**Bài thực hành số 5. Sử dụng DataFrame**

**1. Giới thiệu về bộ dữ liệu:** bộ dữ liệu bán hàng gồm các tệp:

Customers.csv - chứa thông tin khách hàng gồm các trường

customerid,companyname,contactname,contacttitle,address,city,country,phone Productions.csv - chứa thông tin sản phẩm gồm các trường

productid,productname,quantityperunit,unitprice,unitsinstock

Orders.csv - chứa thông tin đơn hàng gồm các trường

orderid,customerid,orderdate

Order-details.csv - chứa thông tin chi tiết đơn hàng hàng gồm các trường orderid,productid,unitprice,quantity,discount

**2. Một số kiến thức bổ sung**

Một số hàm thao tác với dữ liệu kiểu date:

from pyspark.sql.functions import year, month, dayofmonth

**3. Hướng dẫn một số thao tác**

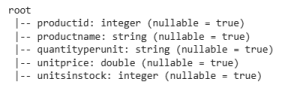
Đọc dữ liệu từ tệp:

filepath = "/content/drive/MyDrive/Northwind/"

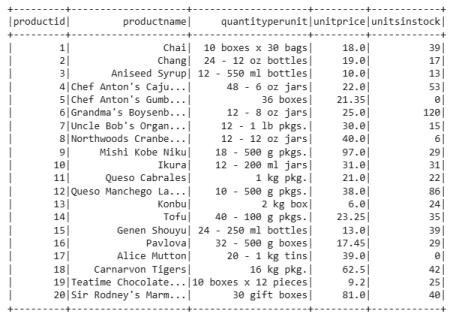
#Đọc tệp Productions.csv

df\_productions = spark.read.csv(filepath+"Productions.csv", header=True, inferSchema=True)

df\_productions.printSchema()



df\_productions.show()



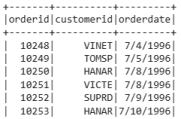
#Đọc tệp Orders.csv

df\_orders = spark.read.csv(filepath+"Orders.csv", header=True, inferSchema=True)

df\_orders.printSchema()



df\_orders.show()



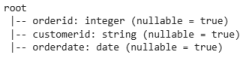
#Chuyển cột orderdate thành kiểu date

from pyspark.sql.functions import to\_date

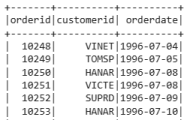
df\_orders = df\_orders.withColumn('orderdate',

to\_date(df\_orders['orderdate'], 'M/d/yyyy'))

df\_orders.printSchema()



df\_orders.show()

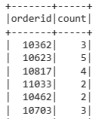


#Liệt kê những đơn hàng trước ngày 1996-07-20

df\_orders.filter(df\_orders['orderdate']<"1996-07-20").show() #Liệt kê những đơn hàng năm 1996

from pyspark.sql.functions import year

df\_orders.filter(year(df\_orders['orderdate'])==1996).show() #Đếm số mặt hàng trong từng đơn hàng

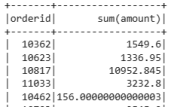
df\_orderdetails.groupBy('orderid').count().show() 

#Tính tổng tiền của từng đơn hàng

from pyspark.sql.functions import col

#Thêm cột amount tính thành tiền từng mặt hàng trong đơn hàng df\_orderdetail\_withamount = df\_orderdetails.withColumn("amount", col('unitprice')\*col('quantity')\*(1-col('discount'))) #Phân nhóm và tính tổng tiền của từng đơn hàng

df\_orderdetail\_withamount.groupBy('orderid').sum('amount').show()



#Tính tổng tiền bán hàng trong từng tháng của năm 1999

#Nối đơn hàng với chi tiết đơn hàng và lọc những đơn hàng năm 1996 df\_order\_full =

df\_orderdetail\_withamount.join(df\_orders.filter(year(col('orderdate'))==19 96), 'orderid','inner')

#Thêm cột month

df\_order\_full = df\_order\_full.withColumn('month',month(col('orderdate'))) #Phân nhóm theo tháng và tính tổng thành tiền

sum\_amount\_by\_month =

df\_order\_full.groupby('month').sum('amount').orderBy('month').collect()

print(sum\_amount\_by\_month)

[Row(month=7, sum(amount)=27861.894999999997), Row(month=8, sum(amount)=25485.275000000005), Row(month=9,

sum(amount)=26381.399999999998),…]

#Lấy danh sách tháng và tổng thành tiền từng tháng

m = [tuple(row)[0] for row in sum\_amount\_by\_month]

a = [tuple(row)[1] for row in sum\_amount\_by\_month]

print(m)

print(a)

[7, 8, 9, 10, 11, 12]

[27861.894999999997, 25485.275000000005, 26381.399999999998, 37515.72499999999, 45600.045, 45239.63000000001]

#Vẽ biểu đồ

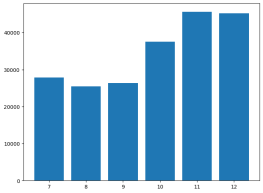
%matplotlib inline

import matplotlib.pyplot as plt

fig=plt.figure()

ax=fig.add\_axes([0,0,1,1])

ax.bar(m, a)



**4. Nội dung thực hành**

4.1 Cho biết những khách hàng ở USA.

4.2 Cho biết tên, đơn giá, số lượng tồn kho của những sản phẩm tồn kho hơn 100 sản phẩm 4.3 Cho biết những đơn hàng trong quý 2 năm 1997

4.4 Cho biết tên khách hàng (công ty) của đơn hàng có mã 10252

4.5 Đếm số mặt hàng trong từng đơn hàng

4.6 Tính số đơn hàng và tổng tiền của khách hàng có mã 10252

4.7 Cho biết số lượng đơn hàng và tổng tiền bán từng ngày trong tháng 1 năm 1997. 4.8 Tính số lượng, tổng tiền, đơn giá bán bình quân của từng mặt hàng

4.9 Cho biết tên những mặt hàng chưa bán được

4.10 Cho biết tên những mặt hàng bán nhiều nhất theo số lượng

4.11 Vẽ biểu đồ cột biểu diễn thành tiền bán của 10 mặt hàng bán nhiều nhất trong năm 1997 4.12 Vẽ biểu đồ cột so sánh số tiền bán được theo tháng của từng năm

4.13 Vẽ biểu đồ đường biểu diễn số lượng bán được từng tháng/năm của mặt hàng có nhiều đơn hàng nhất

4.14 Cho biết những cặp khách hàng mua giống nhau từ 5 mặt hàng trở lên, đó là những mặt hàng nào?

4.15 Cho biết trong những cặp khách hàng (c1, c2) mua giống nhau hơn 5 mặt hàng thì những mặt hàng nào khách hàng c1 mua mà khách hàng c2 chưa mua.

-------------