Course duration

- 3 days

Course Benefits

- Store, manage, and analyze unstructured data
- Select the correct big data stores for disparate data sets
- Process large data sets using Hadoop to extract value
- Query large data sets in near real time with Pig and Hive
- Plan and implement a big data strategy for your organization

Course Outline

1. Introduction to Big Data
    1. Defining Big Data
        1. The four dimensions of Big Data: volume, velocity, variety, veracity
        2. Introducing the Storage, MapReduce and Query Stack
    2. Delivering business benefit from Big Data
        1. Establishing the business importance of Big Data
        2. Addressing the challenge of extracting useful data
        3. Integrating Big Data with traditional data
2. Storing Big Data
    1. Analyzing your data characteristics
        1. Selecting data sources for analysis
        2. Eliminating redundant data
        3. Establishing the role of NoSQL
    2. Overview of Big Data stores
        1. Data models: key value, graph, document, column–family
        2. Hadoop Distributed File System
        3. HBase
        4. Hive
        5. Cassandra
        6. Amazon S3
        7. BigTable
        8. DynamoDB
        9. MongoDB
        10. Redis
        11. Riak
        12. Neo4J
    3. Selecting Big Data stores
        1. Choosing the correct data stores based on your data characteristics

2. Moving code to data
3. Messaging with Kafka
4. Implementing polyglot data store solutions
5. Aligning business goals to the appropriate data store

3. Processing Big Data
    1. Integrating disparate data stores
        1. Mapping data to the programming framework
        2. Connecting and extracting data from storage
        3. Transforming data for processing
        4. Subdividing data in preparation for Hadoop MapReduce
    2. Employing Hadoop MapReduce
        1. Creating the components of Hadoop
        2. MapReduce jobs
        3. Executing Hadoop
        4. MapReduce jobs
        5. Monitoring the progress of job flows
    3. The building blocks of Hadoop MapReduce
        1. Distinguishing Hadoop daemons
        2. Investigating the Hadoop Distributed File System
        3. Selecting appropriate execution modes: local, pseudo–distributed and fully distributed
        4. Accelerating process with Spark
    4. Handling streaming data
        1. Comparing real–time processing modelsLeveraging Storm to extract live events
        2. Leveraging Spark Streaming to extract live events
        3. Combining streaming and batch processing in a Lambda architecture
4. Tools and Techniques to Analyze Big Data
    1. Abstracting Hadoop MapReduce jobs with Pig
        1. Communicating with Hadoop in Pig Latin
        2. Executing commands using the Grunt Shell
        3. Streamlining high–level processing
    2. Performing ad hoc Big Data querying with Hive
        1. Persisting metadata in the Hive Metastore
        2. Performing queries with HiveQL
        3. Investigating Hive file formats
    3. Creating business value from extracted data
        1. Mining data with Mahout
        2. Visualizing processed results with reporting tools
        3. Querying in real time with Impala
5. Developing a Big Data Strategy
    1. Defining a Big Data strategy for your organization
        1. Establishing your Big Data needs
        2. Meeting business goals with timely data
        3. Evaluating commercial Big Data tools
        4. Managing organizational expectations
    2. Enabling analytic innovation

    1. Focusing on business importance
    2. Framing the problem
    3. Selecting the correct tools
    4. Achieving timely results
3. Implementing a Big Data Solution
    1. Selecting suitable vendors and hosting options
    2. Balancing costs against business value
    3. Keeping ahead of the curve

Class Materials

Each student will receive a comprehensive set of materials, including course notes and all the class examples.

Class Prerequisites

Experience in the following *is required* for this Hadoop class:

- Working knowledge of the Microsoft Windows platform and basic database concepts.