ADaSci Certified Data Engineer

Program Curriculum

The ADaSci Certified Data Engineer program is a 30-hour self-paced certification designed for comprehensive training in modern data engineering. It covers core concepts, tools, and practices essential for building and managing data systems at scale.

Participants will explore topics such as data pipelines, cloud platforms, and data governance through video lectures, reading materials, case studies, and hands-on exercises. The program concludes with a rigorous certification exam, enabling learners to demonstrate their knowledge, technical skills, and problem-solving abilities in real-world data engineering scenarios.

Course Structure

The curriculum is divided into **five key modules**. Each module addresses a critical domain of data engineering, ensuring a holistic and practical understanding of the field.

Module 1: Foundations of Data Engineering

This module introduces the core concepts and workflows fundamental to any data engineering role.

- **Data Engineering Lifecycle**
  Understand the complete flow from data generation, ingestion, transformation, storage, to consumption.
- **Database Systems and Concepts**
  Learn about relational and non-relational databases, indexing, normalization, and query optimization.
- **ETL and ELT Pipelines**
  Explore both traditional and modern approaches to data movement and transformation.
- **Batch vs Stream Processing**
  Compare and apply batch and real-time processing methods for different business use cases.
- **Big Data Ecosystem Overview**
  Get an overview of the technologies and frameworks (like Hadoop, Spark, Hive) used in big data environments.

Why it Matters
This module builds the foundation required to understand how data flows in modern systems and introduces key tools and concepts used across all industries.

Module 2: Scalable Data Storage and Modeling Techniques

This module focuses on strategies for organizing and storing data at scale.

- **Data Warehousing and Data Lakes**
  Learn differences, use cases, and best practices for structured and unstructured data storage.
- **File Formats for Big Data**
  Understand efficient storage formats such as Parquet, Avro, and ORC.
- **Data Partitioning**
  Explore techniques to optimize data storage and access through intelligent partitioning.
- **Bucketing and Clustering**
  Learn advanced file organization methods that enhance processing speed and storage efficiency.
- **Data Modeling Techniques**
  Study conceptual and physical data models including star schema, snowflake schema, and dimensional modeling.

Why it Matters
A data engineer must design systems that scale efficiently. This module teaches how to structure and manage data to ensure performance and maintainability.

Module 3: Data Ingestion and Processing

This module introduces ingestion pipelines, real-time data tools, and processing engines.

- **SQL and NoSQL for Data Engineers**
  Learn to work with traditional SQL and document-based NoSQL systems like MongoDB.
- **Apache Spark**
  Gain practical knowledge in distributed processing using Spark for ETL and analytics.
- **Real-Time Data Streaming with Kafka**
  Understand the use of Kafka for building high-throughput, real-time data ingestion pipelines.
- **Workflow Orchestration with Apache Airflow**
  Learn to schedule, monitor, and manage complex workflows using Directed Acyclic Graphs (DAGs).
- **Streaming with Apache Flink (Introductory)**
  Get introduced to real-time event-driven processing using Apache Flink.

Why it Matters
Modern data systems require the ability to ingest, process, and deliver data in real time. This module prepares learners to handle high-velocity data pipelines effectively.

Module 4: Cloud Data Engineering

This module covers key cloud platforms and tools that power today's data infrastructure.

- **Cloud Data Warehouses**
  Explore the architecture and capabilities of BigQuery, Snowflake, and Redshift, including performance and pricing models.
- **Managed Spark and Data Pipelines**
  Learn about Databricks, AWS Glue, and GCP Dataflow to build scalable managed ETL pipelines.
- **Serverless Data Engineering**
  Understand how AWS Lambda, Google Cloud Functions, and event-driven architecture support scalable automation.
- **Cloud-native Storage Solutions**
  Study object storage solutions like Amazon S3, Google Cloud Storage, and Azure Blob, with focus on lifecycle policies and cost management.
- **Modern Data Stack Tools**
  Get familiar with tools like Stitch, Airbyte, and dbt Cloud that are transforming the modern data pipeline.

Why it Matters
Cloud platforms are now the default for enterprise data engineering. This module ensures engineers can work with the tools and architectures used in real-world production environments.

Module 5: Data Quality, Lineage, and Governance

This final module emphasizes trust, traceability, and compliance in data systems.

- **Data Quality Frameworks**
  Implement rules and checks to maintain data integrity and trustworthiness.
- **Data Lineage and Observability**
  Learn to track the origin and transformation path of data across systems.
- **Data Privacy and Compliance**
  Understand regulations like GDPR and HIPAA and how to engineer systems that remain compliant.
- **Data Contracts and Schema Evolution**
  Study methods to manage changes in data structure while maintaining system stability.

Why it Matters
High-quality, traceable, and compliant data systems are essential for decision-making and risk management. This module ensures engineers can build data platforms that meet enterprise standards.

Certification Exam

- **Format**: Multiple Choice Questions (MCQs)
- **Number of Questions**: 60
- **Duration**: 60 minutes
- **Passing Score**: 70 percent

The exam tests participants on both conceptual knowledge and practical application across all five modules. A successful candidate will demonstrate problem-solving ability, technical proficiency, and understanding of modern data systems.