Course Outline for

Data Sciences/Optimization

Cloud Data Services – AWS, GCP and Azure with governance

Visualization and Charting

KG and Vector DBs

RDBMS and NoSQL

Contents

Stream 1

Cloud Data Services for AWS, GCP, and Azure with Governance and Catalog

Course Description:

This course provides a comprehensive overview of data services offered by the three major cloud providers: AWS, Google Cloud Platform (GCP), and Microsoft Azure. It covers essential data services, data governance, and data cataloging, emphasizing practical applications and hands-on experience. Students will learn to leverage cloud-based data services for storage, processing, and analysis while ensuring data governance and effective cataloging.

1: Introduction to Cloud Data Services

Overview of Cloud Computing

- Cloud Computing Models (IaaS, PaaS, SaaS)
- Benefits of Cloud Data Services

Introduction to AWS, GCP, and Azure

- Key Differences and Similarities
- Use Cases and Industry Adoption

Setting Up Cloud Accounts

- Creating AWS, GCP, and Azure Accounts
- Navigating Cloud Consoles

2: Data Storage Services

AWS Data Storage Services

- Amazon S3 (Simple Storage Service)
- Amazon EBS (Elastic Block Store)
- Amazon Glacier

GCP Data Storage Services

- Google Cloud Storage
- Google Persistent Disk
- Google Cloud Filestore

Azure Data Storage Services

- Azure Blob Storage
- Azure Disk Storage
- Azure File Storage

Hands-On Labs

- Setting Up and Managing Storage Services

3: Data Processing and Analytics

AWS Data Processing and Analytics

- Amazon RDS (Relational Database Service)
- Amazon Redshift
- Amazon EMR (Elastic MapReduce)

GCP Data Processing and Analytics

- Google BigQuery
- Google Cloud SQL
- Google Dataproc

Azure Data Processing and Analytics

- Azure SQL Database
- Azure Synapse Analytics
- Azure HDInsight

Hands-On Labs

- Deploying and Managing Data Processing Services

4: Big Data and Machine Learning

AWS Big Data and ML Services

- AWS Glue
- Amazon SageMaker
- Amazon Kinesis

GCP Big Data and ML Services

- Google Dataflow
- Google AI Platform
- Google Pub/Sub

Azure Big Data and ML Services

- Azure Data Factory
- Azure Machine Learning
- Azure Stream Analytics

Hands-On Labs

- Implementing Big Data and ML Solutions

5: Data Integration and ETL

AWS Data Integration Services

- AWS Glue

- AWS Data Pipeline

GCP Data Integration Services

- Google Cloud Dataflow
- Google Cloud Dataprep

Azure Data Integration Services

- Azure Data Factory

ETL Concepts and Tools

- Extract, Transform, Load (ETL) Processes
- Comparing ETL Tools

Hands-On Labs

- Creating ETL Pipelines

6: Data Governance Fundamentals

Introduction to Data Governance

- Importance of Data Governance
- Key Components of Data Governance

AWS Data Governance

- AWS Lake Formation
- AWS IAM (Identity and Access Management)
- AWS Audit Manager

GCP Data Governance

- Google Cloud Data Catalog
- Google Cloud IAM
- Google Cloud Audit Logs

Azure Data Governance

- Azure Purview
- Azure Active Directory
- Azure Policy and Blueprints

7: Data Security and Compliance

Security Best Practices

- Encryption and Key Management
- Identity and Access Management

AWS Security and Compliance

- AWS KMS (Key Management Service)
- AWS Shield and WAF

GCP Security and Compliance

- Google Cloud KMS
- Google Cloud Armor

Azure Security and Compliance

- Azure Key Vault
- Azure Security Center

Hands-On Labs

- Implementing Security Measures

8: Data Cataloguing and Metadata Management

Introduction to Data Cataloging

- Importance of Data Cataloging
- Key Features of Data Catalogs

AWS Data Cataloging

- AWS Glue Data Catalog

GCP Data Cataloging

- Google Cloud Data Catalog

Azure Data Cataloging

- Azure Data Catalog (Deprecated)
- Azure Purview

Hands-On Labs

- Creating and Managing Data Catalogs

9: Advanced Data Governance and Policy Management

Data Lineage and Ancestry

- Tracking Data Movement and Transformation

Policy Management

- Defining and Enforcing Data Policies

Data Quality Management

- Ensuring Data Accuracy and Consistency

Case Studies and Best Practices

- Implementing Governance Frameworks

10: Data Lake and Data Warehouse Solutions

Building Data Lakes

- Concepts and Architectures
- AWS Lake Formation
- Google Cloud Storage

- Azure Data Lake Storage

Data Warehousing

- Data Warehouse Concepts
- Amazon Redshift
- Google BigQuery
- Azure Synapse Analytics

Hands-On Labs

- Setting Up Data Lakes and Warehouses

11: Monitoring and Optimization

Monitoring Data Services

- AWS CloudWatch
- Google Cloud Monitoring
- Azure Monitor

Performance Optimization

- Scaling and Tuning Data Services
- Cost Management and Optimization

Hands-On Labs

- Implementing Monitoring Solutions

12: Capstone Project and Case Studies

Capstone Project

- Designing a Comprehensive Cloud Data Solution
- Implementing Data Governance and Cataloging
- Ensuring Security and Compliance
- Optimizing Performance and Cost

Case Studies

- Real-World Applications of Cloud Data Services
- Success Stories and Lessons Learned

Final Presentations

- Presenting Capstone Projects
- Peer Reviews and Feedback

Software and Tools:

- AWS Management Console
- Google Cloud Console
- Azure Portal

  - Data Integration and ETL Tools
  - Data Catalog and Governance Tools

Assessment:

- **Assignments:**
  - ly assignments on cloud data services and governance
- **Projects:**
  - Midterm project on a specific cloud platform
  - Capstone project integrating all three platforms
- **Exams:**
  - Midterm and final exams testing theoretical and practical knowledge

Prerequisites:

- Basic knowledge of cloud computing concepts
- Familiarity with data management principles
- Understanding of basic programming concepts (preferably in Python)

Real-Time Data Processing

Introduction to Real-Time Data Processing

- Overview of Real-Time Data Processing
- Use Cases and Applications
- Key Concepts: Stream Processing vs. Batch Processing

Apache Kafka

- Introduction to Kafka
  - Architecture and Components
  - Producers, Consumers, and Brokers
- Setting Up Kafka
  - Installation and Configuration
- Hands-On Lab: Building a Simple Kafka Producer and Consumer

Real-Time Data Processing with Cloud Services

AWS Kinesis

- Overview of AWS Kinesis
  - Kinesis Data Streams
  - Kinesis Data Firehose
  - Kinesis Data Analytics

- Setting Up Kinesis
    - Creating Data Streams
- Hands-On Lab: Stream Data with Kinesis

GCP Pub/Sub and Azure Event Hub

- Introduction to GCP Pub/Sub
    - Architecture and Components
    - Setting Up Pub/Sub
    - Hands-On Lab: Publish and Subscribe to Messages
- Introduction to Azure Event Hub
    - Architecture and Components
    - Setting Up Event Hub
    - Hands-On Lab: Stream Data with Event Hub

Big Data Processing and Data Governance

Introduction to Spark and Databricks

- Overview of Apache Spark
    - Spark Architecture and Components
    - RDDs and DataFrames
- Introduction to Databricks
    - Databricks Environment
    - Integration with Spark
- Hands-On Lab: Basic Data Processing with Spark and Databricks

Data Governance and Classification

- Basic Concepts of Data Classification
    - Types of Data: PII, Sensitive Data, etc.
    - Importance of Data Classification
- Introduction to Data Annotation
    - Tools and Techniques
- Using Atlas Tool for Data Governance
    - Overview and Features
    - Hands-On Lab: Data Annotation and Classification with Atlas Tool

Data Warehousing and Data Lakes

Introduction to Data Warehousing and Data Lakes

- Overview of Data Warehousing and Data Lakes
    - Key Concepts and Differences
- Introduction to Azure Synapse
    - Features and Architecture
    - Setting Up and Using Synapse
- Hands-On Lab: Basic Operations with Azure Synapse

GCP Spanner and Alloy DB

- Introduction to GCP Spanner
    - Features and Architecture
    - Setting Up and Using Spanner
- Introduction to Alloy DB
    - Features and Architecture
    - Setting Up and Using Alloy DB
- Hands-On Lab: Basic Operations with GCP Spanner and Alloy DB

Visualization and Charting with Data

Course Description:

This course provides a comprehensive overview of data visualization and charting techniques using various tools and programming languages. Students will learn to create effective visualizations that convey data insights clearly and compellingly. The course covers foundational principles, advanced techniques, and best practices in data visualization.

Course Outline:

1: Introduction to Data Visualization

- **Overview of Data Visualization**
  - Importance of Data Visualization
  - History and Evolution
- **Principles of Effective Visualization**
  - Visual Perception
  - Data-Ink Ratio
  - Choosing the Right Chart
  - Introduction to Co-pilot in Power BI
  - Overview and Capabilities
  - Setting Up Co-pilot in Power BI
  - Creating Dashboards with Co-pilot
  - Automated Insights and Suggestions
  - Customizing Visualizations with Co-pilot Assistance
  - Hands-On Lab: Building an Interactive Dashboard with Co-pilot
  - Importing Data
  - Creating Visuals and Charts
  - Applying Co-pilot Recommendations

- **Tools and Libraries Overview**
  - Overview of GCP Looker
  - Key Features and Advantages over Tableau
  - Integration with GCP Ecosystem
  - Setting Up GCP Looker
  - Account Setup and Configuration

  - Connecting Data Sources
  - Creating Visualizations in GCP Looker
  - Building Dashboards and Reports
  - Advanced Visuals and Interactivity
  - Python (Matplotlib, Seaborn)
  - JavaScript (D3.js)

2: Basic Chart Types

- **Bar Charts**
  - Vertical and Horizontal Bars
  - Grouped and Stacked Bars
- **Line Charts**
  - Single and Multiple Lines
  - Smoothing Techniques
- **Pie and Donut Charts**
  - When to Use and When to Avoid
  - Alternatives to Pie Charts
- **Hands-On Labs**
  - Creating Basic Charts with Different Tools

3: Advanced Chart Types

- **Scatter Plots**
  - Adding Trend Lines
  - Bubble Charts
- **Histograms and Density Plots**
  - Distribution Analysis
  - Kernel Density Estimation
- **Box Plots and Violin Plots**
  - Comparative Analysis
- **Hands-On Labs**
  - Creating Advanced Charts with Python and R

4: Interactive Visualizations

- **Introduction to Interactivity**
  - Importance and Applications

- **Interactive Tools and Libraries**
  - Plotly, Bokeh
  - Dash, Shiny
- **Building Interactive Dashboards**
  - Linking Multiple Visuals
  - Adding Filters and Controls
- **Hands-On Labs**
  - Creating Interactive Visualizations with Plotly and Dash

5: Data Visualization with D3.js

- **Introduction to D3.js**
  - Understanding the Basics
  - Selections and Data Binding
- **Creating Basic Visualizations**
  - Bar Charts, Line Charts, Scatter Plots
- **Advanced D3.js Techniques**
  - Transitions and Animations
  - Custom Visualizations
- **Hands-On Labs**
  - Building Visualizations with D3.js

6: Geospatial Visualization

- **Introduction to Geospatial Data**
  - Types and Sources of Geospatial Data
- **Mapping Tools and Libraries**
  - Leaflet, Folium
  - Mapbox, Google Maps API
- **Creating Maps and Geospatial Charts**
  - Choropleth Maps
  - Heatmaps and Marker Clusters
- **Hands-On Labs**
  - Visualizing Geospatial Data with Python and JavaScript

7: Visualization Best Practices

- **Design Principles**

  - Color Theory and Usage
  - Typography and Layout
- **Avoiding Common Pitfalls**
  - Misleading Visualizations
  - Overcomplicating Charts
- **Storytelling with Data**
  - Crafting a Narrative
  - Annotating Visualizations
- **Hands-On Labs**
  - Redesigning Poor Visualizations

8: Data Visualization in Business Intelligence

- **Introduction to Business Intelligence (BI) Tools**
  - Tableau, Power BI
- **Connecting to Data Sources**
  - Databases, APIs, Spreadsheets
- **Creating Dashboards and Reports**
  - Best Practices for BI Dashboards
  - Real-Time Data Integration
- **Hands-On Labs**
  - Building BI Dashboards with Tableau and Power BI

9: Advanced Visualization Techniques

- **Time Series Visualization**
  - Line Charts, Area Charts, and Slope Graphs
- **Network Visualization**
  - Node-Link Diagrams, Sankey Diagrams
- **Multidimensional Data Visualization**
  - Parallel Coordinates, Radar Charts
- **Hands-On Labs**
  - Implementing Advanced Techniques with Python and R

10: Capstone Project and Case Studies

- **Capstone Project**
  - Selecting a Dataset

  - Designing and Implementing Visualizations
  - Presenting Findings and Insights
- **Case Studies**
  - Analysis of Real-World Visualization Projects
  - Success Stories and Lessons Learned
- **Final Presentations**
  - Presenting Capstone Projects
  - Peer Reviews and Feedback
- **Software and Tools:**
  - Tableau, Power BI
  - Python (Matplotlib, Seaborn, Plotly, Dash)
  - JavaScript (D3.js)
  - R (ggplot2, Shiny)
- **Graph Visualization using D3.js**
- **Introduction to Graph Visualization**
- Overview of Graphs and Networks
- Nodes and Edges
- Types of Graphs (Directed, Undirected, Weighted, etc.)
- Importance of Graph Visualization
- Use Cases and Applications (Knowledge Graphs, Social Networks, etc.)
- Introduction to D3.js
- Features and Capabilities
- Setting Up D3.js
- **Creating Graph Visualizations with D3.js**
- Basic Concepts in D3.js
- Selections and Data Binding
- SVG Elements (Circles, Lines, etc.)
- Building a Simple Graph
- Representing Nodes and Edges
- Adding Labels and Tooltips
- Hands-On Lab: Creating a Simple Graph Visualization
- Loading Data

- Rendering Nodes and Edges
- Adding Interactivity
- **Advanced Graph Visualization and Introduction to Graph Databases**
- **Advanced Graph Visualization Techniques**
- Force-Directed Layouts
- Understanding Force Simulations
- Configuring Forces (Link, Charge, Center, etc.)
- Enhancing Visualizations
- Styling and Animations
- Handling Large Graphs
- Hands-On Lab: Creating an Interactive Knowledge Graph
- Implementing Force-Directed Layout
- Adding Interactivity and Animations
- **Introduction to Azure Cosmos DB and AWS Neptune**
- Overview of Graph Databases
- What are Graph Databases?
- Use Cases and Applications
- Introduction to Azure Cosmos DB
- Features and Capabilities
- Data Model and API (Gremlin API)
- Introduction to AWS Neptune
- Features and Capabilities
- Data Model and API (Gremlin and SPARQL)
- Hands-On Lab: Setting Up and Querying Graph Databases
- Setting Up Azure Cosmos DB
- Setting Up AWS Neptune
- Basic CRUD Operations
- Writing and Executing Graph Queries

Assessment:

- **Assignments:**
  - ly assignments on creating various types of visualizations

Projects:

- Midterm project on interactive visualizations
- Capstone project integrating multiple visualization techniques

Knowledge Graphs and Vector Databases

Course Description:

This course provides a comprehensive understanding of Knowledge Graphs (KG) and Vector Databases (Vector DBs). Students will learn the principles, techniques, and applications of these technologies in managing and querying complex, interconnected data. The course covers foundational concepts, practical implementation, and advanced topics, emphasizing hands-on experience with relevant tools and platforms.

Course Outline:

1: Introduction to Knowledge Graphs

- **Overview of Knowledge Graphs**
    - Definition and History
    - Importance and Applications
- **Fundamental Concepts**
    - Nodes, Edges, and Properties
    - Ontologies and Taxonomies
    - RDF (Resource Description Framework) and SPARQL
- **Tools and Platforms**
    - Neo4j, Apache Jena, GraphDB
- **Hands-On Labs**
    - Creating a Simple Knowledge Graph

2: Data Modeling with Knowledge Graphs

- **Data Modeling Principles**
    - Entity-Relationship Modeling
    - Schema Design for Knowledge Graphs
- **Ontology Development**
    - Creating and Using Ontologies
    - OWL (Web Ontology Language)
- **Case Studies**
    - Real-World Examples of Knowledge Graphs
- **Hands-On Labs**
    - Designing and Implementing a Knowledge Graph Schema

3: Querying Knowledge Graphs

- **SPARQL Fundamentals**
  - Basic SPARQL Queries
  - Filtering and Aggregation
- **Advanced SPARQL Techniques**
  - Joins and Subqueries
  - SPARQL Extensions (SPIN, SHACL)
- **Hands-On Labs**
  - Writing and Executing SPARQL Queries

4: Knowledge Graph Construction and Integration

- **Data Ingestion Techniques**
  - ETL (Extract, Transform, Load) Processes
  - Data Integration from Various Sources
- **Knowledge Graph Enrichment**
  - Entity Linking and Resolution
  - Semantic Annotation
- **Hands-On Labs**
  - Building and Enriching a Knowledge Graph

5: Introduction to Vector Databases

- **Overview of Vector Databases**
  - Definition and Applications
  - Key Differences from Traditional Databases
- **Fundamental Concepts**
  - Vector Representation of Data
  - Embedding Spaces and Similarity Search
- **Tools and Platforms**
  - FAISS, Milvus, Pinecone
- **Hands-On Labs**
  - Setting Up and Querying a Vector Database

6: Data Modelling with Vector Databases

- **Vector Representation Techniques**
  - Word Embeddings (Word2Vec, GloVe)

  - Sentence Embeddings (BERT, GPT)
- **Indexing and Retrieval**
  - Approximate Nearest Neighbor (ANN) Search
  - Index Structures (LSH, HNSW)
- **Case Studies**
  - Real-World Examples of Vector Databases
- **Hands-On Labs**
  - Creating and Querying Embeddings in a Vector Database

7: Advanced Topics in Knowledge Graphs

- **Graph Algorithms**
  - Pathfinding and Shortest Path
  - Centrality Measures
  - Community Detection
- **Scalability and Performance**
  - Distributed Knowledge Graphs
  - Optimization Techniques
- **Hands-On Labs**
  - Implementing and Analyzing Graph Algorithms

8: Advanced Topics in Vector Databases

- **Dimensionality Reduction**
  - PCA, t-SNE, UMAP
- **Hybrid Search Techniques**
  - Combining Vector and Traditional Queries
- **Scalability and Performance**
  - Distributed Vector Databases
  - Optimization Techniques
- **Hands-On Labs**
  - Applying Dimensionality Reduction and Hybrid Search

9: Integration of Knowledge Graphs and Vector Databases

- **Complementary Use Cases**
  - Enhancing Knowledge Graphs with Embeddings
  - Leveraging Vector Databases for Semantic Search

- **Integration Techniques**
  - Data Pipelines and ETL for Combined Use
  - Querying Across Knowledge Graphs and Vector Databases
- **Hands-On Labs**
  - Integrating and Querying Combined Systems

10: Capstone Project and Case Studies

- **Capstone Project**
  - Selecting a Real-World Problem
  - Designing and Implementing a Solution Using Knowledge Graphs and Vector Databases
  - Presenting Findings and Insights
- **Case Studies**
  - Analysis of Successful Implementations
  - Challenges and Best Practices
- **Final Presentations**
  - Presenting Capstone Projects
  - Peer Reviews and Feedback
- **Software and Tools:**
  - Neo4j, Apache Jena, GraphDB
  - FAISS, Milvus, Pinecone
  - Python Libraries: NetworkX, Gensim, scikit-learn

Assessment:

- **Assignments:**
  - ly assignments on various aspects of knowledge graphs and vector databases
- **Projects:**
  - Midterm project on constructing a knowledge graph
  - Capstone project integrating knowledge graphs and vector databases

Prerequisites:

- Basic knowledge of data structures and algorithms
- Familiarity with database concepts
- Understanding of basic programming concepts (preferably in Python)

Relational Database Management Systems (RDBMS) and NoSQL Databases

Course Description:

This course provides a comprehensive overview of Relational Database Management Systems (RDBMS) and NoSQL databases. Students will learn the fundamentals of database design, implementation, and management for both relational and non-relational databases. The course covers SQL and NoSQL query languages, data modeling, and the advantages and use cases of each database type.

Course Outline:

1: Introduction to Databases

- **Overview of Databases**
  - History and Evolution of Databases
  - Types of Databases
- **Introduction to RDBMS**
  - Key Concepts: Tables, Rows, Columns, Primary Keys, Foreign Keys
  - ACID Properties
- **Introduction to NoSQL**
  - Key Concepts: Collections, Documents, Key-Value Pairs
  - BASE Properties

2: Database Design and Data Modeling (RDBMS)

- **Entity-Relationship (ER) Modeling**
  - Entities, Attributes, and Relationships
  - ER Diagrams
- **Normalization**
  - Normal Forms (1NF, 2NF, 3NF, BCNF)
  - De-normalization
- **SQL Data Definition Language (DDL)**
  - Creating and Modifying Tables
  - Constraints and Indexes

3: SQL Basics

- **SQL Data Manipulation Language (DML)**
  - SELECT, INSERT, UPDATE, DELETE

- **SQL Query Syntax**
    - Basic Queries
    - Filtering and Sorting Data
- **Joins and Subqueries**
    - INNER JOIN, LEFT JOIN, RIGHT JOIN, FULL JOIN
    - Nested Queries

4: Advanced SQL

- **Advanced Query Techniques**
    - Aggregate Functions (SUM, AVG, COUNT, etc.)
    - GROUP BY and HAVING Clauses
- **Stored Procedures and Functions**
    - Creating and Using Stored Procedures
    - Creating and Using Functions
- **Transactions and Concurrency Control**
    - BEGIN, COMMIT, ROLLBACK
    - Isolation Levels

5: Introduction to NoSQL Databases

- **Types of NoSQL Databases**
    - Document Stores (e.g., MongoDB)
    - Key-Value Stores (e.g., Redis)
    - Column Stores (e.g., Cassandra)
    - Graph Databases (e.g., Neo4j)
- **NoSQL vs. RDBMS**
    - Scalability and Performance
    - Use Cases and Trade-offs

6: Document-Oriented Databases (MongoDB)

- **MongoDB Basics**
    - Collections and Documents
    - BSON Format
- **CRUD Operations**
    - Inserting, Updating, Deleting Documents
    - Querying Documents

- **Indexing and Aggregation**
    - Creating Indexes
    - Aggregation Framework

7: Key-Value Stores (Redis)

- **Redis Basics**
    - Key-Value Data Model
    - Data Types in Redis (Strings, Lists, Sets, Hashes, etc.)
- **Redis Commands**
    - Basic CRUD Operations
    - Transactions and Pipelining
- **Use Cases**
    - Caching, Session Management, Real-Time Analytics

8: Column-Oriented Databases (Apache Cassandra)

- **Cassandra Basics**
    - Data Model (Keyspace, Column Families, Rows)
    - Consistency Levels
- **Cassandra Query Language (CQL)**
    - Basic CRUD Operations
    - Advanced CQL Queries
- **Data Replication and Consistency**
    - Replication Strategies
    - Handling Consistency

9: Graph Databases (Neo4j)

- **Graph Database Basics**
    - Nodes, Relationships, Properties
    - Graph Data Model
- **Cypher Query Language**
    - Basic CRUD Operations
    - Traversing Graphs
- **Use Cases**
    - Social Networks, Recommendation Systems, Fraud Detection

10: Data Management and Security

- **Backup and Recovery**
    - Strategies for Backup and Recovery
    - Tools and Techniques
- **Security**
    - Authentication and Authorization
    - Encryption and Data Masking
- **Performance Tuning**
    - Indexing and Query Optimization
    - Monitoring and Diagnostics

11: Database Integration and Migration

- **Integrating RDBMS and NoSQL**
    - Hybrid Database Systems
    - Data Sync and ETL Processes
- **Database Migration**
    - Migrating from RDBMS to NoSQL
    - Tools and Best Practices
- **Case Studies**
    - Real-World Examples of Database Integration and Migration

12: Capstone Project and Case Studies

- **Capstone Project**
    - Designing and Implementing a Database Solution
    - Integrating RDBMS and NoSQL Components
    - Presenting Project Findings
- **Case Studies**
    - Analysis of Industry Use Cases
    - Lessons Learned from Real-World Implementations
- **Final Presentations**
    - Presenting Capstone Projects
    - Peer Reviews and Feedback
- **Software and Tools:**
    - MySQL, PostgreSQL

- MongoDB, Redis, Apache Cassandra, Neo4j
- SQL Management Tools (e.g., MySQL Workbench, pgAdmin)
- NoSQL Management Tools (e.g., MongoDB Compass, RedisInsight)

Assessment:

- **Assignments:**
  - ly assignments on database design, SQL queries, and NoSQL operations
- **Projects:**
  - Midterm project on RDBMS design and implementation
  - Capstone project integrating RDBMS and NoSQL solutions

Prerequisites:

- Basic knowledge of programming concepts
- Familiarity with data structures and algorithms
- Understanding of basic computer science principles

Hadoop

Day 1: Introduction to Hadoop and HDFS

- Overview of Big Data
- History and Evolution of Hadoop
- Hadoop Ecosystem Components

Hadoop Distributed File System (HDFS)

HDFS Architecture

    - NameNode and DataNode
    - Block Storage and Replication

HDFS Commands

    - Basic File Operations (put, get, list, etc.)

- Hands-On Lab: Setting Up a Hadoop Environment
    - Installing Hadoop
    - Configuring HDFS
    - Running Basic HDFS Commands

Introduction to MapReduce

- MapReduce Framework
    - Map and Reduce Functions
    - Data Flow in MapReduce
- MapReduce Example
    - Word Count Program

YARN (Yet Another Resource Negotiator)

- YARN Architecture
    - ResourceManager and NodeManager
    - ApplicationMaster and Containers
- Resource Allocation and Job Scheduling

Hands-On Lab: Writing and Running MapReduce Jobs

- Setting Up a Sample Data Set
- Writing a Simple MapReduce Program
- Submitting and Monitoring Jobs

Introduction to Spark and PySpark

Introduction to Apache Spark

Spark Overview

- Differences Between Hadoop and Spark
- Spark Ecosystem Components

Spark Architecture

- Driver and Executors
- RDDs (Resilient Distributed Datasets)
- DAG (Directed Acyclic Graph)

Getting Started with PySpark

Setting Up PySpark

- Installing and Configuring PySpark
- PySpark Shell and Notebooks

Basic PySpark Operations

- Creating RDDs
- Transformations and Actions

DataFrame API in PySpark

- Introduction to DataFrames
    - Creating DataFrames
    - Performing Operations on DataFrames
- SQL Queries with DataFrames
    - Using Spark SQL
- Hands-On Lab: Working with DataFrames in PySpark
    - Loading Data into DataFrames
    - Applying Transformations and Actions

Spark Streaming

- Introduction to Spark Streaming
    - Streaming Architecture
    - DStreams (Discretized Streams)
- Processing Real-Time Data

  - Window Operations
  - Integrating with Data Sources (Kafka, HDFS, etc.)

Machine Learning with MLlib

- Introduction to Spark MLlib
  - MLlib vs. Spark ML
  - Pipelines and Estimators
- Example Use Case: Predictive Modeling
  - Loading and Preparing Data
  - Training and Evaluating Models

Hands-On Lab: End-to-End Data Processing with PySpark

- Loading Data from HDFS
- Transforming Data with DataFrame API
- Running a Machine Learning Pipeline
- Streaming Data Processing

Assessment:

- **Hands-On Labs:**
  - Practical exercises on HDFS, MapReduce, and PySpark
- **Quizzes:**
  - Short quizzes to test understanding of key concepts
- **Capstone Project:**
  - An end-to-end data processing project using Hadoop and Spark

Prerequisites:

- Basic knowledge of programming (preferably Python)
- Familiarity with Linux command line
- Understanding of basic data processing concepts

Stream 2

Data Science and Python

Course Outline:

This course aims to provide an in-depth understanding of data science principles and optimization techniques using Python. Students will learn to manipulate and analyze data, build predictive models, and implement optimization algorithms. The course combines theoretical concepts with practical applications, ensuring that students can apply their knowledge to real-world problems.

1: Introduction to Data Science and Python

Introduction to Data Science

- What is Data Science?
- Applications of Data Science
- The Data Science Process

Python Basics

- Introduction to Python
- Data Types, Variables, and Operators
- Control Structures (Loops and Conditionals)
- Functions and Modules

Python Libraries for Data Science

- NumPy, Pandas, Matplotlib, and Seaborn

2: Data Manipulation and Analysis with Pandas

Introduction to Pandas

- Series and DataFrames
- Importing and Exporting Data

Data Cleaning and Preparation

- Handling Missing Data
- Data Transformation
- Data Aggregation and Grouping

Exploratory Data Analysis (EDA)

- Descriptive Statistics
- Data Visualization Techniques

3: Data Visualization

Introduction to Data Visualization

- Importance of Data Visualization
- Types of Visualizations

Matplotlib and Seaborn

- Basic Plotting with Matplotlib
- Advanced Visualization with Seaborn
- Customizing Plots

Interactive Visualization with Plotly

- Introduction to Plotly
- Creating Interactive Dashboards

4: Probability and Statistics

- **Introduction to Probability**
  - Basic Probability Concepts
  - Probability Distributions
- **Statistical Inference**
  - Descriptive vs. Inferential Statistics
  - Hypothesis Testing
  - Confidence Intervals

5: Machine Learning Basics

- **Introduction to Machine Learning**
  - Types of Machine Learning
  - Supervised vs. Unsupervised Learning
- **Supervised Learning**
  - Linear Regression
  - Logistic Regression
  - Evaluation Metrics (RMSE, Accuracy, Precision, Recall)

6: Advanced Machine Learning

- **Unsupervised Learning**
  - Clustering (K-Means, Hierarchical)
  - Dimensionality Reduction (PCA)
- **Advanced Algorithms**
  - Decision Trees and Random Forests

  - Support Vector Machines
  - Ensemble Methods

7: Optimization Fundamentals

- **Introduction to Optimization**
  - What is Optimization?
  - Applications of Optimization
- **Mathematical Foundations**
  - Linear Algebra and Calculus for Optimization
  - Convex and Non-Convex Functions
- **Linear Programming**
  - Formulating Linear Programs
  - Simplex Method
  - Duality in Linear Programming

8: Non-Linear and Integer Programming

- **Non-Linear Optimization**
  - Gradient Descent
  - Constrained Optimization (Lagrange Multipliers)
- **Integer Programming**
  - Formulating Integer Programs
  - Branch and Bound Method

9: Heuristic and Metaheuristic Methods

- **Introduction to Heuristics**
  - Basic Concepts
  - Greedy Algorithms
- **Metaheuristic Algorithms**
  - Genetic Algorithms
  - Simulated Annealing
  - Particle Swarm Optimization

10: Optimization in Machine Learning

- **Optimization in Model Training**
  - Loss Functions
  - Gradient-Based Optimization (SGD, Adam)

- **Hyperparameter Tuning**
  - Grid Search
  - Random Search
  - Bayesian Optimization

11: Advanced Topics in Data Science and Optimization

- **Time Series Analysis**
  - Introduction to Time Series
  - ARIMA and Exponential Smoothing
- **Natural Language Processing (NLP)**
  - Text Preprocessing
  - Sentiment Analysis
- **Deep Learning**
  - Introduction to Neural Networks
  - Convolutional Neural Networks (CNNs)
  - Recurrent Neural Networks (RNNs)

12. Introduction to Probability and Statistics (30 minutes)

- Basic Concepts of Probability
- Overview of Statistical Inference

Bayes Theorem

- Introduction to Bayes Theorem
  - Definition and Formula
  - Conditional Probability
- Applications of Bayes Theorem
  - Diagnostic Testing
  - Spam Filtering
  - Bayesian Inference
- Worked Examples
  - Calculating Posterior Probabilities
  - Examples in Real-World Contexts
- Hands-On Lab: Implementing Bayes Theorem
  - Practical Exercises with Data Sets
  - Using Python/R for Bayesian Calculations

Markov Theorem

- Introduction to Markov Processes
  - Definition and Characteristics
  - Types of Markov Chains (Discrete and Continuous)
- Understanding the Markov Property
  - Memoryless Property
  - Transition Matrices
- Applications of Markov Theorem
  - Predictive Modeling
  - PageRank Algorithm
  - Stock Market Analysis
- Worked Examples
  - Constructing Transition Matrices
  - Solving for Steady-State Probabilities
- Hands-On Lab: Implementing Markov Chains
  - Practical Exercises with Transition Matrices
  - Using Python/R for Markov Chain Simulations

. Advanced Topics in Bayesian and Markov Theorems

- Bayesian Networks
  - Structure and Inference
  - Applications in Machine Learning
- Hidden Markov Models (HMM)
  - Definition and Applications
  - Viterbi Algorithm
- Hands-On Lab: Advanced Implementations
  - Building Bayesian Networks
  - Implementing HMMs in Python/R

Case Studies and Real-World Applications Case Study: Medical Diagnosis Using Bayesian Methods

- Case Study: Predictive Text Input Using Markov Chains
- Group Discussion: Practical Applications in Industry
- **Software and Tools:**

  - Python (with libraries such as NumPy, SciPy, and PyMC3)
  - R (with packages such as BayesianTools, markovchain)
  - Jupyter Notebooks/RStudio

Assessment:

- **Hands-On Labs:**
  - Practical exercises on implementing Bayes Theorem and Markov Chains
  - Simulating Bayesian and Markov models using Python/R
- **Quizzes:**
  - Short quizzes to test understanding of key concepts
- **Final Project:**
  - An end-to-end project involving the application of Bayes Theorem and Markov Chains to a real-world data set

Prerequisites:

- Basic knowledge of probability and statistics
- Familiarity with Python or R programming
- Understanding of basic data analysis techniques

12: Project and Case Studies

- **Real-World Data Science Project**
  - Project Planning and Data Collection
  - Data Analysis and Model Building
  - Optimization and Results Presentation
- **Case Studies**
  - Case Study 1: Optimization in Supply Chain Management
  - Case Study 2: Predictive Maintenance using Machine Learning
  - Case Study 3: Optimizing Marketing Campaigns
- **Software and Tools:**
  - Python (Anaconda Distribution)
  - Jupyter Notebooks
  - Libraries: NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Plotly

Prerequisites:

- Basic knowledge of Python programming
- Understanding of basic mathematical concepts (algebra, calculus, probability)