PES UNIVERSITY

ELECTIVE 1: DATABASE TECHNOLOGIES (UE18CS315)

ASSIGNMENT 3

NAME: SHAAZIN SHEIKH SHUKOOR

SRN: PES1201801754

SEMESTER: 5

SECTION: J

Write a simple Map-Reduce program

Map-Reduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment. Map-Reduce consists of two distinct tasks – Map and Reduce.

- As the name Map-Reduce suggests, the reducer phase takes place after the mapper phase has been completed.
- So, the first is the map job, where a block of data is read and processed to produce key-value pairs as intermediate outputs.
- The output of a Mapper or map job (key-value pairs) is input to the Reducer.
- The reducer receives the key-value pair from multiple map jobs.
- Then, the reducer aggregates those intermediate data tuples (intermediate key-value pair) into a smaller set of tuples or key-value pairs which is the final output.

Problem statement:

i. To print the count of number of books in the library which were published before 2000.
ii. To print the name of the latest publisher.

Input: A library dataset of csv format containing ISBN, Book-Title, Book-Author, Year-Of Publication, Publisher, Image-URL-S, Image-URL-M, and Image-URL-L as its fields.

Output: 2 lines; 1st line is the count of the number of books published before 2000 and 2nd line is the name of the latest publisher.

Mapper code:

```
#!/usr/bin/python3.8

import sys

first=1

for line1 in sys.stdin:
        for line2 in line1.splitlines():
                list1 = [x.strip('"') for x in line2.split(';')]
                if(first==1):
                        idx1=list1.index('Year-Of-Publication')
                        idx2=list1.index('Publisher')
                        first=0
                else:
                        print(list1[idx1],'\t',list1[idx2])
```

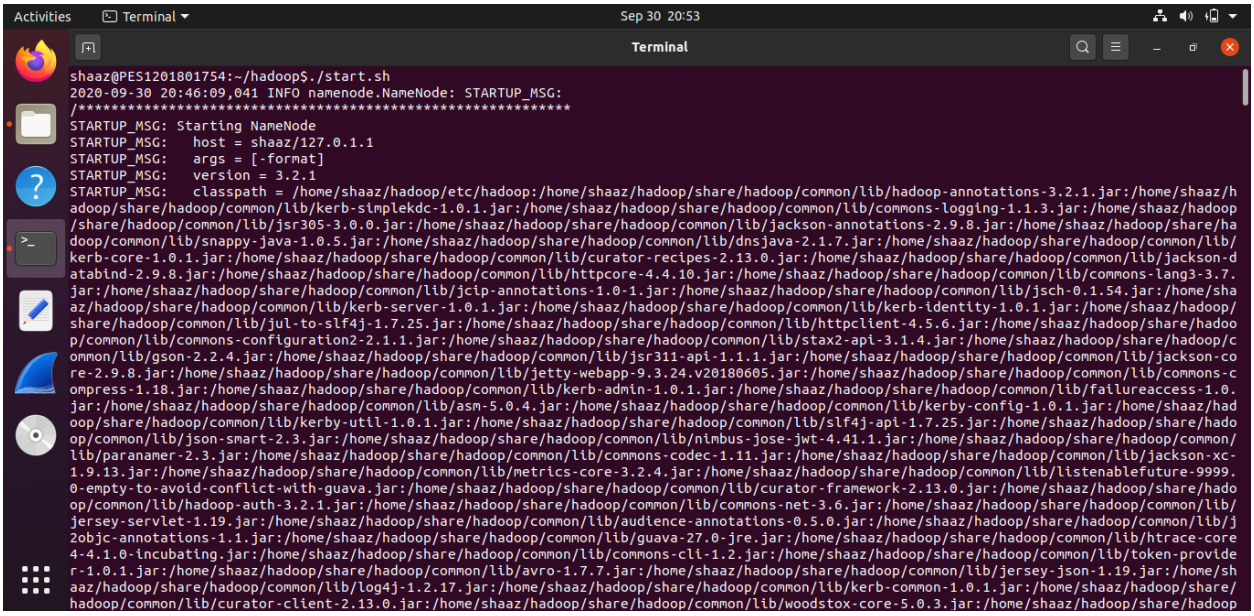
Reducer code:

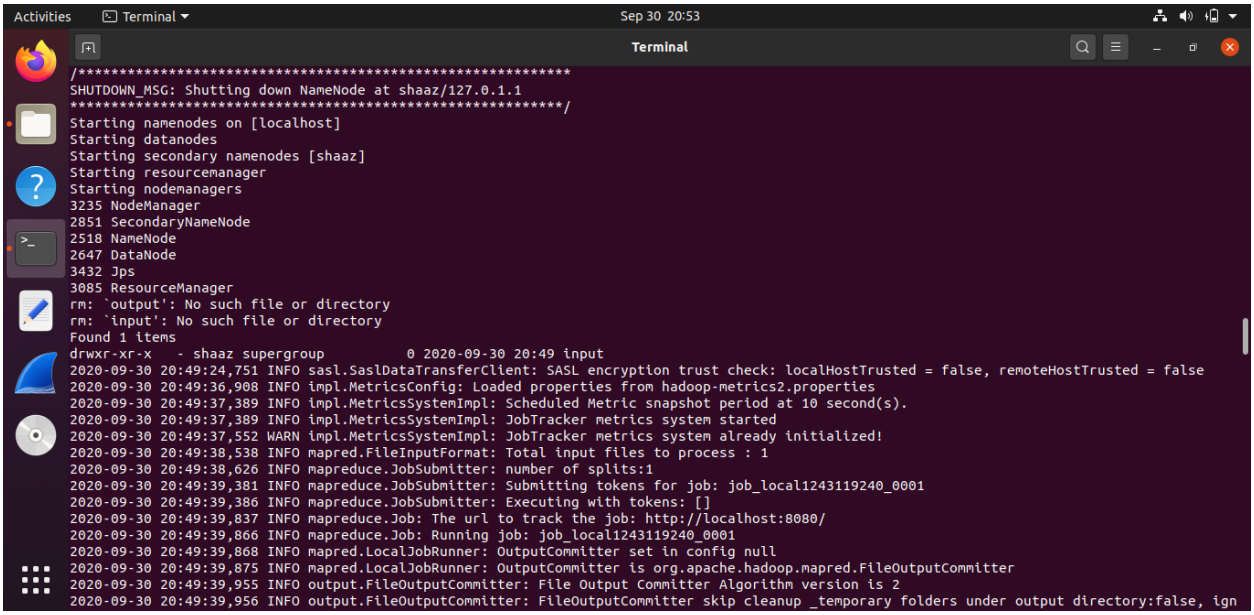
```
#!/usr/bin/python3.8

import sys

count=0

for line1 in sys.stdin:
        line1=line1.strip().split('\t',1)
        if(int(line1[0]) < 2000):
                count=count+1
        name=line1[1].strip()
print(count)
print(name)
```

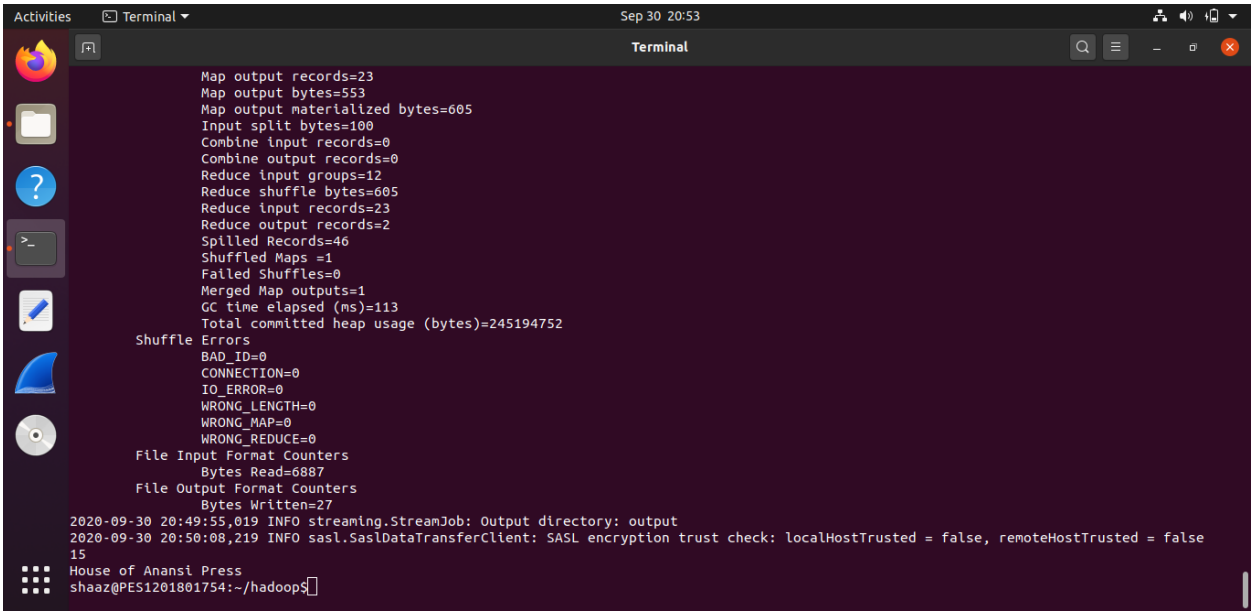
Running the map-reduce program in hadoop:

The commands for running the mapper and reducer and to print the contents of the output file are include in the bash script (start.sh).
(The bash script is ran using the command ./start.sh)

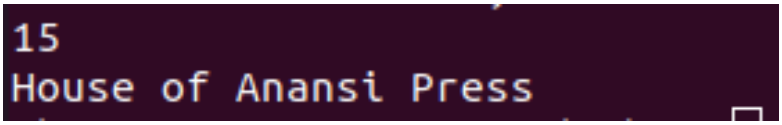
```
#!/bin/sh
bin/hdfs namenode -format
sbin/start-dfs.sh
sbin/start-yarn.sh
jps
bin/hdfs dfs -rm -r output
bin/hdfs dfs -rm -r input
bin/hdfs dfs -mkdir /user
bin/hdfs dfs -mkdir /user/shaaz
bin/hdfs dfs -mkdir input
bin/hdfs dfs -ls
bin/hdfs dfs -put '/home/shaaz/dbt3/input.csv' input
bin/hadoop jar '/home/shaaz/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar' -mapper '/home/shaaz/dbt3/mapper.py'
-reducer '/home/shaaz/dbt3/reducer.py' -input input/ -output output
bin/hdfs dfs -cat output/part-00000
```


Activities
Terminal
Sep 30 20:53
Terminal
shaaz@PES1201801754:~/hadoop$./start.sh
2020-09-30 20:46:09,041 INFO namenode.NameNode: STARTUP_MSG:
/**
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = shaaz/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.2.1
STARTUP_MSG: classpath = /home/shaaz/hadoop/etc/hadoop:/home/shaaz/hadoop/share/hadoop/common/lib/hadoop-annotations-3.2.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerb-simplekdc-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jsr305-3.0.0.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jackson-annotations-2.9.8.jar:/home/shaaz/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/home/shaaz/hadoop/share/hadoop/common/lib/dnsjava-2.1.7.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerb-core-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/curator-recipes-2.13.0.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jackson-databind-2.9.8.jar:/home/shaaz/hadoop/share/hadoop/common/lib/httpcore-4.4.10.jar:/home/shaaz/hadoop/share/hadoop/common/lib/commons-lang3-3.7.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jcip-annotations-1.0-1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jsch-0.1.54.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerb-server-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jul-to-slf4j-1.7.25.jar:/home/shaaz/hadoop/share/hadoop/common/lib/httpclient-4.5.6.jar:/home/shaaz/hadoop/share/hadoop/common/lib/commons-configuration2-2.1.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/stax2-api-3.1.4.jar:/home/shaaz/hadoop/share/hadoop/common/lib/gson-2.2.4.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jsr311-api-1.1.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jackson-core-2.9.8.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jetty-webapp-9.3.24.v20180605.jar:/home/shaaz/hadoop/share/hadoop/common/lib/commons-compress-1.18.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerb-admin-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/failureaccess-1.0.jar:/home/shaaz/hadoop/share/hadoop/common/lib/asm-5.0.4.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerby-config-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerby-util-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/slf4j-api-1.7.25.jar:/home/shaaz/hadoop/share/hadoop/common/lib/json-smart-2.3.jar:/home/shaaz/hadoop/share/hadoop/common/lib/nimbus-jose-jwt-4.41.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/paranamer-2.3.jar:/home/shaaz/hadoop/share/hadoop/common/lib/commons-codec-1.11.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/home/shaaz/hadoop/share/hadoop/common/lib/metrics-core-3.2.4.jar:/home/shaaz/hadoop/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/shaaz/hadoop/share/hadoop/common/lib/curator-framework-2.13.0.jar:/home/shaaz/hadoop/share/hadoop/common/lib/hadoop-auth-3.2.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/commons-net-3.6.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jersey-servlet-1.19.jar:/home/shaaz/hadoop/share/hadoop/common/lib/audience-annotations-0.5.0.jar:/home/shaaz/hadoop/share/hadoop/common/lib/j2objc-annotations-1.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/guava-27.0-jre.jar:/home/shaaz/hadoop/share/hadoop/common/lib/htrace-core4-4.1.0-incubating.jar:/home/shaaz/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar:/home/shaaz/hadoop/share/hadoop/common/lib/token-provider-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/avro-1.7.7.jar:/home/shaaz/hadoop/share/hadoop/common/lib/jersey-json-1.19.jar:/home/shaaz/hadoop/share/hadoop/common/lib/log4j-1.2.17.jar:/home/shaaz/hadoop/share/hadoop/common/lib/kerb-common-1.0.1.jar:/home/shaaz/hadoop/share/hadoop/common/lib/curator-client-2.13.0.jar:/home/shaaz/hadoop/share/hadoop/common/lib/woodstox-core-5.0.3.jar:/home/shaaz/hadoop/share/hadoop


Activities
Terminal
Sep 30 20:53
Terminal
/**
SHUTDOWN_MSG: Shutting down NameNode at shaaz/127.0.1.1
**/
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [shaaz]
Starting resourcemanager
Starting nodemanagers
3235 NodeManager
2851 SecondaryNameNode
2518 NameNode
2647 DataNode
3432 Jps
3085 ResourceManager
rm: `output': No such file or directory
rm: `input': No such file or directory
Found 1 items
drwxr-xr-x - shaaz supergroup 0 2020-09-30 20:49 input
2020-09-30 20:49:24,751 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-09-30 20:49:36,908 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-09-30 20:49:37,389 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-09-30 20:49:37,389 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-09-30 20:49:37,552 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2020-09-30 20:49:38,538 INFO mapred.FileInputFormat: Total input files to process : 1
2020-09-30 20:49:38,626 INFO mapreduce.JobSubmitter: number of splits:1
2020-09-30 20:49:39,381 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1243119240_0001
2020-09-30 20:49:39,386 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-09-30 20:49:39,837 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2020-09-30 20:49:39,866 INFO mapreduce.Job: Running job: job_local1243119240_0001
2020-09-30 20:49:39,868 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2020-09-30 20:49:39,875 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2020-09-30 20:49:39,955 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-09-30 20:49:39,956 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ign



The cat command at the end prints the output of the map-reduce program in the terminal (output of the reducer):



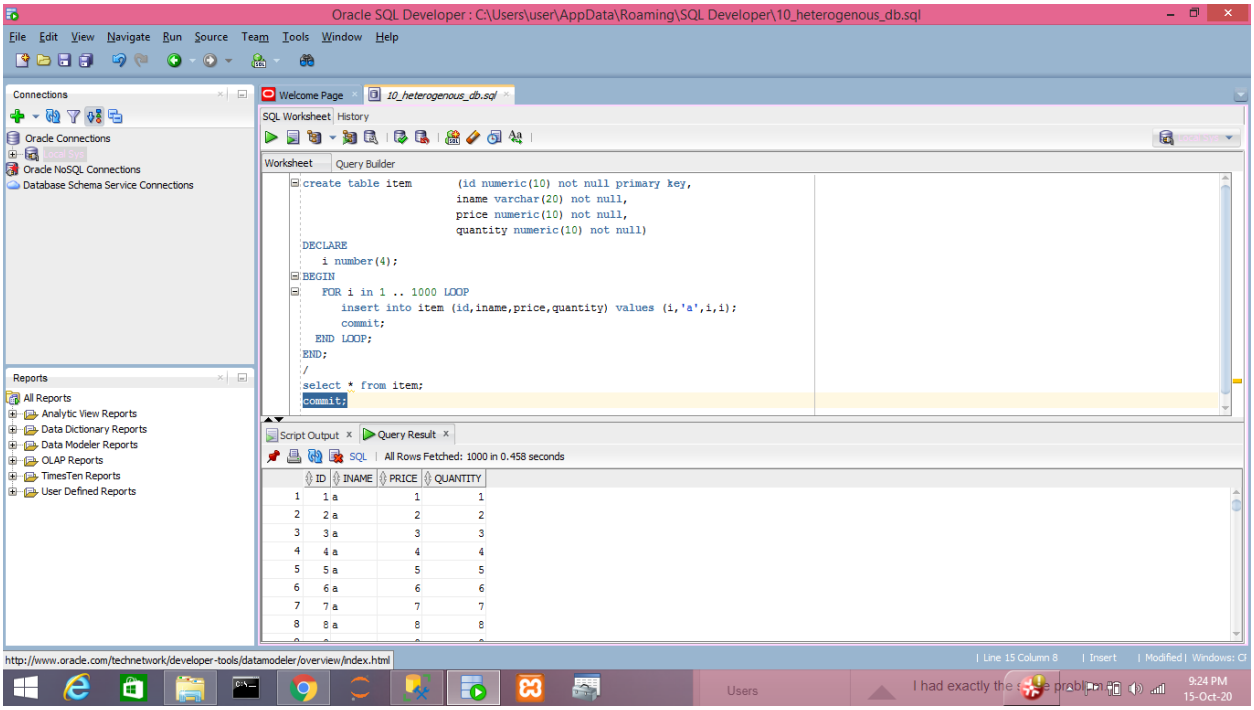
The intermediate key-value pair of the mapper (sorts it internally):

```
shaaz@PES1201801754:~/dbt3$cat output_mapper
1988    Emblem Editions
1991    HarperPerennial
1991    Putnam Pub Group
1993    Audioworks
1994    Plume
1994    River City Pub
1996    Random House
1998    Brilliance Audio - Trade
1998    Health Communications
1999    Dell
1999    Farrar Straus Giroux
1999    Kensington Publishing Corp.
1999    Mira Books
1999    Three Rivers Press
1999    W. W. Norton &amp
2000    Berkley Publishing Group
2000    Citadel Press
2001    HarperFlamingo Canada
2001    Ryland Peters &amp
2002    Oxford University Press
2002    Scribner
2003    Cypress House
2004    House of Anansi Press
shaaz@PES1201801754:~/dbt3$
```

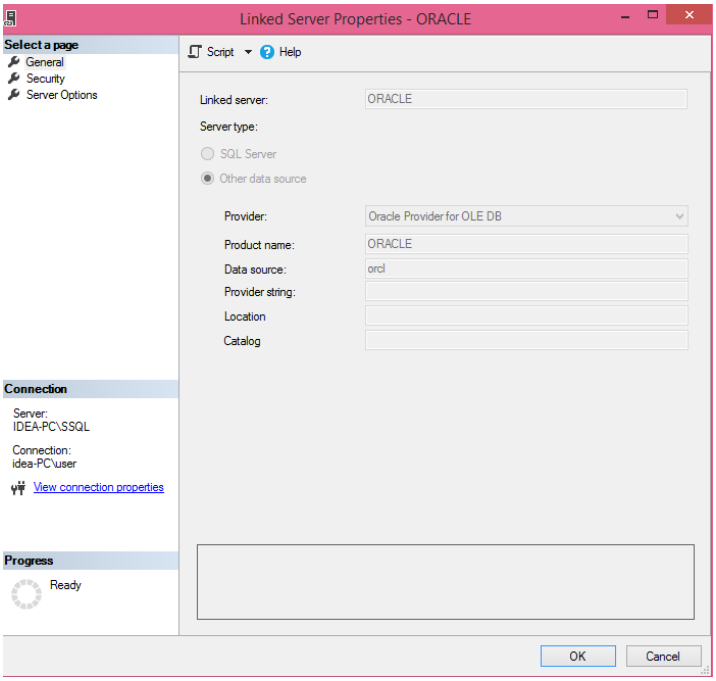
Demonstrate a heterogeneous distributed query.

A distributed database is a set of interconnected databases that is distributed over the computer network or internet. A Distributed Database Management System (DDBMS) manages the distributed database and provides mechanisms so as to make the databases transparent to the users.

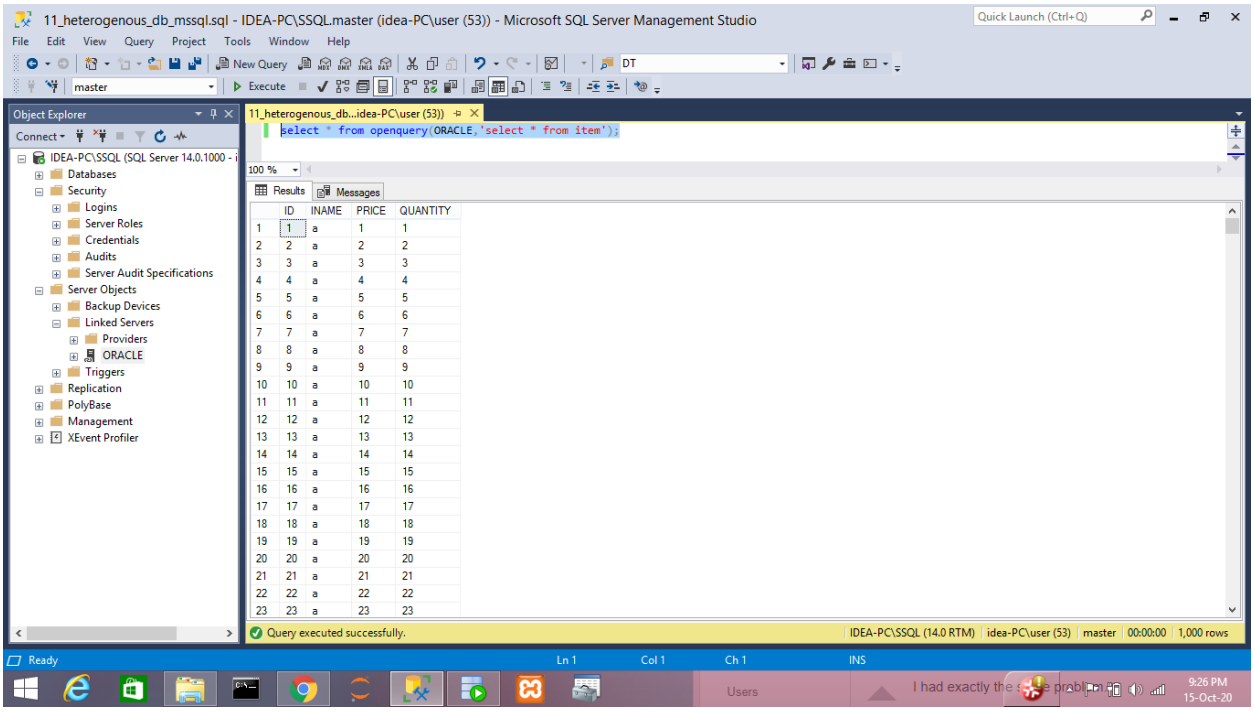
Item table located in oracle:



A linked server is created in Microsoft SQL Server Management Studio by using the provider ‘Oracle Provider for OLE DB’



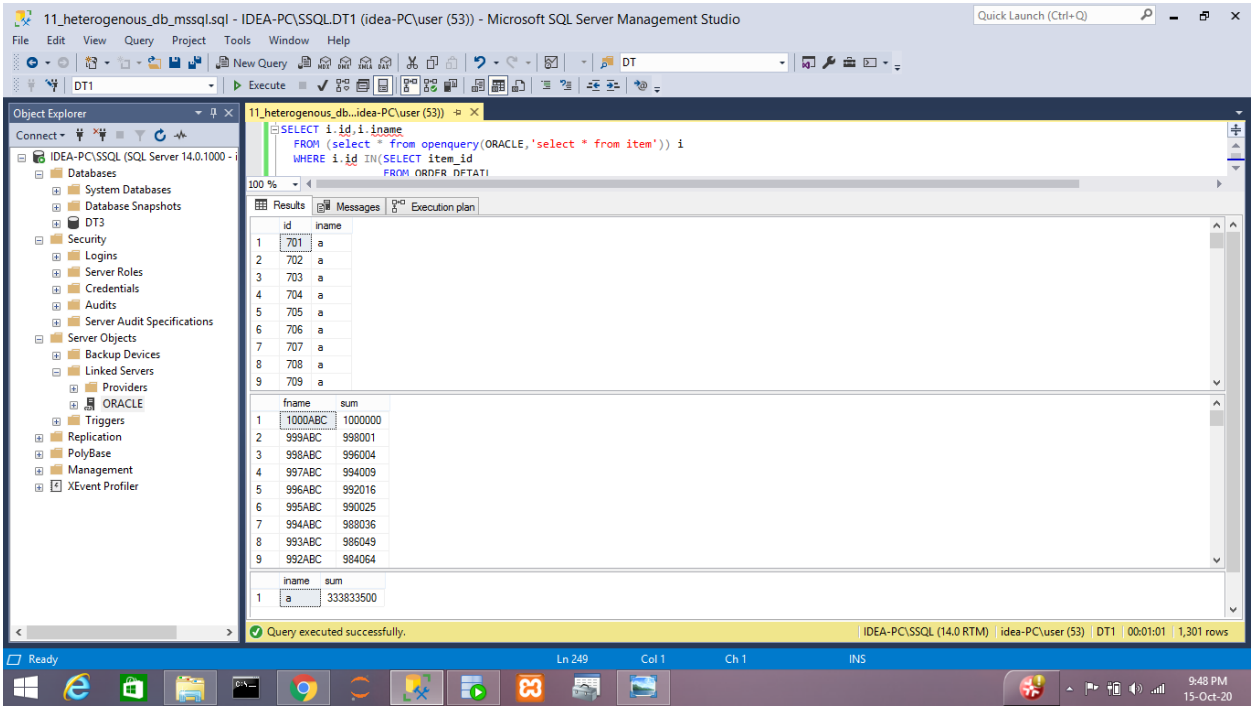
The contents of table ‘item’ can be viewed using openquery:



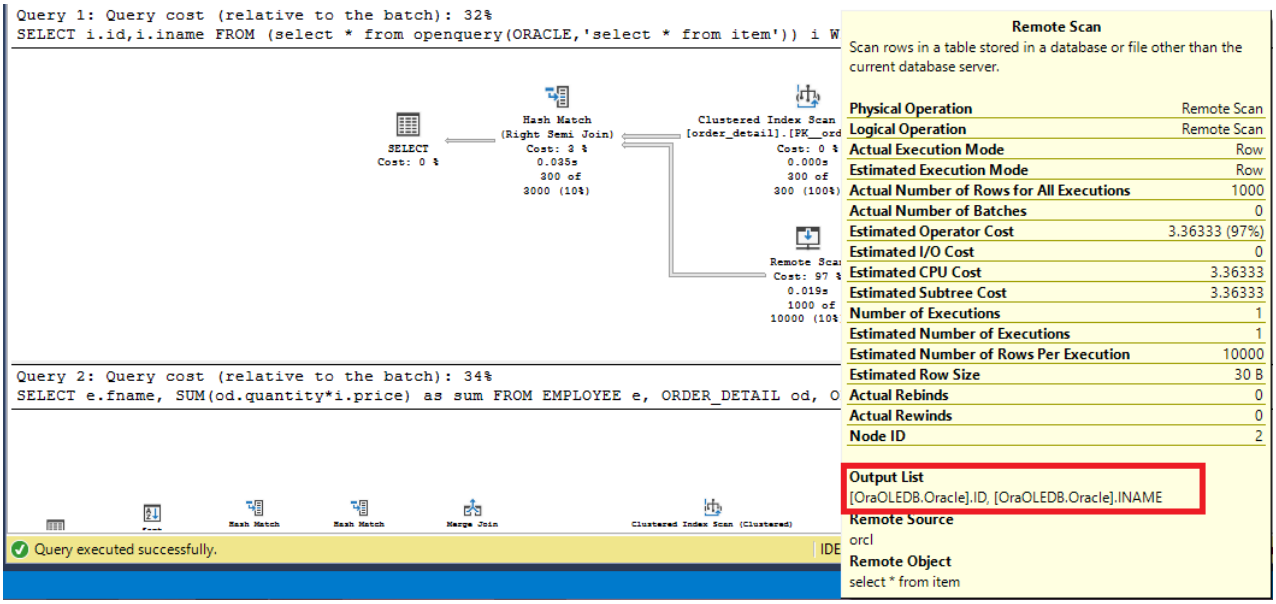
Queries using the item table:

```
commit transaction T6

--query
SET STATISTICS IO ON
SET STATISTICS TIME ON

SELECT i.id,i.iname
  FROM (select * from openquery(ORACLE,'select * from item')) i
  WHERE i.id IN(SELECT item_id
                FROM ORDER_DETAIL
                WHERE quantity>700
               );

SELECT e.fname, SUM(od.quantity*i.price) as sum
  FROM EMPLOYEE e, ORDER_DETAIL od, ORDERS o, (select * from openquery(ORACLE,'select * from item')) i
  WHERE e.id=o.emp_id AND o.id=od.order_id AND od.item_id=i.id
  GROUP BY e.fname
  ORDER BY SUM(od.quantity*i.price) DESC

SELECT top 3 i.iname,SUM(od.quantity*i.price) as sum
  FROM ORDER_DETAIL AS od LEFT OUTER JOIN (select * from openquery(ORACLE,'select * from item')) AS i ON od.item_id=i.id
  GROUP BY i.iname
  ORDER BY SUM(od.quantity*i.price) ASC

SET STATISTICS IO OFF
SET STATISTICS TIME OFF
```

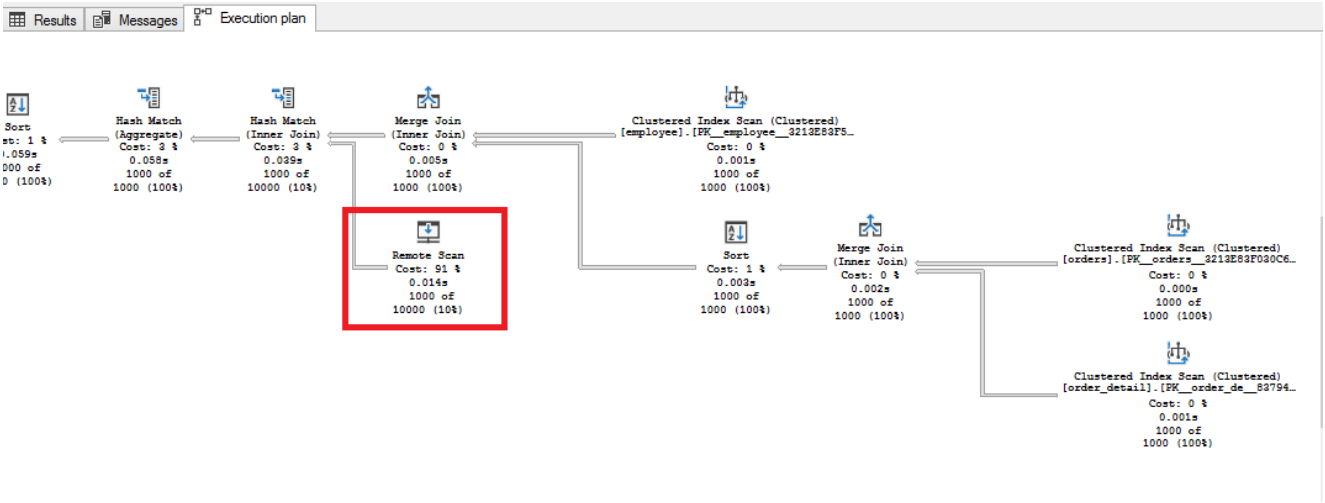
Output of the above queries:



We see that in the first query, only those rows are retrieved from the ‘item’ table that needs to be projected in the output:

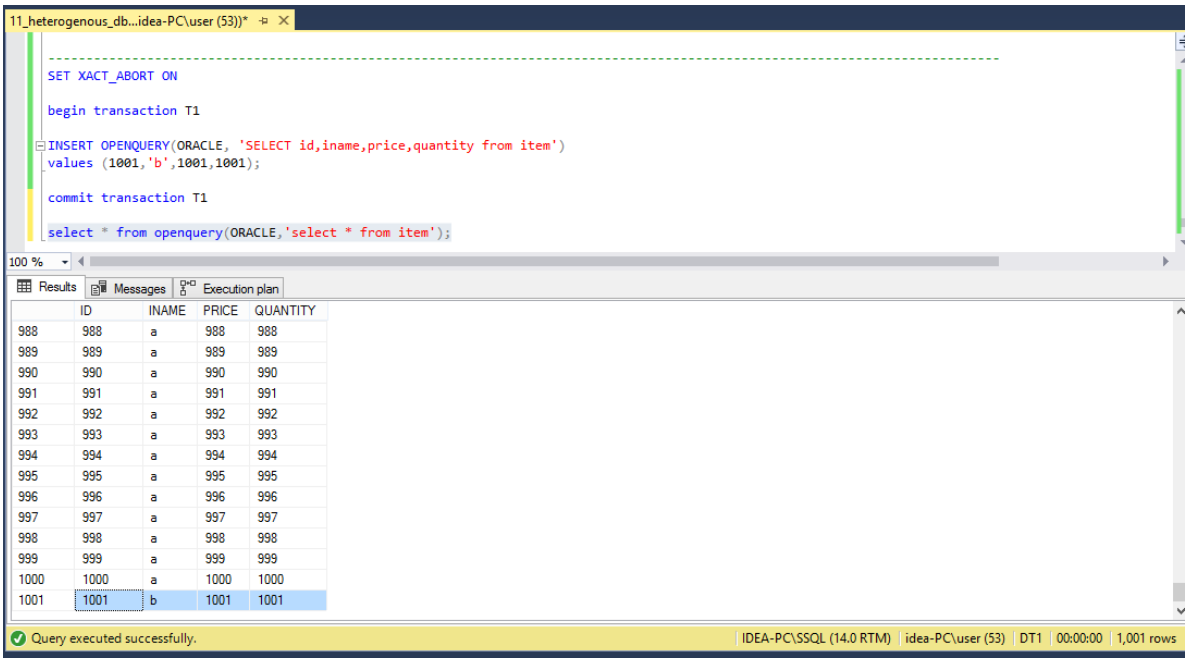


One of the disadvantages of distributed database is that the query optimizer is not so efficient. Below is an example showing the order of join of tables. We see that the ‘item’ table is joined at the end:



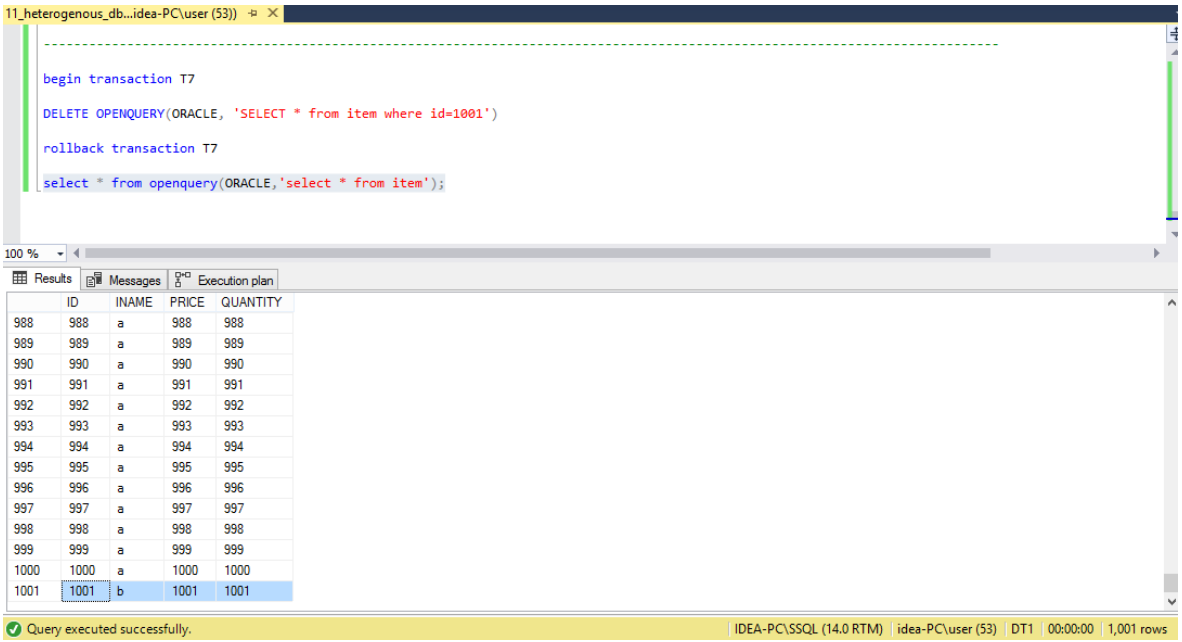
Use of transaction and commit:

A record is inserted using ‘insert openquery’. We see that the changes get reflection using ‘commit’.



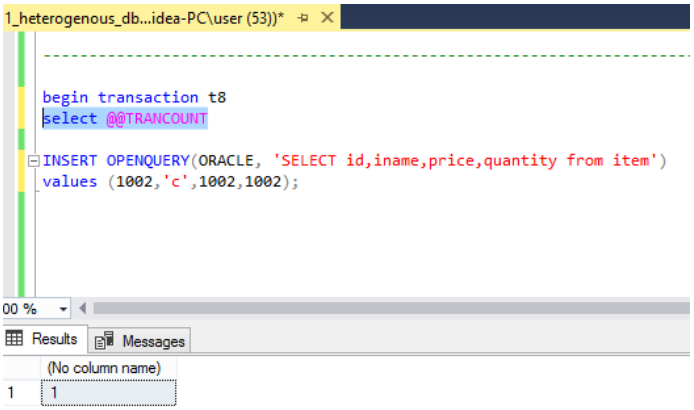
Use of transaction and rollback:

A record is deleted using ‘delete openquery’. We see that the change of the transaction does not get reflected using ‘rollback’.

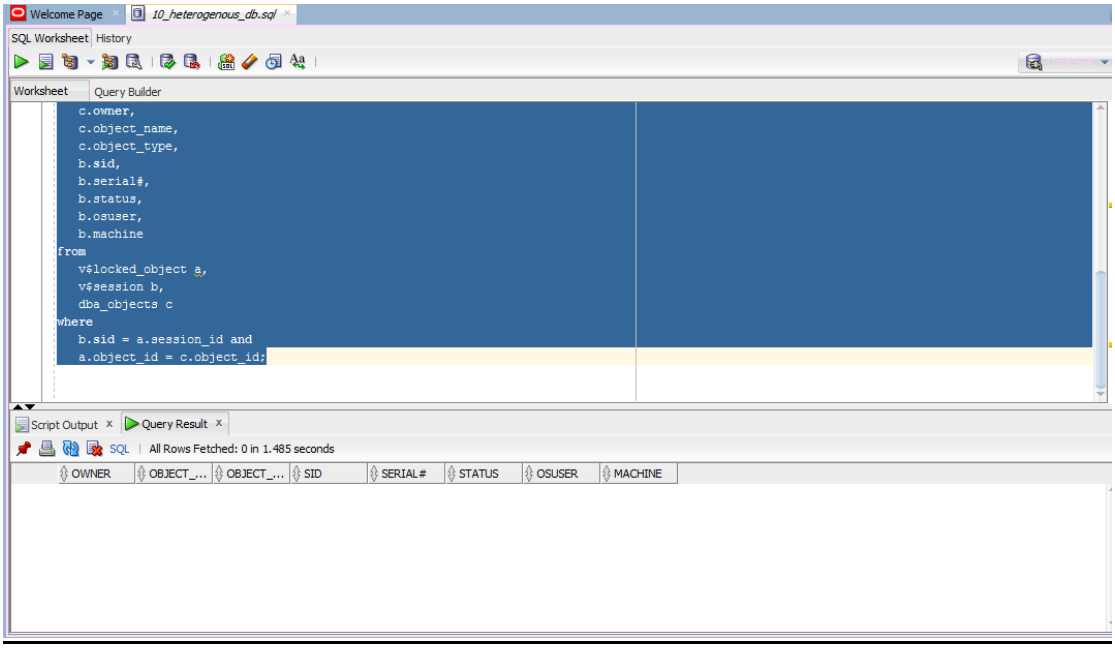


Nested transactions:

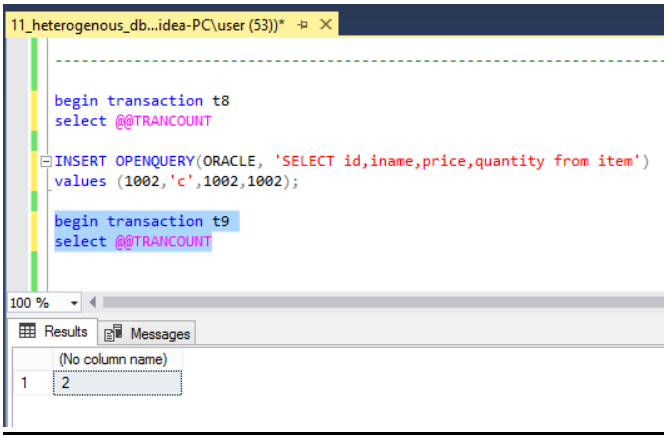
Begin the outer transaction ‘t8’:

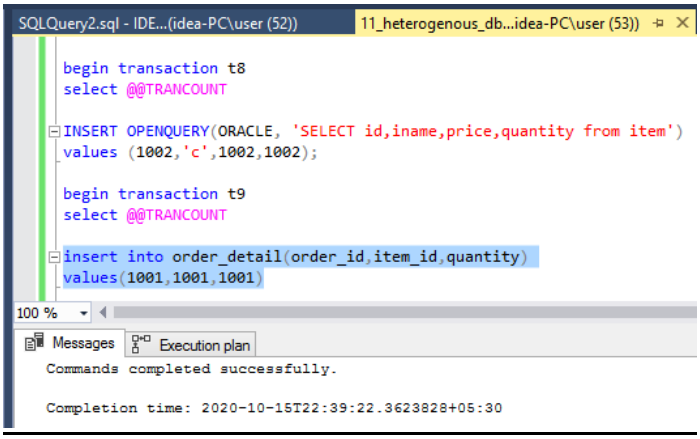


Currently there are no locks in oracle:

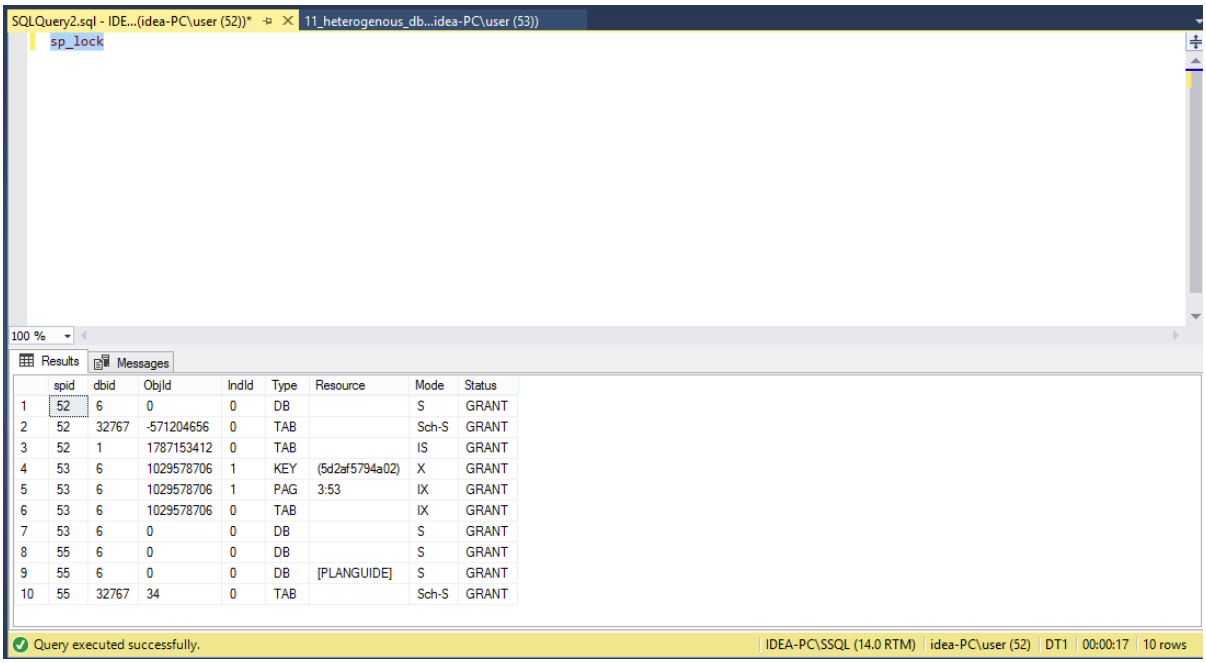


Begin the inner transaction ‘t9’:



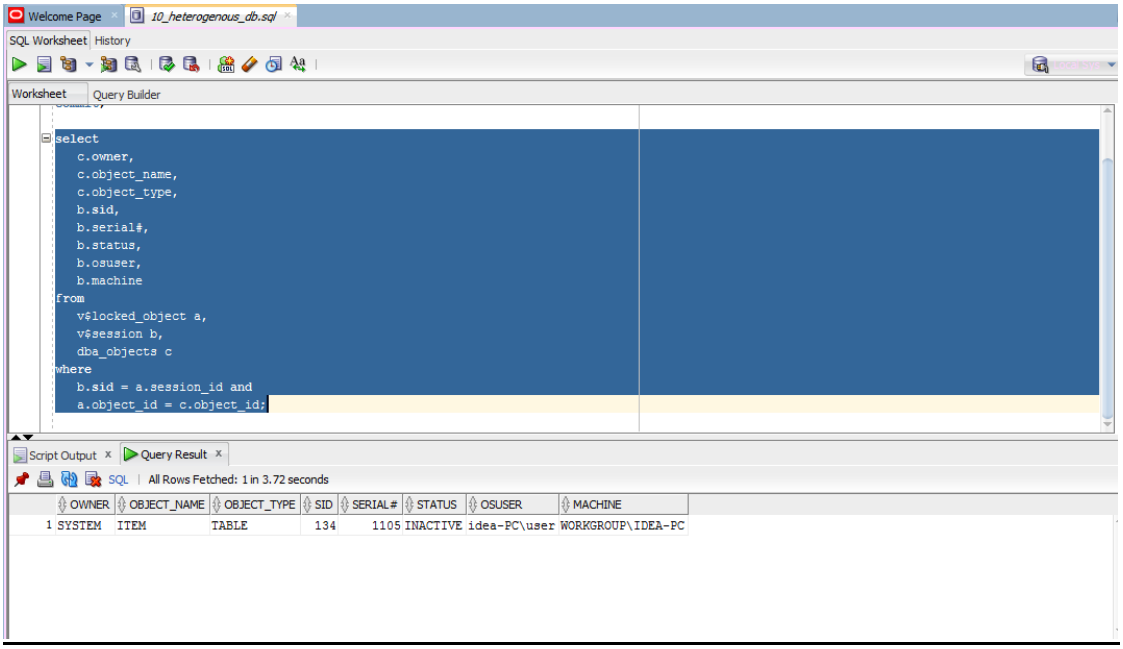


We see a lock on (53) in MS SQL:

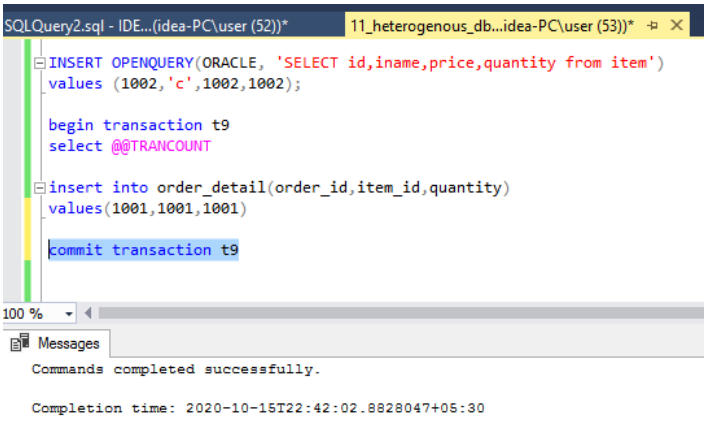


	spid	dbid	ObjId	IndId	Type	Resource	Mode	Status
1	52	6	0	0	DB		S	GRANT
2	52	32767	-571204656	0	TAB		Sch-S	GRANT
3	52	1	1787153412	0	TAB		IS	GRANT
4	53	6	1029578706	1	KEY	(5d2af5794a02)	X	GRANT
5	53	6	1029578706	1	PAG	3:53	IX	GRANT
6	53	6	1029578706	0	TAB		IX	GRANT
7	53	6	0	0	DB		S	GRANT
8	55	6	0	0	DB		S	GRANT
9	55	6	0	0	DB	[PLANGUIDE]	S	GRANT
10	55	32767	34	0	TAB		Sch-S	GRANT

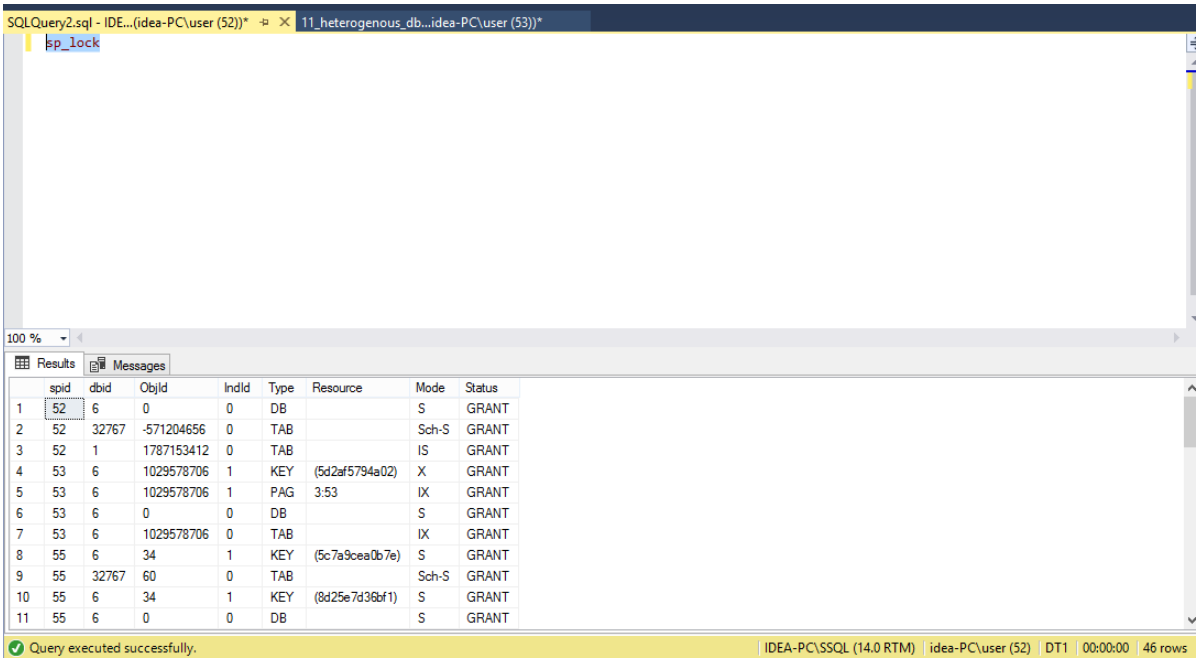
We also see a lock in oracle:



We commit the inner transaction ‘t9’:

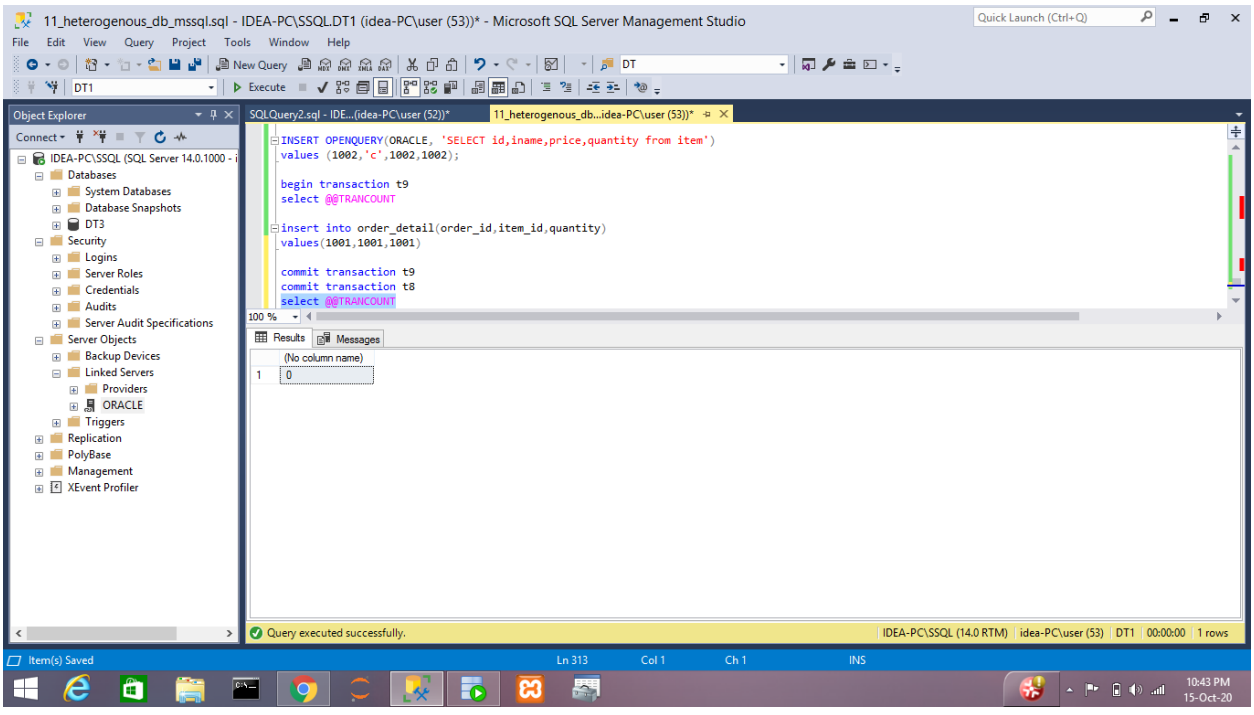


We still find the locks on (53) in MS SQL:

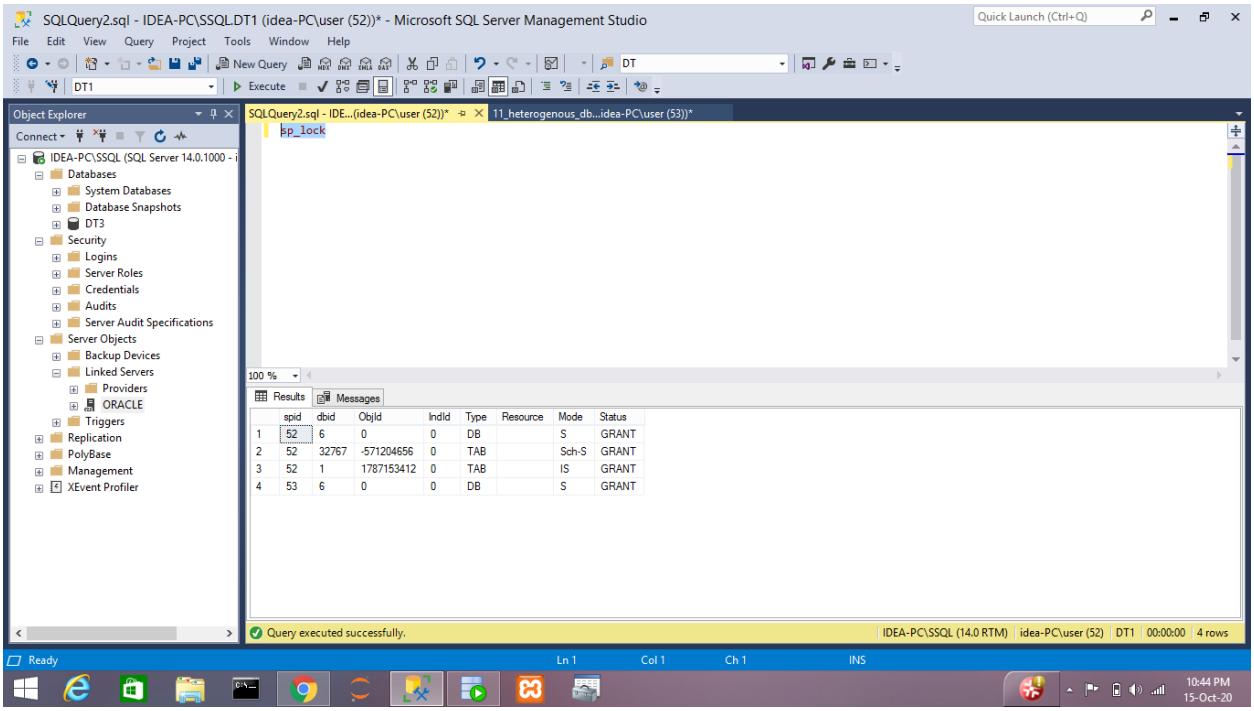


	spid	dbid	ObjId	IndId	Type	Resource	Mode	Status
1	52	6	0	0	DB		S	GRANT
2	52	32767	-571204656	0	TAB		Sch-S	GRANT
3	52	1	1787153412	0	TAB		IS	GRANT
4	53	6	1029578706	1	KEY	(5d2af5794a02)	X	GRANT
5	53	6	1029578706	1	PAG	3:53	IX	GRANT
6	53	6	0	0	DB		S	GRANT
7	53	6	1029578706	0	TAB		IX	GRANT
8	55	6	34	1	KEY	(5c7a9cea0b7e)	S	GRANT
9	55	32767	60	0	TAB		Sch-S	GRANT
10	55	6	34	1	KEY	(8d25e7d36bf1)	S	GRANT
11	55	6	0	0	DB		S	GRANT

We commit the outer transaction ‘t8’:



The locks are removed:



The locks are also removed in oracle: