BÀI THỰC HÀNH MAPREDUCE - 1

Bài toán

Cho bộ dữ liệu liên quan đến mức tiêu thụ điện của một tổ chức. Dữ liệu chứa mức tiêu thụ điện hàng tháng và mức trung bình hàng năm trong các năm khác nhau.

Yêu cầu: tìm ra các năm có lượng điện tiêu thụ trung bình > 30

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
1979	23	23	2	43	24	25	26	26	26	26	25	26	25
1980	26	27	28	28	28	30	31	31	31	30	30	30	29
1981	31	32	32	32	33	34	35	36	36	34	34	34	34
1984	39	38	39	39	39	41	42	43	40	39	38	38	40
1985	38	39	39	39	39	41	41	41	00	40	39	39	45

Link download dữ liệu: file sample.txt

Hướng dẫn thực hiện

Về cơ bản, toàn bộ chương trình có thể được chia thành ba phần:

- **Mapper**: đọc mỗi dòng dữ liệu từ tệp dữ liệu đầu vào, phân tách dữ liệu thành năm và giá trị tiêu thụ điện trung bình. Năm được sử dụng làm khóa và giá trị là số điện tiêu thụ trung bình.
- **Reducer**: Reducer nhận các cặp key-value từ mapper. Trong phương thức reduce, nó so sánh giá trị tiêu thụ điện trung bình với giá trị tối đa được đặt là 30. Nếu giá trị nào lớn hơn 30, nó sẽ xuất ra cặp key-value tương ứng.
- **Main**: Phần chính của chương trình cấu hình các thông số của công việc MapReduce như lớp mapper và reducer, định dạng dữ liệu đầu vào và đầu ra, đường dẫn đến các tệp dữ liệu đầu vào và đầu ra. Sau đó, nó chạy công việc MapReduce.

Hướng dẫn thực hiện

- Kết quả của chương trình:

```
1981    34
1984    40
1985    45
```

Lưu ý: Tại bước chạy kết quả, sinh viên lưu kết quả vào thư mục đặt tên là BT1_MSV (VD: BT1_22012005)

BÀI THỰC HÀNH MAPREDUCE - 2

Bài toán

Dữ liệu đầu vào được sử dụng là tệp SalesJan2009.csv, chứa thông tin liên quan đến việc Bán hàng như: Tên sản phẩm, giá cả, phương thức thanh toán, thành phố, quốc gia, v.v. Mục tiêu của bài tập là tìm số lượng sản phẩm được bán ở mỗi quốc gia.

1	Transaction_date	Product	Price	Payment_Type	Name	City	State	Country	Account_Created	Last_Login
2	1/2/09 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	1/2/09 6:00	1/2/09 6:08
3	1/2/09 4:53	Product1	1200	Visa	Betina	Parkville	MO	United States	1/2/09 4:42	1/2/09 7:49
4	1/2/09 13:08	Product1	1200	Mastercard	Federica e Andrea	Astoria	OR	United States	1/1/09 16:21	1/3/09 12:32
5	1/3/09 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/05 21:13	1/3/09 14:22
6	1/4/09 12:56	Product2	3600	Visa	Gerd W	Cahaba Heights	AL	United States	11/15/08 15:47	1/4/09 12:45
7	1/4/09 13:19	Product1	1200	Visa	LAURENCE	Mickleton	NJ	United States	9/24/08 15:19	1/4/09 13:04
8	1/4/09 20:11	Product1	1200	Mastercard	Fleur	Peoria	IL	United States	1/3/09 9:38	1/4/09 19:45
9	1/2/09 20:09	Product1	1200	Mastercard	adam	Martin	TN	United States	1/2/09 17:43	1/4/09 20:01

Link download dữ liệu: File SalesJan2009.csv

Hướng dẫn thực hiện

Về cơ bản, toàn bộ chương trình có thể được chia thành ba phần:

- **Mapper Code:** Mapper đọc mỗi dòng dữ liệu từ tệp đầu vào, phân tách dữ liệu thành các trường bằng dấu phẩy và sau đó gửi ra cặp key-value, trong đó key là quốc gia (được lấy từ trường thứ 7 trong dữ liệu đầu vào), và value là số 1 (đại diện cho một giao dịch bán hàng).
- **Reducer Code:** Reducer nhận các cặp key-value từ mapper. Đối với mỗi quốc gia, nó tổng hợp tất cả các giá trị số 1 (tương ứng với số lượng giao dịch) và gửi ra một cặp key-value mới, trong đó key vẫn là quốc gia và value là tổng số lượng giao dịch.
- **Driver Code:** Đây là lớp chính của chương trình. Nó cấu hình các thông số của công việc MapReduce như các lớp mapper và reducer, định dạng dữ liệu đầu vào và đầu ra, đường dẫn đến các tệp dữ liệu đầu vào và đầu ra. Sau đó, nó chạy công việc MapReduce.

Hướng dẫn thực hiện

- ❑ Bước 1: Chuẩn bị dữ liệu:
  - Sinh viên download bộ dữ liệu sau đó đẩy lên HDFS.
- ❑ Bước 2: Tạo Java Project:
  - Sinh viên tạo Java Project như đã thực hiện với ví dụ về WordCount.
- ❑ Bước 3: Thực hiện viết mã cho chương trình

- ❑ Bước 4: Chạy chương trình và đọc kết quả

Minh họa kết quả của chương trình

Lưu ý: Tại bước chạy kết quả, sinh viên lưu kết quả vào thư mục đặt tên là BT2_MSV (VD: BT2_22012005)