מנחה: ראובן יגל  
סטודנט: אביתר גולן  
לקוח: דר' שי אקו, יחידת הניסויים, חיל האוויר

אגירה וסיווג של מסלולים של גופים פיזיקליים באמצעות למידת מכונה – הצעת הפרויקט

רקע ההצעה

מעבדת התעשייה הבטחונית אוספת במשך שנים רבות מסלולי גופים, מעוניינים באמצעים של למידת מכונה על מנת לעקוב בזמן אמת אחרי מסלולם ולחזות שינויים שיכולים לשנות את נקודת הגעתו (לדוג').

ההצעה

הקמת מערכת הכוללת בסיס נתונים ומערכת לאגירת נתונים בזמן-אמת מגופים פיסיקליים כגון טילים ורקטות, הכולל תיעוד מסלולם ותכונותיהם. המאגר יאחסן ויקטלג את הנתונים בעזרת הכלים של למידת מכונה ויאפשר חיזוי בזמן אמת

תכנון ראשוני של הפרוייקט

הפרוייקט יבנה ממספר' שלבים, כאשר במבט עליון השלבים הם:

1. למידת הנושאים:
   1. בניית מערכת אחסון מרובת מחשבים
   2. בניית מערכת קליטה ותיעוד מקבילית
2. איתור בסיסי נתונים מתאימים למסד הנתונים העתידי ובחירת המסד המתאים
3. ניתוח דרישות ראשוני
4. הקמת מסד הנתונים הראשוני
5. כתיבת דוח אלפא
6. הקמת מערכת שאילתות שתבוסס על המסד, ותאפשר שימוש במודלים ואלגוריתמים רבים
7. מדידת ביצועי מערכת
8. ניתוח דרישות מפורט
9. תיכון מערכת
10. כתיבת דוח מסכם

סקר שוק

בדיקת מערכות לאגירת מסלולים במערכות אחרות, לדוג' מערכות רכבים אוטונומיות, מטוסים וכו'.

בדיקת סוגי מוסדות הנתונים השונים

ראובן ואני הסכמנו כי שנינו מעדיף לעבוד עם מסד נתונים לא-יחסי, כלומר מסוג NoSQL.  
הבסיסים אותם אני בודק למען הפרוייקט הם MongoDB, Cassandra ו-HBase.

**Cassandra:**בסיס נתונים מבוזר הכתוב בשפת ג'אווה, מיועד לטיפול בכמות נתונים גדולה העולה על מאות טרה-בייטים.  
הבסיס אינו משתמש במערכת Master-Slave אלא כל מחשב בצומת יוכל הפנות לכל צומת אחר. בנוסף, כל צומת מכיל שכפולים של מס' צמתים אחרים. המסד מבוסס שיטת "אחסון עמודה רחבה", כלומר משתמש בטבלאות עם שורות ועמודות, אך בניגוד מסדי SQL מאפשר ששמות ופורמטי העמודות ישתנו משורה לשורה.   
*חסרונות* הבסיס כוללים חוסר עקביות בעת שחזור מידע לצומת שנפלה, זמן קריאה ארוך מהשרת במידה והמפתח אינו ידוע למשתמש (לדוג' חיפוש של דוגמא ספציפית), וחוסר של תיעוד רשמי מאת הארגון המפתח, קרן Apache. שימושים ראשיים למסד כוללים תוכנות מסרים מיידים, מסחר אלקטרוני ועיבוד מידע מחיישנים בזמן אמת.

**HBase**בסיס נתונים הכתוב גם הוא בשפת ג'אווה, מבוסס על מערכת Bigtable של חברת גוגל, ומשתמש במספר תכונות שלו לרבות ביצוע פעולות על-גבי הזכרון בזמן אמת (לעומת גישה למידע הנמצא על גבי מחשבים רבים, פעולה הלוקחת הרבה יותר זמן), כיווץ מידע במוסד ושימוש במבני נתונים מבוססי Bloom Filter, אשר יכולים לקצר את זמן ההמתנה למשאבים מהדיסקים באופן משמעותי. בדומה ל-Cassandra, גם HBase מבוסס שיטת "אחסון עמודה רחבה". אחת מחוזקות המסד היא השימוש המובנה שלו ב-HDFS, מערכת הקבצים המבוזרת של Hadoop. המסד משתמש במערכת הקבצים HDFS של Hadoop, כך שאין צורך להשתמש במערכת MapReduce. חסרונות עיקריים של המסד כוללים שימוש במודל Master/Slave, דבר התורם לסיכוי לנקודת כשלון יחידה, אין ל-HBase שפת שאילתות מובנית (צריך להשתמש בממשק חיצוני, לדוג' Apache Hive), ותלות-יתר גבוהה בטכנולוגיות המובנות של Hadoop. שימושים שכיחים ל-HBase כוללים אנליטיקת לוגים מקוונת ואתרים מרובי תוכן (כגון נטפליקס, פינטרסט, פייסבוק).

**MongoDB**בסיס נתונים הכתוב בשפות JS, C++ ופייתון, מבוסס שיטת אחסון מסמכים, המציע גרסת קהילה וגרסת עסקים הכוללת גישה מהירה וקלה לנתונים על גבי הרשת, ביקורת נתונים מובנית, אבטחת מידע והצפנה.  
MongoDB מאחסן את הנתונים על גבי המערכת כמסמכי JSON בינאריים, שיטה המאפשרת גישה מהירה למידע. גרסת העסקים גם כוללת זמינות גבוהה ע"י שכפול המידע ליצירת עודפות וריבוי צמתים לאחסון. חסרונות של המסד כוללות חוסר תמיכה רשמי לגרסת הקהילה (לדוג' תיקוני טלאי צריכים להתבצע באופן ידני), וגרסאות סקלביליות גוזלות יותר משאבי זכרון.

ספרות:  
[**https://logz.io/blog/nosql-database-comparison/**](https://logz.io/blog/nosql-database-comparison/)[**https://www.mongodb.com/compare/mongodb-hbase**](https://www.mongodb.com/compare/mongodb-hbase)[**https://db-engines.com/en/system/Cassandra%3BHBase%3BMongoDB**](https://db-engines.com/en/system/Cassandra%3BHBase%3BMongoDB)[**https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0045-4**](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0045-4)