# NoSQL Architecture Comparison (2 Days)

The NoSQL (Not Only SQL) persistence systems space offers a great variety of solutions that may be overwhelming. This class aims at helping the attendees understand the challenges of the emerging world of Big Data as well as identify suitable use cases for a variety of NoSQL systems such as Pig, Hive, HBase, Cassandra and MongoDB.

The attendees will also be given some underlying architecture details of those NoSQL systems to enable them make informed decisions about using NoSQL systems when they return to work

### *Objectives*

- Introduce students to the core concepts of Big Data
- Provide a general overview of the most common NoSQL stores
- Explain how to choose the correct NoSQL database for specific use cases
- General overview of the architecture of Hadoop and MongoDB

### *Course Outline*

**Chapter 1. Defining Big Data**

- Transforming Data into Business Information
- Quality of Data
- Gartner's Definition of Big Data
- More Definitions of Big Data
- Processing Big Data
- Challenges Posed by Big Data
- The Cloud and Big Data
- The Business Value of Big Data
- Big Data: Hype or Reality?
- Big Data Quiz
- Big Data Quiz Answers

***Chapter 2. NoSQL and Big Data Systems Overview***

- Limitations of Relational Databases
- Limitations of Relational Databases (Con't)
- What are NoSQL (Not Only SQL) Databases?
- The Past and Present of the NoSQL World
- NoSQL Database Properties
- NoSQL Benefits
- NoSQL Database Storage Types
- The CAP Theorem
- Mechanisms to Guarantee a Single CAP Property
- NoSQL Systems CAP Triangle
- Limitations of NoSQL Databases
- Big Data Sharding
- Sharding Example
- Amazon S3
- Amazon Storage SLAs
- Amazon Glacier
- Amazon S3 Security
- Data Lifecycle Management with Amazon S3
- Amazon S3 Cost Monitoring
- OpenStack
- Object Store (Swift)
- Components of Swift
- Google BigTable
- BigTable-based Applications
- BigTable Design

***Chapter 3. Adopting NoSQL***

- Hype Cycle and Technology Adoption Model
- Barriers to Adoption

- Dismantling Barriers to Adoption
- Use Cases for NoSQL Database Systems
- Example Applications
- Industry trends
- Enterprise Hadoop Solutions Offerings
- Enterprise Big Data / NoSQL Offerings
- IBM InfoSphere Platform
- Oracle Big Data Appliance
- NoSQL Technology Adoption Action Plan

**Chapter 4. MapReduce Overview**

- MapReduce Defined
- Google's MapReduce
- MapReduce Explained
- MapReduce Word Count Job
- MapReduce Shared-Nothing Architecture
- Similarity with SQL Aggregation Operations
- Example of Map & Reduce Operations using JavaScript
- Problems Suitable for Solving with MapReduce
- Typical MapReduce Jobs
- Fault-tolerance of MapReduce
- Distributed Computing Economics
- MapReduce Systems

**Chapter 5. Introduction to MongoDB**

- MongoDB
- MongoDB Features (Cont'd)
- MongoDB's Logo
- Positioning of MongoDB
- Sharding in MongoDB
- Data Replication
- A Sample Sharded Cluster Diagram
- MongoDB Security
- Authentication
- Data and Network Encryption
- MongoDB Limitations
- MongoDB Operational Intelligence
- MongoDB Use Cases
- MongoDB Data Model
- The _id Primary Key Filed Considerations
- Terminology
- MongoDB Data Model
- Data Modeling in RDBMS
- Data Modeling in MongoDB
- MongoDB Data Modeling
- A Sample JSON Document Matching the Schema
- Data Lifecycle Management
- Data Lifecycle Management: TTL
- Data Lifecycle Management: Capped Collections
- MongoDB Query Language (QL)
- The find and findOne Methods
- The find and findOne Methods
- A MongoDB QL Example
- Data Inserts
- Creating an Index
- MongoDB vs Apache CouchDB

**Chapter 6. Hadoop Overview**

- Apache Hadoop
- Apache Hadoop Logo
- Typical Hadoop Applications
- Hadoop Clusters
- Hadoop Design Principles
- Hadoop's Core Components
- Hadoop Simple Definition
- High-Level Hadoop Architecture
- Hadoop-based Systems for Data Analysis

**Chapter 7. Hadoop Distributed File System Overview**

- Hadoop Distributed File System
- Data Blocks
- Data Block Replication Example
- HDFS NameNode Directory Diagram
- Accessing HDFS
- Examples of HDFS Commands
- Client Interactions with HDFS for the Read Operation
- Read Operation Sequence Diagram
- Client Interactions with HDFS for the Write Operation
- Communication inside HDFS

**Chapter 8. MapReduce with Hadoop**

- Hadoop's MapReduce
- MapReduce v1 ("Classic MapReduce")
- JobTracker and TaskTracker
- YARN (MapReduce v2)
- MapReduce Programming Options
- Java MapReduce API
- The Structure of a Java MapReduce Program

- The Mapper Class
- The Reducer Class
- The Driver Class
- Compiling Classes
- Running the MapReduce Job
- The Structure of a Single MapReduce Program
- Combiner Pass (Optional)
- Hadoop's Streaming MapReduce
- Python Word Count Mapper Program Example
- Python Word Count Reducer Program Example
- Setting up Java Classpath for Streaming Support
- Streaming Use Cases
- The Streaming API vs Java MapReduce API
- Amazon Elastic MapReduce

**Chapter 9. Apache Pig Scripting Platform**

- What is Pig?
- Pig Latin
- Apache Pig Logo
- Pig Execution Modes
- Local Execution Mode
- MapReduce Execution Mode
- Running Pig
- Running Pig in Batch Mode
- What is Grunt?
- Pig Latin Statements
- Pig Programs
- Pig Latin Script Example
- SQL Equivalent
- Differences between Pig and SQL
- Statement Processing in Pig
- Comments in Pig
- Supported Simple Data Types
- Supported Complex Data Types
- Arrays
- Defining Relation's Schema
- The bytearray Generic Type
- Using Field Delimiters
- Referencing Fields in Relations

**Chapter 10. Apache Pig HDFS Interface**

- The HDFS Interface
- FSShell Commands (Short List)

**Chapter 11. Apache Pig Relational and Eval Operators**

- Pig Relational Operators
- Example of Using the JOIN Operator
- Example of Using the Order By Operator
- Caveats of Using Relational Operators
- Pig Eval Functions
- Caveats of Using Eval Functions (Operators)
- Example of Using Single-column Eval Operations
- Example of Using Eval Operators For Global Operations

**Chapter 12. Hive**

- What is Hive?
- Apache Hive Logo
- Hive's Value Proposition
- Who uses Hive?
- Hive's Main Systems
- Hive Features
- Hive Architecture
- HiveQL
- Where are the Hive Tables Located?
- Hive Command-line Interface (CLI)

**Chapter 13. Hive Command-line Interface**

- Hive Command-line Interface (CLI)
- The Hive Interactive Shell
- Running Host OS Commands from the Hive Shell
- Interfacing with HDFS from the Hive Shell
- The Hive in Unattended Mode
- The Hive CLI Integration with the OS Shell
- Executing HiveQL Scripts
- Comments in Hive Scripts
- Variables and Properties in Hive CLI
- Setting Properties in CLI
- Example of Setting Properties in CLI
- Hive Namespaces
- Using the SET Command
- Setting Properties in the Shell
- Setting Properties for the New Shell Session

**Chapter 14. Hive Data Definition Language**

- Hive Data Definition Language
- Creating Databases in Hive
- Using Databases
- Creating Tables in Hive
- Supported Data Type Categories
- Common Primitive Types
- Example of the CREATE TABLE Statement
- The STRUCT Type
- Table Partitioning
- Table Partitioning
- Table Partitioning on Multiple Columns
- Viewing Table Partitions
- Row Format
- Data Serializers / Deserializers
- File Format Storage
- More on File Formats
- The EXTERNAL DDL Parameter
- Example of Using EXTERNAL
- Creating an Empty Table
- Dropping a Table
- Table / Partition(s) Truncation
- Alter Table/Partition/Column
- Views
- Create View Statement
- Why Use Views?
- Restricting Amount of Viewable Data
- Examples of Restricting Amount of Viewable Data
- Creating and Dropping Indexes

**Chapter 15. Apache HBase**

- What is HBase?
- HBase Design
- HBase Features
- The Write-Ahead Log (WAL) and MemStore
- HBase vs RDBS
- HBase vs Apache Cassandra
- Interfacing with HBase
- HBase Thrift And REST Gateway
- HBase Table Design
- Column Families
- A Cell's Value Versioning
- Timestamps
- Accessing Cells
- HBase Table Design Digest
- Table Horizontal Partitioning with Regions
- HBase Compaction
- Loading Data in HBase
- HBase Shell
- HBase Shell Command Groups
- Creating and Populating a Table in HBase Shell
- Getting a Cell's Value
- Counting Rows in an HBase Table