**פרויקט מסכם בקורס Big Data בר אילן**

**שמות מבצעי הפרויקט:**

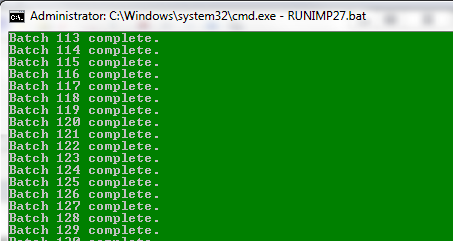
**שם סטודנט א מור מרקוביץ ת.ז. 203957980**

**שם סטודנט ב נתניאל פרייבסקי ת.ז. 211375167**

**שם סטודנט ג אפרת קנדלקר ת.ז. 209073204**

**הסטודנט המופיע כאן כסטודנט א' צריך להעלות את קובץ פתרון הפרויקט למודל (רק הוא)**

בפרויקט זה נחקור נתוני טיסות משנים שונות, באמצעות Neo4J ו MongoDB.

1. ניתן לבצע את הפרויקט בקבוצות של עד 3 סטודנטים לקבוצה. את התשובות יש לכתוב במסמך זה בצבע אדום. נסו גם להסביר/לנמק מדוע בחרתם לכתוב שאילתה בצורה שבחרתם – על מנת שגם אם ביצעתם טעות, עדיין יהיה אפשר לתת ניקוד חלקי.
2. נתחיל ב **Neo4J**:
3. התקנת תוכנות וספריות נדרשות:
   1. בספריה שקיבלתם יש ספריה פנימית בשם python27 המכילה קובץ התקנה – יש לבצע את ההתקנה. בסיומה, אמורים לקבל התקנה חדשה של python 2.7 בספריה c:\python27 *(לא לשנות מיקום כי אני מסתמך עליו בקבצים אחרים בהמשך)*
   2. בספריה שקיבלתם, יש לפתוח command line (cmd) ולהריץ את הקובץ runinst27
      1. *קובץ זה יתקין עבורכם מספר ספריות נדרשות עבור הפרויקט*
4. פתיחת בסיס נתונים ב Neo4J:
   1. להריץ Neo4J desktop
   2. לפתוח בו בסיס נתונים חדש (לוחצים על Add Graph)
   3. לבחור Create a Local Graph
   4. לתת לו שם – Graph, וסיסמא 12345 (חובה את אותה סיסמא כי הסקריפט שמעלה את הנתונים מסתמך על זה)
   5. ללחוץ על כפתורStart ואז על Manage ואז להריץ את בסיסי הנתונים ולפתוח ממשק באמצעות הכפתור "Open Browser"
5. העלאת הנתונים ל Neo4J:
   1. בספריה שקיבלתם, יש לפתוח command line (cmd) ולהריץ את הקובץ runimp27
      1. *קובץ זה יעלה עבורכם את נתוני הפרויקט, זה ייקח קצת זמן ותוכלו לראות שהוא סופר בזמן ההעלאה כמו בתצלום המסך הבא:*
      2. *ייתכן שכשתנסו להריץ את הקובץ תקבלו שגיאה:*

*httplib.BadStatusLine: No status line received –*

*the server has closed the connection*

*במקרה זה יש פשוט להריץ את Neo4Jמחדש ושוב להריץ את runimp27*

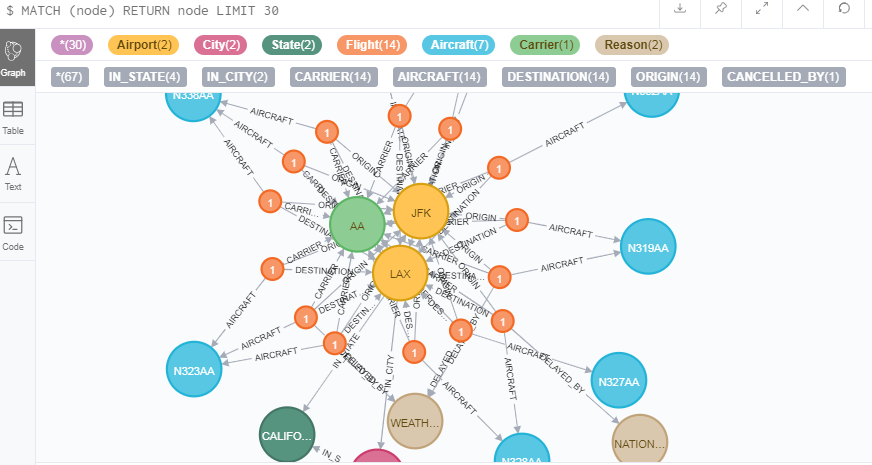
1. על מנת להכיר את הנתונים טוב יותר, בצעו את השאילתה:

MATCH (node)

RETURN node

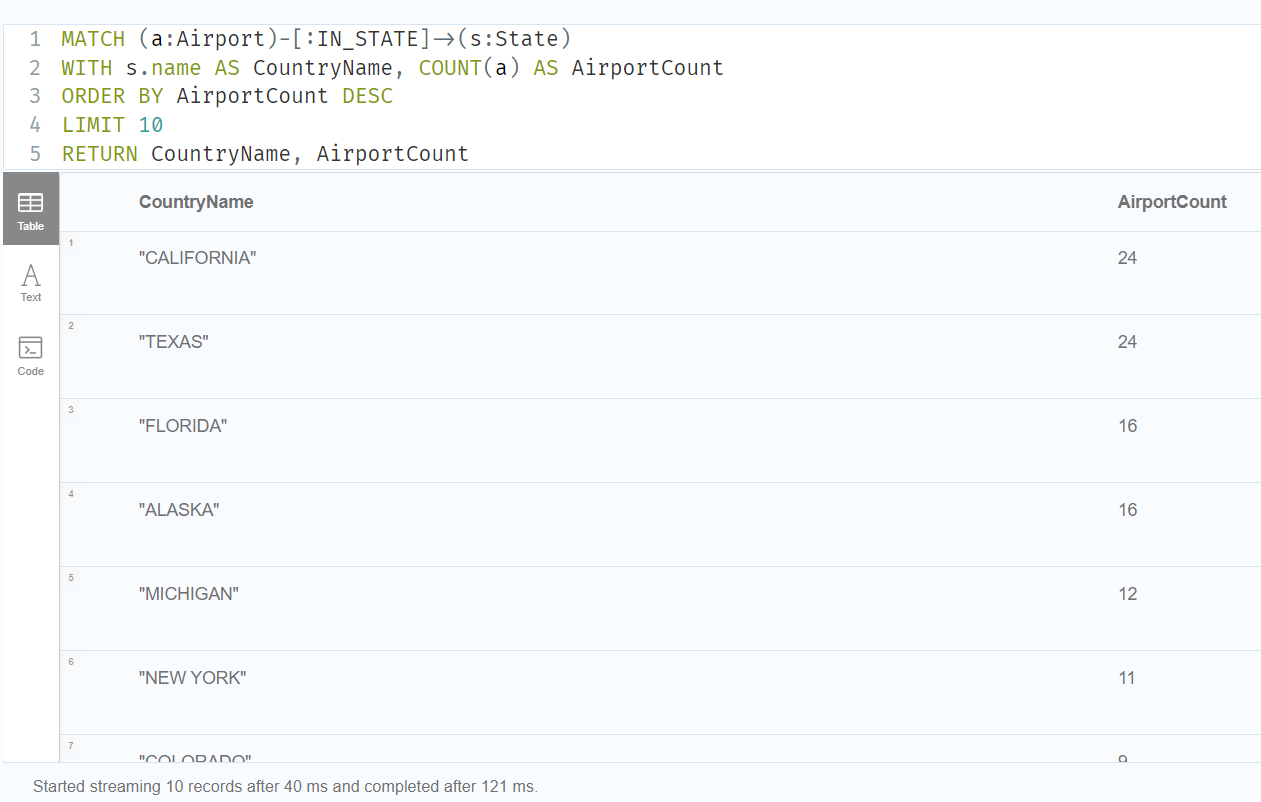
LIMIT 30

אשר תחזיר לכם נתונים של טיסות המכילים: שדות תעופה שונים, ערים, ארצות, מספרי טיסות, סוג/סימול המטוס, חברת התעופה וסיבות לאיחור בטיסה. הסבר על השדות תוכלו לקבל בלינק של המקור ממנו נלקחו הנתונים: <https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time>



1. שאילתות לביצוע:
   1. הציגו טבלה של 10 הארצות/מדינות אשר בהן הכי הרבה שדות תעופה. יש להציג את שם הארץ ואת מספר שדות התעופה, ממוינים מהגדול לקטן לפי מספר שדות התעופה.

MATCH (a:Airport)-[:IN\_STATE]->(s:State)  
WITH s.name AS CountryName, COUNT(a) AS AirportCount  
ORDER BY AirportCount DESC  
LIMIT 10  
RETURN CountryName, AirportCount



* 1. הציגו את הערים במדינת Alaska אשר יש בהן שדות תעופה

MATCH (a:Airport)-[:IN\_CITY]->(c:City)-[:IN\_STATE]->(s:State {name: "ALASKA"})  
RETURN c.name AS CityName

תמונה שמכילה טקסט, צילום מסך, תוכנה, תצוגה

התיאור נוצר באופן אוטומטי

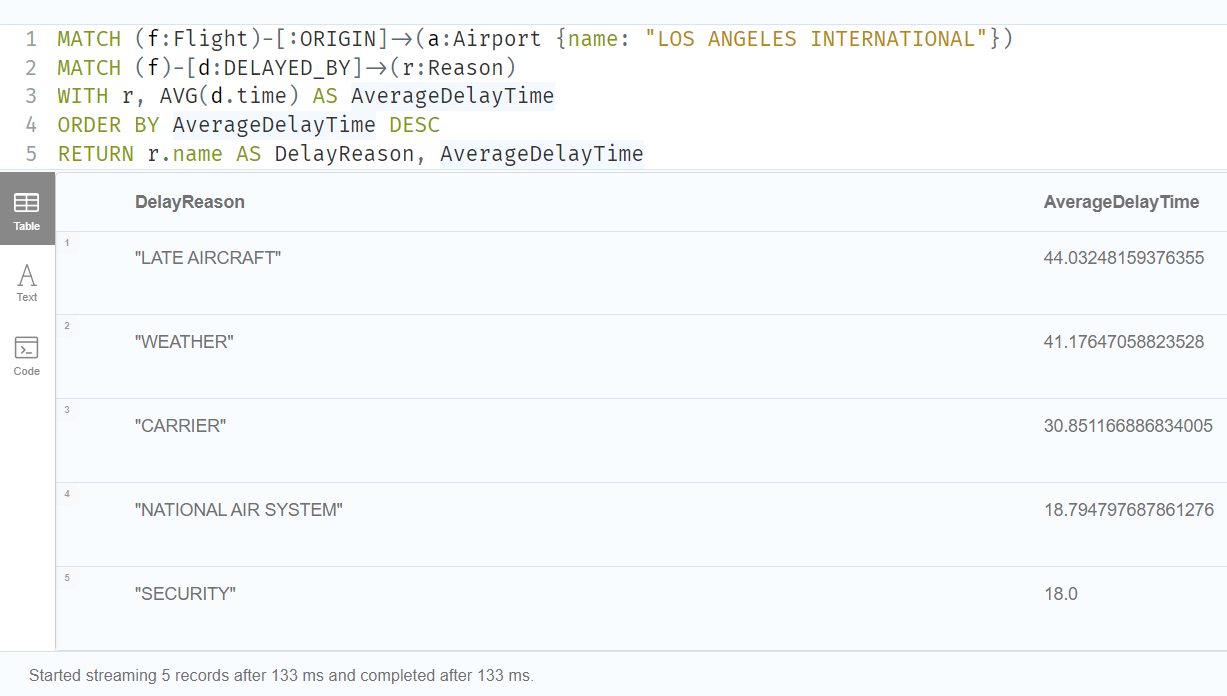
* 1. הציגו ערים אשר בהן יש יותר משדה תעופה אחד

MATCH (a:Airport)-[:IN\_CITY]->(c:City)  
WITH c, COUNT(a) AS AirportCount  
WHERE AirportCount > 1  
RETURN c.name AS CityName, AirportCount  
ORDER BY AirportCount DESC



* 1. הציגו טבלה המפרטת את הסיבות המובילות לעיכובים בטיסות בשדה התעופה LAX, ממוינת לפי ממוצע זמן העיכוב (מהעיכוב הכבד ביותר לקל) עבור כל סיבה.

MATCH (f:Flight)-[:ORIGIN]->(a:Airport {name: "LOS ANGELES INTERNATIONAL"})  
MATCH (f)-[d:DELAYED\_BY]->(r:Reason)  
WITH r, AVG(d.time) AS AverageDelayTime  
ORDER BY AverageDelayTime DESC  
RETURN r.name AS DelayReason, AverageDelayTime



* 1. הצג מסלול (ויזואלית) בין שדה התעופה JFK לשדה התעופה SFO, אשר לפחות עובר בשדה התעופה DFW בתאריך ה 8 בינואר (עשוי בדרך לעבור גם בשדות נוספים).
     1. על מנת להבין איך מבקשים מסלול, ניתן לקרוא כאן: <https://graphaware.com/graphaware/2015/05/19/neo4j-cypher-variable-length-relationships-by-example.html> במיוחד שימו לב לסימון כוכבית **\*** ומשמעותו

שימו לב כי על מנת לשפר ביצועים עדיף לקבוע מספר לצד הכוכבית המגביל את היקף החיפוש

MATCH (jfk:Airport{abbr:'JFK'})<-[:ORIGIN]-(F1:Flight{day:8, month:1})  
MATCH (sfo:Airport{abbr:'SFO'})<-[:DESTINATION]-(F2:Flight{day:8, month:1})  
MATCH (dfw:Airport{abbr:'DFW'})  
MATCH path = (F1)-[:ORIGIN|DESTINATION\*2]-(F2)  
WHERE dfw IN NODES(path)  
RETURN jfk, path, sfo

תמונה שמכילה טקסט, צילום מסך

התיאור נוצר באופן אוטומטי

* 1. יצרו קשתות חדשות בין שדות תעופה השייכים לאותה עיר עם שם הקשר: IN\_SAME\_CITY

MATCH (a:Airport)-[:IN\_CITY]->(c:City)<-[:IN\_CITY]-(a2:Airport)

WHERE a <> a2

MERGE (a)-[:IN\_SAME\_CITY]->(a2)

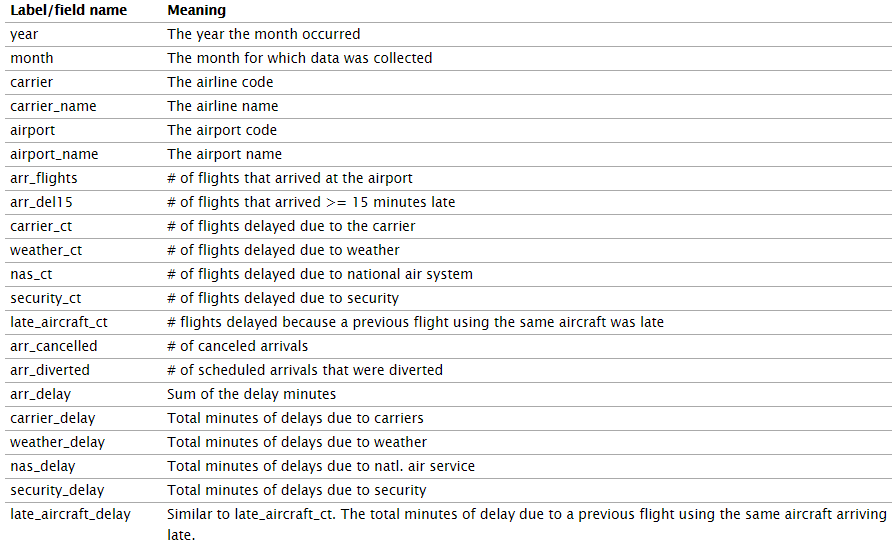
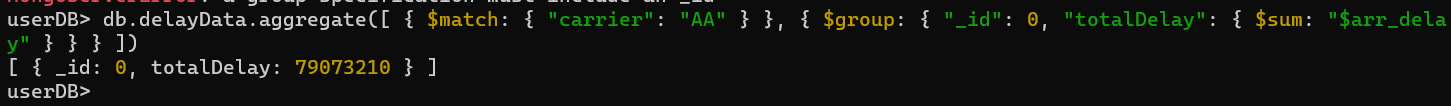
RETURN a, a2, c

תמונה שמכילה צילום מסך, טקסט

התיאור נוצר באופן אוטומטי

1. כעת נעבור ל **MongoDB**:
2. בספריה תוכלו למצוא קובץ בשם mongoAirlineDelay.csv המכיל מדגם של נתונים משנת 1987 ועד 2014
3. יש להעלות mongodb כפי שלמדנו באמצעות הפקודה mongod –dbpath=c:\tmp (ניתן לשנות את שם הספריה לספריה שיש לכם בה מקום)
4. כעת נייבא את הנתונים אל mongo באמצעות הפקודה:

mongoimport -d userDB -c delayData --type csv --file mongoAirlineDelay.csv --headerline

1. בזמן ייבוא הנתונים תוכלו לקבל חיווי כדוגמת: 
2. אחרי כן, יש כמובן לפתוח חלון cmd נוסף ולהעלות mongo.exe על מנת לכתוב פקודות.
3. על מנת לוודא שהנתונים הוכנסו כמו שצריך, עברו לחלון ה mongo ובחרו בבסיס הנתונים באמצעות הפקודה: use userDB . לאחר מכן כתבו db.delayData.findOne() אשר אמורה להציג לכם את המסמך הראשון ב collection שלנו.
4. לנוחיותכם, מובאת רשימת השדות:
5. שאילתות לביצוע:
   1. מצאו את מספר הדקות המצטבר אשר חברת American Airlines איחרה בה לאורך כל התקופה  
      

db.delayData.aggregate([{$match: {'carrier': 'AA'}}, {$group:{ '\_id': 0, 'totalDelay': { $sum: '$arr\_delay'}}}])

* 1. מצאו את מספר הדקות המצטבר של האיחור **עבור כל אחת** מחברות התעופה

תמונה שמכילה טקסט, צילום מסך, תוכנה, תכונות מולטימדיה

התיאור נוצר באופן אוטומטי

db.delayData.aggregate([{ $group: { \_id: "$carrier\_name", total\_delay: { $sum: "$arr\_delay" } } }, { $sort: { total\_delay: -1 } } ])

* 1. מצאו את שדה התעופה אשר בו היו הכי הרבה הסטות מסלול (arr\_diverted)

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

db.delayData.aggregate([ { $group: { \_id: "$airport\_name", total\_diversions: { $sum: "$arr\_diverted" } } }, { $sort: { total\_diversions: -1 } }, { $limit: 1 } ])

* 1. מצאו את שדה התעופה אשר בו היו הכי הרבה ביטולי טיסות בשנת 2012

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

db.delayData.aggregate([ { $match: { year: 2012 } }, { $group: { \_id: "$airport\_name", total\_cancellations: { $sum: "$arr\_cancelled" } } }, { $sort: { total\_cancellations: -1 } }, { $limit: 1 } ])

* 1. עבור **כל אחת** מחברות התעופה, מצאו את קטגוריית העיכוב התורמת את העיכוב הכי קטן ומצא את שדה התעופה שבו יש לחברה זו הכי הרבה עיכובים בגיל סיבה/קטגוריה זו

תמונה שמכילה טקסט, צילום מסך

התיאור נוצר באופן אוטומטי

db.delayData.aggregate([ { $group: { \_id: "$carrier", carrierName: { $first: "$carrier\_name" }, carrierDelay: { $sum: "$carrier\_delay" }, weatherDelay: { $sum: "$weather\_delay" }, nasDelay: { $sum: "$nas\_delay" }, securityDelay: { $sum: "$security\_delay" }, lateAircraftDelay: { $sum: "$late\_aircraft\_delay" } } }, { $project: { carrierName: 1, minDelayCategory: { $arrayElemAt: [ { $filter: { input: [ { delay: "carrier", value: "$carrierDelay" }, { delay: "weather", value: "$weatherDelay" }, { delay: "nas", value: "$nasDelay" }, { delay: "security", value: "$securityDelay" }, { delay: "lateAircraft", value: "$lateAircraftDelay" } ], as: "item", cond: { $eq: ["$$item.value", { $min: ["$carrierDelay", "$weatherDelay", "$nasDelay", "$securityDelay", "$lateAircraftDelay"] }] } } }, 0 ] } } }, { $lookup: { from: "delayData", let: { carrierCode: "$\_id", minCategory: "$minDelayCategory.delay" }, pipeline: [ { $match: { $expr: { $and: [ { $eq: ["$carrier", "$$carrierCode"] } ] } } }, { $group: { \_id: "$airport", totalDelay: { $sum: "$$minCategory" } } }, { $sort: { totalDelay: -1 } }, { $limit: 1 } ], as: "maxDelayAirport" } } ])

1. ממשו אתר ב Node.js (כפי שראינו בשיעור האחרון) המאפשר:
   1. לצפות ברשימת כל הטיסות בטבלה ב HTML
   2. הממשק יאפשר למחוק טיסות קיימות מבסיס הנתונים ולהוסיף טיסות חדשות
   3. הממשק יאפשר לפלטר/לסנן טיסות שיש להן:
      1. טיסות של carrier מסוים
      2. טיסות שנחתו בשדה תעופה מסוים

קישור גוגל דרייב לפרויקט nodeJS – לחץ [כאן](https://drive.google.com/file/d/1ZCJm9Y9P6oi5rA0VnduK4DGwEfABowA-/view?usp=sharing)

**בהצלחה בפרויקט ובתקופת הבחינות!**