Introduction to Cassandra for JPMC

Course: Introduction to Cassandra
Course ID: TTLCASS2-GKJ
Duration: 2 full days or 4 half days
Audience: **This introductory-level course is for attendees new to Cassandra**. Students should be experienced SQL and ETL developers with some database familiarity.
Hands-on Learning: This course is *approximately* **50% hands-on lab to 50% lecture ratio**, combining engaging lecture, demos, group activities and discussions with machine-based student labs and exercises. Student machines are required.

The Cassandra (C*) database is a massively scalable NoSQL database that provides high availability and fault tolerance, as well as linear scalability when adding new nodes to a cluster. It has many powerful capabilities, such as tunable and eventual consistency, that allow it to meet the needs of modern applications, but also introduce a new paradigm for data modeling that many organizations do not have the expertise to use in the best way.

Introduction to Cassandra (for JPMC) is a two-day (or four half-day) hands-on course designed to teach attendees the basics of how to create good data models with Cassandra. This technical course has a focus on the practical aspects of working with C*, and introduces essential concepts needed to understand Cassandra, including enough coverage of internal architecture to make good decisions. It is hands-on, with labs that provide experience in core functionality. Students will also explore CQL (Cassandra Query Language), as well as some of the “anti-patterns” that lead to non-optimal C* data models and be ready to work on production systems involving Cassandra.

Learning Objectives

The goal of this course is to enable technical students new to Cassandra to begin working with Cassandra in an optimal manner. Throughout the course students will learn to:

Understand the Big Data needs that C* addresses
Be familiar with the operation and structure of C*
Be able to install and set up a C* database
Use the C* tools, including cqlsh, nodetool, and ccm (Cassandra Cluster Manager)
Be familiar with the C* architecture, and how a C* cluster is structured
Understand how data is distributed and replicated in a C* cluster
Understand core C* data modeling concepts, and use them to create well-structured data models
Be familiar with the C* eventual consistency model and use it intelligently
Be familiar with consistency mechanisms such as read repair and hinted handoff
Understand and use CQL to create tables and query for data
Know and use the CQL data types (numerical, textual, uuid, etc.)
Be familiar with the various kinds of primary keys available (simple, compound, and composite primary keys)
Be familiar with the C* write and read paths
Understand C* deletion and compaction
Get introduced to using Cassandra and IntelliJ

Audience & Pre-Requisites

Attendees should have incoming experience with and knowledge of SQL. Some familiarity with distributed systems is also helpful.

Course Topics / Agenda

NOTE: The topics, tools and skills in this course have been selected by JPMC management to align

with the skills and technologies utilized by your overall organization. Timing may be adjusted by the instructor during live delivery based on audience skill-level, needs and participation.

Session 1: Cassandra Overview

- Why We Need Cassandra - Big Data Challenges vs RDBMS
- High level Cassandra Overview
- Cassandra Features
- Basic Cassandra Installation and Configuration

Session 2: Cassandra Architecture and CQL Overview

- Cassandra Architecture Overview
- Cassandra Clusters and Rings
- Nodes and Virtual Nodes
- Data Replication in Cassandra
- Introduction to CQL
- Defining Tables with a Single Primary Key
- Using cqlsh for Interactive Querying
- Selecting and Inserting/Upserting Data with CQL
- Data Replication and Distribution
- Basic Data Types (including uuid, timeuuid)

Session 3: Data Modeling and CQL Core Concepts

- Defining a Compound Primary Key
  - CQL for Compound Primary Keys
  - Partition Keys and Data Distribution
  - Clustering Columns
  - Overview of Internal Data Organization
- Overview of Other Querying Capabilities
  - ORDER BY, CLUSTERING ORDER BY, UPDATE , DELETE, ALLOW FILTERING
  - Batch Queries
- Data Modeling Guidelines
  - Denormalization
  - Data Modeling Workflow
  - Data Modeling Principles
  - Primary Key Considerations
- Composite Partition Keys
  - Defining with CQL
  - Data Distribution with Composite Partition Key
  - Overview of Internal Data Organization
  - Lab: Composite Partition Key (Substantial lab)

Session 4: Additional CQL Capabilities

- Indexing
  - Primary/Partition Keys and Pagination with token()
  - Secondary Indexes and Usage Guidelines
- Cassandra collections
  - Collection Structure and Uses
  - Defining and Querying Collections (set, list, and map)
- Materialized View
  - Overview
  - Usage Guidelines

Session 5: Data Consistency In Cassandra

- Overview of Consistency in Cassandra
- CAP Theorem
- Eventual (Tunable) Consistency in C* - ONE, QUORUM, ALL
- Choosing CL ONE
- Choosing CL QUORUM
- Achieving Immediate Consistency
- Overview of Other Consistency Levels
- Supportive Consistency Mechanisms
  - Writing / Hinted Handoff
  - Read Repair
  - Nodetool repair

Session 6: Internal Mechanisms

- Ring Details
  - Partitioners
  - Gossip Protocol
  - Snitches
- Write Path
  - Overview / Commit Log
  - Memtables and SSTables
  - Write Failure
    - Unavailable Nodes and Node Failure
    - Requirements for Write Operations
- Read Path Overview
  - Read Mechanism
  - Replication and Caching
  - Deletion/Compaction Overview
  - Delete Mechanism
  - Tombstones and Compaction

Session 7: Working with IntelliJ

- Configuring JDBC Data Source for Cassandra
- Reading Schema Information
- Querying and Editing Tables