Northeastern University
College of Engineering

Special Topics: Data Integration DAMG7374

Course (Future name): *Hybrid Data Integration for Data Engineering, Machine Learning, BI, Analytics and Data Warehousing*

Course Information

Course Title: Special Topics (DT): Data Integration
Course Number: DAMG7374
Term and Year: 2022 Summer
Credit Hour: 4

Instructor Information

Richard Sherman, ri.sherman@northeastern.edu

Course Description

Data engineering and data integration are critical activities for data science, analytics, business intelligence, data lake and data warehousing along with an increasing number of software applications. All these initiatives require data to be ingested, transformed, integrated, curated, cleansed, and modeled to enable analytics. This course explores the fundamental concepts and provides hands-on assignments applying the concepts to real-world scenarios.

The course explores hybrid data integration technologies such as data pipelines, data streaming, data preparation, Integration Platform as a Service (iPaaS), ETL (Extract, Transform & Load) and ELT (Extract, Load & Transform) implemented on cloud, on-premises, and hybrid environments. The course examines the fundamental concepts regarding data architecture; integration use cases; integration processes, workflows, techniques, and best practices; data quality; and data governance.

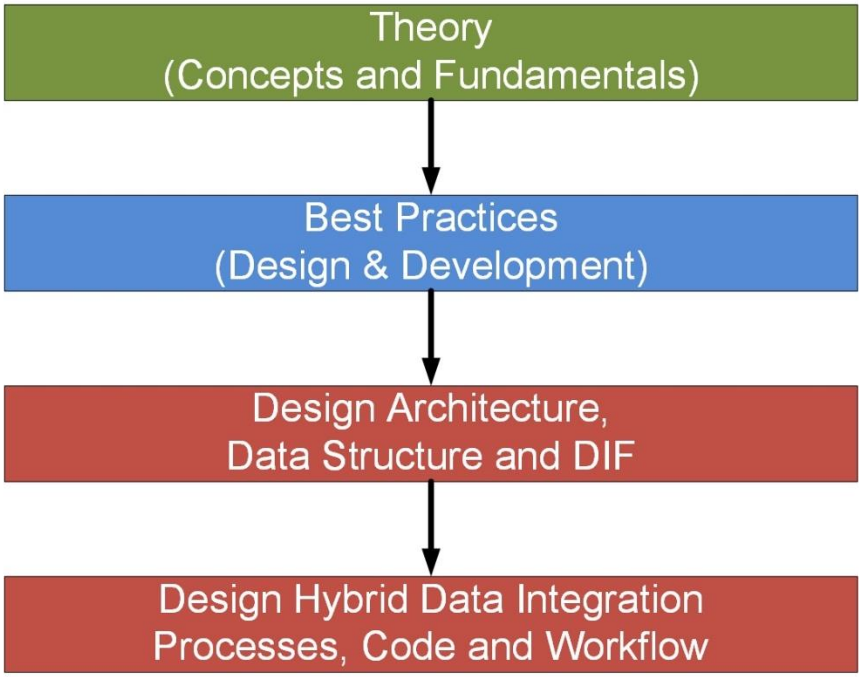
Students acquire hands-on development experience using various data sources (structured, semi-structured and unstructured) and file formats.

Learning Outcomes

- Understand concepts and theory regarding integration use cases, integration processes, integration personas, data quality, data governance and data management.
- Work with real-world integration use cases that present data silos, data fragmentation, data inconsistency (hint: the underlying causes are not technology nor new data types) and data quality challenges. Assignments and projects contain data that is dirty, inconsistent, and incomplete from various data silos with different contexts, granularities, and data structures.
- Examine architectures (data, information, technology, and product) needed to design, develop, and implement a Hybrid Integration Platform (HIP) in cloud, on-premises and hybrid environments.
- Review various integration applications technologies in the context of use cases and the P's (above).
- Leverage real-world data and integration case studies to learn how to apply theory and best & pragmatic practices with various technologies to implement Hybrid Integration Platform (HIP) solutions providing businesses with consistent, conformed, clean, comprehensive, and current (information 5 C's)

Learning approach

- Learn the concepts, fundamentals, and best practices
- Learn how to implement above with technology and tools (below)
- Implement solutions through your assignments, workshops, and projects to understand


Theory
(Concepts and Fundamentals)
Best Practices
(Design & Development)
Design Architecture,
Data Structure and DIF
Design Hybrid Data Integration
Processes, Code and Workflow

Lecture & Workshop Topics:

- Hybrid Data Integration Categories
    - Data pipelines
    - Data streaming or event data delivery
    - ELT (Extract, Load & Transform) vs ETL (Extract, Transform & Load))
    - Integration Platform as a Service (iPaaS)
    - Enterprise Service Bus (ESB) & Message-oriented middleware (MOM)
    - API management
    - Data Preparation
- Cloud Data Architecture
    - Cloud Databases & Files
    - Cloud Data Warehousing, Data Lakes & Analytical Data Architecture (ADA)
    - Cloud Data Integration
- Analytical Data Architecture (ADA)
    - Recap (or Review via recorded session)
    - Cloud, on-premise and hybrid implementations
- Integration Needs
    - Integration domains: Applications, Data, Processes, Internet of Things (IoT), Business to Business (B2B)
    - Integration use cases
- Data Management
    - Data quality & cleansing
    - Change data capture (CDC)
    - Slowly Changing Dimensions (SCD) with CDC to track historical changes
    - Master Data Management (MDM)
    - Data Governance & Analytical Governance
    - Error processing and recovery for above
- Designing & implementing a Hybrid Integration Platform (HIP)
    - Determining use cases and requirements
    - Identifying and defining the applicable architectures, processes, and data structures
    - Designing the architectures
    - Creating and operating a Data Integration Framework (DIF)
        - Data integration process, standards, and best practices
        - Data Ops (data operations), Process metadata and tagging

Tools used in this course:

- Data pipeline
    - Fivetran (Cloud)
    - Talend Pipeline Manager (Cloud)
    - Talend Stitch (Cloud)

- Data Integration (ETL, ELT)
    - Talend Data Fabric (Data Integration, DQ, MDM)
    - Microsoft Data Factory (Cloud) & Microsoft SSIS

- Data Preparation:
    - Alteryx or Trifacta - to-be-determined (TBD)

- Streaming: to-be-determined (TBD)

- Databases (data sources and targets):
    - Microsoft Azure SQL & Cosmos, Google BigQuery & BigTable, Oracle Cloud ADW
    - Microsoft SQL Server, Oracle, MySQL, PostgreSQL, Mongo
    - Data lakes, files (json, csv, tsv)

- Cloud Platform (data sources and targets):
    - Microsoft Azure, Google Cloud Platform (GCP), Oracle Cloud Infrastructure (OCI)

Lecture & Workshop Topics:

95-100%	A	87-89.9%	B+	77-79.9%	C+	69.9% or below	F
		84-86.9%	B	74-76.9%	C		
90-94.9%	A-	80-83.9%	B-	70-73.9%	C-		

Lecture & Workshop Topics:

Course Breakdown	Grade %
Assignments	10%
Workshops	10%
Team Projects	25%
Quizzes	15%
Exams	40%

Appendix:

Hybrid Data Integration for Data Engineering, Machine Learning, BI, Analytics and Data Warehousing

Problem statement:

Enterprises are increasingly adding, expanding, and altering data sources. Although much of the hype surrounds Big Data and machine learning, the proliferation of new data sources (and data silos) is primarily from adding applications that automate data collection and management of data in business processes and events that were previously either not captured or manually collected using spreadsheets or files. While gathering data from the data sources is often straightforward, enterprises struggle to integrate, cleanse, curate, transform and govern data to deliver a comprehensive and consistent view of their customers, prospects, patients, students, employees, products, services, suppliers and business processes or functions.

In addition to integrating data from a myriad of data source, information is data in context so even when data is collected it typically needs to be transformed for each context it will be used.

Too often integration is merely viewed as a data pipeline where data is collected, basic data transformations are applied and then the data is loaded into its target. Integration, cleansing, curation, transforming, consistency, and governing data involves applying extensive data and business rules, algorithms and relationships that vary based on context. Data management including metadata management, data & analytical governance and privacy & security needs to be appropriately applied during integration. Master data management (MDM), such as customer data integration (CDI), is an example of data management and needs to be applied when data needs to consistent, current, conformed, comprehensive and current (and historically defined).

Hybrid data integration encompasses the policies, practices (best & pragmatic), procedures (& standards), people and platform (architectures) needed to integrate data across data silos and provide information, i.e. data in context. Technology and tools need to be selected based on use cases (data, integration, analytics and personas) and apply the P’s.