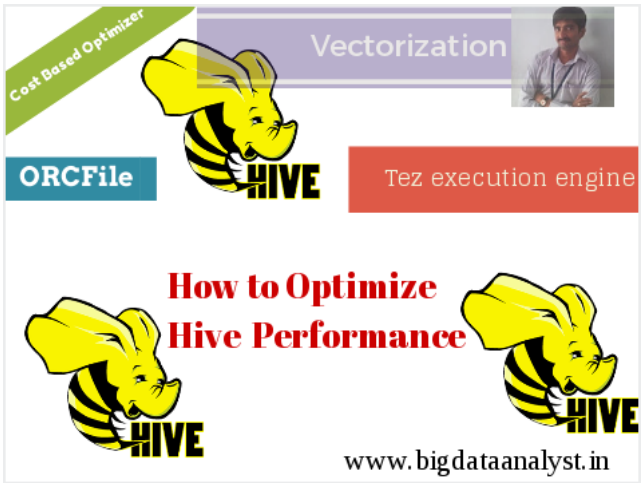


Hive Optimize query performance

19
AUG 2015

by Venu A+ve | posted in: Hive | 3

The Default configuration not suitable for all applications. Few changes can optimize your Hadoop and Hive queries. In this post I am explain about different ways to optimize Hive and few Hive Technical interview questions.

Is SQL scalable? How to run SQL queries in the Hadoop?

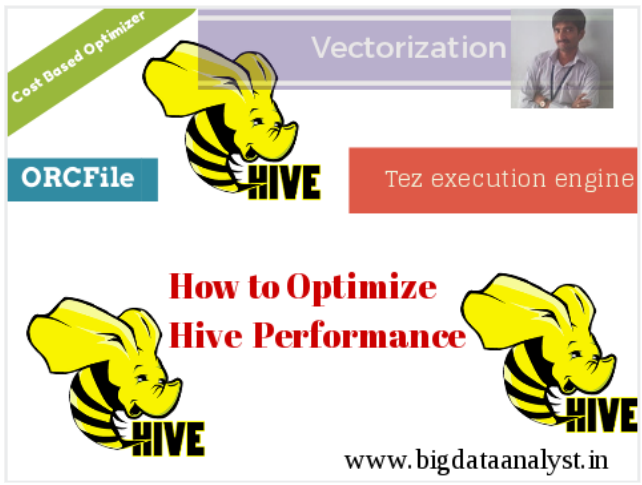
By default SQL is not scalable, SQL databases are vertically scalable. Hadoop is scalable and horizontally scalable. Hive is a Hadoop component which allows programmers to run SQL queries on the top of Hadoop.

Why Scalable is most too important in Hadoop?

Let example: job perform on 3 nodes out of 5 nodes. If one node is failed, automatically job run in 4th node. Everyone knows job failed in distributed process. So job process run without any fail/interruption.

What is the Pros and Cons of Broadcast Join?
In BroadCast Join, small tables are loaded into memory in all nodes. Mappers scans through the large table and joins. It's the best suitable for small tables. It's fast and single scan through a largest table, but if table is more than Ram, it's not process. So when we join two tables, that tables must smaller than RAM. If table more than RAM size, use SortMergeBucket Join.

How Cost Based Optimizer (CBO) optimize Hive?
CBO introduced in Hive 0.14, It's main goal is generate efficient execution plans. Cost Based Optimization (CBO) that leverages statistics collected on Hive tables and optimize hive query optimizer. set two parameters:
set hive.compute.query.using.stats=true;
set hive.stats.dbclass=fs;



How Tez execution engine optimize Hive performance?
If you are using hadoop 2.x use Tez execute engine for better performance. Run following in the terminal to enable Tez execution engine.
set hive.execution.engine=tez;

What is Skewed tables in Hive?
When certain values are appear very often, Skewed tables highly recommendable. Skewed tables split frequently appeared values into a separate table and rest of the values separate to some other file. So when user queried, most appeared values skipped to process again. As a result Hive optimize the performance.

```
create table TableName (column1 string, column2 string) skewed by (column1) on ('x_separate_file')
```

What is Vectorization?
Use Vectorization to improve query performance. It combines multiple rows instead of single row each time. Use given code on terminal to enable it.
set hive.vectorized.execution.enabled = true;
set hive.vectorized.execution.reduce.enabled = true;

What is ORCFile? How it's optimize Hive Query performance?
use ORCFile format to optimize query performance. SNAPPY is the best compression technique use it with ORC format.
CREATE TABLE ORC_table (EmpID int, Emp_name string, Emp_age int, address string) STORED AS ORC tblproperties ("orc.compress" = "SNAPPY");

Other Tips to optimize hive performance:

If you join two tables, one table is smaller another is too big, use Map-side join to optimize the task. If select has multiple fields, leverage to multiple queries format. SELECT count(1) FROM (SELECT DISTINCT column_field FROM table_name) All Imported data automatically partitioned into hourly buckets based on time. Where clause must be used to prevent unnecessary data.

Eg: select name, age, cell from biodata where time > 1349393020 Use hive, ORDER BY use one reducer, SORT BY use multiple reducer.
So if you process a large amount of data, dont' use ORDER BY, prefer SORT BY.

Eg: SELECT name, location, voterid FROM aadhar_card DISTRIBUTE BY name SORT BY age.

Increase Parallelism: Please add given lines to compress the data. ensure maximum split size 265Mb

SET hive.exec.compress.output=true;
SET mapred.max.split.size=256000000;
SET mapred.output.compression.type=BLOCK;
SET mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCodec;

If you are process/joining a small table and Large table use Map side Join.

If you enable "Set hive.auto.convert.join=true"
It can optimize Job performance when you performing Join Operation.

Paste given code in mapred-site.xml to decrease burden on Namenode during sort & shuffling.

It can compress output of Mapper & reduce
<property>
<name>mapred.compress.map.output</name>
<value>true</value>
</property>
<property>
<name>mapred.map.output.compression.codec</name>
<value>org.apache.hadoop.io.compress.SnappyCodec</value>
</property>

If possible Apply SMB map join.
Sort Merge Bucket Join is faster than map join. It's very efficient if applicable, but it's used when you have sorted & bucketed the table.
To enable, use this configuration settings.
set hive.auto.convert.sortmerge.join=true;
set hive.optimize.bucketmapjoin = true;
set hive.optimize.bucketmapjoin.sortedmerge = true;

Hive, hive interview questions

About Author Latest Posts

Venu A+ve

Hive
Interview
Pig
Scala
Spark
Sqoop

I love exploring new technologies especially Hadoop and BigData ecosystems. I would like to share my knowledge through online.

Follow Venu A+ve:

Similar Posts

Previous Post

Next Post

3 Responses

Abhay Kumar 20/08/2015 | Reply

Great Work Venu !! Extremely helpful !!

Naren 05/11/2015 | Reply

Excellent info…really helpful …Thank you Venu

Bhabesh 15/07/2016 | Reply

Great Venu . Very informative and helpfull .. Thanks

Leave a Reply

Comment

Name * **Email (will not be published)** * **Website**

Submit Comment