Course Syllabus Part I
DSC 650 Big Data

3 Credit Hours

Course Description

This course covers the fundamentals of data infrastructure and how technologies fit together to form a process, or pipeline, to refine data into usable datasets. This course focuses on building a predictive modeling pipeline used by the various types of projects that are called, “big data.”

Course Prerequisites

Recommend DSC 540

Course Objectives

Students who successfully complete this course should be able to:

1. Explain big data architecture and the engineering trade-offs of different data storage and data processing paradigms
2. Process real-time data streams from multiple input sources
3. Integrate datasets from multiple disparate sources and systems using batch and real-time data processing
4. Construct data processing and machine learning pipelines using directed acyclic graphs (DAG) workflows

Grading Scale

93 – 100% = A	87 – 89% = B+	77 – 79% = C+	67 – 69% = D+
90 – 92% = A-	83 – 86% = B	73 – 76% = C	63 – 66% = D
	80 – 82% = B-	70 – 72% = C-	60 – 62% = D-
			0 – 59% = F

Topic Outline

1. Data Models
    A. Fact-based
    B. Graph schemas
    C. Relational data
    D. Document models
    E. Schema on-read vs. schema on-write
    F. Unstructured data

© 2016 Bellevue University. Unless otherwise noted, the contents of this course, including the text and images and how they are arranged and presented, are owned by Bellevue University and protected by U.S. Copyright law. You may use this course content for your own personal, educational, informational and non-commercial use.

2. Data Storage
    A. Distributed file systems
    B. Immutable data
    C. Column-oriented vs. row-oriented storage
    D. Serialization, compression, and data types
3. Batch Data Processing
    A. MapReduce paradigm
    B. Batch processing pipelines using DAGs
    C. Joins and aggregations
4. Realtime Views
    A. CAP Theorem
    B. Scalable big data stores
    C. Caching and data expiration
5. Stream Processing
    A. Queues, sinks, and sources
    B. Structured Streaming
    C. Stateful processing
    D. Micro-batch processing
6. Analytics and Machine Learning
    A. Classification
    B. Regression
    C. Recommendations
    D. Graph Analytics

© 2016 Bellevue University. Unless otherwise noted, the contents of this course, including the text and images and how they are arranged and presented, are owned by Bellevue University and protected by U.S. Copyright law. You may use this course content for your own personal, educational, informational and non-commercial use.