שנה״ל **תשע״ט**, סמסטר **ב**, מועד ב
שאלון בחינה בקורס : **בסיסי נתונים וניתוח נתוני עתק**
מספר קורס: **240517**

מס׳ תלמיד: _____

קמפוס: _____
למילוי ע"י הסטודנט

- שם המרצה: ד"ר מאיר גולדנברג, ד"ר אריאלה ריכרדסון
- תאריך הבחינה: 06/08/2019
- משך הבחינה (בדקות): 150
- חומר עזר מותר לשימוש: לא
- מחשבון: לא
- יש לענות על כל החלקים וכל השאלות (אין בחירה). הציון המקסימלי במבחן הוא 100
- את התשובות יש לכתוב **ע"ג השאלון**, דפי הטיוטה לא ייבדקו.
- פירוט ניקוד: מצוין בכל שאלה

תלמיד יקר,

1. **נוהל הבחינות של המרכז האקדמי לב מחייב אותך**, באחריותך לקוראו ולהכירו - בחינה עלולה להיפסל על כל חריגה מהנוהל.
2. אם אינך מבין את כוונת המרצה בשאלה כלשהי, עליך לכתוב בראש התשובה כיצד הינך מבין את השאלה ולפתור בהתאם. המרצה ישקול האם יש מקום להבנה זו ואז ינקד בהתאם.
3. **חובה להחזיר את השאלון**. מחברת שלא יצורף לה השאלון, לא תיבדק!
4. לידיעתך, תורדנה נקודות לא רק על שגיאות, אלא גם על תוספות לא רלוונטיות, העדר נימוק הולם לתשובה, חוסר סדר ותשובה דו-משמעית, כאשר נדרשת תשובה חד משמעית.

בהצלחה רבה !

נוהל לכל השאלות:

עבור כל שאלה, יש לסמן על ידי V תשובה אחת בלבד בטבלה המופיעה למטה. אין צורך לנמק את התשובות. התשובות בתוך המחברת לא ייבדקו.

מספר	תשובה	ציון
.1	A B C D E	
.2	A B C D E	
.3	A B C D E	
.4	A B C D E	
.5	A B C D E	
.6	A B C D E	
.7	A B C D E	
.8	A B C D E	
.9	A B C D E	
.10	A B C D E	
סה"כ	**לא ניתן לכתוב כאן**	

1. נניח שפיתחנו חישוב מורכב משני תהליכי MapReduce. ה- Mapper בשם `map1.py` ו ה- Reducer בשם `red1.py` של התהליך הראשון נמצאים בתיקייה הנוכחית בשם `mr1` שהיא תת-תיקיית של תיקיית `comput` של תיקיית בית ב `Linux`. ה- Mapper בשם `map2.py` ו ה- Reducer בשם `red2.py` של התהליך השני נמצאים בתיקייה בשם `mr2` שהיא תת-תיקיית של תיקיית `comput` הנ"ל. הקלט עבור החישוב נמצא בקובץ `in.txt` בתיקייה הנוכחית. בעזרת איזו מהפקודות הבאות נבצע סימולציה לוקלית של החישוב ללא שימוש ב HDFS ונשמור את פלט החישוב בקובץ `out.txt` שבתיקיית `output` בתיקיית בית? ניתן להניח שהפלט של Mapper בשני השלבים הוא מילון עם מפתח מחרוזתי וערך מספרי.

A.
```
./map1.py | sort | ./red1.py < in.txt | \
../mr2/map2.py | sort | ../mr2/red2.py \
> ~/output/out.txt
```

B.
```
cat in.txt | ./map1.py | sort | ./red1.py | \
../mr2/map2.py | sort | ../mr2/red2.py \
> ~/output/out.txt
```

C.
```
cat in.txt | ./map1.py | sort | ./red1.py | \
mr2/map2.py | sort | mr2/red2.py \
> output/out.txt
```

D.
```
cat in.txt | ./map1.py | ./red1.py | \
../mr2/map2.py | ../mr2/red2.py | sort \
> ../../output/out.txt
```

E.
```
cat in.txt | ./map1.py | sort -grk 2 | ./red1.py | \
../mr2/map2.py | sort -grk 2 | ../mr2/red2.py \
> ~/output/out.txt
```

2. איזה מהמשפטים הבאים נכון לגבי חישוב בעזרת שיטת MapReduce:

- A. הפלט של תהליך MapReduce נשמר בזיכרון מרכזי של הקדקודים באשכול. לכן, אם יש צורך לשרשר מספר תהליכי MapReduce, אז צריך רק לשמור את הנתונים על דיסק קשיח לוקלי של כל קדקוד.
- B. יש צורך ב Mapper נפרד עבור כל קובץ קלט.
- C. כאשר משרשרים מספר תהליכי MapReduce, אז החל מהתהליך שני, יש צורך ב Mapper נפרד עבור כל קובץ קלט בגלל שכל קובץ כזה הוא פלט של התהליך הקודם.
- D. השלב של Sort & Shuffle מבצע מיון לפי ערך בנוסף למיון לפי מפתח.
- E. שימוש ב Combiner יכול להקטין את כמות ההעברות דרך הרשת בשלב של Sort & Shuffle.

3. אחד התרגילים ביקש לפתור את השאלה הבאה:

הקובץ ads.txt מכיל זוגות מורכבות משם מוצר ושם ערוץ בו המוצר מתפרסם, מופרדים ע"י פסיק. הנה חלק הקובץ של ads.txt :

```
Amazon Kindle,VBE
Amazon Kindle,CPI
Amazon Kindle,LYE
Apple AirPort,QWW
Apple AirPort,HWX
Apple AirPort,QZR
Apple AirPort,YBJ
```

הקובץ channels.txt מכיל זוגות מורכבות משם ערוץ וכמות הצופים שלו. הנה ההתחלה של channels.txt :

```
GID,14312
PAY,71628
BRG,71971
YCR,96656
FRV,44468
JEA,50557
IAZ,5080
```

השאלה: (בתרגיל זה יש שאלה אחת בלבד)

עליכם לפתח תהליך או שרשור תהליכי MapReduce ש- יחשב, עבור כל מוצר, את מספר הצופים הפוטנציאליים של הפרסומת שלו (כלומר, סה"כ צופים של כל הערוצים בהם המוצר מתפרסם). **כל תהליך MapReduce צריך להשתמש ב- 2 reducers לכל הפחות.**

הנה קוד אפשרי עבור שלב אחד של החישוב הנדרש:

```
def strIsInt(s):
    try:
        int(s)
        return True
    except ValueError:
        return False
def myPrint(gadgets, nViewers):
    for gadget in gadgets:
        print '%s\t%d' % (gadget, nViewers)

# maps words to their counts
lastChannel = ""
gadgets = []
nViewers = -1
for line in sys.stdin:
    parts = line.strip().split('\t')
    channel = parts[0]
    if channel != lastChannel:
        myPrint(gadgets, nViewers)
        lastChannel = channel
        gadgets = []
    if strIsInt(parts[1]):
        nViewers = int(parts[1])
    else:
        gadgets.append(parts[1])

if lastChannel != "": myPrint(gadgets, nViewers)
```

איזה מהמשפטים הבאים נכון עבור הקוד הזה?

- A. זה קוד של ה- Mapper בפתרון עם תהליך MapReduce אחד. בפרט, רואים שהוא מדפיס מילון לצורך השלב של Sort & Shuffle.
- B. זה קוד של ה- Reducer בפתרון עם תהליך MapReduce אחד. בפרט, רואים שהוא מסתמך על העובדה שהקלט שלו ממוין לפי שם ערוץ.
- C. זה קוד של Mapper שני בפתרון עם שני תהליכי MapReduce. בפרט, רואים שהוא מסתמך על העובדה שהקלט שלו ממוין לפי שם ערוץ, כי הקלט שלו הוא פלט של תהליך ה- MapReduce הראשון.
- D. זה קוד של ה- Reducer השני בפתרון עם שני תהליכי MapReduce. בפרט, רואים שהוא מסתמך על העובדה שהקלט שלו ממוין לפי שם ערוץ.
- E. זה קוד עבור פתרון עם מספר תהליכי MapReduce גדול משניים.

4. הנה קוד אפשרי (ללא `import`) עבור השאלה על פרסומות מוצרים הנ"ל בעזרת שיטת Spark:

```
def adsMap(el):
    els = el.split(',')
    return (els[1], els[0])

def channelsMap(el):
    els = el.split(',')
    return (els[0], els[1])

def printRes(rdd):
    for el in rdd.collect():
        print '%s\t%d' % (el[0].encode('utf-8'), el[1])

sc = SparkContext()
ads =
sc.textFile('/user/cloudera/targil5/ads.txt').map(adsMap)
channels =
sc.textFile('/user/cloudera/targil5/channels.txt').map(ch
annelsMap)
joined = ads.join(channels)
second = joined.map(lambda el: (el[1][0], int(el[1][1])))
res = second.reduceByKey(lambda mySum, el: mySum + el)
printRes(res)
```

איזה מהמשפטים הבאים נכון עבור הקוד הזה?

- A. הפונקציה `adsMap` תתבצע פעם אחת ותעבור בלולאה על כל הפרסומים.
- B. הפונקציה `adsMap` תתבצע פעם אחת עבור כל פרסום.

C. אם הקובץ `ads.txt` אינו קיים, אז המחשב לא יגיע בכלל לשורה של חישוב ה- RDD שנשמר במשתנה `channels`. זה אחד היתרונות של שיטת Spark.

D. אפילו אם הקובץ `ads.txt` אינו קיים, המחשב יגיע לשורות הבאות כדי לחשב Lineage. זה אחד החסרונות של שיטת Spark.

E. הקריאה ל `collect` תגרום לשמירת הפלט בדיסק קשיח.

5. איזה מבין המשפטים הבאים נכון:

A. אחד החסרונות של MapReduce הוא שלא ניתן לבצע חישוב בעזרת מחשבים רבים עם כמות זיכרון קטנה בכל מחשב.

B. לא ניתן להשתמש ב MapReduce ו Tableau בתהליך אנליזה אחד.

C. Spark תמיד קורא נתונים מ HDFS, מה שלא נכון על MapReduce.

D. MapReduce תמיד קורא נתונים מ HDFS, מה שלא נכון על Spark.

E. לא תיתכן עבודה עם נתונים שמורים ב HDFS בעזרת Tableau.

6. נניח שרוצים למצוא בתוך קובץ את כל המקומות שמתחילות שורה ברווח מסוג כלשהו, לאחר מכן יש להן אות A שאחריה מספר כלשהו של אותיות או ספרות או רווחים. איזו מהביטויים הרגולריים עשוי לבצע זאת?

A. `^[\s]A[d|\s|\w]?`
B. `^A[d|\s]*`
C. `^[\s]A[\w]*`
D. `^\sA[\s\w]*`
E. `^\s[\s\w]*`

7. למדנו שיש חשיבות רבה לניקוי של נתונים. איזו טענה מבין הטענות הבאות היא נכונה?

A. ניקוי הנתונים הוא שלב שמבצעים לאחר הניתוח שלהם, כאשר רואים אנומליות בתרשימים שמתקבלים בניתוח.

B. שלב ה ETL הוא אחד השלבים המעניינים ביותר שמבצע מנתח BigData למרות שהוא אינו דורש זמן רב.

C. פורמט של תאריכים יכול להיות רגיש בזמן הניתוח, ויש לשים לב שהפורמט אחיד בכל מקורות הנתונים, על מנת למנוע טעויות בניתוח.

D. כאשר מוחקים שורה בגלל שהיא פגומה, כדאי לשמור אותה. זה חשוב כי נרצה לעשות אנליזה על כל השורות הפגומות בנפרד.

E. ניתן לבצע בקלות את ניקוי הנתונים באופן ידני כי בדרך כלל ניתן לסמוך על כך שמקורות המידע נתקבלו מתוכנות אוטומטיות ולכן יש בהם מעט מאד שגיאות ובעיות.

8. בתמונה למטה רואים תרשים שנתקבל מתוך Tableau. מה מהטענות הבאות נכון? הגוון של האפור מייצג צבעים שונים (המבחן לא מודפס צבעוני, ניאלץ להסתפק בכך). TotalSales הוא שדה מחושב שיצרנו, שמכיל את המחיר הכולל של מכירה אחת.

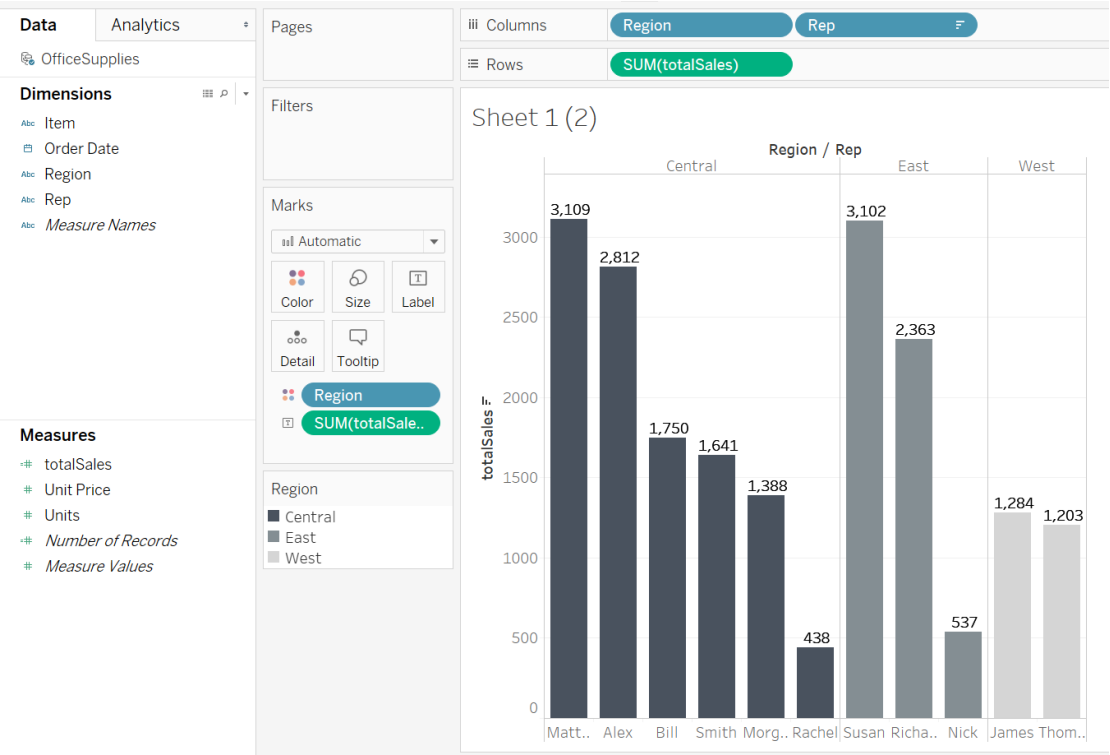
A. ניתן לראות בתרשים את מספר המוצרים המוצעים למכירה בכל אזור.

B. על מנת להוריד את הרישום המספרי שיש בעמודות בתרשים, יש ללחוץ על המספר בעכבר ימני.

C. הצביעה שישנה כאן אינה מועילה לצופה, והיה עדיף לצבוע כל עמודה בצבע אחר כדי שיהיה קל להפריד בין העמודות.

D. אם נגרור את Units ל Label נוכל לראות גם את מספר המוצרים שכל מוכר מכר בנוסף לסכום המכירות הכולל.

E. אם נגרור את Units ל Label נוכל לראות רק את מספר המוצרים שכל מוכר מכר ולא את סכום המכירות הכולל.



9. בתמונה למטה רואים תרשים של מפה שנתקבל מתוך Tableau. מה מהטענות הבאות נכון? הגוון של האפור מייצג צבעים שונים (המבחן לא מודפס צבעוני, ניאלץ להסתפק בכך).

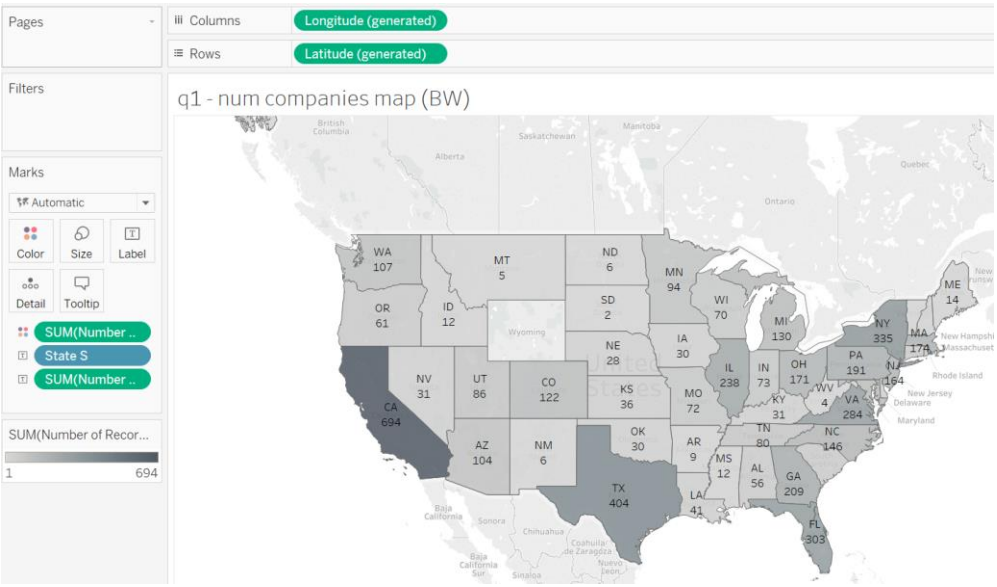
A. על מנת לייצר מפה כזאת, Longitude חייב להיות שדה שקיים בבסיס הנתונים המקורי.

B. אם נשים את Longitude ו Latitude שניהם ב Rows נקבל את המפה בעברית.

C. שם כל מדינה מופיע על המפה מכיוון שאחרי שצבענו את המפה גררנו את State למלבן שרשום עליו Color.

D. הצבעים במפה מייצגים את מספר הרשומות שיש בכל מדינה.

E. הסימון של שם המדינה ומספר הרשומות מופיע בכל אזור כיוון שגררנו ל Tooltip את הממדים הללו.



10. בעבודה ב Tableau מה מהטענות הבאות נכון?

A. החלוקה ל Measures ו Dimensions מבוצעת באופן אוטומטי ע"י Tableau בזמן טעינת הנתונים ואינה ניתנת לשינוי.

B. גרירה של אחד ה Dimensions ל- Color יכולה לגרום לצביעה של תרשים.

C. ניתן לסמן Label רק בראש העמודות של תרשים עמודות (bar chart, כמו בשאלה 8).

D. הנתונים המוצגים בתרשים קיימים תמיד גם במסך הכניסה בטבלה של הנתונים בצורה מפורשת.

E. Tableau מבצעת עבורנו בדיקות סטטיסטיות אוטומטיות, ואם רואים משהו בתרשים, זה סימן שקיימת עבורו מובהקות סטטיסטית.