CS6545/CS4545 – Big Data Systems
Winter 2018 - Course Outline

Course Description:

Data systems are going through a major transition due to the challenges of Big Data processing. The outcome of this shift is the emergence of a new breed of systems that can handle data at massive scales. This course presents some of these systems, along with the principles of query processing, through a combination of lectures and research paper reviews. Specifically, it compares Relational vs. NoSQL data models and covers the foundations of query processing, including index-based access and join processing. It presents the principles of parallel databases, and explores batch processing frameworks, as well as iterative processing frameworks. It also covers SQL interfaces over these frameworks. It introduces update-intensive systems and graph data stores. It includes the special topics including time-series.

Note: Hands-on sessions provide students the opportunity to use several Big Data Systems such as Hadoop, Hive, Stado, HBase, Spark etc.

Topics:

1. **Foundations of access methods and query processing**
   a. Data models: relational vs. NOSQL
   b. Different indexing techniques
   c. Query processing overview
   d. Join processing

2. **Parallel database**
   a. Parallel algorithms
   b. Partitioning
   c. Stado (a parallel database)

3. **Batch processing frameworks and SQL interface on MapReduce**
   a. MapReduce, Hadoop, Hive

4. **Iterative processing frameworks**
   a. Spark, SparkSQL

5. **Update-intensive data processing frameworks**
   a. Transaction
   b. Consistency models
   c. Cassandra, HBase

6. **Graph data processing**
   a. Neo4J

7. **In-Memory Data Management**

   a. Compression
   b. Storage layout
   c. Update handling

8. **Special topics in Big Data**

   a. Time-series database (InfluxDB)
   b. Systems for data analytics and machines learning