Module - 1

Syllabus :

Introduction to Big Data, Big Data characteristics, Types of Big Data, Traditional vs. Big Data business approach, Case Study of Big Data Solutions, Concept of Hadoop, Core Hadoop Components, Hadoop Ecosystem

Chapter 1 : Introduction to Big Data 1-1 to 1-11

1.1 Introduction to Big Data Management 1-1
1.2 Big Data 1-1
1.3 Big Data Characteristics - Four Important V of Big Data 1-2
1.4 Types of Big Data 1-4
1.5 Big Data vs. Traditional Data Business Approach 1-6
1.6 Tools used for Big Data 1-9
1.7 Data Infrastructure Requirements 1-10
1.8 Case Studies of Big Data Solutions 1-10

Chapter 2 : Introduction to Hadoop 2-1 to 2-16

2.1 Hadoop 2-1
2.1.1 Hadoop - Features 2-2
2.1.2 Hadoop and Traditional RDBMS 2-2
2.2 Hadoop System Principles 2-2
2.3 Hadoop Physical Architecture 2-3
2.4 Hadoop Core Components 2-5
2.4.1 HDFS (Hadoop Distributed File System) 2-5
2.4.2 MapReduce 2-6
2.4.3 Hadoop - Limitation 2-8
2.5 Hadoop - Ecosystem 2-8
2.6 ZooKeeper 2-10
2.7 HBase 2-11
2.7.1 Comparison of HDFS and HBase 2-11
2.7.2 Comparison of RDBMS and HBase 2-11
2.7.3 HBase Architecture 2-12
2.7.4 Region Splitting Methods 2-12
2.7.5 Region Assignment and Load Balancing 2-13
2.7.6 HBase Data Model 2-13
2.8 HIVE 2-14
2.8.1 Architecture of HIVE 2-15
2.8.2 Working of HIVE 2-15
2.8.3 HIVE Data Models 2-16

Module - 2

Syllabus :

Distributed File Systems : Physical Organization of Compute Nodes, Large-Scale File-System Organization, MapReduce : The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures, Algorithms Using MapReduce : Matrix-Vector Multiplication by MapReduce, Relational-Algebra Operations, Computing Selections by MapReduce, Computing Projections by MapReduce, Union, Intersection, and Difference by MapReduce, Hadoop Limitations

Chapter 3 : Hadoop HDFS and Map Reduce 3-1 to 3-13

3.1 Distributed File Systems 3-1
3.1.1 Physical Organization of Compute Nodes 3-2
3.1.2 Large-Scale File-System Organization 3-2
3.2 MapReduce 3-3
3.2.1 The Map Tasks 3-4
3.2.2 Grouping by Key 3-4
3.2.3 The Reduce Tasks 3-4
3.2.4 Combiners 3-5
3.2.5 Details of MapReduce Execution 3-6
3.2.6 Coping with Node Failures 3-7
3.3 Algorithms using MapReduce 3-7
3.3.1 Matrix-Vector Multiplication by MapReduce 3-8
3.3.2 Relational-Algebra Operations 3-9
3.3.3 Computing Selections by MapReduce 3-10
3.3.4 Computing Projections by MapReduce 3-10
3.3.5 Union, Intersection and Difference by MapReduce 3-11
3.4 Hadoop Limitations 3-12

Module - 3

Syllabus :

Introduction to NoSQL, NoSQL Business Drivers, NoSQL Data Architecture Patterns: Key-value stores, Graph stores, Column family (Bigtable)stores, Document stores, Variations of NoSQL architectural patterns, NoSQL Case Study, NoSQL solution for big data, Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models : master-slave versus peer-to-peer; NoSQL systems to handle big data problems.

Chapter 4 : NoSQL 4-1 to 4-27

4.1 NoSQL (What is NoSQL?) 4-1
4.2 NoSQL Basic Concepts 4-2
4.3 Case Study NoSQL (SQL vs NoSQL) 4-3
4.4 Business Drivers of NoSQL 4-4
4.5 NoSQL Database Types 4-5
4.6 Benefits of NoSQL 4-9
4.7 Introduction to Big Data Management 4-9
4.8 Big Data 4-10
4.8.1 Tools Used for Big Data 4-10
4.8.2 Understanding Types of Big Data Problems 4-11
4.9 Four Ways of NoSQL to Operate Big Data Problems 4-11
4.10 Analyzing Big Data with a Shared-Nothing Architecture 4-15
4.10.1 Shared Memory System 4-15
4.10.2 Shared Disk System 4-16
4.10.3 Shared Nothing Disk System 4-17
4.10.4 Hierarchical System 4-18
4.11 Choosing Distribution Models : Master-Slave versus Peer-to-Peer 4-18
4.11.1 Big Data NoSQL Solutions 4-20
4.11.1(A) Cassendra 4-20
4.11.1(B) Dynamo DB 4-21

Module - 4

Syllabus :

The Stream Data Model: A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing, Sampling Data techniques in a Stream, Filtering Streams: Bloom Filter with Analysis, Counting Distinct Elements in a Stream, Count-Distinct Problem, Flajolet-Martin Algorithm, Combining Estimates, Space Requirements, Counting Frequent Items in a Stream, Sampling Methods for Streams, Frequent Itemsets in Decaying Windows, Counting Ones in a Window: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.

Chapter 5 : Mining Data Streams 5-1 to 5-17

5.1 The Stream Data Model 5-1
5.1.1 A Data-Stream-Management System 5-2
5.1.2 Examples of Stream Sources 5-4
5.1.3 Stream Queries 5-4
5.1.4 Issues in Stream Processing 5-5
5.2 Sampling Data Techniques in a Stream 5-6
5.3 Filtering Streams 5-7
5.3.1 Bloom Filter with Analysis 5-8
5.4 Counting Distinct Elements in a Stream 5-10
5.4.1 Count – Distinct Problem 5-10
5.4.2 The Flajolet- Martin Algorithm 5-10
5.4.3 Combining Estimates 5-12
5.4.4 Space Requirements 5-12
5.5 Counting Frequent Items in a Stream 5-12
5.5.1 Sampling Methods for Streams 5-13
5.5.2 Frequent Itemsets in Decaying Windows 5-13
5.6 Counting Ones in a Window 5-14
5.6.1 The Cost of Exact Counts 5-14
5.6.2 The DGIM Algorithm (Datar – Gionis – Indyk - Motwani) 5-14
5.6.3 Query Answering in the DGIM Algorithm 5-15
5.6.4 Decaying Windows 5-16

Module - 5

Syllabus :

Distance Measures : Definition of a Distance Measure, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance., CURE Algorithm, Stream-Computing, A Stream-Clustering, Algorithm, Initializing & Merging Buckets, Answering Queries

Chapter 6 : Finding Similar Items 6-1 to 6-8

6.1 Distance Measures....6-1
6.1.1 Definition of a Distance Measure....6-1
6.1.2 Euclidean Distances....6-2
6.1.3 Jaccard Distance....6-3
6.1.4 Cosine Distance....6-4
6.1.5 Edit Distance....6-5
6.1.6 Hamming Distance....6-7

Chapter 7 : Clustering 7-1 to 7-7

7.1 Introduction....7-1
7.2 CURE Algorithm....7-1
7.2.1 Overview of CURE (Cluster Using REpresentative)....7-2
7.2.2 Hierarchical Clustering Algorithm....7-2
7.2.2(A) Random Sampling and Partitioning Sample....7-3
7.2.2(B) Eliminate Outlier's and Data Labelling....7-3
7.3 Stream Computing....7-5
7.3.1 A Stream - Clustering Algorithm....7-5
7.4 Initializing and Merging Buckets....7-6
7.5 Answering Queries....7-6

Module - 6

Syllabus :

PageRank Overview, Efficient computation of PageRank : PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector, A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering, Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities in a social graph

Chapter 8 : Link Analysis 8-1 to 8-22

8.1 Page Rank Definition....8-1
8.1.1 Importance of Page Ranks....8-2
8.1.2 Links in Page Ranking....8-2
8.1.3 Structure of the Web....8-6
8.1.4 Using Page Rank in a Search Engine....8-12
8.2 Efficient Computation of Page Rank....8-13
8.2.1 Representation of Transition Matrix....8-14
8.2.2 Iterating Page Rank with MapReduce....8-14
8.2.3 Use of Combiners to Aggregate the Result Vector....8-15
8.3 Link Spam....8-16
8.3.1 Spam Farm Architecture....8-16
8.3.2 Spam Farm Analysis....8-17
8.3.3 Dealing with Link Spam....8-18
8.4 Hubs and Authorities....8-19
8.4.1 Formalizing Hubs and Authority....8-20

Chapter 9 : Recommendation Systems 9-1 to 9-9

9.1 Recommendation System....9-1
9.1.1 The Utility Matrix....9-1
9.1.2 Applications of Recommendation Systems....9-2
9.1.3 Taxonomy for Application Recommendation System....9-2
9.2 Content Based Recommendation....9-3
9.2.1 Item Profile....9-3
9.2.2 Discovering Features of Documents....9-3
9.2.3 Obtaining Item Features from Tags....9-4
9.2.4 Representing Item Profile....9-4
9.2.5 User Profiles....9-5
9.2.6 Recommending Items to Users based on Content....9-5
9.2.7 Classification Algorithm....9-5
9.3 Collaborative Filtering....9-6
9.3.1 Measuring Similarity....9-7
9.3.2 Jaccard Distance....9-7
9.3.3 Cosine Distance....9-7
9.3.4 Rounding the Data....9-8
9.3.5 Normalizing Rating....9-8
9.4 Pros and Cons in Recommendation System....9-8
9.4.1 Collaborative Filtering....9-8
9.4.2 Content-based Filtering....9-9

Chapter 10 : Mining Social Network Graph 10-1 to 10-14

10.1 Introduction....10-1

10.2 Social Network as Graphs 10-2

10.2.1 Parameters Used in Graph (Social Network) 10-3

10.2.2 Varieties of Social Network....10-4

10.2.2(A) Collaborative Network....10-4

10.2.2(B) Email Network 10-4

10.2.2(C) Telephone Network....10-4

10.3 Clustering of Social Network Graphs 10-4

10.3.1 Distance Measure for Social-Network Graphs 10-5

10.3.2 Applying Standard Cluster Method 10-5

10.3.3 Betweenness....10-7

10.3.4 The Girvan - Newman Algorithm....10-7

10.3.5 Using Betweenness to Find Communities 10-9

10.4 Direct Discovery of Communities....10-10

10.4.1 Bipirate Graph 10-10

10.4.2 Complete Bipirate Graph 10-10

10.5 Simrank....10-10

10.5.1 Random Walker on Social Network 10-11

10.5.2 Random Walks with Restart 10-11

10.6 Counting Triangles using MapReduce 10-12

❑❑❑