Large Scale Data Processing
CSE3025

Dr. Ramesh Ragala

School of Computer Science and Engineering
VIT Chennai

February 2, 2021

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

1. Introduction
2. UNIT - I
3. UNIT - II
4. UNIT - III
5. UNIT - IV
6. UNIT - V
7. UNIT - VI
8. UNIT - VII
9. UNIT - VIII

Course Objective:

- Understand different characteristics of big data
- Understand the requirement of big data frameworks
- Learn the concepts of distributed file system
- Provide MapReduce programming environment
- Understand need of inverted indexing and graph data analytics

Expected Course Outcome:

- Define the characteristics of big data and explain the data science life cycle
- Differentiate between conventional and contemporary distributed framework
- Characterize storage and processing of large data
- Implement and demonstrate the use of the Hadoop eco-system
- Compare scalable frameworks for large data
- Identify independent tasks in a program that may be parallelized
- Decompose a problem into map and reduce operations for implementation
- Recognize different input output formats for map reduce programs
- Design programs to analyze large scale text data
- Identify problems suitable for use of graph mining in large data processing

1. Introduction
2. UNIT - I
3. UNIT - II
4. UNIT - III
5. UNIT - IV
6. UNIT - V
7. UNIT - VI
8. UNIT - VII
9. UNIT - VIII

Introduction to Big Data and Analytics

- Big Data Overview
- Characteristics of Big Data
- Business Intelligence vs Data Analytics

1 Introduction

2 UNIT - I

3 UNIT - II

4 UNIT - III

5 UNIT - IV

6 UNIT - V

7 UNIT - VI

8 UNIT - VII

9 UNIT - VIII

Need of Data Analytics

- Data Analytics Life Cycle
- Data Analytics in Industries
- Exploring Big Data
- Challenges in handling Big Data

1. Introduction
2. UNIT - I
3. UNIT - II
4. UNIT - III
5. UNIT - IV
6. UNIT - V
7. UNIT - VI
8. UNIT - VII
9. UNIT - VIII

Big Data Tools

- Need of Big Data Tools
- Understanding Distributed System
- Overview of Hadoop
- Comparing SQL databases and Hadoop
- Hadoop Eco System
- HDFS: Distributed File System
- Design of HDFS
- Writing Files to HDFS
- Reading Files from HDFS

1. Introduction
2. UNIT - I
3. UNIT - II
4. UNIT - III
5. **UNIT - IV**
6. UNIT - V
7. UNIT - VI
8. UNIT - VII
9. UNIT - VIII

Hadoop Architecture

- Hadoop Daemons
- Hadoop Cluster Architecture
- YARN Yet Another Resource Negotiator
- Advantages of YARN

1. Introduction
2. UNIT - I
3. UNIT - II
4. UNIT - III
5. UNIT - IV
6. **UNIT - V**
7. UNIT - VI
8. UNIT - VII
9. UNIT - VIII

Introduction to MapReduce

- Developing MapReduce Program
- Anatomy of MapReduce Code
- Simple MapReduce Code : Counting Things
- Map Phase
- Shuffle and Sorting Phase
- Reduce Phase
- Master Slave Architecture
- Job Processing in Hadoop
- MapReduce Pipelining

1. Introduction
2. UNIT - I
3. UNIT - II
4. UNIT - III
5. UNIT - IV
6. UNIT - V
7. **UNIT - VI**
8. UNIT - VII
9. UNIT - VIII

MapReduce Programming Concepts

- Use of Combiner
- Block Vs Split Size
    - Key
    - Text
    - Sequence
    - Nline File Format
    - XML File Format

1 Introduction

2 UNIT - I

3 UNIT - II

4 UNIT - III

5 UNIT - IV

6 UNIT - V

7 UNIT - VI

8 UNIT - VII

9 UNIT - VIII

Inverted Indexing and Graph Analytics

- Web Crawling
- Inverted Index
- Baseline and revised Implementation
- Graph Representation
- Parallel Breath First Search
- Page Rank
- Issues with graph Processing

1. Introduction
2. UNIT - I
3. UNIT - II
4. UNIT - III
5. UNIT - IV
6. UNIT - V
7. UNIT - VI
8. UNIT - VII
9. UNIT - VIII

Recent Trends

Guest Lecture

Guest Lecture from Industry Expert