PAF KIET

PAF-Karachi Institute of Economics & Technology

Big Data and Analytics (Fall 2015)

Core/Elective: MS CS Elective – MS SPM Elective

Instructor: Dr. Tariq MAHMOOD

Credit Hours: 3

Pre-requisite:

- **All Core CS Courses (Knowledge of Data Mining is a Plus)**
- **Every student must have his/her own laptop for the experiments**
- **Those students who want to avoid implementation and coding shouldn't take this course.**

Course Website: https://sites.google.com/site/bigdatabolt/courses/bdaf15

Why is it Important to Offer this Course?

- Big Data is the hottest buzz word of the global industry,
- The requirement for analytics of Big Data is on a very steep rise, and has applications are continuing to rise
- In its October 2012 issues, Harvard Business Review stressed the need for managing Big Data in its frontline article.
- Many companies are analyzing their big data to predict their future trends (Predictive Analytics).
- MS level courses on Big Data being formalized extensively in USA and other countries, e.g., MS in Predictive Analytics offered by Northwestern University.

What is the Contribution of this Course?

- First course of its type which will deal with state of the art big data and analytics technologies, primarily, document databases, columnar stores, Hadoop infrastructure and Apache Spark

- Each technology will be accompanied with rigorous experiments on different big and small data sets

- Imparting of extremely useful knowledge that is spread out across many different types of books, websites, articles, case studies, research papers etc.

Course Description:

This course will focus on more commonly applied technologies for storing, processing, managing and analyzing big data. These are: document databases (MongoDB, CouchDB), columnar stores (Cassandra, MonetDB) and Apache Spark on Apache Hadoop infrastructure. The following questions answer basic queries about the course:

What is Big Data?: Big Data refers to a collection of large and complex data sets that become difficult to process using traditional database management tools or data processing applications. The challenges include the collection, storage, search, sharing, analysis and visualization of this data.

Why is Big Data being Generated? The sources of data collection have become increased and also more sophisticated, e.g., ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks.

Some Examples of Big Data: Meteorological, genomics, physics simulations, biological research, astronomical, geological, internet search, banking, E-Commerce etc.

Why is Analyzing Big Data Important? These days analyzing data has become a necessity. Analytics of Big Data supports Business Intelligence by providing insights into business processes which allows managers to evaluate current strategies and design new ones if needed. This is having a very large positive impact in very important domains, e.g., national security, geographical analysis, clinical healthcare etc.

What are Document Databases?: These allow big data to be stored as documents (XML, JSON, BSON etc.) rather than in traditional tables. More common examples are MongoDB, CouchDB, and RavenDB.

What are Columnar Stores? These store big data in huge tables which are called columnar stores. These tables have flexible column and row sizes along with helpful meta-data information for retrieval and storage.

What is Apache Hadoop? Apache Hadoop is the well-known infrastructure for managing big data. Starting from a small Apache open-source project, Hadoop is now a more sophisticated and widely applied database in the corporate industry.

What is Apache Spark? Apache Spark is the latest, state of the art in-memory data processing framework which has shown considerable performance improvement as compared to Hadoop’s processing technology. Coupled with Hadoop infra, Apache Spark is seeing remarkable applications in the industry.

Specific Outcomes – What will the Students Learn?

Upon successful completion of this course, the students should be able to:

- Understand the impact of Big Data and Analytics in running organizations effectively and efficiently.
- Understand and implement the Hadoop and MapReduce distributed system frameworks, which are crucial components of Big Data
- Understand and implement document databases

- Understand and implement columnar stores
- Configure and use Hadoop clusters
- Configure and use Apache Spark
- Understand and implement all technologies and methods related to ETL (Extract, Transform, Load) and data pre-processing
- Understand and implement data mining techniques in the context of Big Data, e.g., large-scale association rule mining, large-scale regression and large-scale clustering
- Implement a project that employs several open-source tools and APIs related to Big Data and Analytics

Employability Skills:

This course will also teach the students one or more of the following employability skills:

- Apply a systematic approach to solve problems
- Use a variety of thinking skills to anticipate and solve problems
- Locate, select, organize and document information using appropriate technology and information systems
- Analyze, evaluate, and apply relevant information from a variety of sources
- Interact with others in group or teams in ways that contribute to effective working relationships and the achievement of goals
- Manage the use of time and other resources to complete projects
- Take responsibility for one's own actions, decisions, and consequences

Weekly Plan:

Week	List of Topics
Week 1 and 2	**<u>An Introduction to Big Data and Analytics</u>** • BI Life Cycle

	• Data Mining • Big Data • NoSQL Movement • Ecosystems and Evolution • Applications, Technologies, Tools, Implementations • Progress, Need and the Future • Research Directions
Week 3	**MapReduce and Hadoop** • Distributed File systems • MapReduce explained • Algorithms using MapReduce • Extensions to MapReduce • Hadoop API explained • Hadoop and MapReduce explained
Week 4	**Simulations with Hadoop Framework**
Week 5 6 7	**Document Databases** • MongoDB (Methodology and Experiments)
Week 8 9 10	**Columnar Stores** • MonetDB and Cassandra (Methodology and Experiments)
Week 11 12 13	**Apache Spark with Hadoop** • Framework, Methodology • Experiments

Textbooks:

- Mining Massive Datasets, by Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman – Ebook provided
- Big Data: Principles and best practices of scalable realtime data systems, by Nathan Marz and James Warren, January 2012

- Hadoop the definitive guide, by Tom White, OReilly, 2009, 3rd Edition – Ebook provided.

Grading

Assignments	8 %
Quizzes	8 %
Project	17 %
Hourly (only one exam)	17 %
Final	50 %

Project:

The project is the most important part of the course. Implementation is essential (no marks to be assigned for simple literature reviews). Potential topics and datasets will be discussed and given.