

Big Data Essentials Bootcamp

Course Outline:

Introduction to Hadoop

Hadoop history, concepts
ecosystem
distributions
High-level architecture
Hadoop myths
Hadoop challenges
hardware / software *HDFS Overview*
concepts (horizontal scaling, replication, data locality, rack awareness)
architecture (Namenode, Secondary NameNode, DataNode)
data integrity
future of HDFS : Namenode HA, Federation
lab exercises *MapReduce Overview*
MapReduce concepts
phases : driver, mapper, shuffle/sort, reducer
thinking in MapReduce
future of mapreduce (yarn)
lab exercises *Pig*
pig vs java vs MapReduce
pig latin language
user defined functions
understanding pig job flow
basic data analysis with Pig
complex data analysis with Pig
multi datasets with Pig
advanced concepts
lab exercises *Hive*
hive concepts
architecture
data types
Hive data management
hive vs sql
lab exercises *Spark Basics*
Spark Background and history
Spark and hadoop
Spark concepts and architecture
Spark eco system (core, spark sql, mlib, streaming)
First look at Spark
Spark in local mode
Spark web UI

Spark shell
Analyzing dataset – part 1
Inspecting RDDs *RDDs In Depth* Partitions
RDD Operations / transformations
RDD types
MapReduce on RDD
Caching and persistence
Sharing cached RDDs *Spark API programming*

Introduction to Spark API / RDD API
Submitting the first program to Spark
Debugging / logging
Configuration properties
Spark Streaming
Streaming overview
Streaming operations
Sliding window operations
Writing spark streaming applications
NoSQL
Introduction to Big Data / NoSQL
NoSQL overview
CAP theorem
When is NoSQL appropriate
NoSQL ecosystem
Cassandra Basics
Cassandra nodes, clusters, datacenters
Keyspaces, tables, rows and columns
Partitioning, replication, tokens
Quorum and consistency levels
Labs
Cassandra drivers
Introduction to Java driver
CRUD (Create / Read / Update, Delete) operations using Java client
Asynchronous queries
Labs
Data Modeling – part 1
introduction to CQL

CQL Datatypes

creating keyspaces & tables

Choosing columns and types

Choosing primary keys

Data layout for rows and columns

Time to live (TTL), create, insert, update

Querying with CQL

CQL updates

Labs

Data Modeling – part 2

Creating and using secondary indexes

Denormalization and join avoidance

composite keys (partition keys and clustering keys)

Time series data

Best practices for time series data

Counters

Lightweight transactions (LWT)

Data Modeling Labs : Group design sessions

multiple use cases from various domains are presented

students work in groups to come up designs and models

discuss various designs, analyze decisions

Lab : implement 'Netflix' data models, generate data