|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC CÔNG NGHỆ GTVT  KHOA CÔNG NGHỆ THÔNG TIN  BỘ MÔN HỆ THỐNG THÔNG TIN | CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM  Độc lập - Tự do - Hạnh phúc  *Hà Nội, ngày tháng năm 2021* |

**ĐỀ TÀI BÀI TẬP LỚN**

**Tên học phần:** BigData

Mã học phần: DC2HT17 Số tín chỉ: 3

Hình thức đánh giá: Bài tập lớn

Thời gian thực hiện: 2 tuần Thời gian trả lời: 30 phút

**Ngành/chuyên ngành:** Công nghệ thông tin

**Trình độ đào tạo:** Đại học

**I. NỘI DUNG ĐỀ TÀI**

**Đề tài số 1.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày về định nghĩa và các đặc trưng của BigData?

**Câu 2.** Trình bày các thành phần của Spark (khái niệm, mô tả các thành phần)?

**Câu 3.** Cho tập dữ liệu về danh sách các cầu thủ bóng đá Fifa 2020 có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau:

* 1. Đếm số lượng cầu thủ?
  2. Tính số lượng câu lạc bộ trong danh sách?
  3. Tính độ tuổi trung bình của các cầu thủ? Cầu thủ có tuổi lớn nhất và nhỏ nhất là bao nhiêu?
  4. Tính chiều cao trung bình của các cầu thủ? Cầu thủ có chiều cao lớn nhất và nhỏ nhất là bao nhiêu?
  5. Liệt kê số lượng cầu thủ tương ứng với từng Quốc Gia?
  6. Tìm cầu thủ người Brazil có giá trị cao nhất?
  7. Tính số lượng cầu thủ Brazil trong các câu lạc bộ?
  8. Hiển thị danh sách top 5 các cầu thủ ghi nhiều bàn thắng nhất (Lưu ý trường hợp số lượng bàn thắng ghi được của các cầu thủ bằng nhau)
  9. Tạo một cột mới nếu: Age < 25 điền là Y, 25 <= Age <= 30 điền là T, 30 < Age điền là O
  10. Tạo một cột mới và điền thông tin là 2 ký tự đầu của cột Nationality tương ứng.



**Câu 4.** Sử dụng dữ liệu giải đấu Premier League từ năm 2015 đến 2020 (file đính kèm) để thực hành streaming kết quả 3 đội bóng có thứ hạng cao nhất theo năm.

**Đề tài số 2.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Nêu khái niệm và so sánh RRD, DataFrame và DataSet?

**Câu 2.** So sánh chế độ TemporaryView và Global Temporary View trong SparkSQL?

**Câu 3.** Cho tập dữ liệu về mùa giải Premier League năm 2020 - 2021 có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau:

3.1. Có bao nhiêu đội tham gia giải bóng?

3.2. Có tổng số bao nhiêu trận đấu thực hiện trong mùa giải?

3.3. Có bao nhiêu trận có tỷ số hòa và tỷ số chênh lệch trong mùa giải?

3.4. Thời gian diễn ra trận đấu muộn nhất trong ngày là các trận đấu nào?

3.5. Đội Arsenal dành được bao nhiêu trận thắng trong mùa giải?

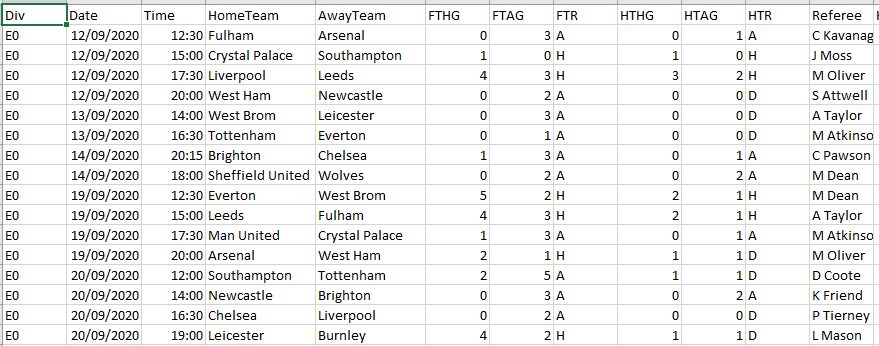
3.6. Đội Arsenal dành được bao nhiêu điểm trong toàn bộ mùa giải?

3.7. Đội bóng nào vô địch mùa giải?

3.8. Xếp thứ hạng của các đội bóng mùa giải dựa theo các tiêu chí điểm số, kết quả đối đầu, số bàn thắng ghi được?

3.9. Tạo một cột mới điền thông tin 3 chữ cái đầu tiên từ cột HomeTeam tương ứng

3.10. Tạo một cột mới nếu: Time từ 00:00 đến trước 12:00 điền là “Time1”, từ 12:00 đến trước 18h00 điền là “Time2”, còn lại điền là “Time3”



**Câu 4.** Sử dụng số liệu thống kê thu được từ giá ngoại tệ trong các năm 2018, 2019, 2020 (file đính kèm) để thực hành streaming hiển thị kết quả 3 ngoại tệ có độ chênh lệch trung bình giữa các ngày liền kề trong năm cao nhất.

**Đề tài số 3.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày về khái niệm RDD và các cách thức tạo ra RDD?

**Câu 2.** Trình bày về khái niệm và đặc điểm của Apache Spark?

**Câu 3.** Cho tập dữ liệu về danh sách hành khách trên chuyến tàu có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau:

3.1. Tổng cộng có bao nhiêu hành khách trên chuyến tàu?

3.2. Trong số các hành khách bao nhiêu người là Nam, bao nhiêu người là Nữ?

3.3. Số lượng hành khách là trẻ em (Tuổi < 13) ?

3.4. Tính số lượng hành khách là trẻ em nam và trẻ em nữ?

3.5. Tuổi lớn nhất và nhỏ nhất của tất cả hành khách trên chuyến tàu là bao nhiêu?

3.6. Có bao nhiêu hành khách có cùng tuổi lớn nhất, có cùng tuổi nhỏ nhất?

3.7. Tính tổng số tiền vé thu được? Giá vé trung bình ?

3.8. Liệt kê danh sách hành khách theo điều kiện giới tính là nam, mua vé có giá lớn hơn hoặc bằng mức giá trung bình, có tuổi lớn hơn hoặc bằng độ tuổi trung bình

3.9. Tạo một cột mới nếu: Pclass = 1 điền là “VIP”, Pclass = 2 điền là “Luxury”, còn lại điền là “Normal”

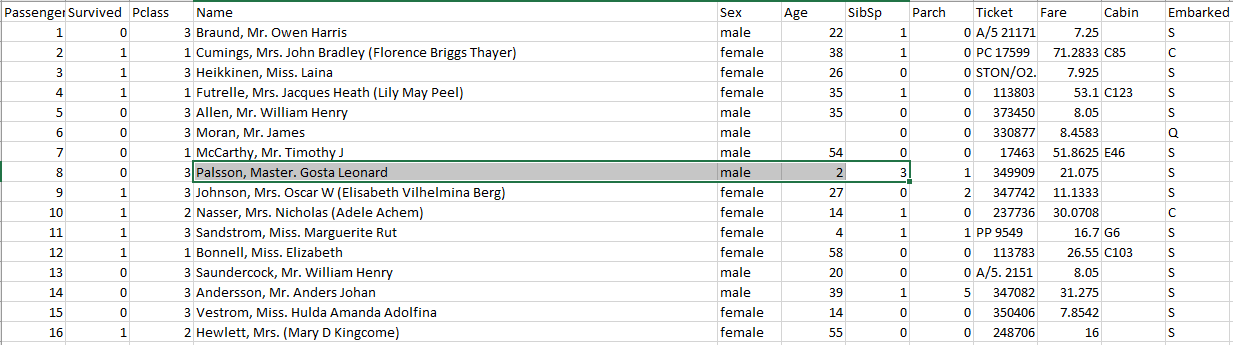
3.10. Tạo một cột mới nếu thông tin cột Sex là female và Sibsp là 1 điền là “wife”, nếu Sex là male điền là “husband”, còn lại điền là “Other”

**Trong đó:**

Sibsp: Số anh / chị / em / vợ / chồng trên tàu)

Parch: Số cha mẹ / con cái trên tàu

Fare: Giá vé hành khách

****

**Câu 4.** Sử dụng dữ liệu giải đấu Championship từ năm 2015 đến 2020 (file đính kèm) để thực hành streaming kết quả tổng số bàn thắng và hiệu suất ghi bàn của đội Arsenal theo năm.

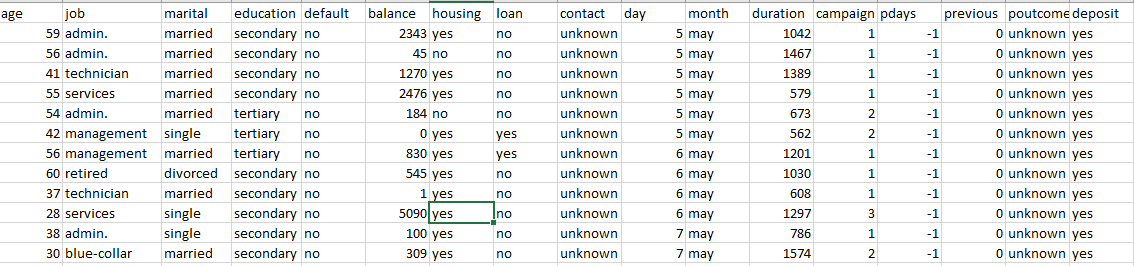
**Đề tài số 4.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** So sánh Spark và Map Reduce?

**Câu 2.** Trình bày về DStream? (Khái niệm, cách thức hoạt động)

**Câu 3.** Cho tập dữ liệu về các phỏng vấn một số khách hàng của ngân hàng có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau:

1. Có bao nhiêu loại nghề nghiệp trong thống kê?
2. Liệt kê danh sách và số lượng các loại công việc trong thống kê?
3. Có bao nhiêu người đã kết hôn và độc thân trong thống kê?
4. Có bao nhiêu loại tình trạng giáo dục trong thống kê và tương ứng số lượng người trong mỗi loại?
5. Tính độ tuổi trung bình của các khách hàng trong danh sách?
6. Khách hàng có độ tuổi lớn nhất, nhỏ nhất là bao nhiêu?
7. Liệt kê danh sách các khách hàng còn độc thân, đồng thời có khoản vay ngân hàng và được giáo dục ở mức “secondary”
8. Liệt kê danh sách các khách hàng có tuổi lớn hơn độ tuổi trung bình, đồng thời không có khoản nợ ngân hàng và chưa vay mua nhà?
9. Tạo một cột mới nếu: Age < 25 điền là Y, 25 <= Age <= 30 điền là T, 30 < Age điền là O
10. Tạo một cột mới nếu giá trị tại cột “Housing” là Yes và cột “Loan” là Yes điền là “Type1”, cột “Housing” là Yes và cột “Loan” là No điền là “Type2”, cột “Housing” là No và cột “Loan” là No điền là “Type3”, còn lại điền là “Type4”



**Câu 4.** Sử dụng số liệu thống kê số lần click vào các đường dẫn phim trên NetFlix trong năm 2017, 2018, 2019 (file đính kèm) để thực hành streaming kết quả 3 bộ phim được quan tâm nhất theo năm.

Trong đó: datetime – thời gian click vào tên phim

**Đề tài số 5.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày mô hình map – reduce trên Hadoop và giải thích?

**Câu 2.** Nêu các khái niệm, đặc trưng của HDFS?

**Câu 3.** Sử dụng 2 file seedsource1.csv và treeSpec.csv

3.1. Đọc dữ liệu từ file seedsource1.csv và TreeSpec.csv vào dataframe hoặc RDD

3.2. Có bao nhiêu AcceptCode? Có bao nhiêu TreeSpecCode

3.3. Tìm những AcceptCode có TreeSpec bắt đầu bằng chữ “Eucalyptus”

3.4. Tìm 10 diện tích lớn nhất xếp từ cao xuống thấp

3.5. Có bao nhiêu AcceptedCode có diện tích > 3

3.6. Có bao nhiêu Xã(communeCode) trong một Tỉnh(ProvinceCode)

3.7. Tính trung bình diện tích trong một xã

3.8. Tạo một cột mới có giá trị = diện tích \* 10

3.9. Tạo một cột mới nếu: AcceptCode có kí tự là SD thì điền giá trị SD. Nếu AcceptCode có kí tự SM thì điền SM, SC điền SC…. (2 kí tự đầu của AcceptCode)

3.10. Tạo một cột mới: Nếu Area <= 1 thì điền small, 1< m < 3 : Normal, >= 3: large

**Câu 4.** Sử dụng dữ liệu giải đấu Laliga từ năm 2018 đến 2021 (file đính kém) để thực hành stream tổng hợp kết quả thẻ vàng và đỏ nhận được trong một mùa giải của đội Real Madrid

**Đề tài số 6.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Nêu các đặc điểm của MapReduce Hadoop – so sánh với spark?

**Câu 2.** Thế nào là tính toán phân tán, tính toán song song? Nếu một vài ví dụ thực tiễn sử dụng tính toán phân tán và tính toán song song?

**Câu 3.** Sử dụng file census\_1000.csv (Điều tra dân số)

3.1. Đọc dữ liệu từ file census\_1000.csv vào dataframe hoặc RDD./

3.2. Có bao nhiêu nhóm tuổi được đề cập trong dữ liệu? Liệt kê các nhóm tuổi./

3.3. Tìm những nhóm tuổi của bằng cấp là “Bachelors”./

3.4. Thay đổi định dạng dữ liệu của cột education-num, age thành Integer

3.5. Tìm nhóm tuổi có số lượng “Education-num” lớn nhất

3.6. Tìm nhóm tuổi là nữ chưa kết hôn “never-married” là người da trắng

3.7. Có bao nhiêu nhóm tuổi có người da trắng? Liệt kê các nhóm tuổi là người da trắng

3.8. Tìm những nhóm tuổi có marital-status bắt đầu bằng “Married”

3.9. Xoay giá trị trong cột gender thành nhiều cột với mỗi cột là một giá trị theo yêu cầu sau: dữ liệu trong cột là tổng số education-num của nhóm dữ liệu nghề nghiệp (Occupitation)

|  |  |  |
| --- | --- | --- |
| Occupitation | Male | Femail |
| Adm-clerical | 10 | 50 |
| Other-service | Null | 100 |

3.10. Tạo một cột mới có giá trị là: Nếu gender = Male thì điền giá trị là M, Nếu gende = Female thì điền giá trị là

**Câu 4.** Sử dụng dữ liệu giải đấu laliga từ năm 2018 đến 2021 (file đính kèm) để thực hành stream tổng hợp tìm ra đội có số trận thua trên sân nhà nhiều nhất mùa giải, cùng với đó là tổng số bàn thua của họ.

FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win)

FTHG and HG = Full Time Home Team Goals,

FTAG and AG = Full Time Away Team Goals)

**Đề tài số 7.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày Các thành phần của YARN trong Hadoop?

**Câu 2.** Trình bày ưu điểm, nhược điểm của Hadoop – MapReduce?

**Câu 3.** Sử dụng file king-county-house-sales.csv.bz (Dữ liệu bất động sản bán)

3.1. Đọc dữ liệu từ file vào dataframe hoặc RDD

3.2. Có bao nhiêu nhà đang được bán

3.3. Có bao nhiêu vùng được liệt kê trong dữ liệu (dựa vào zipcode)? Liệt kê các vùng

3.4. Thay đổi định dạng của cột price, sqft\_living, floors sang Integer

3.5. Tính giá trung bình nhà theo vùng

3.6. Tìm những nhà có 3 tầng được xây dựng sau năm 1970 (yr\_built)

3.7. Tìm những nhà <= 3 phòng ngủ có giá trị <= 250000$. Liệt kê tổng số nhà theo giá từng loại tầng và giá trung bình của nó

3.8. Tìm những nhà được đăng bán vào năm 2015

3.9. Xoay giá trị trong cột floors thành nhiều cột. Với mỗi cột chứa các giá trị là số lượng căn nhà có giá trị floors tương ứng - nhóm theo số lượng phòng ngủ:

Ví dụ: Căn 3 phòng ngủ - có số lượng 1 tầng là 30 – 2 tầng: 15: 3 tầng:20

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bedrooms | 1 | 2 | 3 | … |
| 3 | 10 | Null | 20 | … |
| 2 | 3 | 3 | null |  |

3.10. Tạo một cột mới với tên cột tuỳ chọn với giá trị là: Nếu diện tích (sqft\_living) <= 1000 thì điền giá trị “small”, nếu 1000 < x < 2000 thì điền giá trị “medium”, còn lại điền “large”

**Câu 4.** Sử dụng dữ liệu giải đấu laliga từ năm 2018 đến 2021 (file đính kém) để thực hành stream tổng hợp tìm ra 3 đội có số trận thắng trên sân khách nhiều nhất mùa giải, cùng với đó là số cú sút trúng đích trong những trận thắng đó (sử dụng dữ liệu cột:

AST: Away team shots on target,

FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win).

**Đề tài số 8.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Lợi ích của BigData? Trình bày lợi ích của bigdata ứng dụng trong giáo dục?

**Câu 2.** Độ đa dạng dữ liệu là gì? Theo bạn chúng ta cần dữ liệu lớn hay dữ liệu có giá trị? Các công ty cần làm gì để có thể tận dụng lợi thế của dữ liệu lớn?

**Câu 3.** Sử dụng file E1.csv (Dữ liệu mùa giải championship 2011)

3.1. Đọc dữ liệu vào dataframe hoặc RDD

3.2. Thay đổi định dạng cột (FTHG, FTAG, HS,AS) thành Integer

3.3. Có bao nhiêu đội đá trong mùa giải, liệt kê các đội

3.4. Tìm số trận có kết quả hoà

(FTR= Full time result: H = homewin, D = Draw, A= Awaywin)

3.5. Tìm tổng số bàn thắng các đội đá trên sân nhà ghi được

(FTHG = Fulltime hometeam goal – FTAG: Fulltime AwayGoal)

3.6. Tìm những trận có tổng số bàn thắng > 3

3.7. Tìm những trận của Burnley được thi đấu trên sân nhà và có số bàn thắng >=3 (Tính cả của đội khách)

3.8. Tìm những trận mà Reading thua (Không được sử dụng cột FTR)

3.9.

, với mỗi cột chứa số lượng FTR tương ứng. nhóm theo HomeTeam

|  |  |  |  |
| --- | --- | --- | --- |
| HomeTeam | A | D | H |
| Hull | 1 | Null | 1 |
| Derby | Null | 1 | 1 |
| .. | .. | .. | .. |

3.10. Tạo một cột mới với tên cột tuỳ chọn: Nếu tổng số bàn thắng 2 đội ghi được trong trận <2 thì điền “well” , nếu số bàn thắng 2 < x < 4 thì điền “very good”, nếu số bàn thắng >= 4 thì điền “amazing”.

**Câu 4.** Sử dụng dữ liệu giải đấu laliga từ năm 2018 đến 2021 (file đính kém) để thực hành stream tổng hợp tìm ra các trận đấu được kết thúc trong một hiệp (nghĩa là hiệp 2 không có bàn được ghi). Và ghi ra số bàn thắng ghi được trong trận đấu đó

HTHG = Half Time Home Team Goals

HTAG = Half Time Away Team Goals

FTHG and HG = Full Time Home Team Goals

FTAG and AG = Full Time Away Team Goals

**Đề tài số 9.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày các thành phần của Hadoop (Khái niệm, đặc điểm)?

**Câu 2.** Trình bày về DStream Window?

**Câu 3.** Cho tập dữ liệu về danh sách 100 Tiktoker hàng đầu tại Pakistan có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau:

3.1. Tiktoker có thời gian tham gia sớm nhất?

3.2. Tiktoker có số lượng Fans nhiều nhất?

3.3. Liệt kê các năm gia nhập của các Tiktoker?

3.4. Liệt kê danh sách 3 Tiktoker hàng đầu của Pakistan theo số lượng Fan?

3.5. Liệt kê danh sách 3 Tiktoker hàng đầu của Pakistan theo Tổng số lượng Fan và số lượng View? Trong đó: số lượng Fan nhân hệ số 10

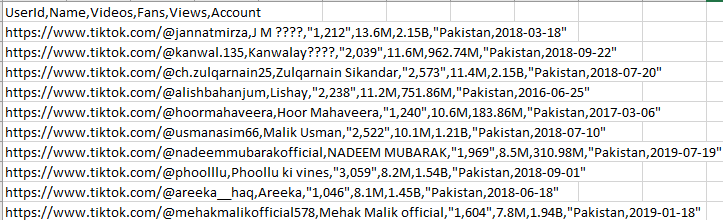
3.6. Liệt kê tỷ lệ giữa số lượng Fan và số lượng View của 3 Tiktoker có số lượng Fan thấp nhất?

3.7. Liệt kê tỷ lệ giữa số lượng Video và số lượng View của 3 Tiktoker có số lượng View cao nhất?

3.8. Tính số lượng View trung bình của tất cả các Tiktoker?

3.9. Tạo một cột mới nếu số lượng Video < 1000 điền là “Type1”, 1000 <= Số Video < 2000 điền là “Type2”, còn lại điền là “Type3”

3.10. Tạo một cột mới điền thông tin tỷ lệ giữa số lượng View và Fan, Video theo công thức Tỷ lệ = Số lượng View/(Số lượng Fan \* Số lượng Video)

****

**Câu 4.** Sử dụng số liệu thống kê các chuyến bay trong tháng 1,2,3 năm 2015 (file đính kèm) để thực hành streaming kết quả 3 hãng hàng không bay nhiều nhất. Trong đó cột Airline Identifier là mã định danh của các hãng hàng không.

**Đề tài số 10.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày về Hadoop Cluster (Khái niệm, các thành phần)?

**Câu 2.** Trình bày về các Transformation thường được sử dụng trên DStream?

**Câu 3.** Cho tập dữ liệu về danh sách 500 người giàu nhất thế giới năm 2021 có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau:

3.1. Hiển thị thông tin Tên người giàu nhất thế giới?

3.2. Hiển thị thông tin tên người có thứ hạng 100 trong danh sách?

3.3. Liệt kê danh sách 3 tỷ phú có sự thay đổi về tài sản tăng lớn nhất theo danh sách?

3.4. Liệt kê danh sách 3 tỷ phú giàu nhất nước Mỹ?

3.5. Liệt kê danh sách 3 quốc gia có số lượng tỷ phú nhiều nhất? (Lưu ý trường hợp các quốc gia có số lượng tỷ phú bằng nhau)

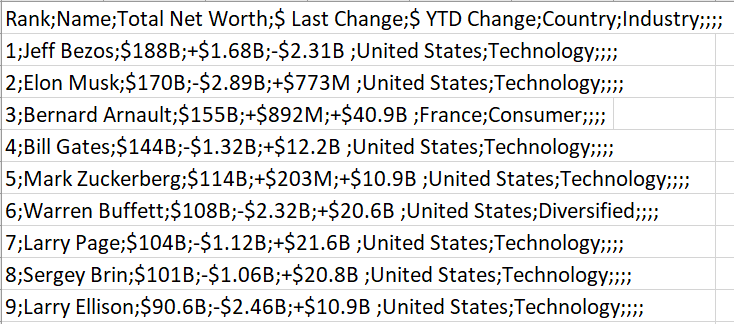
3.6. Liệt kê danh sách 3 ngành nghề có số lượng tỷ phú nhiều nhất? (Lưu ý trường hợp các ngành nghề có số lượng tỷ phú bằng nhau)

3.7. Tính tài sản trung bình của 500 tỷ phú?

3.8. Tính số lượng tỷ phú có tài sản giảm trong năm 2021?

3.9. Tạo một cột mới nếu tài sản >= 60 tỷ USD điền là “A”, 60 < tài sản < 30 điền là B, còn lại điền là “C”

3.10. Tạo một cột mới điền thông tin tỷ lệ thay đổi theo công thức Tỷ lệ = YTD Change/Total Net World

****

**Câu 4.** Sử dụng số liệu thống kê lưu lượng giao thông tại một thành phố năm 2020, 2021 (file đính kèm) để thực hành streaming kết quả 3 tuyến đường được thống kê dữ liệu nhiều nhất, trong đó cột street\_id lưu trữ ID của mỗi tuyến đường.

**Đề tài số 11.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày các bước quan trọng để triển khai nền tảng big data?

**Câu 2.** Cơ sở dữ liệu phi quan hệ là gì? So sánh cơ sở dữ liệu quan hệ và cơ sở dữ liệu phi quan hệ?

**Câu 3.** Sử dụng file phim.thongtin.csv và phim.binhchon.csv

3.1. Cho biết schema của bộ dữ liệu. Hiển thị số bản ghi của bộ dữ liệu.

3.2. Liệt kê 20 phim có tiêu đề gốc là “Harry Potter”.

3.3. Đểm số lượng phim theo mỗi thể loại và sắp xếp theo thứ tự giảm dần.

3.4. Tính độ dài trung bình của các phim “Star wars” theo số phút.

3.5. Hiển thị các thông tin về tiêu đề gốc, năm bắt đầu, năm kết thúc, điểm bình chọn trung bình, số lượt bình chọn của 10 phim có điểm bình chọn cao nhất.

3.6. Hiển thị các thông tin về tiêu đề gốc, năm bắt đầu, năm kết thúc, điểm bình chọn trung bình, số lượt bình chọn của 7 phim có thể loại là “tvSeries” và số lượt bình chọn hơn 150000 lượt và có điểm bình chọn cao nhất.

3.7. Đếm số phim từng năm có hơn 100000 lượt bình chọn và năm bắt đầu > 2000.

3.8. Liệt kê phim hay nhất theo từng thể loại.

3.9. Tạo một cột mới: nếu điểm bình chọn >= 8.5 thì điền A, từ 7 đến 8.4 thì điền B, từ 5.5 đến 6.9 điền C, từ 4 đến 5.4 điền D, còn lại điền F.

3.10. Tạo một cột mới “time” được định nghĩa bằng công thức năm kết thúc trừ năm bắt đầu, sau đó hiển thị danh sách các phim có time lớn nhất theo từng thể loại.

**Câu 4.** Sử dụng dữ liệu về phim từ trong thập kỷ 2010 (file đính kèm) để thực hành stream tổng hợp tìm ra phim có thời lượng nhiều nhất.

**Đề tài số 12.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày các loại cơ sở dữ liệu phi quan hệ. Mỗi loại lấy 1 ví dụ.

**Câu 2.** Dữ liệu được biểu diễn như thế nào trong Spark.

**Câu 3.** Sử dụng file player\_football\_2021.csv

3.1. Cho biết schema của bộ dữ liệu. Hiển thị số bản ghi của bộ dữ liệu.

3.2. Hiển thị danh sách cầu thủ có chỉ số trung bình tốt nhất tại mỗi vị trí.

3.3. Liệt kê danh sách cầu thủ đá ở vị trí hậu vệ phải và có quốc tịch Tây Ban Nha.

3.4. Liệt kê danh sách 5 cầu thủ Brazil đá ở vị trí tiền đạo trung tâm có cân nặng lớn nhất.

3.5. Đếm số lượng cầu thủ theo từng độ tuổi.

3.6. Liệt kê danh sách 10 câu lạc bộ có quỹ lương lớn nhất.

3.7. Liệt kê danh sách 7 câu lạc bộ cho mượn nhiều cầu thủ nhất.

3.8. Liệt kê danh sách 15 đội bóng có tổng phí chuyển nhượng các cầu thủ đang ở năm cuối hợp đồng cao nhất.

3.9. Tạo một cột mới: nếu chỉ số tổng thể >= 85 thì điền excellent, từ 70 đến 84 thì điền good, từ 55 đến 69 điền normal, từ 40 đến 54 điền weak, còn lại điền unrank.

3.10. Tạo một cột mới: Tính giá trị thực của cầu thủ bằng giá trị cầu thủ cộng với lương nhân với số tuần còn lại trong hợp đồng.

**Câu 4.** Sử dụng dữ liệu về cầu thủ năm 2021 (file đính kèm) để thực hành stream tổng hợp tìm ra cầu thủ nhận lương cao nhất thế giới.

**Đề tài số 13.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Trình bày sự khác biệt giữa cơ sở dữ liệu quan hệ và Hadoop File System?

**Câu 2.** Bộ nhớ đệm phân tán trong Hadoop là gì? Lợi ích của nó là gì?

**Câu 3.** Sử dụng file baocao-toipham.csv, tramcanhsat.csv

3.1. Cho biết schema của bộ dữ liệu, hiển thị số bản ghi của bộ dữ liệu.

3.2. Có bao nhiêu trạm cảnh sát.

3.3. Cho biết Id quận, tên quận, địa chỉ và mã zip cho đồn cảnh sát có id=7.

3.4. Liệt kê 10 loại hình thức tội phạm xuất hiện nhiều nhất theo thứ tự giảm dần.

3.5. Cho biết tỷ lệ tội phạm đã bị bắt.

3.6. Liệt kê 3 địa điểm có số lần xuất hiện tội phạm nhiều nhất.

3.7. Hoạt động phi tội phạm (non-criminal) xuất hiện nhiều nhất là gì?

3.8. Tìm ngày có nhiều tội phạm được báo cáo nhất?

3.9. Thêm cột “Day of week” cho biết vụ việc diễn ra vào ngày thứ mấy trong tuần.

3.10. Thêm cột “District name” cho biết tên của quận nơi diễn ra vụ việc.

**Câu 4.**

Sử dụng dữ liệu báo cáo về tội phạm (file đính kèm) để thực hành streaming các vụ phạm tội diễn ra tại các địa điểm Street, Residence, Apartment.

**Đề tài số 14.** Thực hiện các nội dung yêu cầu sau

**Câu 1.** Tại sao MapReduce xử lý dữ liệu chậm hơn các nền tảng xử lý dữ liệu khác?

**Câu 2.** Trình bày hạn chế của Apache Spark

**Câu 3.** Sử dụng file vdv\_olympics, qg\_noc.csv

3.1. Cho biết schema của bộ dữ liệu, hiển thị số bản ghi của bộ dữ liệu.

3.2. Đưa ra thông tin về các vận động viên Việt Nam tham gia Olympic trong thập kỷ 2000.

3.3. Đưa ra thông tin về 7 quốc gia có số lượng vận đông viên đông nhất tham gia thế vận hội 2016

3.4. Cho biết số lượng vận động viên tham dự 5 kỳ thế vận hội gần nhất

3.5. Cho biết số lượng vận động viên nam tại mỗi kỳ thế vận hội trong thế kỷ 21

+----+------------------+

|year|number\_of\_athletes|

+----+------------------+

|2000| XXXXX|

|2002| XXXX|

3.6. Cho biết tổng số vận động viên tại mỗi kỳ thế vận hội trong thập kỷ 1990 của Nga.

+-----------+------+------------------+

| country|season|number\_of\_athletes|

+-----------+------+------------------+

| Australia|Summer| XXX|

3.7. Cho biết chiều cao tối thiểu, trung bình và tối đa của mỗi quốc gia tham gia thế vận hội mùa đông và sắp xếp theo thứ tự giảm dần.

+--------------------+----------+------------------+----------+

| country|min\_height| avg\_height|max\_height|

+--------------------+----------+------------------+----------+

3.8. Đưa ra tên các quốc gia không có thông tin về quốc gia trong qg\_noc (ví dụ: Yugoslavia-Nam Tư không còn tồn tại nữa).

+--------------------+---+

| team|noc|

+--------------------+---+

| Almaz|URS|

| Australasia|ANZ|

3.9. Liệt kê danh sách 10 quốc gia giành nhiều huy chương vàng nhất thế vận hội mùa hè 2012.

3.10. Liệt kê 10 vận động viên có giành nhiều huy chương vàng nhất thế kỷ 20.

**Câu 4.**

Sử dụng dữ liệu các vận động viên tại thế vận hội 2016 (file đính kèm) để thực hành streaming các vận động viên giành huy chương vàng.