HIVE,PIG GROOMING SESSION(23FEB17)

max stock variance using MR

stock volume using MR - changed to

AEA 1254255 (LongWritable)

AEA 1.254255 (DoubleWritable)

stock volume in millions..

What is Pig?

It is wrapper built on top of MR

Pig Latin

Pig Compiler will convert all the pig commands into mapreduce jobs

how many modes u can run pig?

local and mapreduce

Without changing or coming out of the shell,we can switch to one mode to another by passing parameters.

pig -x local -p input=/home/hduser/custs -f A.pig

pig -p input=/user/hduser/custs -f A.pig

pig -param\_file = PigInput -f A.pig

cat PigInput

input = /user/hduser/custs

A.pig

load '$input' using PigStorage();

Different relational operators

1.Foreach....generate- reading each record and generates a new bag

2. Order by-sorting is done in order by

3. Filter-filters the previous bag on some condition/criteria

4. group-Mapper class

5 distinct

6. join

7. limit

8 union

9. Split

Group BY - Mapper class

key, value

key as group, tuples in the bag

customer master

custid, first, last, age, profession

(profession, age)

A = load 'custs';

B = group A by (profession, age);

Actor 30 - 3 tuples

Actor 27 - 2 tuples

Actor ( 5 tuples)

Lawyers

Carpenters

Actor ,30 - (3 tuples)

Actor 27 - (2 tuples)

Lawyers,30

Lawyers,35

Lawyers,45

Sales

------

1,4000001,50.45

2,4000002,70.55

3,4000003,80.00

select sum(amount) from sales;

Ans : 201

in Pig how will u calc total amount

A = Load 'sales';

B = group A all;

C = foreach B generate SUM(A.amount);

dump C;

Flatten

to convert index elements of an array to ind tuples.

SQL Assets

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

Hive = Hadoop + SQL

What are best practises in pig?

100 columns

20 years of data

1. filter the data-fetch the reqd columns as per need

A = load 'data'; (100 columns)

A = foreach A generate col1, col2;

A = filter A by col2>100;

2. give datatypes..

def datatype is bytearray

string - chararray

What is Hive?

Wrapper on mapreduce. For processing structured queries on unstructured data

how to take a particular charcter in pig.....exaple "INDIAN" is given string i only want INDIA....

Id Name Nationality

1 Rajesh Indian

A = Load 'data' ;

B = foreach A generate $0, $1, SUBSTRING(nationality,0,5)

<https://pig.apache.org/docs/r0.13.0/func.html>

Taking user-side input in pig

pig -p input=/user/hduser/custs -p myage=30 -f A.pig

A.pig

X = load '$input' using PigStorage(',');

Y = Filter X by age>$myage

command in pig similar to like clause in hive is regex

Rank-skip the first record,to generate unique key for each line,used to give the rank to the rows not the columns

ex:-

id,name

1,Sanjay

2,John

A = Load 'data'

B = rank A;

dump B;

1,id,name

2,1,Sanjay

3,2,John

describe B;

C = filter B by ranked\_column>1

dump C;

2,1,Sanjay

3,2,John

D = foreach C generate $1,$2;

dump D;

1,Sanjay

2,John

In this way we can get rid of header.

store D into '/home/hduser/folder' using PigStorage(',');

Hive - how many types of tables

Hive Managed Tables or Internal

Hive External Tables

Partitions and Buckets-partition or buckets is usually done to reduce the load and do perform fast search or process from the datasets

Scenario:-

1 million records of sales- one File

each sale transaction is handled by one salesperson

there are 10 sp in the company

10 files - one each for one salesperson

divide one file of each sp into 4 files- bucketing

40 files in total

select \* from tbl where sp=1001

Oozie is for automating the no of tools in one basket.