advanced Programming)
4. Chapter 9 (Spark SQL), and the following paper:
   Spark SQL: Relational Data Processing in Spark
   Michael Armbrust, Reynold Xin, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael Franklin, Ali Ghodsi, Matei Zaharia
   ACM SIGMOD Conference 2015, May. 2015
5. Chapter 11 (Machine Learning with Spark), include Demo
6. Chapter 7 (Running on a Cluster), include Demo
   GraphX: Graph Processing in a Distributed Dataflow, Framework
   Joseph E. Gonzalez, Reynold S. Xin, Ankur , Daniel Crankshaw, Michael J. Franklin, Ion Stoica

B. Graph Databases
System G papers on Graph Database

- **SYSTEM G on top of Hadoop**


- **RDF Support in DB2**


- **GBASE papers**


**Neo4j**

- **Neo4j (Graph Database)**

https://en.wikipedia.org/wiki/Neo4j

http://neo4j.com/docs/1.8.3/index.html

A starting point, you’ll need to gather information, extra credit for demonstrating it live.

**C. Document Store**

**Mongo DB (JSON-like documents)**

https://en.wikipedia.org/wiki/MongoDB
D. Properties of Distributed Systems

**CAP Theorem (Consistency, Availability, Partition Tolerance)**, team of 2

1. **Introduction**
   Eric A. Brewer, PODC '00 Proceedings of the nineteenth annual ACM symposium on Principles of distributed computing, Page 7.


CAP Twelve Years Later: How the "Rules" Have Changed
Posted by Eric Brewer: [http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed](http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed)

2. **Proof of the conjecture**


   [http://markburgess.org/blog_cap.html](http://markburgess.org/blog_cap.html) - interesting reading
E. High Availability Distributed Processing

**Cassandra (Apache)**
Source: Cassandra High Availability, book, Robbie Strickland

1. Chapters 1-3 (Approach, Data Distribution, Replication)
2. Chapters 4-5 (Data Centers, Scaling)
3. Chapter 7 (Storage Layer, Query Execution)

F. Key-value Store

**Riak**

Riak is a distributed NoSQL key-value data store that offers high availability, fault tolerance, operational simplicity, and scalability. In addition to the open-source version, it comes in a supported enterprise version and a cloud storage version that is ideal for cloud computing environments.

G. Large Scale Distributed Processing: Apache Hadoop
Source: Hadoop: The Definitive Guide, Tom White

This is a framework that allows for the distributed processing of large data sets across clusters of computers. It uses a simple programming model. It is designed to scale up from single servers to thousands of machines. The library itself is designed to detect and handle failures at the application layer.

1. Chapter 1-2, Overall introduction
2. Chapter 3 – HDFS
3. Chapters 4-5 – Yarn, I/O
4. Chapters 6-9, MapReduce in depth, team of 2, also

MapReduce: A major step backwards by David J. DeWitt and Michael Stonebraker

Databases are hammers; MapReduce is a screwdriver by Mark C. Chu-Carroll.
HBase (Apache)
An open source, non-relational, distributed database modeled after Google's BigTable and written in Java.
1. Chapter 20, Hadoop book

H. Application level concurrency control

Feral Concurrency Control: An Empirical Investigation of Modern Application Integrity

Peter Bailis, Alan Fekete, Michael Franklin, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica
ACM SIGMOD Conference 2015, May. 2015.

The following is a draft schedule, there will likely be changes!
<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Date</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Course Overview, Introduction to Spark</strong></td>
<td>21/10/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td>2</td>
<td>a. Spark, Chapters 1-3 (Introduction, Installation, RDDs)</td>
<td>28/10/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. Spark, Chapters 4-5 (Key-Value pairs, Loading/Storing data)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>a. GPUs and graph queries/Cloud Computing</td>
<td>4/11/15</td>
<td>Avraham</td>
</tr>
<tr>
<td></td>
<td>b. Review of key concepts in DB concurrency control and recovery/Cloud Computing</td>
<td></td>
<td>Marzeh</td>
</tr>
<tr>
<td>4</td>
<td>a. Neo4j or Mongo DB</td>
<td>11/11/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. Neo4j or Mongo DB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>a. Spark Chapter 6, include DEMO</td>
<td>18/11/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. Spark Chapter 9 (Spark SQL), and Spark SQL: Relational Data Processing in Spark</td>
<td></td>
<td>Marzeh</td>
</tr>
<tr>
<td>6</td>
<td>a. Spark, GraphX</td>
<td>25/11/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. System G paper/Spark Chapter 7 or 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>a. System G paper/Spark Chapter 7 or 11</td>
<td>2/12/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. System G paper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td><strong>CAP theorem: Team of 2</strong></td>
<td>16/12/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td>9</td>
<td>Cassandra (Chapters 1-5) – Team of 2</td>
<td>23/12/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td>10</td>
<td>a. Cassandra Chapter 7</td>
<td>30/12/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. Riak</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>a. Hadoop, Chapter 1-3, Overall introduction, HDFS</td>
<td>6/1/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. Hadoop, Chapters 6-9, MapReduce in depth-Team of 2, include DEMO</td>
<td></td>
<td>Marzeh</td>
</tr>
<tr>
<td>12</td>
<td>a. Hadoop, Chapters 4-5 – Yarn, I/O</td>
<td>13/1/15</td>
<td>Marzeh</td>
</tr>
<tr>
<td></td>
<td>b. HBase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td><strong>Reserved for spillovers</strong></td>
<td>20/1/15</td>
<td>Marzeh</td>
</tr>
</tbody>
</table>
Hadoop: The Definitive Guide
by Tom White

Learning Spark: Lightning-Fast Big Data Analysis
Karau, Holden

Learn Python in One Day and Learn It Well: Python for Beginners with Hands-on Project. The only book you need to start coding in Python immediately
Chan, Jamie

Retrieving tracking information…
Cassandra High Availability
Strickland, Robbie

JSON: Learn JSON In A DAY! - The Ultimate Crash Course to Learning the Basics of JSON In No Time (JSON, JSON Course...