Guía 2: Spark

1- Se tiene un RDD con el registro de notas de los alumnos de la forma (padrón, materia, nota, fecha). Se pide resolver utilizando PySpark:

A. Cuántos alumnos aprobaron al menos 1 materia en los últimos 2 años.
B. Un RDD conteniendo el promedio de notas de cada alumno de la forma (padrón, promedio).
C. El nombre y apellido del alumno con mejor promedio. Para esto puede utilizarse un segundo RDD alumnos con registros (padron, nombre y apellido).

2- Se tiene un RDD registros de ventas de producto con la forma (fecha de venta, código de producto, precio de venta) y en otro RDD detalle de los productos con (código de producto, descripción del producto, categoría). Se pide resolver utilizando PySpark:

A. Cuál es el producto más vendido.
B. Cuál es la categoría de productos más vendida.
C. Cuál es el top5 de productos más vendidos generando un RDD con (código de producto, descripción, cantidad de ventas)
D. Cuál es el producto que registró mayor aumento de precio en el último año, tomando para este análisis solo los productos que cuenten con al menos 50 ventas en el último año.
E. Idem anterior, pero calculando la categoría de productos que registró mayor variación de precios en el último año.

3- Se tiene un RDD con información de vuelos programados con la forma (número de vuelo, código de aerolínea, código de aeropuerto de salida, código de aeropuerto de llegada, fecha de salida AAAAMMDD, hora de salida HH:MM, fecha de llegada AAAAMMDD, hora de llegada HH:MM). A su vez, se cuenta con el registro actualizado del estado de los vuelos que fueron ocurriendo, con la forma (número de vuelo, aerolínea, fecha de salida AAAAMMDD, hora de salida HH:MM, fecha de llegada AAAAMMDD, hora de llegada HH:MM, estado). En base al estado, podría contar con algún dato en blanco, por ejemplo si el vuelo fue cancelado no tendrá información de fechas y horas, si el vuelo se encuentra aún en curso, no contendrá información de la llegada. Se pide resolver utilizando PySpark:

A. Cuál es el aeropuerto con mayor tránsito.
B. Cuál es la aerolínea con mayor cantidad de vuelos.
C. Cuál es la aerolínea con mayor cantidad de cancelaciones.
D. Cuál es el vuelo (numero de vuelo + fecha) con mayor retraso en el horario de salida.
E. Cuál es el vuelo (numero de vuelo + fecha) con mayor retraso en el horario de llegada.
F. Cuál es la aerolínea más puntual.

G. Cuál es el aeropuerto que registra mayor desviación con respecto a los horarios coordinados.

4- Se tiene un RDD con las coordenadas de rectángulos de la forma (x1,x2,y1,y2). Se pide programar en PySpark un programa que encuentre el rectángulo de superficie mínima que contiene al punto (w,z)

5- Se tiene un RDD con libros en donde cada registro es un texto. Se pide obtener todos los anagramas de mas de 7 letras que puedan encontrarse. El formato de salida debe ser una lista de listas en donde cada lista tiene un conjunto de palabras que son anagramas. Ejemplo: [[discounter,introduces,reductions],[percussion,supersonic]...]

6- UBER almacena en un cluster todos los datos sobre el movimiento y viajes de todos sus vehículos. Existe un proceso que nos devuelve un RDD llamado trip_summary con los siguientes campos: (driver_id, car_id, trip_id, customer_id, date (YYYYMMDD), distance_traveled), Programar usando Py Spark un programa que nos indique cual fue el conductor con mayor promedio de distancia recorrida por viaje para Abril de 2016.

7- Una red social almacena el contenido de los chats entre sus usuarios en un RDD que tiene registros con el siguiente formato: (chat_id, user_id, nickname, text). Queremos saber cuál es el usuario (user_id) que mas preguntas hace en sus chats, contabilizamos una pregunta por cada caracter “?” que aparezca en el campo text. Programar en Spark un programa que identifique al usuario preguntón.

8- Contamos con un cluster que tiene 4 computadoras. Queremos aprovechar el paralelismo del cluster para calcular los números primos entre 2 y 20.000.000. Para esto usaremos el conocido algoritmo de la criba de Eratóstenes. Por ejemplo si empezamos con una lista de tipo (2,3,4,5,6,7,8...) en un primer paso eliminamos todos los que son mayores a 2 y divisibles por 2 y nos queda (2,3,5,7,9,11,13…) luego eliminamos todos los mayores a 3 divisibles por 3 y nos queda (2,3,5,7,11,13….etc) luego todos los divisibles por 5 y así sucesivamente. El resultado final es una lista de números que son primos. Programar este programa usando PySpark.

9- Se cuenta con un RDD con información sobre patentamientos de autos con la siguiente información (patente, marca, modelo, versión, tipo_vehiculo, provincia, fecha), donde tipo_vehiculo indica si la unidad patentada es auto, pickup, camión o moto. Se pide generar un programa en pySpark que indique la marca y modelo del auto más patentado por tipo de vehículo en la provincia de Buenos Aires en el mes de Abril de 2017.

10- A partir de la plataforma online (e-shop) de los países en los que opera, Nintendo tiene información de ventas de videojuegos diarias digitales por país en el siguiente RDD: (id_videojuego, codigo_pais, fecha, visitas_diarias, total_ventas_diarias). Por otro lado se tienen otro RDD que tiene información de todos los videojuegos que se venden en su plataforma con el siguiente formato (id_videojuego, titulo, rating_pegi, rating_esbr). Tener en cuenta que un mismo videojuego se puede vender en distintos países y esos nos permitirá obtener métricas a nivel global. Con esta información escribir un programa en pySpark que permita: a) Obtener el videojuego con más ventas digitales globales (es decir en todos los países) en un RDD con el siguiente formato: (id_videojuego, titulo, total), siendo total la cantidad total de ventas digitales globales b) Para el videojuego con mas ventas, obtener cual es el país para el cual ser registra una mayor tasa de conversión (es decir, mayor total_ventas_diarias / visitas_diarias)