

Big data mini project

AD1511

Problem statement

Apache Hive is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale. A data warehouse provides a central store of information that can easily be analyzed to make informed, data driven decisions.

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark

The problem we have choosed here is 'subject'

With this problem I have explained the commands in this project.some are create , insert,sqrt,floor etc.

And I have load ,dump the commands using the apache pig language.

Pig comments:

Start-all.sh

Cd \$HIVE_HOME

Hive

Create database subject;

Use database;

Create table subject(id int, subname string,marks1 int,marks2 int);

Descibe subject;

Insert into subject values(01,'bda',78,65);

Insert into subject values(02,'network',87,78);

Insert into subject values(03,'soft computing '78,65);

Insert into subject values(04,'maths',78,65):

Insert into subject values(05,'Java',78,95);

Select * from subject;

Select subname,marks1 from subject;

Select count (*) from subject;

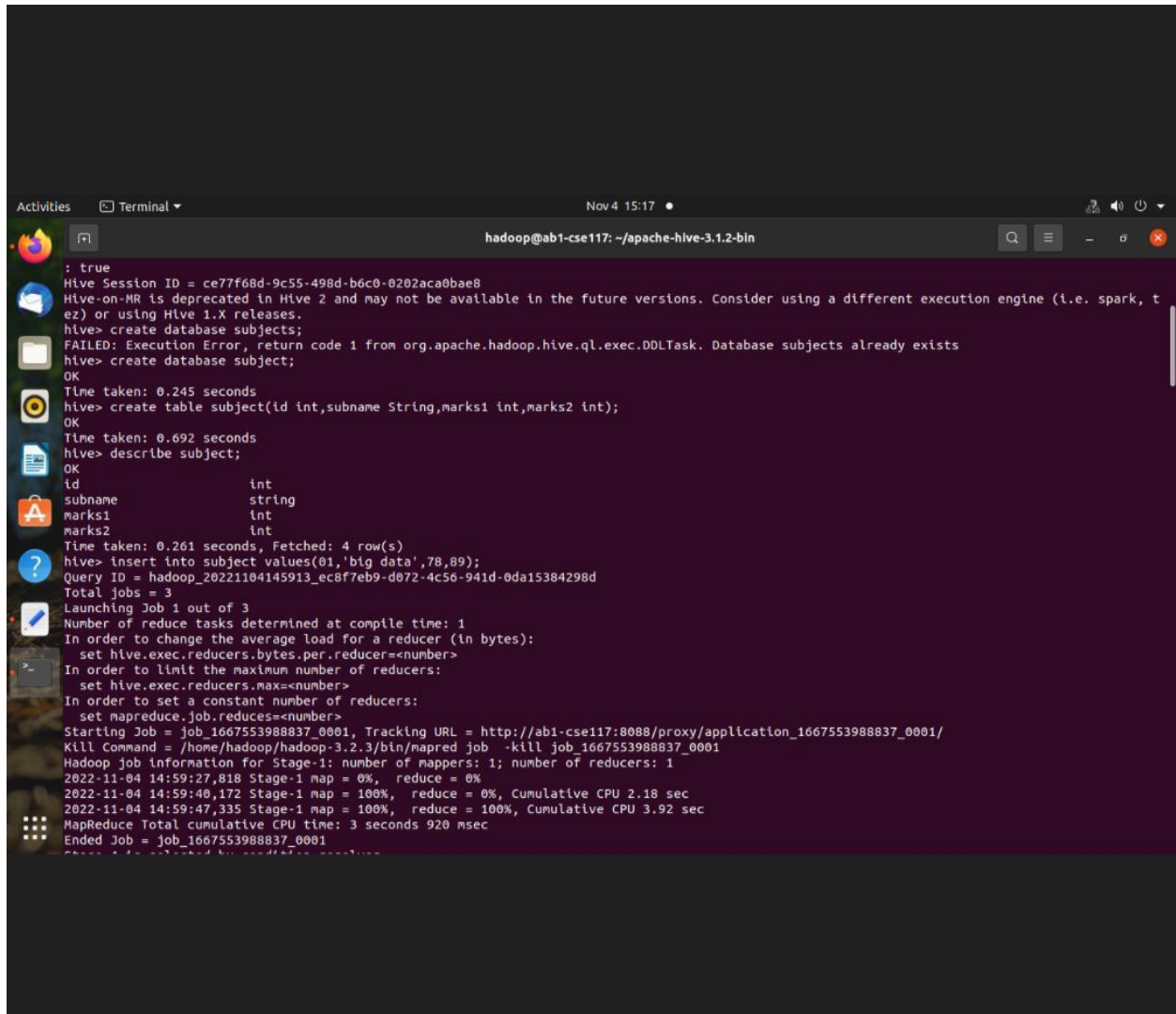
Select sum(marks1),avg(marks1) from subject;

Select max (marks1) from subject;

Select sqrt(marks1) from subject;

Select floor(marks1) from subject;

- Output



```
hadoop@ab1-cse117: ~/apache-hive-3.1.2-bin
: true
Hive Session ID = ce77f68d-9c55-498d-b6c0-0202aca0bae8
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> create database subjects;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.q1.exec.DDLTask. Database subjects already exists
hive> create database subject;
OK
Time taken: 0.245 seconds
hive> create table subject(id int,subname String,marks1 int,marks2 int);
OK
Time taken: 0.692 seconds
hive> describe subject;
OK
id                int
subname           string
marks1            int
marks2            int
Time taken: 0.261 seconds, Fetched: 4 row(s)
hive> insert into subject values(01,'big data',78,89);
Query ID = hadoop_20221104145913_ec8f7eb9-d072-4c56-941d-0da15384298d
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1667553988837_0001, Tracking URL = http://ab1-cse117:8088/proxy/application_1667553988837_0001/
Kill Command = /home/hadoop/hadoop-3.2.3/bin/mapred job -kill job_1667553988837_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-04 14:59:27,818 Stage-1 map = 0%, reduce = 0%
2022-11-04 14:59:40,172 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.18 sec
2022-11-04 14:59:47,335 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.92 sec
MapReduce Total cumulative CPU time: 3 seconds 920 msec
Ended Job = job_1667553988837_0001
```

Activities Terminal Nov 4 15:17
hadoop@ab1-cse117: ~/apache-hive-3.1.2-bin

```
Time taken: 37.024 seconds
hive> insert into subject values(02,'maths',89,90);
Query ID = hadoop_20221104150022_228dde5d-d20f-4cd2-bb87-3518ec193f8e
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1667553988837_0002, Tracking URL = http://ab1-cse117:8088/proxy/application_1667553988837_0002/
Kill Command = /home/hadoop/hadoop-3.2.3/bin/mapred job -kill job_1667553988837_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-04 15:00:32,486 Stage-1 map = 0%, reduce = 0%
2022-11-04 15:00:49,743 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.19 sec
2022-11-04 15:01:01,985 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.24 sec
MapReduce Total cumulative CPU time: 8 seconds 240 msec
Ended Job = job_1667553988837_0002
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://127.0.0.1:9000/user/hive/warehouse/subject/.hive-staging_hive_2022-11-04_15-00-22_814_1661498037838617385-1/-e
xt-10000
Loading data to table default.subject
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.24 sec HDFS Read: 17605 HDFS Write: 312 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 240 msec
OK
Time taken: 42.93 seconds
hive> insert into subjects values(03,'data mining',89,98);
FAILED: SemanticException [Error 10001]: Line 1:12 Table not found 'subjects'
hive> insert into subject values(03,'data mining',89,98);
Query ID = hadoop_20221104150135_6b33e150-bda9-4485-a346-5eb71ac12a7a
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
```

Pig commands:

```
Subject = LOAD '/home/hadoop/mini.txt' USING PigStorage(',') as (Subname:chararray,marks1 :  
int,marks2:int,id:int);
```

```
Dump subject;
```

```
Data=GROUP subject BY marks1;
```

```
Dump data;
```

```
Order=ORDER subject by marks asc;
```

```
Dump order;
```

```
Sub1 = JOIN subject by marks1,subject1 by marks1;
```

```
Dump limitt;
```

```
Limitt = LIMIT food 2;
```

```
Dump limitt;
```

OIN file:/tmp/tenp-3029687/tmp-1588622677,

Input(s):

Successfully read 6 records from: "/home/hadoop/minl.txt"

Successfully read 6 records from: "/home/hadoop/minl.txt"

Output(s):

Successfully stored 8 records in: "file:/tmp/tenp-3029687/tmp-1588622677"

Counters:

Total records written : 8

Total bytes written : 0

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job_local1671699273_0008

2022-11-04 15:04:39,093 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system

2022-11-04 15:04:39,094 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system

2022-11-04 15:04:39,095 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system

2022-11-04 15:04:39,096 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher

2022-11-04 15:04:39,097 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated

bytes-per-checksum

2022-11-04 15:04:39,097 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been

2022-11-04 15:04:39,098 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to

2022-11-04 15:04:39,098 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input

(idly,20,20,5,idly,20,20,5)

(chapathi,30,10,25,chapathi,30,10,25)

(puri,50,10,30,puri,50,10,30)

(puri,50,10,30,dosa,120,10,30)

(dosa,120,10,30,puri,50,10,30)

(dosa,120,10,30,dosa,120,10,30)

(friedrice,100,100,60,friedrice,100,100,60)

(briyani,120,170,110,briyani,120,170,110)

grunt> █

