שנה״ל **תשע״ט**, סמסטר **ב**, מועד א
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: 08/07/2019
- 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 שקיים קובץ בשם `a000.txt` בתיקיית `output` של HDFS. הקובץ מכיל פלט של תהליך MapReduce מאורגן לשתי עמודות – מפתח (מחרוזת) וערך (מספר). ברצוננו להוסיף 12 שורות בהם ה**מפתח** הוא גדול ביותר (במובן לקסיקוגרפי) לקובץ `out.txt` שקיים בתיקיית הורה של התיקייה הנוכחית במערכת הקבצים הלוקלית של המחשב ממנו אנחנו נגשים ל- HDFS. איזו מפקודות `shell` הבאות עלולה לבצע זאת? שימו לב שלא למדנו את פקודת `head` ולכן עליכם להשתמש גם בהגיון ואינטואיציה כדי לקבוע איזו מהפקודות היא נכונה.

A.
```
hdfs dfs -cat output/a.txt | sort -rgk 2 | \
head >> ../out.txt
```
B.
```
hdfs dfs -ls output/a.txt | sort -k 1 | \
head -n 12 >> ./out.txt
```
C.
```
hdfs dfs -cat output/a.txt | sort -rk 1 | \
head > ../out.txt
```
D.
```
hdfs dfs -cat output/a.txt | sort -rk 1 | \
head -n 12 >> ../out.txt
```
E.
```
hdfs dfs -cat output/a.txt | sort -rgk 2 | \
head >> /user/cloudera/out.txt
```

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

- A. הפלט של תהליך MapReduce נשמר בקובץ במערכת קבצים לוקלית (לא HDFS). לכן, אם יש צורך לשרשר מספר תהליכי MapReduce, אז צריך כל פעם להעביר את הנתונים ל HDFS והעברה זאת אמורה להיות יקרה בזמן.
- B. השלב של Map של MapReduce אינו דורש העברות נתונים דרך הרשת.
- C. השלב של Reduce של MapReduce אינו דורש העברות נתונים דרך הרשת.
- D. השלב של Reduce אינו יכול לעבוד נכון ללא השלב של Sort & Shuffle.
- E. השלב של Mapper חייב להכין מילון עבור השלב 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 (ללא import) עבור שלב אחד של החישוב הנדרש:

```
last = ""
nViewers = 0

for line in sys.stdin:
    parts = line.strip().split('\t')
    gadget = parts[0]
    if gadget != last:
        if last != "": print '%s\t%d' % (last, nViewers)
        last = gadget
        nViewers = 0
    nViewers += int(parts[1])

if last != "": print '%s\t%d' % (gadget, 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 מוצר.
- C. זה קוד של Mapper שני בפתרון עם שני תהליכי MapReduce. בפרט, רואים שהוא מסתמך על העובדה שהקלט שלו ממוין לפי שם מוצר, 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 מוצר. בנוסף, הוא מדפיס כל מוצר עם מספר הצופים שלו. אע"פ שתהליך ה- MapReduce הזה הוא שני ואחרון, מידע עבור שני מוצרים שונים יכול להופיע בקבצים שונים בפלט.

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. העובדה שקוראים את הקבצים מ HDFS היא אחד העקרונות של שיטת Spark.

B. ה RDD ששמרנו במשתנה `second` היה אפשר לקבל גם בעזרת `flatMap`, כי היא פועלת כמו `map` במקרה הזה.

C. אפשר לשפר את הפתרון הזה על ידי החלפת `reduceByKey` ב- `groupByKey` וטרנספורמציה נוספת.

D. אפשר להגיע לפתרון אחר נכון על ידי החלפת `reduceByKey` ב- `groupByKey` וטרנספורמציה נוספת, אבל הפתרון המופיע למעלה הינו מהיר יותר.

E. הקוד הזה אינו נכון, כי הוא מאבד את `el[0]` בשורה שמחשבת את ה- RDD הנשמר במשתנה `second`.

5. איזה מבין המשפטים הבאים נכון:

A. שיטת MapReduce היא חשובה רק בגלל הכלים המבוססים עליה.

B. שיטת Spark היא שיפור משמעותי על MapReduce, אבל שימוש בה יכול לדרוש משאבים יקרים.

C. ראינו דוגמאות של חישובים שניתן לבצע בעזרת MapReduce, אבל לא ניתן לבצע בעזרת Spark.

D. ראינו דוגמאות של חישובים שניתן לבצע בעזרת Spark, אבל לא ניתן לבצע בעזרת MapReduce.

E. קיימים חישובים שניתן לבצע בעזרת Hive, אבל לא ניתן לבצע בעזרת MapReduce.

6. נניח שרוצים למצוא בתוך קובץ את כל השורות שמתחילות בספרה ואינן מסתיימות באות A, ובאמצע יש אותיות ורווחים בלבד. איזו מהביטויים הרגולריים עשוי לבצע זאת?

A. `^\d[\w\d]+[^A]$`
B. `$\d[\d\s]+[^A]^`
C. `^\d[\w\s]+[^A]$`
D. `^[\0\9][\w\s]*[^A]$`
E. `^\d[\w\s]+[^A]*`

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

A. פתיחת קובץ נתונים ב Excel ושמירתו היא דרך בטוחה לטיפול בניקוי נתונים. השינויים שמוכנסים ע"י Excel תמיד מועילים לניקוי.

B. פתיחת קובץ נתונים ב Excel, ביצוע שינויים רצויים ושמירתו היא דרך אפשרית לטיפול בניקוי נתונים, אבל השינויים שמוכנסים ע"י Excel לפעמים פוגמים בנתונים.

C. כדאי לבצע את שלב ה ETL ישירות על קובץ הנתונים המקורי, ככה יובטח לנו שנעבוד רק עם נתונים נקיים.

D. בגלל שעבודה עם גרשיים (") יכולה מאד לבלבל את Excel כדאי להשתמש בביטויים רגולריים ולנקות את כל הגרשיים מקובץ הנתונים לפני תחילת העבודה.

E. בעיות של שיבושים בנתונים הם תמיד מקומיות ואינן גורמות לכמה שדות להתמזג לשדה אחד.

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.

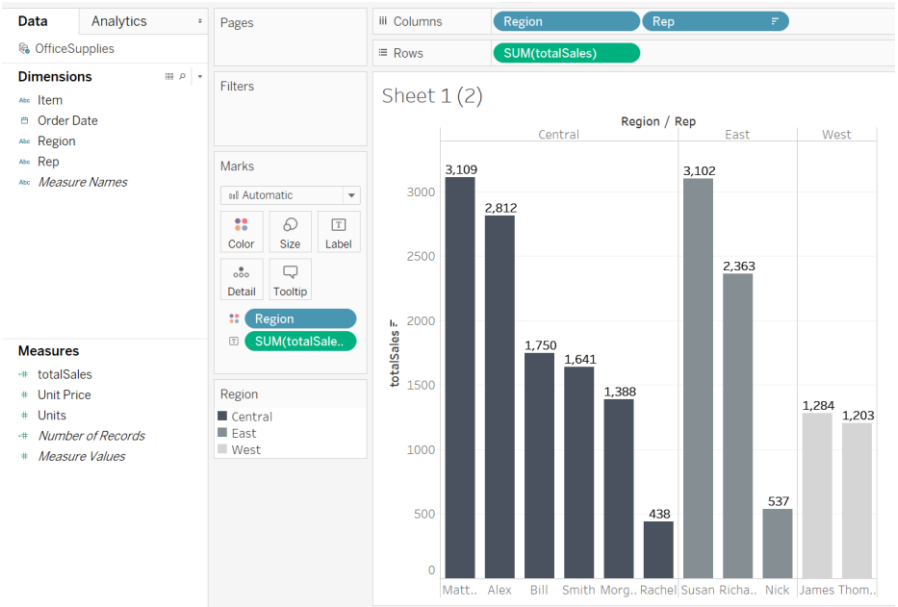
A. ניתן לראות בתרשים את סכום המכירות שביצע כל מוכר, כאשר המוצרים מחולקים לפי סוג המוצר.

B. מטרת הצביעה (בגווני אפור) בתרשים היא כדי להבדיל בין מי שמכר הרבה פריטים למי שמכר מעט.

C. לו רצינו לראות את סכום המכירות הכללי בכל אזור, השינוי שעלינו לבצע הוא, לגרור את הממד Region ל Label.

D. לו רצינו לראות את סכום המכירות הכללי בכל אזור, השינוי שעלינו לבצע הוא, להוריד הממד Rep מ Columns.

E. על מנת לשנות את הצביעה של התרשים, חייבים לגרור את המלבן שרשום בו Color עד שיגיע מעל התרשים.



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

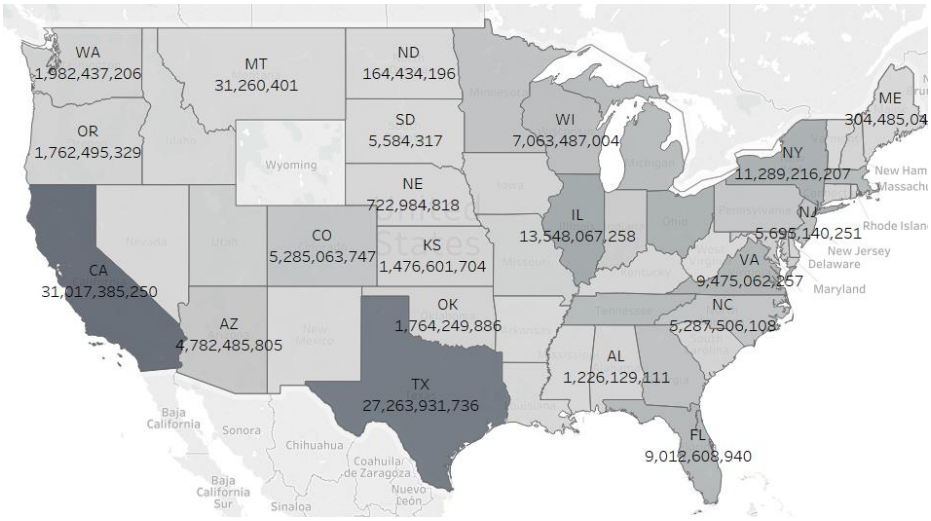
A. יש כאן מידה וממד (Dimension and Measure) שמשמשות ל Label ומידה (Measure) אחת ל Color.

B. יש כאן שתי מידה וממד (Dimension and Measure) שמשמשות ל Size ומידה (Measure) אחת ל Color.

C. על מנת לבנות מפה כזאת יש לגרור לכל אזור במפה את הסוג של הנתונים שרוצים להציג באותו האזור.

D. הצביעה במפה מטרתה ליצור שונות בין האזורים במפה כדי שיהיה קל להפריד ביניהם בצפייה מהירה.

E. אם נגרור את המידה Workers ל Size נקבל תווית בה רשום מספר העובדים בכל מדינה.



10. בעבודה ב Tableau מה מהטענות הבאות **אינו** נכון?

A. החלוקה ל Measures ו Dimensions מבוצעת ע"י המשתמש בזמן של יצירת תרשים.

B. ישנה אפשרות ליצור שדה מחושב.

C. צבעים יכולים לשמש לחידוד ויזואלי של המידע שרוצים להדגיש.

D. שימוש ב ShowMe מאפשר לשנות באופן מהיר את צורת התרשים.

E. שכפול של גליון ושינוי רכיב מצומצם ממנו הוא מאד נוח כדי להשיג אנליזה דומה עם שדות אחרים.