סיכום מפגש שולחן-עגול

חידושים טכנולוגיים בתחום הביג דאטה ולקחים use cases

מנחים

עינת שמעוני

פיני כהן

לקוחות נכבדים שלום,

תודה על השתתפותכם במפגש שולחן עגול Round Table בנושא חידושים טכנולוגיים, לקחים ו use cases בעולם הביג דאטה.

מצ"ב סיכום עיקרי הדברים שעלו במהלך המפגש. במפגש עלו נושאים מהותיים שתומצתו בסיכום כפי שעלו. אין בסיכום זה המלצה גורפת ללקוחות אלא מתן פרספרטיבה והצגה של ההתלבטויות שעלו במפגש כלומר "מהשטח".

מכיוון שכמות הנוכחים במפגש הייתה גבוהה מדי מכדי לקיים דיון פתוח בנושא, נבחרו מספר "סיפורי לקוח" אשר שיתפו מנסיונם עם הנוכחים וסביבם התפתחו מספר דיוני-משנה.

האווירה הכללית במפגש הייתה ש"עלינו מדרגה" ברמת בשלות התחום בישראל. יש כבר התחלות, יש לקחים, יש מעט "עשו ואל תעשו", ויש מסר שעלה לאורך כל המפגש והוא שלא ניתן לתכנן, להתנהל ולפעול בתחום הטכנולוגי הזה בפרקטיקות הקלאסיות והידועות מתחומים טכנולוגיים אחרים.

קריאה מהנה,
עינת ופיני.

תוכן

רקע כללי ומגמות (תמצות המצגת שהועברה על ידי STKI)

במצגת אשר הועברה על ידי עינת ופיני מ-STKI עלו המגמות הבאות:

- אנו עולים לשלב האבולוציוני ב"סוגי השאלות" שאנו שואלים באמצעות כלי BI ואנליטיקה. אם רוב הארגונים כבר מזמן סיגלו לעצמם את היכולת לענות על שאלת ה"מה קרה" ואף "למה זה קרה"? כיום מאוד טבעי לשלב גם פרדיקציה ולחזות מה "עוד מעט יקרה". אנחנו כבר מתקרבים למדרגה הבאה שהיא היכולת להיעזר בטכנולוגיה לדעת אילו שאלות ראוי לשאול (שכן הדאטה עליו אנו מסתמכים הופך לפחות "מאורגן" ושאלות המחקר פחות ברורות ומוגדרות מראש).
- עמדנו על הכשל העיקרי ביישום טכנולוגיות אנליטיקה מתקדמת וביג דאטה, שהוא כמובן לא טכנולוגי, אלא היכולת של הארגון לעכל טכנולוגיות אלה (וגם "לדמיין" מה בכלל אפשר לעשות איתו).
- סקרנו את ההתפתחות המדהימה (!) בעולם הבינה המלאכותית AI, ענף ה Machine Learning ובתוכו תחום הרשתות הנוירונים ו- Deep learning ומהי המשמעות עבור ארגונים בטווח הביניים? (לא התאפקנו והזכרנו גם את תחום הרובוטים והבוטים אשר מקבלים דחיפה משמעותית מהתפתחות טכנולוגית זו).
- דיברנו על "שכבה חדשה" שכדאי לשים אליה לב ב Stack הטכנולוגי – והיא שכבת נתונים וירטואלית: Data virtualization (לא כ"כ חדשה טכנולוגית, אבל הצורך בה הולך ועולה לאחרונה בשל אותן מגמות "ביזור הדאטה", מקורות הנתונים והצורך לזוז יותר מהר בפתרונות הדאטה הארגוניים).
- סקרנו את המגמות העיקריות גם בשכבת הדאטה הבסיסית; בסיסי הנתונים החדשים, תצורות עבודה חדשות כדוגמת Microservices ש"מבזרת" את ארכיטקטורת הדאטה שלנו; נושא ריבוי הטכנולוגיות Polyglot ועוד.

ניתן לקרוא בהרחבה על מגמות אלה בפוסטים האלה בבלוג שלנו:

חלק א'
חלק ב'
וכן לצפות במצגת המלאה כאן.

Use case: הקמת מעבדת מידע

במפגש תואר והוצג תהליך בארגון פיננסי להקמת מעבדת מעבדת דאטה – Data Lab (סביבת ביג דאטה זמינה ומוכנה לתחקור אשר מיועדת לבחינה מהירה של רעיונות ליישומים אנליטיים חדשים):

תצורת העבודה הקלאסית עד היום: דרישה עסקית –> אפיון –> פיתוח -> ייצור.

בעולם הביג דאטה, הגישה של "תראה לי קודם כל מה ה case use ואז אקים את הסביבה" לא עובד, הגיעו למסקנה שצריך קודם להתנסות.

לאחר הקמת המעבדה, התהליך נראה אחרת לגמרי: למידה טכנולוגית –> תשתיות תוכנה -> ניסוי טעיה -> ורק בסוף דרישה עסקית!

האתגרים העיקריים בבניית Lab Data ובכניסה לעולם הביג דאטה:

- חינוך התרבות הארגונית (הרבה sessions ומצגות פנים ארגוניות) וכן חינוך מקצועי (קורסים בCoursera – , לדוגמה של Learning Machine)
- הכנסת תהליכי עבודה תומכים. בין כלי הפיתוח הבולטים, spark R python scala וגם JAVA, כאשר SPARK הוא ה"בריון החדש בשכונה" וכולם מתחילים להתחבר אליו.
- התמודדות עם המון כלים והבנת המרחב הטכנולוגי המורכב
- הגדרת ארכיטקטורה ארגונית: איך יראה lake data? איזה כלי NoSQL? איך ישנעו נתונים? האם אנליסטים יעבדו על שכבת BIG DATA או שיקבלו שכבה "באמצע"?

ברמה הארגונית ישנם שינויים מהותיים:

מקצועות חדשים: במקום הפרדה קלאסית של תשתיות פיתוח יש משהו באמצע - Devops.

ב- BI- במקום אנליסטים צריך scientists data – שהם אנשי BI שיותר קרובים לעולם הפיתוח/

Chief data office – אמור למנוע את תופעת ה Data swamp = הרבה דאטה ברמת איכות ירודה. אותו/ה CDO אמור לעסוק באיכות, הגדרה וניהול הנתונים.

צורת העבודה במעבדת המידע:

עבור כל רעיון עסקי מתבצעת בחינה האם קיים ROI, ואז מחליטים האם ללכת על זה. לוקח מספר שבועות לבחון כל תהליך. איך מגיעים לרעיונות?

מתכננים לקיים datathon (בדומה ל Hackathon) – סדנה של DATA במהלכה מעלים רעיונות ומקבלים תמיכה, במסגרת תחרותית.

איך מתמודדים עם החשש של אבטחת מידע? יש פאראנויה שנובעת לעתים מחוסר ידע. לדוגמה, כלי קוד פתוח באופן טבעי נחשדים ללא בטוחים (אולם להפצות המסודרות יש כלי אבטחת מידע).

ישנן גם חששות מכיוון הרגולציה והייעוץ המשפטי – האם בכלל מותר לעשות שימוש בתובנות החדשות? מה לגבי מידע ציבורי? כל עוד המידע הוא Public ניתן לעשות בו שימוש, אולם האם ניתן להכניס מידע פייסבוקי ציבורי לתוך גבולות הארגו וולהשתמש בו לצד הדאטה הארגוני? זו נקודה פחות ברורה.

יישום MongoDB לטובת "תמונת לקוח"

במפגש תואר יישום "תמונת לקוח" בארגון פיננסי. מטרת הפרויקט: יצירת מאגר וחשיפה לצרכני מידע (הן משתמש הקצה והן אפליקציות), לא לשימוש אנליטי (לפחות לא בשלב ראשון).

לצורך זה, הוחלט על בניית שכבת data virtualization (פעם קראנו לזה ODS) שמאחדת מידע ממגוון פלטפורמות (חלקן מתעדכנות באונליין וחלקן ב BATCH) לפלטפורמת מידע אחת. "צרכניות" המידע המאוחד: אפליקציות, ערוצים המנגישים מידע בזמן אמיתי ללקוחות הקצה, לשותפים וגם לנציגים

למה הDW לא פותר את זה? ה-DW מתעדכן ברמה יומית ולעתים זה לא מספיק. צרכני המידע (עובדים, לקוחות, שותפים) צריכים לקבל מידע בריל טיים (לדוגמה, כשלקוח חדש רוכש מוצר רוצים כבר לראות זאת באתר). כמו כן, בניית פתרון בDW ותחזוקה שלו לוקח הרבה יותר זמן.

הפתרון הנבחר: NoSQL – במקרה זה, MongoDB.

אחת התועלות החשובות – התקדמות מאוד מהירה ותכנון זריז, כל מודל הנתונים שנשמר בתמונת הלקוח הוא לא ממודל וגמיש.

את מבנה הנתונים בנו בתחילה על בסיס הנחות יסוד שלהם (בדומה למערכות התפעוליות) ואותו מפתחים כל שבועיים. היתרון הגדול במונגו הוא שניתן לבצע טרנספורמציה של המידע בתוך מונגו, ולהרחיב את מבנה הנתונים בלי ששאר המערכות נפגעות ("צרכני המידע"). אם פעם כל פעולה כזו הייתה לוקחת שבועות, היום עושים את זה בשעה (!).

בכל שבוע מוסיפים מקורות מידע נוספים, מעשירים את בסיס המידע. התחילו במספר מקורות מידע על פי צרכים עסקיים שהיו קיימים באותו זמן.

מודל הנתונים במונגו מבוסס על collections של Jsonים, שזה מבנה XMLי גמיש לחלוטין. השאילתות לוקחות את המכנה המשותף של הנתונים. אם חסרים נתונים השאילתה עובדת, היא לא נכשלת! אם יש מידע מיותר, היא עדיין עובדת.

האפליקציות (צרכני המידע) לא ניגשות למונגו ישירות, בארגון בנו שכבת ביניים שלוקחת את השאילתות מהאפליקציות ומתרגמות ל query language וגם נותנות סטטיסטיקות (הערת STKI – כיום יש כלים שמספקים יכולות דומות). האפליקציות פונות עם פרמטרים ואותה "שכבת ריכוך" שהם בנו ומתרגמת אותם.

מהי מערכת היחסים בין המערכות התפעוליות ל DW והמונגו? ה DW והמונגו – שניהם צורכים מידע מהמערכות התפעוליות במקביל, בהתחלה התחילו בכוונה ללא תלות בכלל, בנו את כל הקשרים מחדש (היה יותר מהר מלבנות אותם לDW).
משתמשים בטכנולוגיות CDC (Change data capture) – שמזהים שינויים בדטהבייס של המקור ולהשלים אותם (משמתמשים במוצר). ה CDC עובר דרך תשתית של תורים, כל שינוי נכנס לתורים. מתוך התורים יש להם רכיבים שטוענים את המידע כפי שהוא לתוך המונגו.

פרויקט ההקמה לקח בין 2-3 חודשים בלבד! בנו את זה לקראת פרויקט CRM (בניית תמונת לקוח) כאשר ההערכה היא שכל חלופה אחרת הייתה לוקחת 12 חודשים צפונה.
את כל ההטענה למונגו, הסנכרון ההעשרה – את הכל פיתחו אולם כיום ישנם פתרונות, כמו לדוגמה K2view ו- DBS-H שתיהן חברות ישראליות.
יש 2 סוגי צרכנים למידע ב NOSQL – משתמש קצה שרוצה לראות תמונה בעיניים, והצרכן השני – מערכות מידע שאמורות להציג אותו.
לצורך הזה פותח FRONT END בדוט נט על בסיס FRAMEWORK גמיש, מאפשר להציג כל נתון שיושב ב NOSQL – תצוגת לקוח ויזואלית.

מחזיקים כיום בתוך מונגו 3-4 טרה של מידע (הודות ליכולות דחיסה בגרסאות המתקדמות זה יכול לרדת לפחות מ 1TB). באופן עקרוני אין צורך לתשתיות חזקות (בשל ארכיטקטורת הפתרון), הארכיטקטורה מאפשרת לרוץ על תשתית מאוד זולה ועדיין תתן ביצועי אחסון טובים מאוד. במקרה שלהם זה כן רץ על פלטפורמות מאוד חזקות. גם בשליפות נותן ביצועים טובים (נציג מארגון נוסף סיפר על בעיות ביצועים אצלם בשליפות מעל סביבת האדופ).

עוד נקודה הממחישה את אופי העבודה במוצרי קוד פתוח: מונגו זז בגרסאות בקצב רצחני, בארגון באופן קלאסי רגילים שגל שדרוג גרסה הוא פרויקט עם בדיקות. למעשה ראו שאין צורך, המוצר תומך בצורת עבודה של שדרוג מתגלגל, אפשר להעביר צרכנים בצורה מתגלגלת, כמו כן יש לו backward compatability. כלי קוד פתוח מובילים לעבוד בצורה שונה ממה שעושים בתשתיות סטנדרטיות.
דוגמה לתועלת מוחשית: לאחרונה נכנסה תקנה חדשה שמחייבת לספק גישה לסוג מסוים של לקוחות. אם פעם עשו את זה כשאילתות מול המערכות תפעוליות (יכול להגיע למיליוני רשומות), על ה BI זה לקח דקות, במונגו זה שניות וגם מייתר את הגישה למערכות התפעוליות.

יוזמת ביג דאטה בתהליך בארגון בריאות:

המטרה הנה לייצר פלטפורמת מחקר שתשמש גופים וחוקרים שונים מכל הסוגים לגשת למידע לצרכי מחקר. אחד מהאתגרים המורכבים כאן הנם מצד אחד מתן גישה לנתונים על מנת שיהיו מספקים למחקר, אך באותו הזמן לשמור באופן אדוק על צנעת הפרט תוך התייחסות ומענה להיבטים המשפטיים: איך עושים de-identification לנתונים, אבל שעדיין יהיו מספיק טובים למחקר?

מעבר לכלי מיסוך והרעשה קלאסיים, הכוונה לעשות שימוש בכלים שעושים ניתוחי what if (ניתוח סיכונים לחדירה למאגר) ולהחזיר את הנוסחה שאומרת כמה המאגר הזה בטוח.

מבחינת הפתרונות הטכנולוגיים, הולכים על פתרונות קוד פתוח וכלים proprietary שיש להם ערך מוסף או יחודי למחקרים בתחום הבריאות.

הרעיון הוא שכל חוקר יקבל עותק ספציפי שמתאים למחקר שרוצה לעשות, הכולל תת קבוצה מנתוני המאגר, שכוללת הצלבה ממקורות שונים. פלטפורמת המחקר תכלול שרותים וכלים בענן וכן גישת bring your own license (לכלים מועדפים ע"י החוקרים שאינם זמינים בפלטפורמה המוצעת).

בתשובה לשאלה שנשאלה על נושא ניתוח טקסט בעברית, מדווחים שכלי של Text Analysis נותנים מענה טוב ברמה הבסיסית גם בשפה העברית, ללא כיול מיוחד. זאת, בניגוד לדעה הרווחת.

כמה הסתייגויות מתחום הביג דאטה

אחד ממשתתפי המפגש שיתף את ההסתייגויות שלו עם תחום הביג דאטה:

הרגשה שכולם רק מדברים ולא עושים, והשאלה היא – מה הערך המוסף? ואיך עושים מזה כסף?

אחד מהשימושים שמדברים עליהם הוא לעשות העשרה של מידע, מתוך הנחת מוצא שכיום המידע נמצא באיים.

מצד שני, בארגון קיימת תשתית אינטגרציה - ESB שעושה ENRICHMENT ומסוגל להביא מידע מ 5 מקורות ויודע לערבב אותם. כלומר, עוד לא נתקלו בדרישה שמישהו זרק על השולחן שאי אפשר ליישם ב ESB (יכול להביא מידע מכל מקור, גם מערכת תפעולית וגם מהDW).

נקודה נוספת היא שכיום הם חיים בתוך "כלוב זהב" של מערכות יקרות מדי – באותו הארגון קיים אורקל EXADATA יש שם over power (כמות השאילתות כיום די נמוכה מול העוצמה של המכונות) אז גם נושא הביצועים אינו ממש "כואב".

שימוש באלסטיק לצורך ניתוח לוגים

ארגון בתחום הבטחוני סיפר על תהליך שהחל כבר לפני שנה וחצי.
בארגון התחילו עם Splunk – המוצר נועד לניתוח לוגים, מנתח מידע מובנה ולא מובנה.
לSplunk יש קשת של אפליקציות מובנות למספר רב של פתרונות ומוצרים בתחום ה .IT.
קשת הפתרונות למוצר נעה מפתרונות עבור הIT (בעיקר) - חיפוש תקלות, ועד לפעולות מול הרכש הארגוני. התחום בארגון זה יושב בצוות שליטה ובקרה ב IT.

מחולקים בביג דאטה ל-3:
בסיסי נתונים ותשתיות ביג דאטה
BI הצד העסקי
מתן פתרונות בנושאים טכנולוגיים ותשתיתיים
מוצר נוסף בתחום זה הנו Elastic – אשר מספק סל פתרונות די דומה. לוג אנליטיקס, חיפוש...
אחד היתרונות הבולטים של אלסטיק זו הקהילה הרחבה, אשר מפתחת ומעבה את היכולות.
בתשובה לשאלת ה"מוצר אל מול קוד פתוח", לכל מוצר במודל קוד פתוח זמן ועקומת למידה הרבה יותר איטית, אין תורה מובנית, פחות תיעודים. בארגון זה רוצים לעשות שימוש בשני הכלים, וכיום מנסים לסגל לעצמם את היכולת כשמגיע USE CASE לנווט אותו לכאן או לכאן.
הבדל אחד שצויין בין הכלים הוא החיבור למילונים – לפי מה שנאמר, SPLUNK אינו יודע לעבוד עם מילונים, לעומת אלסטיק שיודע להתחבר למלינגו, Hebmorph, בייסיס.
אלסטיק מהניסיון שלו יותר מתאים למצב בו יודעים מראש איך הדאטה שלך מובנה (עובד על JSON ויש סכימה של JSON, כרגע שאינדקסת פעם אחת...) אך זה גם קצת "חוטא למטרה".
יש ציפייה שנוכל לזרוק לשם את כל המידע, כל אחד יעשה חיתוך בצורה אחרת, ולא נצטרך לדעת מראש איך הדאטה בנוי. על כך אחד ממשתתפי הדיון גם כן עם ניסיון באלסטיק צציין כי מניסיונם, הם לא מחוייבים לסכימת נתונים קבועה, אפשר לשנות, להוסיף שדות, וזה לא מוחל על מה שכבר אנדקסת קודם. הבדל נוסף שצויין הוא שנושא ה-JOINS לוקה בחסר באלסטיק.
נושא העלויות הנו משמעותי ולכן השאיפה שלהם היא להטות את הכף לכיוון OS.
בהקשר לשאלת ה ROI והערך מתחום ביג דאטה, צוין כי הערך מאוד גדול! ברגע שמתחילים ונותנים את המשהו הראשוני הזה, מהר מאוד מתווספות עוד דרישות ומגיע התיאבון.
ROI מהיר, לא צריך פיתוח ארוך, זה Agile הלכה למעשה, נותנים תוצרים תוך כדי הפיתוח.
בהקשר לחשש אבטחת מידע בקוד פתוח – אכן קיים חשש, אבל יש קהילה גדולה שעוזרת. תיקונים ובקרות על מנגנוני פריצה.

לסיכום - גישת ה "Just do it":

יש תחושה כי הארגונים שהתקדמו יותר בתחום הביג דאטה האם אלה שהצליחו למצוא את קיצור הדרך, או ה"תירוץ" לקפוץ למים (גם אם סיבת הקפיצה ההתחלתית לאו דווקא הייתה מוצדקת ב-ROI מוכח), כאשר "עם האוכל בא התיאבון" ואותם ארגונים ממליצים כיום להיכנס לטכנולוגיות אלה, להקים סביבה מוכנה, וכך לגרום לאנשים להבין למעשה מה הם בכלל יכולים להרשות לעצמם לבקש.

זה אינו תנאי מספיק כמובן. נדרשת בניית "מפעל" תהליכי ארגוני, עם תהליכי עבודה מוגדרים מראש (איזה רעיון נכנס לבדיקה, איך בוחנים אותו, מה יש לרגולציה להגיד על זה...) הסברה ארגונית כל הזמן, גיבוש שיטות הדרכה בארגון כדי לגדל את ה"דור האנליטי" הבא ועוד ועוד.

אנו ב-STKI בהחלט חושבים שזו התקדמות מבורכת, גם אם אנחנו בפער אל מול גופים מקבילים בעולם, יש שלבים ראשונים בעקומת הגדילה שצריך לעבור ומוטב לעבור אותם כעת ולא כשכבר נהיה "מוכרחים".

תגובות ספקים ויועצים:

תגובת חברת HPE:

תחום ה Machine Learning -מזוהה כשלב הבא בתחום הביג דאטה .לא רק לאסוף נתונים ולהפיק דוחות ,כי אם לפתח יישומים שמבצעים פעולות בהתאם לנתונים . חברת HPE מציעה למפתחים פלטפורמה בענן לפיתוח יישומים עשירים במידע ,אשר מסייעת לפתח את יישומי ה Machine Learning -של הדור הבא. הפלטפורמה כוללת יותר מ -70 ממשקי , API שאותם יכולים מפתחים לשלב ביישומים .בין ה APIs -הללו נכללים כלים לחיפוש מידע ,חיזוי אנליטי ,ניתוח קונטקסטואלי,זיהוי פנים ,ניתוח תמונות ווידיאו ,הפיכת קול לטקסט ,אינדוקס אוטומטי של מידע לא מובנה ועוד.

במהלך התקופה מאז הושקה הפלטפורמה עשו בה שימוש יותר מ 15,000 מפתחים וכיום מבוצעות כמיליון קריאות API ביום.עד לאחרונה סיפקה HPE את הפלטפורמה ללא עלות לבתי תוכנה וסטארט-אפים, ובחודש אפריך הכריזה החברה על הפתרון המסחרי שלה עבור ארגונים גדולים וקטנים כאחד.

קיימים מספר חברות שעושים שימוש בפלטפורמה, לדוג' חברת Blink, היא חברת סטארט אפ-אמריקאית ,אשר מציעה יישום ספיד דייטינג וירטואלי .המנוי לשירות צופה בוידיאו של שתי דקות .החברה משתמשת ב- APIשל זיהוי הפנים בוידיאו על מנת לזהות אם הפנים המצולמות בוידיאו הן אנושיות .אם לא מדובר בפנים אנושיות הוידיאו ייסגר אוטומטית. המשמעות היא שבאפליקציה צרכנית פשוטה זו,משתמשים באנליטיקה מתקדמת מאוד כדי לבנות אמון עבור המשתמש .לפני שנה או שנתיים לא היה אפשרי לקבל אנליטיקה כזו במכשיר נייד, האנליטיקה שלנו יכולה לזהות לא רק אדם ,אלא גם גיל ,או מין.

דוגמא לשימוש בפלטפורמה הינו ההאקתון "היסטוריה פוגשת חדשנות" – ביוזמת חברת HPEו"יד ושם", המרכז העולמי לזכר השואה. הזוכים, אשר נבחרו על ידי פאנל מכובד של שופטים, ובנוכחות שר החינוך, נפתלי בנט, השתמשו בפתרונות הביג דאטה של HP על מנת לפתח דרכים חדשות לחשיפת, למעלה מפטה בייט של מידע בלתי-מובנה, המאוחסן בארכיוני "יד ושם"

ניתן לקרוא בקישור הבא:
http://www.nrg.co.il/online/13/ART2/734/691.html

וכן לצפות בסרטון הבא:

https://www.youtube.com/watch?v=sgoVF5qKuPQ

איש קשר: עמית מנור, HPE, amit.manor@hpe.com

תגובת חברת SAS - מיה:

פתרונות SAS מסייעים לארגונים העוברים לעולמות ה big data במגוון דרכים. החל מניהול נתונים פשוט יותר החוסך זמן יקר של הכנת נתונים, דרך ויזואליזצית מידע ותובנות כדי להבין מהר מה רלוונטי ואפקטיבי, ועד אנליטיקה in-memory וכן machine learning המאפשרים לאנליסט או data scientist לשאול את השאלות הנכונות ולקבל תשובות מדוייקות יותר.
SAS פיתחה פתרון ייעודי בשם SAS In-memory Statistics אשר מיועד לניתוח נתונים בעולמות ה big data. הפתרון הינו ממשק משתמש מאוחד, הנותן ל data scientist את כל הכלים הנחוצים לקבלת תובנות ממקורות מידע כגון Hadoop, בזמן הקצר ביותר.
הפתרון מאפשר:

- ניהול כל התהליכים מסביבת משתמש אחת – מניפולציות על המידע, הכנת נתונים, תחקור, בניית מודלים והפצתם. כל מחזור חיי המודל מנוהל ב-SAS
- בניית מודלים טובים יותר עם תוצאות טובות יותר – שימוש במודלים הסטטיסטים החדשניים ביותר וכן טכניקות של machine learning הרצים ישירות על נתוני Hadoop
- ניתוח בקצב המחשבה שלנו – סביבה אינטראקטיבית, in-memory מאפשרת לכתוב קוד לקבלת תובנות מיידיות באופן היצירתי ביותר
- סקלביליות מלאה – יותר נתונים ומורכבות גבוהה יותר אינן חסם בשום שלב בסביבה של sas

נושא חשוב בעולם ה Big Data הוא IoT Analytics. הרעיון הוא לנתח אנליטית נתוני IoT תוך כדי תנועה מבלי לשמור אותם ולנתח אותם בדיעבד.
ניתן לראות כיצד מידע מוזרם בתנועה לתוך מסננים וטרנספורציות אשר מזהים אנומליות בנתונים הגולמיים ומעבירים אותם למודל אנליטי שינתח אותם. על סמך תוצאת המודל ניתן להחליט האם להתעלם, ליצור התראה למשתמש, לעדכן דוח (דשבורד לדוגמה) או להפעיל תהליך קבלת החלטה מורכב יותר.
התהליך כולל היזון חוזר שמאפשר לבצע כיוונון של המודל ו deployment שלו בחזרה לסביבת ה Streaming Analytics.
ניתן לראות שאפשר לשמור חלק מהנתונים שמוזרמים למערכת במחסן נתונים כגון Hadoop לצורך טיוב המודלים האנליטיים בעתיד.