Introduction to Massive Data Analysis

Term Project

107065529 何沛臻

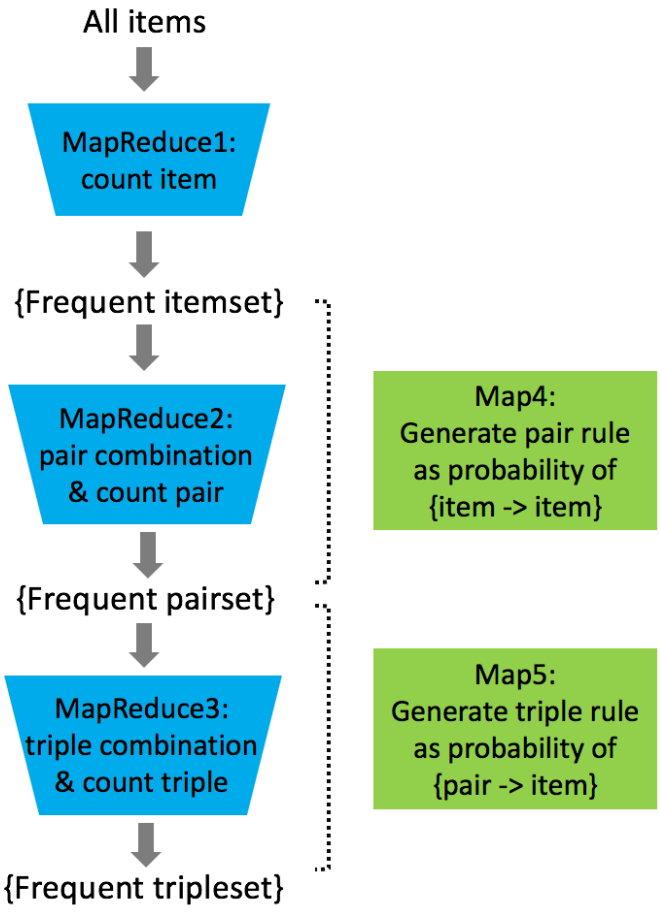
- 實作題目：關聯規則分析

- 資料集說明：共有 31101 筆購物資料，每筆購物資料中有多個產品，如下圖

```
FRO11987 ELE17451 ELE89019 SNA90258 GRO99222
GRO99222 GRO12298 FRO12685 ELE91550 SNA11465 ELE26917 ELE52966 FRO90334 SNA30755 ELE17451 FRO84225 SNA80192
ELE17451 GRO73461 DAI22896 SNA99873 FRO86643
ELE17451 ELE37798 FRO86643 GRO56989 ELE23393 SNA11465
ELE17451 SNA69641 FRO86643 FRO78087 SNA11465 GRO39357 ELE28573 ELE11375 DAI54444
ELE17451 GRO73461 DAI22896 SNA99873 FRO18919 DAI50921 SNA80192 GRO75578
ELE17451 ELE59935 FRO18919 ELE23393 SNA80192 SNA85662 SNA91554 DAI22177
ELE17451 SNA69641 FRO18919 SNA90258 ELE28573 ELE11375 DAI14125 FRO78087
ELE17451 GRO73461 DAI22896 SNA80192 SNA85662 SNA90258 DAI46755 FRO81176 ELE66810 DAI49199 DAI91535 GRO94758 ELE94711 DAI22177
ELE17451 SNA69641 DAI91535 GRO94758 GRO99222 FRO76833 FRO81176 SNA80192 DAI54690 ELE37798 GRO56989
ELE17451 GRO73461 DAI22896 GRO99222 SNA47306 GRO36567 ELE82555 SNA17715 SNA94781 DAI87514 GRO48282 GRO12935 SNA55952 DAI93692
GRO99222 DAI48891 GRO36567 ELE82555 SNA17715 SNA94781 DAI87514 SNA55952 DAI93692 GRO12935 GRO48282 DAI92253 FRO82427 ELE17451
```

http://snap.stanford.edu/class/cs246-data/browsing.txt

- 運算架構：使用 A-Priori 演算法，運算架構如下圖



- 程式說明：
  - 使用 **Python Hadoop Streaming**
  - Mapper & Reducer 設計：
    1. 計算 Frequent item：此部分用 MapReduce 進行，先在 mapper 計算每個 item 的出現次數，傳到 reducer 後將相同編號的 item 在每個 mapper 的出現次數加總，只留下出現次數超過 threshold 的 item（threshold 在此設為出現 100 次），並存成 Frequent item set。
    2. 計算 Frequent pair：此部分用 MapReduce 進行，在 mapper 將每個購物籃中的 item 兩兩組合成一個 pair，剔除任一 item 不包含在 frequent item set 的 pair 並計算 pair 的出現次數，傳到 reducer 後將相同 pair 的出現次數加總，只留下出現次數超過 threshold 的 pair（threshold 在此亦設為 100 次），並存成 Frequent pair set。
    3. 計算 Frequent Triple：此部分用 MapReduce 進行，在 mapper 將每個購物籃中的任三個 item 組合成一個 triple，剔除任一 item 不包含在 frequent item set 的 triple 並計算 triple 的出現次數，傳到 reducer 後將相同 triple 的出現次數加總。最後只留下出現次數超過 threshold 的 triple（threshold 在此設為 50 次），並存成 Frequent triple set。
    4. 計算｛Item A -> Item B｝的關聯規則：此部分僅用 Mapper 進行。Frequent pair 中，Item A -> Item B 及 Item B -> Item A 的兩種關聯規則分別拆開計算。
       Output：{ Item A -> Item B } : probability
    5. 計算｛Item A, Item B -> Item C｝的關聯規則：此部分僅用 Mapper 進行。Frequent triple 中，Item A, Item B -> Item C、Item A, Item C -> Item B 及 Item B, Item C -> Item A 的三種關聯規則分別拆開計算。
       Output：{ Item A, Item B -> Item C } : probability

  - Sequential 方法驗證：這次的 project 有另寫 sequential 方法檢查結果是否相同。觀察到用 sequential 方式較 MapReduce 方法快很多，不過因為是將所有資料讀入，在資料量大時會有 Memory 不足的問題。

- 結果說明：
  - **Top 10 Pair rule：**

```
DAI93865->FRO40251 : 1.0
GRO85051->FRO40251 : 0.999176276771
GRO38636->FRO40251 : 0.990654205607
ELE12951->FRO40251 : 0.990566037736
DAI88079->FRO40251 : 0.986725663717
FRO92469->FRO40251 : 0.983510011779
DAI43868->SNA82528 : 0.972972972973
DAI23334->DAI62779 : 0.954545454545
ELE92920->DAI62779 : 0.732664995823
DAI53152->FRO40251 : 0.717948717949
SNA18336->DAI62779 : 0.713681241185
ELE55848->GRO32086 : 0.709459459459
```

  - **Top 10 Triple rule**

```
GRO38814 GRO85051->FRO40251 : 1.0
ELE20847 FRO92469->FRO40251 : 1.0
DAI31081 GRO85051->FRO40251 : 1.0
ELE20847 GRO85051->FRO40251 : 1.0
DAI75645 GRO85051->FRO40251 : 1.0
DAI55911 GRO85051->FRO40251 : 1.0
DAI23334 ELE92920->DAI62779 : 1.0
GRO21487 GRO85051->FRO40251 : 1.0
ELE17451 GRO85051->FRO40251 : 1.0
ELE26917 GRO85051->FRO40251 : 1.0
FRO53271 GRO85051->FRO40251 : 1.0
GRO85051 SNA45677->FRO40251 : 1.0
DAI62779 DAI88079->FRO40251 : 1.0
GRO73461 GRO85051->FRO40251 : 1.0
GRO85051 SNA80324->FRO40251 : 1.0
DAI62779 GRO85051->FRO40251 : 0.997382198953
DAI75645 DAI88079->FRO40251 : 0.993288590604
DAI88079 GRO73461->FRO40251 : 0.993103448276
DAI88079 ELE17451->FRO40251 : 0.991935483871
FRO92469 GRO73461->FRO40251 : 0.990610328638
DAI62779 FRO92469->FRO40251 : 0.98347107438
ELE17451 FRO92469->FRO40251 : 0.981818181818
```

  - Performance：31101 筆資料，7 mins 執行完成