Assignment

Question 1 :

Given retail dataset

hdfs path : /public/trendytech/retail_db

we need to consider 3 datasets

orders
========
order_id,order_date,order_customer_id,order_status

customers
==========
customer_id,customer_fname,customer_lname,customer_email,customer_password,customer_street,customer_city,customer_state,customer_zipcode

order_items
==============
order_item_id,order_id,order_item_product_id,order_item_quantity,order_item_subtotal,order_item_product_price

Note - one customer can have multiple orders in the orders dataset.
one order can have multiple order_items in the order_items table

1. we need to find top 10 customers who have spent the most amount (premium customers)
2. top 10 product id's with most quantities sold
3. how many customers are from Caguas city
4. Top 3 states with maximum customers
5. how many customers have spent more than $1000 in total
6. which state has most number of orders in CLOSED status
7. how many customers are active (active customers are the one's who placed atleast one order)
8. What is the revenue generated by each state in sorted order.

you need to perform all of the above using RDD's in Apache Spark.

Try using any optimizations like broadcast join or caching if possible.

Question 2 :

Consider the Covid19 Dataset

hdfs path - /public/trendytech/covid19

We need to consider the 2 datasets

Cases
=====

date,state,positive,negative,pending,hospitalizedCurrently,hospitalizedCumulative,inIcuCurrently,inIcuCumulative,onVentilatorCurrently,onVentilatorCumulative,recovered,dataQualityGrade,lastUpdateEt,dateModified,checkTimeEt,death,hospitalized,dateChecked,totalTestsViral,positiveTestsViral,negativeTestsViral,positiveCasesViral,deathConfirmed,deathProbable,fips,positiveIncrease,negativeIncrease,total,totalTestResults,totalTestResultsIncrease,posNeg,deathIncrease,hospitalizedIncrease,hash,commercialScore,negativeRegularScore,negativeScore,positiveScore,score,grade

States
=====

state,notes,covid19Site,covid19SiteSecondary,covid19SiteTertiary,twitter,covid19SiteOld,name,fips,pui,pum

1. Find the top 10 states with the highest no.of positive cases.
2. Find the total count of people in ICU currently.
3. Find the top 15 States having maximum no.of recovery
4. Find the top 3 States having least no.of deaths
5. Find the total number of people hospitalized currently.
6. List the twitter handle and fips code for the top 15 states with the highest number of total cases.

Question 3 :

Given Trendytech Students Google Reviews

hdfs path : /public/trendytech/reviews/trendytech-student-reviews.csv

You need to find the top 20 words

This is done with a thought process that the final answer would represent the central theme of the reviews, like what most of the students are saying.

But there would be boring words like "the", "is", "are", "a" etc.. which do not have any significance.

Your solution should not consider such boring words, so you need to remove the boring words and give the top 20 keywords which come the most of the time.

The list of boring words is kept in a file called boringwords.txt in your edge/gateway node:
/data/trendytech/boringwords.txt

hint: try to broadcast the boring words for better performance.

you need to solve this using Spark RDD

TRENDYTECH
UPLIFT YOUR CAREER!