תכנות מערכות מבוזרות - עבודה 2

ולדימיר שרגורודסקי - 310016621
דור שטרקר 204467757
שמות משתמש: shargvla dorsht

Results Bucket

https://s3.amazonaws.com/vladi-dsp-191-ass2-bucket/local-aggregation-output/Result-14-26-45.814/part-r-00000

Running instructions (linux)

1. **ליצור תיקיה ".aws" ב-home, ובתוכה ליצור קובץ credentials בפורמט הבא:**

```
[default]
aws_access_key_id = ABCD
aws_secret_access_key = ABCD
```

2. **לפתוח bucket שבו תישמר התוצאה ב s3.**
3. **ליצור רולים: AmazonElasticMapReduceRole ו AmazonElasticMapReduceforEC2Role ב IAM**
4. **להעתיק את הקובץ userInfo שבתיקיה הזאת לתיקיה שממנה תריצו את הפרוייקט, ולשנות את הפרטים בו בהתאם.**
5. **לקמפל את הפרויקט בתיקיה DeletedEstimationsRunner באמצעות הפקודה mvn package**
6. **להריץ באמצעות java -jar DeletedEstimationsRunner-1.0.jar**
7. **התוצאה תמצא ב bucket/deleted-estimations-output**
8. **הלוג ימצא ב bucket/deleted-estimations-logs**

Map-Reduce Steps Explenation

התוכנית שלנו מורכבת מ 7 שלבים.

- שלב ה-Split:
  - הקורפוס מחולק לשני חלקים.
  - עבור כל שלשה שומרים את מספר המופעים שלו בכל חלק
  - סוכמים את סך כמות המופעים בקורפוס (N)

- שלב ה-Nr :
  - לכל מספר מופעים, סוכמים כמה פעמים הוא מופיע בשני החלקים.
- שלב ה-Tr:
  - לכל מספר מופעים, סוכמים כמה פעמים הוא מופיע בחלק השני.

- שלב ה-MergeNr:
  - עבור שני קבצי ה-Nr, מבצעים איחוד לקובץ אחד.
- שלב ה-MergeTr:
  - עבור שני קבצי ה-Tr, מבצעים איחוד לקובץ אחד.

- שלב ה-Merge:
  - עבור קבצי ה-Nr וה-Tr מבצעים איחוד על מנת לחשב את ההסתברויות.
- שלב ה-Sort:
  - עוברים על הקובץ שהתקבל בשלב הקודם (Merge) וממיינים את השלשות על פי הדרישה בעבודה.

Statistics

Local Aggregation	With	Size (bytes)	Without	Size (bytes)
Split	4,734,708	139935580	100,085,231	929608375
Nr0	23,266	73700	2,367,318	1593292
Nr1	23,223	73937	2,367,318	1593424
Tr01	23,266	137820	2,367,318	7271914
Tr10	23,223	137447	2,367,318	7271909
MergeNr0Intermediate	2,371,923	41908157	2,371,923	41907796
MergeNr1Intermediate	2,371,942	41910734	2,371,942	41911367
MergeNr	4,734,636	56696922	4,734,636	56697513
MergeTr01Intermediate	2,371,923	41921384	2,371,923	41921384
MergeTr10Intermediate	2,371,942	41910398	2,371,942	41910594
MergeTr	4,734,636	58687692	4,734,636	58687692
Merge	4,734,636	73675326	4,734,636	73675326
Sort	2,367,318	81383625	100,085,231	81411671

מסקנות:
תעבורת הרשת עם local aggregation נמוכה בהרבה, אך מבחינת זמן ריצה שתי השיטות רצו בערך בזמן שקול (36 לעומת 38 דקות)

Analysis

- הקרן הקיימת:
  1. לישראל
  2. את
  3. וקרן
  4. לא
  5. על

- ומטעם זה:
  1. לא
  2. אין
  3. עצמו
  4. גם
  5. היה

- לתת אמון:
  1. רב
  2. מלא
  3. בכל
  4. בדברי
  5. באיש

- היה לו:
  1. לומר
  2. לכתוב
  3. כל
  4. שום
  5. מה

- שר האוצר:
  1. של
  2. על
  3. ושר
  4. הראשון
  5. היה

- לאחר סיום:
  1. המלחמה
  2. לימודיו
  3. מלחמת

4. הלימודים
5. הקרבות

- היה יכול:
  1. להיות
  2. לעמוד
  3. לעשות
  4. לומר
  5. לסבול

- בכל מקצועות:
  1. התורה
  2. החיים
  3. היהדות
  4. הספרות
  5. העבודה

- לו בנים:
  1. זכרים
  2. ובנות
  3. ממנה
  4. קטנים
  5. או

- האדם מן:
  1. העולם
  2. הבהמה
  3. העבירה
  4. החטא
  5. הטבע

התשובות שקיבלנו אכן הגיוניות, ברוב המקרים ניתן לראות כי המילה השלישית מתוך חמשת המילים בעלות ההסתברות הגבוהה ביותר עבור הזוג היא זו שהיינו מצפים לקבל.