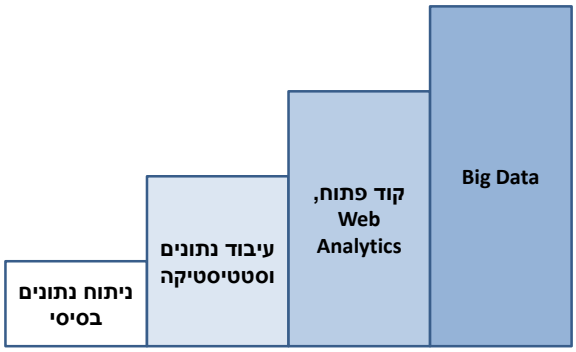
כיצד תהפוך מאנליסט למדען נתונים – ותקפיץ לעצמך את המשכורת

הקדמה

מדען נתונים הוא אחד מהמקצועות המגוונים ביותר הקיימים כיום. בעוד שמטרת המקצוע היא להפיק תובנות עסקיות מתוך ים הנתונים המקיף את כולנו, הדרך לעשות זאת כוללת מרחב עצום של מתודולוגיות, טכנולוגיות ויכולות – מה שמקשה מאוד למקד את תהליך הלמידה והרחבת הידע המקצועי, וגורם מצד אחד למדעני נתונים רבים לחוש מתוסכלים וחסרי סיפוק מקצועי, ומצד שני למנהלים לחוש פספוס מאחר ואינם מממשים את הפוטנציאל האדיר הטמון בחברה שלהם.

במקביל, השוק יודע להבחין בין רמות שונות של מדעני נתונים – בעיקר בהיבט מגזר הפעילות (Industry) בו הם פעילים, ובעיקר בהתאם לכלים בהם הם שולטים. בעוד שבמגזרים מסורתיים (פיננסיים, קמעונאות, Telco, תעשייה וכיו"ב) אנליסטים ירוויחו שכר מסוים, במגזרים עתירי הידע נדרשות יכולות נוספות – טכניות יותר – ובהתאם להן גם רמת השכר גדלה.

מדריך זה ממפה את הרמות השונות של מקצוע אנליסט/מדען הנתונים, כולל תחומי האחריות והידע המתווספים בכל רמה, וכן כלים המייצגים את הידע הזה.
המדריך בנוי בצורה של מדרגות – כל אחת מציגה 'קפיצת מדרגה' בהיבט רמת היכולות של האנליסט והכלים אותם הוא מכיר, ובהתאם לכך עולה גם רמת השכר. רמת השכר הספציפית בכל מדרגה לא מוצגת ברמה המספרית, אלא באופן איכותי בלבד ביחס ליתר המדרגות.



ארבע המדרגות במקצוע מדען הנתונים

חשוב לציין, כי הכלים המצוינים במסמך זה – אמורים לייצג באופן הולם את התחומים השונים בהם מדען נתונים נדרש לשלוט; ייתכן, ואף רצוי לבחור במקרים מסוימים בכלים תחליפיים/ מתחרים. רשימה של כלים מעין אלה מופיעה בעמוד הקישורים המומלצים בבלוג.
בנוסף, בעוד שהתפיסות המיוצגות ע"י הכלים השונים המפורטים במדריך מהווים בסיס הכרחי לכל פעילות בתחום מדע הנתונים במדרגה הספציפית, בהחלט ייתכן כי יידרשו לך כלים נוספים – בהתאם לצרכי הארגון הספציפי בו את/ה פועל/ת.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

<u>מדרגה ראשונה - ניתוח נתונים בסיסי (אנליסט/כלכלן)</u>

בארגונים רבים, לרוב – ארגונים גדולים אשר יחידת האנליזה בהם חדשה, או ארגונים בעלי מערכות מידע פשוטות יחסית, תפקיד אנליסט הנתונים יכוון באופן בלעדי להבנה בסיסית של עולם התוכן העסקי והסקת מסקנות ברמת המאקרו. בתרחיש כזה האנליסטים יתמקדו בעיקר בשימוש בכלי ניתוח נתונים גנריים – בעלי תפוצה רחבה בארגון, ופחות בעיבוד נתונים מסיבי (תפקיד זה לרוב יושת על מערך מערכות המידע בארגון).
הידע הטכני הנדרש עבור שימוש בכלים אלה הוא יחסית בסיסי, והמיקוד הוא באופן החיתוך של הנתונים והצגתם באופן שיאפשר הסקת מסקנות מהירה ואינטואיטיבית.

להלן הכלים העיקריים במדרגה:

1. **Excel**

גם עם כל הטכנולוגיה חסרת התקדים שאנו מוקפים בה, ה-Excel הוא עדיין הכלי המוביל והשימושי ביותר עבור כל מדען נתונים. החל מתכנון משאבים וחישובים פשוטים, דרך נוסחאות מורכבות וכלה בטבלאות ציר, Analysis Toolpack, Solver וחתירה למטרה – ה-Excel הוא עדיין זקן השבט והוא חי ובועט ביתר שאת. לאחרונה הכניסו ב-Microsoft מספר פיצ'רים מעניינים ל-Excel, ביניהם Power Pivot המאפשר שיפור ביכולות עיבוד הנתונים של טבלאות ציר והצגתן, ו-Power Query המאפשר ממשק נוח לקליטת נתונים חיצוניים ואינטגרציה שלהם לטובת ניתוח והצגת דוחות.
בעבודה היומיומית, ה-Excel נותן מענה לרוב דרישות הנתונים מצד אנשי הביזנס בארגון - באפשרו מעבר מהיר בין מימד החישובים למימד התצוגה (גרפים וטבלאות) – ולכן זהו הכלי הראשון שעל אנליסט הנתונים להכיר.

2. **Power Point**

Power Point אינו כלי נתונים מובהק, אלא כלי משלים להצגה ויזואלית של אלמנטים שנעשו עם כלי ניתוח נתונים אחרים. זהו אחד מהכלים החשובים ביותר בארסנל של מדען הנתונים, מאחר והוא מאפשר לתקשר את הממצאים והתובנות שזיהה עם דרג המנהלים – לטובת שיפור תהליכי קבלת ההחלטות. על כל מדען נתונים להכיר על בוריה את סביבת ה-Power Point (או כלי אחר ליצירת מצגות), ולדעת כיצד להכין מצגת אפקטיבית - בעלת אלמנטים ויזואליים דוגמת תרשימים, גרפים ותמונות, אשר יעבירו בצורה הברורה והחדה ביותר את המסרים והתובנות האנליטיים למנהלים.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

3. **Business Objects**

אחד הכלים הקריטיים לעבודתו של אנליסט הנתונים הוא ממשק לנתוני הארגון. בארגונים רבים מוטמעים כלי BI שונים, המאפשרים למספר רב של משתמשים לגשת לבסיס הנתונים בארגון ולאחזר ממנו מידע – ואנליסט הנתונים הוא אחד המשתמשים הכבדים בממשק זה – בעיקר לצורך הכנת דוחות אד-הוק שאינם סטנדרטיים, זיהוי תובנות חדשות, תכנון ובקרת ביצועים בארגון ועוד.

אחד הכלים הנפוצים ביותר בקטגוריה זו הוא Business Objects (כיום שייך ל-SAP). הכלי מאפשר שליפה וחיתוך של דוחות ברמת גמישות גבוהה יחסית על פני נתונים רבים במחסן הנתונים הארגוני – וניתן לייצא את תוצאות הדוחות לאקסל להמשך ניתוח נתונים והסקת מסקנות.

חשוב לציין, כי בעוד שהכלי הוא הכרח במדרגה זו של מקצוע האנליסט, הוא גם זה שמגביל את האנליסט מלמצות את יכולותיו המקצועיות והאנליטיות – מאחר והוא נועד למשתמשי קצה ולא לאנליסטים; לכן חשוב לכל אנליסט להכיר גם את המדרגה הבאה שתוכל להביא לידי ביטוי את יכולותיו בצורה טובה יותר.

בשוק קיימים מספר כלים נוספים שכדאי להכיר במדרגה זו, דוגמת Congnos, כלי קוביות דוגמת Panorama וכלים להצגת Dashboards עבור משתמשי הקצה דוגמת Qlikview.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

מדרגה שניה – אחזור נתונים, עיבוד נתונים וסטטיסטיקה (אנליסט נתונים)

המדרגה הבאה של אנליסט הנתונים מתמקדת באחזור ועיבוד נתונים ברמה מורכבת על בסיס הנתונים הארגוני. במדרגה זו המיקוד בעבודת האנליסט הופך מעסקי לטכני יותר, מה שמתבטא בשני היבטים:

- יכולות אחזור ועיבוד נתונים מורכב מתוך בסיס הנתונים הארגוני.
- הפעלת שיטות סטטיסטיות שונות להצפת תובנות עמוקות יותר החבויות בנתונים (Predictive Modelling ,Predictive Analytics).

רמה זו היא הנפוצה ביותר כיום במגזרי הפעילות המסורתיים.

4. **SQL**

בסיס הנתונים הוא מקור חומר הגלם עבור רובנו. אם Excel מאפשר ביצוע ניתוחים מורכבים על סט מצומצם יחסית של נתונים, אזי בסיס הנתונים מאפשר עבודה במסות, וכנובע מכך גם זיהוי מגמות ותופעות מעניינות ביותר. יחד עם Excel, SQL הוא כלי העבודה הנפוץ ביותר בקרב מדעני נתונים, והוא מהווה את אבן היסוד לרוב הפעילות של מדען הנתונים – עיבוד הנתונים והכנתו לאנליזה דרך שיטות סטטיסטיות. גם בעידן ה-Big-Data ה-SQL נשאר רלוונטי, מאחר ועל אף כל הטכנולוגיות החדשות – הוא עדיין היחידי המסוגל לשמור על עקביות ודיוק הנתונים ברמת הטרנזאקציה. בסיס טוב של ידע ב-SQL – על המחוללים השונים הקיימים לו (SQL Server, MySQL, SAS SQL, Queryman, SQLNavigator ועוד רבים אחרים) – הוא קריטי עבור כל מדען נתונים, וחשוב להכיר את הכלי לפני ולפנים – כולל שימוש בלולאות, משתנים אינדקסים ופונקציות מובנות. מומלץ מאוד להוריד את אחד מהכלים החינמיים ובאמצעותם לתרגל את השפה.

במשפחה זו רצוי גם להכיר את Microsoft Access – המשמש כמעין פלטפורמת ביניים בין אקסל לבין כלי תכנון בסיסי נתונים מבוססי SQL ומאפשר 'נחיתה רכה' יחסית בעולם בסיסי הנתונים הרלציוניים (Relational DB).

5. **SAS**

גם לאחר כמעט ארבעה עשורים מאז החלה הפצתו, הכלי מהווה את ה-Best Practice לפלטפורמה סטטיסטית לניתוח נתונים. סט הכלים של SAS נפוץ מאוד בארגונים גדולים, וסביר להניח שיישאר בהם לאור הניסיון הרב בשימוש בו. למי שעוסק בניתוחים סטטיסטיים, בכריית מידע וב-Predictive Analytics בארגונים גדולים – חובה להכיר את הכלי הבסיסי של ניתוח הנתונים הסטטיסטי (SAS Enterprise Guide), מאחר והוא הופך ללב תהליך ניתוח הנתונים. בנוסף, בארגונים רבים מהווה הכלי האמור אמצעי נגיש מאוד לעיבוד וניתוח נתונים בסביבת עבודה אחת.

מצד שני, חשוב לדעת באילו פרוצדורות בכלי להשתמש ומתי, ולהעדיף עבודה שיטתית לפי ה-Best Practice בענף על פני שימוש בפרוצדורות שקשה לתעד או להבין (לדוגמה: להעדיף את Proc SQL לשלב עיבוד הנתונים על פרוצדורות ב-Data Step).

בנוסף לכלי הסטטיסטי, SAS מחזיקה בסל פתרונות רחב מאוד שכדאי להכיר, גם בהיבט הסקטורים הספציפיים אליהם היא פונה, וגם בהיבט השלב בתהליך המידע – בהם היא תומכת, החל ממחסן נתונים, דרך כלי ETL, ניתוח נתונים סטטיסטי, כריית מידע וטקסט, ויזואליזציה ודשבורדים לניטור ביצועי המודלים.

כלים מתחרים/ תחליפיים: SPSS.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

6. Tableau

היבט חשוב של תחום ניתוח הנתונים הוא ויזואליזציה. עבור חלק גדול מהאנשים, בכללם מדעני נתונים, תצוגה גרפית מצליחה להעביר מידע באופן אפקטיבי יותר מנתונים יבשים. שימוש במנוע גרפי להצגת הנתונים הוא אחד מהכלים החזקים בארגז הכלים של מדען הנתונים, והוא מאפשר הצפה מהירה מאוד של תובנות ומגמות. בנוסף, מאפשר Tableau יכולת תקשורת טובה למדי עם דרג המנהלים דרך הפיכת תוצר ניתוח הנתונים לדשבורד.

Tableau מאפשר התממשקות למגוון רחב מאוד של מקורות נתונים, כולל למקורות נתונים השייכים לעולם ה-Big-Data, דוגמת Hadoop.

חשוב לציין, כי קיימים בשוק שני סוגי כלי ויזואליזציה: האחד מכוון לשיפור תהליך ניתוח הנתונים עצמו (דוגמת Tableau, Spotfire ו-Sisense), והשני הוא כלי BI להצגת דשבורדים סגורים עבור משתמשי הקצה (דוגמת Qlikview) – שצוינו במדרגה הקודמת.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

<u>מדרגה שלישית – קוד פתוח ו-Web\Mobile Analytics (אנליסט דיגיטל)</u>

המדרגה השלישית ביכולותיו של מדען הנתונים היא קוד פתוח (Open Source). קוד פתוח מהווה את אחת המהפכות הגדולות ביותר בעשור האחרון, ומציב בפני מדען הנתונים את אחד האתגרים המשמעותיים ביותר בקריירה שלו – מעבר מלימוד כלים ותיקים באמצעות קורס מובנה או הכשרה מצד החברה – ללימוד עצמי בקצב מהיר. השינוי הזה מהותי מאוד – מאחר ומדען הנתונים נדרש באחת להימדד לפי תוצאות – המחייבות אותו ללמוד כלים חדשים בקצב יחסית גבוה, תוך יציאה מאיזור הנוחות שלו. במילים אחרות – 'ללמוד איך ללמוד'.

למגמה זו ניתן לצרף גם את עולם ה-Web Analytics, אשר חולק תכונות רבות עם הקוד פתוח באמצעות תפיסת ה-SaaS (Software as a Service), כולל חבילות חינמיות מסוימות – והדרישה ללימוד כלים רבים במהירות גבוהה.

מדרגת הקוד הפתוח מהווה לרוב את נקודת הקפיצה בין המגזרים המסורתיים ובין מגזר ההיי-טק – קפיצה המתבטאת גם באופן מהותי בשכרו של מדען הנתונים.

7. **R**

אם SAS הוא המלך הסטטיסטי של הארגונים הגדולים – R הוא הכלי המועדף על סטטיסטיקאים ביתר הארגונים. עובדת היותו קוד פתוח הביאה לצמיחת קהילת משתמשים מפותחת מאוד ולפיתוח תוספים רבים. עד לפני כ-5-10 שנים לימוד השפה היה מעט מורכב, אך עם כניסת אתרי המידע למשוואה (Stackoverflow ודומיו), כמו גם פיתוח סביבות משתמש ידידותיות יותר (דוגמת RStudio), הפך לימוד הכלי לנפוץ יותר ויותר, וכיום הוא כלי חובה לכל מדען נתונים.

8. **Stackoverflow**

אחד מהחידושים המהותיים ביותר של העשור האחרון הוא אתרי שיתוף הידע והפורומים המקצועיים. התחום כל כך מפותח היום, כך שסביר להניח שכמעט עבור כל שאלה מקצועית העולה במוחנו – ככל הנראה היה מישהו אחר בעולם שכבר נתקל בבעיה דומה, ופתר אותה בצורה זו או אחרת. במילים אחרות - התשובה לכל שאלה נמצאת אי שם ברחבי הרשת, נגישה לכל דורש ורק מחכה שנמצא אותה. תוסיפו על זה אתרי לימוד שפות תכנות דוגמת Coursera, W3Schools, CodeAcademy, ואפילו Youtube, ותקבלו יכולת בלתי מוגבלת ללמוד ולתרגל כמעט כל נושא, כל שפה וכל טכניקה, בערוץ המועדף עליכם (טקסט, תמונות, מצגות, וידאו) – ובחינם.

בעוד שבחרתי באתר Stackoverflow כמייצג את המגמה הזו, מאחר ולטעמי הוא אחד מהאפקטיביים ביותר לקבלת תשובות בתחומים הרלוונטיים אלינו, נכללים בתחום זה אתרים רבים אחרים, דוגמת Quora, Github ועוד בלוגים, פורומים ואתרי קורסים רבים.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

9. Google Analytics

בעולם האינטרנט, בו התפעול של החברה מתבצע בעזרת אפליקציות/מערכות אונליין, אין אנשים הנמצאים בחזית מול המשתמשים, ו'חיים' את התהליך העסקי; במציאות כזו, עיקר הפידבק מהמשתמשים מגיע בצורת נתונים – והפיכתם לתובנות ולהחלטות עסקיות נופל על כתפיהם של אלה היודעים לקרוא אותם, לסנן מהם רעשים ולזקק מהם תובנות עסקיות והצעות לפעולה.

בעולם המדובר, Google Analytics הוא כיום כלי ברירת המחדל, ומאפשר ניטור של התנהגות המשתמשים באתר מצד אחד, וניתוח בסיסי שלה – מצד שני. בעוד שהכלי נוח מאוד להטמעה, הוא טומן בחובו בעיה מהותית, של 'הצפה' מסוימת של המשתמש בנתונים שקשה להבין. כאן בדיוק נכנס לתמונה מדען הנתונים – שכן בעוד שחלק ניכר מהשימוש בכלי BI מבוצע ע"י משתמשים, עדיין 'עין מקצועית' של מדען נתונים מיומן תהיה עדיפה כמעט תמיד במשימה זו על פני אדם בעל אוריינטציה עסקית. על מדען הנתונים להבין זאת, ולסגל לעצמו את היכולת להבין מתי להעדיף שימוש בכלי פשוט כמו Google Analytics על פני הפעלת אלגוריתמים מורכבים על בסיסי נתונים מסיביים – ולממש את יכולותיו האנליטיות בקריאת נתונים והבנתם.

10. כלי אופטימיזציה – Google Optimize

אחד המאפיינים הייחודיים של עסקים אינטרנטיים בכלל, וסטארטאפים בפרט, הוא חוסר הודאות המובנה לגבי המוצר והלקוחות. בתחילת הדרך, אין ממש ידע לגבי מיהו הלקוח האידאלי, ומהו הערך המדויק אשר הוא מקבל מהמוצר.

האתגר המובנה הזה מצריך פתרונות אחרים מאשר רק דיווח וניתוח נתונים כמו בחברות גדולות.

הפתרון הוא בתהליך מסודר של ניסוי וטעייה – A/B Testing - על מה עובד יותר טוב או פחות טוב ולאילו משתמשים. זה תהליך ארוך של למידה – אשר הכרחית לצורך זיהוי קהל היעד המדויק, ומיקוד הפיצ'רים של המוצר עבור קהל היעד הזה.

התהליך נקרא גם 'אופטימיזציה'.

בשוק קיימים מספר לא מבוטל של כלים. בתחילת 2017 נכנסה Google לעולם הזה עם כלי חדש בשם Optimize , אשר נותן קרב קשה מאוד לכלים אחרים וטובים כמו Optimizely, VWO ועוד.

הכלי פשוט מאוד ללמידה ותפעול, ומאפשר כניסה מאוד קלה לעולם הניסויים.

כמובן שהדבר החשוב באמת הוא לדעת כיצד לתכנן ניסויים כמו שצריך, על מנת למנוע הטיות של מדידה, תוצאות לא מובהקות ושאר מרעין בישין.

את זה עליך ללמוד דרך הכרת התיאוריה מאחורי הניסויים, ואפילו יותר חשוב – הרבה מאוד התנסות בשביל להכיר את כל הפינות.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

מדרגה רביעית – Big Data (מדען נתונים)

מדרגה זו היא הגבוהה ביותר בקרב מדעני הנתונים – ועיקרה – יכולת לעבד נתונים לא מובנים בנפחים גבוהים מאוד ובקצב מהיר, ולהפיק מהם תובנות אשר תאפשרנה שיפור בביצועי הארגון עתיר הידע.

המדרגה מחייבת את מדעני הנתונים לפתח יכולות טכנולוגיות ברמה גבוהה מאוד – ידע תכנותי מהותי והפעלת אלגוריתמים מורכבים – על מנת לאפשר להם לחקור טכנולוגיות נתונים חדשות ומידע חדש ממקורות נתונים חיצוניים, וזאת בנוסף לניתוח הנתונים המסיביים הנשמרים בארגון.

בהתאם לרמת היכולות הנדרשת במדרגה זו ממדעני הנתונים – רמת השכר כאן היא אחת הגבוהות בכל עולם מערכות המידע כיום, ולעיתים עשויה להיות כפולה מאשר בשתי המדרגות הנמוכות.

11. **Python**

שפת Python היא אחד מהמאפיינים הבולטים במהפך שחל בעולם הנתונים. אם לפני 10 שנים על מנת לפתח כלי/רעיון טכנולוגי מסוים נדרש היה להיות מתכנת מיומן (C#, C++, Java וכיו"ב), הרי שכיום פיתוח התוכנה הפך לנגיש מאוד באמצעות שפות High Level. שפות אלו הן בעלות Syntax פשוט למדי, וכנובע מכך החלו לאפשר גם לבעלי מקצוע שאינם מתכנתים להיכנס לעולם הפיתוח. מעט אחרי שגילו זאת האנליסטים/ סטטיסטיקאים – הם החלו לשלב ידע סטטיסטי עם ידע תכנותי וכך נולד המקצוע שלנו – מדען נתונים. כיום, השימוש ב-Python כולל את כל שלבי המחקר בעבודתו של מדען הנתונים, החל מהתממשקות וקליטת נתונים מגוונים דרך API's, דרך עיבוד נתונים, ניתוחם ואף הצגתם באופן גרפי.
Python היא שפה קלה מאוד להתקנה ותחילת עבודה, והקהילה הרחבה שלה בעולם מוציאה מספר אינסופי של תוספים, עבור כל יישום העולה על הדעת. באופן זה, מדען הנתונים יכול לאפיין את המחקר ואת תוצריו, ולממש אותם באופן מודולרי בעזרת התוספים – מה שמקצר בסדרי גודל את סבב המחקר-פיתוח-הטמעה, ומאפשר הצפת תובנות חדשות למנהלים בקצב שיא.

12. **Web Api's**

אחד מהמאפיינים הבולטים ביותר של השנים האחרונות, הוא שיתוף מסיבי של נתונים מכל רחבי האינטרנט. כיום, כל פלטפורמה אשר מכבדת את עצמה מציעה API (Application Programming Interface) לשימוש בנתוניה לצרכים שונים ומשונים כאשר מטרת הממשק היא להנגיש את השימוש בפלטפורמה עבור ספקי שירותים רבים – דוגמת Facebook, Ebay וכיו"ב. עבור מדען נתונים, ממשק חיצוני – משמעו עוד נתונים אשר יכולים להשתלב עם נתוני הארגון בו הוא פועל – ולהוות פוטנציאל עצום לזיהוי תובנות חדשות. כמה דוגמאות לכך: נתוני משתמשים מ-Facebook, איתור לקוחות פוטנציאליים ב-LinkedIn, מחקר טכנולוגיות מ-Github, מוצרים בעלי פוטנציאל גבוה ב-Ebay, ניתוח טקסט של ציוצים מ-Twitter ועוד רבים אחרים. מאחר וקיימים כל כך הרבה API's באינספור תחומים, חלק ניכר מתפקידו של מדען הנתונים כולל מחקר אילו מבין הנתונים החיצוניים יניבו תועלת רבה - ופיתוח אבטיפוס הבוחן את התועלת משילובם בארגון. דרישה זו מחייבת את מדען הנתונים לסגל לעצמו יכולות תכנות בהתאם, על מנת שיהיה מסוגל ללמוד API's של כלים חדשים. למרבה המזל, כלי Script דוגמת Python מאפשרים הורדת תוספי קוד פתוח (לדוגמה: חבילת Httplib) המאפשרים פעילות קלה למדי מול API's, מה שמקל על מדען הנתונים לקלוט נתונים חיצוניים באופן עצמאי.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

13. **Hadoop**

אם תשאלו 10 אנשים מהי המילה הראשונה שעולה בראשם כשהם שומעים את הביטוי "Big-Data" – רובם ככל הנראה יענה "Hadoop"; זהו הכלי המזוהה בצורה הברורה ביותר עם התחום, ולא בכדי. השם Hadoop מתייחס לארגז כלים שלם, אשר מטרתו העיקרית היא יכולת אחסון, עיבוד וניתוח נתונים בהיקפים אדירים. הכלי הבסיסי העונה לשם Hadoop הוא פלטפורמה לאגירת נתונים ועיבוד שלהם באופן מקבילי – כלומר במקום אחסון, שליפת ועיבוד הנתונים בשרת אחד מרכזי – אלה מפוזרים באופן מובנה על פני מספר רב של מכונות – מה שמאפשר עבודה על היקפי נתונים מסיביים בזמן קצר יחסית. רכיב האחסון המקבילי קרוי HDFS (Hadoop File System), ורכיב השליפות ועיבוד הנתונים המקבילי נקרא Map-Reduce.

בנוסף לרכיבים אלה, כולל ארגז הכלים כלים נוספים, אשר מהם מספר כלים קריטיים עבור מדען נתונים: שפת Script בשם Pig, ממשק דמוי SQL בשם Hive, וכלי ניתוח נתונים ו- Machine Learning בשם Mahout. עוד חשוב לציין את בסיסי הנתונים מסוג NoSQL, אשר מאפשרים אחסון נתונים מסיבי – ברוב המקרים על גבי מערכת הקבצים של Hadoop.

בעוד שעד לפני כשנתיים היה Hadoop נגיש רק בגרסת הקוד הפתוח שלו, מה שהקשה על הנגישות שלו, בשנים האחרונות הוקמו לא מעט פתרונות המציעים ממשק פשוט למדי לשימוש בפלטפורמה. מספר דוגמאות לכך: Cloudera, Azure HDInsights ו- Amazon EMR (Elastic Map-Reduce).

חשוב עבור כל מדען נתונים להכיר את מהות סט הכלים של Hadoop, גם אם בארגון בו הוא פעיל אין עדיין שימוש בו, מאחר ובמוקדם או במאוחר סביר להניח שהטכנולוגיה תהפוך לסטנדרטית במרבית הארגונים.

14. **AWS\Azure**

קשה להתייחס למונח ה-Big-Data מבלי להזכיר באותה הנשימה את תפיסת מחשוב הענן (Cloud Computing). לכאורה המגמה למעבר לעבודה בענן היתה שקופה יחסית עבור מדעני הנתונים, אך יחד עם זאת השלב הנוכחי במחשוב הענן היא לא פחות ממהפכנית. השוני המהותי הוא שכיום לא נדרש מחשב/ שרת פיזי במקום מסוים (On-Peremisis) על מנת להריץ שאילתות ו'לטחון נתונים'; כל שנדרש הוא להרים מכונה וירטואלית 'מפלצתית' היושבת בחוות שרתים אי שם בעולם ולהתחיל 'להתעלל' בה... בסיום העבודה – אפשר לכבות את המכונה בדיוק באותה הקלות שבה היא הוקמה – ולשלם רק עבור זמן העיבוד.

הסיפור הופך אפילו למהנה יותר כשעובדים עם Hadoop ומפעילים מספר מכונות – מה שעד לפני מספר שנים היה בגדר חלום. כיום, כל מדען נתונים מסוגל להרים בתוך זמן קצר ביותר מערכת בעלת כוח חישוב קרוב למדי לזה של חברת ענק כמו Facebook – ובעלויות השוות לכל נפש.

בעולם ה-Cloud קיימים שני מתחרים עיקריים – Microsoft ו-Amazon.

ה-Azure של Microsoft מאפשר ממשק נוח למדי לתחילת עבודה ב-Cloud לצרכי ניתוח נתונים, ולכן מומלץ לבחון אותו בשלבים הראשונים של הכניסה לתחום.

ה-AWS המתחרה (Amazon Web Services), מעט יותר 'מורכב לעיכול', אך מאפשר תכונות רבות ומגוונות אשר עשויות להתאים יותר לצרכי נתונים ייחודיים.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.

סיכום

במדריך פורטו ארבע מדרגות אשר הופכות אנליסט למדען נתונים, כולל סט היכולות והכלים שעליו להכיר על מנת להתפתח מקצועית.

אם אתה אנליסט – גם אתה יכול לדעת היכן אתה עומד ביחס לכלל התעשייה, להגדיר לעצמך מטרות - ולהתחיל ללמוד מתודולוגיות וכלים חדשים אשר יאפשרו לך להשיג קפיצה הן ברמת האתגר המקצועי, הן ברמת ההכנסה שלך והן בידיעה שאתה עוסק באחד המקצועות הנחשקים ביותר כיום.

את חלק מהכלים והטכניקות תוכל להתחיל ללמוד בעצמך כבר כעת.
חלק אחר עשוי לדרוש ממך קצת יותר מאמץ.

- **אז קודם כל, אשמח מאוד לקבל פידבק שלך לגבי המדריך – האם וכיצד הוא סייע לך.**
- **בנוסף, אשמח אם תוכל להמליץ עליו לאנשים נוספים אותם אתה מכיר, אשר יוכלו להפיק ממנו ערך כמוך.**

ולבסוף,

אם הגעת לכאן, סימן שהתחומים שפורטו במדריך עניינו אותך, ושאתה רוצה להבין בהם יותר.
מצד שני, ייתכן מאוד וקיימת אצלך אי בהירות לא מועטה.
קיימים כמה סימפטומים לאי הבהירות הזו:

- אם חלק **מהדברים נראים לך רחוקים** או מורכבים מדי.
- אם אתה **לא יודע כיצד לגשת לשלב הראשון** בלימוד של הנושאים.
- אם אתה **לא בטוח אילו מהכלים אכן מתאימים עבורך**.
- אם אתה **בספק אם אכן תוכל להרוויח יותר** במידה ותלמד את התחומים האלה.
- אם אתה **עדיין לא בטוח לגבי מסלול ההתפתחות שלך** כאיש נתונים.

אם הנושא בוער בך, ואתה רוצה לקבל יותר מושג על היכולות שלך, הכיוון המקצועי שלך ומה ברצונך לממש בקריירה כאיש נתונים – אשמח לסייע.

במקרה זה, השלב הבא עבורך הוא פגישת אבחון למציאת הנישה האנליטית הייחודית עבורך.

התהליך יחסית קצר, וכולל שיחה אישית להתרשמות והערכת יכולות אנליטיות.
בסיום הפגישה, תבין אילו מבין ההיבטים אשר פורטו במדריך הם הכי רלוונטיים עבורך, ותקבל מושג לגבי הצעדים הראשונים שלך בלמידה שלהם.

במקצוע שלנו הלמידה היא אחד הדברים הכי חשובים – מאחר והתחום משתנה ומתפתח כל העת.
לאחר שתדע בדיוק באיזו נישה ברצונך להיות – תוכל לפתח את המומחיות הייחודית שלך בה.
זה יאפשר לך מתן ערך גבוה יותר לחברות איתן תעבוד – וכנובע מכך גם צמיחה ברמת ההכנסה.

והכי חשוב – תוכל לחוות קפיצה משמעותית ברמת הסיפוק והאתגר המקצועי שלך.

צור קשר כעת לקבלת מידע נוסף ותיאום פגישת האבחון.

© 2015, כל הזכויות שמורות לאייל גורפינקל ו'אנשי נתונים' Datappl.com.