Big Data and Hadoop InfoSphere BigInsights Foundation (DW612)

Course ID#: 0370-427-ZZ-W

Hours: 21

Course Content

Course Description:

This two and one-half day hands-on training course is for those who want a foundation of IBM InfoSphere BigInsights. It will give you an overview of IBM's Big Data strategy as well as a more detailed information on Apache Hadoop. It presents concepts required by a system administrator to work with the Hadoop Distributed File System and concepts of MapReduce that are required by a developer. It gives an introduction to the scheduling capabilities of Hadoop and how to use Oozie to control workflows and use Flume to load data into HDFS.

This material has been updated to InfoSphere BigInsights 2.1 level.

After completing this course, you should be able to:

- Describe functions and features of InfoSphere BigInsights
- List the capabilities of Hadoop and HDFS
- Administer HDFS
- Describe the use of MapReduce
- Set up a Hadoop cluster
- Manage job execution
- Explain the Oozie workflows
- Describe some scenarios for loading data into HDFS
- DW72* - Programming for InfoSphere Streams V3 with SPL
- DW73* - InfoSphere Streams Administration
- DW64* - IBM InfoSphere BigInsights Analytics for Business
- Analysts
- DW65* - IBM InfoSphere BigInsights Analytics for Programmers

Audience:

System administrators and developers.

Prerequisites:

None, however, knowledge of Linux would be beneficial.

Big Data and Hadoop InfoSphere BigInsights Foundation (DW612)

Course ID#: 0370-427-ZZ-W

Hours: 21

Topics:

Module 1:
Introduction to Big Data

Lesson 1:

- System of Units / Binary System of Units
- The scale
- There is an explosion in data and real world events
- Some examples of Big Data
- ... And organizations need deeper insights
- Example: The perception gap surrounding social media
- The challenge: bring together a large volume and variety of data to find new insights
- Is there really a need for Big Data?
- Streams and oceans of information
- Big Data presents big opportunities
- Merging the traditional and Big Data approaches
- Enterprise information architecture
- IBM Big Data platform strategy
- Enterprise class
- Different BigInsights editions for varying needs
- InfoSphere Streams

Module 2:
An Introduction to InfoSphere BigInsights

Lesson 1:

- InfoSphere BigInsights open source components
- BigInsights: Value Beyond Open Source
- BigInsights Content
- BigInsights Content (cont.)
- What is Hadoop
- Open source programming
- Open source control
- Open source other
- Topic Summary

Lesson 2:

- InfoSphere BigInsights IBM Components
- Web-based Installation
- A rich management Big Data tool
- Running Applications from the Web Console
- BigInsights and Text Analytics
- BigInsights Text Analytics Development
- BigSheets – Spreadsheet – Style Analysis
- GPFS-FSO
- Performance Enhancements
- Topic Summary

Module 3:
Apache Hadoop and HDFS Overview

Lesson 1:

- Why Hadoop?
- How about Technology?
- How long it will take to read 1TB of data?
- Parallel Data Processing is the answer!
- What do we care about when we process data?
- Why Hadoop when we have relational databases?
- RDMS and Hadoop – complementary, not competing.
- Topic Summary

Lesson 2:

- Working with Hadoop
- HDFS - Hadoop Distributed File System
- Design principles of Hadoop
- More details about HDFS

- Hadoop system components overview
- Topic Summary

Lesson 3:

- MapReduce
- MapReduce programming abstraction overview
- NameNode
- NameNode directory structure
- Secondary NameNode
- DataNode
- JobTracker and TaskTrackers
- HDFS file blocks
- Storing file blocks into HDFS from client machine
- Rack Awareness (1 of 2)
- Rack Awareness (2 of 2)
- Topic Summary

Lesson 4:

- HDFS commands
- HDFS file commands
- File commands in HDFS (1 of 10)
- File commands in HDFS (2 of 10)
- File commands in HDFS (3 of 10)
- File commands in HDFS (4 of 10)
- File commands in HDFS (5 of 10)
- File commands in HDFS (6 of 10)
- File commands in HDFS (7 of 10)
- File commands in HDFS (8 of 10)
- File commands in HDFS (9 of 10)
- File commands in HDFS (10 of 10)
- Topic Summary

Lesson 5:

- Web Console Data Management
- Web Console Data View
- Working with Files and Directories
- Changing Permissions
- Hadoop Shell Command
- Application Status
- Workflows Tab
- Application Status
- Workflows Tab
- Application Running Status
- BigSheets
- BigSheets Workbooks
- Manipulation of Data in BigSheets
- Topic Summary

Module 4:
GPFS-FPO: Motivation

Lesson 1:

- GPFS-FPO: Architecture
- Locality Awareness
- Allows Applications to Define Own Logical Block Size
- Write Affinity: Allow Applications to Dictate Layout
- Pipelined Replication: Efficient Replication of Data
- Fast Recovery
- Hybrid Allocation: Treat Metadata and Data Differently
- Information Lifecycle Management (ILM)
- Comparison with HDFS and MapR
- BigInsights Interface to GPFS-FPO (1 of 3)
- BigInsights Interface to GPFS-FPO (2 of 3)
- BigInsights Interface to GPFS-FPO (3 of 3)
- BigInsights Interface to GPFS-FPO – URI Access
- Things to Note
- GPFS Cluster and File System Concepts
- Cluster Topology – Pool Stanza File
- Cluster Topology – NSD Stanza File
- GPFS – FPO – File System for BigInsights

Big Data and Hadoop InfoSphere BigInsights Foundation (DW612)

Course ID#: 0370-427-ZZ-W

Hours: 21

Module 5: BigInsights Web Console Security

Lesson 1:

- Installation Type
- File System
- Web Console Security
- Web Console Roles
- Assigning Groups to Roles
- Flat File Authentication
- LDAP or PAM Authentication
- Web Console Welcome
- Module Summary

Module 6: Introduction to MapReduce Programming

Lesson 1:

- MapReduce Overview
- MapReduce
- An SQL Example of MapReduce
- The Map Function
- Sort Phase
- The Reduce Function
- Combiner and Partition Functions
- Streaming and Pipes
- MapReduce example: Wordcount
- MapReduce co-location with HDFS
- MapReduce Processing
- MapReduce Processing (cont.)
- Speculative Execution
- Topic Summary

Lesson 2:

- MapReduce Programming
- MapReduce – a Tale of Two APIs
- MapReduce Anatomy (1 of 4)
- Basic Map Code
- MapReduce Anatomy (2 of 4)
- Basic Reduce Code
- MapReduce Anatomy (3 of 4)
- MapReduce Anatomy (4 of 4)
- Main()
- MapReduce Summary
- Topic Summary

Lesson3:

- MapReduce Programming using BigInsights
- Create a BigInsights Project
- Create a BigInsights Program
- Mapper Class
- Reducer and Driver Classes
- Generated Code
- Topic Summary
- Exercise Introduction

Module 7: Adaptive MapReduce

Lesson 1:

- Emerging Workload Patterns
- Adaptive MapReduce Features
- Workload and Resource Management Architecture
- Adaptive MapReduce Architecture
- Optimized Shuffling
- User Interface for Adaptive MapReduce
- Administrative Tasks

Module 8: Setup, Configuration, and Administration of a Hadoop Cluster

Lesson 1:

- Setup of Hadoop Clusters
- Starting Points
- What can be compressed in Hadoop?
- Should I use compression with Hadoop?
- Compression with BigInsights?

- Enabling Map Output Compression
- Enabling Job Output Compression
- Working with SEQ Files
- Capacity Calculations
- Capacity Planning
- Disks and File System (1 of 3)
- Disks and File System (2 of 3)
- Disks and File System (3 of 3)
- Hardware Considerations (1 of 3)
- Hardware Considerations (2 of 3)
- Hardware Considerations (3 of 3)
- Networking Considerations
- OS Considerations
- Topic Summary

Lesson 2:

- Configuration of Hadoop Clusters
- Configuration Management
- Configuration Files
- Preventing Configuration Property Override
- hadoop-env.sh Settings
- hdfs-site.xml settings
- hdfs-site.xml Settings (cont.)
- core-site.xml Settings
- core-site.xml Settings (cont.)
- mapred-site.xml Configuration (1 of 6)
- mapred-site.xml Configuration (2 of 6)
- mapred-site.xml Configuration (3 of 6)
- mapred-site.xml Configuration (4 of 6)
- mapred-site.xml Configuration (5 of 6)
- mapred-site.xml Configuration (6 of 6)
- Topic Summary

Lesson 3:

- Administration of Hadoop Clusters with BigInsights
- Setting Rack Topology (rack awareness)
- Example of Rack Awareness Script
- Setting Rack Topology (cont.)
- ibm-hadoop Properties
- ibm-hadoop Properties (cont.)
- Cluster Status
- Node Administration
- Balancer
- Safemode at Startup
- Safemode Commands
- Dashboards
- Dashboards (cont.)
- Topic Summary

Module 9: Overview of Oozie

Lesson 1:

- Oozie Workflows
- Oozie Workflows (1 of 2)
- Oozie Workflows (2 of 2)
- Action Nodes
- Action Nodes (cont.)
- Effect of the MapReduce APIs
- Control Flows at a High Level
- Control Flows Nodes (1of 2)
- Control Flows Nodes (2 of 2)
- Expression Language Functions
- Workflow EL Functions
- Hadoop EL Constants
- HDFS EL Functions
- Workflow Job
- Job Properties
- Topic Summary

Lesson 2:

- BigInsights Workflow Editor
- BigInsights Application Publishing
- Publishing an Application (1 of 5)
- Publishing an Application (2 of 5)
- Publishing an Application (3 of 5)
- Publishing an Application (4 of 5)

- Publishing an Application (5 of 5)
- Deploy the Application
- Schedule the Application
- Link Multiple Applications
- Link Output to Input
- Deploy the Linked Application
- Topic Summary
- Exercise Introduction

Module 10: Managing Job Execution

Lesson 1:

- FIFO Scheduler
- Job Execution
- Some Terminology
- FIFO Scheduler – First In First Out (Default)
- Priorities in FIFO
- Topic Summary

Lesson 2:

- Fair Scheduler
- FAIR scheduler – Pools Allocation
- FAIR scheduler – Pools
- FAIR scheduler – Minimum Share
- FAIR scheduler – Minimum Share, No Demand
- FAIR scheduler – Minimum Share Exceeds Slots
- FAIR scheduler – Minimum Share Less Than Fair Share (1 of 4)
- FAIR scheduler – Minimum Share Less Than Fair Share (2 of 4)
- FAIR scheduler – Minimum Share Less Than Fair Share (3 of 4)
- FAIR scheduler – Minimum Share Less Than Fair Share (4 of 4)
- FAIR scheduler – weights
- FAIR scheduler – weights example (1 of 4)
- FAIR scheduler – weights example (2 of 4)
- FAIR scheduler – weights example (3 of 4)
- FAIR scheduler – weights example (4 of 4)
- Multiple Jobs per Pool
- Configuring FAIR scheduler
- Example of an Allocation File
- BigInsights Scheduler
- InfoSphere BigInsights Scheduler Priorities
- Topic Summary

Module 11: Moving Data into Hadoop

Lesson 1:

- Loading Scenarios
- Load Scenarios
- Data is at Rest
- Data in Motion
- Streaming Data
- Solution if Data is from a Data Warehouse
- Load Solution using Flume
- Data from a Web Server
- Topic Summary

Lesson 2:

- Workings of Sqoop
- Overview of Sqoop
- Sqoop Connection
- Sqoop Import
- Sqoop Import Examples
- Sqoop Exports
- Sqoop Exports Examples
- Additional Export Information
- Topic Summary

Lesson3:

- Workings of Flume
- How Flume Works (1 of 3)
- How Flume Works (2 of 3)
- How Flume Works (3 of 3)
- Consolidation
- Replicating and Multiplexing
- Topic Summary

Lesson 4:

- Configuration of Flume
- Configuration
- Configuration Example
- Flume Sources
- Interceptors
- Flume Sinks
- Flume Channels
- Flume Channel Selectors
- Configuration Details – Components
- Configuration Details – Properties
- Configuration Details – Bindings
- Flume Example
- Working with an Agent
- Topic Summary
- Exercise Introduction