

PARALLEL AND DISTRIBUTED SYSTEMS

SQL assignment

Martí Municoy, Jorge Pardillos

Q0. Can you describe the series of steps to open a database for querying?

1. Open *MySQL* and type your password:

```
mysql -u root -p
```

2. In order to explore the available databases, type:

```
mysql> show databases;
```

3. To choose a specific database, for instance called *A*, type:

```
mysql> use A;
```

Q1. What is the purpose of this query?

```
mysql> SELECT * from Sources;
```

It provides us with all the table entries of the table called *Sources*.

```
+-----+-----+
| exptId | source   |
+-----+-----+
| 1      | Pancreas |
| 2      | Liver    |
| 4      | Human Liver |
+-----+-----+
3 rows in set (0,00 sec)
```

Q2. Get 5 *GenBank* ids and corresponding *descriptions*.

```
mysql> SELECT gbId, description FROM Descriptions LIMIT 5;
```

```
+-----+-----+
| gbId  | description |
+-----+-----+
| A00142 | granulysin |
| A00146 | lyase, gastric |
| A03911 | seryne (or cysteine) proteinase inhibitor |
| A06977 | albumin |
| A12027 | S100 calcium binding protein A8 |
+-----+-----+
5 rows in set (0,00 sec)
```

Q3. What is the purpose of this query?

```
mysql> SELECT count(*) from LocusLinks;
```

It counts the total number of entries in the table *LocusLinks*.

```
+-----+
| COUNT(*) |
+-----+
|      22 |
+-----+
1 row in set (0,00 sec)
```

Q4. How many different *Affy* ids are in the expression data?

```
mysql> SELECT COUNT(affyId) FROM Data;
```

```
+-----+
| COUNT(affyId) |
+-----+
|          37 |
+-----+
1 row in set (0,00 sec)
```

Q5. What is the expression level of *Affy* id *U95-32123_at* in experiment number 1?

```
mysql> SELECT level FROM Data WHERE affyId="U95-32123_at" AND exptId=1;
```

```
+-----+
| level |
+-----+
|    128 |
+-----+
1 row in set (0,00 sec)
```

Q6. Find all the gene *descriptions*, along with their *GenBank* ids containing the word "*Human*"?

```
mysql> SELECT * FROM Descriptions WHERE description LIKE "%Human%";
```

```
+-----+-----+
| gbId  | description                                     |
+-----+-----+
| A12345 | HSLFBPS7 Human fructose-1, 6-biphosphatase      |
| A12346 | HSU30872 Human mitosin mRNA                     |
| A12347 | HSU33052 Human lipid-activated protein kinase   |
| A12348 | HSU33053 Human lipid-activated protein kinase   |
| A12349 | Human clone lambda 5 semaphorin mRNA            |
| A22124 | Human rearranged immunoglobulin lambda light chain mRNA |
| A22127 | Human rearranged immunoglobulin lambda light chain mRNA |
+-----+-----+
7 rows in set (0,00 sec)
```

Q7. What *Gene Ontology descriptions* (and corresponding *accession*) contain the phrase "*protein kinase*"? Answer should be provided in ascending order of accessions.

```
mysql> SELECT * FROM GO_Descr WHERE description LIKE "%protein kinase%"
mysql> ORDER BY goAcc ASC;
```

```
+-----+-----+
| goAcc  | description                                     |
+-----+-----+
| 0001236 | protein kinase |
| 0001237 | protein kinase |
| 1112222 | protein kinase |
| 4442222 | protein kinase |
+-----+-----+
4 rows in set (0,00 sec)
```

Q8. Which *AffyId* of table *Data* correspond to sequences in *Targets* table with the phrase "*kinase*" in their *description*?

```
mysql> SELECT Data.affyId FROM Data, Targets, Descriptions
mysql> WHERE Data.affyId=Targets.affyId
mysql> AND Targets.gbId=Descriptions.gbId
mysql> AND Descriptions.description LIKE "%kinase%";
```

```
Empty set (0,00 sec)
```

Use the following command:

```
LOAD DATA INFILE 'file.tsv' INTO TABLE Targets;
```

To add a new entry in *Descriptions* with the string "kinase" and the *gbId*="M18228". Now repeat the query again.

Repeating the same query again is still showing 0 results. There is an *affyId* corresponding to the *gbId* that we have introduced (M18228), but that *affyId* is not in the *Data* table.

```
Empty set (0,00 sec)
```

Q9. Get two *affyId*, *uId* and *descriptions* in *LocusDescr* in reverse alphabetical order of *descriptions*.

```
mysql> SELECT Targets.affyId, UniSeqs.uId, LocusDescr.description
mysql> FROM LocusDescr, LocusLinks, Targets, UniSeqs
mysql> WHERE UniSeqs.gbId=LocusLinks.gbId AND Targets.gbId=LocusLinks.gbId
mysql> AND LocusLinks.linkId=LocusDescr.linkId ORDER BY description DESC LIMIT 2;
```

```
+-----+-----+-----+
| affyId      | uId      | description |
+-----+-----+-----+
| U95_40474_at | Hs1691   | Glucan      |
| U95_32123_at | Hs1640   | Collagen    |
+-----+-----+-----+
2 rows in set (0,01 sec)
```

Q10. How would you find the average expression level of each experiment in *Data*?

```
mysql> SELECT exptId, AVG(level) FROM Data GROUP BY exptId;
```

```
+-----+-----+
| exptId | AVG(level) |
+-----+-----+
| 1      | 125.3333   |
| 2      | 95.3333    |
| 3      | 126.3333   |
| 4      | 83.5000    |
| 5      | 92.7500    |
| 6      | 18.3333    |
| 7      | 20.0000    |
| 8      | 40.0000    |
| 9      | 20.0000    |
+-----+-----+
9 rows in set (0,00 sec)
```

Q11. What is the average expression level of each array probe (*affyId*) across all experiments?

```
mysql> SELECT affyId, AVG(level) FROM Data GROUP BY affyId;
```

affyId	AVG(level)
31315_at	250.0000
31324_at	91.0000
31325_at	89.0000
31356_at	91.0000
31362_at	260.0000
31510_s_at	257.0000
5321_at	90.0000
5322_at	90.0000
5323_at	90.0000
5324_at	73.5000
5325_at	90.0000
AFEX-BioB-3_at	97.0000
AFEX-BioB-5_at	20.0000
AFEX-BioB-M_at	62.8000
AFEX-HSAC07/X00351_M_at	86.0000
AFEX-HUMBAPDH/M33197_3_st	277.0000
AFEX-HUMTFFR/M11507_at	90.0000
AFEX-M27830_3_at	271.0000
AFEX-MurIL10_at	6.6667
AFEX-MurIL2_at	20.0000
AFEX-MurIL4_at	49.0000
U95-32123_at	128.0000
U98-40474_at	57.0000

23 rows in set (0.00 sec)

Q12. What is the purpose of the following query?

```
mysql> SELECT Data.affyId, Data.level, Data.exptId, DataCopy.affyId,
mysql> DataCopy.level, DataCopy.exptId
mysql> FROM Data, Data DataCopy
mysql> WHERE Data.level > 10 * DataCopy.level
mysql> AND Data.affyId=DataCopy.affyId
mysql> AND Data.affyId LIKE "AFEX%",
mysql> LIMIT 10;
```

From the table called "*Data*" it takes the entries with an *affyId* beginning with the string "*AFEX*". Then, it selects those whose levels are ten times higher than the levels of other experiments of the same *affyId*. Finally, it shows up the first ten matches.

affyId	level	exptId	affyId	level	exptId
AFFX-BioB-M_at	214	5	AFFX-BioB-M_at	20	3
AFFX-BioB-M_at	214	5	AFFX-BioB-M_at	20	7
AFFX-BioB-M_at	214	5	AFFX-BioB-M_at	20	9

3 rows in set (0,00 sec)

Q13. Write a query to provide three different descriptions for all *gbId* in table *Targets*.

There are four tables containing descriptions. Therefore, there are 4 possible combinations and 4 possible outputs, depending on which three tables are chosen (*Descriptions*, *LocusDescr*, *GO_Descr* or *UniDescr*). One of the possibilities is the following one:

```
mysql> SELECT Targets.gbId, Descriptions.description AS General_Description,
mysql> LocusDescr.description AS Locus_Description,
mysql> GO_Descr.description AS GO_Description
mysql> FROM Targets, Descriptions, LocusDescr, LocusLinks, GO_Descr, Ontologies
mysql> WHERE Descriptions.gbId=Targets.gbId
mysql> AND Targets.gbId=LocusLinks.gbId AND LocusLinks.linkId=LocusDescr.linkId
mysql> AND LocusLinks.linkId=Ontologies.linkId
mysql> AND Ontologies.goAcc=GO_Descr.goAcc;
```

gbId	General_Description	Locus_Description	GO_Description
S75295	Glucan	Glucan	Glucan Enz

1 row in set (0,00 sec)

Q14. Write a query to provide all gene ontology (*GO_descr*) descriptions related with all species in table *Species* sorted alphabetically and providing the first five results. Export the query to a tab-separated-file with the command:

```
mysql> SELECT * FROM TABLE INTO OUTFILE 'data.out';

mysql> SELECT GO_Descr.description, Targets.species
mysql> FROM GO_Descr, Ontologies, LocusLinks, Targets
mysql> WHERE GO_Descr.goAcc=Ontologies.goAcc
mysql> AND Ontologies.linkId=LocusLinks.linkId
mysql> AND LocusLinks.gbId=Targets.gbId
mysql> ORDER BY GO_Descr.description ASC LIMIT 5
mysql> INTO OUTFILE '14.out';
```

```
+-----+-----+
| description                | species |
+-----+-----+
| extracellular space        | Hs      |
| fructose-2, 6-biophosphate 2-phosphatase | Hs      |
| Glucan Enz                 | Hs      |
| protein kinase             | Hs      |
| protein kinase             | Hs      |
+-----+-----+
5 rows in set (0,00 sec)
```