HIVE

Duration: 2 Days

Pre-requisite:

- Knowledge of SQL, data modeling, and scripting is also helpful. No prior **Hadoop** Knowledge is needed

Hadoop Distributed File System (HDFS)
• HDFS overview and design
• HDFS architecture
• HDFS file storage
• Component failures and recoveries
• Block placement

Map-Reduce Abstraction
• What MapReduce is and why it is popular
• The Big Picture of the MapReduce
• MapReduce process and terminology
• MapReduce components failures and recoveries
• Working with MapReduce
• Lab: Working with MapReduce

Hive

Overview of Hive

- Why Hive?
- Use cases
- Hive architecture – building blocks
- Hive CLI and language (exercise)
- Variables and Properties
- Executing Hive Queries from Files
- Primitive Data Types
- Collection Data Types
- Difference between HivelQL and SQL92
- Custom UDF

HiveQL Queries

- SELECT … FROM Clauses
    - Specify Columns with Regular Expressions
    - Computing with Column Values
    - Arithmetic Operators
    - Using Mathematical, Aggregate, Table generating and other built-in functions
    - LIMIT Clause
    - Column Aliases
    - Nested SELECT Statements
    - CASE … WHEN … THEN Statements
- WHERE Clauses
    - Predicate Operators
    - Gotchas with Floating-Point Comparisons
    - LIKE and RLIKE
- GROUP BY Clauses
    - HAVING Clauses

HiveQL: Data Manipulation

- Hive Variables
- Partitioned, Managed Tables
- External Partitioned Tables
- Loading Data into Managed Tables
- Inserting Data into Tables from Queries
- Dynamic Partition Inserts
- Creating Tables and Loading Them in One Query
- Bucketing Data

Honds-On

1. **Running map reduce program**
2. **Creating Static, Dynamic Partition and temporary table in hive**
3. **Loading data into tables**
4. **Running queries and storing the results**
5. **ORC file , Parquet file example**

- JOIN Statements
    - Inner JOIN
    - Join Optimizations
    - LEFT OUTER JOIN
    - OUTER JOIN Gotcha
    - RIGHT OUTER JOIN
    - FULL OUTER JOIN
    - LEFT SEMI-JOIN
    - Cartesian Product JOINs
    - Map-side Joins
- ORDER BY and SORT BY
- DISTRIBUTE BY with SORT BY
- CLUSTER BY
- Casting
- Queries that Sample Data
- Union All

Hive Views

- Creating Views to reduce query complexity
- Views that restrict data based on condition
- Lateral view and explode method

Hive Indexes

- Creating index
- Rebuilding the index
- Showing an index
- Dropping an index

Debugging and troubleshooting Hive queries

Customizing Hive File and Record Formats

- RC file

- Sequence File

Record Formats: SerDes

- CSV and TSV SerDes
- JSON SerDe
- Avro Hive SerDe