Università degli studi di Torino

Modelli e Architetture Avanzate di Basi di Dati

14 Dicembre 2021

Progetto di

Pierandrea Morelli

Giulio Piazza

Enrico Rizzello

Project Goals

- Creare uno strumento per l'analisi dei tweets per il conteggio delle parole, degli hastag e delle emoji/emoticon
- Implementare due differenti soluzioni, una mediante un database relazionale e una con un database non relazionale
- Generazione delle world cloud per entrambe le soluzioni e confronto delle parole, risorse lessicali e tweet

Scelte implementative

- MariaDB è uno dei database relazionali open source più popolari. È realizzato dagli sviluppatori originali di MySQL e garantito per rimanere open source. Fa parte della maggior parte delle offerte cloud ed è l'impostazione predefinita nella maggior parte delle distribuzioni Linux.

- E’ un database di tipo documentale non relazionale che permette l’annidamento di oggetti in altri e rende persistente questi documenti di tipo testuale. Permette un linguaggio su questi documenti molto più ricco e di andare a ricercare anche elementi annidati senza avere i vincoli di uno schema rigido

Schema dei dati

- Emoticon

Field	Type	Null	Key	Default	Extra
ID	int(11)	NO	PRI		auto_increment
Code	varchar(30)	NO		0	
Polarity	smallint(6)	NO		0	

- EmoticonCount

Field	Type	Null	Key	Default	Extra
Emotion	varchar(15)	NO	PRI		
IDEmoticon	int(11)	NO	PRI		
Count	int(11)	NO			

- HashtagCount

Field	Type	Null	Key	Default	Extra
Emotion	varchar(15)	NO	PRI		
Hashtag	varchar(100)	NO	PRI		
Count	int(11)	NO		0	

- Tweet

Field	Type	Null	Key	Default	Extra
ID	int(11)	NO	PRI		auto_increment
Emotion	varchar(15)	NO		0	
Text	varchar(1000)	NO		0	

Schema dei dati

- NegativeWord

Field	Type	Null	Key	Default	Extra
Word	varchar(15)	NO	PRI		

- Slang

Field	Type	Null	Key	Default	Extra
Slang	varchar(15)	NO	PRI		
Traduction	varchar(50)	NO			

- StopWord

Field	Type	Null	Key	Default	Extra
Word	varchar(50)	NO	PRI	0	

- WordCount

Field	Type	Null	Key	Default	Extra
Emotion	varchar(15)	NO	PRI		
Word	varchar(200)	NO	PRI		
Count	int(11)	NO		0	
FlagSentisense	tinyint(4)	NO		0	
FlagNRC	tinyint(4)	NO		0	
FlagEmoSN	tinyint(4)	NO		0	

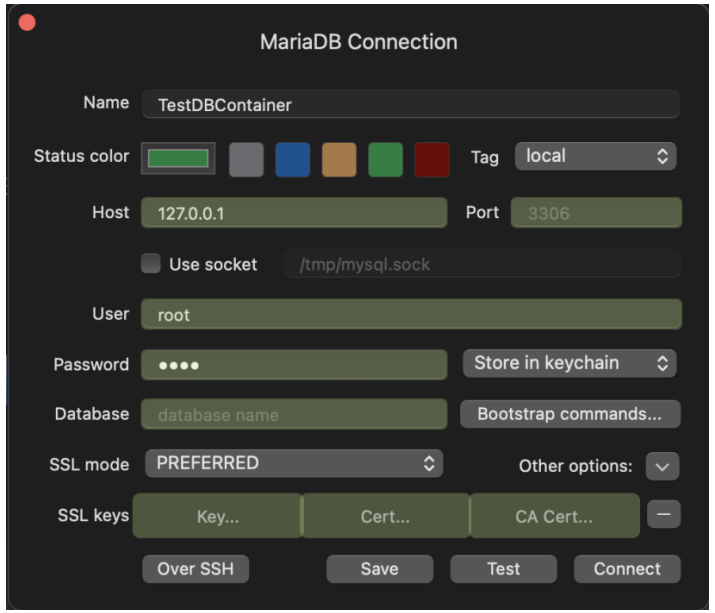
Docker - MariaDB

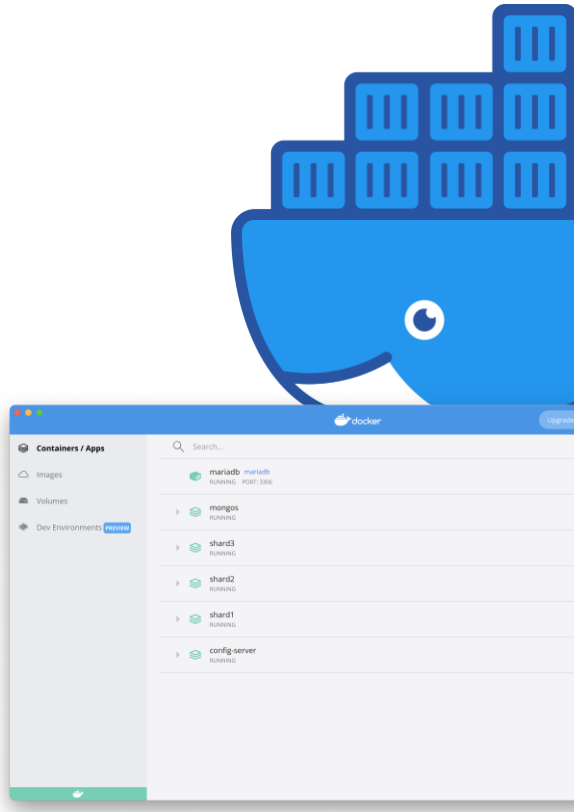
- **Pull image da DockerHub**

```
docker pull mariadb
```

- **Running container**

```
docker run --name mariadb –e MYSQL_ROOT_PASSWORD=admin -p 3306:3306 -d mariadb
```





Docker - MongoDB

- **Config Server**

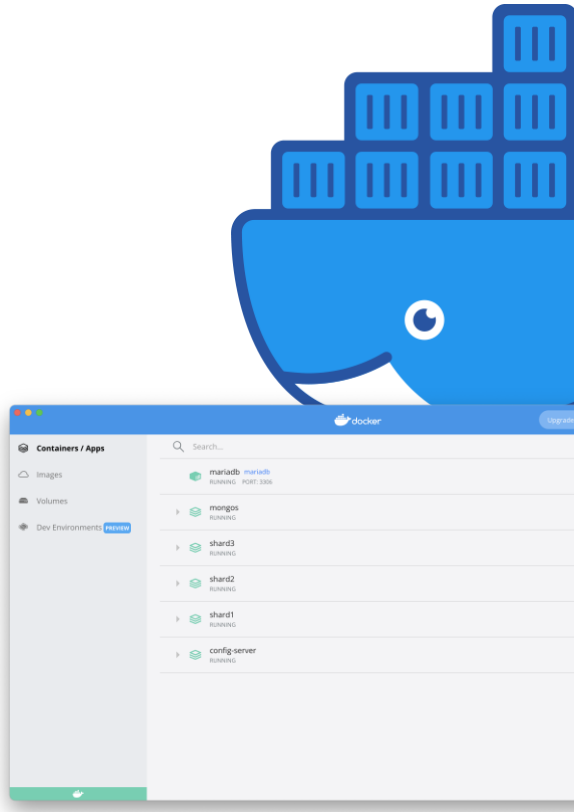
Starting config Server

```
docker-compose -f config-server/docker-compose.yaml up -d
mongo mongodb://192.168.1.140:40001
rs.initiate(
 {
  _id: "cfgrs",
  configsvr: true,
  members: [
   { _id : 0, host : "192.168.1.140:40001" }
  ]
 }
)
```

- **Shard** ✖3

Start shard 1 servers

```
docker-compose -f shard1/docker-compose.yaml up -d
mongo mongodb://192.168.1.140:50001
rs.initiate(
 {
  _id: "shard1rs",
  members: [
   { _id : 2, host : "192.168.1.140:50001" }
  ]
 }
)
```



Docker - MongoDB

- **Mongos Router**

Start mongos query router

docker-compose -f mongos/docker-compose.yaml up –d

Connect to mongos

mongo mongodb://192.168.1.140:60000

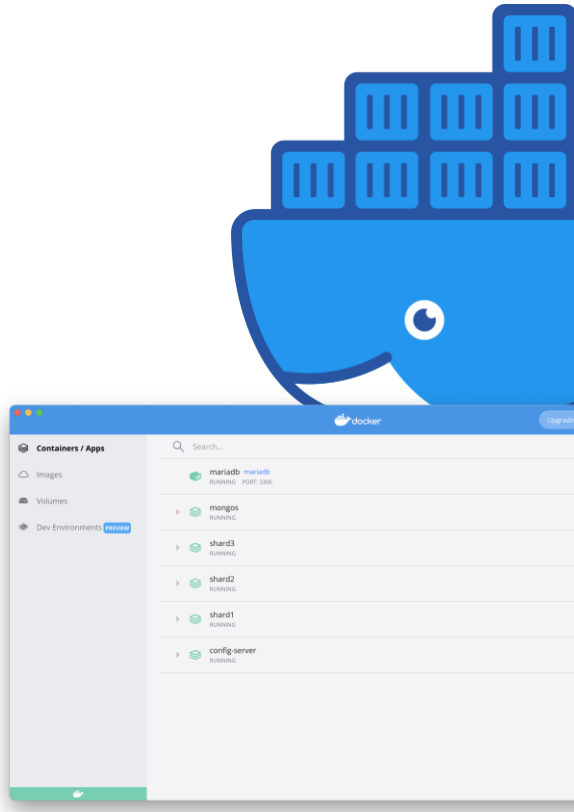
Add shard

mongos> sh.addShard("shard1rs/192.168.1.140:50001")

mongos> sh.addShard("shard2rs/192.168.1.140:50002")

mongos> sh.addShard("shard3rs/192.168.1.140:50003")

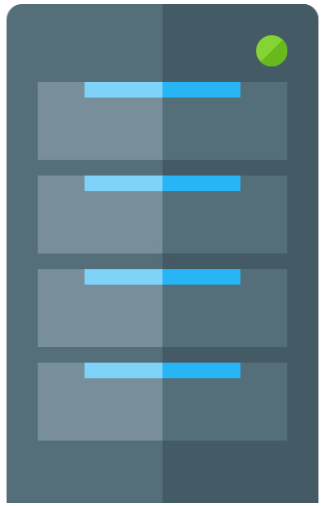
- Le istanze **mongos** di MongoDB instradano le query e scrivono le operazioni sugli shard in un cluster sharded. mongos fornisce l'unica interfaccia a un cluster partizionato dal punto di vista delle applicazioni. Mongos tiene traccia di quali dati si trovano su quale frammento memorizzando nella cache i metadati dai server di configurazione. Mongos utilizza i metadati per instradare le operazioni dalle applicazioni e dai client alle istanze mongod. Un mongos non ha uno stato persistente e consuma risorse di sistema minime.

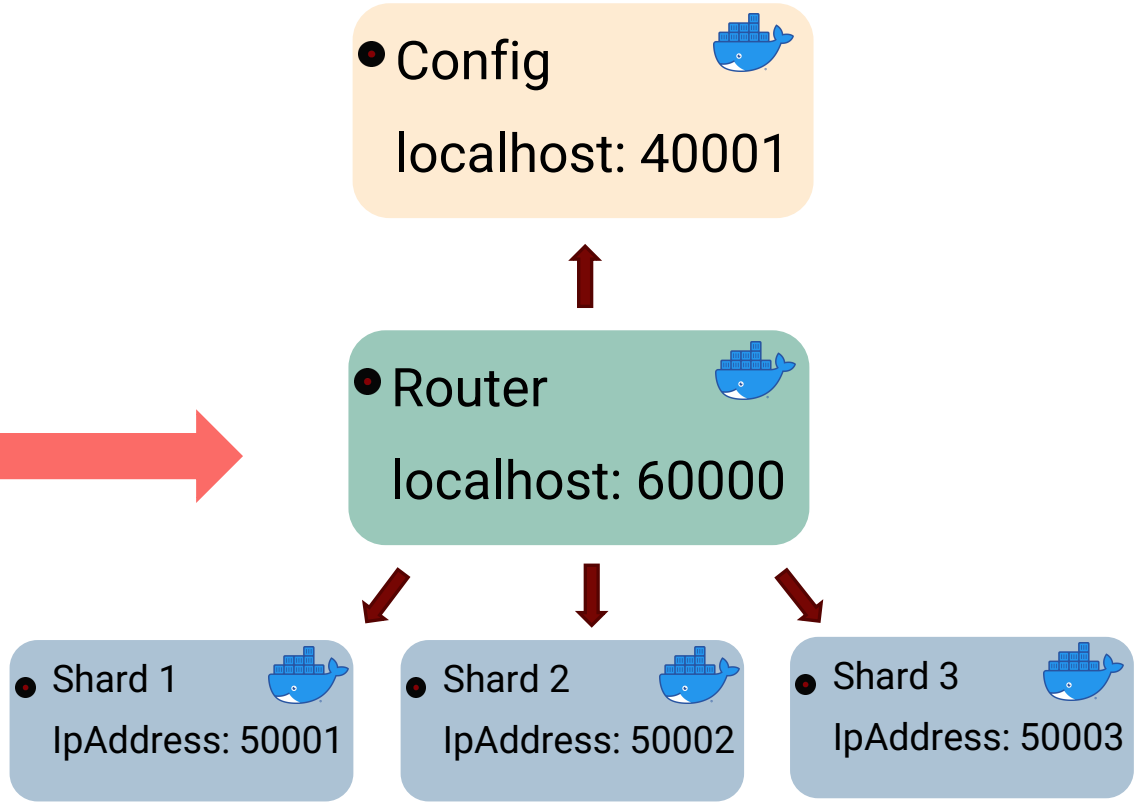


- **Enable Sharding**

```
sh.enableSharding("TwitterEmotions")
db.adminCommand({movePrimary : "TwitterEmotions", to: "shard1rs"})
sh.shardCollection("TwitterEmotions.Tweet", { _id : "hashed" } )
```

Sharded Cluster





Map Reduce

- Essendo deprecata l’operazione di map reduce da mongo 5.0 abbiamo utilizzato un aggregation pipeline che offre prestazioni e usabilità migliorate rispetto alle operazioni di map reduce tradizionali.

  Le operazioni di Map Reduce possono essere riscritte mediante gli operatori di aggregation pipeline come **$group** e **$merge**

- Un’aggregation pipeline è costituita da una o più fasi che elaborano i documenti:

  Ogni fase esegue un'operazione sui documenti di input. Ad esempio, una fase può filtrare documenti, raggruppare documenti e calcolare valori. I documenti in output da una fase vengono inseriti nella fase successiva.Un’aggregation pipeline può restituire risultati per gruppi di documenti. Ad esempio, i valori totale, medio, massimo e minimo.

- **Pipeline delle Words**

```
    pipeline = [
     {"$unwind": "$Words"},
     {"$group": {"_id": {"Emotion": "$Emotion", "Word": "$Words"}, "Count": {"$sum": 1}}}
]
db.Tweet.aggregate(pipeline, allowDiskUse=True)
```

Preprocessing Tweets

Pulizia dei messaggi

Cancellazione degli URL e USERNAME

Processing degli

Hastag, emoticon ed emoji

Normalizzazione

Di slang e parole

Tokenizzazione e steamming

Eliminazione delle stop word

Calcolo della

Stem frequency

Emojicon Clouds

Anger

Disgust

Joy

Fear

Emojicon Clouds

Trust

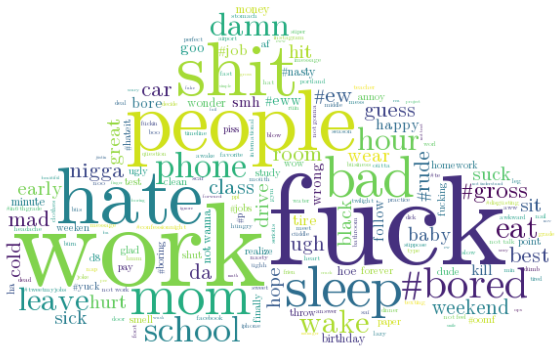
Anticipation

Sadness

Surprise

Word Clouds

Anger



Disgust

Joy

Fear

Word Clouds

Trust

Anticipation

Sadness

Surprise

Results tables

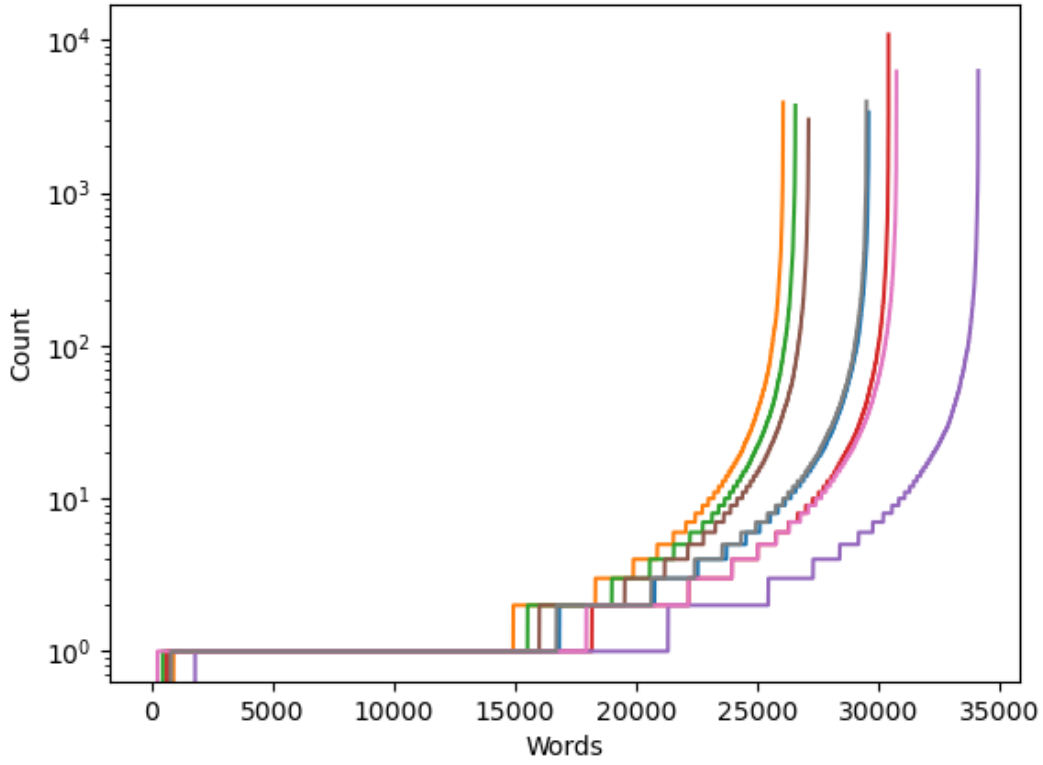
Emotion	Sentisense	NRC	EmoSN
anger	0.8	15.2	5.5
anticipation	3.0	13.1	N.D.
disgust	7.4	12.4	N.D.
fear	3.0	16.8	N.D.
joy	1.6	11.6	47.9
sadness	1.8	15.0	N.D.
surprise	0.4	8.3	N.D.
trust	N.D.	16.8	N.D.

Emotion	Sentisense	NRC	EmoSN
anger	56.3	46.9	56.2
anticipation	63.1	49.8	N.D.
disgust	49.6	44.6	N.D.
fear	62.5	40.0	N.D.
joy	43.7	59.2	59.3
sadness	40.9	41.6	N.D.
surprise	51.7	54.6	N.D.
trust	N.D.	48.3	N.D.

Emotion	anger	New Words:	2978
Emotion	anticipation	New Words:	2531
Emotion	disgust	New Words:	3000
Emotion	fear	New Words:	2685
Emotion	joy	New Words:	2145
Emotion	sadness	New Words:	2577
Emotion	surprise	New Words:	2986
Emotion	trust	New Words:	2693

Istogramma

- A ogni colore è associata un’emozione diversa



Confronto MariaDB e MongoDB

- Tempo di esecuzione relativo alla Tweeter analysis

00:20:13

- Tempo di esecuzione relativo alla Tweeter analysis con sharding

00:03:46

Grazie per l’attenzione