a8

pranav

November 12, 2017

Objective

Given the Million Song dataset,the task is to write a spark porgram that computes some queries on the dataset.- * number of distinct songs, artists, and albums * top 5 loudest songs * top 5 longest songs * top 5 fastest songs * top 5 most familiar artists * top 5 hottest songs * top 5 hottest artists * top 5 hottest genres (mean artists hotness in artist_term) * top 5 most popular keys (must have confidence > 0.7) * top 5 most prolific artists (include ex-equo items, if any) * top 5 most common words in song titles (excluding articles, prepositions, conjunctions)

Dataset

The dataset given is the million Song dataset. The dataset is not completely clean and needs to be handled.The output of the subset Million Song dataset is in the output/small folder.

Execution environment

- Operating System - Windows 10
- Java version - 1.8.0_131
- Processor -2
- 4 logical processors 2GHz

Problem 1

Problem 1 consists of 3 queries. Hence,the total running time of problem 1 is slightly higher than the other queries since 3 different queries are fired to get the distinct outputs.Runtime is ** 65727 ms** The output of problem 1 is-

Distinct Artists Count: 44745

Distinct Songs Count: 999056

Distinct Album Count: 221753

Problem 2

Problem 2 outputs the loudest songs in descending order. This is a basic map and sort query on the RDD to get the result which is same for problems 3-7. Runtime ** 15759ms**

	Title	Song.Id	Song.Loudness
1	Modifications	SOCMYZF12AB0186FF4	4.318
2	Track 02	SONHDXQ12AB018C1F1	4.300
3	Hey You Fuxxx	SOZCIHV12AC46894E3	4.231
4	War Memorial Exit	SOEJMJF12AC90715D1	4.166
5	Meta Abuse	SOBEMPS12A8C13BD46	4.150

Problem 3

Simliar to problem 2,but with different parameters to output. The query runs efficiently. Runtime: ** 24568ms**The top 5 longest songs are-

	Title	Song.Id	Song.Duration
1	Grounation	SOOUBST12AC90977B6	3034.9058
2	Raag Shuddha Kalyan Language Hindustani Raga Shuddha Kalyan	SOXUCQN12A6D4FC451	3033.5996
3	Discussion 2	SOOMVZJ12AB01878EB	3033.4429
4	Chapitre Un a Toutes Les Histoires	SOTNVEE12A8C13F470	3032.7637
5	Der Geist des Llano Estacado Ein Spion	SOGFXNB12A8C137BE5	3032.5808

Problem 4

Runtime: ** 35953ms**.The top 5 fastest songs are-

	Title	Song.Id	Tempo
1	Beep Beep	SOVVTEZ12AB0184AAB	302.300
2	Late Nite Lounge WVIP	SOMSJWX12AB017DB99	296.469
3	A Place Called Hope	SOUTBKH12A8C136286	285.157
4	Bellas Lullaby Perrier Citron	SOEVQJB12AC960DA2C	284.208
5	Troubled Times	SOTUXOB12AB0188C3A	282.573

Problem 5

Runtime:** 22725ms**.The top 5 most familiar artists are-

	ArtistId	Name	Familiarity
1	ARCGJ6U1187FB4D01F	Akon Styles P	1
2	ARCGJ6U1187FB4D01F	Akon Sweet Rush	1
3	ARCGJ6U1187FB4D01F	Akon	1
4	ARCGJ6U1187FB4D01F	Akon Eminem	1
5	ARCGJ6U1187FB4D01F	Akon Red Caf	1

Problem 6

Runtime:** 37124ms**.The top 5 hottest songs are-

	Title	Song.Id	Hottness
1	Many Of Horror	SOVWHPM12AB017DABB	1
2	Halfway Gone	SOMFTNM12A8C1434C3	1
3	Stronger	SOWFAZC12A8C13D4B9	1
4	Cooler Than Me	SOBQYCF12AC909726F	1
5	Seven Nation Army Album Version	SOAUFOF12AB0180C65	1

Problem 7

Runtime:** 21005ms**.The top 5 hottest artists are-

	Title	Artist.Id	Hottness
1	ARRH63Y1187FB47783	Kanye West Kid Cudi	1.0825026
2	ARRH63Y1187FB47783	Kanye West John Legend Consequence	1.0825026
3	ARRH63Y1187FB47783	Kanye West Nas Really Doe	1.0825026
4	ARRH63Y1187FB47783	Kanye West JayZ	1.0825026
5	ARRH63Y1187FB47783	Kanye West Talib Kweli QTip Common Rhymefest	1.0825026

Problem 8

This query takes the longest time to execute. This is because the query joins two tables which are in MB of data. Since my machine does not have enough space,the parallel execution is very limited and many mempry blocks fail because of memory restrictions.Even with a better machine,the query time would be longer than the other queries since joinig is an expensive process compared to others. Runtime:** 756041ms**. the top 5 hottest genres are-

	Artist.Term	Average.Hottness
1	christmas songs	0.6083985
2	kotekote	0.6022205
3	musical soundtrack	0.5857621
4	girl rockers	0.5737627
5	alternative latin	0.5737568

Problem 9

This query first filters out data based on confidence and then performs standard map task to get the count of the highest keys.Runtime:** 9386ms**.The top 5 most popular keys are-

	Key	Count
1	7	30420
2	0	28333
3	2	25845
4	9	21283
5	4	15214

Problem 10

Runtime: ** 18092ms** .The top 5 most prolific artists are grouped by the artistId and name and then taking the count-

	Artist.Id	Name	Count
1	ARXPPEY1187FB51DF4	Michael Jackson	194
2	ARH861H1187B9B799E	Johnny Cash	191
3	ARLHO5Z1187FB4C861	Beastie Boys	187
4	AR60ODO1187FB4D9AB	Joan Baez	181
5	AR40GSU1187FB3AA01	Neil Diamond	176

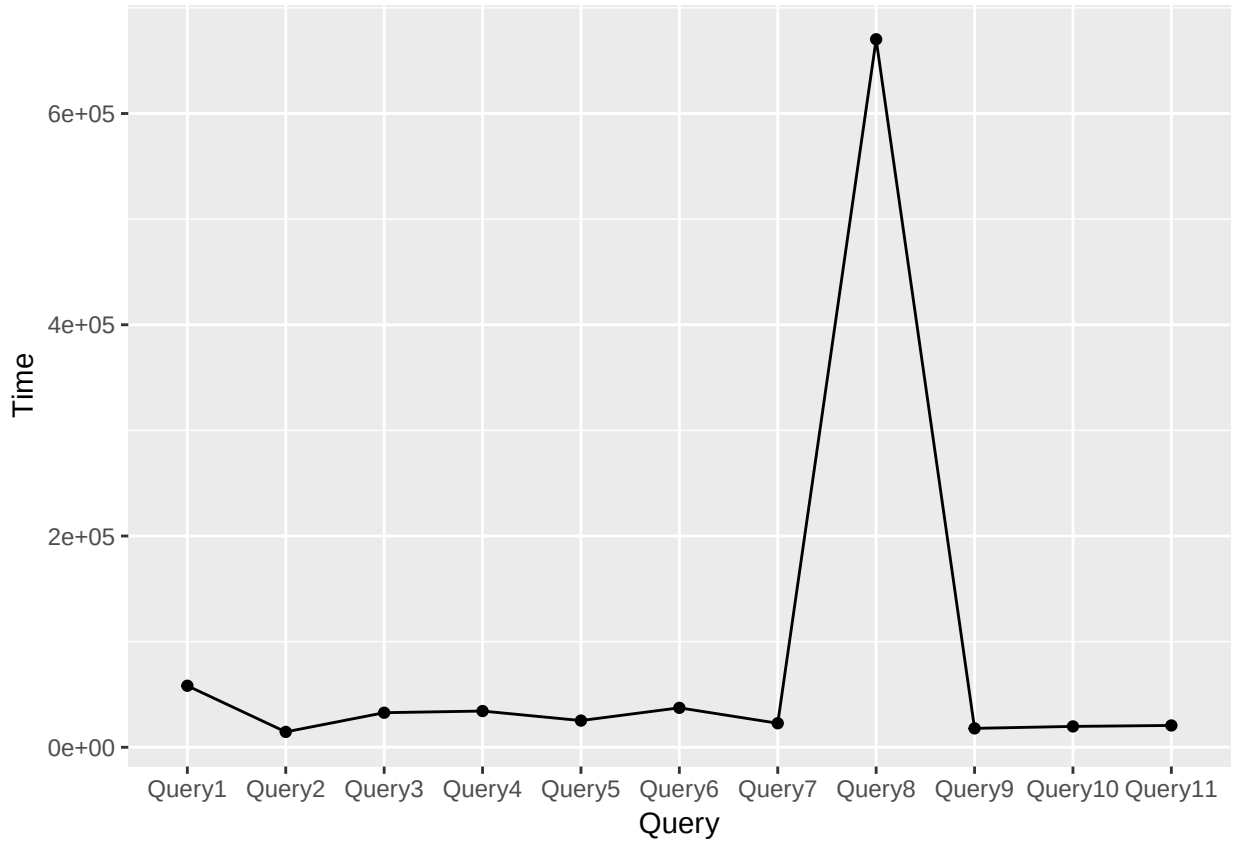
Problem 11

This first converts the RDD to a flatmap structure and filters out the blanks after which it calculates the count of words.Runtime: ** 11975ms**.

	Word	Count
1	version	59385
2	album	35663
3	love	28666
4	live	17297
5	lp	17165

Runtime

The actual runtimes can be found in the output/runtime.txt files. Because of limited space issues,there were many failed RDD's which increased the duration of the execution. The 1st and the 3rd queries take longest because of join/multiple queries.



Conclusion

Spark in general is faster than map-reduce and better for interactive analysis and data analytics. This is because spark uses in-memory execution which eliminates loading times.After the first time,RDD's are persisted and hence can be easily operated upon. However,there are better ways to achieve the tasks done in this assignment with the use of SPARK-SQL since they reduce the complexity allowing the use of normal SQL-queries.