Big Data Pipeline Engineering

ה"צנרת" הארגונית
ניהול, ניקוי ואבטחת זרימת המידע

<u>תיאור הקורס:</u>

בעולם שבו מידע זורם מכל כיוון - אפליקציות, אתרים, מערכות מכירות וחיישנים - האתגר הגדול ביותר הוא לאסוף את הכל למקום אחד בצורה מסודרת ובטוחה.
קורס זה נועד ללמד אתכם איך לבנות את ה"אינסטלציה" של עולם הנתונים: מערכת של צינורות דיגיטליים (Data Pipelines) שיודעים לקחת מידע גולמי ומבולגן ולהפוך אותו למידע איכותי שניתן לקבל ממנו החלטות עסקיות. נתמקד בתהליכי ה ETL-המודרניים (שליפה, עיבוד וטעינה), שהם הלב הפועם של כל ארגון טכנולוגי.

נלמד כיצד לוודא שהנתונים בדרך אינם "מזוהמים" וכיצד להפעיל כלים טכנולוגיים שחושפים את האמת מאחורי המספרים.

נבין איך בונים מערכות שמזהות חריגות ודפוסים חשודים בזמן אמת בתוך מיליוני עסקאות דיגיטליות. נלמד להשתמש בחתימות דיגיטליות ובמתודולוגיות אימות כדי להבטיח שהמידע שמגיע למחסן הנתונים הוא מקורי ולא עבר מניפולציה.

<u>קהל יעד:</u>

- מפתחי Frontend ו Full-Stack
- אנליסטים ואנשי דאטה המעוניינים לשדרג את יכולותיהם הטכניות בבניית תשתיות.
- בוגרי קורסי פיתוח ו-JavaScript שרוצים להשתלב בצוותי הנדסת נתונים מודרניים.
- מנהלי פרויקטים טכנולוגיים המבקשים להבין את מחזור חיי הנתון בארגון.

סילבוס נושאים

מודול 1: מבוא להנדסת נתונים (Data Engineering)

- מהי "צנרת נתונים" (Data Pipeline) ולמה הארגון לא יכול לתפקד בלעדיה.
- הכרת המבנה של מחסן נתונים (Data Warehouse) לעומת בסיס נתונים רגיל.
- תכנון ראשוני של מסלול הנתון: מהמקור ועד ללוח הבקרה של המנהל.
- סקירת כלי העבודה המרכזיים בשוק והבנת תפקיד מהנדס הנתונים בצוות.

מודול 2: שליפה (Extraction): איסוף נתונים ממקורות מרובים

- איך מושכים מידע ממקורות שונים: בסיסי נתונים, קבצים חיצוניים וממשקי API.
- התמודדות עם סוגי נתונים שונים (טקסט, מספרים, תאריכים) ואיחודם לשפה אחת.
- שיטות לאיסוף מידע ברשת בצורה אוטומטית ובטוחה.
- מניעת עומסים על המערכות המקוריות בזמן שליפת המידע.

מודול 3: שלב הטרנספורמציה (Transformation): ניקוי ועיבוד הנתונים

- טכניקות לניקוי דאטה: טיפול במידע חסר, כפילויות ושגיאות הקלדה.
- המרת נתונים גולמיים למידע עסקי (למשל: הפיכת תאריך לידה לקבוצת גיל).
- שימוש ב-AI לזיהוי חריגות (Anomaly Detection) כבר בשלב העיבוד הראשוני.
- בניית "מסכים" חכמים שמוודאים שהנתונים עומדים בסטנדרט האיכות של הארגון.

מודול 4: שלב הטעינה (Loading): אחסון בטוח במחסן הנתונים

- שיטות טעינה שונות: טעינה מלאה לעומת טעינה של שינויים בלבד (Incremental Load).
- ארגון המידע בתוך מחסן הנתונים בצורה שתאפשר שליפה מהירה בעתיד.
- ניהול גרסאות והיסטוריה של נתונים: איך יודעים מה היה המצב בכל נקודת זמן.
- הבטחת שקיפות הנתונים ויכולת מעקב אחרי המקור של כל נתון (Data Lineage).

מודול 5: אבטחת הצנרת ואימות נתונים

- הגנה על הנתונים בזמן המעבר: הצפנה ומניעת דליפות מידע.
- שימוש בחתימות דיגיטליות לאימות מקוריות של תוכן וזהות בתוך הצינור.
- בניית מערכות התראה בזמן אמת על תהליכים שנכשלו או נתונים חשודים.
- פרוטוקולי אבטחה למניעת "הזרקת" מידע מזויף למערכות הארגוניות.

מודול 6: אוטומציה וניטור של תהליכי הנתונים

- איך גורמים ל-Pipeline לעבוד לבד בכל יום ובכל שעה ללא מגע יד אדם.
- בניית לוחות בקרה (Dashboards) שעוקבים אחרי "בריאות" הצנרת ונתונים.
- טיפול בתקלות: מה עושים כשהמידע מפסיק לזרום או מגיע בצורה משובשת.
- סיכום: בניית "פרוטוקול אימות" חסין לתקלות וזיופים עבור תשתית הנתונים.