Must-Have Skills for Big Data Practitioners

Course Outline

Part 1 Big Data Fundamentals: Creating Value from Non-Traditional Data Sets

Big Data Basics

- What Is Big Data?
  - Definitions
  - Characteristics (3 V's plus 2)
  - Types of Big Data
- Why Big Data Analytics – Extending Advanced Analytics Capabilities
- Big Data Use Cases
  - Customer Understanding and Targeting
  - Business Process Optimization
  - Healthcare Advances
  - Law Enforcement and Public Safety
  - Sports Performance Improvement
  - Public Transportation and Infrastructure Advances
- Why Big Data Now? – The Driving Forces
- Kinds of Big Data – Data Variety
- Sources of Big Data
  - Web and Social Media
  - Machine to Machine
  - Other Sources (Big Transaction Data, Biometrics, Human Generated Data, Publicly Available Data, Legacy Documents)
- Working with Big Data – The Big Picture

Big Data Processes

- Business Case
  - Business Needs and Opportunities
  - Areas of Insight
  - Expected Outcomes
  - Business Value Projection
- Technical Case – Big Data Rationale
- Data Sourcing – Getting Big Data
- Data Preparation and Storage
  - Data Selection
  - Data Cleansing
  - Data Integration
  - Data Reduction
- Big Data Analytics
  - Problem Framing
  - Analytic Purpose
  - Analytic Modeling
  - Data Visualization

- Consumption and Application

Big Data Architecture

- The Role of Architecture
  - What and why
- Data Architecture
  - Data Storage
  - Data Access
  - Data Analysis
  - Data Consumption
- Process Architecture
  - Data Governance Processes
  - Data Integration and Quality
- Analytics Architecture
  - Machine Learning
  - Predictive Analytics
  - Prescriptive Analytics
  - Descriptive Analytics
  - Reporting
- Technology Architecture
  - Search and Visualization
  - Data Management and Data Access
  - Hadoop and NoSQL
- Big Data ... Big Architecture – Summary

Big Data Technology

- The Technology Landscape – Overview
- Infrastructure
  - Databases
  - Development and Deployment Environment
- Analytics – Data Analysis
- Data Sources – Big Data Providers
- The Core Technologies
  - MapReduce
  - Hadoop

Getting Started with Big Data

- Readiness Assessment – Check Your Position
- Planning and Preparation – Charting the Course
- Execution – Navigating the Course
- Post-Project Activities – At the Destination
- Best Practices – Lessons Learned
- Mistakes to Avoid – More Lessons Learned

Summary and Conclusion

- Summary of Key Points – A Quick Review
- References and Resources – To Learn More

Part 2 Big Data Roadmap

Big Data and Business Case

- What is Big Data? – Creating a definition
- What will it solve? – Potential solutions for an organization (this will be specific for onsite courses)
- Business Users and Big Data
- Understanding Roles and Skiils – What does the business user bring to the table? Why should IT look to getting the business users own and drive the initiative?
- Business User Ownership – What does this entail?
- Challenges – Issues and Risks
- Building the Business Case
- Components of Big Data Business Case
- How to build the Appropriate Business Case
- Next Generation of Business Intelligence
- Analytics and Metrics – What do we derive new?
- Visualization Requirements – What are the changes and associated challenges?
- Mashups – Understanding multi-dimensional data management. Metadata is critical and why?

Semantics and Ontologies

- Introduction to Semantic Frameworks – Future of Visualization and Analytics
- Understanding Semantic integration for Big Data – Where and How? Business Benefits.
- Using Ontologies for Metadata Management – Case Study
- Managing Business Rules for Processing - Case Study

Big Data and the Data Warehouse

- The New Landscape
- What Can We Solve
- How to Assess and Manage Data For Today and Future
- Technology overview
- Hadoop, NoSQL, Cassandra, Big Query, Drill, Redshift, AWS (S3, EC2)
- Programming with MapReduce
- Understanding analytical requirements

Self-Service Discovery Platforms

- Challenges of Data Management and Processing
- MDM, Metadata and More – Have we moved over this?

Workloads

- Data Management
- Infrastructure Limitations

Next-Generation Data Warehouse

- Solution architectures
- The three s's: scalability, sustainability, and stability
- People skills
- Critical success factors

Big Data Road map

- Building A Road map
- Risks and Mitigations
- Business Driven Objectives
- Solving A Million Dollar Puzzle
- Readying The Organization

Part 3 Machine & Deep Learning: Delivering Insights from Big Data

Chaos Theory

- Definition
- Characteristics of Systems
- Chaos Theory Applications

Game Theory

- Definition
- Players
- Strategies
- Payoffs
- Equilibrium Concepts
    - Nash Equilibrium
    - An Illustrative Advertising Game
    - Dominant Strategies and Nash Equilibria
    - Hotelling's Beach
    - Television Scheduling

Machine Talk

- Techniques
    - kNN algorithm
    - Winnow algorithm
    - Naïve Bayes classifier
    - Decision trees
    - Reinforcement learning (Rocchio algorithm)
    - Genetic algorithm
- Neural Networks
    - Input layer, hidden layer, output layer
    - Forward pass
    - Back Propagation

- Classification
    - Connections
    - Learning
- Hopfield Network
- Self-Organization
- Self-Organizing Networks
- How Is CNN/ConvNets different?
- LeNet-5 Architecture
- AlexNet Architecture - ImageNet 2012
- Case Study: GoogLeNet

Search Algorithm

- Types of Search Algorithm
    - PageRank Algorithm
    - Penguin Algorithm
    - Panda Algorithm
    - Hummingbird Algorithm
- Differences between old and new search engine methods
- Knowledge Graph
- Applications of Search Algorithm
    - Enterprise Search
    - Target Marketing
    - Performance Optimization
    - List Optimization
    - Indexing

Google TensorFlow

- Definition
- Data Flow Graph
- Google TensorFlow Basic Elements
    - Variable
    - Operation
    - Session
    - Placeholder
    - TensorBoard
- TensorBoard: Visual Learning
- MNIST Dataset
- TensorBoard
- Applications
    - Natural Language Processing
    - Image Processing
    - Geo-Coding Processing
    - Gaming Simulators

  - Real-World Game Data Processing
  - Intermittent Recurrent Data Processing

Machine Learning Implementations

- R
- IBM Watson
- Microsoft
- Oracle Advanced Analytics DB Option
- Enterprise Solutions

Machine Learning with Hadoop

- Tools for Data Preparation/Feature Engineering
- Apache Mahout
- More Machine Learning Interfaces for Hadoop

Visualization

- Key Points
  - Mashups
  - Lat-Long Processing
  - Semantic Processing
  - Machine Learning
  - Algorithms
  - Iterative Processing of Data
- Polymaps
- Data-Driven Documents
- Apache Tajo
  - Architecture
  - Query Federation
  - Storage and Data Format Support

Presto

- History
- Architecture
- Connectors
- Extensibility – plug-ins

Workshop

- Working with Your People, Projects, Processes, and Data

- Choose from a provided list of topics for a tailored workshop. Workshop topic selection to be determined during pre-course discussion and planning with instructor. *(It is recommended that you select two topics for approximately 4.5 hours of workshop activity.)*