

Analytics su Pig e Hive

Antonio Cola, Maristella Simonetti, Rosario Urso

1 Descrizione dataset

Il dataset preso in esame contiene informazioni sui Giochi olimpici a partire dall'anno in cui sono iniziati (1896) fino alle Olimpiadi tenutesi a Tokyo nel 2020 includendo sia i Giochi estivi che quelli invernali. Le variabili presenti nel dataset sono: Nome, Sesso, Età, Squadra, Giochi, Anno, Sport e Medaglie.

È opportuna una specificazione sulla variabile Medaglie: essa è una variabile categorica le cui modalità sono "0", "1", "2", "3" che corrispondono rispettivamente a: nessuna medaglia vinta, medaglia di bronzo, medaglia d'argento, medaglia d'oro.

La quantità di dati presenti nel dataset è 2480787 divisa in 9 colonne, più una colonna di valori ordinali. Nel dataset sono presenti: 139157 atleti (Name) che hanno un'età compresa tra i 10 e i 97 anni (Age), 1203 diverse squadre (Team), 52 edizioni (Games) in 36 anni (Year), 2 Stagioni (Season) e 79 sport diversi (Sport).

2 Descrizione analytics

Con la prima query effettuata sul dataset intendiamo ottenere una classifica delle prime 10 nazioni ad aver ottenuto il maggior numero di medaglie d'oro nel Judo tra gli anni 2000 e 2020. Con la seconda e la terza query vogliamo avere un'idea di quali sono gli sport nei quali l'età media dei partecipanti è più elevata e di quelli dove l'età media è più bassa, rispettivamente. In entrambi i casi ci limitiamo ai primi dieci sport in classifica. Con la quarta query invece abbiamo calcolato per ogni sesso sia l'età media che la frequenza assoluta dei partecipanti. Con la quinta query cerchiamo una classifica delle nazioni che hanno vinto il maggior numero di ori, in assoluto. In questo caso, abbiamo considerato solo la top 5. La sesta query riguarda i 5 sport che vedono il maggior numero di partecipanti donne. Con la settima query, infine, vogliamo ottenere il nome dei 10 atleti, con le rispettive nazioni, i quali hanno vinto il maggior numero di medaglie d'oro.

3 Implementazione analytics

3.1 Hive

Elimina la tabella dataset se esiste.

```
1 DROP TABLE IF EXISTS dataset;
```

Crea la tabella dataset dal file csv caricato.

```
CREATE TABLE dataset
USING csv
OPTIONS (path "/FileStore/tables/Olympic_Dataset.csv",header="TRUE");
```

Visualizza l'intera tabella.

```
1 SELECT *
2 FROM dataset;
```

	_c0	Name	Sex	Age	Team	Games	Year	Season	Sport	Medal
1	0	A Dijiang	M	24.0	China	1992 Summer	1992	Summer	Basketball	0
2	1	A Lamusi	M	23.0	China	2012 Summer	2012	Summer	Judo	0
3	2	Gunnar Nielsen Aaby	M	24.0	Denmark	1920 Summer	1920	Summer	Football	0
4	3	Edgar Lindenu Aabye	M	34.0	Denmark/Sweden	1900 Summer	1900	Summer	Tug-Of-War	3
5	4	Christine Jacobsa Aaftink	F	21.0	Netherlands	1988 Winter	1988	Winter	Speed Skating	0

Figura 1: Visualizzazione tabella dataset su Hive.

Top 10 nazioni per medaglie d'oro vinte nel Judo.

```
1 SELECT Team AS Nazione, Count(*) AS 'Medaglie Vinte'
2 FROM(
3     SELECT Team, Medal, Sport, 'Year'
4     FROM dataset
5 )
6 WHERE Medal==3 AND Sport=="Judo" AND 'Year'>=2020
7 GROUP BY Team
8 ORDER BY 'Medaglie Vinte' DESC
9 LIMIT 10;
```

	Nazione	Medaglie Vinte
1	Japan	20
2	Azerbaijan	7
3	France	6
4	China	6
5	Russia	5
6	South Korea	4
7	Italy	3
8	Brazil	3
9	Cuba	3
10	Georgia	3

Figura 2: Output prima query Hive.

Top 10 sport con atleti più vecchi.

```

1 SELECT Sport, ROUND(Mean(Age),2) AS 'Eta Media'
2 FROM dataset
3 WHERE Age>0
4 GROUP BY Sport
5 ORDER BY 'Eta Media' DESC
6 LIMIT 10;

```

	Sport ▲	Età Media ▲
1	Roque	53.33
2	Art Competitions	45.9
3	Alpinism	38.81
4	Equestrian	37.88
5	Cycling Road	37.05
6	Powerlifting	35.51
7	Polo	35.46
8	Sitting Volleyball	34.61
9	Equestrianism	34.4
10	Wheelchair Fencing	34

Figura 3: Output seconda query Hive.

Top 10 sport con atleti più giovani.

```

1 SELECT Sport, ROUND(Mean(Age),2) AS 'Eta Media'
2 FROM dataset
3 WHERE Age>0
4 GROUP BY Sport
5 ORDER BY 'Eta Media'
6 LIMIT 10;

```

	Sport ▲	Età Media ▲
1	Rhythmic Gymnastics	18.74
2	Swimming	20.68
3	Figure Skating	22.24
4	Synchronized Swimming	22.37
5	Diving	22.48
6	Gymnastics	22.73
7	Short Track Speed Skating	22.81
8	Boxing	23.06
9	Alpine Skiing	23.21
10	Ski Jumping	23.35

Figura 4: Output terza query Hive.

Età media e numerosità per sesso.

```
1 SELECT Sex AS Sesso, round(Mean(Age),2) AS 'Eta Media', Count(*) AS 'n'
2 FROM dataset
3 WHERE Sex=="M" OR Sex=="F"
4 GROUP BY Sex;
```

	Sesso	Età Media	n
1	F	23.74	74357
2	M	26.28	196086

Figura 5: Output Quarta query Hive.

Top 5 nazioni per medaglie d'oro vinte.

```
1 SELECT Team AS Nazione, COUNT (*) AS 'Numero Ori'
2 FROM dataset
3 WHERE Medal==3
4 GROUP BY Team
5 ORDER BY 'Numero Ori' DESC
6 LIMIT 5;
```

	Nazione	Numero Ori
1	United States	2377
2	Soviet Union	1058
3	Germany	690
4	Great Britain	574
5	Italy	550

Figura 6: Output Quinta query Hive.

Top 5 sport per numero di partecipanti.

```
1 SELECT Sport, COUNT(Name) AS 'Numero Partecipanti'
2 FROM dataset
3 GROUP BY Sport, Sex
4 HAVING Sex=='F'
5 ORDER BY 'Numero Partecipanti' DESC
6 LIMIT 5;
```

	Sport	Numero Partecipanti
1	Athletics	11614
2	Swimming	9816
3	Gymnastics	9097
4	Alpine Skiing	3387
5	Cross Country Skiing	3385

Figura 7: Output Sesta query Hive.

Top 10 atleti per medaglie d'oro vinte e nazione di appartenenza.

```
1 SELECT 'Nome Atleta', 'Medaglie Vinte', Nazione
2 FROM(
3 SELECT Name AS 'Nome Atleta', Count(*) AS 'Medaglie Vinte'
4 FROM dataset
5 WHERE Medal==3
6 GROUP BY Name
7 ) RIGHT JOIN (
8 SELECT DISTINCT Name AS Nome, Team AS Nazione
9 FROM dataset
10 )
11 ON 'Nome Atleta'==Nome
12 ORDER BY 'Medaglie Vinte' DESC, 'Nome Atleta'
13 LIMIT 10;
```

	Nome Atleta ▲	Medaglie Vinte ▲	Nazione ▲
1	Michael Fred Phelps, II	23	United States
2	"Raymond Clarence ""Ray"" Ewry"	10	United States
3	"Frederick Carlton ""Carl"" Lewis"	9	United States
4	Larysa Semenivna Latynina (Diriy-)	9	Soviet Union
5	Mark Andrew Spitz	9	United States
6	Paavo Johannes Nurmi	9	Finland
7	"Jennifer Elisabeth ""Jenny"" Thompson (-Cumpelik)"	8	United States
8	"Matthew Nicholas ""Matt"" Biondi"	8	United States
9	Birgit Fischer-Schmidt	8	Germany
10	Birgit Fischer-Schmidt	8	East Germany

Figura 8: Output Settima query Hive.

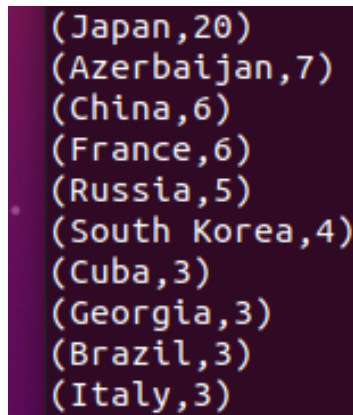
3.2 Pig

Crea la tabella dataset dal file csv caricato.

```
dataset = LOAD '/home/antonio/Olympic_Dataset.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(' ','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER')
AS (conto:int,Name:chararray,Sex:chararray,Age:int,Team:chararray,Games:chararray,
Year:int,Season:chararray,Sport:chararray,Medal:int)
```

Top 10 nazioni per medaglie d'oro vinte nel Judo.

```
1 select = FOREACH dataset GENERATE Team, Medal, Sport, Year;
2 filtro = FILTER select BY Medal==3 AND Sport=='Judo' AND Year>=2000;
3 gruppo = GROUP filtro BY Team;
4 conto = FOREACH gruppo GENERATE group, COUNT(filtro);
5 ordina = ORDER conto BY $1 DESC;
6 limita = LIMIT ordina 10;
7 DUMP limita;
```



```
(Japan,20)
(Azerbaijan,7)
(China,6)
(France,6)
(Russia,5)
(South Korea,4)
(Cuba,3)
(Georgia,3)
(Brazil,3)
(Italy,3)
```

Figura 9: Output prima query Pig.

Top 10 sport con atleti più vecchi.

```
1 select = FOREACH dataset GENERATE Sport, Age;
2 filtro = FILTER select BY Age is not null;
3 gruppo = GROUP filtro BY Sport;
4 conto = FOREACH gruppo GENERATE group, AVG(filtro.Age);
5 ordina = ORDER conto BY $1 DESC;
6 limita = LIMIT ordina 10;
7 DUMP limita;
```

```
(Roque,53.33333333333336)
(Art Competitions,45.90100944317812)
(Alpinism,38.8125)
(Equestrian,37.883116883116884)
(Cycling Road,37.04694835680751)
(Powerlifting,35.51123595505618)
(Polo,35.333333333333336)
(Sitting Volleyball,34.6096256684492)
(Equestrianism,34.39083075922614)
(Wheelchair Fencing,34.0)
```

Figura 10: Output seconda query Fig.

Top 10 sport con atleti più giovani.

```
1 select = FOREACH dataset GENERATE Sport, Age;
2 filtro = FILTER select BY Age is not null;
3 gruppo = GROUP filtro BY Sport;
4 conto = FOREACH gruppo GENERATE group, AVG(filtro.Age);
5 ordina = ORDER conto BY $1;
6 limita = LIMIT ordina 10;
7 DUMP limita;
```

```
(Rhythmic Gymnastics,18.737082066869302)
(Swimming,20.68395273899033)
(Figure Skating,22.232189973614776)
(Synchronized Swimming,22.36685082872928)
(Diving,22.48144064682102)
(Gymnastics,22.733038232528987)
(Short Track Speed Skating,22.804432855280314)
(Boxing,23.054808867167043)
(Alpine Skiing,23.20546223288767)
```

Figura 11: Output terza query Fig.

Età media e numerosità per sesso.

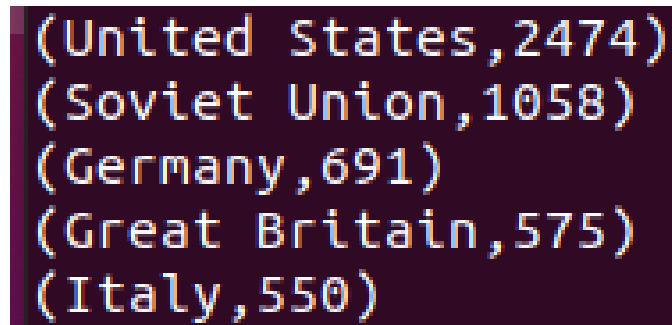
```
1 select = FOREACH dataset GENERATE Sex, Age;
2 filtro = FILTER select BY Sex=='M' OR Sex=='F';
3 gruppo = GROUP filtro BY Sex;
4 conto = FOREACH gruppo GENERATE group, AVG(filtro.Age), COUNT(filtro);
5 DUMP conto;
```

```
(F,23.732880779508218,74522)
(M,26.277561532227104,196594)
```

Figura 12: Output quarta query Fig.

Top 5 nazioni per medaglie d'oro vinte.

```
1 select = FOREACH dataset GENERATE Team, Medal;  
2 filtro = FILTER select BY Medal==3;  
3 gruppo = GROUP filtro BY Team;  
4 conto = FOREACH gruppo GENERATE group, COUNT(filtro);  
5 ordina = ORDER conto BY $1 DESC;  
6 limita = LIMIT ordina 5;  
7 DUMP limita;
```

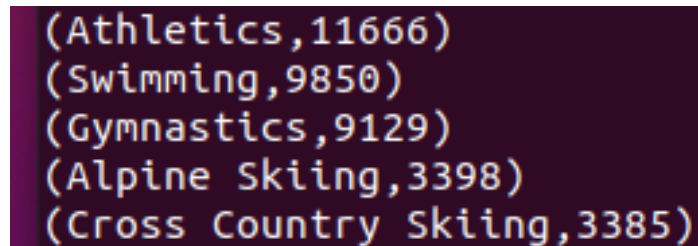


(United States,2474)
(Soviet Union,1058)
(Germany,691)
(Great Britain,575)
(Italy,550)

Figura 13: Output quinta query Fig.

Top 5 sport per numero di partecipanti.

```
1 select = FOREACH dataset GENERATE Sport, Sex;  
2 filtro = FILTER select BY Sex=='F';  
3 gruppo = GROUP filtro BY Sport;  
4 conto = FOREACH gruppo GENERATE group, COUNT(filtro);  
5 ordina = ORDER conto BY $1 DESC;  
6 limita = LIMIT ordina 5;  
7 DUMP limita;
```



(Athletics,11666)
(Swimming,9850)
(Gymnastics,9129)
(Alpine Skiing,3398)
(Cross Country Skiing,3385)

Figura 14: Output sesta query Fig.

Top 10 atleti per medaglie d'oro vinte e nazione di appartenenza.

```
1 select = FOREACH dataset GENERATE Name, Medal;
2 filtro = FILTER select BY Medal==3;
3 gruppo = GROUP filtro BY Name;
4 conto = FOREACH gruppo GENERATE group, COUNT(filtro);
5 ordina = ORDER conto BY $1;
6 select2 = FOREACH dataset GENERATE Name, Team;
7 filtro2 = DISTINCT select2;
8 unione = JOIN ordina BY $0 RIGHT, filtro2 BY $0;
9 ordina2 = ORDER unione BY $1 DESC;
10 final = FOREACH ordina2 GENERATE $0, $1, $3;
11 limita = LIMIT final 10;
12 DUMP limita;
```

```
(Michael Fred Phelps, II,23,United States)
(Raymond Clarence "Ray" Ewry,10,United States)
(Frederick Carlton "Carl" Lewis,9,United States)
(Larysa Semenivna Latynina (Diriy-),9,Soviet Union)
(Mark Andrew Spitz,9,United States)
(Paavo Johannes Nurmi,9,Finland)
(Jennifer Elisabeth "Jenny" Thompson (-Cumpelik),8,United States)
(Birgit Fischer-Schmidt,8,Germany)
(Ole Einar Bjrndalen,8,Norway)
(Usain St. Leo Bolt,8,Jamaica)
```

Figura 15: Output settima query Pig.