1. באיזה מהמצבים הבאים **לא** נשתמש ב- Hadoop:
2. הנתונים שמורים בדיסקים קשיחים של עשר אלף מחשבים.
3. למחשבים העומדים לרשותנו זיכרון מרכזי קטן.

Hadoop can be used. Hadoop stores data in a distributed file system that allows it to be processed in parallel, making it less dependent on the machine's central memory.

1. אפשר לבטא חישוב בצורת שאילתא בשפת SQL.
2. דרוש שימוש במחשב על.

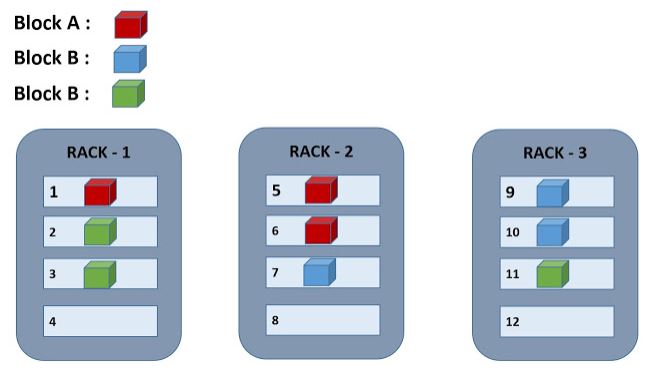
Supercomputers are highly specialized systems designed to handle complex computational tasks very quickly. If such resources are available, it might be more efficient to use these instead of a distributed system like Hadoop, depending on the task.

1. המחשבים העומדים לרשותנו אינם נמצאים באותו מקום גאוגרפי.
2. המחשבים העומדים לרשותנו משתמשים בסוגים שונים של חומרה (מעבד, לוח אם וכו').

Hadoop can be used. Hadoop can run on different hardware configurations. As long as the machines meet Hadoop's minimum requirements, it should be able to function across various types of hardware.

ANSWER: D

1. איזה מהמשפטים הבאים נכון על MapReduce:
2. MapReduce היא הרחבה של Python.
3. נניח שהאיור המופיע למטה (נלקח מהמצגת) מתאר שמירה של קובץ שלם ב- HDFS. עבור הקובץ הזה, MapReduce יריץ שלוש Mappers.



1. MapReduce מונע באופן מוחלט את הצורך להעביר נתונים דרך הרשת, כי אנחנו מעבירים חישובים לנתונים ולא להפך.

This statement is partially true. MapReduce does aim to reduce the amount of data that needs to be transferred over a network by performing calculations near where the data is stored, which is a principle known as data locality. The idea is to move the computation close to where the data is located rather than moving large amounts of data over the network.

However, it's not completely accurate to say that MapReduce prevents the need to transfer data through the network entirely. During the shuffle and sort phase of MapReduce, intermediate data (the output from the Map phase) is transferred across the network to the Reduce tasks. Thus, there still is some network I/O involved.

So, the statement is mostly true, but not entirely accurate. MapReduce significantly reduces the need for network data transfers but does not eliminate it completely.

1. כאשר משתמשים ב- MapReduce, המתכנת צריך לכתוב שלוש תכניות – Mapper, Sorter ו- Reducer.

The "Sort and Shuffle" phase that happens between the Map and Reduce stages is handled by the MapReduce framework itself, and the programmer typically does not need to explicitly write any code for this phase. This phase ensures that all data belonging to a single key is directed to the same reducer, which is essential for the reduce step to work correctly.

1. ישנם כלים ב- Hadoop שעושים שימוש ב- MapReduce, אך עדיף לכתוב תכנית ב- Python.
2. MapReduce היא השיטה היחידה של חישוב מבוזר שיכולה לעשות שימוש ב- HDFS.

ANSWER: B

1. נניח שרוצים למצוא בתוך קובץ את כל התאריכים במבנה dd-MM-yyyy איזו מהביטויים הרגולריים עשוי לבצע זאת?
2. \d{2}-\M{2}-\y{4}
3. ^\d{2}-\d{2}-\d{4}
4. \*^\d{2}-\d{2}-\d{4}\*
5. \d\d-\d(2)-\d{4}
6. \d{2}-\d{2}-\d{4}
7. \d[2]-\d[2]-\d[4]

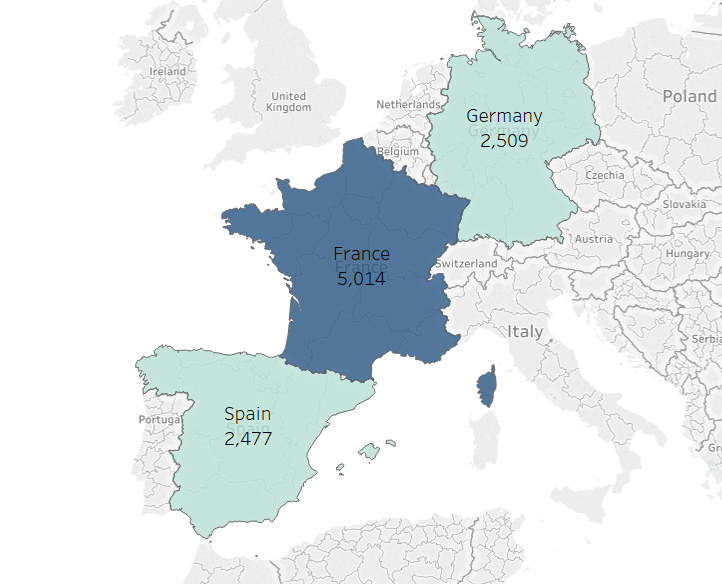
ANSWER: E

1. נניח שרוצים למצוא בתוך קובץ את כל השורות שמסתיימות במחרוזת המורכבת מ- בין 2 ל 6 הופעות של A ברצף, אחר כך XX או YY ואחר כך תווים כלשהם אך לא ספרות. דוגמה של שורה: AAAAYYa%d. איזו מהביטויים הרגולריים עשוי לבצע זאת?
2. ^A+[\d]+[XY]$
3. A{2,6}(XX|YY)[^\d]\*$
4. A{2,6}(XX|YY)[^\w]\*$
5. A{2-6}(XX|YY)[\d]\*^
6. AAA{2,6}(XX|YY)[0-9]\*$
7. A\*(XX|YY)[^0-9]\*$
8. ^A+[\w]+[X\*Y\*]$

ANSWER: B

1. בתמונה למטה רואים תרשים של מפה שמתקבלת מתוך Tableau. איזו מהטענות הבאות נכונה?
2. עבור 3 המדינות הצבועות גררנו את שם המדינה ל- Color ולכן מוצגות דווקא מדינות אלו
3. יש מדינות שמופיעות בלבן (לדוגמא Italy), כי לא הגדרנו עבורן צבע.
4. על מנת לבנות מפה כזאת יש לגרור לכל אזור במפה את המידות שרוצים להציג באותו האזור.
5. יתכן שהצביעה במפה מטרתה להציג את גודל הערכים של מידה מסוימת, אך בתמונה אין מידע מפורש שאומר מהי המידה.
6. הגודל של הגופן במפה נקבע לפי מספר התושבים.

ANSWER: D

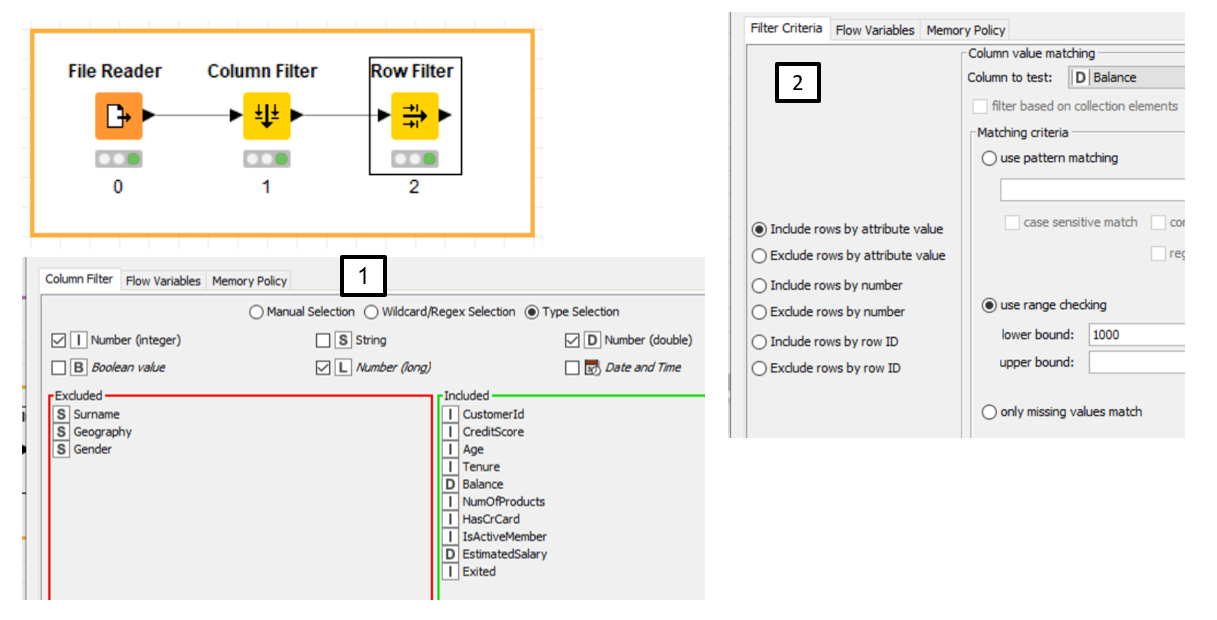


1. בעבודה ב Knime איזו מהטענות הבאות נכונה?
2. לכל קודקוד יש לפחות שתי כניסות (קלטים) ולפחות שתי יציאות (פלטים).
3. לא ניתן לעשות configure עבור כל קודקוד בנפרד, אלא צריך לעשות configure עבור כל הקודקודים ביחד.
4. ישנם הרבה סוגי קודקודים, כל אחד מבצע פעולה מסוג אחר.משתמשים ב- configure כדי להגדיר פרטים של הפעולה.
5. לכל קודקוד יש כניסה (קלט) אחת ויציאה (פלט) אחת.
6. ניתן לראות את הנתונים רק בהתחלה (כלומר, אחרי קריאה מהקובץ) או בסוף (כלומר, בפלט של הקודקוד האחרון) התהליך, אלא אם כן הוספנו קודקוד מסוג "data" שמאפשר ציפייה בנתונים באמצע התהליך.

ANSWER: C

1. על פי המוצג בתרשים Knime הבא עבור אנליזה של נתונים על לקוחות בנק, איזו מהטענות הבאות נכונה?

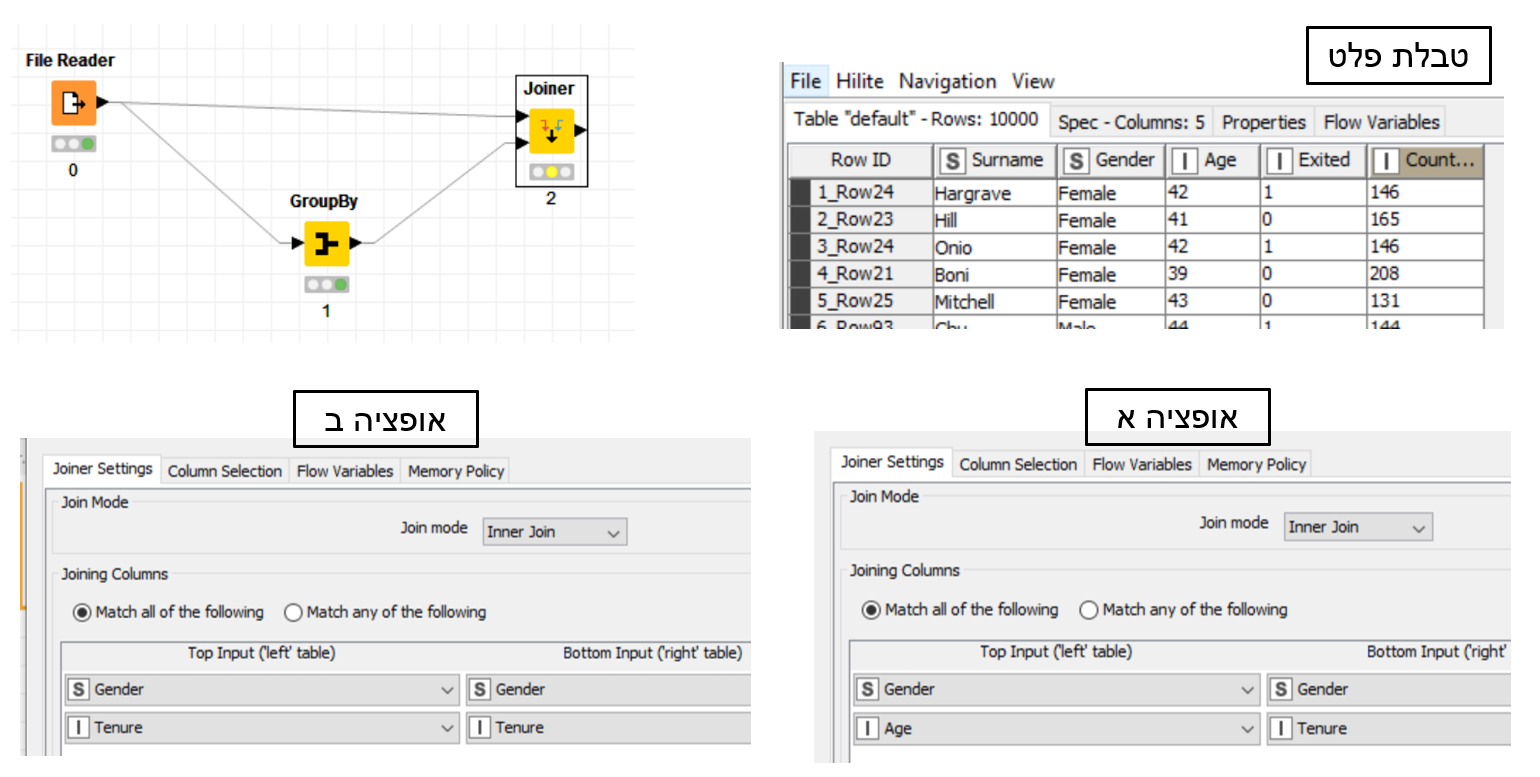
שימו לב שכל חלון של מאפיינים מופיע עם מספר של הקודקוד המתאים לו.



1. לאחר הפעלת קודקוד 1 נקבל שורות של לקוחות שעבורם הגיל (age) מוגדר אבל חסר מידע על שם המשפחה שלהם (Surname) בקובץ הנתונים.
2. לאחר הפעלת קודקוד 2 נקבל לקוחות שבהם חסר מידע בעמודות שמופיעות בריבוע האדום עם התג Excluded בקודקוד 1.
3. לאחר הפעלת קודקוד 2 נקבל שורות של לקוחות שיש להם יתרה (Balance) של מעל 1000, אבל לא יופיעו העמודות שמופיעות בריבוע האדום עם התג Excluded בקודקוד 1.
4. אנחנו רואים שכל הקודקודים מוכנים להפעלה כי לכולם יש אור ירוק. והפלט של ההפעלה יכיל את השורות שיש במלבן הירוק בקודקוד 1.
5. בשדה ColumnToTest בקודקוד 2 מופיע Balance כי הוא שדה מחרוזת. אבל שדות של מספר לא יופיעו שם באפשרויות הבחירה.

ANSWER: C

1. על פי המוצג בתרשים Knime הבא איזו מהטענות הבאות נכונה?



1. בהנחה שהפעלנו את קודקוד 2 עם הפרמטרים לפי אופציה **ב** נקבל את טבלת הפלט המוצגת.
2. בהנחה שהפעלנו את קודקוד 2 עם הפרמטרים לפי אופציה **א או ב** נקבל את טבלת הפלט המוצגת.
3. אף על פי שלאופציה **א** אין משמעות כל-כך, היא תרוץ ללא שגיאה.
4. טבלת הפלט המוצגת שייכת בוודאות לקודקוד 1, ולא לקודקוד 2.
5. אופציה **ב** אינה אופציה חוקית.

ANSWER: C

1. למדנו דוגמת שימוש ב- MapReduce עבור השאלה הבאה. קיים קובץ קלט של שורות, כאשר בכל שורה מידע על מילה (שמציגה שירות מסוימת), שם משתמש ותאריך. קובץ קלט לדוגמה:

Apple 12.03.2018 John

Banana 10.03.2018 Rick

Banana 12.03.2018 Patrick

Apple 12.03.2018 Rodger

המטרה היא לחשב שני קבצי פלט. הקובץ הראשון יכיל, עבור כל תאריך ש- יש בו שימוש במילה ועבור כל מילה, שורה עם התאריך, המילה ומספר השימושים במילה הזאת בתאריך הזה. עבור הקלט לדוגמה, הקובץ הראשון ייראה כך:

10.03.2018 Banana 1

12.03.2018 Apple 2

12.03.2018 Banana 1

הקובץ השני יכיל, עבור כל מילה שהיה בה שימוש, שורה עם המילה ומספר סה"כ השימושים במילה הזאת בכל התאריכים. עבור הקלט לדוגמה, הקובץ השני ייראה כך:

Apple 2

Banana 2

השיטה שלמדנו פתרה את השאלה ע"י שרשור שני שלבים, כאשר השלב הראשון מחשב את תוכן קובץ הפלט הראשון והשלב השני משתמש בנתונים אלה כדי לחשב את תוכן קובץ הפלט השני. השאלה הבאה מתייחסת לשיטה הזאת ו- מניחה שהקלט מורכב ממיליוני שורות. איזה מהמשפטים הבאים נכון?

1. לא ניתן ליישם את השיטה הזאת ב- Spark.
2. ניתן ליישם את השיטה הזאת ב- Spark. השלב השני של החישוב יצטרך לקרוא את קובץ הפלט הראשון מזיכרון משנה (כלומר, דיסקים קשיחים, HDFS וכו').
3. ניתן ליישם את השיטה הזאת ב- Spark. נצטרך לכתוב תכנית נפרדת עבור כל שלב.
4. ניתן ליישם את השיטה הזאת ב- Spark. יתר מזה, הדרישה לטפל במיליוני שורות דורשת שימוש ב- Spark דווקא (ולא ב- MapReduce).
5. ניתן ליישם את השיטה הזאת ב- Spark. השלב השני של החישוב **לא** יצטרך לקרוא את קובץ הפלט הראשון מזיכרון משנה (כלומר, דיסקים קשיחים, HDFS וכו').
6. ניתן ליישם את השיטה הזאת ב- Spark. שום גודל קלט לא יהווה בעיה בהינתן שטח מספיק בדיסקים קשיחים (או HDFS).

1. איזו מהפקודות הבאות של לינוקס תיצור תיקייה חדשה בשם myDir בתוך התיקייה הנוכחית ואז תעביר (כלומר, תזיז) אל התיקייה החדשה את כל הקבצים והתיקיות בתיקייה הנוכחית שהשם שלהם מתחיל באות z?
2. mkdir myDir; mv z\* myDir
3. mv z\* myDir
4. mv ./z\* myDir
5. mkdir myDir; mv z\* > myDir
6. mkdir myDir; cp z\* > myDir
7. cat z? >> myDir

ANSWER: A

1. בעבודה ב Tableau איזו מהטענות הבאות נכונה?
2. צבעים בתרשימים משמשים בעיקר ליופי ועל מנת למשוך את הקורא בעזרת אסתטיקה.
3. הכלי של Tableau הוא מאד שימושי אבל דורש תכנון מדויק לפני שמתחילים לעבוד על התרשים כי קשה לשנות את פורמט התרשים.
4. גרירה של הסימן "color" למפה תצבע את המפה לפי ארצות.
5. על מנת לבצע מבחן AB חייבים לייצר Alias.
6. Reference Line משמש כדי להשוות תרשים אחד לתרשים אחר.
7. אם מתעורר הצורך, אפשר לייצר שדות (אטריבוטים) חדשים שמחושבים מתוך שדות קיימים.

ANSWER: F

1. בעבודה ב Knime איזו מהטענות הבאות נכונה?
2. Knime הוא כלי שמיועד לעבודה על מבנה נתונים. את הניתוח עושים ב Tableau.
3. אחד היתרונות של Knime שראינו, הוא שאת הנתונים הנקיים שהתוכנה מייצרת אפשר להעביר בקלות ל Weka או לממשק מבוסס Python לצורך ביצוע כריית מידע.
4. Knime מאפשר ביצוע של כריית מידע במלואו החל מתהליך קריאת וניקוי הנתונים דרך בנייה ובדיקה של מודל כריית המידע ועד להצגת תוצאות הבדיקה.
5. אחד היתרונות של Knime שראינו הוא שאת הנתונים הנקיים שהתוכנה מייצרת אפשר להתממשק בקלות ל Weka או לממשק מבוסס Python לצורך ביצוע כריית שבסופה התוצאות מוחזרות ל Knime להצגה.
6. Knime מאפשר אפיון של תהליך כריית מידע ע"י האפשרות לשים קודקודים שמייצגים, אך לא מבצעים, את התהליך של קריאה וניקוי הנתונים, הכרייה והצגת תוצאות הבדיקה.

ANSWER: C