

Impala – An Open Source SQL Engine for Hadoop

Course Summary

Description

This is an ideal course package for individuals who want to understand the basic concepts of Massively Parallel Processing or MPP SQL query engine that runs on Apache Hadoop. On completing this course, learners will be able to interpret the role of Impala in the Big Data Ecosystem.

The course focuses on the basics of Impala. It further provides an overview of the superior performance of Impala, against other popular SQL-on-Hadoop systems.

Objectives

By the end of this course, participants will be able to:

- Describe Impala and its role in Hadoop Eco-system
- Explain how to query data using impala SQL
- Discuss partitioning of Impala tables and explain its benefits
- List the factors affecting the performance of Impala
- Describe the complete flow of a SQL query execution in Impala.

Topics

- Course Introduction
- Introduction to Impala
- Querying with Hive and Impala
- Data Storage and File Format
- Working with Impala

Audience

- Analysts
- Data scientists
- Hadoop administrator and developers
- SQL developers
- Data warehouse developers
- Database administrators and developers

Prerequisites

Fundamental Knowledge of programming language and Hadoop components is the basic course prerequisite. However, participants are expected to have knowledge of SQL commands.

Duration

Two days

Impala – An Open Source SQL Engine for Hadoop

Course Outline

I. Course Overview

- A. About Impala-An Open Source SQL Engine for Hadoop course

II. Introduction to Impala

- A. What is Impala?
- B. Benefits of Impala
- C. Exploratory Business Intelligence
- D. Impala Installation
- E. Demo: Using Cloudera Manager for Impala
- F. Starting Impala
- G. Demo: Starting Impala from Command Line
- H. Data Storage
- I. Managing Metadata
- J. Controlling Access to Data
- K. Impala shell Commands
- L. Demo: Launching Impala Shell and Shell Command

III. Querying with Hive and Impala

- A. SQL Language Statements
- B. DDL Statements
- C. DML Statements
- D. Create Database
- E. Create Table
- F. Internal and External Table
- G. Loading Data into Impala Table
- H. Alter and Drop Table
- I. Drop Database
- J. Describe, Explain, Show Table statement
- K. Insert and Select Statement
- L. Data Type
- M. Operators
- N. Functions
- O. Create View in Impala
- P. Hive and Impala Query Syntax
- Q. Demo: Using Impala Shell for DDL and DDML SQL Statements

IV. Data storage and File Format

- A. Partitioning Tables
- B. SQL Statement for Partitioned Tables
- C. File Format and Performance Considerations
- D. Choosing File Type and Compression Techniques
- E. Demo: Choosing File Type and Compression Techniques

V. Working with Impala

- A. Impala Architecture
- B. Impala Daemon
- C. Impala Statestore
- D. Impala Catalog Service
- E. Query Execution Flow in Impala
- F. User Defined Functions
- G. Hive UDFs with Impala
- H. Demo: UDF in Impala
- I. Improving Impala Performance