1. **Introduction to Linux and Big Data Virtual Machine (VM)**
   Introduction/ Installation of VirtualBox and the Big Data VM Introduction to Linux – Why Linux? – Windows and the Linux equivalents – Different flavors of Linux – Unity Shell (Ubuntu UI) – Basic Linux
   Commands (enough to get started with Hadoop)
2. **Understanding Big Data**
   • 3V (Volume- Variety- Velocity) characteristics
   • Structured and Unstructured Data
   • Application and use cases of Big Data
   Limitations of traditional large Scale system s
   How a distributed way of computing is superior (cost and scale) Opportunities and challenges with Big Data
3. **HDFS (The Hadoop Distributed File System)**
   HDFS Overview and Architecture
   • Deployment Architecture
   • Nam e Node, Data Node and Checkpoint
   Node (aka Secondary Name Node)
   • Safe mode
   • Configuration files
   • HDFS Data Flows (Read v/s Write)
4. **How HDFS addresses fault tolerance?**
   • CRC Check Sum
   • Data replication
   • Rack awareness and Block placement policy
   • Small files problem
5. **HDFS Interfaces**
   • Command Line Interface
   • File System
   • Administrative
   • Web Interface
6. **Advanced HDFS features**
   • Load Balancer
   • Dist Cp
   • HDFS Federation
   • HDFS High Availability
   • Hadoop Archives
7. **Map Reduce – 1 (Theoretical Concepts)**
   MapReduce overview
   • Functional Programming paradigm s
   • How to think in a MapReduce way?
8. **MapReduce Architecture**
   • Legacy MR v/s Next Generation MapReduce
   (aka YARN/ MRv2)
   • Slots v/s Containers
   • Schedulers

- Shuffling, Sorting
- Hadoop Data Types
- Input and Output Formats
- Input Splits – Partitioning (Hash Partitioner v/s Customer Partitioner)
- Configuration files
- Distributed Cache

9. **MR Algorithm and Data Flow**
   - Word Count
10. **Alternatives to MR – BSP (Bulk Synchronous Parallel)**
    - Adhoc querying
    - Graph Computing Engines
11. **Map Reduce – 2 (Practice) Developing, debugging and deploying MR programs**
    - Stand alone mode (in Eclipse)
    - Pseudo distributed mode (as in the Big Data VM)

    – Fully distributed mode (as in Production)

    MR API
    - Old and the new MR API
    - Java Client API
    - Hadoop data types and custom Writable
12. **WritableCom parables**
    - Different input and output formats
    - Saving Binary Data using SequenceFiles and Avro Files

    Hadoop Streaming (developing and debugging non Java MR program s – Ruby and Python)
13. **Optimization techniques**
    - Speculative execution
    - Combiners
    - JVM Reuse
    - Compression
14. **MR algorithms (Non- graph)**
    - Sorting
    - Term Frequency
    - Inverse Document Frequency
    - Student Data Base
    - Max Temperature
    - Different ways of joining data
    - Word Co- Occurrence
15. **MR algorithms (Graph)**

    – PageRank

    – Inverted Index
16. **Higher Level Abstractions for MR (Pig)**

    Introduction and Architecture

    Different Modes of executing Pig constructs

Data Types
Dynamic invokers Pig streaming Macros
Pig Latin language Constructs (LOAD, STORE,
DUMP, SPLI T, etc) User Defined Functions
Use Cases

17. **Higher Level Abstractions for MR (Hive)**
    Introduction and Architecture
    Different Modes of executing Hive queries
    Metastore Implementations
    HiveQL (DDL & DML Operations) External v/s
    Managed Tables Views
    Partitions & Buckets User Defined Functions
    Transformations using Non Java Use Cases
18. **Comparison of Pig and Hive**
    NoSQL Databases – 1 (Theoretical
    Concepts)
    NoSQL Concepts
    • Review of RDBMS
    • Need for NoSQL
    • Brewers CAP Theorem
    • ACI D v/s BASE
    • Schema on Read vs. Schema on Write
    • Different levels of consistency
    • Bloom filters
19. **Different types of NoSQL databases**
    • Key Value
    • Columnar
    • Document
    • Graph
20. **Columnar Databases concepts NoSQL Databases – 2 (Practice)**
    HBase Architecture
    • Master and the Region Server
    • Catalog tables (ROOT and META)
    • Major and Minor compaction
    • Configuration files
    • HBase v/s Cassandra
21. **Interfaces to HBase (for DDL and DML operations)**
    • Java API
    • Client API
    • Filters
    • Scan Caching and Batching
    • Command Line Interface
    • REST API
22. **Advance HBase Features**
    • HBase Data Modeling
    • Bulk loading data in HBase

- HBase Coprocessors – EndPoints (similar to Stored Procedur es in RDBMS)
- HBase Coprocessors – Observers (similar to Triggers in RDBMS)

23. **Spark**
    - Introduction to RDD
    - Installation and Configuration of Spark
    - Spark Architecture
    - Different interfaces to Spark
    - Sample Python program s in Spark
24. **Introduction to YARN**
    - Usecase of YARN
    - YARN Architecture
    - YARN Demo
25. **Introduction to Oozie**
    - Usecase of Oozie
    - Oozie Architecture
    - Oozie Demo
26. **Introduction to Flume**
    - Usecase of Flume
    - Flume Architecture
    - Flume Demo
27. **Introduction to Sqoop**
    - Usecase of Sqoop
    - Sqoop Architecture
    - Sqoop Demo
28. **Setting up a Hadoop Cluster using Apache Hadoop**
    Cloudera Hadoop cluster on the Amazon Cloud (Practice)
    - Using EMR (Elastic Map Reduce)
    - Using EC2 (Elastic Compute Cloud)
29. **SSH Configuration**
    Stand alone mode (Theory) Distributed mode (Theory)
    - Pseudo distributed
    - Fully distributed
30. **Hadoop Ecosystem and Use Cases**
    Hadoop industry solutions
    - Importing/ exporting data across RDBMS and HDFS using Sqoop Getting real- time events into HDFS using Flume
    - Creating workflows in Oozie Introduction to Graph processing Graph processing with Neo4J

    •Using the Mongo Document Database
    •Using the Cassandra Columnar Database
    •Distributed Coordination with ZooKeeper
31. **Proof of concepts and use cases**
    Click Stream Analysis using Pig and Hive
    Analyzing the Twitter data with Hive
    Further ideas for data analysis