Expertise Training for Big Data (Hadoop + Spark)

Key Points of Our Training:

- ✓ A Single Course covers all the Hadoop components and apache Spark.
- ✓ Overall 60+ Assignments
- ✓ Java Refresher for Non Java Candidates
- ✓ No Pre-Configured VMs.
- ✓ 24 Hours SLA for email Support
- ✓ Refresher Classes
- ✓ Vendor Neutral. Using Apache versions.
- ✓ Offering Big Data Course not Hadoop Alone.
- ✓ Course taken by trainers from real time professionals who have taken for 800+ professionals

Course Outline

Introduction

Big Data (What, Why, Who) – 3++Vs – Overview of Hadoop EcoSystem and Apache Spark - Role of Hadoop and in Big data – Overview of other Big Data Systems – Who is using Hadoop – Hadoop integrations into Exiting Software Products - Current Scenario in Hadoop Ecosystem - Installation - Configuration - UseCases of Hadoop (HealthCare, Retail, Telecom)

HDFS

Concepts - Architecture – Data Flow (File Read , File Write)–Fault Tolerance - Shell Commands – Java Base API – Data Flow Archives – Coherency - Data Integrity – Role of Secondary NameNode – HDFS Federation – High Availability – HDFS Caching – HDFS SnapShots

YARN & MapReduce

Yarn Architecture – MR in YARN – Changes wrt MR1 – Data Flow (Map – Shuffle - Reduce) – MapRed vs MapReduce APIs - Programming [Mapper, Reducer, Combiner, Partitioner] –Writables – InputFormat – Outputformat - Streaming API using python – Inherent Failure Handling using Speculative Execution – Magic of Shuffle Phase –FileFormats – Sequence Files

Advanced Mapreduce Programming

Counters (Built In and Custom) – CustomInputFormat – Distributed Cache – Joins(MapSide, Reduce Side) – Sorting - Performance Tuning –GenericOptionsParser - ToolRunner – Debugging(LocalJobRunner)

Administration

Multi Node Cluster Setup using AWS Cloud Machines –Hardware Considerations –Software Considerations - Commands (fsck, job, dfsadmin) – Schedulers in Job Tracker - RackAwareness Policy - Balancing - NameNode Failure and Recovery - commissioning and Decommissioning a Node – Compression Codecs

HBase

Introduction to NoSQL – CAP Theorem – Classification of NoSQL – Hbase and RDBMS – HBASE and HDFS- Architecture (Read Path, Write Path, Compactions, Splits) - Installation – Configuration - Role of Zookeeper – HBase Shell - Java Based APIs (Scan, Get, other advanced APIs)– Introduction to Filters- RowKey Design - Map reduce Integration – Performance Tuning –What's New in HBase 0.98 – Backup and Disaster Recovery - Hands On

Hive

Architecture – Installation –Configuration – Hive vs RDBMS - Tables – DDL – DML – UDF – UDAF – Partitioning – Bucketing – MetaStore - Hive-Hbase Integration – Hive Web Interface – Hive Server(JDBC,ODBC, Thrift) – File Formats (RCFile - ORCFile) – Other SQL on Hadoop

Pig

Architecture –Installation - Hive vs Pig - Pig Latin Syntax –Data Types –Functions (Eval, Load/Store, String, DateTime) - Joins - Pig Server –Macros- UDFs- Performance - Troubleshooting – Commonly Used Functions

Sqoop

Architecture , Installation, Commands(Import , Hive-Import, EVal, Hbase Import, Import All tables, Export) – Connectors to Existing DBs and DW

Flume

Why Flume ? - Architecture, Configuration (Agents), Sources(Exec-Avro-NetCat), Channels(File,Memory,JDBC, HBase), Sinks(Logger, Avro, HDFS, Hbase, FileRoll), Contextual Routing (Interceptors, Channel Selectors) - Introduction to other aggregation frameworks

Oozie

Architecture, Installation, Workflow, Coordinator, Action (Mapreduce, Hive, Pig, Sqoop) – Introduction to Bundle – Mail Notifications

Apache Spark

Introduction to Apache Spark - Role of Spark in Big data - Who is using Spark - Installation of SparkShell and StandAlone Cluster – Configuration - RDD Operations (Transformations and Actions) - Cluster Components (Master , Workers, Executors) – Introduction to SparkSQL and Spark Streaming

SOLR**

Introduction to Information Retrieval - common usecases - Introduction to Solr and Lucene – Installation – Concepts (Cores,Schema , Documents, fields, Inverted Index,) - Configuration - CRUD operation requests and responses – Java Based APIs – Introduction to SolrCloud

Versions used in our Hadoop Training are

- Hadoop-.2.5.0 & Hadoop-1.2.1
- Hive -0.14.1
- Pig-0.14.0
- Oozie-4.0.1
- Sqoop-1.4.5
- HBase-0.94.0 & HBase-1.0.0
- Flume-1.5.0
- Spark – 1.3.0

** Covered in Self Paced Training