**BÀI TẬP THỰC HÀNH**

**MÔN HỌC: DỮ LIỆU LỚN**

**LAB 6**

**PYSPARK và JSon**

***Mục tiêu:***

* PySpark SQL
* SparkSession: kết nối giữa các views và thao tác truy vấn
* Json với SparkSession

**BÀI TẬP THỰC HÀNH**

1. **Với dữ liệu:**

* Tạo file Khoa.csv lưu thông tin table Khoa gồm: MaKhoa, TenKhoa có các records như sau:

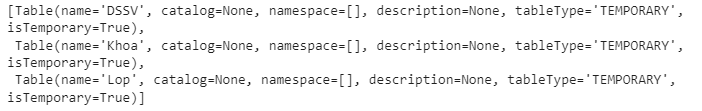
|  |  |
| --- | --- |
| MaKhoa | TenKhoa |
| CNTT | Công nghệ thông tin |
| NN | Ngoại ngữ |
| QTKD | Quản trị kinh doanh |

* Tạo file Lop.csv lưu thông tin table Lop gồm: MaLop, MaKhoa, SiSo có các records như sau:

|  |  |  |
| --- | --- | --- |
| MaLop | MaKhoa | SiSo |
| CNTT\_L1 | CNTT | 40 |
| NN\_L1 | NN | 25 |
| NN\_L2 | NN | 28 |
| QTKD\_L1 | QTKD | 39 |
| CNTT\_L2 | CNTT | 38 |
| CNTT\_L3 | CNTT | 20 |

Tiến hành thêm file DSSV.csv như dữ liệu mẫu đã gửi. Sau đó tiến hành thực hiện các công việc và truy vấn sau:

* Tiến hành gộp 3 views Khoa, Lop, DSSV vào cùng catalog.



1. Rút toàn bộ danh sách sinh viên nào thuộc khoa CNTT thành file DSSV\_CNTT.csv.
2. Liệt kê mã lớp nào của khoa CNTT chưa có sinh viên thành file DSLopChuaCoSV.csv.
3. Cho biết số lượng sinh viên của từng lớp (với lớp nào đã có sinh viên).
4. Liệt kê số lượng sinh viên còn thiếu ở mỗi lớp của từng khoa. Ghi thành file SoLuongThieu.csv.
5. Có bao nhiêu lớp đã có sinh viên?
6. Tạo danh sách sinh viên ứng với từng lớp của từng khoa. Ghi lần lượt thành file có tên theo quy định “Khoa\_MaLop\_DSSV.csv”.
7. Hiển thị đầy đủ thông tin của sinh viên nào nhỏ tuổi nhất ở khoa CNTT.
8. Tạo danh sách Birthday.csv chứa thông tin toàn bộ sinh viên nào có sinh nhật trong tháng 3.
9. Tìm xem sinh viên nào chưa có email theo quy định (quy định là mã sinh viên @gmail.com) để cập nhật lại cho đúng. Lưu thành danh sách mới mang tên DSSV\_.csv.
10. Lọc ra danh sách sinh viên nào sinh trong năm 2003 và ghi vào file DSSV\_2003.csv.
11. **Tạo các file json sau và thực hiện các truy vấn bên dưới:**

*house\_data.json:* lưu thông tin toàn bộ các căn nhà đang rao bán

{

"id": 1,

"address": "123 Main St, Anytown, USA",

"price": 250000,

"bedrooms": 3,

"bathrooms": 2,

"square\_feet": 1500,

"description": "This is a great house",

"year\_built": 2010

},

{

"id": 2,

"address": "456 Elm St, Anytown, USA",

"price": 150000,

"bedrooms": 2,

"bathrooms": 1,

"square\_feet": 1000,

"description": "This is a good house",

"year\_built": 2015

},

{

"id": 3,

"address": "789 Oak St, Anytown, USA",

"price": 350000,

"bedrooms": 4,

"bathrooms": 3,

"square\_feet": 2000,

"description": "This is a very nice house",

"year\_built": 2005

},

{

"id": 4,

"address": "321 Pine St, Anytown, USA",

"price": 200000,

"bedrooms": 2,

"bathrooms": 2,

"square\_feet": 1200,

"description": "This is a decent house",

"year\_built": 2012

},

{

"id": 5,

"address": "654 Maple St, Anytown, USA",

"price": 400000,

"bedrooms": 3,

"bathrooms": 2

},

{

"id": 6,

"address": "987 Cedar St, Anytown, USA",

"price": 180000,

"bedrooms": 1,

"bathrooms": 1,

"square\_feet": 800,

"description": "This is a small house",

"year\_built": 2018

}

*house\_deposit\_data.json*: lưu thông tin các căn nhà đã được đặt cọc mua và số tiền cọc.

{

"id": 1,

"deposit": 10000

},

{

"id": 2,

"deposit": 5000

},

{

"id": 3,

"deposit": 20000

},

{

"id": 5,

"deposit": 30000

},

{

"id": 6,

"deposit": 8000

}

*house\_loan\_data.json*: lưu thông tin các căn nhà đã được đặt cọc và số tiền vay để mua nhà.

{

"id": 1,

"loan": 50000

},

{

"id": 2,

"loan": 20000

},

{

"id": 3,

"loan": 30000

}

1. Đọc toàn bộ 3 file json trên vào cùng catalog.
2. Hiển thị toàn bộ thông tin của ngôi nhà nào đã được đặt cọc.
3. Căn nhà nào chưa được đặt cọc thì lưu thông tin gồm id, address, price ra file house\_no\_deposit.json.
4. Hiển thị toàn bộ thông tin căn nhà nào đã được đặt cọc nhưng không vay.
5. Cập nhật tăng thêm 10% tiền vay cho toàn các căn nhà đã vay. Ghi thành file house\_loan\_data\_updated.json.
6. Tính số tiền phải trả trước của các căn nhà đã vay. Biết rằng tiền trả trước = tổng tiền nhà (price) - tiền cọc (deposit) – tiền vay (loan) đã được cập nhật thêm 10% ở câu trên. Ghi thành file house\_loan\_.json.
7. Hiển thị thông tin ngôi nhà nào đã trả trước hơn 80% khi mua.
8. Tính lãi suất hằng tháng cho các căn nhà đang vay. Biết rằng lãi suất hằng tháng = 3.2% \* giá trị căn nhà + tiền vay \* 1.1 %.

**--- HẾT ---**