College of San Mateo
Official Course Outline

1. **COURSE ID:** CIS 364 **TITLE:** From Data Warehousing to Big Data
   Units: 4.0 units **Hours/Semester:** 48.0-54.0 Lecture hours; 48.0-54.0 Lab hours; and 96.0-108.0 Homework hours
   Method of Grading: Grade Option (Letter Grade or Pass/No Pass)
   Recommended Preparation:
   Completion of or concurrent enrollment in CIS 132.

2. **COURSE DESIGNATION:**
   Degree Credit
   Transfer credit: CSU; UC

3. **COURSE DESCRIPTIONS:**
   Catalog Description:
   Introduction to data warehousing architecture, data extraction, management, and load. Also covered are metadata management, dimensional modeling, data aggregation, data mining and Business Intelligence. Both SQL and NoSQL databases will be employed. Introduction to Big Data architecture, technologies and analytics. Selection, processing and querying of Big Data stores for disparate data sets are also covered. Other topics such as Cloud computing, security management, machine learning, Agile methodology and Big Data tools will be introduced.

4. **STUDENT LEARNING OUTCOME(S) (SLO'S):**
   Upon successful completion of this course, a student will meet the following outcomes:
   1. Determine the best data warehouse architecture using proven analytical modeling concepts
   2. Design and develop a data warehouse and model dimensions for it
   3. Query and manage the data warehouse
   4. Extract, transform and load operational data
   5. Define and describe Big Data and its role
   6. Give examples of Big Data usage in areas such as science and data warehouse augmentation
   7. Create an advanced project using Big Data Analytics and tools

5. **SPECIFIC INSTRUCTIONAL OBJECTIVES:**
   Upon successful completion of this course, a student will be able to:
   1. Determine the best data warehouse architecture using proven analytic modeling concepts
   2. Design and develop a data warehouse and model dimensions for it
   3. Query and manage the data warehouse
   4. Extract, transform and load operational data
   5. Define and describe Big Data and its role
   6. Give examples of Big Data usage in areas such as science and data warehouse augmentation
   7. Create an advanced project using Big Data Analytics and tools

6. **COURSE CONTENT:**
   Lecture Content:

1. Basic Elements of the Data Warehouse

A. Source System
B. Data Staging Area
C. Presentation Server
D. Dimensional Model
E. Business Process
F. Big Data
G. Apply Data Preprocessing Techniques for Cleaning, Integration, Reduction and Transformation of Data.
H. DataMart/Data Warehouse
I. Operational Data Store (ODS)
J. OLAP (On-Line Analytic Processing)
K. ROLAP (Relational OLAP)
L. MOLAP (Multidimensional OLAP)

2. Project Management and Requirements

A. The Business Dimensional Lifecycle
  a. Lifecycle Evolution
  b. Lifecycle Approach
B. Project Planning and Management
  a. Define and Plan the Project
  b. Collect the Requirements
C. Prepare and Publish the Requirements Deliverables

3. Data Design

A. Dimensional Modeling
B. The Data Warehouse Bus Architecture
C. Basic Dimensional Modeling Techniques
  a. Fact Tables and Dimension Tables
  b. Foreign Keys, Primary Keys, and Surrogate Keys
  c. Additive, Semiadditive, and Nonadditive Facts
D. Extended Dimension Table Designs
  a. Many-to-Many Dimensions
  b. Many-to-One-to-Many Traps
E. Extended Fact Table Designs
F. Build Dimensional Models

4. Data Warehouse Architecture

A. Architectural Framework
B. Logical Models and Physical Models
C. Back Room Technical Architecture
  a. Back Room Data Stores
  b. Back Room Services
    i. Extract Services
    ii. Data Transformation Services
    iii. Data Loading Services
  c. Backup and Archive Planning
  d. Architecture for the Front Room
    i. Front Room Data Stores
    ii. Front Room Services for Data Access
      a. Warehouse Browsing
      b. Access and Security Services
      c. Activity Monitoring Services
      d. Query Management Services

5. Security Management in a Data Warehouse Environment

A. Security: Vulnerabilities
  a. Physical Assets
  b. Information Assets: Data, Financial Assets, and Reputation
  c. Software Assets
  d. Network Threats
B. Security: Solutions
  a. Routers and Firewalls
  b. The Directory Server
  c. Encryption

6. Introduction to Big Data

A. Defining Big Data
B. The four dimensions of Big Data: volume, velocity, variety, veracity
C. Integrating Big Data with traditional data
D. Storing Big Data

E. Overview of Big Data stores
F. Data models: key value, graph, document, column–family
G. Hadoop Distributed File System

7. Processing Big Data

A. Integrating disparate data stores
B. Mapping data to the programming framework
C. Connecting and extracting data from storage
D. Transforming data for processing
E. Subdividing data in preparation for Hadoop MapReduce
F. Creating the components of Hadoop MapReduce jobs
G. Executing Hadoop MapReduce jobs

Lab Content:

1. Project Management Requirements
2. Data Design and Build Dimensional Models
3. Develop Data Warehouse Architecture
4. Implement the Data Warehouse and Query it
5. Manage Security in a Data Warehouse Environment
6. Select the correct Big Data stores for disparate data sets
7. Process large data sets using Hadoop to extract value
8. Query large data sets in near real time with Pig and Hive
9. Integrate key Big Data components to create a Big Data platform
10. Load unstructured data into Hadoop Distributed File System
11. Query Hadoop MapReduce jobs using Hive
12. Use Data Visualization tools

7. **REPRESENTATIVE METHODS OF INSTRUCTION:**

Typical methods of instruction may include:

A. Lecture
B. Lab
C. Activity
D. Discussion
E. Other (Specify): The course will include the following instructional methods as determined appropriate by the instructor: Lecture will be used to introduce new topics; Teacher will model problem-solving techniques; Class will solve a problem together, each person contributing a potential "next step"; Students will participate in short in-class projects (in teacher-organized small groups) to ensure that students experiment with the new topics in realistic problem settings; Teacher will invite questions AND ANSWERS from students, generating discussion about areas of misunderstanding; Teacher will create and manage an Internet conference for discussion of course topics; and Students will work in small groups to solve significant programming assignments.

8. **REPRESENTATIVE ASSIGNMENTS**

Representative assignments in this course may include, but are not limited to the following:

Writing Assignments:

The primary writing opportunity for students in this course is documentation supporting their lab and programming projects. This includes both technical documentation and end-user documentation. The technical documentation describes the problem to be solved, the scope of the project, an overview of the solution, and any limitations of the solution. User documentation will be provided to the client.

Reading Assignments:

Weekly textbook readings support all learning objectives.

Other Outside Assignments:

Weekly exercises from the textbook and lab/database programming assignments comprise the majority of the assignments. The lab and data warehouse design and development assignments support all learning objectives. In addition, students will create several substantial data warehouse and Big Data programming projects consisting of 500-600 lines of code.

9. **REPRESENTATIVE METHODS OF EVALUATION**

Representative methods of evaluation may include:

A. Class Participation
B. Class Work
C. Exams/Tests
D. Group Projects
E. Homework
F. Lab Activities
G. Oral Presentation
H. Projects
I. Quizzes
J. Written examination
K. Bi-weekly quizzes to provide feedback to students and teacher; (Short answer--from textbook material) Assessment of student contributions during class discussion and project time; Individual database development assignments to assess objectives 4-5; Midterm and Final exams; and (Short answer (textbook material), general problem solving (similar to in-class work), short program segments (similar to database development assignments) Assessment of group participation on course projects, including peer-assessment of participation and contribution to the group effort.

10. **REPRESENTATIVE TEXT(S):**

Possible textbooks include:

A. Gorelik. *The Enterprise Big Data Lake: Delivering on the Promise of Hadoop and Data Science in the Enterprise* , 1st ed. O'Reilly, 2019
B. Kleppmann. *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*, 1st ed. O'Reilly, 2017
C. Morley. *Data Science Design Patterns*, 1st ed. Addison-Wesley, 2019
D. Kane. *Ultimate Big Data Application Development*, 1st ed. Packt Publishing, 2019
E. Michael Mannino. *Database Design, Application Development & Administration*, 7th ed. ed. Chicago Business Press, 2019

Origination Date: November 2020
Curriculum Committee Approval Date: January 2021
Effective Term: Fall 2021
Course Originator: Mounjed Moussalem