CDP Certified Data Developer

Course outline

Module 1: Introduction to Data Science

Module 1: Introduction to Data Science is an introductory course designed to provide an overview of the fundamentals of data science. It covers topics such as data collection, data cleaning, data analysis, data visualization, and machine learning. This module will provide a foundation for the rest of the CDP Certified Data Developer course and equip students with the necessary skills to become successful data scientists.

Lessons

- Introduction to Data Science Concepts
- Data Collection and Preparation
- Exploratory Data Analysis
- Data Visualization
- Data Cleaning and Pre-processing
- Feature Engineering
- Machine Learning Algorithms
- Model Evaluation and Optimization
- Big Data and Cloud Computing
- . Data Science in Business and Industry

After completing this module, students will be able to:

- Understand the fundamentals of data science and its applications in the real world.
- Develop an understanding of the different types of data and how to analyze them.
- Learn how to use data visualization techniques to present data in a meaningful way.
- Develop the skills to create and interpret data models and algorithms.

Module 2: Data Modeling and Analysis

Module 2 of the CDP Certified Data Developer course focuses on data modeling and analysis. Students will learn how to create data models, analyze data, and use data to make decisions. They will also learn how to use data visualization tools to present data in a meaningful way. Additionally, they will gain an understanding of the different types of data analysis techniques and how to apply them to real-world scenarios.

Lessons

- Introduction to Data Modeling
- Data Modeling Concepts and Techniques
- Data Modeling Tools and Techniques
- Data Modeling Best Practices
- Data Modeling for Big Data
- Data Modeling for Data Warehousing
- Data Modeling for Business Intelligence
- Data Modeling for Data Lakes
- Data Modeling for Data Visualization
- . Data Modeling for Predictive Analytics
- . Data Modeling for Machine Learning
- . Data Modeling for Natural Language Processing
- . Data Modeling for Text Mining
- . Data Modeling for Graph Databases
- . Data Modeling for NoSQL Databases
- . Data Modeling for Cloud Computing
- . Data Modeling for Data Governance
- . Data Modeling for Data Quality
- . Data Modeling for Data Security
- . Data Modeling for Data Integration

After completing this module, students will be able to:

- Understand the fundamentals of data modeling and analysis.
- Develop an understanding of the different types of data models and their uses.
- Utilize data modeling techniques to create and analyze data models.
- Apply data analysis techniques to identify patterns and trends in data.

Module 3: Data Warehousing and Data Lakes

Module 3 of the CDP Certified Data Developer course covers the fundamentals of data warehousing and data lakes. It provides an overview of the different types of data warehouses and data lakes, their architectures, and the tools and techniques used to build and maintain them. It also covers the best practices for designing and managing data warehouses and data lakes, as well as the security and privacy considerations associated with them.

Lessons

- Introduction to Data Warehousing
- Data Warehousing Architecture
- Data Modeling for Data Warehousing
- ETL Processes for Data Warehousing
- Data Warehousing Performance Tuning
- Data Warehousing Security
- Introduction to Data Lakes
- Data Lake Architecture
- Data Modeling for Data Lakes
- . ETL Processes for Data Lakes

- . Data Lake Performance Tuning
- . Data Lake Security

After completing this module, students will be able to:

- Understand the fundamentals of data warehousing and data lakes, including their architecture, components, and use cases.
- Design and implement data warehouses and data lakes to store and analyze large volumes of data.
- Utilize ETL tools to extract, transform, and load data into data warehouses and data lakes.
- Develop and deploy data models and queries to analyze data stored in data warehouses and data lakes.

Module 4: Data Visualization and Reporting

Module 4 of the CDP Certified Data Developer course focuses on data visualization and reporting. Students will learn how to create effective visualizations and reports to communicate data insights. They will also explore the use of various tools and techniques to create interactive dashboards and reports. Additionally, they will learn how to use data to create meaningful stories and insights.

Lessons

- Introduction to Data Visualization
- Exploring Data Visualization Tools
- Designing Effective Visualizations
- Creating Dashboards and Reports
- Working with Advanced Visualization Techniques
- Creating Interactive Visualizations
- Automating Reports and Dashboards
- Best Practices for Data Visualization
- Troubleshooting Data Visualization Issues
- . Advanced Data Visualization Techniques

After completing this module, students will be able to:

- Understand the fundamentals of data visualization and reporting, including the principles of effective data visualization and the importance of data storytelling.
- Utilize various data visualization tools and techniques to create meaningful visualizations from data.
- Develop and present data-driven reports to stakeholders.
- Analyze and interpret data to identify trends and patterns.

Module 5: Data Governance and Security

Module 5 of the CDP Certified Data Developer course focuses on data governance and security. It covers topics such as data privacy, data security, data access control, data quality, and data auditing. It also provides an overview of the different types of data governance frameworks and how to implement them. Additionally, it provides an introduction to data security best practices and how to protect data from unauthorized access.

Lessons

- Understanding Data Governance Principles
- Data Security Best Practices
- Data Access Control and Authorization
- Data Encryption and Masking
- Data Auditing and Monitoring
- Data Privacy and Compliance
- Data Quality Assurance
- Data Risk Management
- Data Backup and Recovery
- . Data Protection Strategies

After completing this module, students will be able to:

- Understand the importance of data governance and security in the context of data engineering.
- Implement data security measures such as encryption, authentication, and authorization.
- Develop data governance policies and procedures to ensure data integrity and compliance.
- Utilize data security tools and technologies to protect data from unauthorized access.

Module 6: Big Data Technologies

Module 6 of the CDP Certified Data Developer course covers the fundamentals of Big Data technologies, including Apache Hadoop, Apache Spark, and Apache Kafka. Students will learn how to use these technologies to process and analyze large datasets, as well as how to design and implement distributed data processing pipelines. Additionally, students will gain an understanding of the various components of the Hadoop ecosystem and how to use them to build data-driven applications.

Lessons

- Introduction to Big Data Technologies
- Overview of Hadoop and its Ecosystem
- Understanding HDFS and MapReduce
- Working with Apache Spark
- Exploring Apache Hive and Impala
- Analyzing Data with Apache Pig
- Introduction to NoSQL Databases
- Working with Apache Cassandra
- Understanding Apache Flume and Kafka
- . Introduction to Apache Storm
- . Using Apache Oozie for Scheduling
- . Introduction to Apache Mahout
- . Big Data Security and Governance

After completing this module, students will be able to:

- Understand the fundamentals of big data technologies such as Hadoop, Spark, and NoSQL databases.
- Develop and deploy big data applications using the Hadoop ecosystem.
- Utilize NoSQL databases to store and query large datasets.
- Analyze and visualize big data using tools such as Apache Pig and Apache Hive.

Module 7: Data Engineering

Module 7 of the CDP Certified Data Developer course focuses on data engineering. It covers topics such as data modeling, data warehousing, data integration, data quality, and data security. Students will learn how to design and implement data pipelines, create data models, and develop data-driven applications. They will also gain an understanding of the principles of data governance and the importance of data security.

Lessons

- Introduction to Data Engineering
- Data Warehousing
- Data Modeling
- Data Integration
- Data Quality Management
- Data Security and Privacy
- Data Governance
- Big Data Technologies
- Data Lakes and Data Warehouses
- . Data Visualization
- . Data Analysis and Machine Learning
- . Cloud Computing for Data Engineering
- . Data Pipelines and ETL
- . Data Streaming
- . Data Warehouse Automation
- . Data Warehouse Performance Tuning
- . Data Warehouse Optimization
- . Data Warehouse Design Patterns
- . Data Warehouse Best Practices
- . Data Warehouse Testing and Validation

After completing this module, students will be able to:

- Understand the fundamentals of data engineering and its role in the data science process.
- Develop the skills to design and implement data pipelines for data ingestion, transformation, and storage.
- Utilize various data engineering tools and technologies such as Apache Spark, Apache Kafka, and Apache Airflow.
- Implement best practices for data engineering such as data quality assurance, data security, and data governance.

Module 8: Data Quality and Data Cleansing

Module 8 of the CDP Certified Data Developer course focuses on data quality and data cleansing. It covers topics such as data validation, data cleansing techniques, data profiling, and data quality metrics. It also provides hands-on exercises to help students gain practical experience in data quality and data cleansing.

Lessons

- Introduction to Data Quality and Data Cleansing
- Data Quality Assessments
- Data Profiling and Data Cleansing
- Data Validation and Data Transformation
- Data Standardization and Data Enrichment
- Data Quality Metrics and Measurement
- Data Quality Tools and Techniques
- Data Quality Governance and Best Practices
- Data Quality Automation and Monitoring
- . Data Quality Improvement Strategies

After completing this module, students will be able to:

- Understand the importance of data quality and data cleansing in the data development process.
- Identify and apply techniques to detect and correct data quality issues.
- Utilize data cleansing tools to automate data cleansing processes.
- Develop strategies to ensure data quality and accuracy in data development projects.

Module 9: Data Mining and Machine Learning

Module 9 of the CDP Certified Data Developer course covers the fundamentals of data mining and machine learning. It provides an introduction to the concepts and techniques used in data mining and machine learning, including supervised and unsupervised learning, decision trees, clustering, and neural networks. It also covers the use of Python and R for data mining and machine learning.

Lessons

- Introduction to Data Mining
- Data Pre-processing
- Exploratory Data Analysis
- Classification Algorithms
- Clustering Algorithms
- Association Rule Mining
- Anomaly Detection
- Text Mining
- Time Series Analysis
- . Recommender Systems
- . Deep Learning

- . Natural Language Processing
- . Reinforcement Learning
- . Model Evaluation and Selection
- . Big Data Analytics

After completing this module, students will be able to:

- Understand the fundamentals of data mining and machine learning algorithms.
- Develop and implement data mining and machine learning models.
- Utilize data mining and machine learning techniques to solve real-world problems.
- Interpret and visualize data mining and machine learning results.

Module 10: Advanced Analytics and Predictive Modeling

Module 10 of the CDP Certified Data Developer course covers advanced analytics and predictive modeling techniques. Students will learn how to use machine learning algorithms to build predictive models, as well as how to evaluate and interpret the results. They will also explore the use of data mining and text mining techniques to uncover hidden patterns and trends in data. Finally, they will learn how to use advanced analytics to make data-driven decisions.

Lessons

- Introduction to Predictive Modeling
- Exploring Data for Predictive Modeling
- Feature Engineering for Predictive Modeling
- Model Selection and Evaluation
- Advanced Modeling Techniques
- Time Series Analysis
- Text Mining and Natural Language Processing
- Recommender Systems
- Ensemble Modeling
- . Deploying Predictive Models

After completing this module, students will be able to:

- Understand the fundamentals of predictive modeling and advanced analytics techniques.
- Develop and implement predictive models using various machine learning algorithms.
- Utilize data mining techniques to identify patterns and trends in data.
- Interpret and communicate the results of predictive models to stakeholders.

Module 11: Cloud Computing and Data Platforms

Module 11 of the CDP Certified Data Developer course covers the fundamentals of cloud computing and data platforms. It provides an overview of the different types of cloud computing and data platforms, their advantages and disadvantages, and how to use them to store, process, and analyze data. It also covers topics such as data security, scalability, and cost optimization.

Lessons

- Introduction to Cloud Computing
- Cloud Computing Architectures
- Cloud Computing Services
- Cloud Computing Security
- Cloud Computing Storage
- Cloud Computing Networking
- Cloud Computing Platforms
- Cloud Computing Automation
- Cloud Computing Orchestration
- . Cloud Computing Monitoring
- . Cloud Computing Cost Optimization
- . Cloud Computing Performance Optimization
- . Cloud Computing Disaster Recovery
- . Cloud Computing Data Platforms
- . Cloud Computing Data Warehousing
- . Cloud Computing Data Lakes
- . Cloud Computing Data Pipelines
- . Cloud Computing Data Governance
- . Cloud Computing Data Security
- . Cloud Computing Data Analytics

After completing this module, students will be able to:

- Understand the fundamentals of cloud computing and data platforms
- Develop and deploy applications on cloud platforms
- Utilize cloud services to store, process, and analyze data
- Design and implement data pipelines for data ingestion, transformation, and analysis

Module 12: Data Science Project Management

Module 12 of the CDP Certified Data Developer course focuses on data science project management. It covers topics such as project planning, risk management, and communication strategies. It also provides guidance on how to effectively manage data science projects from start to finish.

Lessons

- Understanding the Project Lifecycle
- Developing a Project Plan
- Estimating Project Resources
- Managing Project Risks
- Implementing Quality Assurance
- Managing Project Teams
- Communicating Project Progress
- Managing Project Change

- Managing Project Documentation
- . Evaluating Project Success

After completing this module, students will be able to:

- Understand the fundamentals of project management and how to apply them to data science projects.
- Develop a project plan and timeline for data science projects.
- Utilize project management tools and techniques to manage data science projects.
- Identify and manage risks associated with data science projects.