Data Engineering - Daywise Schedule

Total Hours: 176

Total Duration: 22 days (11 weeks, Weekly 2 days session)

Week #1 - Day 1	Week #1 - Day 2	Week #2 - Day 3
Introduction to Data Engineering	Fast API Integration	Data Virtualization
Pre-assessment	Recap Day 1	Recap Day 2
- Overview of Data Engineering - Setting up a Data Engineering Environment - Role and Responsibilities in Data Analytical Engineering - Pandas in Data Engineering	- Importance in Business - Introduction to Fast API for building APIs - Role of Fast API in Data Engineering Tasks	Introduction to Data Virtualization - Understanding the concept of data virtualization - Advantages and use cases of data virtualization. - Key components and architecture of data virtualization
➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Kickstart**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**
Week #2 - Day 4	Week #3 - Day 5	Week #3 - Day 6
Data Product & Starburst	Data Storage Basics & Introduction to Data Warehousing	Data Pipeline Development
Recap Day 3	Recap Day 4	Recap Day 5
Data Product - What are data products? - Data product's Components - Value proposition of data products - Lifecycle of a data product Introduction to Starburst - Overview of Starburst - Key features and capabilities - Use cases for Starburst in building Data Products	Data Storage Basics - File systems. - Storage formats (CSV, JSON, Parquet) - File Formats – Parquet Introduction to Data Warehousing - Overview of cloud-based data warehouses - Comparison of Redshift, Big Query, and Snowflake - Best practices for migrating data to cloud data warehouses	- Data Processing with Apache Spark and Flink - Cluster Management and Optimization for Spark and Flink - Files to write to Kafka Producer - Kafka Consumer & Stream
➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project -Milestone #1 Demo** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**

Week #4 - Day 7	Week #4 - Day 8	Week #5 - Day 9
Data Pipeline Development	Data Pipeline Development	Data Pipeline Development
Recap Day 6	**Recap Day 7**	**Recap Day 8**
▪ Integration of Spark and Flink ▪ Integration of Apache Airflow	▪ ETL vs ELT ▪ Change Data Capture (CDC) using Oracle Golden Gate	▪ Integration of Oracle Golden Gate with Informatica ▪ Use cases for combining Golden Gate and Informatica ▪ Best practices for data integration using these tools
➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**
Week #5 - Day 10	**Week #6 - Day 11**	**Week #6 - Day 12**
Data Storage Solutions	Data Storage Solutions	Data Storage Solutions & Data Modeling with Apache Iceberg
Recap Day 9	**Recap Day 10**	**Recap Day 11**
ECS, NAS, Share: Read and Write • Understanding ECS (Elastic Cloud Storage), NAS (Network Attached Storage), and file sharing. • Configuring read and write access for different storage solutions. • Security considerations for file sharing Introduction to Data Lakehouse	▪ Characteristics of a Data Lakehouse ▪ Comparison with Data Warehouse and Data Lake ▪ Components of a Data Lakehouse ▪ Implementing Data Lakehouse with Spark and Flink	▪ Introduction to vector databases for storing and querying high dimensional data. ▪ Overview of pgvector and its features ▪ Role of Feast in feature management and serving Data Modeling with Apache Iceberg • Introduction to Apache Iceberg • Overview of Iceberg for data lake management • Advantages of Iceberg
➢ **Demo & Hands-on Exercises** ➢ **Capstone Project -Milestone #2 Demo** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**
Week #7 - Day 13	**Week #7 - Day 14**	**Week #8- Day 15**
Data Modeling with Apache Iceberg	Data Modeling with Apache Iceberg & Automation and Scripting	Automation and Scripting
Recap Day 12	**Recap Day 13**	**Recap Day 14**

▪ Advantages of Iceberg for Data Lake Management
  ▪ Use cases and scenarios.
  ▪ Implementing Iceberg tables in data engineering projects with Spark and Flink
▪ Practical implementation examples
	▪ Best practices for data modeling with Iceberg, integrating Spark and Flink
	▪ Cataloging in Apache Iceberg
	  • Importance of cataloging in data engineering
	  • Utilizing Apache Iceberg for cataloging, considering Spark and Flink integration
	▪ Flask/API for automation
		▪ Introduction to Docker and Kubernetes
		  • Overview of Docker containers and Kubernetes orchestration
		  • Benefits of containerization and container orchestration
		  • Getting started with Docker and Kubernetes

		Basics of Flask framework for web development
		Creating APIs with Flask
		Automation use cases for Flask based APIs

➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project -Milestone #3 Demo** ➢ **Capstone Project Discussion**

Week #8 - Day 16	**Week #9 - Day 17**	**Week #9 - Day 18**
Automation and Scripting	ML Ops Integration	ML Ops Integration
Recap Day 15	**Recap Day 16**	**Recap Day 17**

▪ ControlM for scheduling
  ▪ Introduction to ControlM job scheduling software
  ▪ Configuring and managing jobs with ControlM
  ▪ Integration with other data engineering tools and workflows
▪ Feature Store
  ▪ Concept of feature stores in machine learning and data engineering
  ▪ Benefits of centralizing feature storage and management
  ▪ Implementing a feature store in data engineering pipelines
	▪ Introduction to ML Ops in Data Engineering
	  ▪ Overview of ML Ops (Machine Learning Operations)
	  ▪ Role of data engineering in ML Ops workflows
	  ▪ Best practices for integrating ML Ops into data engineering processes.
	▪ Integration of Machine Learning Models
	  ▪ Incorporating machine learning models into data pipelines
	  ▪ Model deployment strategies and considerations
		▪ Integration of Machine Learning Models
		  ▪ Monitoring and updating ML models in production.
		  ▪ Continuous training
		▪ Model Deployment
		  ▪ Deployment options for machine learning models
		  ▪ Containerization of models for deployment

➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**

Week #10 - Day 19	**Week #10 - Day 20**	**Week #11 - Day 21**
ML Ops Integration	Performance Optimization	Collaboration and Monitoring
Recap Day 18	**Recap Day 19**	**Recap Day 20**

▪ Model Deployment
  Scaling and managing model deployments in production.
Monitoring and Maintenance
  ▪ Importance of monitoring data pipelines and systems
  ▪ Tools and techniques for data pipeline monitoring
  ▪ Model observability.
	▪ Performance optimization strategies
	  ▪ Techniques for improving performance of data processing workflows.
	  ▪ Optimization considerations for different data storage and processing technologies
	  ▪ Monitoring and tuning performance in production environments
		▪ Collaboration with Industry Experts
		  ▪ Importance of collaboration in data engineering projects
		  ▪ Working with domain experts and stakeholders
		  ▪ Knowledge sharing and continuous learning initiatives
		▪ Monitoring data quality and performance

Maintenance best practices for data engineering infrastructure	• Best practices for Nonprod to Prod migration.
	   • Strategies for migrating data pipelines from non
	   • production to production environments
		Techniques for assessing data quality and performance metrics

➢ **Demo & Hands-on Exercises** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises** ➢ **Capstone Project – Final Demo & Presentation** ➢ **Capstone Project Discussion**	➢ **Demo & Hands-on Exercises**

Week #11 - Day 22

Collaboration and Monitoring

Recap Day 21

- Data Observability with Monte Carlo
  - Introduction to data observability and its importance
- Data Observability with Monte Carlo
  - Overview of Monte Carlo data observability platform
  - Using Monte Carlo for detecting anomalies and ensuring data reliability

➢ **Demo & Hands-on Exercises**
➢ **Post Assessment**
➢ **Final Feedback**