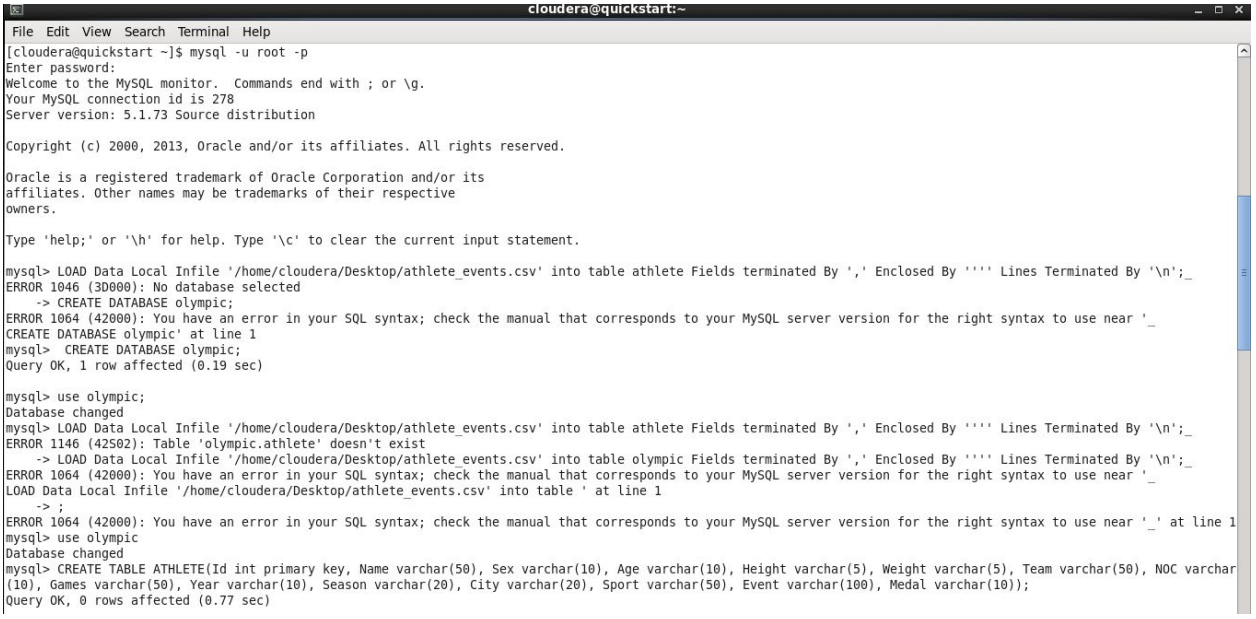
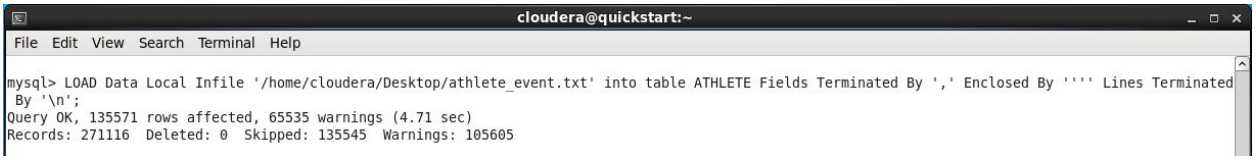
PROGRAM CODE:

LOADING DATA IN HADOOP, CREATING DATABASE AND TABLE:

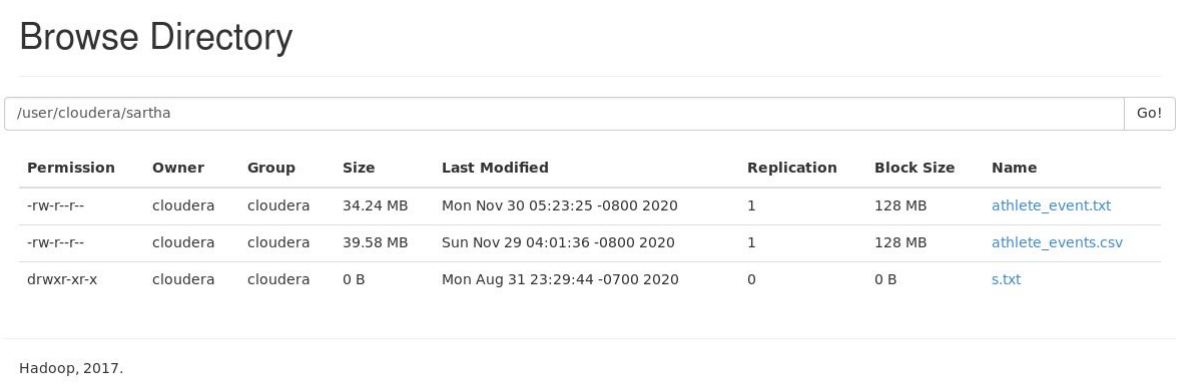


```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 278
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> LOAD Data Local Infile '/home/cloudera/Desktop/athlete_events.csv' into table athlete Fields terminated By ',' Enclosed By ''' Lines Terminated By '\n';_
ERROR 1046 (3D000): No database selected
    -> CREATE DATABASE olympic;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '_
CREATE DATABASE olympic' at line 1
mysql>  CREATE DATABASE olympic;
Query OK, 1 row affected (0.19 sec)

mysql> use olympic;
Database changed
mysql> LOAD Data Local Infile '/home/cloudera/Desktop/athlete_events.csv' into table athlete Fields terminated By ',' Enclosed By ''' Lines Terminated By '\n';_
ERROR 1146 (42S02): Table 'olympic.athlete' doesn't exist
    -> LOAD Data Local Infile '/home/cloudera/Desktop/athlete_events.csv' into table olympic Fields terminated By ',' Enclosed By ''' Lines Terminated By '\n';_
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '_
LOAD Data Local Infile '/home/cloudera/Desktop/athlete_events.csv' into table ' at line 1
    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '_' at line 1
mysql> use olympic
Database changed
mysql> CREATE TABLE ATHLETE(Id int primary key, Name varchar(50), Sex varchar(10), Age varchar(10), Height varchar(5), Weight varchar(5), Team varchar(50), NOC varchar
(10), Games varchar(50), Year varchar(10), Season varchar(20), City varchar(20), Sport varchar(50), Event varchar(100), Medal varchar(10));
Query OK, 0 rows affected (0.77 sec)
```



```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help

mysql> LOAD Data Local Infile '/home/cloudera/Desktop/athlete_event.txt' into table ATHLETE Fields Terminated By ',' Enclosed By ''' Lines Terminated
 By '\n';
Query OK, 135571 rows affected, 65535 warnings (4.71 sec)
Records: 271116  Deleted: 0  Skipped: 135545  Warnings: 105605
```

```
mysql> SELECT * from ATHLETE limit 10;
+----+-----------------------------------+-----+-----+--------+--------+---------------+-----+------------+------+--------+---------------+---------------------+-----------------------------------------+---
-----+
| Id | Name                              | Sex | Age | Height | Weight | Team          | NOC | Games      | Year | Season | City          | Sport               | Event                                   | M
edal |
+----+-----------------------------------+-----+-----+--------+--------+---------------+-----+------------+------+--------+---------------+---------------------+-----------------------------------------+---
-----+
|  1 | A Dijiang                         | M   | 24  | 180    | 80     | China         | CHN | 1992 Summer | 1992 | Summer | Barcelona     | Basketball          | Basketball Men's Basketball             | N
   |
|  2 | A Lamusi                          | M   | 23  | 170    | 60     | China         | CHN | 2012 Summer | 2012 | Summer | London        | Judo                | Judo Men's Extra-Lightweight            | N
   |
|  3 | Gunnar Nielsen Aaby               | M   | 24  | NA     | NA     | Denmark       | DEN | 1920 Summer | 1920 | Summer | Antwerpen     | Football            | Football Men's Football                 | N
   |
|  4 | Edgar Lindenau Aabye              | M   | 34  | NA     | NA     | Denmark/Sweden | DEN | 1900 Summer | 1900 | Summer | Paris         | Tug-Of-War          | Tug-Of-War Men's Tug-Of-War             | G
 |d
|  5 | Christine Jacoba Aaftink          | F   | 21  | 185    | 82     | Netherlands   | NED | 1988 Winter | 1988 | Winter | Calgary       | Speed Skating       | Speed Skating Women's 500 metres        | N
   |
|  6 | Per Knut Aaland                   | M   | 31  | 188    | 75     | United States | USA | 1992 Winter | 1992 | Winter | Albertville   | Cross Country Skiing | Cross Country Skiing Men's 10 kilometres | N
   |
|  7 | John Aalberg                      | M   | 31  | 183    | 72     | United States | USA | 1992 Winter | 1992 | Winter | Albertville   | Cross Country Skiing | Cross Country Skiing Men's 10 kilometres | N
   |
|  8 | Cornelia Cor Aalten (-Strannood) | F   | 18  | 168    | NA     | Netherlands   | NED | 1932 Summer | 1932 | Summer | Los Angeles   | Athletics           | Athletics Women's 100 metres            | N
   |
|  9 | Antti Sami Aalto                  | M   | 26  | 186    | 96     | Finland       | FIN | 2002 Winter | 2002 | Winter | Salt Lake City | Ice Hockey          | Ice Hockey Men's Ice Hockey             | N
   |
| 10 | Einar Ferdinand Einari Aalto      | M   | 26  | NA     | NA     | Finland       | FIN | 1952 Summer | 1952 | Summer | Helsinki      | Swimming            | Swimming Men's 400 metres Freestyle     | N
   |
+----+-----------------------------------+-----+-----+--------+--------+---------------+-----+------------+------+--------+---------------+---------------------+-----------------------------------------+---
-----+
10 rows in set (0.01 sec)
```



Browse Directory

/user/cloudera/sartha Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	cloudera	34.24 MB	Mon Nov 30 05:23:25 -0800 2020	1	128 MB	athlete_event.txt
-rw-r--r--	cloudera	cloudera	39.58 MB	Sun Nov 29 04:01:36 -0800 2020	1	128 MB	athlete_events.csv
drwxr-xr-x	cloudera	cloudera	0 B	Mon Aug 31 23:29:44 -0700 2020	0	0 B	s.txt

Hadoop, 2017.

IMPORTING TABLE FROM HDFS TO HIVE:

```
[cloudera@quickstart ~]$ sqoop import-all-tables --connect jdbc:mysql://localhost/olympic --username=root --password=cloudera --compression-codec=snap
py --as-parquetfile --warehouse-dir=/user/hive/warehouse --hive-import
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/11/30 05:37:14 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
20/11/30 05:37:14 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/11/30 05:37:14 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
20/11/30 05:37:14 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
20/11/30 05:37:14 WARN tool.BaseSqoopTool: It seems that you're doing hive import directly into default
20/11/30 05:37:14 WARN tool.BaseSqoopTool: hive warehouse directory which is not supported. Sqoop is
20/11/30 05:37:14 WARN tool.BaseSqoopTool: firstly importing data into separate directory and then
20/11/30 05:37:14 WARN tool.BaseSqoopTool: inserting data into hive. Please consider removing
20/11/30 05:37:14 WARN tool.BaseSqoopTool: --target-dir or --warehouse-dir into /user/hive/warehouse in
20/11/30 05:37:14 WARN tool.BaseSqoopTool: case that you will detect any issues.
20/11/30 05:37:14 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
20/11/30 05:37:15 INFO tool.CodeGenTool: Beginning code generation
20/11/30 05:37:15 INFO tool.CodeGenTool: Will generate java class as codegen_ATHLETE
20/11/30 05:37:15 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ATHLETE` AS t LIMIT 1
20/11/30 05:37:15 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ATHLETE` AS t LIMIT 1
20/11/30 05:37:15 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/e1461a7685148e0eb2889895c12317cf/codegen_ATHLETE.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/11/30 05:37:20 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/e1461a7685148e0eb2889895c12317cf/codegen_ATHLETE.jar
20/11/30 05:37:20 WARN manager.MySQLManager: It looks like you are importing from mysql.
20/11/30 05:37:20 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
20/11/30 05:37:20 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
20/11/30 05:37:20 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
20/11/30 05:37:20 INFO mapreduce.ImportJobBase: Beginning import of ATHLETE
20/11/30 05:37:20 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
20/11/30 05:37:21 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
20/11/30 05:37:22 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ATHLETE` AS t LIMIT 1
20/11/30 05:37:22 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ATHLETE` AS t LIMIT 1
20/11/30 05:37:25 INFO hive.metastore: Trying to connect to metastore with URI thrift://127.0.0.1:9083
```

cloudera@quickstart:~

File Edit View Search Terminal Help

```
20/11/30 05:38:33 INFO mapreduce.Job: Job job_1606741526175_0001 running in uber mode : false
20/11/30 05:38:33 INFO mapreduce.Job:  map 0% reduce 0%
20/11/30 05:40:07 INFO mapreduce.Job:  map 100% reduce 0%
20/11/30 05:40:16 INFO mapreduce.Job: Job job_1606741526175_0001 completed successfully
20/11/30 05:40:17 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=984952
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=71872
                HDFS: Number of bytes written=4177546
                HDFS: Number of read operations=192
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=40
        Job Counters
                Launched map tasks=4
                Other local map tasks=4
                Total time spent by all maps in occupied slots (ms)=371909
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=371909
                Total vcore-milliseconds taken by all map tasks=371909
                Total megabyte-milliseconds taken by all map tasks=380834816
        Map-Reduce Framework
                Map input records=135571
                Map output records=135571
                Input split bytes=424
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=7999
                CPU time spent (ms)=26470
                Physical memory (bytes) snapshot=639549440
                Virtual memory (bytes) snapshot=6105960448
                Total committed heap usage (bytes)=243531776
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
20/11/30 05:40:17 INFO mapreduce.ImportJobBase: Transferred 3.984 MB in 165.4493 seconds (24.6579 KB/sec)
20/11/30 05:40:17 INFO mapreduce.ImportJobBase: Retrieved 135571 records.
[cloudera@quickstart ~]$
```

HIVE:

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> Show Tables;
OK
athlete
sales
Time taken: 1.017 seconds, Fetched: 2 row(s)
```

```
Time taken: 0.759 seconds, Fetched: 10 row(s)
hive> select * from ATHLETE limit 10;
OK
33894   Craig Dean Falkman      M       24      180     95      United States   USA     1968 Winter     1968    Winter  Grenoble        Ice Hockey      Ice Hockey Men's Ice Hockey     NA
33895   Grigory Alekseyevich Falko      M       17      187     72      Russia  RUS     2004 Summer     2004    Summer  Athina  Swimming        Swimming Men's 200 metres Breaststroke  NA
33896   Leif Roar Falkum        M       27      186     72      Norway  NOR     1976 Summer     1976    Summer  Montreal        Athletics       Athletics Men's High Jump       NA
33897   Brre Erik Falkum-Hansen M       32      NA      NA      Encore  NOR     1952 Summer     1952    Summer  Helsinki        Sailing Sailing Mixed 5.5 metres        Silver
33898   Abdou Fall      M       NA      173     60      Senegal SEN     1972 Summer     1972    Summer  Munich  Boxing  Boxing Men's Light-Welterweight NA
33899   Adama Fall      M       25      172     72      Senegal SEN     1976 Summer     1976    Summer  Montreal        Athletics       Athletics Men's 100 metres      NA
33900   Aicha Bilal Fall        F       18      160     46      Mauritania      MTN     2012 Summer     2012    Summer  London  Athletics       Athletics Women's 800 metres    NA
33901   Ada Fall        F       29      193     95      Senegal SEN     2016 Summer     2016    Summer  Rio de Janeiro  Basketball      Basketball Women's Basketball   NA
33902   Assane Dame Fall        M       24      NA      78      Senegal SEN     2008 Summer     2008    Summer  Beijing Canoeing        Canoeing Men's Kayak Singles    500 metre
33903   Cheikh Amadou Fall      M       22      182     70      Senegal SEN     1968 Summer     1968    Summer  Mexico City     Basketball      Basketball Men's Basketball     NA
Time taken: 0.057 seconds, Fetched: 10 row(s)
hive>
```

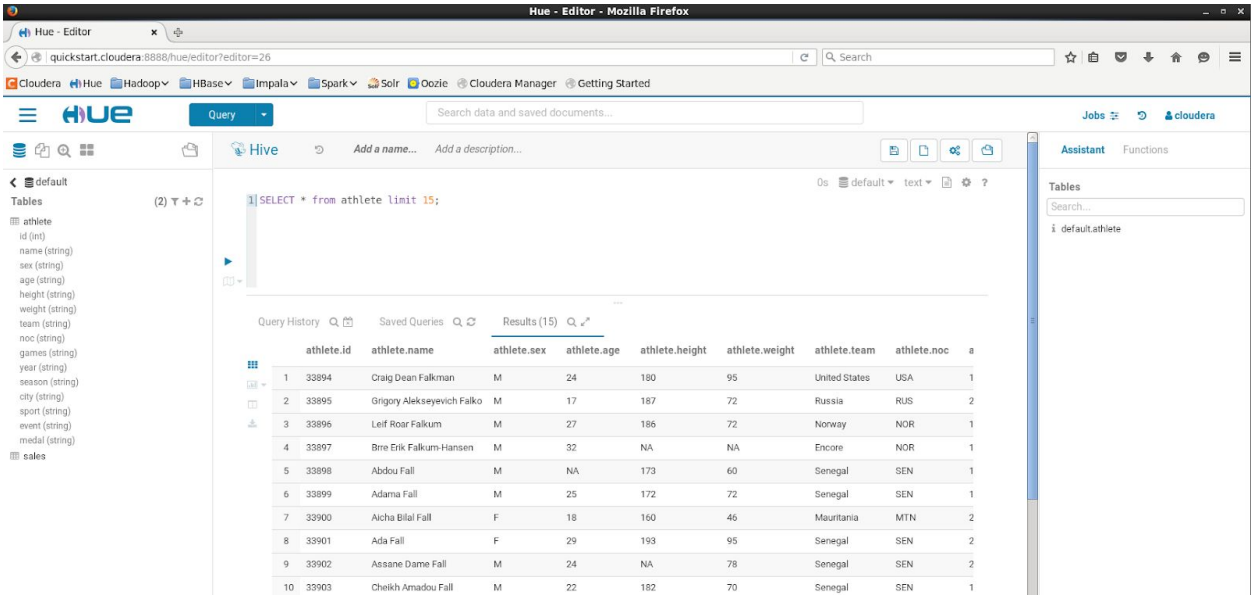
OUTPUT:

1. Year wise participation of each country in Winter season:

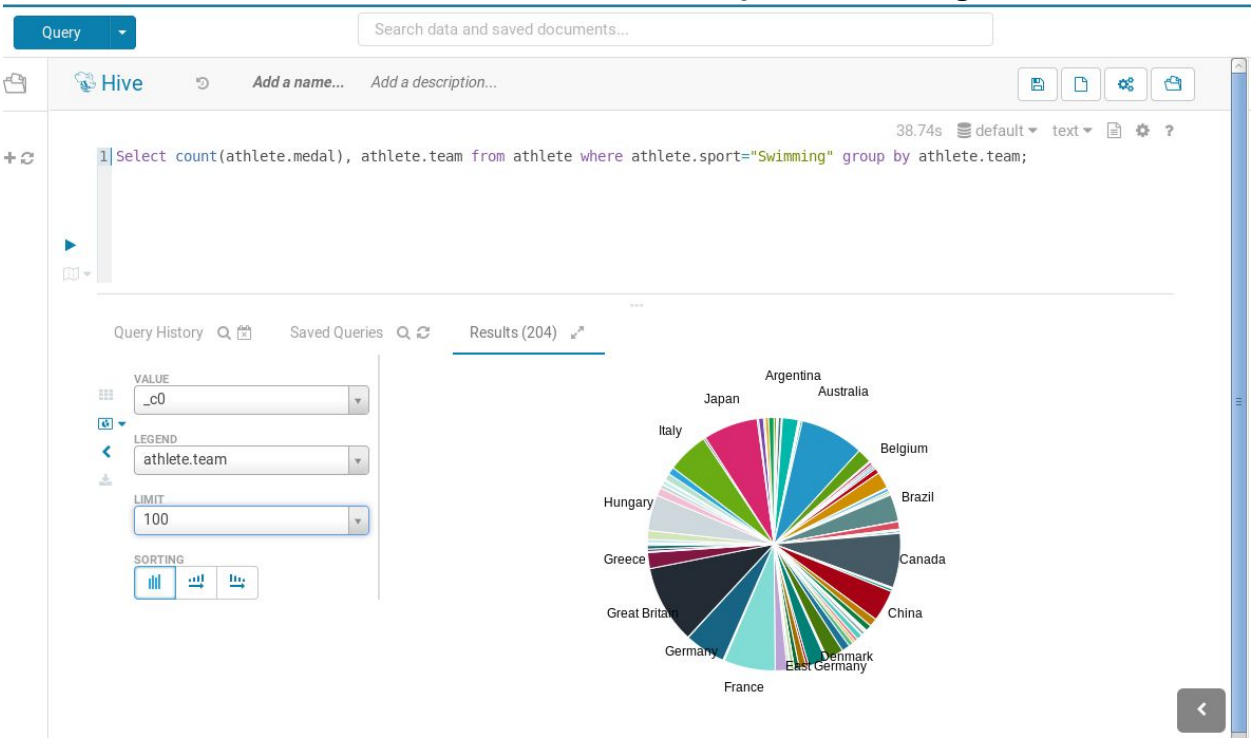
```
hive> select count(athlete.team),athlete.year from athlete where athlete.season="Winter" group by athlete.year;
Query ID = cloudera_20201202030707_d3b73d95-288a-4e22-9cd3-e279db024e24
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In 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_1606903707630_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1606903707630_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1606903707630_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-12-02 03:08:25,870 Stage-1 map = 0%,  reduce = 0%
2020-12-02 03:09:04,483 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.15 sec
2020-12-02 03:09:32,079 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.2 sec
MapReduce Total cumulative CPU time: 9 seconds 200 msec
Ended Job = job_1606903707630_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.2 sec   HDFS Read: 376467 HDFS Write: 204 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 200 msec
OK
```

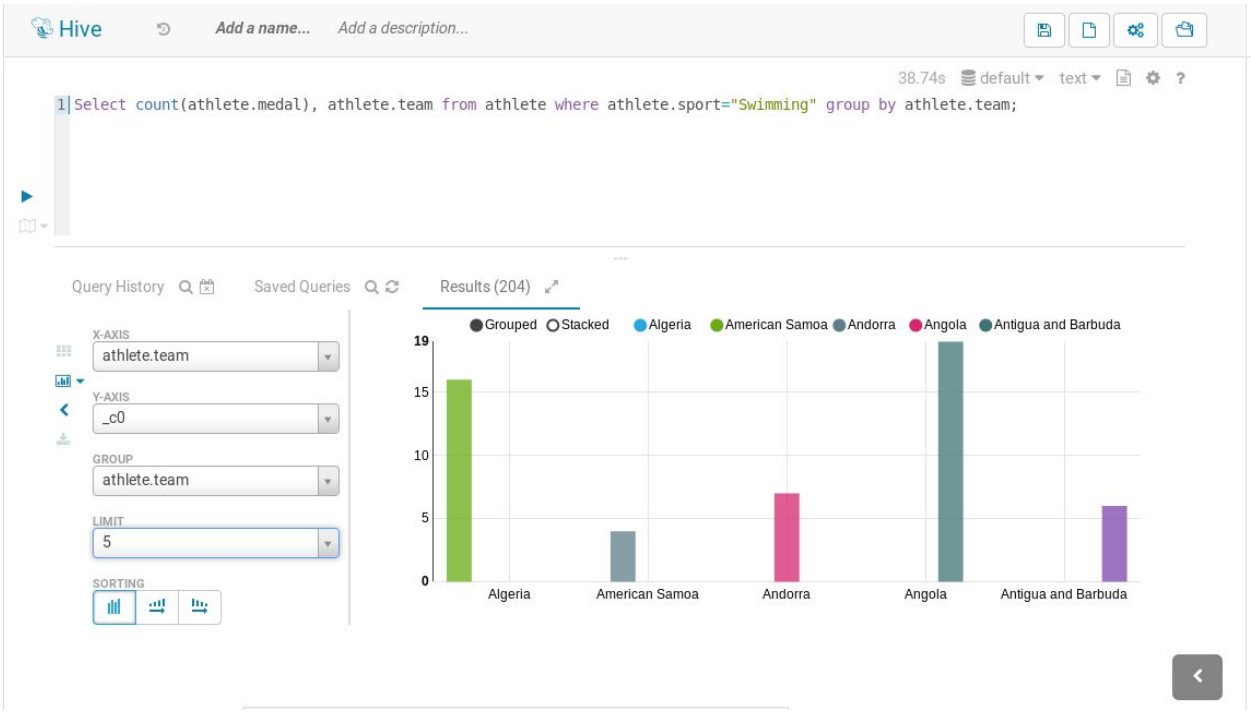
```
280     1924
386     1928
187     1932
546     1936
597     1948
527     1952
604     1956
474     1960
806     1964
817     1968
668     1972
810     1976
717     1980
939     1984
970     1988
1298    1992
858     1994
1388    1998
1383    2002
1433    2006
1416    2010
1551    2014
Time taken: 102.895 seconds, Fetched: 22 row(s)
hive>
```

2. Displaying table contents

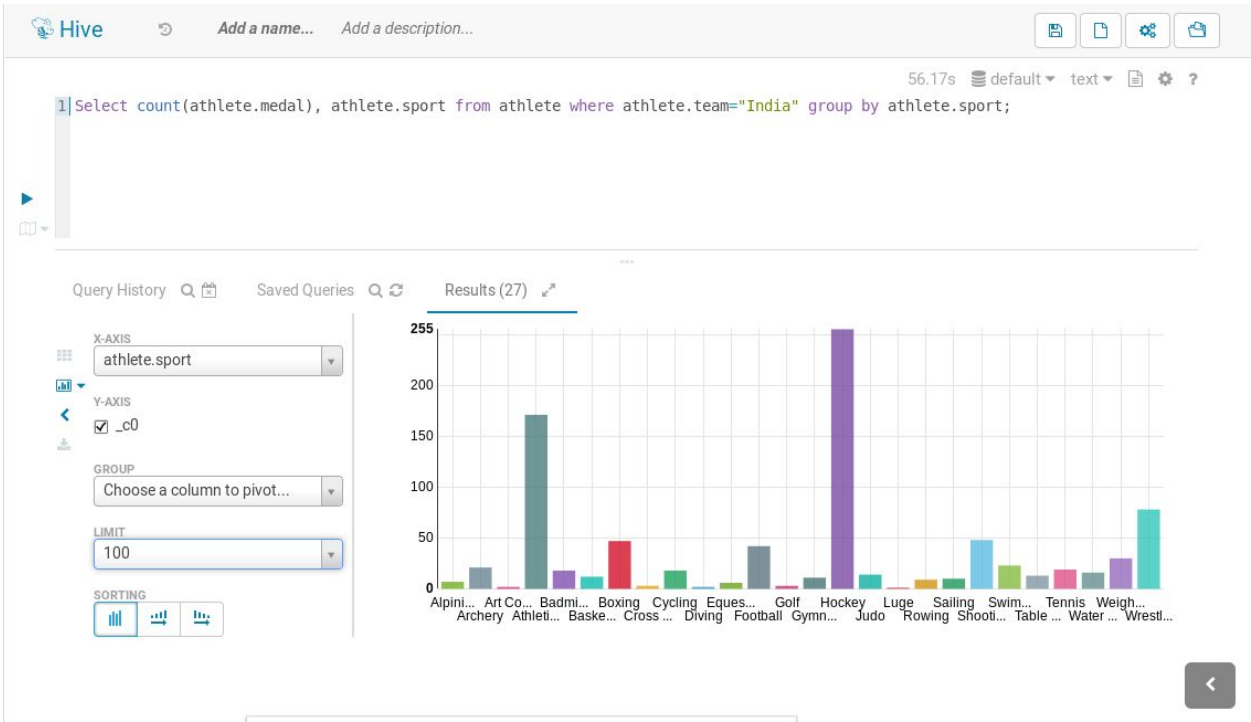


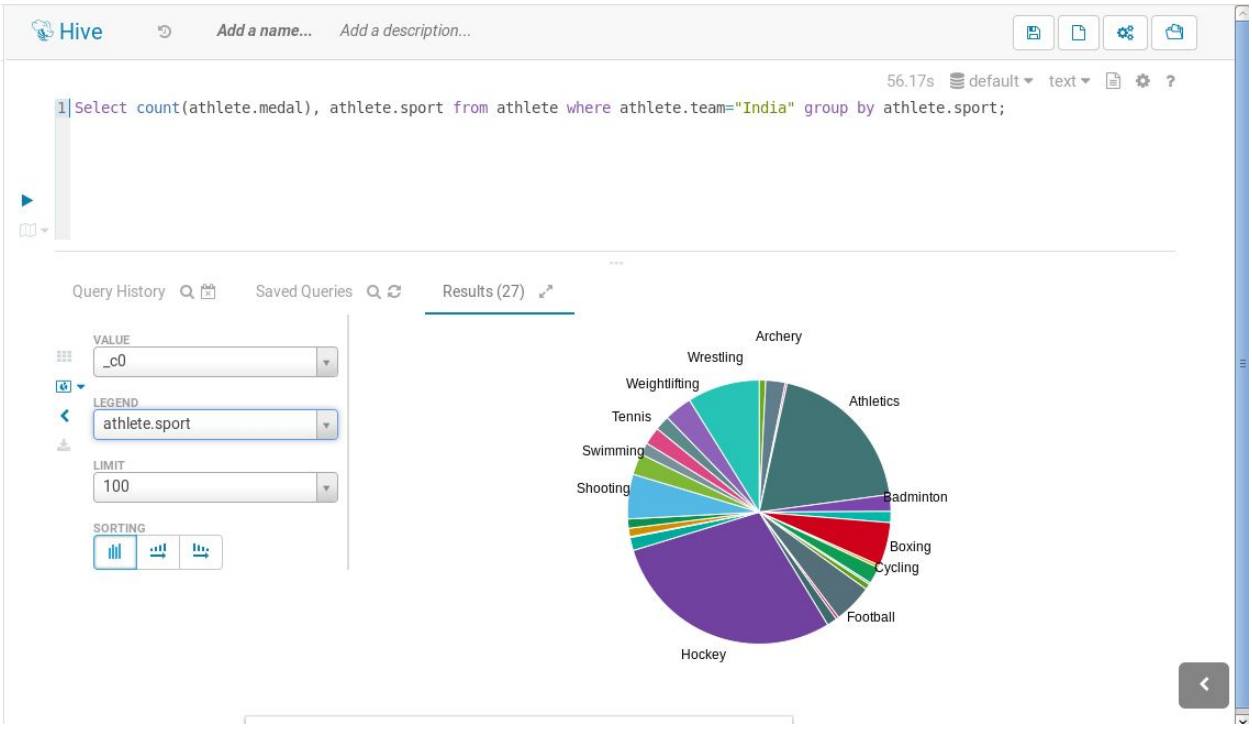
3. Performance of different countries in the sport-Swimming.



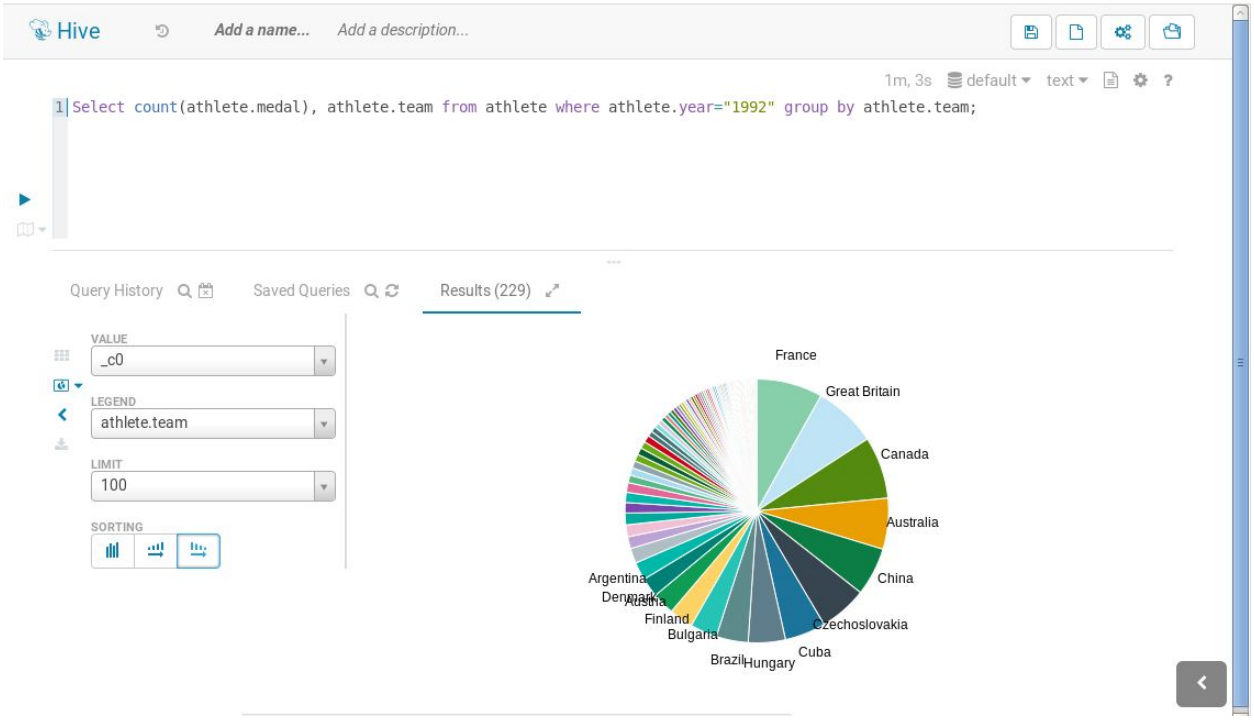


4. Number of medals won by Team India in different sports over the years.

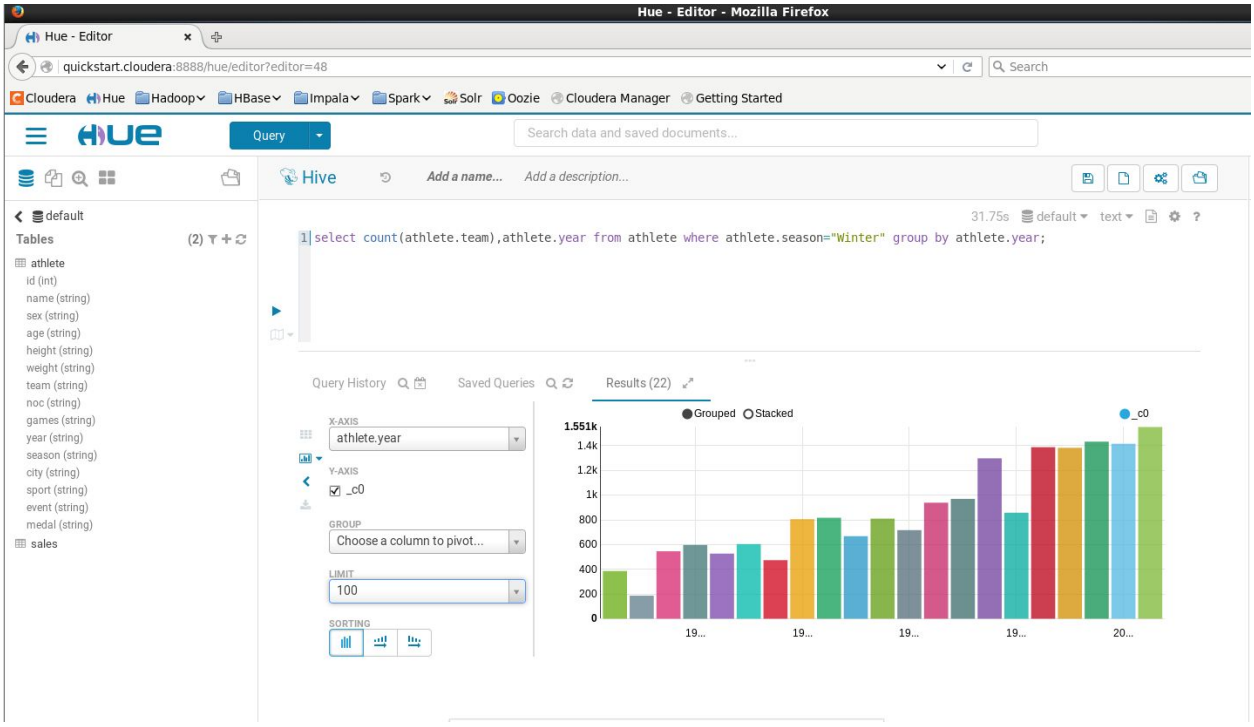




5. Performance of different countries in the year 1992 in descending order.

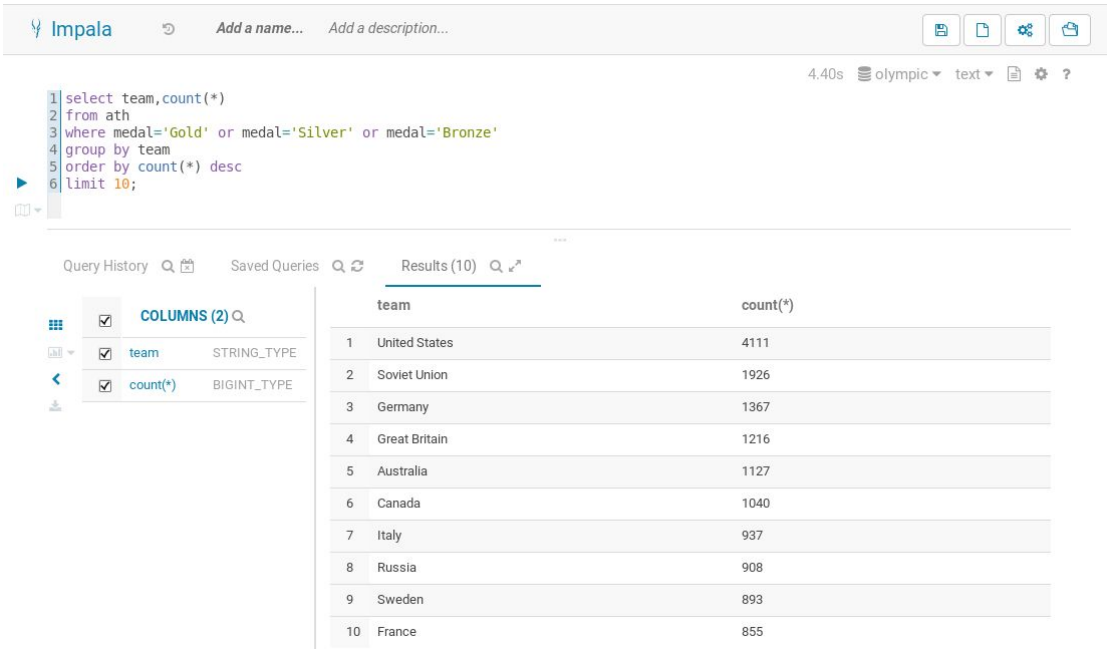


6. **Number of countries who participated in the Olympics in the winter season.**

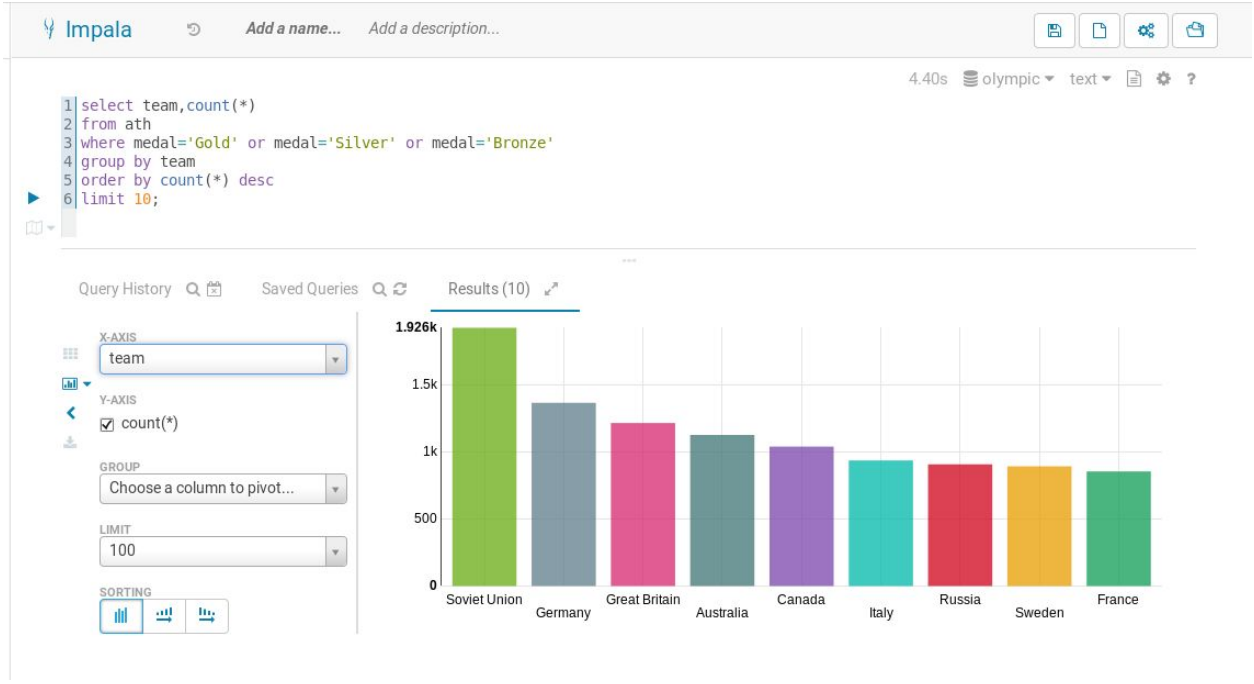


IMPALA:

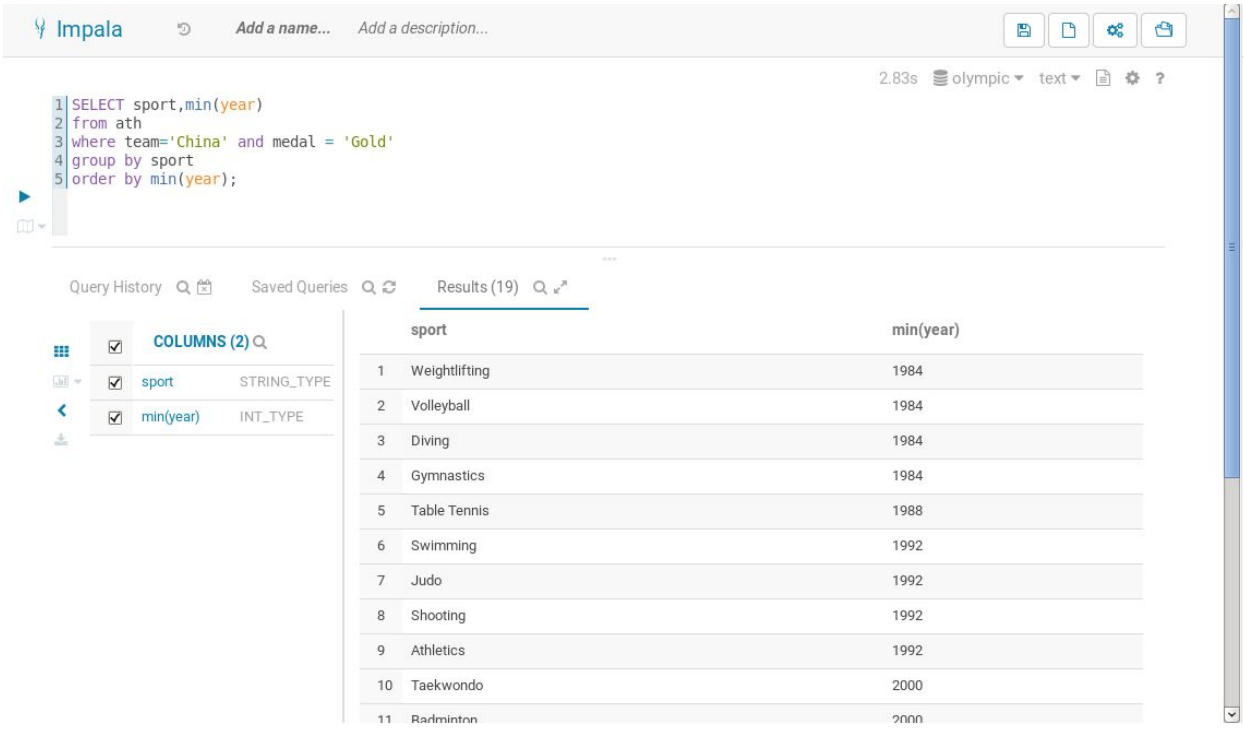
1. **Top ten countries with the highest Olympic medals**



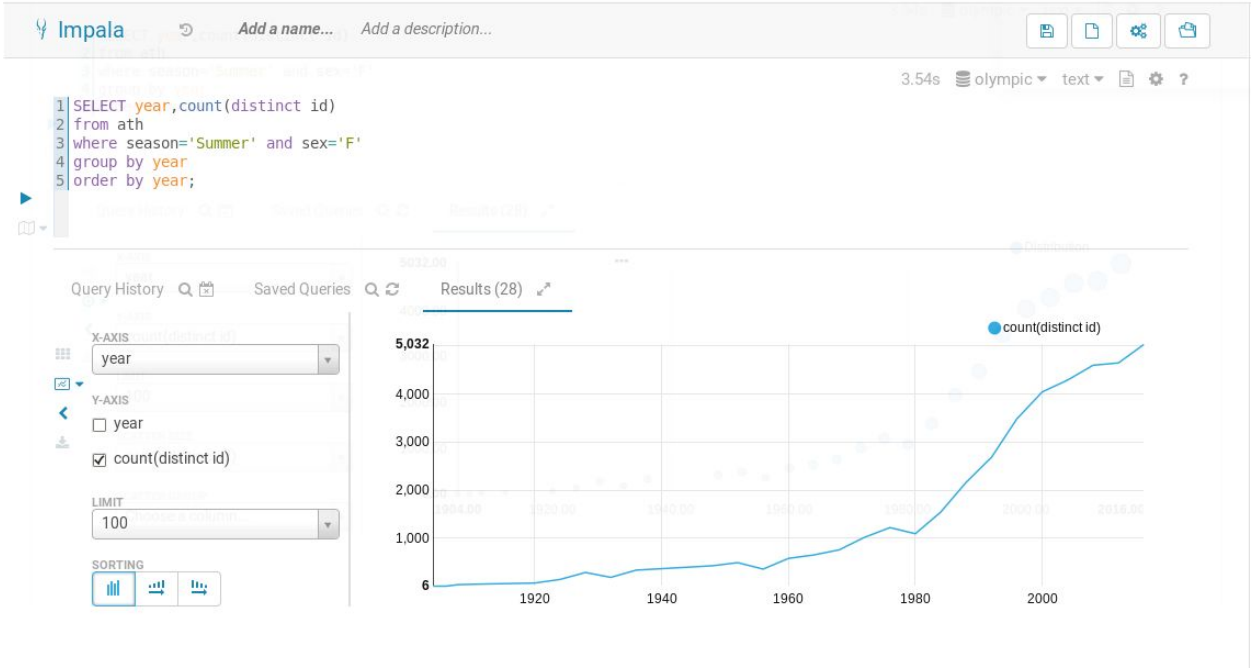
	team	count(*)
1	United States	4111
2	Soviet Union	1926
3	Germany	1367
4	Great Britain	1216
5	Australia	1127
6	Canada	1040
7	Italy	937
8	Russia	908
9	Sweden	893
10	France	855

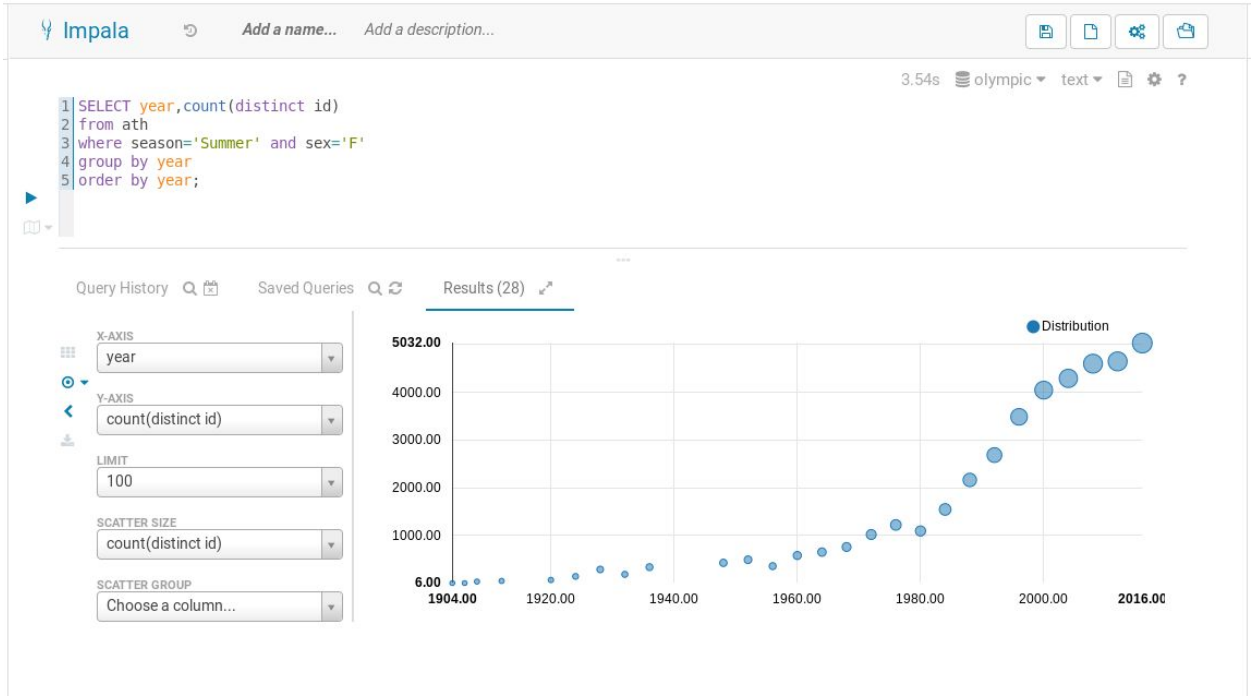


2. First Gold of various sports in the history of Chinese Olympics

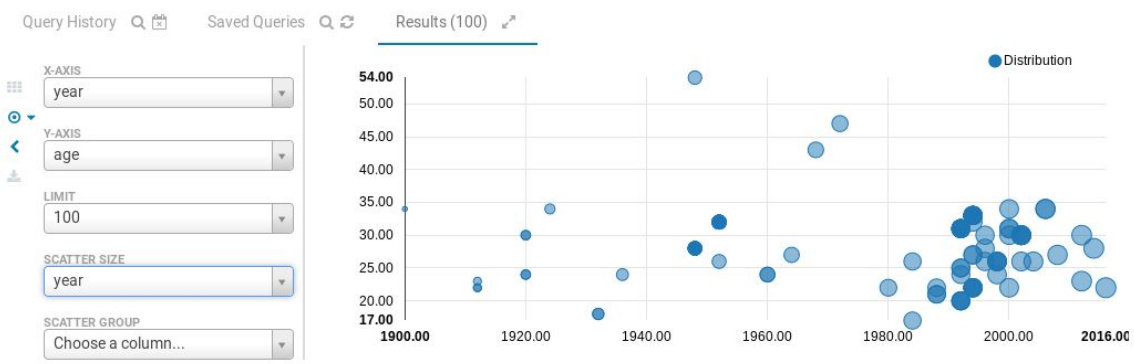


3. Participation of women in Olympic Games over the years.





4. Comparison of age of athletes distributed over the years



Pyspark:

```
from __future__ import division
from pyspark import SparkConf, SparkContext, SQLContext
import pyspark.sql.functions as F

conf = SparkConf().setMaster("local[*]")
sc = SparkContext(conf=conf)
sqlContext = SQLContext(sc)

df = sqlContext.read.csv('athlete_events.csv', header=True)

df_performance = df.select(['NOC',
'Medal']).filter(~df['Medal'].isin(['Gold', 'Silver', 'Bronze']) ==
False).
groupby(['NOC', 'Medal']).count().orderBy('NOC', 'Medal')

df_performance.show()
```

```
+---+------+-----+
|NOC| Medal|count|
+---+------+-----+
|AHO|Silver|    1|
|ALG|Bronze|    2|
|ALG|  Gold|    1|
|ALG|Silver|    2|
|ANZ|Bronze|    2|
|ANZ|  Gold|    3|
|ANZ|Silver|    3|
|ARG|Bronze|   16|
|ARG|  Gold|   11|
|ARG|Silver|   19|
|ARM|Bronze|    2|
|ARM|  Gold|    1|
|ARM|Silver|    1|
|AUS|Bronze|   87|
|AUS|  Gold|   47|
|AUS|Silver|   72|
|AUT|Bronze|    6|
|AUT|  Gold|    8|
|AUT|Silver|   15|
|AZE|Bronze|    5|
+---+------+-----+
```

```
df_bmi = df.select(['Year','Weight','Height'])
df_bmi = df_bmi.filter(~df_bmi['Height'].isin(['NA']) == True)
df_bmi = df_bmi.filter(~df_bmi['Year'].isin([x for x in range(1900,
2017)]) == False)
df_bmi_avg = df_bmi.withColumn("BMI",
F.col('Weight')/(F.col('Height')/100)**2).select(['Year','BMI']).groupby(
'Year').avg()
df_bmi_avg = df_bmi_avg.orderBy('Year')
df_bmi_avg.show()
```

```
+----+------------------+
|Year|          avg(BMI)|
+----+------------------+
|1900|27.757487216946675|
|1904| 21.63186790149742|
|1906|22.882972491691703|
|1908|23.650597084429958|
|1912|22.724233045635465|
|1920| 23.17853025170637|
|1924|23.370463021632947|
|1928|22.280582934439167|
|1932|23.438062458132315|
|1936|22.791167518844507|
|1948|23.063659726311343|
|1952| 23.45656549567396|
|1956|23.611457971043507|
|1960|22.950775690842296|
|1964|22.824364923825573|
|1968|22.618188985743274|
|1972|22.692729816825086|
|1976|22.790800677838096|
|1980| 22.80151509668998|
|1984|22.687699253011786|
+----+------------------+
```