Big Data and Cloud Administration

הכשרה מתקדמת תחום הדאטה לאנשי IT, DevOps ו-DBA

250 שעות לימוד אקדמיות

תיאור התפקיד:

מידע (Data) הוא המשאב היקר והחשוב ביותר של ארגונים בעולם של היום. ארגונים אוספים ומנהלים יותר מידע מאי פעם בעבר, מבצעים בו שימושים מתקדמים לצרכי תפעול, בינה עסקית (BI) ולימוד מכונה (Data Science), והטכנולוגיות המשמשות לניהול המידע עוברות התפתחות מואצת.

מנהלי מסדי נתונים (DBA), אנשי תשתיות (IT), אנשי DevOps שצמחו ועוסקים בעולם ה-Data המסורתי וצברו ידע ונסיון בניהול מידע במערכות רלציוניות מסחריות (RDBMS) הממוקמות בחוות השרתים של הארגון (On-premise), עומדים כיום מול האתגר של שלל טכנולוגיות ניהול מידע חדשות ומגוונות שארגונים מאמצים כמענה לצרכי ניהול המידע המתקדמים והאתגרים העסקיים העומדים בפניהם.

שלושת המגמות הטכנולוגיות המובילות בעולם ה-IT של היום הן:

- **אימוץ הולך וגובר של *טכנולוגיות קוד פתוח* (OSS)**, ובכללן בסיסי נתונים רלציוניים המבוססים על קוד פתוח, כגון **MySQL** ו-**PostgreSQL**. בסיסי נתונים אלה מציעים פתרון טכנולוגי מתקדם, העונה על צרכי ניהול המידע של הארגון בעלויות נמוכות משמעותית מאלו של בסיסי נתונים מסחריים הדורשים רישוי יקר.
- **מעבר של ארגונים ומערכות *לענן הציבורי* (Public Cloud)** ושימוש בשירותים מנוהלים (Managed Services) של ספקים כגון **Amazon (AWS), Google (GCP) או Microsoft (Azure)** עבור טווח רחב ביותר של שירותים, כגון אחסון וניהול מידע, אירוח אפליקציות, מהלכי עיבוד וניתוח מידע וכך הלאה. שימוש בטכנולוגיית ענן מאפשר לארגונים להשתמש בטכנולוגיות המתקדמות ביותר באופן גמיש ודינמי תוך חסכון משמעותי בעלויות חומרה וכח אדם.
- **שימוש נרחב *בטכנולוגיות Big Data*** לשם אחסון ותחקור מידע בנפחים עצומים, לעיתים לא מובנה (Unstructured), המגיע בקצבים גבוהים של עד מליוני אירועים בשניה, שבסיס נתונים רלציוני לא מסוגל להתמודד איתם. טכנולוגיות אלו משתמשות במערכות מבוזרות ומרובות שרתים כדי לתמוך בדרישות הקצה של מערכות מהדור החדש, כגון מערכות **IOT**, מחסני נתונים גדולים ו-**Data lakes**, מערכות **Web** גלובליות ושלל מערכות נוספות בארגון הזקוקות לאחסון ועיבוד מידע בנפחים גדולים וקצבים מהירים.

ככל שתחום ניהול המידע מתרחב ומתפתח, כך צומחים גם תחומי הידע והמקצועיות הדרושים ומצופים מאנשי דאטה המעוניינים לתת מענה רלוונטי ומשמעותי לארגון, ולצעוד יחד עם מערכות המידע אל העתיד.

למרות כל זאת, מקומם של בסיסי הנתונים הרלציוניים המסחריים הותיקים, כגון Oracle ו-SQL Server, עדיין מובטח והם אינם צפויים להיעלם בעתיד הנראה לעין. הם עדיין פופולריים ונפוצים ברוב הארגונים, ולכן הכשרה זו כוללת מודול של נושאים מתקדמים בעבודה מול בסיסי נתונים אלו, כגון SQL מתקדם ושיפור ביצועים, כך שבוגרי הקורס ידעו לעבוד ברמת המקצועיות הגבוהה ביותר גם מול בסיסי נתונים אלו.

תיאור ההכשרה:

קורס **Big Data and Cloud Administration** של נאיה קולג' "צמח מהשטח" ופותח בהתאם לדרישות הקיימות כיום בשוק העבודה עבור מנהלי מסדי נתונים (DBA), אבל גם אנשי IT ו-DevOps העוסקים הרבה בתחום הדאטה. הכשרה זו מותאמת לאנשי מקצוע מנוסים בתחום הדאטה המגיעים מעולם בסיסי נתונים רלציוניים מסחריים, כגון Oracle, SQL Server ו-DB2, ומעוניינים להרחיב את כישוריהם לטכנולוגיות של עולם המידע החדש. ההכשרה מותאמת לדרישות התפקיד ולכן הינה מגוונת וכוללת היכרות עם טכנולוגיות מידע וכלים שונים.

הקורס מתחיל במפגש היכרות עם עולמות ה- Data החדשים, ומתאר את הסיבות למעבר אליהן מהטכנולוגיות המסורתיות. נסקור את המגמות המובילות בעולם ניהול ה-Data של היום, ונפגוש את הטכנולוגיות הבולטות בעולמות הקוד הפתוח (Open Source), הענן (Cloud) ונְתוּנֵי העָתֵק (Big Data) השולטות כיום בשוק המידע.

נמשיך **למודול המבינה לשם "יישור קו"**, בו נסקור את כל נושאי הליבה של **ניהול בסיסי נתונים מסורתיים, גם מההיבט האפליקטיבי וגם מההיבט התשתיתי**. זהו בסיס חשוב שכן בהמשך הקורס נלמד כיצד הטכנולוגיות החדשות נותנות מענה לדרישות מגוונות אלו הקיימות בכל מערכת לניהול מידע. בנוסף, נצלול לנושאים מתקדמים בעבודה מול בסיסי נתונים רלציוניים, כגון עיצוב נכון של סכמה, SQL מתקדם, שיפור ביצועים ו-New features של הטכנולוגיות המובילות בשוק. חלק נוסף ממודול הכנה זה הוא היכרות עם מערכת ההפעלה Linux, המותאם גם למתחילים המגיעים מעולמות Windows, והיכרות עם

הטכנולוגיה של **Dockers, Containers ו-Kubernetes**, שהיא נפוצה ופופולרית מאד כיום בעולם מערכות המידע.

המודול השני של הקורס מתמקד בניהול בסיסי נתונים רלציוניים מבוססים קוד פתוח (OSS), תוך התמקדות בבסיס הנתונים MySQL. מודול זה מבוסס על **קורס MySQL** המלא של נאיה קולג' וכולל את כל נושאי הליבה האפליקטיביים והתשתיתיים הדרושים למנהל בסיס נתונים בסביבה זו. לאחר שנכיר לעומק את העבודה עם MySQL, נקדיש שיעור נוסף גם לבסיס הנתונים הפופולרי **PostgreSQL**.

המודול השלישי צולל לעולמות ה-**Big Data** ומספק היכרות טכנולוגית ומעשית (Hands-on) עם הטכנולוגיות המובילות בעולם ה- Big Data של היום – Apache Hadoop, NoSQL Databases ומנועי חיפוש (Search engines). מעבר להיכרות עם העולמות הטכנולוגיים הכלליים, נתמקד בכלים המובילים היום בשוק שהם **Cloudera Hadoop, Apache Cassandra, MongoDB** ו-**Elasticsearch**. בסוף המודול נפגוש ונתרגל את הטכנולוגיה הפופולרית של Apache Kafka, שנמצאת בלב הרבה מתשתיות ה-Data כיום, ומשמשת כפלטפורמה מרכזית להזרמת מידע בין מערכות בארגון.

המודול הרביעי עוסק בטכנולוגיות מחשוב ענן (**Cloud computing**). כאן נכיר את מושגי היסוד של התחום ונתוודע לשירותים השונים הזמינים לנו בענן הציבורי. מודול זה מתמקד ב-**Amazon Web Services**, שהוא הנפוץ ביותר כיום, ובשירותי הליבה שהוא מציע בגון S3, EMR, RDS, Athena, Redshift וביוצא באלה. בסוף המודול נקדיש שיעור נוסף להיכרות מעמיקה יותר גם עם שירותי הענן של Google ושל Microsoft ושיעור נוסף המתמקד במתודולוגיות וכלים עבור הסבת בסיסי נתונים לענן הציבורי (**Cloud Migration**), נושא שחברת נאיה טכנולוגיות מתמחה בו באופן מיוחד (כולל פיתוח כלי בשם MigVisor).

המודול החמישי והאחרון מרחיב את היריעה ומספק היכרות עם תחומי התמחות משיקים לזה של ה-DBA. כאן נתוודע למושגי יסוד, מתודולוגיות עבודה, טכנולוגיות וכלים של שלושה מבעלי התפקידים המרכזיים איתם עובד ה-DBA, ופעמים רבות לוקח על עצמו משימות מעולם התוכן שלהם: מהנדס המידע (**Data Engineer**), ארכיטקט המידע (**Data Architect**) ומפתח הקוד בסביבת Big Data (**Big Data Developer**).

בנוסף ללימוד התיאורטי והתרגולים השוטפים בקורס במודולים השונים, הקורס כולל פרוייקט Hands-on מקיף בו תוכלו להתנסות במכלול הטכנולוגיות הנלמדות. הפרוייקט מסייע לבוגרי הקורס לבצע אינטגרציה לישורים והידע שאספו במודולים השונים באמצעות תפירת פתרון מהעולם האמיתי מקצה לקצה. באופן זה ההכשרה מבינה את בוגריה באופן מיטבי להתמודדות עם האתגרים בהם יתקלו בהמשך דרכם המקצועית.

קהל יעד:

DBA מנוסים, וכן אנשי IT , DevOps, ומפתחים בעלי ידע עשיר ונסיון רב במערכות מידע ומסדי נתונים רלציוניים.

דרישות קדם:

- ידע וניסיון בתחום ה-DATA.
- נדרשות מספר שנות ניסיון לפחות בתחום **הכולל היכרות טובה מאוד עם בסיסי נתונים ושפת SQL.**
- הקבלה לקורס מותנית במעבר ראיון אישי ומבחן קבלה.

תכנית הלימודים:

Data Technologies Introduction for DBAs

- **Introduction to NextGen DBA**
  - History of Database Systems and the traditional on-premise RDBMS
  - What has changed and why
- **OSS and Open Source databases**
  - MySQL, PostgreSQL and others
- **Big Data and the Internet of Things**

- Introduction to Big Data technologies
- Apache Hadoop – Core and Ecosystem
- NoSQL Databases – Key/Value, Wide-column, Object and Graph
- Leading NoSQL technologies – Redis, Cassandra, MongoDB and Neo4j.
- Search engines and Elasticsearch

- **Cloud computing**
  - Concepts and service models
  - Leading cloud providers – Amazon, Google and Microsoft
  - Core cloud technologies and services

Module I – Preparation Module – RDBMS DBA and Linux

- **Application DBA core concepts**
  - ERD and Schema design concepts – Tables, data types and constraints
  - Advanced table design – Partitions, compression and storage structures
  - Advanced SQL tools and techniques (Including analytical functions)
  - SQL Statement tuning
  - Using Indexes to increase performance
  - The optimizer and execution plans
  - Database programming and code objects - Procedures, Functions and Triggers
- **Infrastructure DBA core concepts (Overview)**
  - Security - Users, privileges, roles and encryption
  - Automation - Jobs and alerts
  - Backup and Recovery - tools and techniques
  - Performance monitoring and tuning - tools and techniques
  - High availability - tools and techniques
- **Basic Linux**
  - History & introduction
  - Connection & Man Pages
  - Files & Directories
  - Files Contents
  - Commands and Arguments
  - Shell Variables
  - Introduction to vi
  - Bash Scripting
  - Pipes and Commands
- **Using Dockers and Containers, including Kubernetes**
  - Understanding Dockers and Containers architecture and usage
  - Common use-cases for Dockers and Containers
  - Common operations with Dockers and Containers
  - Working with Kubernetes

Module II – Open Source Databases and MySQL

- **Introduction to OSS** (Open Source Software)

- **MySQL - General Overview and Basic Elements**
  - **Overview**
    - Overview
    - Real Examples and Use Cases of Customers Using MySQL Database
  - **Architecture Overview**
    - Sever
    - Client
  - **Storage Engines**
    - Engines Types
    - Engines Characteristics
    - InnoDB and MyISAM Extensive Overview
  - **Networking and Security**
    - Operating System
    - File System Security
    - Network Security
  - **MySQL with Linux Overview**
    - Basic Commands
  - **Installation (Linux Only)**
    - Linux RPM Installation
    - Linux Binary Installation
    - Source Installation
  - **Users Management**
    - User Account Management
    - User Privileges
    - Administrative Privileges
    - Database Access Privileges
- **Managing MySQL databases**
  - **Database Management**
    - System Variables
    - Altering Database Variables at Session / Global Level
  - **Configuration File Best Practices (my.cnf)**
    - InnoDB Memory Buffers
    - Log file size and log buffer
    - Memory Consumption
    - Memory Limitations
  - **Using MySQL Built-In Tools**
    - MySQLAdmin
    - MySQLBinlog
    - MySQLDump
    - MySQLImport
  - **Table Maintenance**
    - Analyze Table
    - Check Table
    - Checksum Table
    - Optimize Table

- Repair Table
- Moving Tables between Schemas / Servers / Tablespaces

- **Locking in MySQL**
  - MySQL Locking Overview
  - Table Level Locking
  - Row Level Locking
  - Internal / External Locking
  - Dead Locks
- **Backup and Recovery**
  - MySQL Backup
  - MySQL Backup Types
  - MySQL Backup and Recovery Tools
  - Load data infile & select into outfile
- **Basic Replication**
  - Replication Architecture
  - Replication Basics
  - Replication Types
  - Master-Slave Replication Setup
  - Master-Slave Replication Filters
- **MariaDB and Galera Replication Architecture**
  - Master-Master Replication
- **Upgrading**
  - Pre Upgrade Steps
  - Upgrade Steps and Check List
  - Post Upgrade Commands

- **Monitoring and performance tuning in MySQL**
  - **Monitoring Database Activity**
    - Working with Information Schema and Performance Schema
  - **Various Logs in MySQL**
    - Error Log
    - Query Log
    - Slow Query Log
    - Binary Log
    - Relay Log
    - General Query Log
  - **Identifying Slow Queries**
    - Slow-Query-Log
    - ProcessList
    - InnoDB_TRX
    - Using Show Warnings
  - **Optimizing Queries**
    - Explaining the EXPLAIN
    - Server Variables Check
  - **Development**
    - Basic MySQL Queries

    - Join / Sub-Query / Union
    - Build-In Functions
    - User Defined Functions
    - Procedures
- **MySQL 8.0 – Selected New Features**
  - Document Store (NoSQL)
  - InnoDB Clusters
  - Additional new features for better usability and performance
- **Managing PostrgreSQL databases – Extended Overview**
  - Introduction to PostgreSQL
  - Administrating PostgreSQL
  - Programming in PostgreSQL

Module III – Big Data and NoSQL Technologies for DBAs

- **Apache Hadoop – Core, ecosystem and analytics**
  - **Hadoop core** – Distributed storage (HDFS) and processing (MapReduce/Spark)
  - **Ingesting data into Hadoop**
    - Using Flume to ingest data from various sources in real time
    - Using Sqoop to ingest data from Relational databases
  - **Using Hive to query HDFS data with SQL**
    - Designing Hive tables
    - Using HQL (Hive SQL language)
    - File types and storage options
    - Using SerDes and Querying unstructured data
    - Implementing partitioned tables
  - **Using Impala to query HDFS data with SQL**
  - **Additional ecosystem technologies – Overview**
    - Using Cloudera Data Flow (Apache Nifi) to integrate data from various sources
    - Using Apache Kafka to stream data into Hadoop
    - Implementing Kudu as a columnar storage solution
    - Using Oozie as a workflow manager and job scheduler
- **Administrating a Hadoop cluster**
  - **Hadoop Cluster Installation (Overview)**
    - Working with Cloudera Manager – the leading Hadoop cluster management tool
    - Deploying a Cloudera Hadoop cluster – Steps and best practices
  - **The Hadoop Distributed File System (HDFS)**
    - HDFS architecture
    - Main components of HDFS
    - Working with HDFS, including WebUIs and the Hadoop File Shell.
    - HDFS Security (Overview)
    - HDFS High Availability
  - **MapReduce and Spark on YARN**
    - Understanding Hadoop's computation frameworks
    - MapReduce – core concepts and architecture
    - Apache Spark – core concepts and architecture

        - Working with YARN: Hadoop's resource manager
    - **Hadoop Configuration and parameters**
        - Finding parameters and applying configuration changes
        - Understanding services, roles, role instances and role groups
        - Important parameters to know about
    - **Hadoop Security (Overview)**
        - Core security concepts in Hadoop
        - Applying authentication using Kerberos
        - Authorization using ACLs and roles
    - **Managing Cluster Resources**
        - Prioritizing tasks and groups for cluster resources
        - The Fair Scheduler
        - Static service pools and Dynamic Resource Pools
    - **Cluster Maintenance**
        - Checking HDFS and node status
        - Entering and leaving SafeMode
        - Rebalancing the cluster
        - Upgrading to cluster
    - **Backup and recovery**
        - Copying Data Between Clusters
        - Working with HDFS snapshots
        - HDFS Replication
    - **Cluster Monitoring**
        - Monitoring Hadoop Clusters
        - Health and configuration alerts
        - Cloudera manager monitoring tools – dashboards, charts and metrics
- **MongoDB Document store (NoSQL databases)**
    - Introduction to MongoDB and use cases
    - Understand documents and collections
    - Querying MongoDB database
    - Creating, Reading and Updating Data (CRUD)
    - Understanding MongoDB replication and sharding
    - Performance monitoring and tuning
    - Scalability
    - Backup and Recovery

- **Apache Cassandra wide-column store (NoSQL databases)**
    - **Introduction to Cassandra and wide-column stores**
        - Use cases for Cassandra
        - Apache Cassandra vs. DataStax Cassandra
    - **Cassandra architecture and configuration**
        - The Cassndra cluster (ring)
        - Sharding and replication in Cassandra
        - Memtables , SSTables and Compaction

        - Configuration files and main settings
    - **Cassandra Schema design**
        - Keyspaces and tables
        - Primary keys, Partition keys and Clustering columns
        - Single-row and multi-row partitions
        - Design principals and Chebotko diagrams
    - **Creating, Reading and Updating Data (CRUD)**
        - CQL (Cassandra Query Language)
        - Inserting, updating and deleting rows
        - Consistency levels and eventual consistency
        - Using Light-Weight transactions
    - **Performance monitoring and tuning**
        - Monitoring performance statistics
        - Bloom filters
        - Secondary indexes
        - Materialized views
    - **Cassandra Tools**
        - Nodetool CLI
        - GUI tools
- **Elasticsearch and the Elastic Stack (Search Engines)**
    - **Introduction to search engines and Elastic Stack overview**
        - Apache Lucene and search engines
        - Elasticsearch and the Elastic Stack
        - Real-world use-cases
    - **Basic setup and configurations**
        - Installing Elasticsearch
        - Configuring Elasticsearch
        - Securing Elasticsearch
    - **Elasticsearch architecture**
        - Node types – Master and Data nodes
        - Sample architectures
        - Shards
        - Distributed writes and search
    - **Schema design**
        - Using Lucene indexes
        - Text analysis and custom mapping configurations
        - Understanding Analyzers and Filters
    - **Querying Elasticsearch**
        - APIs
        - Simple queries and complex queries
        - Using Aggregations
        - Tips and tricks for better search performance
    - **Troubleshooting Elasticsearch**
        - Monitoring and diagnosing cluster health
        - Diagnosing Shard issues

    - Best practices
  - **The Elastic Stack - Overview**
    - Using Logstash to ingest data into Elasticsearch
    - Using Kibana to visualize data
    - Using Beats for send data from edge machines
- **Apache Kafka**
  - Introduction to Kafka
  - Installation & configuration
  - Kafka components – Consumer, Producer and Brokers
  - Configuring Kafka operations on topics
  - Kafka integration with various consumers
  - Kafka APIs
  - Kafka cluster configuration and balancing

Module IV - Cloud Technologies for DBAs

- **Cloud Computing – Core concepts**
  - Introduction to cloud technologies
  - Service models (Iaas, Paas, Saas)
  - Leading public cloud vendors – Amazon Web Services, Microsoft Azure and Google Cloud Platform.
  - Decoupling Storage and Compute
- **Amazon Web Services (AWS)**
  - Introduction to AWS
  - Introduction to AWS Products
  - Networking and security in AWS
  - AWS RDS & Aurora
  - Amazon Redshift
  - Amazon EMR
  - Amazon Athena
  - Migrating on-premise databases to the cloud using Amazon SCT and DMS
- **Cloud migration**
  - Migrating database from on-premise to the public cloud
  - Tools, techniques, methodologies and best practices
  - Real-world examples and projects
  - Naya Technologies' MigVisor and Amazon SCT/DMS (Schema Conversion Tool & Data Migration Service)
- **Google Cloud Platform (GCP) and Microsoft Azure – Overview**
  - Identify parallel (similar) services
  - Map main differences and advantages of different cloud providers

Module V – Additional technologies and skills

- **Data Engineer**
  - Concepts, methodologies and tools
  - Building data pipelines between various data sources
  - Leveraging Python for Data Engineering
  - Cloudera Data Flow and Apache NiFi

- Leading technologies and best practices
- Real world examples

- **Data Architect**
  - Concepts, methodologies and tools
  - Designing a complex next-generation systems
  - Analyzing needs, requirements and limitations to design an optimal architecture
  - Best practices and real-world examples
- **Big Data Developer**
  - Concepts, methodologies and tools
  - Using Apache Spark for distributed processing and cloud processing
  - Python, Java, R and Scala
  - CI/CD – Continuous Integration / Continuous Deployment