Hadoop Fundamentals

by Krish Krishnan

Module 0. About the Course (7 min)

Module 1. Foundations of Hadoop (45 min)

- *Speed in Compute*
- *Internet*
- *American Online (AOL) – Way to Connect to the Internet*
- *Netscape – Popular Dot-Com Portal*
- *Google – The First Search Engine*
- *Search Process*
- *Nutch*
- *Nutch Architecture*
- *Yahoo*
- *Hadoop Creation History*
- *Hadoop Today*
- *Facebook*
- *LinkedIn*

Module 2. Hadoop Core Modules: Part 1 (87 min)

- *Perspective – Food For Thought*
- *Why Hadoop?*
- *Human Behavior – New Insights*
- *Twitter Example*
- *Forces Shaping Business*
- *Conundrum*
- *The New Data Fabric*
- *Big Data*
  - Big Data Challenges
- *CIO Continuum*
- *Architect's Thinking*
- *User Needs*
- *State of Data*
- *What is Apache Hadoop*
- *Hadoop Design Goals*
- *Stack*
- *Core Components*
- *Storage: Hadoop Distributed File System (HDFS)*
  - Foundational Design Goals
  - HDFS Architecture: Parts 1 & 2
  - Write Operation in HDFS: Parts 1 & 2
  - Read Operation in HDFS: Parts 1 & 2
  - Block Placement
  - Replication Management
  - Using HDFS
  - Fast Changing Data Gaps
  - Kudu Design Goals

Module 3. Hadoop Core Modules: Part 2 (57 min)

- *Compute: MapReduce and Yet Another Resource Navigator (YARN)*

  - The MapReduce Framework
  - Automatic Parallel Execution in MapReduce
  - Hadoop 1.0 Architecture
  - Job Execution
- *Operating System: YARN*
  - Hadoop YARN Requirements
  - YARN Architecture
  - YARN Execution
  - YARN Application Lifecycle
  - YARN Deployed
- *Services Management: Zookeeper*
  - Services Management
  - The Zookeeper Service
  - The Zookeeper Data Model
  - ZNodes
  - ZNode Operations
  - ZNode Watches
  - API Synchronicity
- *Hadoop 3.0: More Enterprise Features*
  - *Hadoop 1 vs. Hadoop 2*
  - *Hadoop 3.0*
  - *Hadoop 3.0: Parts 1 & 2*

Module 4. Hadoop Ecosystem Components: Part 1 (78 min)

- *HBASE: Columnar Database*
  - HBASE
  - HBase Server Architecture
  - Region Assignment
  - HBase Tables
  - HBase Table
  - HBase – How Data is Actually Stored
  - HBase Write-Ahead-Log
- *PIG: Dataflow*
  - PIG
  - Pig Architecture
  - Pig Functions
  - Pig Types
  - Pig User Defined Functions
- *TEZ: Accelerator*
  - Why Do We Need Tez?
  - Where is Tez Available?
- *In-memory: Spark*
  - Motivation
  - AMP Lab – Berkeley Data Analytics
  - What is Spark?
  - Goal: In-Memory Data Sharing
  - Spark Programming Model
  - Example: Log Mining
  - RDDs
  - RDD Operations

    - RDD Fault Tolerance
    - Where is My Data?
    - RDD Benefits & Pitfalls
    - RDDs v. Distributed Shared Memory
    - API
    - Dataframe & Dataset
    - API Typing
    - Plan Optimization
    - Optimization
    - Catalyst + Intelligent Source
    - Spark Review
- *Data Ingestion: AVRO*
    - What Is Avro?
    - Avro Key Features
    - Avro Schema

Module 5. Hadoop Ecosystem Components: Part 2 (SQL on Hadoop) (34 min)

- *Apache Hive*
    - Hive
    - Hive Architecture
    - Hive Today
- *Impala*
    - Impala Origins
    - Google Dremel System
    - Impala Architecture
- *Apache Drill*
    - Apache Drill Brings Flexibility and Performance
    - Drill's Flexible Data Model
    - Enterprise Data Architecture
    - Extending Self Service to Schema-free Data
    - Drill Enables "SQL on Everything"
    - Granular Security Permissions Through Drill Views
    - MapR Optimized Data Architecture
- *Security in Hadoop*
- *Workflow*
- *Hadoop Technical Architecture*
- *Module 4 & 5 Summary*