מהנדס מידע - Big Data Engineer

320 שעות לימוד אקדמיות

תיאור התפקיד:

ככל שהאנושות צועדת עמוק יותר לתוך "עידן המידע" והארגונים מאמצים את הטכנולוגיות המתקדמות ביניהן גם טכנולוגיות ה-Big Data, עולה הצורך באנשי מקצוע שידעו לארגן את המידע הנרחב והמגוון שנאסף מהערוצים השונים ולהתאים אותו לצרכי הארגון.

בשנים אחרונות מחקרים מראים כי ארגונים מפסידים בממוצע 9.7 מיליון דולר בשנה כתוצאה מאיכות נתונים ירודה. נתונים לא מהימנים דורשים זמן רב מאוד לטיוב והכנתם לתחקור ע"י מדעני נתונים ואנליסטים, וזה יכול להגיע לכמעט 80% זמן השקעה מהזמן הכולל – מדובר בשימוש לא מיטבי בכישוריהם.

אבל איך ארגונים יכולים להרגיש בטוחים כי מקור הנתונים שלהם הוא אמין מבלי להבין תחילה מהו סוג הנתונים שנכנסים למערכת וכיצד להוציא אותם.

מהנדס המידע, או ה - **Data Engineer**, ממלא תפקיד מפתח משמעותי ביותר בכל ארגון בו הוא משתלב. מהנדסי המידע מעצבים, מפתחים ומנהלים פתרונות לשינוע הנתונים (בניית Pipelines), תוך כדי עבודה עם טכנולוגיות Big Data מגוונות ומתקדמות. עליהם לשלוט בטכנולוגיות המידע השונות, כולל בסביבות הענן, בכלים המובילים לשינוע ועיבוד מידע בנפחים גדולים ובשפת התכנות המובילה היום בעולמות ה- Data Engineering - שהיא Python.

למעשה מהנדס מידע מבין מהם הנתונים הנכנסים לארגון מכל מקור שהוא, עובד עם נתונים גולמיים המכילים טעויות ושגיאות, או מכילים רשומות חשודות ולא מזוהות בפורמט שלהן, הוא מזהה מאיזה מקורות נכנס המידע, איזה סוג נתונים, הוא עוסק בחילוץ הנתונים בתבניות שמישות, ומוודא שהנתונים נטולי שגיאות וטעינתם תקינה עבור מדעני נתונים ואנליסטים. מהנדסי הנתונים יצטרכו להמליץ ולפעמים ליישם דרכים לשיפור אמינות הנתונים, יעילותם ואיכותם. כדי לעשות זאת, הם יצטרכו להשתמש במגוון שפות וכלים, הם יהיו מומחים בבנייה ותחזוקה של מערכות מבוססות נתונים התומכות בפעילות האנליטית והעסקית של הארגון. אף על פי שהם אינם מתפארים בכישורים מתמטיים שמדען נתונים ישתמש בהם, מהנדסי הנתונים יעשו את רוב העבודה הנדרשת כדי לתמוך בעומס העבודה של מדען הנתונים.

תיאור ההכשרה:

קורס Data Engineer של נאיה קולג' "צמח מהשטח" ופותח בהתאם לדרישות הקיימות כיום בשוק העבודה עבור Data Engineers. הכשרה זו מותאמת לדרישות התפקיד ולכן הינה מגוונת וכוללת היכרות עם טכנולוגיות מידע וכלים שונים.

הקורס מתחיל בסקירה של עולמות ה-Data המסורתיים וטכנולוגיות ענן ו-Big Data חדישות השולטות כיום בשוק המידע. ולאחר מכן , בשיעור נוסף, עוסק במהותו של תהליך ETL המוכר בסביבת בסיסי נתונים רלציוניים במטרה ליישר קו בהבנה נכונה של התהליך והאתגרים שעומדים בפני מהנדסי מידע.

מנקודה הבאה אנחנו נכנסים לשלב מהותי בקורס בעוסק בכלים טכנולוגיות Big Data וענן, במטרה להכיר את הארכיטקטורה של הכלים האלה, אופן קליטה ושמירה של נתונים, שליפה ועיבוד הנתונים ואינטגרציה ביניהם וכן יצירת תהליכי אוטומציה בזרימת הדאטה.

המודול מתחיל קודם כל מלימוד מעשי של מערכת הפעלה Linux ואז מתקדם לטכנולוגיות רלוונטיות וביניהן Hadoop, Hbase, MongoDB, Kafka , כולל תרגול תוך כדי השיעורים. את המודול מסכם נושא הענן AWS – Amazon.

מודול ליבה השני של הקורס – האפליקטיבי - מתמקד בלימוד מעמיק של Python ככלי עבודה מרכזי עבור מהנדס המידע בביצוע עיבודים שונים של נתונים – Data Processing וכתיבת Data Pipeline. נראה כיצד Python והחבילות ה"עוטפות" שלו (Python APIs), מאפשרים לנו לאסוף מידע ממקורות שונים (בסיסי נתונים, אינטרנט וכלי Big Data) ולעבד אותו, לנתח ולהציג אותו בדרכים מגוונות. לאחר שנבין את הממשקים הבסיסיים הקיימים ב- Python לתהליכי עיבוד מידע נתייחס לסביבת Spark ונכיר כיצד PySpark מאפשרת לנהל את ה-Pipeline.

בסופו של דבר, במהלך הקורס הסטודנטים ייחשפו לכ-30 טכנולוגיות שונות, השכיחות והחמות ביותר בארגונים, כך שירכשו ידע מעשי ופרקטי להובלת תהליך בניית תשתית לתעבורת הנתונים באופן מיטבי. בנוסף ללימוד התיאורטי-הדגמתי והתרגולים השוטפים בקורס במודולים השונים, הקורס כולל פרוייקט Hands-on מקיף בו תוכלו להתנסות במכלול הטכנולוגיות הנלמדות. הפרוייקט מסייע לבוגרי הקורס לבצע אינטגרציה לכישורים והידע שאספו במודולים השונים באמצעות תפירת פתרון Big Data

מהעולם האמיתי מקצה לקצה. באופן זה ההכשרה מכינה את בוגריה באופן מיטבי להתמודדות עם האתגרים בהם יתקלו בהמשך דרכם המקצועית. מפגשים אחרונים בקורס יוקדשו לפרויקט מסכם והצגתו בפני המשתתפים.

קהל יעד:

מפתחים, DBA, אנשי BI, אנשי IT, מומחי DevOps, מנהלי מוצר טכנולוגי, מנהלי פרויקטים טכנולוגיים ובעלי ידע ונסיון במערכות מידע ומסדי נתונים רלציוניים.

דרישות קדם:

- ידע וניסיון בתחום ה-DATA של מספר שנים לפחות **הכולל עבודה עם בסיסי נתונים והיכרות מעמיקה עם שפת SQL.**
- הקבלה לקורס מותנית במעבר ראיון אישי ומבחן כניסה.

תוכנית הלימודים:

Big Data Technologies Introduction for Data Engineers

- Hadoop Eco-System
- NoSQL
  - Introduction to NoSQL
  - Main characteristics of NoSQL solutions
  - The four types of NoSQL solutions, leading technologies and real-world use cases
- Technologies and trends in the world of Big Data
  - Search engines (Elastic search, SOLR)
  - Cloud Computing
  - NewSQL databases (Vertica, VoltDB, MemSQL)

Data Modelling Concepts

- Database Modelling for OLTP, DWH and OLAP
- Advanced ETL concepts and techniques

Big Data and NoSQL for Data Engineer

- **Basic Linux**
  - Connection & Man Pages
  - Files & Directories
  - Files Contents
  - Commands and Arguments
  - Shell Variables
  - Introduction to vi
  - Bash Scripting
  - Pipes and Commands
- **Docker Basics**
  - Introduction to Docker technologies
  - Docker Vs Virtual Machines
  - Docker Main Components and Terminology
  - Docker Architecture
  - Introduction to Docker Hub
  - Docker basic commands

- Kubernetes High-level introduction

- **Hadoop Infrastructure and Ecosystem Services**
  - Data Engineering Introduction
  - Hadoop core and Eco-system
  - HDFS - Architecture and Read & Write Patterns
  - Using Sqoop to ingest data from Relational databases
  - Using Hive to query HDFS data with SQL
  - Using Impala to query HDFS data with SQL
- **Apache Kafka**
  - Introduction to Kafka
  - Installation & Configuration
  - Kafka Architecture and Components
  - Kafka APIs and Kafka-Connect
  - Kafka-Connect Workshop
- **NoSQL - HBase & MongoDB**
  - NoSQL Introduction
  - Introduction to HBase
  - Apache HBase introduction and solution overview
  - Apache HBase use-cases and considerations
  - MongoDB introduction and architecture
  - MongoDB components, deployment, indexes
  - MongoDB Transaction, Replication and Sharding
  - MongoDB CRUD and Administrative Commands
- **Search Engines - Elastic Stack (previously known as ELK)**
  - Introduction to search engines
  - Introduction to Elasticsearch and typical use-cases
  - Basic setup and configurations
  - Indexes (Lucene) and mapping configurations
  - Querying: APIs, queries and complex queries, aggregations
  - Introduction to Logstash
  - Introduction to Kibana
  - Introduction to Beats

Cloud Technologies for Data Engineers

- **Cloud Computing**
  - Introduction to cloud technologies and types of cloud computing
  - Introduction to AWS / Azure / Google Cloud
- **AWS**
  - Introduction to AWS
  - Introduction to AWS Services
  - Amazon EC2
  - Amazon S3 as a Data Lake
  - Amazon RDS
  - Amazon Redshift
  - Amazon EMR

- Amazon Athena
- AWS Glue
- AWS Lambda
- Amazon Kinesis (Data Streams, Delivery streams, Analytics applications)

Python Programming and Tools for Data Engineers

- **Python Programming**
  - Working environments (Python, Anaconda, Jupyter, PyCharm, etc.)
  - Data types (numbers, strings, Booleans, etc.)
  - Data collections (lists, dictionaries, etc.)
  - Flow control (if, for, while, etc.)
  - Textual interface (input and formatting)
  - Functions
  - File-like objects
  - Object-Oriented programming (OOP)
  - Error handling (try and except)
  - Beyond the built-in library
  - Python Tools for Data Engineers
    - Pandas
    - Json
    - Regular expressions
- **BigData APIs**
  - PyArrow (HDFS)
  - PyHive (Hive)
  - Ibis (Impala)
  - Kafka-python (Kafka)

Apache Spark and PySpark

- Spark core
- Spark SQL
- Spark streaming and Spark Structured streaming

Extra Topics - Overview

- **Apache AirFlow**
  - Introduction
  - Architecture
  - Implementation
- **Apache NiFi**
  - Introduction
  - Architecture
  - Implementation
- Machine Learning for Data Engineer

Final Project