**IDRP-SCM**

**Optimization Details**

SCPO-DAY

**SCPO-DAY->**

**DP-00-IE-AUTH-KMT-SDAY-> DP-00-IE-DETER\_AUTHO\_KMART\_MSTR**

Date:- 22/11/2018

Developed By :- Heena Salim Shaikh

**JOB DETAILS**

I have worked on the following scripts:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr No** | **Job name** | **Average Run Time** | **Status** |
| 1 | DP-00-IE-MASTERDATA\_GOLDDEP | 0:48:28 | 5 minutes performance improvement but no direct impact on run time |
| 2 | DP-00-IE-DETER\_AUTHO\_KMART\_MSTR | 1:28:17 | 8 minutes performance improvement |
|  |  |  |  |

**DP-00-IE-DETER\_AUTHO\_KMART\_MSTR**

This job have following scripts :-

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Scripts Names** | **Execution Time** |
| 1 | perform\_item\_eligibility\_smith\_\_idrp\_space\_planning\_authorized\_vendor\_package\_stores.sh | 37 |
| 2 | perform\_item\_eligibility\_work\_\_idrp\_vp\_cancarry\_stores.sh | 10 |
| 3 | perform\_item\_eligibility\_work\_\_idrp\_post\_drop\_ship\_proc.sh | 13 |
| 4 | perform\_item\_eligibility\_work\_\_idrp\_vpstores\_after\_order\_dotcom\_edits.sh | 7 |

perform\_item\_eligibility\_smith\_\_idrp\_space\_planning\_authorized\_vendor\_package\_stores.sh takes 37 minutes to execute hence I have worked on this long running script .

**Shell Name** perform\_item\_eligibility\_smith\_\_idrp\_space\_planning\_authorized\_vendor\_package\_stores.sh

**Pig Name**

/appl/hdidrp/pig/scripts/item\_eligibility/smith\_\_idrp\_space\_planning\_authorized\_vendor\_package\_stores/perform\_item\_eligibility\_smith\_\_idrp\_space\_planning\_authorized\_vendor\_package\_stores.pig

**Input Files**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Input Files** | **File Count** |
| 1 | /smith/space\_planning/archive/prev\_\*/SG.GDTU100.DTCITEM | 1217173842 |
| 2 | /smith/idrp/eligible\_loc | 165323 |
| 3 | /smith/idrp/shc\_item\_combined | 8239880 |
| 4 | /smith/idrp/vend\_pack\_combined | 11146269 |
| 5 | /work/idrp/items\_vend\_packs\_can\_carry | 1987953 |
| 6 | /smith/space\_planning/current/sbt\_ord\_pt | 462 |
| 7 | /smith/space\_planning/current/sbt\_vend\_packs | 223371 |
| 8 | /smith/space\_planning/current/sbt\_launch\_str | 300729 |
| 9 | /gold/geographic/model\_store | 36722353 |
| 10 | /smith/idrp/ksn\_attribute/current | 3055858 |

Output Files:-

|  |  |  |
| --- | --- | --- |
| Sr No | Output Files | File Count |
| 1 | /smith/idrp/space\_planning\_vendor\_package\_stores/current | 466242290 |

The final test cases selected out of 150 test cases.

From the following final test cases , when I combined the all the changes ,I got only 6 test cases are working together properly.please find the following changes

|  |  |  |  |
| --- | --- | --- | --- |
| **Script Name** | **Optimize Parameters** | **Time I got** | **Original Timings** |
| the\_original.sh |  | 49 in Minutes 40 Seconds |  |
| sample\_test4 | SET pig.tmpfilecompression.codec com.hadoop.compression.lzo.LzopCodec check it should before compress or after | 44 in Minutes 17 Seconds | 49 in Minutes 40 Seconds |
| sample\_test1 | Without Setting SET mapred.max.split.size 134217728 --SET pig.maxCombinedSplitSize 4000000 | 42 in Minutes 1 Seconds | 45 in Minutes 18 Seconds |
| sample\_test2 | rev\_dtc\_file11 | 42 in Minutes 47 Seconds | 45 in Minutes 18 Seconds |
| sample\_test3 | Setting mappers | 45 in Minutes 23 Seconds | 45 in Minutes 18 Seconds |
| sample\_test5 | Without parameter  --SET mapreduce.map.java.opts: -Xmx3072m --SET mapreduce.reduce.java.opts: -Xmx6144m | Remaining | 45 in Minutes 18 Seconds |
| sample\_test6 | input\_join with replicated keyword | 44 in Minutes 33 Seconds | 49 in Minutes 40 Seconds |
| sample\_test7 | Testing sbt\_vend\_join1,sbt\_ord\_join1 with bracket to Key | 49 in Minutes 21 Seconds | 49 in Minutes 40 Seconds |
| sample\_test8 | filtered\_vend\_packs = FILTER smith\_\_idrp\_vend\_pack\_combined\_data  by (TRIM(ksn\_purchase\_status\_cd) != 'U') ; | 47 in Minutes 36 Seconds | 49 in Minutes 40 Seconds |
| sample\_test9 | store\_num\_join | 49 in Minutes 49 Seconds | 49 in Minutes 40 Seconds |
| sample\_test10 | mdl\_op1\_kns\_attr\_join | 44 in Minutes 13 Seconds | 45 in Minutes 18 Seconds |
| sample\_test11 | Adding Parallel to Union test\_data\_1234 | 48 in Minutes 6 Seconds | 49 in Minutes 40 Seconds |
| sample\_test | Adding Parallel to Union test\_data\_1234 | NA |  |
| sample\_test12 | join2 | 49 in Minutes 36 Seconds | 49 in Minutes 40 Seconds |
| sample\_test13 | op1\_req\_prev | 47 in Minutes 24 Seconds | 49 in Minutes 40 Seconds |
| sample\_test14 | sort factor 10 | 50 in Minutes 41 Seconds | 49 in Minutes 40 Seconds |
| sample\_test15 | set io.sort.mb 2048 | Failed it always fails | 49 in Minutes 40 Seconds |
| sample\_test16 | splited\_rep\_join | 52 in Minutes 58 Seconds | 49 in Minutes 40 Seconds |
| sample\_test17 | union\_data | 56 in Minutes 42 Seconds | 49 in Minutes 40 Seconds |
| sample\_test18 | store\_sarm\_nbr\_union\_final | 48 in Minutes 12 Seconds | 49 in Minutes 40 Seconds |
| sample\_test19 | Set high priority option | 40 in Minutes 35 Seconds | 45 in Minutes 18 Seconds |

**FINAL CODE CHANGES**

**1: Adding the compression technique to the files:-**

**Newer Code:-**

SET io.compression.codec.lzo.class com.hadoop.compression.lzo.LzoCodec

SET pig.tmpfilecompression true

SET pig.tmpfilecompression.codec lzo

**Details:-**

The pig stores the temporary output of the steps like join, filter, group

In the form of the files.

By using the compression technique we can compress this temporary output files.

* Advantages of the LZO compression:-

Reduces the hard disk space occupied by the data.

Reduces the time taken to copy or transfer the data from one location to another.

**2: Setting Mappers in the script:-**

**New Code:-**

SET mapred.min.split.size 524288;

**Details:-**

More number of the mapper more speed.

**3: Changed join rev\_dtc\_file11:-**

**Earlier Code:-**

rev\_dtc\_file11 = JOIN rev\_dtc\_file BY REV\_ITEM, hier\_item\_data\_required BY H\_ITEM\_ID using 'skewed';

**Newer Code:-**

rev\_dtc\_file11 = JOIN hier\_item\_data\_required BY H\_ITEM\_ID, rev\_dtc\_file BY REV\_ITEM;

**Details:-**

I have compared the data of the two files, and kept larger file last.

Generally, we have to adjust the files according to speed.

Also, I have removed the skewed keyword as data was not skewed.

**4: Changed the input\_join :-**

**Earlier Code:-**

input\_join = JOIN input\_req\_gen BY ITEM\_ID, sbt\_ord\_join1 BY sbt\_vend\_join1::filtered\_vend\_packs::shc\_item\_id;

**Newer Code:-**

input\_join = JOIN input\_req\_gen BY ITEM\_ID, sbt\_ord\_join1 BY sbt\_vend\_join1::filtered\_vend\_packs::shc\_item\_id using 'replicated';

**Details:-**

I have added the replicated keyword to the join.

When we have one table small and one table large then a small table can be fit into the memory and we can use the replicated join.

It increases the join speed.

**5: Adding the filtered\_vend\_packs after loading the corresponding file only.**

**Moved Code after:-**

smith\_\_idrp\_vend\_pack\_combined\_data =

LOAD '$SMITH\_\_IDRP\_VEND\_PACK\_COMBINED\_LOCATION'

USING PigStorage('$FIELD\_DELIMITER\_CONTROL\_A')

AS ($SMITH\_\_IDRP\_VEND\_PACK\_COMBINED\_SCHEMA);

filtered\_vend\_packs = FILTER smith\_\_idrp\_vend\_pack\_combined\_data by (TRIM(ksn\_purchase\_status\_cd) != 'U') ;

**Details:-**

If we use the filter after loading the file itself , it increases the speed and performance of the job.

**6: Adding parallel to test\_data\_1234:-**

**Newer Code:-**

test\_data\_1234 = UNION t\_data\_1, t\_data\_2, t\_data\_3 PARALLEL 500;

I have added the PARALLEL keyword here to increase the number of the reducers for the union.

**Details:-**

Reducers do join, union , aggregation and combination operations.

So more number of the reducers more number of data Igt will process and do faster opration.