Introduction to Big Data

Modality: Virtual Classroom

Duration: 3 Days

About this course:

Introduction to Big Data is an intermediate level, Data Science training based course that allows students to learn how to leverage big data analysis tools and techniques to facilitate a better business decision-making. Furthermore, students also acquire hands-on knowledge on storing data in order to regulate efficient processing and analysis, and acquire the expertise to store, manage, process, and analyze large amounts of unstructured data and develop a relevant data lake.

Course Objectives:

- Store, manage, and analyze the unstructured data sets
- Choose the right big data stores covering disparate data sets
- Process large data sets through Hadoop to acquire value
- Query large data sets in almost real time through Pig and Hive
- Craft and execute a big data strategy for a business

Prerequisite:

- A sound expertise of the Microsoft Windows platform

Course Outline:

Defining Big Data

- The four dimensions of Big Data: volume, velocity, variety, veracity
- Introducing the Storage, MapReduce and Query Stack

Delivering business benefit from Big Data

- Establishing the business importance of Big Data
- Addressing the challenge of extracting useful data
- Integrating Big Data with traditional data

Storing Big Data

Analyzing your data characteristics

- Selecting data sources for analysis
- Eliminating redundant data
- Establishing the role of NoSQL

Overview of Big Data stores

- Data models: key value, graph, document, column?family
- Hadoop Distributed File System
- HBase
- Hive
- Cassandra
- Hypertable
- Amazon S3
- BigTable
- DynamoDB
- MongoDB
- Redis
- Riak
- Neo4J

Selecting Big Data stores

- Choosing the correct data stores based on your data characteristics
- Moving code to data
- Implementing polyglot data store solutions
- Aligning business goals to the appropriate data store

Processing Big Data

Integrating disparate data stores

- Mapping data to the programming framework
- Connecting and extracting data from storage
- Transforming data for processing
- Subdividing data in preparation for Hadoop MapReduce

Employing Hadoop MapReduce

- Creating the components of Hadoop MapReduce jobs
- Distributing data processing across server farms
- Executing Hadoop MapReduce jobs
- Monitoring the progress of job flows

The building blocks of Hadoop MapReduce

- Distinguishing Hadoop daemons
- Investigating the Hadoop Distributed File System
- Selecting appropriate execution modes: local, pseudo?distributed and fully distributed

Handling streaming data

- Comparing real?time processing models

- Leveraging Storm to extract live events
- Lightning?fast processing with Spark and Shark

Tools and Techniques to Analyze Big Data

Abstracting Hadoop MapReduce jobs with Pig

- Communicating with Hadoop in Pig Latin
- Executing commands using the Grunt Shell
- Streamlining high?level processing

Performing ad hoc Big Data querying with Hive

- Persisting data in the Hive MegaStore
- Performing queries with HiveQL
- Investigating Hive file formats

Creating business value from extracted data

- Mining data with Mahout
- Visualizing processed results with reporting tools
- Querying in real time with Impala

Developing a Big Data Strategy

Defining a Big Data strategy for your organization

- Establishing your Big Data needs
- Meeting business goals with timely data
- Evaluating commercial Big Data tools
- Managing organizational expectations

Enabling analytic innovation

- Focusing on business importance
- Framing the problem
- Selecting the correct tools
- Achieving timely results

Implementing a Big Data Solution

- Selecting suitable vendors and hosting options
- Balancing costs against business value
- Keeping ahead of the curve