DSC 208: Data Management for Analytics

Course Description and Objectives

Course Description

This course covers the principles, techniques, and tools of organizing, storing, querying, transforming, and using data for analytics and ML computations at scale. Students will learn the basics of data storage, acquisition, governance, organization, principles of the relational data model, relational algebra and its relationship to DataFrames, the Structured Query Language (SQL), relational database system features for faster querying and analytics, and basics of non-relational data systems. It will also cover major data quality issues and methodologies to clean data. An introduction to cluster and cloud computing, MapReduce and Spark, and the use of these tools and SQL to transform data at scale for ML feature engineering will be provided. Finally, methodologies to critically evaluate analytics results will be covered, including debugging ML results and reasoning about bias and fairness issues in the whole data science pipeline.

Course Objectives

At the end of the course, students should be able to:

- Explain the basic principles of managing large and complex datasets for analytics.
- Apply the relational model, relational algebra, SQL, and major relational DBMS features for data querying and analytics.
- Apply SQL and cluster programming techniques such as MapReduce and Spark to perform data transformations for ML at scale.
- Explain major data quality issues in data science pipelines and how to handle them.
- Analyze and evaluate tradeoffs of data science pipelines in terms of data quality, automation, scalability, accuracy, and fairness.