BIGDATA HADOOP FOR DATA SCIENCE (3 DAYS)

By Dr. Vishwanath Rao

Hadoop Overview

- •Apache Hadoop
- •Apache Hadoop Logo
- •Typical Hadoop Applications
- •Hadoop Clusters
- •Hadoop Design Principles
- •Hadoop Versions
- •Hadoop's Main Components
- •Hadoop Simple Definition
- •Side-by-Side Comparison: Hadoop 1 and Hadoop 2
- •Hadoop-based Systems for Data Analysis
- •Other Hadoop Ecosystem Projects
- •Hadoop Caveats
- •Hadoop Distributions
- •Cloudera Distribution of Hadoop (CDH)
- •Cloudera Distributions
- •Hortonworks Data Platform (HDP)
- •MapR
- •Summary

Hadoop Distributed File System Overview

- •Hadoop Distributed File System (HDFS)
- •HDFS High Availability
- •HDFS "Fine Print"
- •Storing Raw Data in HDFS
- •Hadoop Security
- •HDFS Rack-awareness
- •Data Blocks
- •Data Block Replication Example
- •HDFS NameNode Directory Diagram
- •Accessing HDFS
- •Examples of HDFS Commands
- •Other Supported File Systems

- WebHDFS
- Examples of WebHDFS Calls
- Client Interactions with HDFS for the Read Operation
- Read Operation Sequence Diagram
- Client Interactions with HDFS for the Write Operation
- Communication inside HDFS
- Summary

Hive

- What is Hive?
- Apache Hive Logo
- Hive's Value Proposition
- Who uses Hive?
- Hive's Main Sub-Systems
- Hive Features
- The "Classic" Hive Architecture
- The New Hive Architecture
- HiveQL
- Where are the Hive Tables Located?
- Hive Command-line Interface (CLI)
- The Beeline Command Shell
- Summary

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
- Setting Alternative Hive Execution Engines
- The Beeline Shell
- Connecting to the Hive Server in Beeline
- Beeline Command Switches
- Beeline Internal Commands
- Summary

Hive Data Definition Language

- Hive Data Definition Language
- Creating Databases in Hive
- Using Databases
- Creating Tables in Hive
- Supported Data Type Categories
- Common Numeric Types
- String and Date / Time Types
- Miscellaneous Types
- Example of the CREATE TABLE Statement
- Working with Complex Types
- Table Partitioning
- Table Partitioning
- Table Partitioning on Multiple Columns
- Viewing Table Partitions
- Row Format
- Data Serializers / Deserializers
- File Format Storage
- File Compression
- More on File Formats
- The ORC Data Format
- Converting Text to ORC Data Format
- 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
- Describing Data
- Summary

Hive Data Manipulation Language

- Hive Data Manipulation Language (DML)
- Using the LOAD DATA statement
- Example of Loading Data into a Hive Table
- Loading Data with the INSERT Statement
- Appending and Replacing Data with the INSERT Statement
- Examples of Using the INSERT Statement
- Multi Table Inserts
- Multi Table Inserts Syntax
- Multi Table Inserts Example
- Summary

Hive Select Statement

- HiveQL
- The SELECT Statement Syntax
- The WHERE Clause
- Examples of the WHERE Statement
- Partition-based Queries
- Example of an Efficient SELECT Statement
- The DISTINCT Clause
- Supported Numeric Operators
- Built-in Mathematical Functions
- Built-in Aggregate Functions
- Built-in Statistical Functions
- Other Useful Built-in Functions
- The GROUP BY Clause
- The HAVING Clause
- The LIMIT Clause
- The ORDER BY Clause
- The JOIN Clause
- The CASE … Clause
- Example of CASE … Clause
- Summary