Introduction to Big Data and NoSQL Training

Learn what is considered Big Data and what are NoSQL DBs and experience Big Data Solutions including Hadoop HDFS, Map Reduce, NoSQL DBs: Document Based DB (Mongo DB), Big Table DB (Cassandra DB), Redis, Hive, Kafka, Elastic Search and much more, Pick and choose between the most common and fascinating technologies that are able to store and manipulate big data. Understand best practices for each solution, understand high availability and scaling capabilities and when and how to choose each.

This training introduces the popular NOSQL and HADOOP current solutions and provides basic hands on experience on each of the solutions themselves.

Audience

System Administrators, Operations

Database Administrators, Support

DevOps, Developers

Knowledge Requirements

Relational Database SQL experience

Length

5 Days

Bonus

Hands-on lab sessions

Syllabus

Course Introduction
Introduction to Big Data
Introduction to Hadoop and HDFS
Hadoop Map Reduce
Introduction to Apache SPARK
Introduction to Document Databases – Mongo DB
Introduction to Bug Table DBs – Apache Cassandra DB
Introduction to Redis
Introduction to Apache Hive
Apache Kafka
Data Architect for Big Data
NOSQL Comparison
Summary
Optional per customer request (Apache Hive, Gigaspaces XAP, Apache Ignite, Redis Enterprise, Elastic Search and more)

HARDWARE AND SOFTWARE REQUIREMENTS

Computer Requirements

- **RAM: minimum 6 GB of RAM required for exercises and platform to operate, 8 GB and up recommended.**
- **Disk Space: At least 40 GB of free disk space**
- **Internet connection**
- **All machines connected to same Network**

Supported Operating Systems

- **Linux Ubuntu (Can also be a Virtual Machine)**

Additional Software Requirements

- **PDF Reader**
- **Zip software**
- **VMPlayer 14 and up**

Class HW required

- **Projector 1024*768 minimum resolution**
- **White Board**
- **Erasable Markers**
- **Desktops or Laptops (see HW Requirements)**
- **12-24 ports Switch**
- **Internet connectivity**
- **Electricity outlets for all computers/monitors and other equipment.**
- **At least 3 electricity outlets next to instructor location.**

Call us and we will customise a class especialy for your needs +972-54-260-7104

AGENDA

Day 1: Lesson 1: Course Introduction	Duration: 30 minutes

- Course Introduction
- Courseware walkthrough
- Documentation
- Lab

Day 1: Lesson 2: Introduction to Big Data and NoSQL	Duration: 0.5 Day

- What is Big Data?
- Big Data challenges and complexity
- General concepts
- Architecture considerations
- Presenting use cases of internet companies (e.g. Facebook)
- The Data Scientist
- RDBMS: Advantages and disadvantages / Impedance Mismatch
- No-SQL vs. Traditional Enterprise Relational Data:
- CAP theorem vs. ACID / Dynamic schema, sharding, replications and caching / Performance
- Scaling vs consistency
- No-SQL types and use cases: Key/value stores / Document databases / Column oriented databases / Graph databases
- When (not) to use No-SQL?

Day 1 : Lesson 3: Introduction to Hadoop and HDFS	Duration: 0.75 Day

- Hadoop Pseudo-Distributed Mode installation
- Lab – Hadoop Installation
- HDFS Assumptions and Goals
- Scale and Feature requirements
- Architecture
- Data Replication
- Robustness
- Data Organization
- Accessibility
- Space Reclamation
- Lab (Install Configure and Experience)

Day 2: Lesson 4: Hadoop Map Reduce and Hadoop 2.0	Duration: 0.5 Day

- Map Reduce Basics
- Inputs and Outputs
- Map Reduce Code Example
- Map Reduce Additional Details
- Map Reduce V1 vs. V2
- YARN – Yet Another Resource Negotiator.
- Lab (Install Configure and Experience)

Day 2: Lesson 5: Introduction to Apache SPARK	Duration: 0.5 Day

- Apache SPARK Introduction
- Getting Started
- SPARK architecture

- SPARK processing
- Map Reduce
- Example
- Lab (Install Configure and Experience)

Day 3: Lesson 6: Introduction to Document Databases – Mongo DB	Duration: 0.5 Day

- Mongo DB Introduction
- Getting Started - Installation
- Mongo DB basic commands
- Aggregation
- Replication
- Sharding
- Sharded Cluster Requirements and configuration
- Shard Cluster Deployment
- Other considerations
- Lab (Install Configure and Experience)

Day 3: Lesson 7: Introduction to Apache Cassandra DB	Duration: 0.5 Day

- Cassandra DB Introduction
- Getting Started
- Google Big Table
- Amazon Dynamo
- Cassandra Query Language Shell - CQLSH
- Replication & Partitioning
- Basic administration
- Cluster configuration
- Lab (Install Configure and Experience)

Day 4: Lesson 8: Introduction to Redis	Duration: 0.25 Day

- Why in Memory Data Grid?
- IMDG Terminology Comparison to Common Platforms and Servers
- Introduction to Redis
- Redis Data Structures
- Redis CLI
- Use cases
- High availability and scalability
- Redis API
- Lab (Install Configure and Experience)

Day 4: Lesson 9: Introduction to Hive	Duration: 0.75 Day

- Hive Introduction
- Getting Started
- Hive Data Model
- Import Data
- Hive Query Language
- Architecture
- Lab (Install Configure and Experience)

Day 5: Lesson 10: Introduction to Apache Kafka	Duration: 0.25 Day

- Common architecture
- Data Lake
- Streaming
- Introduction to Kafka
- Kafka – Producers, Consumers, Streams, Connectors
- Kafka – Store
- Kafka Consumers
- Kafka Consumer Groups
- Kafka Offsets
- Kafka Transactions
- Lab (install demo and experience)

Day 5: Lesson 11: The Data Architect	Duration: 0.25 Day

- The Data Architect
- Data Analysis
- Data Characteristics (Read rarely, Read once, Read many, Write Once, Write Many (Updated after inserted)
- Design Patterns
- Schema Design for NoSQL DBs
- Final Lab Session (Combine all the pieces)

Day 5: Lesson 12: Summary and NoSQL Comparison	Duration: 0.5 Day

- NO SQL DBs General Comparison
  - Performance Comparison
  - When to use which?
- Summary
- Wrap Up