# Mô tả bài toán

Medicare là chương trình bảo hiểm y tế quốc gia của Mỹ được quản lý trực tiếp bởi chính quyền liên bang. Trong chương trình này, các nhà cung cấp bảo hiểm (provider) sẽ thay mặt bệnh nhân là khách hàng của họ (patient) để đệ trình yêu cầu thanh toán bảo hiểm tới Medicare. Các yêu cầu này sẽ được xem xét, đánh giá để sau đó đưa ra mức chi trả thực tế một cách thích hợp. Thông thường khoản chi trả thực tế chỉ bằng một phần của giá trị yêu cầu thanh toán. Do ngân sách ngày càng bị thắt chặt, Medicare chịu sức ép lớn phải thẩm tra kỹ lưỡng những yêu cầu thanh toán bảo hiểm này nhằm đảm bảo chúng là hợp lý và tránh các trường hợp sai sót hoặc cố tình lừa đảo. Nhiệm vụ của bạn trong project này là sử dụng Hadoop MapReduce để tiến hành phân tích danh sách các yêu cầu thanh toán bảo hiểm của Medicare và tìm ra những nhà cung cấp cũng như những vùng địa lý có dấu hiệu bất thường, đáng nghi vấn nhất.

# Mô tả dữ liệu

<http://www.mediafire.com/view/ho53j2qgccyj6nn/outpatient>

<http://www.mediafire.com/view/y29u47n9ylsgnmq/inpatient>

Dữ liệu đầu vào là 2 file text: inpatient (điều trị nội trú) và outpatient (điều trị ngoại trú). Mỗi file gồm nhiều dòng, mỗi dòng gồm 11 trường thông tin phân cách nhau bởi ký tự TAB. Ý nghĩa 11 trường thông tin đó (từ trái sang phải) cụ thể như sau:

1. Mã thủ tục khám chữa bệnh
2. Mã nhà cung cấp bảo hiểm
3. Tên nhà cung cấp
4. Địa chỉ nhà cung cấp
5. Tên thành phố nơi có nhà cung cấp
6. Viết tắt bang nơi có nhà cung cấp
7. Mã zip code của nhà cung cấp
8. Thông tin địa lý
9. Số lượt yêu cầu
10. Số tiền yêu cầu thanh toán
11. Số tiền chi trả thực tế.

# Yêu cầu phân tích

* **Part 1**: Hãy tìm ra 3 thủ tục khám chữa bệnh có giá trị phương sai tương đối (relative variance) lớn nhất trong số tất cả các thủ tục. Đối với từng thủ tục, giá trị phương sai này được tính trên số tiền các nhà cung cấp yêu cầu được thanh toán. Tham khảo phần phụ lục phía dưới để biết cách tính giá trị phương sai tương đối này.
* **Part 2**: Đối với mỗi một loại thủ tục khám chữa bệnh sẽ có một nhà cung cấp bảo hiểm đứng đầu về số tiền yêu cầu thanh toán. Hãy tìm ra 3 nhà cung cấp bảo hiểm có nhiều lần đứng đầu nhất trong số các nhà cung cấp.
* **Part 3**: Đối với mỗi một loại thủ tục khám chữa bệnh sẽ có một vùng địa lý đứng đầu về số tiền yêu cầu thanh toán tính trung bình cho tất cả các nhà cung cấp bảo hiểm trong vùng đó. Hãy tìm ra 3 vùng địa lý có nhiều lần đứng đầu nhất trong số các vùng địa lý. Một vùng địa lý được xác định bởi một cặp thành phố + bang (trường thông tin số 5 + số 6).
* **Part 4**: Đối với mỗi một loại thủ tục khám chữa bệnh sẽ có một nhà cung cấp bảo hiểm đứng đầu về độ chênh lệch giữa số tiền yêu cầu thanh toán với số tiền được chi trả thực tế. Hãy tìm ra 3 nhà cung cấp bảo hiểm có nhiều lần đứng đầu nhất trong số các nhà cung cấp.

# Phụ lục tham khảo

Mean = X1 + X2 + … + Xn / n

Variance = (X1 – Mean)2 + (X2 – Mean)2 + … + (Xn – Mean)2 / n

Relative Variance = Variance / Mean2

= (X1 – Mean)2 + (X2 – Mean)2 + … + (Xn – Mean)2 / n \* Mean2