Statistical Implementation

PROJECT PHASE 4 PART 1

Srinivas Koushik Kondubhatla (UFID: 69238911)
Shruti Shivani (UFID: 90059477)

Code Execution

Instructions For Code Compilation

- Extract the zip file
- Execute the following to **run test cases.**

```
lin114-00:~> cd a4-1test/
lin114-00:~/a4-1test> make -f Makefile
make: 'a4-1.out' is up to date.
lin114-00:~/a4-1test> ./a4-1.out [0-11]
```

Instructions For Unit Testing

- Extract the zip file
- Execute the following

```
lin114-00:~> cd a4-1test/
lin114-00:~/a4-1test> make -f Makefile
make: 'a4-1.out' is up to date.
lin114-00:~/a4-1test> g++ gtests.cc -lgtest -lpthread
lin114-00:~/a4-1test> ./a.out
```

CODE FUNCTIONALITY EXPLANATION

STATISTICS CLASS

The job of the statistics.cc class is to store estimate how many tuples does the query result to without really running the query. This plays a very pivotal role while we are trying to create an query optimizer. The implementation of each class functionality is explained below.

Data Structure Definition

```
struct RelStructure {
    map<string,int> attributes;
    int n;
};
map<string,struct RelStructure> relMapping;
```

AddRel(relName, numOfTuples)

This method will store the relation name and the number of distinct tuples in that relation.

AddAtt(relName, attName, numDistincts)
This method will store the number of distinct tuples of the attribute **attName** of the relation **relName**.

CopyRel(oldName, newName)
This method will copy the **RelStructure** of the **oldName** relation to the **newName**

Write(fileName)
This method will store the contents of **relMapping** into **<fileName>.txt**. The format of the content is as follows
<relation name>: <att1>-<n1> <att2>-<n2> …. <n> \n
Example:

```
partsupp:ps_suppkey-10000 799999
supplier:s_suppkey-10000 799999
```

Read(fileName)
This method will interpret the **<fileName>.txt** and add the relations and attributes to the **relMapping** of that instance.

Apply(parseTree, relNames, numToJoin)
This method will estimate the number of tuples would be returned after executing CNFs in parseTree and update the number of tuples of attributes and relation in **relMapping**

Estimate(parseTree, relNames, numToJoin)
This method will estimate the number of tuples would be returned after executing CNFs in parseTree and return the estimated result.

HOW THE ESTIMATION WORKS?
There are 2 major operations used in **JOIN.**
estimatedTuples = Factor * (# of tuples in left relation) * (# of tuples in right relation)

- **Equality Constraint with a constant :**
  gamma = 1/(# of tuples in the attribute)

- **Equality Constraint with a foreign key :**
  gamma = 1/max(# of tuples in the attribute in left relation, # of tuples in the attribute in right relation)

- **Non-Equality Constraint(Less than, Greater Than) with a constant:**
  gamma = 1/3

- **If Attributes not in the current Statistics instance**
  gamma = 1/3

- **For all expressions involving same relations combined with OR**, the gammas are combined as follows
  Gamma = ∏(1-Gi)

Gi = gamma of each expression

- **For all expressions involving different relations combined with OR**
  gamma =∑Gi
  Gi = gamma of each expression

- **For all expressions involving different relations combined with OR**
  gamma =∑G
  Gi = gamma of each expression

- **For all expressions combined with AND**
  Factor = ∏(Gi)
  Gi = gamma of OR expressions

Screenshots
output41.txt

```
lineitem:l_discount-11 l_returnflag-3 l_shipmode-7 857316
********************************************************************************************************************************************************
customer:c_custkey-150000 c_nationkey-25 1500000
nation:n_nationkey-25 1500000
orders:o_custkey-150000 1500000
********************************************************************************************************************************************************
customer:c_custkey-150000 c_mktsegment-5 400080
lineitem:l_orderkey-1500000 400080
orders:o_custkey-150000 o_orderkey-1500000 400080
********************************************************************************************************************************************************
customer:c_custkey-150000 c_nationkey-25 2000404
lineitem:l_orderkey-1500000 2000404
nation:n_nationkey-25 2000404
orders:o_custkey-150000 o_orderkey-1500000 2000404
********************************************************************************************************************************************************
customer:c_custkey-150000 c_nationkey-25 2000404
lineitem:l_partkey-200000 l_shipinstruct-4 l_shipmode-7 21432
nation:n_nationkey-25 2000404
orders:o_custkey-150000 o_orderkey-1500000 2000404
part:p_container-40 p_partkey-200000 21432
********************************************************************************************************************************************************
```

Gtests

```
lin114-00:~/a4-1test> ./a.out
[==========] Running 4 tests from 4 test suites.
[----------] Global test environment set-up.
[----------] 1 test from Lineitem
[ RUN      ] Lineitem.Query1
[       OK ] Lineitem.Query1 (1 ms)
[----------] 1 test from Lineitem (1 ms total)

[----------] 1 test from part
[ RUN      ] part.Query2
[       OK ] part.Query2 (0 ms)
[----------] 1 test from part (0 ms total)

[----------] 1 test from Supplier
[ RUN      ] Supplier.Query5
[       OK ] Supplier.Query5 (1 ms)
[----------] 1 test from Supplier (1 ms total)

[----------] 1 test from SupplierPartsup
[ RUN      ] SupplierPartsup.Query10
[       OK ] SupplierPartsup.Query10 (1 ms)
[----------] 1 test from SupplierPartsup (1 ms total)

[----------] Global test environment tear-down
[==========] 4 tests from 4 test suites ran. (3 ms total)
[  PASSED  ] 4 tests.
```