תרגיל מסכם – Big Data

רקע

במסגרת התרגיל המסכם התבקשנו לבנות תהליך MapReduce אשר דרכו נקבל כתוצרי עיבוד נתונים סטטיסטיים על כמות המשקעים בין השנים 1900 – 2018.

קובץ הנתונים שקיבלנו הינו בפורמט csv, כאשר כל שורה בנויה משנה ולאחריה 12 ערכים כאשר כל ערך מבטא את כמות המשקעים שירדה באותו חודש (ינואר, פברואר וכו').

אנו כתבתנו תוכנית Java פשוטה אשר מוסיפה עמודה נוספת של נתונים לאותו קובץ ובנוסף משנה את פורמט הקובץ לטקסט כאשר התו המפריד הוא רווח ולא פסיק. עמודת הנתונים שהוספנו מכילה את ממוצע המשקעים השנתי.

הסבר תהליך העבודה

על מנת לחשב את הנתונים הסטטיסטיים כמתואר לעיל, ביצענו שלושה תהליכי MapReduce שונים, כאשר בכל תהליך חישבנו נתונים שונים על מנת לפשט את התהליך.

שלושת התהליכים אותם ביצענו הם:

1. התהליך הראשון הוא *RainFallMinMaxStatistics*.
בתהליך זה חישבנו את הנתונים הבאים:
   - חודש ושנה בהם ירדה כמות המשקעים החודשית המקסימלית.
   - חודש ושנה בהם ירדה כמות המשקעים החודשית המינימלית.
   - השנה בה ירדה כמות המשקעים הגדולה ביותר.
   - השנה בה ירדה כמות המשקעים הקטנה ביותר.

   תהליך המיפוי מתבצע כאשר קוראים כל שורה מהקובץ ומפרקים אותה ל – tokens.

   בתהליך זה אנו יוצרים זוגות (key, value) בצורה הבאה – המפתח הוא השנה, והערך הוא כמות משקעים חודשית. בנוסף, מכיוון שבתהליך זה אנו עוברים על כל הערכים של כמויות המשקעים אנו מחשבים בנוסף את החודש והשנה בהם ירדה כמות משקעים מינימלית ומקסימלית. אנו מבצעים זאת באמצעות שני משתנים פשוטים ותהליכי השוואה. בסופו של דבר אנו יוצרים שני זוגות נוספים של הנתונים הללו עם מפתחות מיוחדים על מנת לשלוף את הנתונים הללו בתהליך ה – Reduce ולהציג אותם בקובץ כתוצאה סופית.

   בתהליך ה – Reduce מתבצע aggregation וכתוצאה מכך נוצרים זוגות של (key, value) בצורה הבאה – המפתח הוא שנה, והערך הוא כמות שנתית. באמצעות הנתונים הללו שקיבלנו לאחר הסכימה בתהליך ה – Reduce שוב עם שני משתנים פשוטים ותהליכי השוואה אנו מחשבים את השנה בה ירדה כמות המשקעים הגדולה ביותר, והשנה בה ירדה כמות המשקעים הקטנה ביותר.

   בסופו של תהליך מבצעת מתודת ה – cleanup אשר כותבת את התוצאות הסופיות לקובץ הפלט.

2. התהליך שני הוא *RainFallSeasonsStatistics*.
בתהליך זה חישבנו את הנתונים הבאים:
- התקופה בשנה בה ירדה כמות המשקעים הגדולה ביותר.
- התקופה בשנה בה ירדה כמות המשקעים הקטנה ביותר.

תהליך המיפוי מתבצע כאשר קוראים כל שורה מהקובץ ומפרקים אותה ל – tokens.

בתהליך זה אנו יוצרים זוגות (key, value) בצורה הבאה – המפתח הוא השנה משורשרת לעונה, והערך הוא כמות משקעים חודשית. ע"פ הגדרת התרגיל עונה בשנה מגודרת כשלושה חודשים - לדוגמה: חודשים 2-1-12 מהווים את עונת החורף. לכן, במהלך תהליך המיפוי אנו בודקים לאיזה עונה שייך אותו חודש ולפי כך יוצרים את הזוג כמתואר לעיל.

בתהליך ה – Reduce מתבצע aggregation וכתוצאה מכך נוצרים זוגות של (key, value) בצורה הבאה – המפתח הוא שנה משורשרת לעונה, והערך הוא כמות משקעים עונתית לאחר סיכום. באמצעות הנתונים הללו שקיבלנו לאחר הסכימה בתהליך ה – Reduce עם שני משתנים פשוטים ותהליכי השוואה אנו מחשבים את העונה והשנה בה ירדה כמות משקעים מקסימלית, ואת העונה והשנה בה ירדה כמות משקעים מינימלית.

בסופו של תהליך מבצעת מתודת ה – cleanup אשר כותבת את התוצאות הסופיות לקובץ הפלט.

3. התהליך שני הוא *RainFallDroughtStatistics*.
בתהליך זה חישבנו את הנתונים הבאים:
- ממוצע משקעים רב שנתי.
- תקופות הבצורת, כאשר ע"פ התרגיל הן תקופות אשר מכילות 3 שנים לפחות בהן כמות המשקעים קטנה מהממוצע הרב שנתי.

תהליך המיפוי מתבצע כאשר קוראים כל שורה מהקובץ ומפרקים אותה ל – tokens.

בתהליך זה אנו יוצרים זוגות (key, value) בצורה הבאה – המפתח הוא מחרוזת אחידה לכל הזוגות והערך הוא שנה משורשרת לממוצע משקעים שנתי, משורשר לכמות משקעים שנתית. ביצענו את הזוגות עם מפתח זהה לכולם על מנת שתהליך ה – Reduce לאחר מכן נוכל לחשב בקלות את הממוצע הרב שני. בנוסף לשרשורי המידע בכל ערך נבצע parsing ונשתמש במידע בתהליך ה – Reduce.

בתהליך ה – Reduce מתבצע parsing באמצעותו את מחלצים את המידע מכל ערך. לאחר מכן, במתבצע aggregation וכתוצאה מכך אנו מחשבים את הממוצע הרב שנתי. כמו כן, אנו שומרים את הממוצע השנתי של כל שנה בטבלה ממוינת כאשר במפתח הוא השנה, והערך הוא הממוצע. לאחר שיש בידנו את הטבלה הממוינת ואת הממוצע הרב שני, אנו עוברים בלולאה על הטבלה ובכך אנו מאתרים בצורת – 3 שנים לפחות בהן הממוצע קטן מהממוצע הרב שנתי.

בסופו של תהליך אנו כותבים את התוצאות הסופיות לקובץ הפלט.

<u>פקודות הידור, הרצה ומיקומי קבצים</u>

- קבצי הקוד – קבצי הקוד הם שלושת קבצי ה – MapReduce שתוארו לעיל, בנוסף לקבוץ ה – Parser אשר רץ בפני עצמו וממיר את קובץ הנתונים לפורמט תואר לעיל.
מיקום הקבצים הללו הוא ב - cloudera בנתיב user/kobi.cohen123452/MR_Project.

- קובץ הנתונים – קובץ נתונים נמצא גם כן בסביבת ה – cloudera, תחת הנתיב user/kobi.cohen123452/MR_Project/Input/.
השם של הקובץ הוא ParsedDataFile.txt.

- קבצי הפלט – קובץ הפלט של כל תהליך MapReduce נוצר לבד באותה התיקייה שבה רץ התהליך. כאמור, חשוב שלא תהיה תיקיית Output לפני הרצת התהליך, מכיוון שהיא נוצרת לבד וקיומה לפני עלול להכשיל את הרצת התהליך. לדוגמה: עבור התהליך *RainFallMinMaxStatistics* קובץ הפלט נמצא במיקום:
user/kobi.cohen123452/MR_Project/MinMax/Output/

- הסבר תהליך הידור והרצה:
נסביר לדוגמה את ההידור וההרצה של תהליך *RainFallDroughtStatistics*, כאשר הפעולה זהה בכל שאר התהליכים.

1. יצירת classpath עבור Hadoop – *export HADOOP_CLASSPATH=$(hadoop classpath)*
2. בדיקת האם ה – classpath נוצר בהצלחה – *echo $HADOOP_CLASSPATH*

```
kobi.cohen123452@cloudera:~/test$ echo $HADOOP_CLASSPATH
/etc/hadoop/conf:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/libexec/../../hadoop/lib/*:/opt/cloudera/parcels/CDH-5.1
5.1-1.cdh5.15.1.p0.4/lib/hadoop/libexec/../../hadoop/.//*:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/libexec/../../h
adoop-hdfs/./:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/libexec/../../hadoop-hdfs/lib/*:/opt/cloudera/parcels/CDH-5
.15.1-1.cdh5.15.1.p0.4/lib/hadoop/libexec/../../hadoop-hdfs/.//*:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/libexec/
../../hadoop-yarn/lib/*:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/libexec/../../hadoop-yarn/.//*:/opt/cloudera/parc
els/CDH/lib/hadoop-mapreduce/lib/*:/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/.//*
```

3. יצירת תיקייה בשם classes – *mkdir classes*

4. הידור הקובץ –

javac -classpath ${HADOOP_CLASSPATH} -d 'classes' RainFallDroughtStatistics.java

5. יצירת קובץ jar – *jar -cvf RainFallDroughtStatistics.jar -C classes/ .*

6. יצירת קובץ jar הניתן להרצה בשרת בסביבת cloudera –

hadoop jar 'RainFallDroughtStatistics.jar' RainFallDroughtStatistics

/user/kobi.cohen123452/MR_Project/Input

/user/kobi.cohen123452/MR_Project/Drought/Output

```
kobi.cohen123452@cloudera:~/MR_Project/Drought$ hadoop jar 'RainFallDroughtStatistics.jar' RainFallDroughtStatistics /user/kobi.cohen123452/MR_Project/Input /user/kobi.cohen123452/MR_Project/Drought/Output
19/06/26 19:50:45 INFO client.RMProxy: Connecting to ResourceManager at cloudera.kinneret.ac.il/172.16.5.1:8032
19/06/26 19:50:45 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/06/26 19:50:46 INFO input.FileInputFormat: Total input paths to process : 1
19/06/26 19:50:46 INFO mapreduce.JobSubmitter: number of splits:1
19/06/26 19:50:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1556771433160_0044
19/06/26 19:50:46 INFO impl.YarnClientImpl: Submitted application application_1556771433160_0044
19/06/26 19:50:46 INFO mapreduce.Job: The url to track the job: http://cloudera.kinneret.ac.il:8088/proxy/application_1556771433160_0044/
19/06/26 19:50:46 INFO mapreduce.Job: Running job: job_1556771433160_0044
19/06/26 19:50:51 INFO mapreduce.Job: Job job_1556771433160_0044 running in uber mode : false
19/06/26 19:50:51 INFO mapreduce.Job:  map 0% reduce 0%
19/06/26 19:50:56 INFO mapreduce.Job:  map 100% reduce 0%
19/06/26 19:51:02 INFO mapreduce.Job:  map 100% reduce 100%
19/06/26 19:51:02 INFO mapreduce.Job: Job job_1556771433160_0044 completed successfully
19/06/26 19:51:02 INFO mapreduce.Job: Counters: 49
```

<u>הערה: אסור שתהיה קיימת תיקיית Output לפני ההרצה, מכיוון שהיא נוצרת באופן אוטומטי בהרצה. אם תהיה קיימת תיקיה כזו ייזרק חריג.</u>

פירוט הפלט

1. תהליך *RainFallMinMaxStatistics*:

קטע מתוך קובץ הפלט של התהליך.

Home / user / kobi.cohen123452 / MR_Project / MinMax / Output / **part-r-00000**

```
1988    1505
1989    1309
1990    1513
1991    1338
1992    1720
1993    1740
1994    1527
1995    1367
1996    1508
1997    1300
1998    1147
1999    1831
2000    2026
2001    1752
2002    1638
2003    1638
2004    1513
2005    1750
2006    1922
2007    1663
2008    1651
2009    1713
2010    1542
2011    1686
2012    1327
2013    1549
2014    1526
2015    1319
2016    1611
2017    1591
2018    1284
The month and year in which the greatest amount of precipitation fell are 2/1922, the amount is        350
The month and year in which the lowest amount of precipitation fell are 6/1901, the amount is   0
The year in which the biggest amount of precipitation fell is 2000, the amount is       2026
The year in which the smallest amount of precipitation fell is 1945, the amount is      1120
```

2. תהליך *RainFallSeasonsStatistics:*
קטע מתוך קובץ הפלט של התהליך.

Home / user / kobi.cohen123452 / MR_Project / Seasons / Output / **part-r-00000**

```
2010-winter     642
2011-fall       340
2011-spring     623
2011-summer     36
2011-winter     687
2012-fall       296
2012-spring     396
2012-summer     38
2012-winter     597
2013-fall       271
2013-spring     645
2013-summer     48
2013-winter     585
2014-fall       295
2014-spring     618
2014-summer     22
2014-winter     591
2015-fall       281
2015-spring     410
2015-summer     9
2015-winter     619
2016-fall       270
2016-spring     497
2016-summer     28
2016-winter     816
2017-fall       363
2017-spring     502
2017-summer     44
2017-winter     682
2018-fall       239
2018-spring     423
2018-summer     28
2018-winter     594
The season and year in which the largest precipitation fell are 2000-winter     977
The season and year in which the smallest precipitation fell are 1979-summer    6
```

3. תהליך *RainFallDroughtStatistics*:

Home / user / kobi.cohen123452 / MR_Project / Drought / Output / **part-r-00000**

```
The multi annual average is 126
The following years detected as drought [1900 - 120, 1901 - 113, 1902 - 112]
The following years detected as drought [1933 - 115, 1934 - 125, 1935 - 122, 1936 - 113]
The following years detected as drought [1976 - 114, 1977 - 95, 1978 - 112]
The following years detected as drought [1987 - 109, 1988 - 125, 1989 - 109]
The following years detected as drought [1995 - 113, 1996 - 125, 1997 - 108, 1998 - 95]
```