Course Outline

Big Data Essentials Bootcamp

Course Summary

Description

Big Data needs proper tools and skills, and this workshop brings you “from zero to hero,” that is, provides the student with the necessary knowledge of Hadoop, Spark, and NoSQL. With these three fundamentals, you will be able to build systems processing massive amounts of data, in archival, batch, interactive and finally real-time manner. The workshop also lays foundations for proper analytics, allowing to extract insights from data.

Objectives

At the end of this course, students will be able to:

- Hadoop: HDFS, MapReduce, Pig, Hive
- Spark: Spark core, SparkSQL, Spark Java API, Spark Streaming
- NoSQL: Cassandra/HBase Architecture, Java API, Drivers, Data Modeling

Topics

- Hadoop
- HDFS Overview
- MapReduce Overview
- Pig
- Hive
- Spark
- Spark Streaming
- NoSQL
- Cassandra Basics
- Cassandra drivers
- Data Modeling – part 1
- Data Modeling – part 2
- Data Modeling Labs: Group design sessions

Audience

This course was designed for Developers.

Prerequisites

Before taking this course, students should be:

- Comfortable with Java programming language (most programming exercises are in Java)
- Comfortable in Linux environment (be able to navigate Linux command line, edit files using vi / nano)

Duration

Five days

Course Outline

Big Data Essentials Bootcamp

Course Outline

I. Hadoop
- A. Introduction to Hadoop
- B. Hadoop history, concepts
- C. ecosystem
- D. distributions
- E. High-level architecture
- F. Hadoop myths
- G. Hadoop challenges
- H. hardware/software

II. *HDFS Overview*
- A. concepts (horizontal scaling, replication, data locality, rack awareness)
- B. architecture (Namenode, Secondary NameNode, DataNode)
- C. data integrity
- D. future of HDFS: Namenode HA, Federation
- E. lab exercises

III. MapReduce Overview
- A. MapReduce concepts
- B. phases: driver, mapper, shuffle/sort, reducer
- C. thinking in MapReduce
- D. future of MapReduce (yarn)
- E. lab exercises

IV. Pig
- A. pig vs java vs MapReduce
- B. pig Latin language
- C. user-defined functions
- D. understanding pig job flow
- E. basic data analysis with Pig
- F. complex data analysis with Pig
- G. multi datasets with Pig
- H. advanced concepts
- I. lab exercises

V. Hive
- A. hive concepts
- B. architecture
- C. data types
- D. Hive data management
- E. hive vs SQL
- F. lab exercises

VI. Spark
- A. Spark Basics
- B. Background and history
- C. Spark and Hadoop
- D. Spark concepts and architecture
- E. Spark ecosystem (core, spark SQL, mlib, streaming)
- F. First look at Spark
- G. Spark in local mode
- H. Spark web UI
- I. Spark shell
- J. Analyzing dataset – part 1
- K. Inspecting RDDs
- L. RDDs In Depth
- M. Partitions
- N. RDD Operations / transformations
- O. RDD types
- P. MapReduce on RDD
- Q. Caching and persistence
- R. Sharing cached RDDs
- S. Spark API programming
- T. Introduction to Spark API / RDD API
- U. Submitting the first program to Spark
- V. Debugging/logging
- W. Configuration properties

VII. Spark Streaming
- A. Streaming overview
- B. Streaming operations
- C. Sliding window operations
- D. Writing spark streaming applications

VIII. NoSQL
- A. Introduction to Big Data / NoSQL
- B. NoSQL overview
- C. CAP theorem
- D. When is NoSQL appropriate
- E. NoSQL ecosystem

IX. Cassandra Basics
- A. Cassandra nodes, clusters, datacenters
- B. Keyspaces, tables, rows, and columns
- C. Partitioning, replication, tokens
- D. Quorum and consistency levels
- E. Labs

Course Outline (cont'd)

X. Cassandra drivers
- A. Introduction to Java driver
- B. CRUD (Create / Read / Update, Delete) operations using Java client
- C. Asynchronous queries
- D. Labs

XI. Data Modeling – part 1
- A. introduction to CQL
- B. CQL Datatypes
- C. creating keyspaces & tables
- D. Choosing columns and types
- E. Choosing primary keys
- F. Data layout for rows and columns
- G. Time to live (TTL), create, insert, update
- H. Querying with CQL
- I. CQL updates
- J. Labs

XII. Data Modeling – part 2
- A. Creating and using secondary indexes
- B. Denormalization and join avoidance
- C. composite keys (partition keys and clustering keys)
- D. Time series data
- E. Best practices for time series data
- F. Counters
- G. Lightweight transactions (LWT)

XIII. Data Modeling Labs: Group design sessions
- A. multiple use cases from various domains are presented
- B. students work in groups to come up with designs and models
- C. discuss various designs, analyze decisions
- D. Lab: implement 'Netflix' data models, generate data