BIG DATA TECHNOLOGIES

CT 765 07

Lecture : 3 **Year : IV**
Tutorial : 1 **Part : II**
Practical : 3/2

Course Objectives:

To introduce the current scenarios of big data and provide various facets of big data and to be familiar with the technologies playing key role in it and equips them with necessary knowledge to use them for solving various big data problems in different domains.

1. Introduction to Big Data **(7 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
3.2 Fundamentals of functional programming
3.3 Real world problems modeling in functional style
3.4 Map reduce fundamentals
3.5 Data flow (Architecture)
3.6 Real world problems
3.7 Scalability goal
3.8 Fault tolerance
3.9 Optimization and data locality
3.10 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 Big Data **(7 hours)**

5.1 Full text Indexing and Searching
5.2 Indexing with Lucene
5.3 Distributed Searching with elasticsearch

6. Case Study: Hadoop **(8 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, Map Reduce, "Simplified Data Processing on Large Clusters"
2. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google File System"
3. http://wiki.apache.org/hadoop/