Big data analysis on Olympic dataset

By

Bhagyashree Patil

Maximum number of participants per year

```
> db.top15years.find().sort({value:-1}).limit(15)
{ "_id" : { "Year" : "2000" }, "value" : 13682 }
{ "_id" : { "Year" : "1988" }, "value" : 13636 }
{ "_id" : { "Year" : "2016" }, "value" : 13443 }
{ "_id" : { "Year" : "2008" }, "value" : 13402 }
{ "_id" : { "Year" : "2004" }, "value" : 13399 }
{ "_id" : { "Year" : "1992" }, "value" : 13109 }
{ "_id" : { "Year" : "2012" }, "value" : 12524 }
{ "_id" : { "Year" : "1996" }, "value" : 11838 }
{ "_id" : { "Year" : "1972" }, "value" : 11482 }
{ "_id" : { "Year" : "1984" }, "value" : 10868 }
{ "_id" : { "Year" : "1968" }, "value" : 10203 }
{ "_id" : { "Year" : "1976" }, "value" : 9567 }
{ "_id" : { "Year" : "1964" }, "value" : 8711 }
{ "_id" : { "Year" : "1980" }, "value" : 8217 }
{ "_id" : { "Year" : "1960" }, "value" : 8038 }
```

```
Var map1= function()
{
emit({Year:this.Year},1);
}
Var red1= function (key,values)
{
var sum =0;
for(i=0;i<values.length;i++)
                                        {

                sum+=values[i];

                                        }
return sum;
 }
```

Indexing In MONGODB for faster query performance

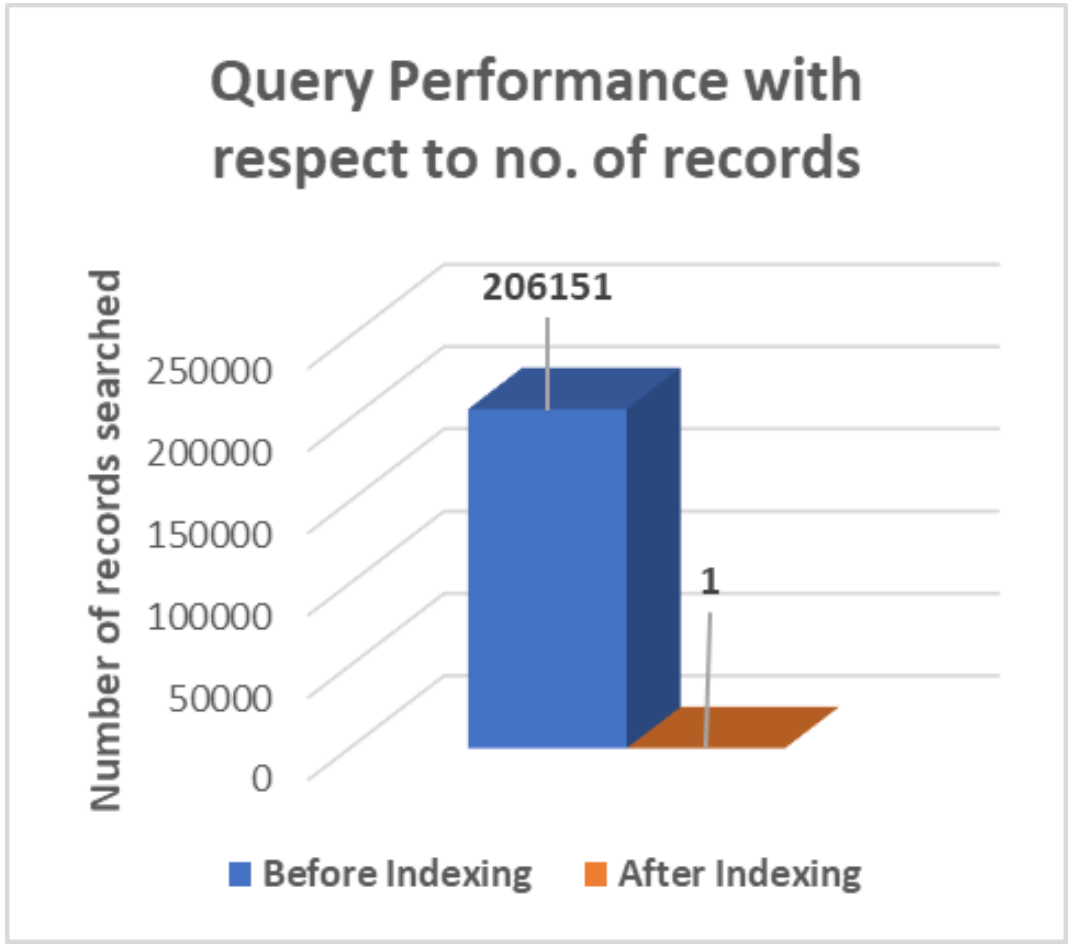
Before Indexing

```
> db.olympic.find({Name:"A Dijiang"}).explain("executionStats")
{
        "queryPlanner" : {
                "plannerVersion" : 1,
                "namespace" : "projectdb.olympic",
                "indexFilterSet" : false,
                "parsedQuery" : {
                        "Name" : {
                                "$eq" : "A Dijiang"
                        }
                },
                "queryHash" : "EBFEE4C5",
                "planCacheKey" : "EBFEE4C5",
                "winningPlan" : {
                        "stage" : "COLLSCAN",
                        "filter" : {
                                "Name" : {
                                        "$eq" : "A Dijiang"
                                }
                        },
                        "direction" : "forward"
                },
                "rejectedPlans" : [ ]
        },
        "executionStats" : {
                "executionSuccess" : true,
                "nReturned" : 1,
                "executionTimeMillis" : 92,
                "totalKeysExamined" : 0,
                "totalDocsExamined" : 206152,
                "executionStages" : {
```

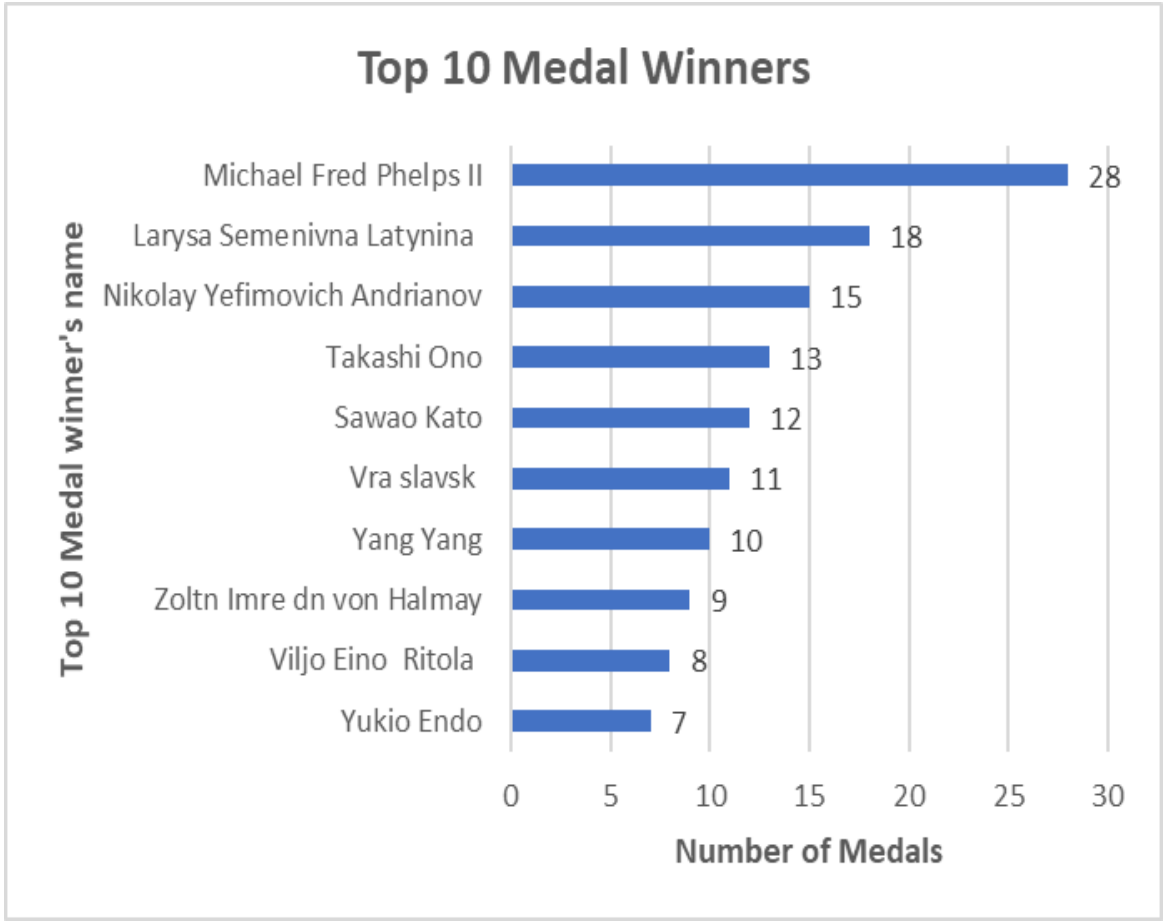
After Indexing

```
> db.olympic.find({Name:"A Dijiang"}).explain("executionStats")
{
        "queryPlanner" : {
                "plannerVersion" : 1,
                "namespace" : "projectdb.olympic",
                "indexFilterSet" : false,
                "parsedQuery" : {
                        "Name" : {
                                "$eq" : "A Dijiang"
                        }
                },
                "queryHash" : "EBFEE4C5",
                "planCacheKey" : "6D446D9E",
                "winningPlan" : {
                        "stage" : "FETCH",
                        "inputStage" : {
                                "stage" : "IXSCAN",
                                "keyPattern" : {
                                        "Name" : 1
                                },
                                "indexName" : "Name_1",
                                "isMultiKey" : false,
                                "multiKeyPaths" : {
                                        "Name" : [ ]
                                },
                                "isUnique" : false,
                                "isSparse" : false,
                                "isPartial" : false,
                                "indexVersion" : 2,
                                "direction" : "forward",
                                "indexBounds" : {
                                        "Name" : [
                                                "[\"A Dijiang\", \"A Dijiang\"]
                                        ]
                                }
                        }
                },
                "rejectedPlans" : [ ]
        },
        "executionStats" : {
                "executionSuccess" : true,
                "nReturned" : 1,
                "executionTimeMillis" : 2,
                "totalKeysExamined" : 1,
                "totalDocsExamined" : 1,
                "executionStages" : {
```

Indexing in MONGODB for faster query performance



Top 10 Medal Winners Overall using MapReduce Chaining




```
bhagyashree@ubuntu:/usr/local/bin/hadoop-2.9.2/bin$ hdfs dfs -cat /top10Me
28      Michael Fred Phelps II
18      Larysa Semenivna Latynina
15      Nikolay Yefimovich Andrianov
13      Takashi Ono
12      Sawao Kato
11      Vra slavsk
10      Yang Yang
9       Zoltn Imre dn von Halmay
8       Viljo Eino  Ritola
7       Yukio Endo
```

Partition records based on Season

File information - part-r-00001

Download | Head the file (first 32K) | Tail the file (last 32K)

Block information -- Block 0

Block ID: 1073743342

Block Pool ID: BP-1321425052-127.0.1.1-1570048576690

Generation Stamp: 2522

Size: 673285

Availability:

- ubuntu

File contents

```
49351,Ji Holk,M,27,178,85,Czechoslovakia,TCH,1972
Winter,1972,Winter,Sapporo,Ice Hockey,Ice Hockey Men's Ice Hockey,Bronze
49351,Ji Holk,M,23,178,85,Czechoslovakia,TCH,1968
Winter,1968,Winter,Grenoble,Ice Hockey,Ice Hockey Men's Ice Hockey,Silver
49351,Ji Holk,M,19,178,85,Czechoslovakia,TCH,1964
Winter,1964,Winter,Innsbruck,Ice Hockey,Ice Hockey Men's Ice
Hockey,Bronze
```

File information - part-r-00000

Download | Head the file (first 32K) | Tail the file (last 32K)

Block information -- Block 0

Block ID: 1073743341

Block Pool ID: BP-1321425052-127.0.1.1-1570048576690

Generation Stamp: 2521

Size: 3349946

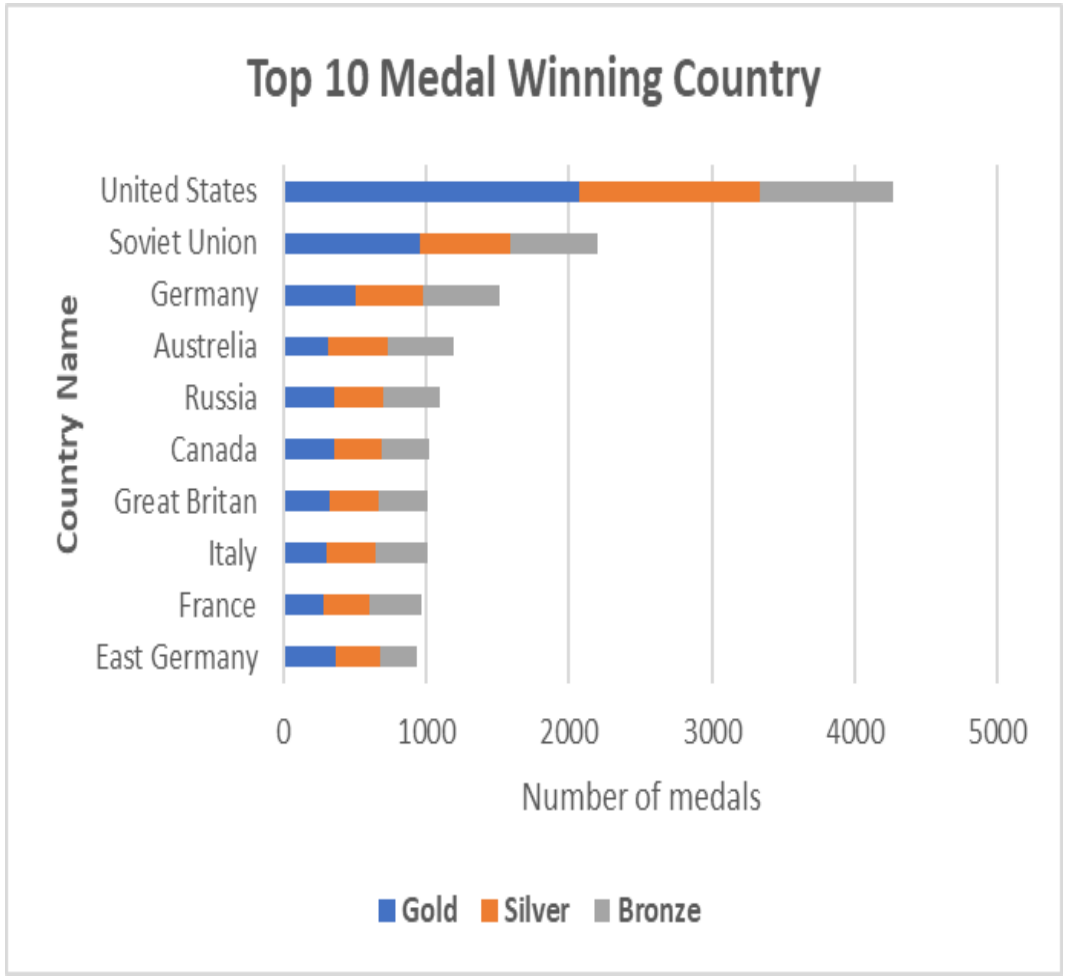
Availability:

- ubuntu

File contents

```
Summer,1956,Summer,Melbourne,Athletics,Athletics Women's Shot
Put,Silver
135553,Galina Ivanovna Zybina ,F,21,168,80,Soviet Union,URS,1952
Summer,1952,Summer,Helsinki,Athletics,Athletics Women's Shot Put,Gold
135545,Henk Jan Zwolle,M,31,197,93,Netherlands,NED,1996
Summer,1996,Summer,Atlanta,Rowing,Rowing Men's Coxed Eights,Gold
135545,Henk Jan Zwolle,M,27,197,93,Netherlands,NED,1992
```

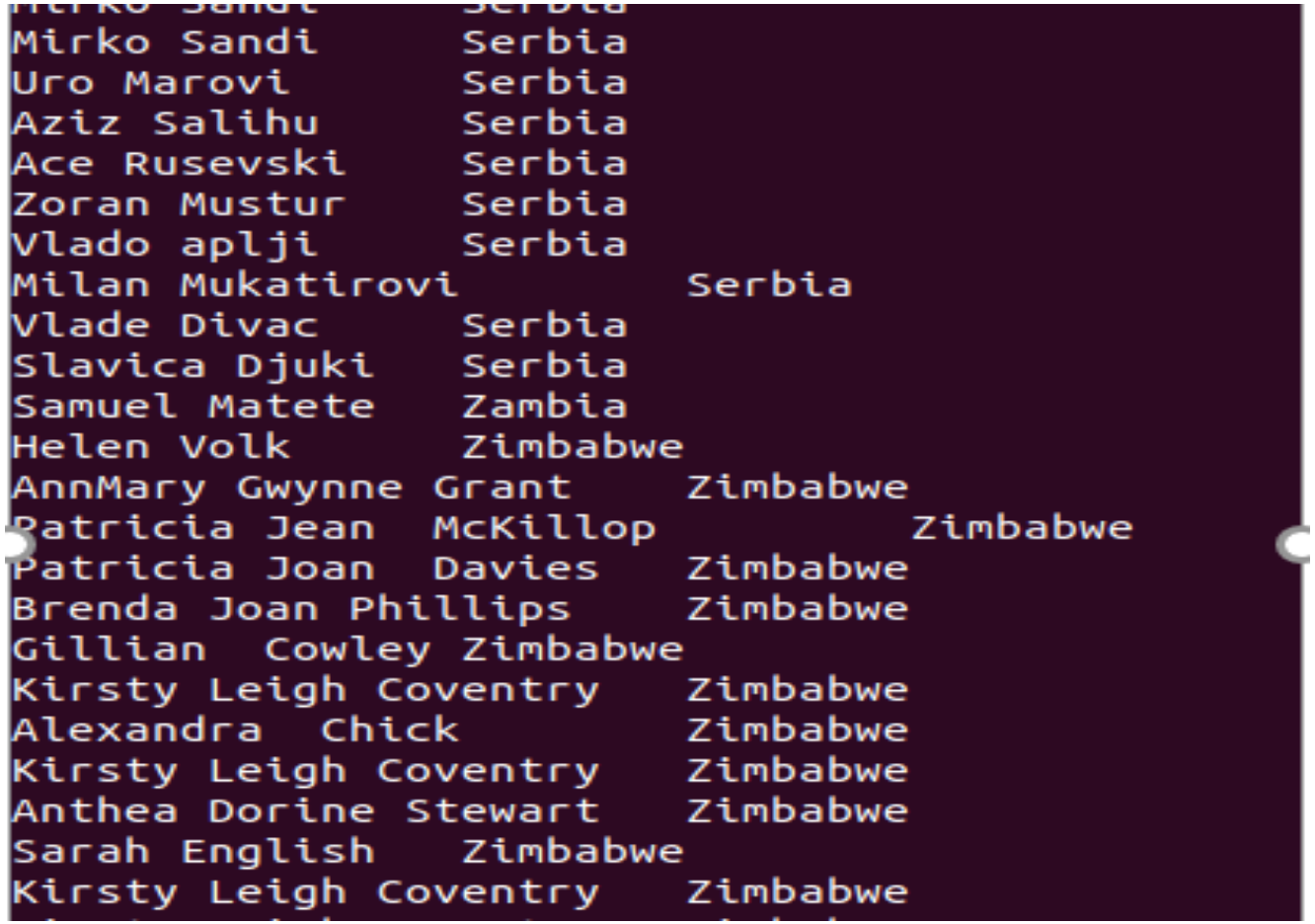
Top 10 Medal Winning Team using top k filtering




```
                Bytes Written=717
bhagyashree@ubuntu:/usr/local/bin/hadoop-2.9.2/bin$ hdfs dfs -cat /top10Teams/part-r-00000
4273    United States   GoldCount=2075  SilverCount=1260        BronzeCount=938 Total=4273
2203    Soviet Union    GoldCount=961   SilverCount=629 BronzeCount=613 Total=2203
1518    Germany GoldCount=508   SilverCount=470 BronzeCount=540 Total=1518
1196    Australia       GoldCount=313   SilverCount=412 BronzeCount=471 Total=1196
1091    Russia  GoldCount=356   SilverCount=343 BronzeCount=392 Total=1091
1024    Canada  GoldCount=350   SilverCount=336 BronzeCount=338 Total=1024
1010    Great Britain   GoldCount=321   SilverCount=343 BronzeCount=346 Total=1010
1008    Italy   GoldCount=302   SilverCount=340 BronzeCount=366 Total=1008
965     France  GoldCount=279   SilverCount=320 BronzeCount=366 Total=965
935     East Germany    GoldCount=368   SilverCount=306 BronzeCount=261 Total=935
```

Players by country using Reduce side join to enrich the dataset

```
bhagyashree@ubuntu:/usr/local/bin/hadoop-2.9.2/bin$ hadoop jar /home/bhagyashree/Desktop/join3.jar ProjectJoin.ProjectJoin.JoinDriver /project /projectInputJoin inner /project_inner_join_output
```



Loading data into partition table from hive table

/user/hive/warehouse/olympicdatabase.db/partition_olympics Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1896
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1900
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1904
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1906
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1908
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1912
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1920
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1924
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1928
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1932
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1936
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1948
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1952
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1956
drwxr-xr-x	bhagyashree	supergroup	0 B	Nov 26 20:11	0	0 B	year=1960

select Team,avg(height) from partition_olympics where year=2016 group by team

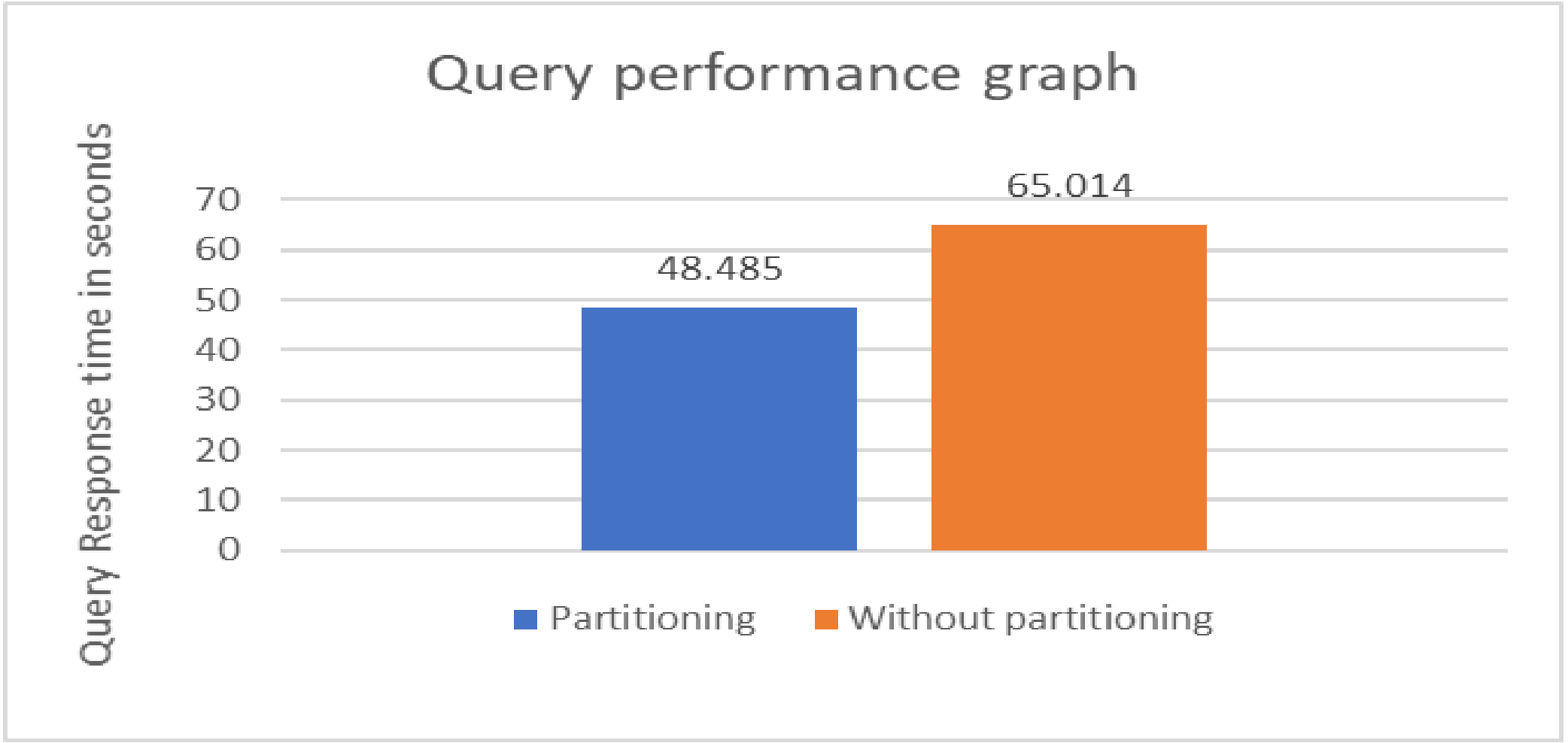
Before partition

```
Afghanistan      173.66666666666666
Albania 176.16666666666666
Algeria 174.02702702702703
American Samoa   176.75
Andorra 171.5
Angola   174.5
Antigua and Barbuda      176.75
Argentina        178.54385964912281
Argentina-1      186.5
Argentina-2      184.0
Armenia 170.3235294117647
Aruba    174.28571428571428
Australia        178.61904761904762
Australia-1      177.0
Australia-2      175.25
Austria 176.30864197530863
Austria-1        184.5
Austria-2        193.0
Azerbaijan       175.66666666666666
Bahamas 176.78125
Bahrain 171.32142857142858
Bangladesh       166.5
Barbados         180.69230769230768
Belarus 176.55633802816902
Belgium 176.2246376811594
Belize   183.66666666666666
Benin    179.33333333333334
Bermuda 174.125
Bhutan   164.0
Bolivia 171.33333333333334
Bosnia and Herzegovina   180.18181818
Botswana         179.4
```

After partition

```
Afghanistan      173.66666666666666
Albania 176.16666666666666
Algeria 174.02702702702703
American Samoa   176.75
Andorra 171.5
Angola   174.5
Antigua and Barbuda      176.75
Argentina        178.54385964912281
Argentina-1      186.5
Argentina-2      184.0
Armenia 170.3235294117647
Aruba    174.28571428571428
Australia        178.61904761904762
Australia-1      177.0
Australia-2      175.25
Austria 176.30864197530863
Austria-1        184.5
Austria-2        193.0
Azerbaijan       175.66666666666666
Bahamas 176.78125
Bahrain 171.32142857142858
Bangladesh       166.5
Barbados         180.69230769230768
Belarus 176.55633802816902
Belgium 176.2246376811594
Belize   183.66666666666666
Benin    179.33333333333334
Bermuda 174.125
Bhutan   164.0
Bolivia 171.33333333333334
```

select Team,avg(height) from partition_olympics where year=2016 group by team;



Finding youngest player[View]

grunt>create view youngestPlayer as select distinct name,age,year,team,sport from olympics where age in (select min(age) from olympics);

```
Beatrice Hutiu   11      1968    Romania Figure Skating
Liana Vicens     11      1968    Puerto Rico     Swimming
Sonja Henie      11      1924    Norway  Figure Skating
Time taken: 105.671 seconds, Fetched: 3 row(s)
```

Indexing in hive

```
hive> CREATE INDEX index_name
    > ON TABLE Olympics (name)
    > AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
    > WITH DEFERRED REBUILD;
OK
Time taken: 0.327 seconds
```

```
hive> show formatted index on olympics;
OK
idx_name        tab_name        col_names       idx_tab_name            idx_type

index_sport     olympics        sport           olympicdatabase__olympics_index_sport__ compact
index_name      olympics        name            olympicdatabase__olympics_index_name__  compact
```

Hbase analysis

Insert data into table

```
hbase(main):001:0> put 'olympicdata','3','cfparticipants:name','Saina Nehawal'
Took 0.8559 seconds
hbase(main):002:0> put 'olympicdata','3','cfparticipants:game','Badminton'
Took 0.0097 seconds
hbase(main):003:0> put 'olympicdata','3','cfparticipants:medal','Gold'
Took 0.0157 seconds
```

Fetching records for particular row

```
hbase(main):004:0> get 'olympicdata','3'
COLUMN                                  CELL
 cfparticipants:game                    timestamp=1574885389523, value=Badminton
 cfparticipants:medal                   timestamp=1574885405456, value=Gold
 cfparticipants:name                    timestamp=1574885352982, value=Saina Nehawal
```

Scan the entire table

```
hbase(main):006:0> scan 'olympicdata'
ROW                                     COLUMN+CELL
 1                                      column=cfparticipants:medal, timestamp=1574883824062, value=Bronze
 1                                      column=cfparticipants:name, timestamp=1574881823982, value=Vikas Yadhav
 1                                      column=cfparticipants:sport, timestamp=1574881834430, value=Boxing
 2                                      column=cfparticipants:medal, timestamp=1574884333549, value=Silver
 3                                      column=cfparticipants:game, timestamp=1574885389523, value=Badminton
 3                                      column=cfparticipants:medal, timestamp=1574885405456, value=Gold
 3                                      column=cfparticipants:name, timestamp=1574885352982, value=Saina Nehawal
3 row(s)
```

Hbase analysis

Alter table to set number of versions we needed

```
hbase(main):010:0> alter 'olympicdata', {NAME => 'cfparticipants', VERSIONS => 3}
Updating all regions with the new schema...
1/1 regions updated.
Done.
```

Check last two version

```
hbase(main):014:0> get 'olympicdata','3',{COLUMN => 'cfparticipants', VERSIONS => 2}
COLUMN                                                 CELL
 cfparticipants:game                                   timestamp=1574885389523, value=Badminton
 cfparticipants:medal                                  timestamp=1574887266721, value=Silver
 cfparticipants:medal                                  timestamp=1574885405456, value=Gold
 cfparticipants:name                                   timestamp=1574885352982, value=Saina Nehawal
 cfparticipants:name                                   timestamp=1574885275934, value=Saina Nehawal
```

Get the record by specifying timestamp

```
hbase(main):015:0> get 'olympicdata','3',{COLUMN => 'cfparticipants', TIMESTAMP => 1574885405456}
COLUMN                                                 CELL
 cfparticipants:medal                                  timestamp=1574885405456, value=Gold
1 row(s)
```

Hbase analysis

Table which will live for only 100 sec, set time to live property to 100

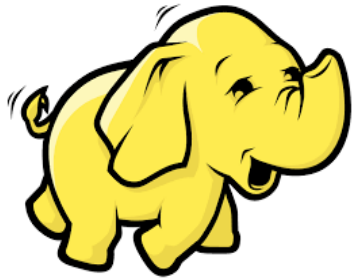
```
hbase(main):030:0> alter 'sampletable', {NAME =>'cfregion', VERSIONS =>1, TTL => 100}
Updating all regions with the new schema...
1/1 regions updated.
```

```
hbase(main):046:0> status
1 active master, 0 backup masters, 1 servers, 0 dead, 5.0000 average load
Took 0.0586 seconds
hbase(main):047:0> status 'summary'
1 active master, 0 backup masters, 1 servers, 0 dead, 5.0000 average load
Took 0.0375 seconds
```

```
hbase(main):053:0> disable  'olympicdata'
Took 1.4098 seconds
```

```
hbase(main):054:0> enable  'olympicdata'
Took 1.4196 seconds
```

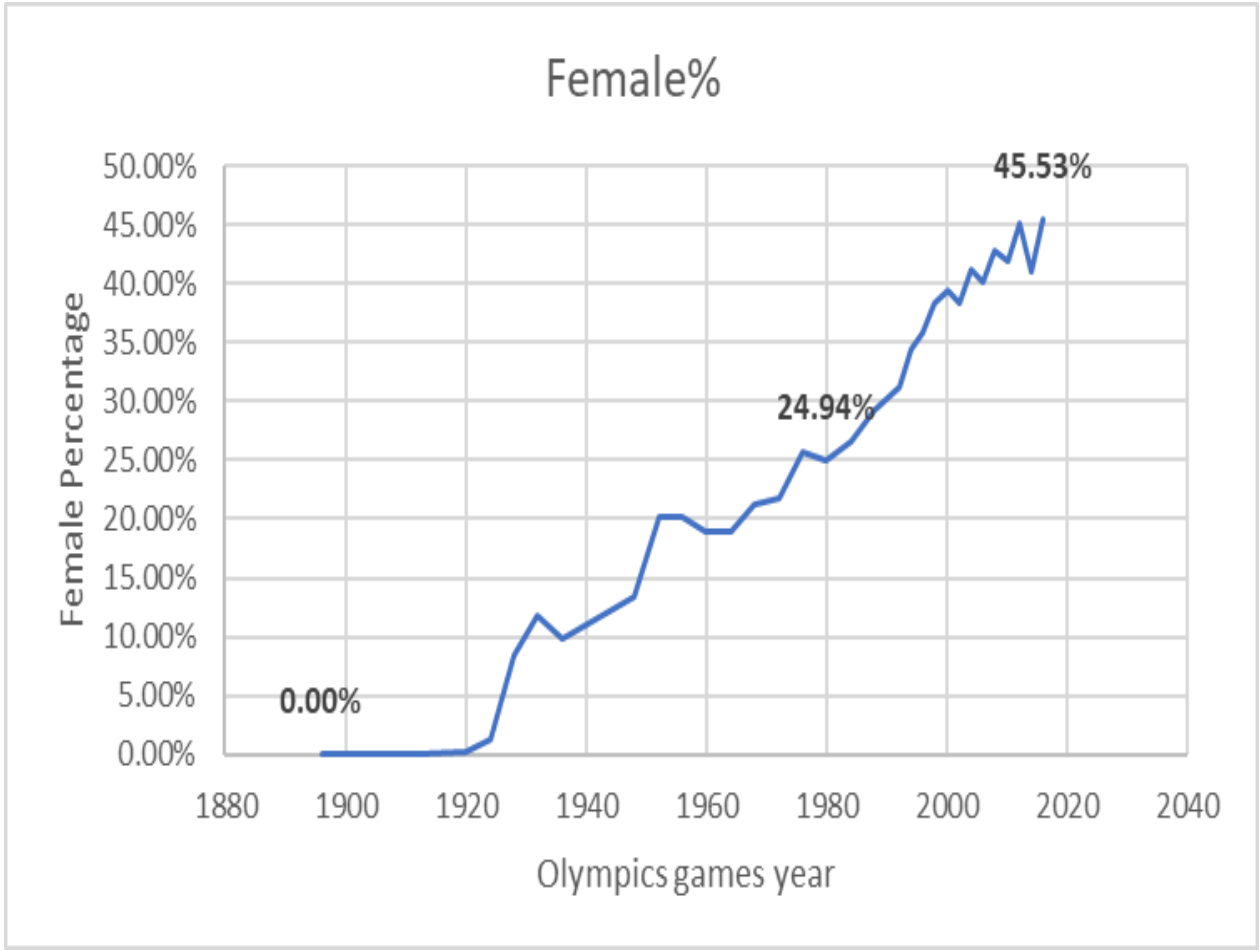
```
hbase(main):055:0> drop  'olympicdata'
```



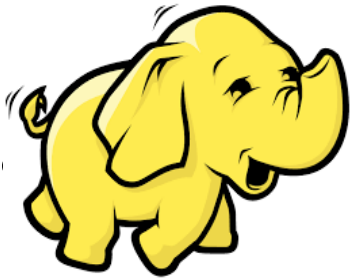
No of Female Participants per year

No Female participants 1896-1912

```
(1920,1)
(1924,7)
(1928,56)
(1932,57)
(1936,88)
(1948,137)
(1952,417)
(1956,525)
(1960,1516)
(1964,1643)
(1968,2169)
(1972,2494)
(1976,2463)
(1980,2049)
(1984,2885)
(1988,4002)
(1992,4085)
(1994,1023)
(1996,4242)
(1998,1350)
(2000,5386)
(2002,1555)
(2004,5536)
(2006,1753)
(2008,5739)
(2010,1837)
(2012,5655)
(2014,1920)
(2016,6121)
```



Finding the distributions of gold medals across different sports in ye 2000 (sport,no. of gold medals)



```
(Swimming,66)
(Athletics,63)
(Rowing,48)
(Hockey,32)
(Football,31)
(Handball,30)
(Canoeing,28)
(Water Polo,26)
(Cycling,25)
(Gymnastics,24)
(Baseball,24)
(Basketball,23)
(Fencing,23)
(Volleyball,19)
(Sailing,18)
(Shooting,17)
(Wrestling,16)
(Equestrianism,15)
(Weightlifting,15)
(Softball,15)
(Judo,14)
(Diving,12)
(Boxing,12)
(Synchronized Swimming,10)
(Badminton,8)
(Taekwondo,8)
(Archery,8)
(Rhythmic Gymnastics,7)
(Tennis,6)
(Table Tennis,6)
(Beach Volleyball,4)
(Trampolining,2)
(Modern Pentathlon,2)
(Triathlon,2)
```