

# Progetto 1

Damiano Nardi

# Processamento delle query



# Query 1

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

# query1

HDFS

2020-02-24,1,4324  
2020-02-25,1,8623  
2020-02-26,3,9587  
2020-02-27,45,12014  
2020-02-28,46,15695  
2020-02-29,50,18661  
2020-03-01,83,21127

flatMapToPair

20209,1,4324

20209,83,21127

reduceByKey

20209,11.714286,2400.4285

ricreazione schema

AnnoSettimana,guariti,tamponi  
202013,799.7143,25508.857

# Tempi Query1

I tempi sono in media su 5 misurazioni

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

# Query 2

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

# Query2



# Query2

top 100

treadlineCoff,

→ mapToPair

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

HDFS

```
Asia,Vietnam  
Asia,Yemen  
Europe,Albania  
Oceania,Tuvalu  
Oceania,Vanuatu
```

mapToPair

```
Vietnam,Asia  
Yemen,Asia  
Albania,Europe  
Tuvalu,Oceania  
Vanuatu,Oceania
```

Join

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



# Query2

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

Key,Continente,Funzione,a2020s5,a2020s6,a2020s7,  
IIIIAsia,Asia,DEV\_WEEK,...,  
ZZZAsia,Asia,MIN\_WEEK,...,  
AAAAAsia,Asia,MAX\_WEEK,...,  
RRRAsia,Asia,AVG\_WEEK,...,  
AAAEurope,Europe,MAX\_WEEK,...,  
IIIEurope,Europe,DEV\_WEEK,...,  
RRREurope,Europe,AVG\_WEEK,...,  
ZZZEurope,Europe,MIN\_WEEK,...

ricreazione schema

IIIIAsia,Asia,DEV\_WEEK,...,  
ZZZAsia,Asia,MIN\_WEEK,...,  
AAAAAsia,Asia,MAX\_WEEK,...,  
RRRAsia,Asia,AVG\_WEEK,...,  
AAAEurope,Europe,MAX\_WEEK,...,  
IIIEurope,Europe,DEV\_WEEK,...,  
RRREurope,Europe,AVG\_WEEK,...,  
ZZZEurope,Europe,MIN\_WEEK,...

flatMap

# Query2

ricreazione schema

```
Key,Continente,Funzione,a2020s5,a2020s6,a2020s7,  
IIIIAsia,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,...,...
```



HDFS

# Tempi Query2

I tempi sono in media su 5 misurazioni

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

# Framework utilizzati

Ingest



REST API



Process



Storage







1 pull dei dati da gitHub





1 pull dei dati da gitHub



2 salvataggio su HDFS in .parquet





1 pull dei dati da gitHub



2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job







1 pull dei dati da gitHub



2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job

4 start spark job





1 pull dei dati da gitHub

2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job

5 prelievo dei dati

4 start spark job



1 pull dei dati da gitHub

2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job

6 inserimento dati processati

5 prelievo dei dati

4 start spark job



1 pull dei dati da gitHub

2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job

6 inserimento dati processati

5 prelievo dei dati

4 start spark job

7 job terminato



1 pull dei dati da gitHub

2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job

8 job terminato

6 inserimento dati processati  
5 prelievo dei dati

4 start spark job

7 job terminato



1 pull dei dati da gitHub

8 prelievo dati processati

2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job

8 job terminato

6 inserimento dati processati

5 prelievo dei dati

4 start spark job

7 job terminato



1 pull dei dati da gitHub

Inserimento su HBASE

8 prelievo dati processati

2 salvataggio su HDFS in .parquet

3 Comunicazione per iniziare il job

8 job terminato

6 inserimento dati processati

5 prelievo dei dati

4 start spark job

7 job terminato

**Grazie  
dell'attenzione**