Blog Home

Data Science

Categories

Courses

Spark – FeaturesData Science Tutorials

- **→ Mrochime rhicag**ning Tutorials
- C Tutorials
- -**BlguData** Hadoop & Spark Scala
- *Beglio Cationials
- +RytchondCottosizilds
- **Bytholiphotics** strem Tutorials
- Blg Thabaild Hutlorials
- **XplaidgellSiptetufkellAbtrieris**ls
- +Agrickavionigals
- **Roder Helitalsinist** brials
- **Japlatkurópjals**kt&r**Sals**la
- **Philipping Application of the Company of the Compa**
- *RngutariassTutorials
- Qloh Seb & Tutorials
- SAS Tutorials
- SAP HANA Tutorials
- SAI Tutorials

with Big Data <u>Ge</u> <u>Exclusive Offers of</u> <u>Big Data Course</u>

1. Objective

The Spark SQL performance can be affected by some tunir consideration. To represent our data efficiently, it uses the knowledge of types vereffectively. Spark SQL

Spark Tutorials ×
Spark – Introduction
◆ Spark – Ecosystem Compo
◆ Spark – Features
◆ Spark – Use Cases
◆ Spark – Install On Ubuntu
* Spark – Install multinode C
◆ Spark – Shell Commands
* Spark – Create Project in E
+ Spark – SparkContext
+ Spark – Stage
+ Spark – Executor
+ Spark – RDD
+ Spark – Ways to Create RDD
+ Spark – RDD Persistence &
+ Spark – RDD Features
Spark Interview Questio +

Spark Quiz

plays a great role in th optimization of querie This blog also covers what is Spark SQL performance tuning ar various factors to tune the Spark SQL performance in **Apache**

Spark.

Before reading this blog I would recommend you to read **Spark**

Performance

Tuning. It will increase your understanding of Spark and help further in this blog.



Spark SQL Performance Tuning – Learn Spark SQL

Stay updated with latest technology trends Join DataFlair on **Telegram!!**

Spark Tutorials ×			
◆ Spark – Introduction			
◆ Spark – Ecosystem Compo			
◆ Spark – Features			
◆ Spark – Use Cases			
◆ Spark – Install On Ubuntu			
◆ Spark – Install multinode C			
◆ Spark – Shell Commands			
◆ Spark – Create Project in E			
◆ Spark – SparkContext	_		
◆ Spark – Stage			
+ Spark – Executor	_		
+ Spark – RDD			
◆ Spark – Ways to Create RDD			
* Spark – RDD Persistence &			
Spark – RDD Features			
Spark Interview Questio +			
Spark Quiz +			

2. What is Spark SQL Performanc e Tuning?

Spark SQL is the module of Spark for structured data processing. The highlevel query language and additional type information makes Spark SQL more efficient. Spark SQL translates commands into codes that are processed by executors. Some tuning consideration can affect the Spark SQL performance. To represent our data efficiently, it also uses the knowledge of types very effectively. Spark SQL plays a great role in the optimization of queries. The Spark SQL makes use of **in-memory** columnar storage while caching data. The inmemory columnar is a feature that allows storing the data in a

Spark Tutorials ×	
◆ Spark – Introduction	
+ Spark – Ecosystem Compo	
→ Spark – Features	
◆ Spark – Use Cases	
◆ Spark – Install On Ubuntu	
◆ Spark – Install multinode C	
◆ Spark – Shell Commands	
◆ Spark – Create Project in E	
◆ Spark – SparkContext	
◆ Spark – Stage	
◆ Spark – Executor	
→ Spark – RDD	
◆ Spark – Ways to Create RDD	
◆ Spark – RDD Persistence &	
◆ Spark – RDD Features	
Spark Interview Questio +	

Spark Quiz

columnar format, rather than row format. The columnar storage allows itself extremely well to analytic queries found in business intelligence product. Using columnar storage, the data takes less space when cached and if the query depends only on the subsets of data, thus Spark SQL minimizes the data read.

3. Options for Performanc e Tuning in Spark SQL

There are several different Spark SQL performance tuning options are available:

i. spark.sql.codegen

The default value of spark.sql.codegen is false. When the value of this is true, Spark SQL will compile each query to Java bytecode very quickly. Thus, improves the performance for large

Spark Tutorials ×		
◆ Spark – Introduction		
◆ Spark – Ecosystem Compo		
+ Spark – Features		
♦ Spark – Use Cases		
◆ Spark – Install On Ubuntu		
◆ Spark – Install multinode C		
+ Spark – Shell Commands		
◆ Spark – Create Project in E		
+ Spark – SparkContext		
+ Spark – Stage		
+ Spark – Executor		
+ Spark – RDD		
+ Spark – Ways to Create RDD		
* Spark – RDD Persistence &		
+ Spark – RDD Features		
Spark Interview Questio +		
Spark Quiz +		

queries. But the issue with codegen is that it slows down with very short queries. This happens because it has to run a compiler for each query.

ii.

spark.sql.inMemorycolumnarStorage.compressed

The default value of spark.sql.inMemorycolumnarStorage.compressed is **true**. When the value is true we can compress the in-memory columnar storage automatically based on statistics of the data.

iii.

spark.sql.inMemoryColumnarStorage.batchSize

The default value of spark.sql.inMemoryCommunation agencies is 10000. It is the batch size for columnar caching. The larger values can boost up memory utilization but causes an out-of-memory problem.

iv.

spark.sql.parquet.compression.codec

The spark.sql.parquet.compression.codec uses default snappy compression. Snappy is a library which for compression/decompression.

It mainly aims at very high speed and

Spark Tutorials ×
◆ Spark – Introduction
◆ Spark – Ecosystem Compo
◆ Spark – Features
◆ Spark – Use Cases
◆ Spark – Install On Ubuntu
◆ Spark – Install multinode C
◆ Spark – Shell Commands
◆ Spark – Create Project in E
◆ Spark – SparkContext
→ Spark – Stage
◆ Spark – Executor
→ Spark – RDD
◆ Spark – Ways to Create RDD
◆ Spark – RDD Persistence &
◆ Spark – RDD Features
Spark Interview Questio+

Spark Quiz

reasonable compression.
In most compression, the resultant file is 20 to 100% bigger than other inputs although it is the order of magnitude faster. Other possible option includes uncompressed, gzip and lzo.

Note:

+

In Spark SQL as more **optimizations** are performed automatically, it is possible that following options can get vanished in the further release:

- sql.files.maxPartitionBytes,
- sql.files.openCostI " '
- sql.autoBroadcastJoinThreshold,
- sql.shuffle.partitions,
- sql.broadcastTimeout.

4. Conclusion

In conclusion to Apache Spark SQL, caching of data in in-memory columnar storage improves the overall performance of the Spark SQL applications. Hence, Using the above mention

Spark Tutorials ×
* Spark – Introduction
+ Spark – Ecosystem Compo
+ Spark – Features
+ Spark – Use Cases
+ Spark – Install On Ubuntu
+ Spark – Install multinode C
+ Spark – Shell Commands
* Spark – Create Project in E
+ Spark – SparkContext
+ Spark – Stage
* Spark – Executor
* Spark – RDD
* Spark – Ways to Create RDD
* Spark – RDD Persistence &
◆ Spark – RDD Features
Spark Interview Questio +

Spark Quiz

operations it's easy to achieve the optimization in Spark SQL.

See Also-

- RDD Persistence and <u>Caching Mechanism</u> in Spark
- <u>Spark SQL</u> <u>DataFrame Tutorial</u>
- <u>Spark SQL DataSet</u> Tutorial

Reference for Spark

You give me 15
seconds I promise
you best tutorials
Please share your happy
experience on Google |
Facebook

Tags: apache spark Spark
Spark SQL optimization
Spark SQL Performance tuning
spark-sql

LEAVE A REPLY

Comment Name * Email *

*	Spark – Introduction	This site is pro
	Spark – Ecosystem Compo	the Google Pri
*	Spark – Features	Policy and Ter
*	Spark – Use Cases	Service apply.
*	Spark – Install On Ubuntu	Post Comme
*	Spark – Install multinode C	
•	Snark - Shell Commands	
	Home About us Contac	t us Terms and Condition
	DataFlair © 2021. All Rights	s Reserved.
*	Spark – KDD	
+	Spark – Ways to Create RDD	
+		
	Spark – RDD Persistence &	

Spark Interview Questio... +

Spark Quiz

×

Spark Tutorials

is protected TCHA and le Privacy d Terms of

mment

ditions Privacy Policy Disclaimer Write For Us Success Stories











