Elective : Big Data Technologies

Evaluation:

	Theory	Practical	Total
Sessional	30	20	50
Final	50	-	50
Total	80	20	100

Course Description:
The growth of information systems has given rise to large amount of data which do not qualify as traditional definition of data. This scenario has given us new possibilities but at same time pose serious challenges. Such challenges lies in effective storage, analysis and search of such large set of data. Fortunately, a number of technologies have been developed that answer such challenges. This course introduces this scenario along with technologies and how they answer these challenges.

Objective of the Course:
To introduce student to current scenarios of big data and provide various facets of big data. It also provides them with technologies playing key role in it and equips them with necessary knowledge to use them for solving various big data problems in different domains.

Course Contents

1 Introduction to Big Data (8 Hours)
- 1.1 Big Data Overview
- 1.2 Background of Data Analytics
- 1.3 Role of Distributed System in Big Data
- 1.4 Role of Data Scientist
- 1.5 Current Trend in Big Data Analytics

2 Google File System (7 Hours)
- 2.1 Architecture
- 2.2 Availability
- 2.3 Fault tolerance
- 2.4 Optimization for large scale data

3 Map-Reduce Framework (10 Hours)
- 3.1 Basics of functional programming
  - 3.1.1 Fundamentals of functional programming
  - 3.1.2 Real world problems modeling in functional style
- 3.2 Map reduce fundamentals
- 3.3 Data flow (Architecture)
- 3.4 Real world problems
- 3.5 Scalability goal
- 3.6 Fault tolerance
- 3.7 Optimization and data locality
- 3.8 Parallel Efficiency of Map-Reduce

4 NoSQL (6 Hours)

4.1 Structured and Unstructured Data
4.2 Taxonomy of NoSQL Implementation
4.3 Discussion of basic architecture of Hbase, Cassandra and MongoDb

5 Searching and Indexing of Big Data (7 Hours)

5.1 Full text Indexing and Searching
5.2 Indexing with Lucene
5.3 Distributed Searching with elastic search
5.4

6 Case Study: Hadoop (7 Hours)

6.1 Introduction to Hadoop Environment
6.2 Data Flow
6.3 Hadoop I/O
6.4 Query languages for Hadoop
6.5 Hadoop and Amazon Cloud

Practical

Student will get opportunity to work in big data technologies using various dummy as well as real world problems that will cover all the aspects discussed in course. It will help them gain practical insights in knowing about problems faced and how to tackle them using knowledge of tools learned in course.

1. HDFS: Setup a hdfs in a single node to multi node cluster, perform basic file system operation on it using commands provided, monitor cluster performance
2. Map-Reduce: Write various MR programs dealing with different aspects of it as studied in course
3. Hbase: Setup of Hbase in single node and distributed mode, write program to write into hbase and query it
4. Elastic Search: Setup elastic search in single mode and distributed mode, Define template, Write data in it and finally query it
5. Final Assignment: A final assignment covering all aspect studied in order to demonstrate problem solving capability of students in big data scenario.

References

1. Jeffrey Dean, Sanjay Ghemawat, MapReduce:Simplified Data Processing on Large Clusters
2. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, The Google File System
3. Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber, Bigtable: A Distributed Storage System for Structured Data
4. http://hadoop.apache.org/
5. http://hbase.apache.org/
6. http://www.elasticsearch.org/guide/
7. Tom White, Hadoop: The Definitive Guide
8. Lars George, Hbase: The Definitive Guide
9. Jason Rutherglen, Ryan Tabora, Jack Krupansky, Lucene and Solr: The Definitive Guide