Nama : Raina Imtiyaz

NIM : 2502010976

Class : LB09

Video Link: <https://binusianorg-my.sharepoint.com/personal/raina_imtiyaz_binus_ac_id/_layouts/15/guestaccess.aspx?docid=1499e510e3abf449ca34d9a19f2eff864&authkey=AaaHXESTUHVa1eN9n4Scrxw&e=QgdaJA>

Hyperlink: [2502010976\_Raina Imtiyaz.mp4](https://binusianorg-my.sharepoint.com/personal/raina_imtiyaz_binus_ac_id/_layouts/15/guestaccess.aspx?docid=1499e510e3abf449ca34d9a19f2eff864&authkey=AaaHXESTUHVa1eN9n4Scrxw&e=QgdaJA)

**Codes**

Text

Description automatically generated with low confidence

A picture containing text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

**Text

Description automatically generated**

Text

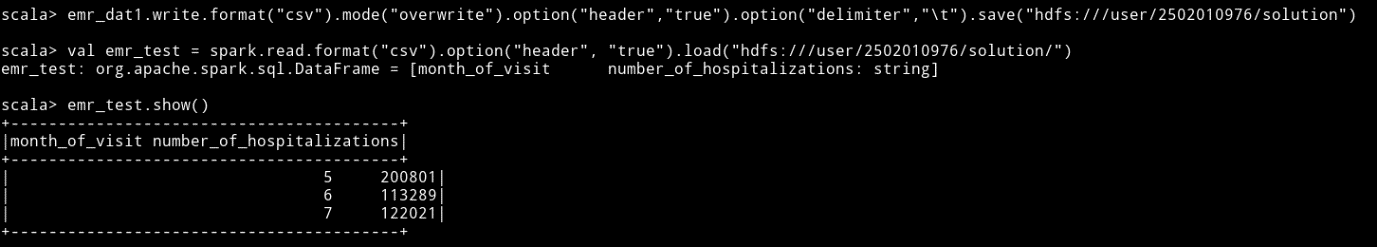
Description automatically generated

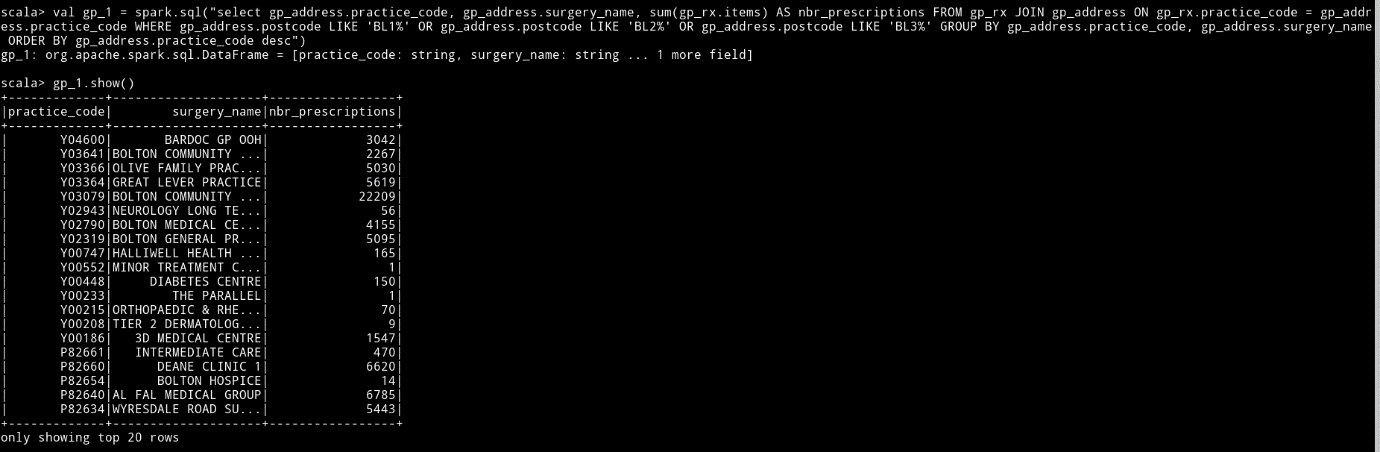
Text

Description automatically generated

Text

Description automatically generated

****

1. 

**Text

Description automatically generated**

**Essay**

1. Suppose you encounter an OOM (Out Of Memory) issue while performing a broadcast join. What is likely cause this happen? How do you deal with this situation?

= Out of Memory error pada saat menampilkan broadcast join disebabkan karena size estimator pada Spark yang terbatas. Ada beberapa cara untuk mengatasi hal tersebut, yang pertama, bisa dengan menggunakan ‘ANALYZE TABLE (AWS | Azure)’ untuk mengumpulkan detail dan menghitung statistik terkait dataframe sebelum bergabung (join). Yang kedua, cache tabel yang akan dibroadcast, jika broadacast join menampilkan BuildLeft, cache sisi kiri pada tabel, jika broadcast join menampilkan BuildRight, cache sisi kanan tabel. Yang ketiga, pada databricks runtime 7.0 ke atas, atur tipe join ke SortMergeJoin dengan mengaktifkan petunjuk join.

1. With Adaptive Query Execution, things become more comfortable for developers as Spark will do the partition coalescing automatically. What this statement means?

= Adaptive Query Execution (AQE) adalah teknik pengoptimalan dalam Spark SQL yang memanfaatkan statistik runtime untuk memilih rencana eksekusi query yang peling efisien. Salah satu fitur utama di AQE adalah menyatukan partisi pasca shuffle. Fitur tersebut menyatukan partisi post-shuffle berdasarkan statistik map output. Fitur tersebut menyederhanakan penyetelan nomor partisi acak saat menjalankan query. Dalam hal ini, Spark dapat memilih nomor partisi acak yang tepat saat runtime setelah jumlah awal pastisi acak yang cukup besar ditetapkan.

1. Explain core parts of a structured streaming application as show in Figure 1. Suppose you are working with a file CSV Source

Diagram

Description automatically generated

= Pertama ada data source, pilihan data type yang disediakan oleh spark adalah file source, kafka source, socket source, dan rate source. Jika bekerja menggunakan sumber CSV berarti data source yang digunakan bertipe file source. Selanjutnya pada processing logic, hal yang dilakukan pertama adalah membuat schema, lalu buat dataframe dari streaming menggunakan metode readStream untuk membaca streaming data sebagai dataframe. Berikutnya terdapat output mode, output mode menentukan cara penulisan data pada sink. Terdapat 3 output mode yaitu append, complete, dan update. Selanjutnya terdapat trigger, trigger digunakan untuk menentukan kapan harus menjalankan logika komputasi streaming yang tersedia pada data streaming yang baru. Terdapat 4 tipe trigger yaitu, not specified (default), fixed interval, one-time, dan continous. Terakhir ada data sink yang digunakan untuk menyimpan output dari streaming application. Tipe sink yang didukung spark adalah file sink, kakfa sink, console sink, dan memory.