FIRST hubIT

Holding your future

Data Engineer

(40h / 5 Days)

Description:

The Data Engineering training program offers an in-depth understanding of data management architectures, data pipelines, and cloud-based analytics using **AWS** services. Beginning with the fundamentals of data engineering and AWS's data services, the course focuses on the design and implementation of robust data architectures, including data lakes and warehouses. Participants will learn to manage data governance, security, and cataloging using tools like AWS Glue, ensuring their data systems are compliant, secure, and scalable.

Throughout the course, participants will also develop expertise in building and orchestrating data pipelines for both batch and streaming data. The training includes the use of key AWS services such as Redshift for data warehousing, Athena for querying, and SageMaker for integrating AI and machine learning into data workflows. The course concludes with a capstone project where trainees apply these concepts to build a comprehensive, scalable data platform that handles real-world data challenges.

Target Audience:

The target audience for this training program includes data engineers, data architects, and IT professionals who are responsible for designing, building, and managing data pipelines and architectures. It is also suitable for individuals who are looking to enhance their skills in leveraging AWS services for data engineering tasks, as well as those interested in incorporating artificial intelligence and machine learning into their data workflows. Prior experience with AWS and basic knowledge of data engineering concepts are recommended for participants.

Training Expected Outcomes:

Trainees will be able to :

Upon completing this 60-hour training program, participants will be able to:

1. **Design and Implement Data Management Architectures:** Understand the principles and best practices for designing data lakes, data warehouses, and data lakehouse architectures using AWS services.
2. **Build and Orchestrate Data Pipelines:** Create scalable and reliable batch and streaming data pipelines, integrating multiple AWS services for data ingestion, processing, and transformation.
3. **Utilise Key AWS Services:** Effectively use AWS services such as Amazon S3, Amazon Redshift, AWS Glue, Amazon Kinesis, and more for data storage, processing, and analytics.
4. **Apply Data Governance and Security Best Practices:** Implement robust data governance frameworks, manage data security, and ensure compliance with

regulatory requirements using AWS IAM, AWS KMS, and AWS Glue Data Catalog.

5. **Optimise Data for Analytics:** Transform and optimise data for analytical workloads using AWS Glue and other AWS services, ensuring data is prepared for accurate and efficient analysis.
6. **Deploy AI and ML Solutions:** Build, train, deploy, and manage machine learning models using Amazon SageMaker, and integrate AWS AI services like Comprehend and Rekognition into data engineering workflows.
7. **Query and Visualise Data:** Use Amazon Athena for querying data stored in Amazon S3 and Amazon QuickSight for creating interactive dashboards and visualising data insights.
8. **Build Transactional Data Lakes and Implement Data Mesh Strategies:** Design and manage transactional data lakes and implement data mesh architectures to support decentralised data management and analytics.
9. **Create Modern Data Platforms:** Design, build, and optimise modern data platforms on AWS Cloud service , leveraging the latest AWS Cloud services technologies to handle large-scale data processing and analytics needs.

By the end of the program, participants will have hands-on experience and a solid understanding of how to leverage Cloud services to build and manage comprehensive data engineering solutions, preparing them to tackle real-world data challenges effectively.

Training Strategy

1. **Module-Based Learning Approach**
   - **Structured Progression**: The training is divided into 10 modules, each building on the previous one to ensure a structured and progressive learning experience.
   - **Cohesive Learning**: Each module covers different aspects of data engineering, ensuring a comprehensive understanding of the entire data engineering lifecycle.
2. **Combination of Theoretical and Practical Learning**
   - **Theoretical Lessons**: Each module begins with theoretical lessons to introduce key concepts, techniques, and best practices. This provides participants with the foundational knowledge required for practical application.
   - **Hands-On Labs**: Practical labs are integrated into each module, allowing participants to apply what they have learned in real-world scenarios. Labs focus on using AWS services for data engineering tasks, reinforcing theoretical knowledge through practice.
3. **Assessments and Real-World Projects**
   - **Quizzes**: Each module includes quizzes to assess participants' understanding of the theoretical concepts.
   - **Hands-On Projects**: Practical projects are included to evaluate participants' ability to apply their knowledge in real-world scenarios. These projects mimic real-life data engineering challenges and require participants to utilize AWS services effectively.
4. **Real-World Case Studies**
   - **Industry Examples**: Each module includes case studies showcasing successful data engineering projects. These case studies provide insights into practical applications and best practices used in the industry.
   - **Lessons Learned**: Participants learn from real-world examples, understanding the challenges and solutions in implementing data engineering projects.
5. **Expert Guidance and Support**
   - **Instructor-Led Sessions**: Modules are delivered through instructor-led sessions, providing expert guidance and facilitating interactive learning.
   - **Q&A Session**s: Regular Q&A sessions are included to address participants' questions and provide additional support.

Community and Peer Learning: Participants are encouraged to engage in discussions and share knowledge with peers, fostering a collaborative learning environment.

Course Modules

1. **An Introduction to Data Engineering and AWS Services –** Overview of data engineering principles and AWS services.
2. **Data Management Architectures for Analytics –** Design data lakes and warehouses for scalable analytics.
3. **Data Governance, Security, and Cataloging –** Implement governance and security using AWS Glue and IAM.
4. **Architecting Data Engineering Pipelines –** Build data pipelines for batch and streaming data with AWS tools.
5. **Ingesting and Transforming Data –** Efficiently ingest and transform data using AWS Glue and Kinesis.
6. **Data Marts and Amazon Redshift –** Create and manage data marts using Amazon Redshift for large-scale analytics.
7. **Orchestrating Data Pipelines and Queries with Athena –** Automate data pipelines and queries using AWS Step Functions and Athena.
8. **Visualising Data with Amazon QuickSight –** Build interactive data visualisations using Amazon QuickSight.
9. **Building Transactional Data Lakes and Implementing Data Mesh –** Implement transactional data lakes and data mesh architecture using AWS services.
10. **Enabling Artificial Intelligence and Machine Learning –** Integrate AI and ML models into data pipelines using AWS SageMaker.

Training Program

<table>
<tr><th colspan="2">Data Engineering</th></tr>
<tr><td colspan="2"><b>Training Objectives:</b>
• Design Scalable Data Architectures
• Develop and Orchestrate Data Pipelines
• Ensure Data Governance and Security
• Integrate AI and ML into Data Workflows:</td></tr>
<tr><td><b><i>Time</i></b></td><td><b><i>Modules</i></b></td></tr>
</table>

Module 1: An Introduction to Data Engineering and AWS Services

5 Hours **Objective**: Provide an overview of data engineering principles and introduce the essential AWS services that support data storage, processing, and analytics in a scalable and efficient manner.

- Introduction to Data Engineering
  - 2h Overview of Data Engineering:
    - Definition and roles of a Data Engineer.
    - Key concepts and principles.
  - Data Engineering Tools and Technologies:
    - Common tools and frameworks (e.g., Apache Hadoop, Apache Spark).
  - Data Engineering in the Cloud:
    - Advantages of cloud-based data engineering.
    - Introduction to AWS for data engineering.
  - Case Studies and Industry Applications:
    - Real-world applications and case studies.
- AWS Services for Data Engineers
  - 3h Compute Services for Data Processing:
    - Amazon EC2, AWS Lambda, Elastic Beanstalk.
  - Storage Services for Data Management:
    - Amazon S3, Amazon EBS, Amazon RDS.
  - Database Services for Data Management:
    - Amazon DynamoDB, Amazon Redshift, Amazon Aurora.

Lab: Setting Up Your Data Engineering Environment

- Provisioning cloud resources on AWS.
- Setting up a local development environment.

Assessment:

4 Hours 1h 1h 1h 1h	**Module 2: Data Management Architectures for Analytics (4 hours)** **Objective**: introduce trainees to the design and the implementing of robust data management architectures, including data lakes and data warehouses, to support advanced analytics using AWS services. ● Introduction to Data Management Architectures (1 hour) ○ Traditional vs. modern data architectures. ○ Key components and principles. ● Data Lakes and Data Warehouses (1 hour) ○ Differences and use cases. ○ Designing a data lake on AWS. ● Data Lakehouse Architecture (1 hour) ○ Combining data lakes and warehouses. ○ Implementing a data lakehouse on AWS. ● Case Studies and Best Practices (1 hour) ○ Real-world examples. ○ Best practices for scalable and efficient data architectures. **Lab**: Building a Data Lake on AWS ○ Creating and configuring an Amazon S3 data lake. ○ Managing data lifecycle and policies. **Assessment**: ○ Quiz on data management architectures and AWS services.

3 Hours	**Module 3: Data Governance, Security, and Cataloging** **Objective**: instruct onthe best practices for data governance, security, and cataloguing to ensure data integrity, compliance, and accessibility within AWS environments.
1h	● Introduction to Data Governance (1 hour) ○ Principles and importance. ○ Key components of a data governance framework.
1h	● Security Best Practices for Data Engineering (1 hour) ○ AWS security services and features. ○ Implementing security best practices.
1h	● Data Cataloging and Metadata Management (1 hour) ○ Importance of data cataloging. ○ Using AWS Glue Data Catalog.
	Lab: Using AWS Glue Data Catalog ○ Creating a data catalog with AWS Glue. ○ Managing and querying metadata.
	Assessment: ○ Quiz on data governance, security, and cataloging.

6 Hours	**Module 4: Architecting Data Engineering Pipelines** **Objective**: Instruct trainees on designing, building, and optimising data engineering pipelines using AWS services, focusing on both batch and real-time data processing.
1h	● Fundamentals of Data Pipelines (1 hour) ○ Key components and principles. ○ Designing scalable and reliable data pipelines.
1h	● Batch vs. Streaming Data Pipelines (1 hour) ○ Differences, use cases, and design considerations. ○ Examples of batch and streaming data pipelines.
1h	● Integrating AWS Services in Data Pipelines (1 hour) ○ Using AWS services for different stages of data pipelines. ○ Best practices for integration and orchestration.
1h	● Monitoring and Optimizing Data Pipelines (1 hour) ○ Tools and techniques for monitoring data pipelines. ○ Optimizing performance and cost-efficiency.
	Labs: ○ Building a Batch Data Pipeline: Designing and implementing a batch data pipeline using AWS services. ○ Building a Streaming Data Pipeline: Designing and implementing a streaming data pipeline using AWS Kinesis.
	Assessment: ○ Quiz on data pipeline architecture and AWS services. ○ Hands-on project: Designing and implementing a data engineering pipeline using AWS.

6 Hours

Module 5: Ingesting and Transforming Data.

Objective: Develop skills for ingesting and transforming data using AWS tools, enabling efficient data preparation and integration for downstream analytics.

- Ingesting Batch and Streaming Data (3 hours)
  - Overview of Data Ingestion Techniques:
    - 3h Batch vs. streaming data ingestion.
    - Key considerations and challenges.
  - Batch Data Ingestion with AWS:
    - Using AWS Data Pipeline and AWS Glue.
  - Streaming Data Ingestion with AWS:
    - Using Amazon Kinesis and AWS Lambda.
- Transforming Data to Optimise for Analytics (3 hours)
  - Introduction to Data Transformation:
    - 3h Key concepts and importance.
    - Common data transformation techniques.
  - Data Transformation with AWS Glue:
    - Using AWS Glue for ETL (Extract, Transform, Load).
  - Optimising Data for Analytics:
    - Best practices for transforming data for analytics.
    - Using AWS services to optimise data transformation.

Labs:

- Batch Data Ingestion with AWS Glue: Creating and running a batch ingestion job with AWS Glue.
- Streaming Data Ingestion with Amazon Kinesis: Setting up and configuring a Kinesis stream.

- Transforming Data with AWS Glue: Designing and implementing an ETL job.

Assessment:

- Quiz on data ingestion and transformation techniques.
- Hands-on project: Implementing a hybrid data ingestion and

5 Hours 1h 1h 2h 1h	**Module 6: Data Marts and Amazon Redshift** **Objective**: Provide advanced knowledge on creating and managing data marts with Amazon Redshift, focusing on optimising performance and cost-efficiency for large-scale data analytics. ● Introduction to Data Marts (1 hour) ○ Definition and use cases. ○ Designing data marts for specific analytics needs. ● Amazon Redshift Overview (1 hour) ○ Key features and benefits. ○ Redshift architecture and components. ● Creating and Managing Data Marts in Redshift (1 hour) ○ Best practices for designing data marts. ○ Loading and querying data in Redshift data marts. ● Optimising Redshift for Performance and Cost (1 hour) ○ Performance tuning and optimization techniques. ○ Managing and reducing Redshift costs. **Lab**: Setting Up a Data Mart in Amazon Redshift ○ Creating and configuring a Redshift cluster. ○ Designing and implementing a data mart. **Assessment**: ○ Quiz on data marts and Amazon Redshift. ○ Hands-on project: Building and optimizing a data mart in Redshift.

Module 7: Orchestrating Data Pipelines and Queries with Athena

5 Hours

Objective: Enable trainees to orchestrate data pipelines and execute efficient queries using Amazon Athena, leveraging serverless technology for big data analytics.

- Orchestrating the Data Pipeline (3 hours)
    - 3h Introduction to Data Pipeline Orchestration:
        - Key concepts and importance of orchestration.
        - Overview of orchestration tools and frameworks.
    - Using AWS Step Functions for Orchestration:
        - Key features and benefits.
        - Designing and implementing workflows with Step Functions.
    - Integrating Multiple AWS Services in a Pipeline:
        - Best practices for integration.
        - Example of an end-to-end data pipeline using Step Functions.
- Queries with Amazon Athena (2 hours)
    - 2h Introduction to Amazon Athena:
        - Key features and benefits.
        - Use cases for querying data in S3.
    - Setting Up and Configuring Athena:
        - Creating and configuring Athena workgroups and databases.
        - Integrating Athena with other AWS services.
    - Writing and Running Queries in Athena:
        - Writing SQL queries to analyze data in S3.
        - Best practices for query optimization.
    - Advanced Querying Techniques:
        - Using complex queries and functions.

Lab: Orchestrating a Data Pipeline with AWS Step Functions

- Designing and implementing a data pipeline.
- Integrating multiple AWS services in the pipeline.

Assessment:

- Quiz on data pipeline orchestration and Athena queries.
- Hands-on project: Building and querying a data pipeline using AWS Step Functions and Athena.

3 Hours	**Module 8: Visualizing Data with Amazon QuickSight** **Objective**: Enable trainees to create interactive and insightful data visualisations using Amazon QuickSight, enabling data-driven decision-making through effective dashboards and reports.
30’	● Introduction to Amazon QuickSight (30 minutes) ○ Key features and benefits. ○ Use cases for data visualisation and business intelligence.
30’	● Setting Up and Configuring QuickSight (30 minutes) ○ Creating and configuring QuickSight accounts and datasets. ○ Integrating QuickSight with other AWS services.
1h	● Creating Visualisations in QuickSight (1 hour) ○ Designing and creating interactive dashboards. ○ Using advanced visualisation features.
1h	● Sharing and Collaborating on Dashboards (1 hour) ○ Best practices for sharing and collaborating on dashboards. ○ Managing user access and permissions.
	Lab: Creating Dashboards in QuickSight ○ Designing and creating interactive dashboards. ○ Using advanced visualisation features. **Assessment**: ○ Quiz on Amazon QuickSight features and visualisation techniques. ○ Hands-on project: Designing and sharing an interactive dashboard using QuickSight.

5 Hours

Module 9: Building Transactional Data Lakes and Implementing Data Mesh

Objective: Instruct on building transactional data lakes and implementing data mesh strategies to support decentralised and scalable data architectures using AWS services.

- Building Transactional Data Lakes (3 hours)
  - Introduction to Transactional Data Lakes (30 minutes)
    - 3h Key concepts and importance.
    - Use cases and benefits.
  - Building a Transactional Data Lake on AWS (1 hour)
    - Key components and architecture.
    - Using AWS services to implement a transactional data lake.
  - Managing Transactions in a Data Lake (1 hour)
    - Best practices for managing transactions.
    - Ensuring data consistency and integrity.
  - Querying and Analyzing Data in a Transactional Data Lake (30 minutes)
    - Using AWS services to query and analyze data.
    - Best practices for data analysis.
- Implementing a Data Mesh Strategy (2 hours)
  - Introduction to Data Mesh (30 minutes)
    - Key concepts and principles.
    - Benefits and challenges.
  - Designing a Data Mesh Architecture (30 minutes)
    - Key components of a data mesh architecture.
    - Best practices for designing a data mesh.
  - Implementing a Data Mesh on AWS (30 minutes)
    - Using AWS services to implement a data mesh.
    - 2h Managing and orchestrating data products.
  - Governance and Security in a Data Mesh (30 minutes)
    - Ensuring data governance and security.
    - Best practices for data governance and security.

Labs:

- Building a Transactional Data Lake on AWS: Creating and configuring a transactional data lake.
- Implementing Governance and Security in a Data Mesh: Implementing governance and security best practices.

3 Hours	**Module 10: Enabling Artificial Intelligence and Machine Learning** **Objective**: Introduce trainees to the integration of AI and machine learning into data engineering workflows, using AWS services like Amazon SageMaker to build, train, and deploy intelligent models.
1h	● Introduction to AI and ML on AWS (1 hour) ○ Overview of AI and ML concepts. ○ AWS services for AI and ML: Amazon SageMaker, AWS Comprehend, Amazon Rekognition. ○ Use cases and benefits of AI and ML on AWS.
1h	● Building and Training ML Models with Amazon SageMaker (1 hour) ○ Key features of Amazon SageMaker. ○ Steps for building and training ML models. ○ Best practices for ML model development.
1h	● Deploying and Managing ML Models on AWS (1 hour) ○ Deploying ML models using SageMaker. ○ Managing model versions and updates. ○ Monitoring and optimising deployed models.
	Lab: Building a Machine Learning Model with SageMaker ○ Setting up a SageMaker environment. ○ Building and training a simple ML model.
	Assessment: ○ Quiz on AI and ML concepts and AWS services. ○ Hands-on project: Building, training, and deploying a machine learning model using Amazon SageMaker.