Garbarino Giacomo
Parmiggiani Manuel

DataWarehousing Project

Olympic Games Business Analysis

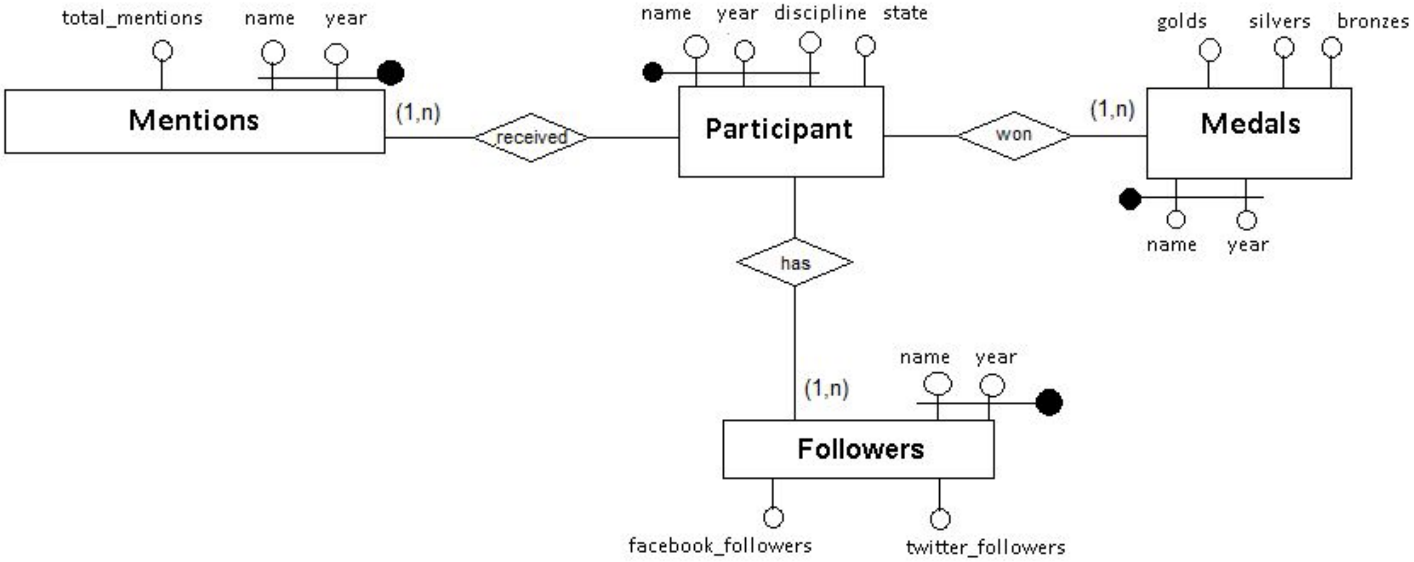
What’s the focus of our analysis?

- Popularity analysis of disciplines.
- Analysis of the athletes for each country to discover the most influent ones.
- Analysis of the reasons behind the popularity of the candidates.
- Comparison between candidates’ mentions and followers.
- Temporal analysis of candidates’ achievements.

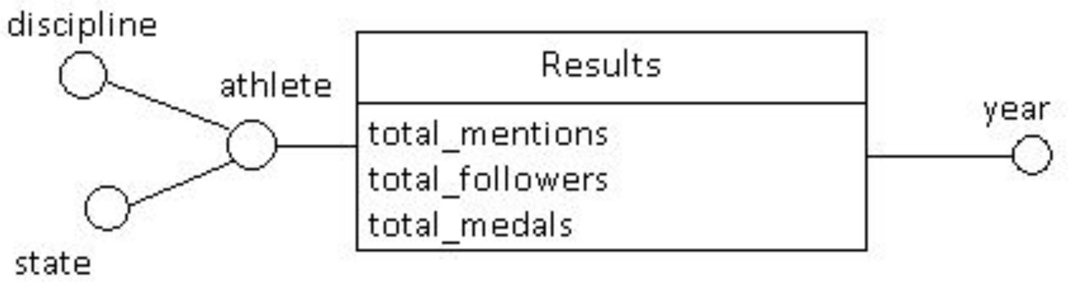
Datasets

- Athletes who participated in the Olympic Games of 2012, 2016 and 2020.
- Medals won by athletes during each edition.
- Received mentions and followers of the athletes during each edition.

Integrated data source's ER diagram

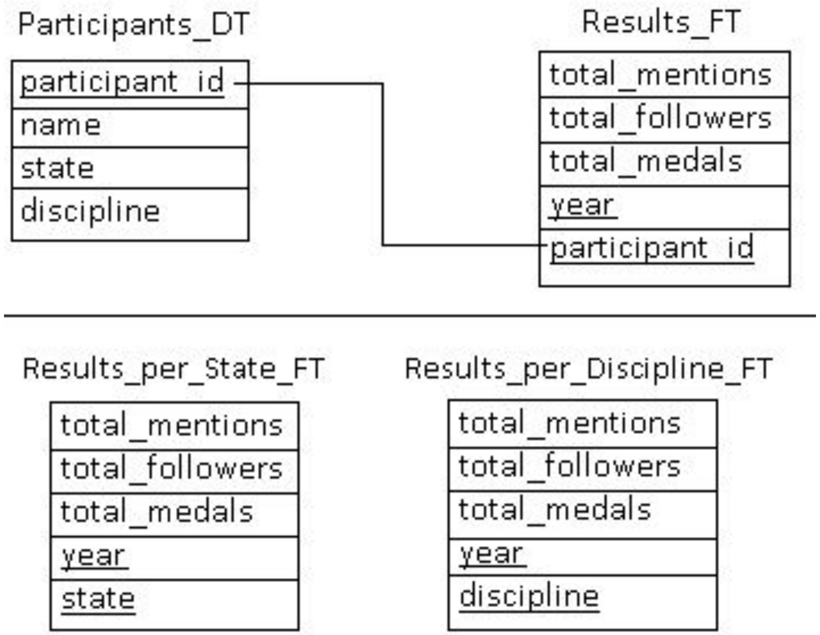


Conceptual design - DFM



- Two dimensions: temporal and geographical.
- Two types of measures: performance and popularity.

Logical design: multiple independent star schemas



Participants_DT
participant_id
name
state
discipline

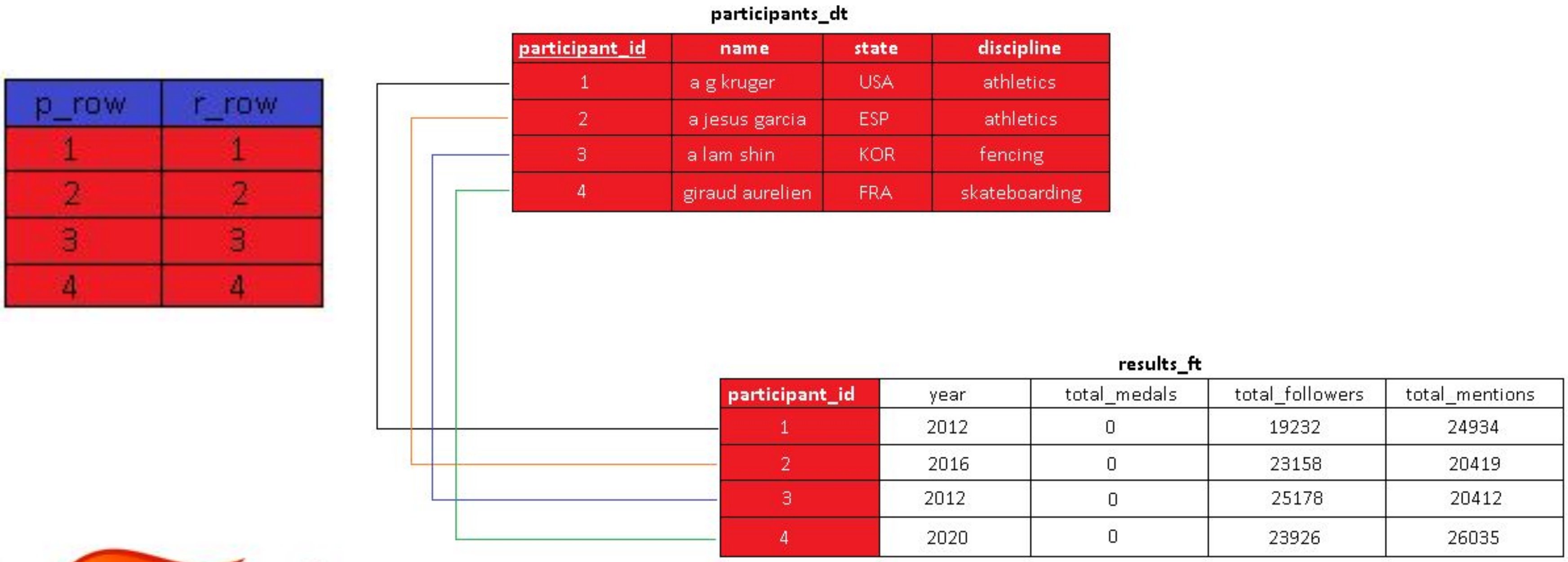
Results_FT
total_mentions
total_followers
total_medals
year
participant_id

Results_per_State_FT
total_mentions
total_followers
total_medals
year
state

Results_per_Discipline_FT
total_mentions
total_followers
total_medals
year
discipline

- Highest level of performance obtained by separating primary and secondary events.

Physical design: join index



p_row	r_row
1	1
2	2
3	3
4	4

participants_dt

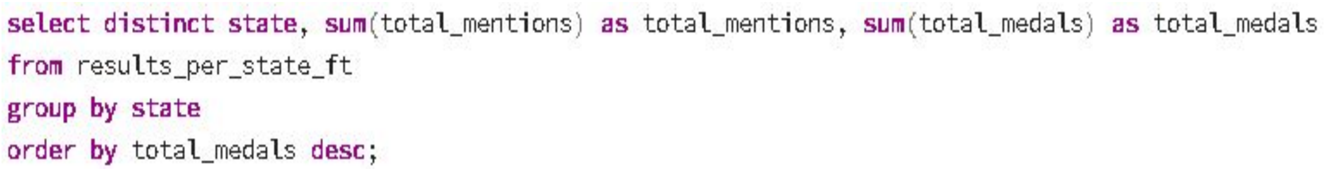
participant_id	name	state	discipline
1	a g kruger	USA	athletics
2	a jesus garcia	ESP	athletics
3	a lam shin	KOR	fencing
4	giraud aurelien	FRA	skateboarding

results_ft

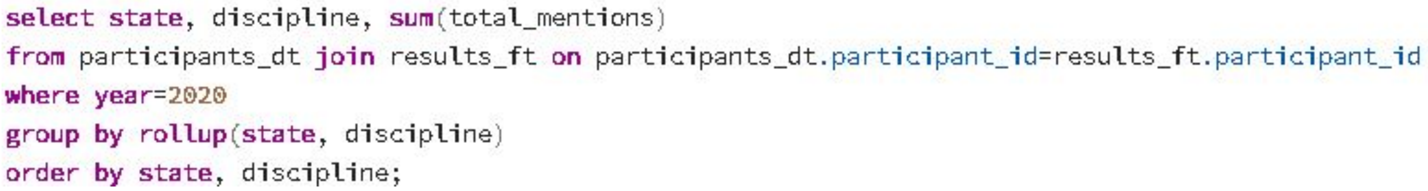
participant_id	year	total_medals	total_followers	total_mentions
1	2012	0	19232	24934
2	2016	0	23158	20419
3	2012	0	25178	20412
4	2020	0	23926	26035

- Optimization of the only one join we have to do.

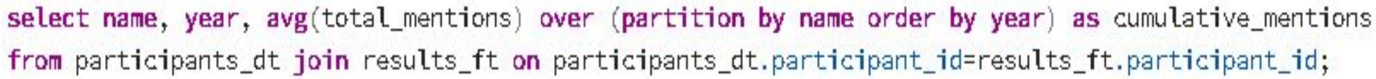
Some queries



```
select distinct state, sum(total_mentions) as total_mentions, sum(total_medals) as total_medals
from results_per_state_ft
group by state
order by total_medals desc;
```



```
select state, discipline, sum(total_mentions)
from participants_dt join results_ft on participants_dt.participant_id=results_ft.participant_id
where year=2020
group by rollup(state, discipline)
order by state, discipline;
```

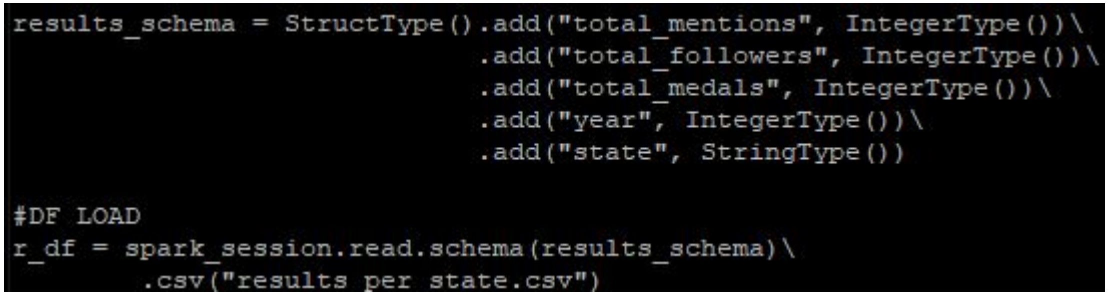


```
select name, year, avg(total_mentions) over (partition by name order by year) as cumulative_mentions
from participants_dt join results_ft on participants_dt.participant_id=results_ft.participant_id;
```

Extra - SparkSQL

- For analyzing the datasets in a large scale environment.
- Use of PySpark and advanced DF concepts.
- Two steps.

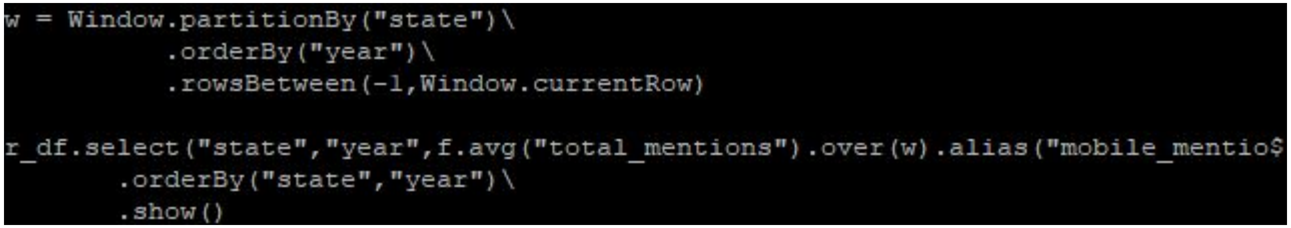
Step 1: schema creation and data loading.



```
results_schema = StructType().add("total_mentions", IntegerType())\
                             .add("total_followers", IntegerType())\
                             .add("total_medals", IntegerType())\
                             .add("year", IntegerType())\
                             .add("state", StringType())


#DF LOAD
r_df = spark_session.read.schema(results_schema)\
        .csv("results per state.csv")
```

Step 2: window definition (optional) and query execution.



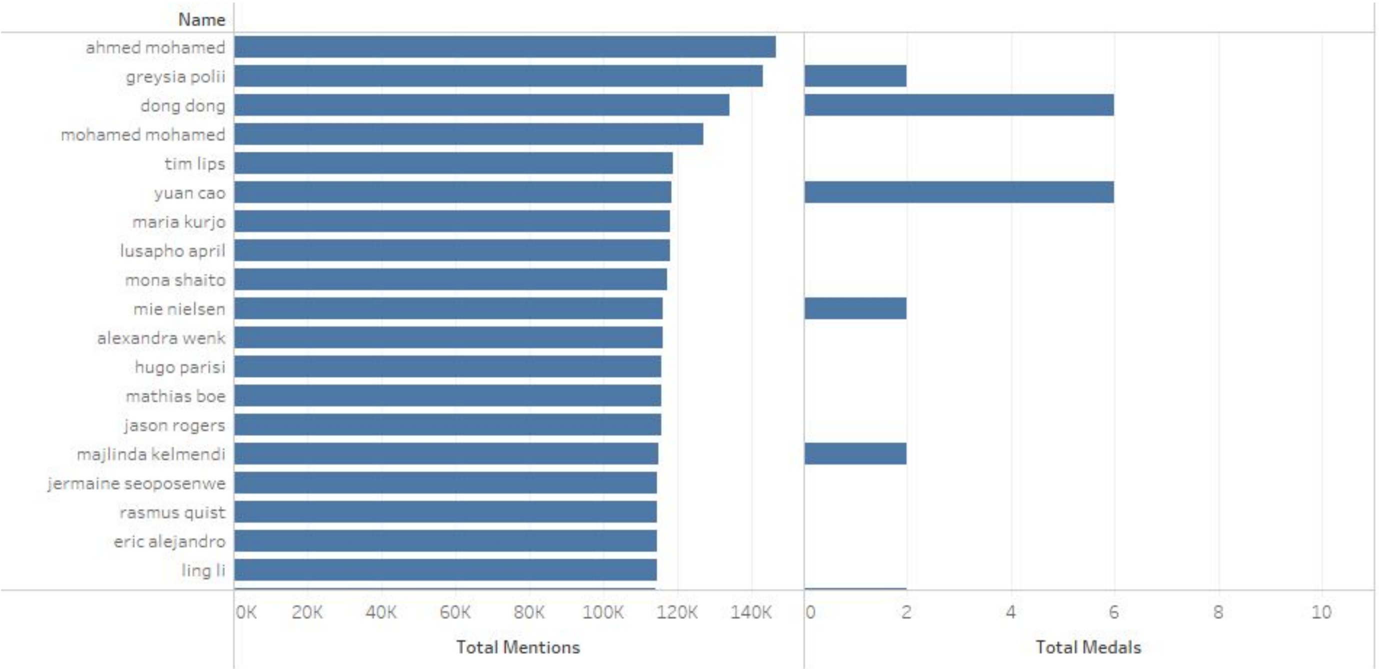
```
w = Window.partitionBy("state")\
        .orderBy("year")\
        .rowsBetween(-1,Window.currentRow)

r_df.select("state","year",f.avg("total_mentions").over(w).alias("mobile_mentio$
       .orderBy("state","year")\
       .show()
```

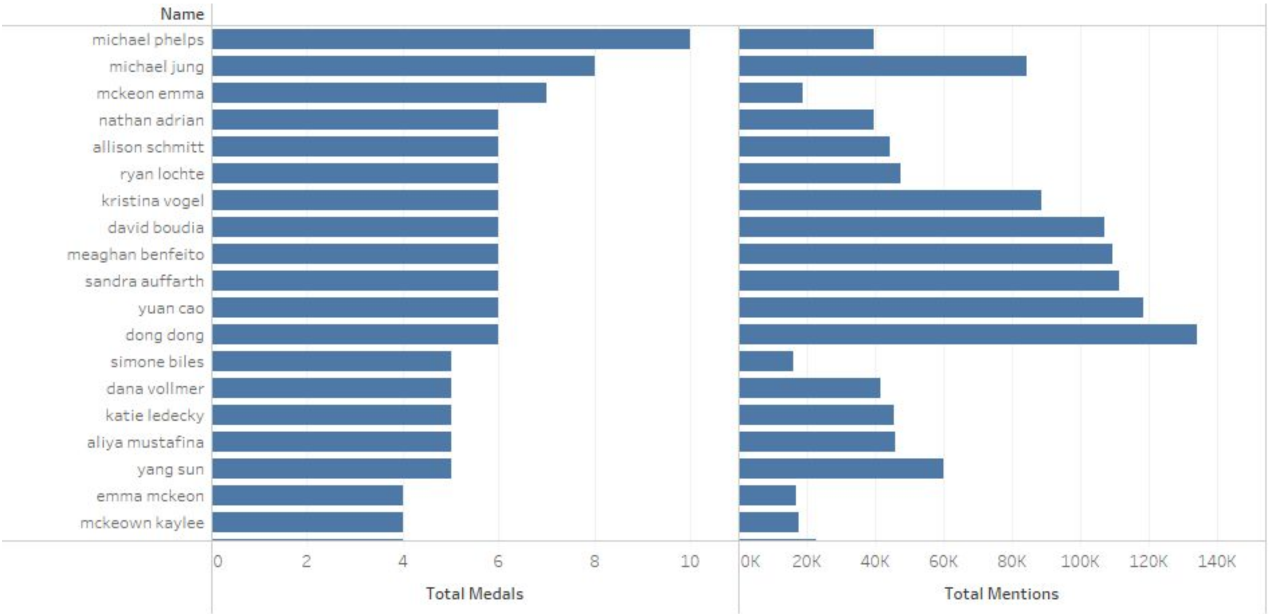
Most relevant outcomes

tableau

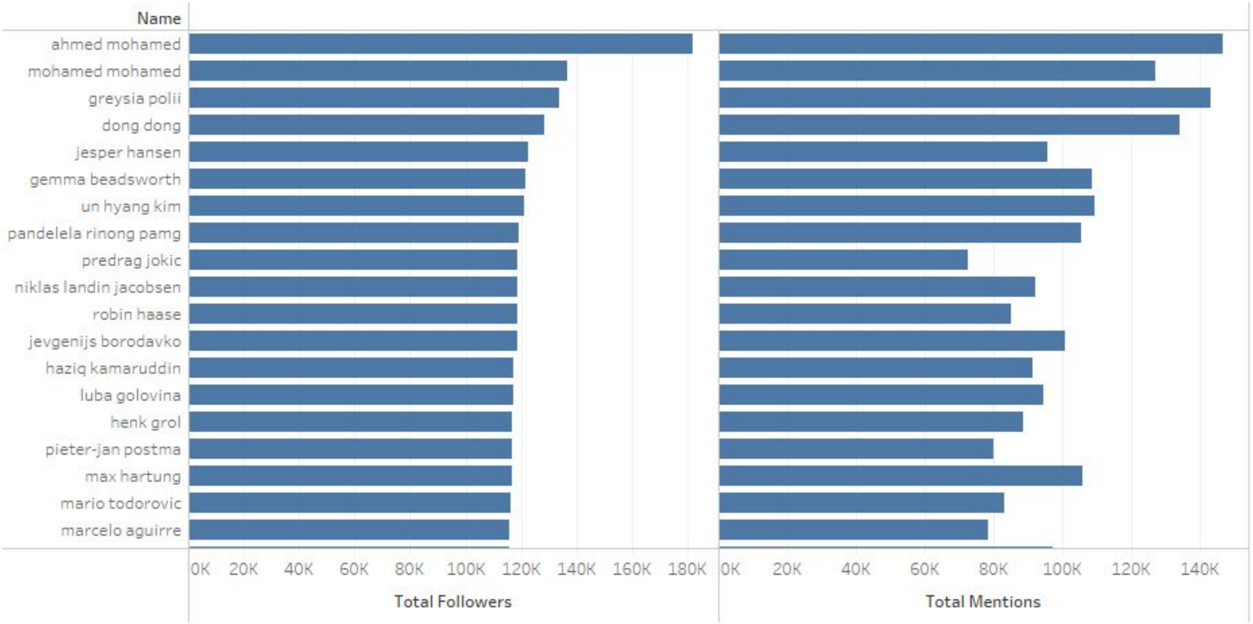
Medals won by the most mentioned athletes



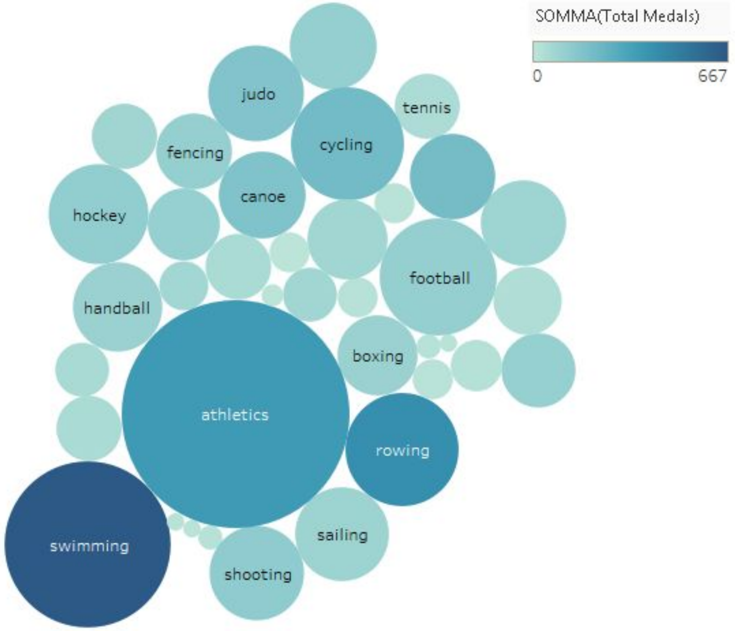
Received mentions of the best performing athletes



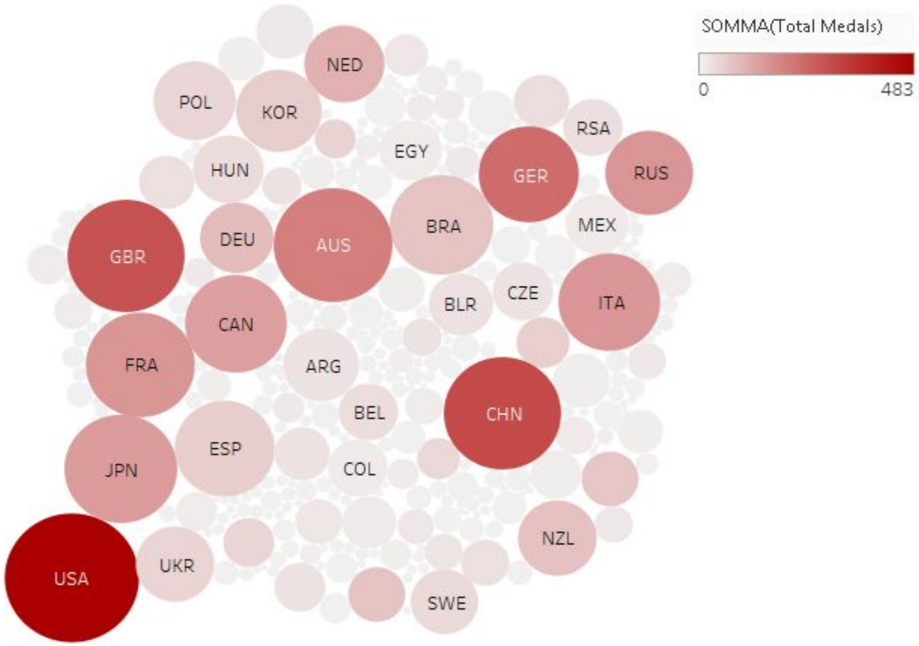
Received mentions of the most followed athletes



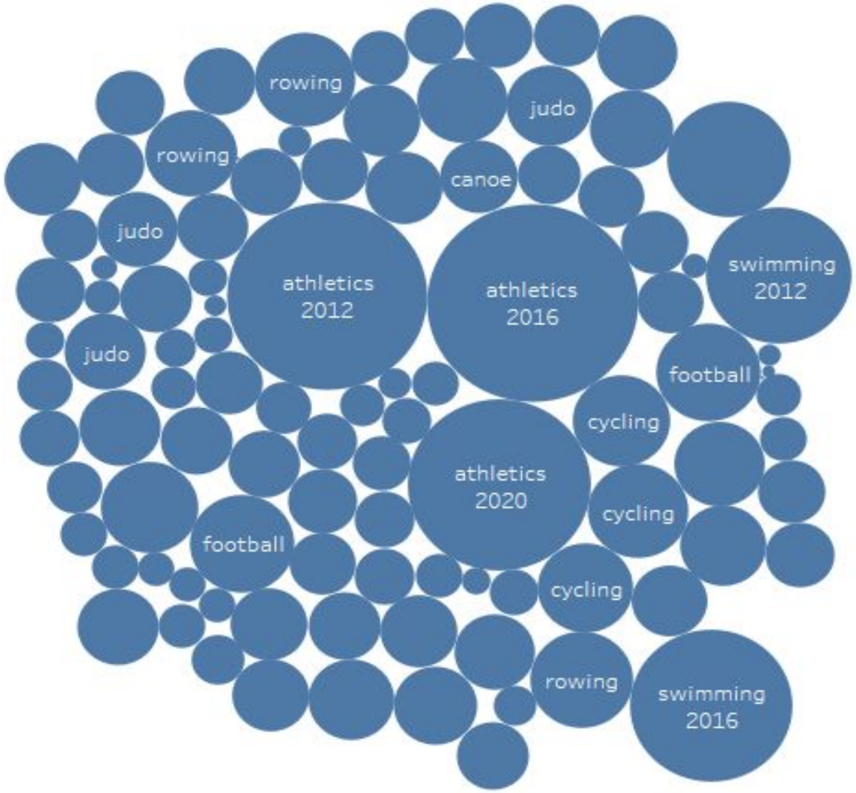
Comparison between mentions and medals per discipline



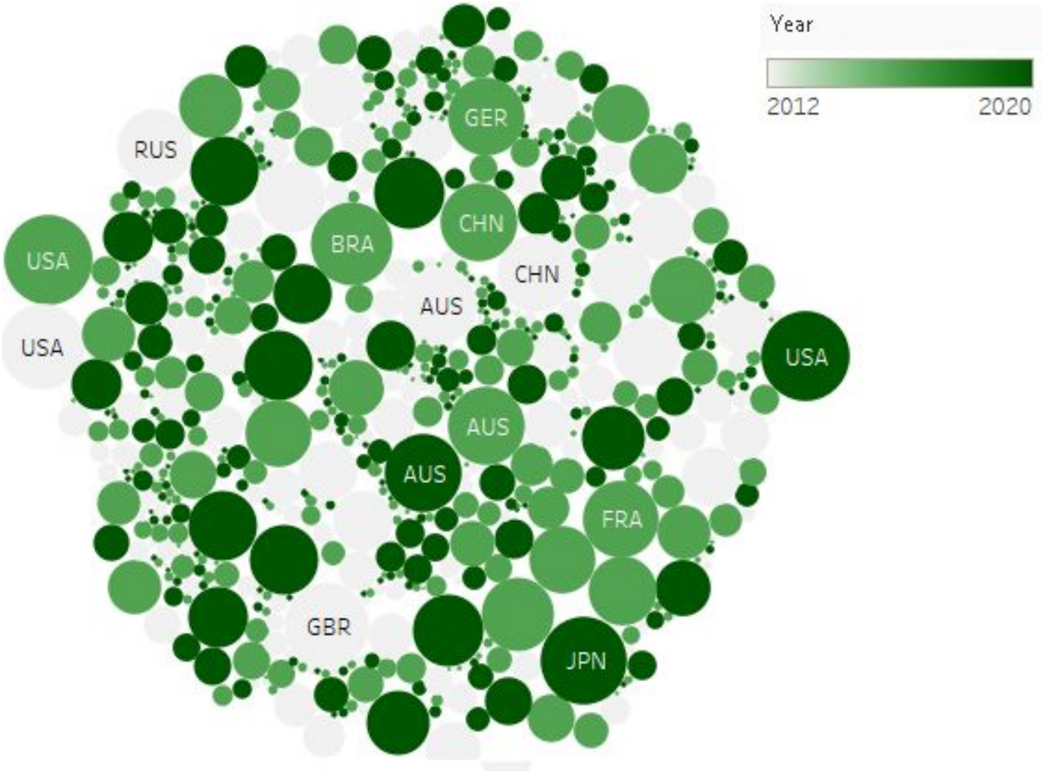
Comparison between mentions and medals per state



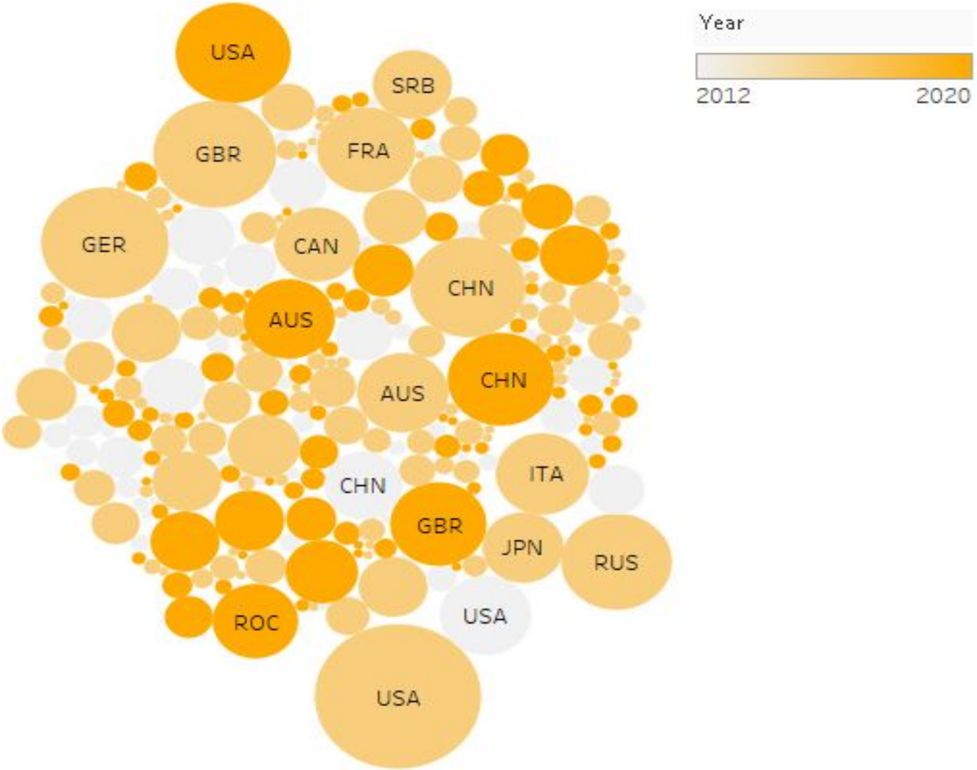
Most discussed disciplines per year



Total mentions per year and state



Total medals per state and year



Overall effort

- Almost 3 hours a day for more than 12 days:
    - About 7 hours for operational data sources inspection.
    - About 4 hours for conceptual design.
    - About 15 hours for logical design whose almost 13 for data cleaning and integration.
    - About 5 hours for OLAP queries.
    - About 3 hours for SparkSQL.
    - About 1 hour for Tableau.

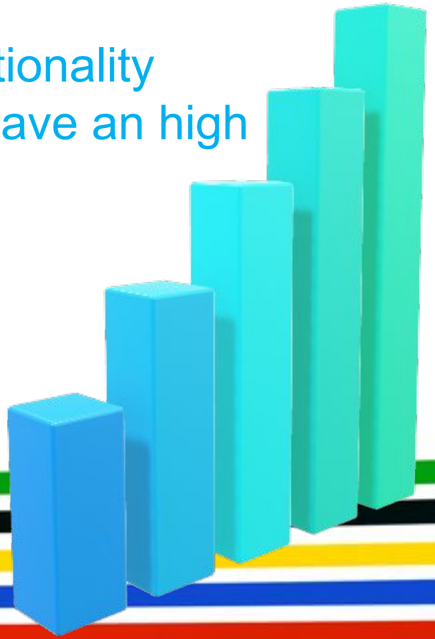
Teamwork

- Pair programming for step (1)-(4).
- Giacomo focused on step (6) and made the powerpoint presentation.
- Manuel focused on step (5) and discussed the presentation.

Conclusions

- We should invest in disciplines that are always popular across the years,then we have to choose a set of candidate testimonials
- For a short-term campaign, it’s better to sign the most popular athletes of the last edition, even if they didn’t win medals.
- For a long-term campaign, it’s better to sign the best performing athletes (in particular those who got significant results in the last two editions)

Finally, the choiche of the state can be done acording to the nationality of the athletes choosen as testimonials, expecially if they have an high audience in their homeland.



THANK YOU FOR THE PARTICIPATION!