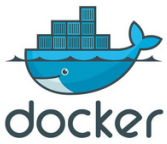
MATRIX

Big Data Əsaslı
DATA ENGINEERING
SYLLABUS | 5 MONTHS























MATRIX

Data Engineering proqramı Matrix Academy məzunlarının Big Data və müasir Data Engineering texnologiyaları əsasında, qlobal standartlara uyğun təhsil aldığı ixtisaslaşmış proqramlarından biridir.

Məzunların çalışdığı şirkətlər :





MATRIX

Big Data Əsaslı
DATA ENGINEERING
SYLLABUS | 5 MONTHS

1.Month

- Introduction to Data Engineering
- Responsibilities of a Data Engineer
- Data Engineering vs Data Science
- NoSQL vs SQL
- RDBMS
- DWH
- ETL
- ELT
- OLTP vs OLAP
- Fundamentals of Data Modeling
- Facts vs Dimensions
- Star Schema
- Snowflake Schema
- Kimball Methodology
- Inmon Methodology
- ER Modeling (1:1, 1:N, M:N)
- Conceptual Data Model
- Logical Data Model
- Physical Data Model
- Normalization (1NF, 2NF, 3NF)
- Denormalization
- Best Practices

MATRIX

2.Month

- Linux File System
- Basic Commands
- User Permissions
- Networking Basics
- Crontab Syntax
- Managing Cron Jobs
- Logging Cron Jobs
- Git Core Commands
- Branching & Merging
- Bash Scripting Basics
- Variables, Conditionals, Loops
- Variables and Data Types
- Conditionals (if, elif, else)
- Loops (for, while, until)
- Script Arguments ($1, $@)
- Best Practices for Script Automation
- Containers vs VMs
- Docker Concepts (Images, Containers, Volumes)
- Essential Docker Commands (run, ps, exec, stop, etc.)
- Writing a Dockerfile (FROM, COPY, RUN, CMD)
- Basic Container Lifecycle

- Image Optimization Techniques
- Multi-stage Builds
- Introduction to Podman
- Docker vs Podman
- Security and Resource Limits in Containers

3.Month

- HDFS Architecture
- NameNode & DataNode
- Replication & Fault Tolerance
- YARN Resource Management
- Job Scheduling
- MinIO Overview
- Buckets and Access
- CSV, JSON, Parquet, ORC
- Delta Lake Features
- Apache Iceberg Features
- Kafka Architecture
- Topics & Partitions
- Producers & Consumers
- Brokers
- Consumer Groups & Offsets
- Message Retention
- Partitioning & Ordering
- Producer Acknowledgments
- Basic Kafka Performance Tuning

MATRIX

4.Month

- Spark Architecture
- RDD vs DataFrame API
- Transformations & Actions
- Spark SQL Basics
- PySpark Setup
- Load & Clean Data
- Transformations & Aggregations
- Write Parquet or Delta
- SQL Queries on DataFrames
- Catalyst Optimizer
- Tungsten Engine
- Explain Plan
- Structured Streaming Intro
- Memory Management
- Read CSV with Spark
- Transformations (filter, groupBy, agg)
- Write Partitioned Parquet
- Structured Streaming: File/Socket Source
- Windowed Aggregations
- Streaming Output
- Spark & Streaming Practice

5.Month

- Airflow Setup
- DAGs & Tasks

MATRIX

- Executors Overview
- BashOperator, PythonOperator, DummyOperator
- Connections Management
- Variables Usage
- Metadata Database
- TimeSensor, FileSensor, ExternalTaskSensor
- EmailOperator, BranchPythonOperator, ShortCircuitOperator
- XCom (push & pull)
- TaskGroups
- Pipeline Design & Modular DAGs
- Dynamic Task Mapping
- What is CI/CD in Data Engineering
- Version Control Integration (Git)
- Deploying DAGs with CI/CD (GitHub Actions, GitLab CI, Jenkins)
- Environment Promotion (Dev → Staging → Prod)
- Best Practices in Airflow DAG Deployment
- Real DAG Creation
- DAG Scheduling & Monitoring
- Grafana Dashboards
- Prometheus Metrics
- Loki Logs
- Complete Data Pipeline
- Scheduling, Streaming, Monitoring
- Presentation & Documentation