Big Data Analysis Using Pig and Hive

Course 3536 – 24 Hours

Overview

Big data analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns and other useful information that can provide competitive advantages and better business results.
Big data analytics can be done with the software tools commonly used - but the unstructured data sources used for big data analytics may not fit in traditional data warehouses. Additionally, traditional data warehouses may not be able to handle the processing demands posed by big data. As a result, a new class of big data technology has emerged and is being used in many big data analytics environments.
In this course participants will learn how to analyze Big Data stored in Hadoop by using Pig and Hive.
This course is designed to introduce you to Hadoop architecture and MapReduce framework, Pig scripting language and Hive

On Completion, Delegates will be able to

- Understand Hadoop architecture and its main components
- Data ETL with Hadoop tools
- Write Pig scripts
- Use Hive's SQL dialect to query and analyze large datasets store in Hadoop

Who Should Attend

- Data analysts, business intelligence, developers and DBA

Prerequisites

- SQL and basic UNIX or Linux commands
- A background in Java is NOT required
- Prior knowledge of Apache Hadoop is NOT required

Course Contents

Module 1: Introduction to Big Data

- RDBMS - Advantages and disadvantages
- Dynamic schema, sharding, replications and caching
- Performance
- Motivation for Hadoop

Module 2: Introduction to Hadoop

- Hadoop overview
- HDFS architecture
- Map/Reduce framework
- Joins with Map-Reduce
    - Map-side join
    - Reduce side join
    - Join with distributed cache
- Hands on: launching Hadoop and a map reduce job on it

Module 3: Pig

- Thinking like a Pig
- Pig vs RDBMS
- Learning Pig Latin
    - Structure
    - Statement
    - Expressions
    - Data Types
    - Schemas
    - Functions
    - Macros
- Data processing operators:
    - Loading
    - Grouping
    - Sorting

Module 4: Pig Advanced

- Data sampling
- Execution plan
- User defined functions
- Case studies

Module 5: Hive

- Hive introduction
  - What is Hive?
  - Hive schema and data storage
  - Hive vs RDBMS
  - Hive vs Pig.
  - When to use Hive?
  - Interacting with hive.
  - Hive services
- The Hive commands
- Hive as relational data
- The Metastore

Module 6: Hive Data Types

- Primitive Data types
- Collection data types (Struct, map and array)
- Text file encoding of data values

Module 7: HiveQL - DDL

- Database and table commands
- External tables
- Partitioned table
- Storage formats

Module 8: HiveQL – Data Manipulation

- Loading data
- Inserting data into table from queries
- Exporting data

Module 9: HiveQL - Queries

- SELECT clauses
- WHERE clauses
- Nested SELECT
- Using functions
- GROUP BY
- JOINS statements
- ORDER BY
- SORT BY
- Queries that sample data

Module 10: HiveQL - Views

- Why view?
- View and map types for dynamic tables

Module 11: HiveQL - Indexes

- Creating an index
- Rebuilding the index

Module 12: Schema design

- Schema on read vs schema on write
- Table by day
- Partitioning
- Unique keys and normalization
- Bucketing
- Compression

Module 13: Hive advanced

- Hive optimization and tuning:
  - Using explain
  - Job execution plans
  - Optimized joins
  - Local mode
  - Parallel execution
  - Tuning with Limit
  - Controlling the number of mappers and reducers
  - Indexing
  - Dynamic partition

Module 14: Functions

- Standard functions.
- Aggregate functions.
- Table generating functions.
- Statistics and data mining function in Hive.
- User defined functions.

Module 15: Advance format types

- File formats:
  - Sequence
  - RCFile
- CSV, TSV and SerDes files.
- XML.
- JSON SerDe.