Big Data Fundamentals

by William McKnight and Jake Dolezal

Module 0. About the Course (8 min)

Module 1. Big Data Definition (34 min)

- *Overview*
- *Big Data Introduction*
  - Business Models are Under Constant Threat
  - What Happens in an Internet Minute?
  - Fueled by Disruptive Technology Factors
  - Big Data is Additive to Systems Data
  - Every Industry Can Leverage Big Data
  - It is Not Easy
  - Top Performers vs. Average
  - Why the Sudden Explosion of Interest
  - Sensor Data Drives Big Data
  - Big Data is Unstructured
  - What's Needed for Big Data
  - Example: Optimizing Transit Duration
  - Example: Aircraft Engines
  - Summary
- *Big Data Technology*
  - Scale Up vs. Scale Out
  - Clusters Change the Game
  - Bringing Data to Processing
  - Bringing Processing to Data
  - ACID
  - Why Not Traditional Platforms for Big Data
- *Enablers for Big Data*
  - Data Integration
  - Data Virtualization
  - Infrastructure Strategy Including Cloud

Module 2. Big Data Drivers (28 mins)

- *Value Density of Data*
- *Before Data was Big…*
- *Once Big Data Grew, Value was Realized*
- *Data is too Valuable to Discard*
- *Data is too Valuable to Ignore*
- *Focus Before Big Data*
- *Focus After Big Data*
- *Performance/Workload Optimization*
- *Cost of Storage*
- *Other Cost Drivers*
- *Analytic Need*
- *Implication for IT Skills*

Module 3. Big Data in the Enterprise (21 mins)

- *The Great Database Thaw*
- *Data Access in the Modern Enterprise*
- *Marz's Lambda Architecture*

- *Row vs. Columnar Stores*
- *In-Memory*
- *Big Data BI & Analytics*
- *Leveraging Hadoop for Analytics*

Module 4. Hadoop Ecosystem (40 mins)

- *Hadoop Overview*
  - Introduction
  - Hadoop, MapReduce and Big Data
  - Who Uses Hadoop
  - Hadoop Nodes
  - Data Node Specification
  - HDFS Block Placement Example
  - File System Summary
  - MapReduce
  - Sample MapReduce Code
- *Hadoop Distributions*
  - What Included with a Distribution Subscription
  - A Hadoop Distribution
  - Hortonworks
  - Cloudera
  - Other Distributions
- *Hadoop Framework*
  - PID – Data Querying
  - Hive – Data Querying
  - HBase – Hadoops NoSQL Database
  - Hcatalog – Metadata Management
  - Mahout – Machine Learning
  - YARN – Resource Management
  - Sqoop – Data Movement
  - Flume – Data Streaming
  - Oozie – Scheduling
  - Spark – Fast Data Query
  - Shark – Hive for Spark
  - BigQuery – Interactive Analysis
  - Cloudera Impala – Data Analytics
  - Hadoop and Data Lifecycle Management
  - Summary

Module 5. NoSQL (31 mins)

- *NoSQL "Schemaless" Data Modeling*
- *NoSQL Heartburn*
- *Key-Value Stores*
- *Key-Value Simple Example*
- *Document-Oriented Database*
- *Document Simple Example*
- *Graph Oriented Database*
- *Graph (Fairly) Simple Example*
- *Stream Processing Engines*

- *Stream Processing Example*
- *NewSQL*

Module 6. Enterprise Architecture with Big Data (45 mins)

- *Module Overview*
- *Modern Components of Information Architecture*
  - Not One Size Fits All
  - Performance is the Top Issue
  - Columnar Databases
  - Data Appliances
  - Hardware Perspective
  - The Relational Database Page
  - The No-Reference Architecture
- *ETL with Big Data Systems*
  - Traditional ETL
  - ETL Alternatives
- *Analytic Patterns with Hadoop*
  - Hadoop and Analytics
  - Hadoop as Distribution Center
  - Hadoop as and Additional Data Store
  - Hadoop on a Data Warehouse Appliance
  - Analytics on Hadoop
  - Parallel DB vs. Hadoop Systems
- *Where Do We go from Here?*
  - The Big Data Challenge
  - What Gives the CIO Heartburn about Big Data
  - What Will Motivate IT to Adopt Big Data?