parquet tool test in bigdataetl02

select a ip to test speed

test repartition in bigdatamr08-10

test in bigdatamr08-10 to use memory

prepare a script to load in parquet instead of loading in mapr

check string min max

pkey file 3days

256m row group

drill cp file

default row group size and block size

------------------------------------------------------------

explain plan for select \* from `dfs`.`/HDS\_VOL\_TMP/el\_oneday/bigdataetl02/el\_oneday/pkey=\*/\*` where srcip\_num='1034180047' and dstip\_num='121202151244';

explain plan for select \* from `dfs`.`/HDS\_VOL\_TMP/el\_oneday/bigdataetl02/el\_oneday` where srcip\_num='1034180047' and dstip\_num='121202151244';

select \* from "dir" where pkey=1 and srcip\_num XXXXX

select \* from "dir" where pkey=1 and srcip\_num XXXXX and trx\_date = 0725

------------------------------------------------------------

matplotlib

Reconfigure bigdatamr08 -10 cluster with 835 gb and 150cores

------------------------------------------------------------

pkey=\_\_HIVE\_DEFAULT\_PARTITION\_\_

------------------------------------------------------------

darrel abc()

------------------------------------------------------------

3day big8-10 02

------------------------------------------------------------

parquet test case phaes II big8-10 02

1) one day fwcdr convert to parquet with no partition by but sortwithinpartition(srcip\_num) only

2) spark write partition by (pkey)

3) reformat the fields back to original 180-132-443 -> 180,130,443

------------------------------------------------------------

------------------------------------------------------------

------------------------------------------------------------

only can hadoop jar parquet tool with hdfs?but not local file?

------------------------------------------------------------

string can check min max

------------------------------------------------------------

[6:49 AM] Kevin Ou

一直怀疑紧etl02 不碎是因为单executor的原因。

​[6:50 AM] Kevin Ou

最烦的是我找不到方法去验证hive 个parquet rowgroup scan

​[6:50 AM] Kevin Ou

感觉它没用谓词下推

------------------------------------------------------------

------------------------------------------------------------

how to drop hive table without rm dir

start=20200723 file name

------------------------------------------------------------

或者我给一个这样的需求。 我要select 2个月的data by src ip ，速度要控制在2mins 以内。

or no sort jost repar and partition by

------------------------------------------------------------

------------------------------------------------------------

('spark.sql.shuffle.partitions',100),\

('spark.default.parallelism',100),\

------------------------------------------------------------

('spark.default.parallelism',100) cluster

------------------------------------------------------------

etl06

------------------------------------------------------------

confirm 8088 schedule space and memory or then

444

dinamy test

spark.dynamicAllocation.minExecutors, spark.dynamicAllocation.maxExecutors, and spark.dynamicAllocation.initialExecutors spark.dynamicAllocation.executorAllocationRatio

8-10 test job

111

[2:04 AM] Kevin Ou

有人有空闲的话研究下有什么方法或者command 生成parquet file 而不需要用spark 的。 例如load firewall cdr awk |split 完之后直接从管道出parquet file

​[2:06 AM] Kevin Ou

因为awk 本身就可以 实现 partition的功能

https://github.com/apache/parquet-mr/tree/master/parquet-cli

8-10 load

222

[6:41 AM] Barry Lu

bigdataetl01

​[6:42 AM] Barry Lu

~/hdsbat/script/dw\_load\_ssl\_udr.sh

​[6:42 AM] Barry Lu

dw\_load\_ssl\_udr.awk

555

matlib panda jupyter

hadoop fs -mkdir maprfs://bigdatamr08-10g:7222/tmp/el/smcin/part\_key=20200720

hadoop fs -chmod 777 maprfs://bigdatamr08-10g:7222/tmp/el/smcin/part\_key=20200720

hadoop distcp /HDS\_VOL\_HIVE/smcin/part\_key=20200720/\* maprfs://bigdatamr08-10g.hksmartone.com/tmp/el/smcin/part\_key=20200720

hadoop distcp maprfs://bigdataetl01g.hksmartone.com/HDS\_VOL\_HIVE/FWCDR/start\_date=20200602 /HDS\_VOL\_TMP/el\_oneday/bigdataetl02/src

------------------------------------------------------------

decimal spark hive

------------------------------------------------------------

hadoop distcp /HDS\_VOL\_HIVE/FWCDR/start\_date=20200602 maprfs://bigdatamr08-10g.hksmartone.com/HDS\_VOL\_TMP/el\_oneday/bigdataetl02/src

hadoop jar /home/hdsbat/parquet-tools-1.11.0.jar convert file:///app/HDSBAT/eloisewu/fwcdr\_ldr\_01203.gz

paper

pkey string or pkey int

alter table h\_fw\_el change srcip\_num srcip\_num bigint;

('spark.default.parallelism',100),\

('spark.sql.shuffle.partitions',100)\

discp

how to check hdfs df

ict job

bigdatamr08 还有10t的位置。可以人手copy 整个8月份过去转了

matlib panda jupyter

ppt

chmod

java parquet

[8/7 6:41 AM] Barry Lu

bigdataetl01

​[8/7 6:42 AM] Barry Lu

~/hdsbat/script/dw\_load\_ssl\_udr.sh

​[8/7 6:42 AM] Barry Lu

dw\_load\_ssl\_udr.awk

my load fwcdr job

dw\_load\_fwcdr.sh

dw\_fwcdr\_convert.awk

jar parquet tool：：：

http://www.openkb.info/2015/02/how-to-build-and-use-parquet-tools-to.html

ultimate -n

crontab

daily run job fwcdr

test hive rowgroup

chmod +x

outlook mail

(base) [hdsbat@bigdataetl02 reload]$ pwd

/app/reload

kelvin ea info

from (select ACCOUNT,MSISDN,USER\_ID,START\_DATE,END\_DATE

from ${etlvar::ADWDB}.NORTON\_ACR\_MAP

where &rpt\_e\_date between START\_DATE and END\_DATE) a

left outer join ${etlvar::ADWDB}.bill\_servs b

on a.ACCOUNT = b.CUST\_NUM and a.MSISDN = b.SUBR\_NUM

and b.BILL\_END\_DATE >= &rpt\_s\_date

and b.BILL\_START\_DATE <= &rpt\_e\_date

define rpt\_s\_date=trunc(add\_months(to\_date('${etlvar::TXDATE}','YYYY-MM-DD'),-1),'mm');

define rpt\_e\_date=last\_day(add\_months(to\_date('${etlvar::TXDATE}','YYYY-MM-DD'),-1));

,case when trunc(b.BILL\_START\_DATE,'mm') = trunc(add\_months(to\_date('${etlvar::TXDATE}','YYYY-MM-DD'),-1),'mm') --&rpt\_s\_date

---------------------------------------------------------------------------------------------

bigdatetl02

[8/14 4:13 AM] Kevin Ou

做个统计by file的日子

​[8/14 4:13 AM] Kevin Ou

count 每日几多个file

​[8/14 4:13 AM] Kevin Ou

话个图看看增长

​[8/14 4:14 AM] Kevin Ou

bigdataetl02

​[8/14 4:15 AM] Kevin Ou

group by ： Jul 20 那两个fields

​[8/14 4:15 AM] Kevin Ou

count files 数

[8/14 4:16 AM] Kevin Ou

pandas read files 做group by 然后出去matplot 话图

​[8/14 4:16 AM] Kevin Ou

参考上次个sample

讲埋个backupground。 我地怀疑 md 有一段时间send 了好多file 过来 可能过万。我想by file 的日子分析下

x 轴time y轴 count

pandas 可以先组合column再group by

import matplotlib.pyplot as plt

import pandas as pd

dtype\_dic= { '1':str, '2':str, '3':str,'4':str, '5':str, '6':str, '7':str, '8':str,'9':str}

df=pd.read\_csv("/app/reload/2t\_list.log",sep='\s+',header=0,dtype = dtype\_dic)

y\_list=df.sort\_values(by=['6','7']).groupby(['6','7']).size()

print (y\_list.to\_list())

df["date"] = df["6"].astype(str) + '.' + df["7"].astype(str)

x\_list=df["date"].sort\_values(ascending=True).unique()

print(x\_list)

plt.rcParams["figure.figsize"] = (16, 8)

plt.title("md files")

plt.xlabel("date")

plt.ylabel("count")

plt.plot(x\_list,y\_list)

plt.show()

import matplotlib.pyplot as plt

import pandas as pd

dtype\_dic= { '1':str, '2':str, '3':str,'4':str, '5':str, '6':str, '7':str, '8':str,'9':str} ???

df=pd.read\_csv("/app/reload/2t\_list.log",sep='\s+',header=0,dtype = dtype\_dic)

df["date"] = df["6"].astype(str) + '.' + df["7"].astype(str)

y\_list=df.sort\_values(by=['date']).groupby(['date']).size()

print (y\_list)

x\_list=df.sort\_values(by=['date']).groupby(['date']).groups.keys()

print(x\_list)

plt.rcParams["figure.figsize"] = (16, 8)

plt.title("md files")

plt.xlabel("date")

plt.ylabel("count")

plt.plot(x\_list,y\_list)

plt.show()

import matplotlib.pyplot as plt

import pandas as pd

df=pd.read\_csv("/app/HDSBAT/eloisewu/mem.log",sep='|',header=None)

#print(df)

y\_list\_mem=df[2].to\_list()

print (y\_list\_mem)

x\_list=range(0,len(y\_list\_mem))

plt.rcParams["figure.figsize"] = (8, 8)

plt.title("Interactive Plot")

plt.xlabel("Cnt")

plt.ylabel("MEM used")

plt.plot(x\_list,y\_list\_mem,label='mem')

plt.show()

support

Stone check

20200824：

new job pandas xlrd

ppt

daily run

support

sub\_eloisewu

dw\_hkrevrpt\_rpt

It doesn't matter

20200831：

M55082069 remove

nsp

remote

irene 5g hpp step email view

norton

ict

mddb bat file log

daily load fwcdr hive script mutt

https://pandas.pydata.org/docs/pandas.pdf

\\GZPC-MASTER7\IS-Dev\GZDWTeam\enhancement\SR000000\_SHK\_REPORT\_AUTOMATE\DOC

交换分区的数据不影响create\_ts

%date:~-4%-%date:~4,2%-%date:~7,2%\_%time:~1%

%date:~-4%-%date:~4,2%-%date:~7,2%\_%time:~1%

darrel kelvin irene

sum(decode(shk\_sale\_tag,'5G\_POTENTIAL',1,0))

count(decode(shk\_sale\_tag,'5G\_POTENTIAL',1,0))

sum(case when cl\_sale\_type='UP\_SELL' then bf\_plan\_tariff else 1 end) as disp\_val

,sum(case when cl\_sale\_type='UP\_SELL' then bf\_plan\_tariff else 0 end) as disp\_val--/

,decode(sum(case when cl\_sale\_type='UP\_SELL' then af\_plan\_tariff else 0 end),0,1,sum(case when cl\_sale\_type='UP\_SELL' then af\_plan\_tariff else 0 end)) as disp\_val2

==

,sum(decode(cl\_sale\_type,'UP\_SELL',bf\_plan\_tariff,0)) as disp\_val--/

,decode(sum(decode(cl\_sale\_type,'UP\_SELL',af\_plan\_tariff,0)),0,1,sum(decode(cl\_sale\_type,'UP\_SELL',af\_plan\_tariff,0))) as disp\_val2

sed -e "s/\[PARAM\]/"$f"/g" ic\_tpl.sql > ./ic\_1.sql

weekly work

case when like

case when reach=0 then 0 else hit/reach

rpt + % .123

.xx (2)

others

\*1000

1

20000

for

VARCHAR2(50 BYTE)

ru\_plan\_mix\_sub\_nature VARCHAR2(100 BYTE) DEFAULT ' ',

--where h.cl\_hit like 'HIT%'

[4:15 AM] Kevin Ou

Oh , sorry. Loop batch\_id and af\_comm\_channel

​[4:15 AM] Kevin Ou

RunRptDate is a job parameter

​[4:16 AM] Kevin Ou

Of course you can treat as 3 parameters and pass to every report generation request.

trx\_date

delear\_name

contractmonth

bmrpt\_ld

notify ea

ea log

billimg.bmrpt\_tmp\_summ a.DEALER\_CD, ld ldmonths

distinct bill\_reference from billimg.bmrpt\_tmp\_summ

must to remove BM\_LD\_HS\_SUBSIDY\_2\_202006.csv

shk table col not null

nvl(tt.ptl,'0%') error

0.00%

,case when sum(t.cnt\_sub\_nature) over(partition by t.batch\_id,t.shk\_tier) = 0

then 0

else t.cnt\_sub\_nature / sum(t.cnt\_sub\_nature) over(partition by t.batch\_id,t.shk\_tier)

end as ptg

nsl\_langue

pandas

3 params

Trim(b.Subr\_Sw\_Off\_Date) || '|' ||

Trim(b.Subr\_Stat\_Cd) || '|' ||

LD.csv

bonnie bill\_reference email compare with kevin logic

m\_bill\_reference = s\_bill\_reference

select \* from all\_tab\_columns where column\_name like 'CHRG\_NATURE' AND TABLE\_NAME LIKE 'NBIOT%'

vm

BILLIMG.bmrpt\_ict\_tmp\_summ\_01T

<https://pandas.pydata.org/docs/pandas.pdf>

？？？

重新整理这部分

[Yesterday 2:04 AM] Kevin Ou

shk  份report 等埋mo 给个 input file 再继续。暂时可以摆一摆。

​

[Yesterday 2:04 AM] Kevin Ou

手头还有咩。

​

[Yesterday 2:04 AM] Kevin Ou

没就安心过国庆了

​

[Yesterday 2:05 AM] Eloise Wu

貌似没有了

​

[Yesterday 2:05 AM] Eloise Wu

(rofl)

​

[Yesterday 2:05 AM] Kevin Ou

当然。得闲个阵去看看前面计数得logic 也好。

​

[Yesterday 2:05 AM] Kevin Ou

国庆做support wor

​

[Yesterday 2:06 AM] Eloise Wu

嗯嗯

​

[Yesterday 2:06 AM] Eloise Wu

计数logic？

​

[Yesterday 2:06 AM] Kevin Ou

01 -04 perl

[Yesterday 2:06 AM] Eloise Wu

噢噢

​

[Yesterday 2:07 AM] Eloise Wu

还有个

​

[Yesterday 2:07 AM] Eloise Wu

"$168 / 8GB å…¨é€Ÿç™»6éžå‡¡è¨ˆåŠƒ",0.34%,0.21%,1.29%,0.00%,0.00%,1.27%,0.00%,0.25%,0.00%,0.00%,1.04%,0.00%,0.00%,0.00%,0.00%,0.00%,0.00%,0.31%,0.00%,0.00%,0.00%,0.00%,0.00%,2.78%,1.35%,0.00%,0.00%

​

[Yesterday 2:07 AM] Eloise Wu

中文乱码

​

[Yesterday 2:07 AM] Eloise Wu

数据库没乱码 出csv乱码

​

[Yesterday 2:08 AM] Kevin Ou

研究下。中文

​

[Yesterday 2:10 AM] Kevin Ou

看看splplus output 中文

​

[Yesterday 2:10 AM] Kevin Ou

记忆中跟nls\_lang 有关。

​

[Yesterday 2:13 AM] Eloise Wu

噢噢

​

[Yesterday 2:15 AM] Eloise Wu

节后我的任务是什么

[Yesterday 2:15 AM] Eloise Wu

有没有什么大任务吖

​

[Yesterday 2:15 AM] Kevin Ou

收尾呢个report

​

[Yesterday 2:15 AM] Eloise Wu

噢噢

​

[Yesterday 2:16 AM] Kevin Ou

然后 etl migration

​

[Yesterday 2:16 AM] Kevin Ou

datastage upgrade 啦。 可能还要带一个report 在身

​

[Yesterday 2:19 AM] Eloise Wu

噢噢噢

[Yesterday 7:19 AM] Eloise Wu

后面有job用到pandas的吗

​

[Yesterday 7:19 AM] Kevin Ou

将来会开放用。

​

[Yesterday 7:19 AM] Kevin Ou

尤其bigdata 那些

​

[Yesterday 7:19 AM] Kevin Ou

manual 那些吧

Edited

​

[Yesterday 7:19 AM] Eloise Wu

噢噢噢

​

[Yesterday 7:27 AM] Kevin Ou

回来有堆datastage job 做

​

[Yesterday 7:27 AM] Eloise Wu

噢噢噢

​

[Yesterday 7:28 AM] Eloise Wu

big data 咧

​

[Yesterday 7:28 AM] Kevin Ou

= 紧upgrade

​

[Yesterday 7:29 AM] Kevin Ou

vendor 搞店后 要搬个fwcdr 过去试

