Progetto 1

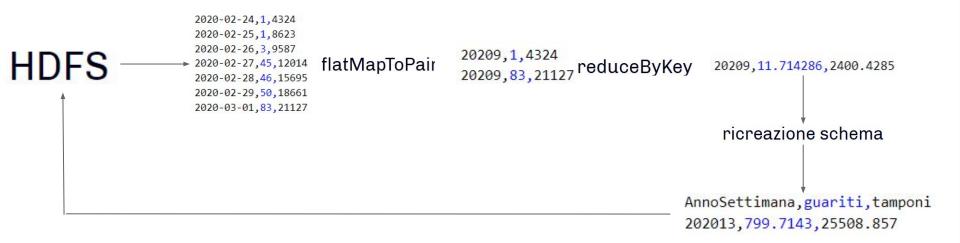
Damiano Nardi

Processamento delle query



- 1. Pull da HDFS
- 2. flatMapToPair
- 3. reduceByKey
- 4. ricreazione schema
- 5. push su HDFS

query1



Tempi Query1

I tempi sono in media su 5 misurazioni

	tempo (sec.)
Query processing	12,5
Spark loading	1,5

- 1. Pull da HDFS
- 2. flatMapToPair
- 3. top 100
- 4. mapToPair
- 5. Join
- 6. mapToPair
- 7. reduceByKey
- 8. flatMap
- 9. ricreazione schema.



top 100 → mapToPair

mapToPair nazione,CovidGlob[contagiati_list;regione;nazione] nazione,CovidGlob[contagiati_list;regione;nazione] nazione,CovidGlob[contagiati_list;regione;nazione]

Asia, Vietnam
Asia, Yemen
Europe, Albani
Oceania, Tuvai

Asia, Yemen
Europe, Albania mapToPair
Oceania, Tuvalu
Oceania, Vanuatu

Vietnam, Asia Yemen, Asia Albania, Europe Tuvalu, Oceania Vanuatu, Oceania nazione, CovidGlob, continente nazione, CovidGlob, continente nazione, CovidGlob, continente

Join

Join

```
nazione, CovidGlob, continente
nazione, CovidGlob, continente
nazione, CovidGlob, continente
```

mapToPair

```
Asia, contagiati list
Asia, contagiati_list
Europa, contagiati list
```

reduceByKey Asia, contagiati_list

```
Europa, contagiati list
```

```
IIIAsia, Asia, DEV WEEK, ..., ....
Key, Continente, Funzione, a2020s5, a2020s6, a2020s7,
                                                                                 ZZZAsia, Asia, MIN_WEEK,...,...
IIIAsia, Asia, DEV_WEEK,...,...
                                                                                 AAAAsia, Asia, MAX WEEK, ..., ...
ZZZAsia, Asia, MIN WEEK, ..., ...
                                                                                 RRRAsia, Asia, AVG_WEEK,...,...
AAAAsia, Asia, MAX WEEK, ..., ...
                               ricreazione schema -
RRRAsia, Asia, AVG WEEK, ..., ....
                                                                                 AAAEurope, Europe, MAX_WEEK,...,...
AAAEurope, Europe, MAX_WEEK, ..., ...
IIIEurope, Europe, DEV_WEEK,...,...
                                                                                 IIIEurope, Europe, DEV WEEK, ..., ...
RRREurope, Europe, AVG WEEK, ..., ...
                                                                                 RRREurope, Europe, AVG WEEK, ..., ...
ZZZEurope, Europe, MIN WEEK, ..., ....
                                                                                 ZZZEurope, Europe, MIN WEEK,...,...
```

flatMap

ricreazione schema

```
Key,Continente,Funzione,a2020s5,a2020s6,a2020s7,
IIIAsia,Asia,DEV_WEEK,...,...

ZZZAsia,Asia,MIN_WEEK,...,...

AAAAsia,Asia,MAX_WEEK,...,...

RRRAsia,Asia,AVG_WEEK,...,...

AAAEurope,Europe,MAX_WEEK,...,...

IIIEurope,Europe,DEV_WEEK,...,...

RRREurope,Europe,AVG_WEEK,...,...

ZZZEurope,Europe,MIN_WEEK,...,...
```



Tempi Query2

I tempi sono in media su 5 misurazioni

	tempo (sec.)
Query processing	16,63
Spark loading	1,3

Framework utilizzati



Storage



















































2 salvataogio su HDFS in parquet

3 Comunicazione per iniziare il job















3 Comunicazione per iniziare il job



4 start spark job









4 start spark job



3 Comunicazione per iniziare il job











4 start spark job





3 Comunicazione per iniziare il job



1 pull dei dati da gitHub







6 inserimento dati proces



4 start spark job

7 job terminato



3 Comunicazione per iniziare il job









6 inserimento dati productati

4 start spark job

7 job terminato

1 pull dei dati da gitHub





3 Comunicazione per iniziare il job

8 job terminato









4 start spark job

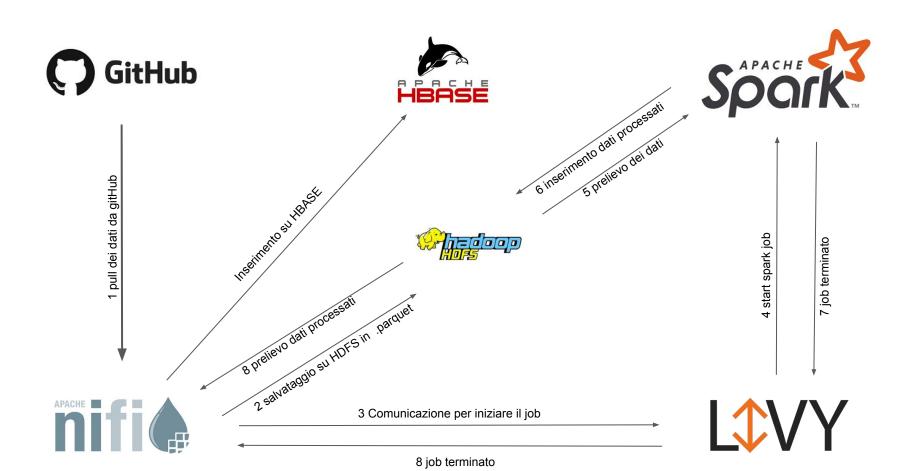
7 job terminato



3 Comunicazione per iniziare il job

8 job terminato





Grazie dell'attenzione