ניהול פרויקטים וניתוח מערכות בעולם ה BIG DATA

קורס 2425 – 40 שעות

אודות הקורס

ארגונים רבים שוקלים לעבור לעולם ה-Big-Data הם נתקלים במושגים חדשים רבים אינספור פלטפורמות, ספקי שירותים, וטכנולוגיות חדשות. פערי הידע של הארגונים אינם מתנקזים לכדי נושא אחד אלא מתפרשים על פני כמה תחומים שכל אחד מהם הוא עולם בפני עצמו. לשם כך על מנהלי הפרויקט/ מנתחי המערכות וראשי הצוותים נדרשים לדעת ולהכיר עולם זה, בשלב הראשון יש צורך להבין את היתרונות העסקיים הטמונים בעולם זה, להכיר ברמה גבוהה את המרכיבים של עולם זה. מידע זה נדרש כיום לכל מי שעוסק בתחום המחשוב הארגוני

האתגרים העיקריים הניצבים כיום בפני ארגונים השוקלים או נמצאים בתהליך של מעבר לעולם ה--Big Data:

- הבנת הצורך והערך העסקי בעולם הביג דטה
- הבנה ממעוף הציפור: מה הן מושגים והמונחים המרכיבים את עולם ה Big-Data
- אפיון הרכיבים המשתייכים לעולם זה לעומת אילו היכולים להיעזר בסביבות המסורתיות
- איתור ניקוי ואיסוף המידע
- אחסון, המידע הרב והלא מובנה בבסיסי נתונים מהסוג החדש
- ניתוח ועיבוד מידע לא-מובנה ורב (non-structured data)
- הצגתו בכלים פשוטים ובעלי משמעות אסטרטגית לארגון
- היכולת לבצע ניתוחים סטטיסטים , כריית נתונים וביצוע למידה ממוחשבת (Machine learning)

מבנה הקורס :

1. הבנת הקשר וההבדל בין עולם ה BI לעולם ה BIG DATA
2. הקמת פרוייקט ביג דטה המורכב מ 5 שלבים:
   - Data exploration
   - Data cleansing
   - Data preparation
   - Data analyze
   - Data visualization

במהלך הקורס נלמד את עולם הביג דטה דרך שלבי ניהול אפיון ומימוש הפרוייקט
פרט לפרויקט הקבוצתי, במהלך כל הקורס יתרגלו החניכים את הנושאים הנלמדים בתרגול מעשי.

תוצרים לחניכי הקורס

- יכולת איתור מקורות המידע אפשריים, כולל מתודולוגיה והכלים לאיסוף הנתונים ממקורות מידע אלו, טיפול בניקיון הנתונים.
- שיקולים לבחירת בסיס נתונים (NoSql) ומידול מודל הנתונים
- הבנה היכולת ניתוח ועיבוד מידע לא-מובנה ורב (non-structured data)
- היכרות וכלים פשוטים להצגת הנתנים ובעלי משמעות אסטרטגית לארגון
- יכולת לבצע ניתוחים סטטיסטים, כריית נתונים וביצוע למידה ממוחשבת (Machine learning)

כלים:

- תבניות אפיון וניהול ככלי עזר למנהל הפרוייקט ומנתח המערכות
- כלי ניהול משימות בעולם נתוני עתק כדוגמת TRELLO

מטרות הקורס

הבנת העולם החדש, הכרת מונחים ומושגים ובהמשך **נלמד לנתח ולנהל פרוייקטי ביג דטה**. נלמד כיצד לנצל מערכות פשוטות וזולות לאגירת נתונים גדולים, כיצד לבחור , להגדיר ולנצל את היתרונות של סוגי מערכות חדשים (NoSQL) וכיצד להתחבר למקורות מידע מתוך הארגון , מהעולם הנייד ומתוך רשתות חברתיות

קהל יעד

אנשי מערכות מידע כדוגמת מפתחים ,מנתחי מערכות, ראשי צוותים, מנהלי פרויקטים, אנשי QA, מומחי יישום ורפרנטים, בעלי ניסיון מעשי בעולם מערכות המידע ומעוניינים להכיר ,להבין להתמחות ולהתמקצע בתחום מערכות big data . הקורס מיועד הן לאנשי IT ולמנהלים המבקשים להיכנס לתחום ה- big data וללמוד כיצד ליישם פרויקטים מעולמות התוכן הרלוונטיים.

דרישות קדם

- ניסיון בעולם מערכות המידע

תכני הקורס

Module Title	Module Description
From BI to Big Data. .	▪ What is Business Intelligence ▪ How Important is BI ▪ From data to knowledge ▪ BI Life cycle ▪ Requirements in BI project ▪ Front end / Dashboards / Data visualization ▪ Data modeling ▪ ETL ▪ What if ▪ Prediction ▪ In memory db ▪ Data mining
Introduction to Big Data	**Introduction to Big Data** ▪ Big Data Characteristics and Use-Cases ▪ Big Data from Business Perspective Challenges and limitations of Big Data Implementation **The Data in Big Data** ▪ Structured vs. Semi-structured vs. Unstructured Data ▪ Various Phases of Data Processing ▪ Problems with O/R Mapping

Module Title	Module Description
	Introduction to NoSQL ▪ NoSQL Definition ▪ Business Drivers for NoSQL ▪ New NoSQL Paradigms ▪ NoSQL Categories: o Key-value Store o Column Family o Document DBs o Graph DBs **Hadoop Overview** ▪ Apache Hadoop Architecture: HDFS and MapReduce ▪ MapReduce and Big Data – Real Life Examples **Introduction to Big Data Analytics** ▪ Data and Analytical Complexity ▪ Real-Time Analytics ▪ Big Data Visualization
Big Data: What it Means to IT Managers on the Front Lines	▪ Making sense of Big Data ▪ Turning Big Data into something useful ▪ Storage ▪ Security ▪ Data reconciliation ▪ Information extraction ▪ Insight distribution ▪ Successfully Navigating Big Data
Project management (project steps) in big data	**1. Data exploration** **2. Data cleansing** **3. Data preparation** **4. Data analyze** **5. Data display** **1+3 data exploration and data preparation** **Introduction** There are no shortcuts for data exploration. If you are in a state of mind, that machine learning can sail you away from every data storm, trust me, it won't. After some point of time, you'll realize that you are struggling at improving model's accuracy. In such situation, data exploration techniques will come to your rescue. **Steps of Data Exploration and Preparation** ▪ Variable Identification

Module Title	Module Description
	▪ Univariate Analysis ▪ Bi-variate Analysis ▪ Missing values treatment ▪ Outlier treatment ▪ Variable transformation ▪ Variable creation **Missing Value Treatment** ▪ Why missing value treatment is required ? ▪ Why data has missing values? ▪ Which are the methods to treat missing value ? **Techniques of Outlier Detection and Treatment** ▪ What is an outlier? ▪ What are the types of outliers ? ▪ What are the causes of outliers ? ▪ What is the impact of outliers on dataset ? ▪ How to detect outlier ? ▪ How to remove outlier ? **The Art of Feature Engineering** ▪ What is Feature Engineering ? ▪ What is the process of Feature Engineering ? ▪ What is Variable Transformation ? ▪ When should we use variable transformation ? ▪ What are the common methods of variable transformation ? ▪ What is feature variable creation
Data Cleansing	▪ Introduction ▪ Motivation ▪ Data quality o Validity: o DE cleansing o Accuracy: o Completeness: o Consistency: o Uniformity: ▪ The process of data cleansing o Data auditing: o Workflow specification: o Workflow execution: o Post-processing and controlling: ▪ De-cleanse o Parsing: o Data transformation

Module Title	Module Description
	○ Duplicate elimination: ○ Statistical methods: ▪ Challenges and problems
Data exploration with ETL, ELT, and ETLT	▪ The ETL Bottleneck in Big Data Analytics ▪ Apache Hadoop for Big Data ▪ ETL, ELT, and ETLT with Apache Hadoop: ▪ Choosing the Physical Infrastructure for ETL with Hadoop: ○ Compute ○ Memory ○ Storage ○ Network ○ Software
Data collection and integration from outside resources	▪ Introduction ▪ Hadoop distributors (Cloudera, Hortonworks, MapR) ▪ Building blocks of Hadoop (NameNode, DataNode…) ▪ Introduction to HDFS ▪ Map-reduce pattern ▪ Distribute cache ▪ Introduction to Hive for ad-hoc queries ○ Hive basics ○ Hive data types ○ HiveQL ▪ Pig: ○ Introduction to Pig as data flow language ○ Pig Latin basic expressions ○ Operators for data processing ▪ YARN (Map-Reduce 2) ○ Motivation for YARN. ○ Architecture. ○ Features.
Data modeling in Big data world	▪ Saving data methods: ○ Hadoop ,Files, XML, CSV, RDBMS,OLAP, Tabular, No-SQL db ▪ RDBMS challenge in Big-data world ▪ No-SQL vs traditional relational data ▪ Scaling vs. consistency ▪ No-Sql database types: ○ Key Value DB ○ Column-Family DB ○ Document DB ○ graph DB

Module Title	Module Description
	▪ Transaction in No-SQL ▪ Applying map-reduce ▪ No-SQL leading implementations
Date inquiry in Big data	▪ Hive o Introduction to Hive for ad-hoc queries o Hive basics ▪ Hbase o Introduction to Pig as data flow language o Introduction to Hbase for processing huge tables o Hbase data model o Hbase vs. RDBMS ▪ Data mining ,using R
Data visualization	▪ 1:2:3 Method to data visualization ▪ What are the changes that should be made in our methods ▪ Alerts & exceptions ▪ Links & networks ▪ Hue – the new gate to big data
	▪ Presentation/ project