

# Relazione Progettino

Gabriele Bisogno

## 1 Introduzione

Link Dataset: <https://www.kaggle.com/donorschoose/io>.

A questo link si trova questa relazione e i vari dataset, che descrivono le donazioni e i donatori di determinati progetti inerenti alle scuole e ai loro insegnanti.

Ho utilizzato i file Donors.csv, Donations.csv. Donors contiene 5 attributi:

- Donor ID
- Donor City
- Donor State
- Donor Is Teacher
- Donor Zip

Donations contiene 7 attributi:

- Project ID
- Donation ID
- Donor ID
- Donation Included Optional Donation
- Donation Amount
- Donor Cart Sequence
- Donation Received Date

Inoltre ho utilizzato come client, per gestire le interrogazioni, Robo3T.

## 2 Import Dataset

Avvio Mongodb con le varie istruzioni, `mongodb --dbpath //path` e `mongo`, salvando i vari dataset nella cartella indicata dal percorso `//path`.

Utilizzo le seguenti istruzioni per importare il dataset:

- `mongoimport -d progettino -c donors --type csv --file /Users/gabrielebisogno/Desktop/Donors.csv --headerline`
- `mongoimport -d progettino -c donators --type csv --file /Users/gabrielebisogno/Desktop/Donators.csv --headerline`

```
MacBook-Pro-di-Gabriele:~ gabrielebisogno$ mongoimport -d progettino -c donationsv --type csv --file /Users/gabrielebisogno/Desktop/progettino/Donationsv.csv --headerline
connected to: localhost
2018-09-18T22:47:33.561+0200 [.....] progettino.donationsv 21.5MB/583MB (3.7%)
2018-09-18T22:47:36.561+0200 [#.....] progettino.donationsv 43.7MB/583MB (7.5%)
2018-09-18T22:47:39.558+0200 [##.....] progettino.donationsv 64.9MB/583MB (11.1%)
2018-09-18T22:47:42.559+0200 [###.....] progettino.donationsv 86.9MB/583MB (14.9%)
2018-09-18T22:47:48.558+0200 [####.....] progettino.donationsv 109MB/583MB (18.7%)
2018-09-18T22:47:51.559+0200 [#####.....] progettino.donationsv 130MB/583MB (22.3%)
2018-09-18T22:47:54.560+0200 [#####.....] progettino.donationsv 151MB/583MB (26.0%)
2018-09-18T22:47:57.560+0200 [#####.....] progettino.donationsv 173MB/583MB (29.7%)
2018-09-18T22:48:00.558+0200 [#####.....] progettino.donationsv 195MB/583MB (33.4%)
2018-09-18T22:48:03.561+0200 [#####.....] progettino.donationsv 217MB/583MB (37.2%)
2018-09-18T22:48:06.563+0200 [#####.....] progettino.donationsv 237MB/583MB (40.6%)
2018-09-18T22:48:09.560+0200 [#####.....] progettino.donationsv 257MB/583MB (44.1%)
2018-09-18T22:48:12.559+0200 [#####.....] progettino.donationsv 278MB/583MB (47.7%)
2018-09-18T22:48:15.559+0200 [#####.....] progettino.donationsv 299MB/583MB (51.3%)
2018-09-18T22:48:18.561+0200 [#####.....] progettino.donationsv 321MB/583MB (55.0%)
2018-09-18T22:48:21.560+0200 [#####.....] progettino.donationsv 342MB/583MB (58.7%)
2018-09-18T22:48:24.561+0200 [#####.....] progettino.donationsv 363MB/583MB (62.3%)
2018-09-18T22:48:27.559+0200 [#####.....] progettino.donationsv 384MB/583MB (65.9%)
2018-09-18T22:48:30.559+0200 [#####.....] progettino.donationsv 406MB/583MB (69.6%)
2018-09-18T22:48:33.562+0200 [#####.....] progettino.donationsv 427MB/583MB (73.2%)
2018-09-18T22:48:36.560+0200 [#####.....] progettino.donationsv 448MB/583MB (76.8%)
2018-09-18T22:48:39.559+0200 [#####.....] progettino.donationsv 468MB/583MB (80.3%)
2018-09-18T22:48:42.560+0200 [#####.....] progettino.donationsv 490MB/583MB (84.1%)
2018-09-18T22:48:45.559+0200 [#####.....] progettino.donationsv 512MB/583MB (87.8%)
2018-09-18T22:48:48.559+0200 [#####.....] progettino.donationsv 532MB/583MB (91.3%)
2018-09-18T22:48:51.559+0200 [#####.....] progettino.donationsv 552MB/583MB (94.7%)
2018-09-18T22:48:54.561+0200 [#####.....] progettino.donationsv 573MB/583MB (98.3%)
2018-09-18T22:48:55.986+0200 [#####.....] progettino.donationsv 583MB/583MB (100.0%)
```

Figure 1:

```
MacBook-Pro-di-Gabriele:~ gabrielebisogno$ mongoimport -d progettino -c donorsv --type csv --file /Users/gabrielebisogno/Desktop/progettino/Donorsfinal.csv --headerline
connected to: localhost
2018-09-18T22:46:10.299+0200 [##.....] progettino.donorsv 12.2MB/118MB (10.3%)
2018-09-18T22:46:13.284+0200 [#####.....] progettino.donorsv 25.2MB/118MB (21.3%)
2018-09-18T22:46:16.284+0200 [#####.....] progettino.donorsv 38.5MB/118MB (32.6%)
2018-09-18T22:46:19.286+0200 [#####.....] progettino.donorsv 51.3MB/118MB (43.4%)
2018-09-18T22:46:22.284+0200 [#####.....] progettino.donorsv 64.2MB/118MB (54.3%)
2018-09-18T22:46:25.285+0200 [#####.....] progettino.donorsv 77.0MB/118MB (65.1%)
2018-09-18T22:46:31.288+0200 [#####.....] progettino.donorsv 89.5MB/118MB (75.7%)
2018-09-18T22:46:34.289+0200 [#####.....] progettino.donorsv 98.3MB/118MB (83.1%)
2018-09-18T22:46:37.285+0200 [#####.....] progettino.donorsv 110MB/118MB (93.1%)
2018-09-18T22:46:39.310+0200 [#####.....] progettino.donorsv 118MB/118MB (100.0%)
2018-09-18T22:46:39.318+0200 imported 2122640 documents
```

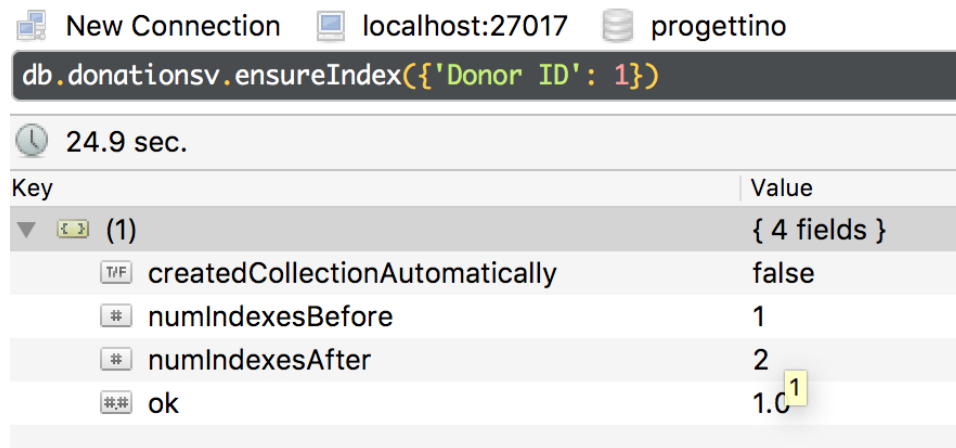
Figure 2:

### 3 Modellizzazione

Ho applicato diverse modifiche/alterazioni ai vari dataset, eliminando campi inutili e aggiungendo nuovi indici.

#### 3.1 Aggiunta indici

Per motivi di performance ho aggiunto alle due collections due indici sull'attributo dove verr eseguita la join/aggregate, per la creazione del documento embed-deb.

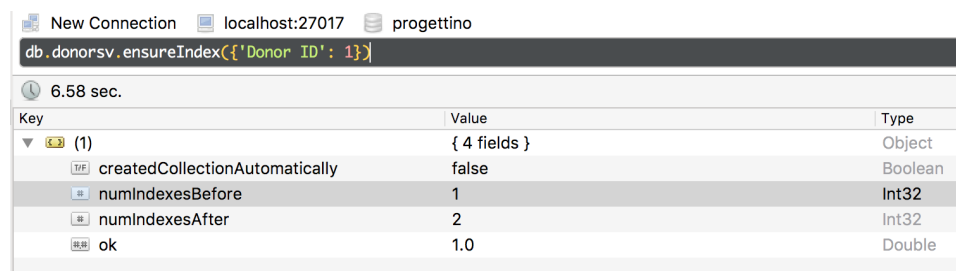


```
db.donationsv.ensureIndex({'Donor ID': 1})
```

24.9 sec.

Key	Value
(1)	{ 4 fields }
createdCollectionAutomatically	false
numIndexesBefore	1
numIndexesAfter	2
ok	1.0

Figure 3:



```
db.donorsv.ensureIndex({'Donor ID': 1})
```

6.58 sec.

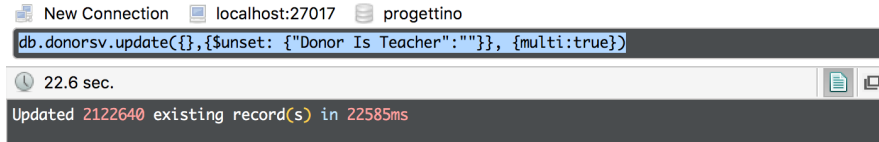
Key	Value	Type
(1)	{ 4 fields }	Object
createdCollectionAutomatically	false	Boolean
numIndexesBefore	1	Int32
numIndexesAfter	2	Int32
ok	1.0	Double

Figure 4:

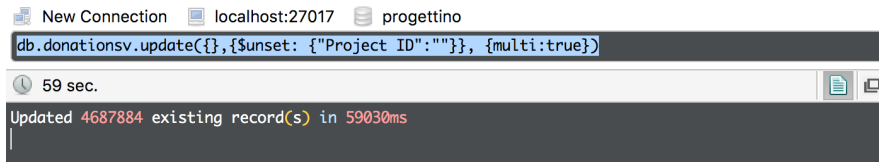
### 3.2 Eliminazione campi futili

Eliminazioni di alcuni campi inutili per favorire una miglior utilizzo dei vari campi e per ridurre grandemente la dimensione del dataset.

- `db.donors.update({},{$unset: {"Donor Is Teacher":""}},{multi:true})`



- `db.donations.update({},{$unset: {"Project ID":""}},{multi:true})`



### 3.3 Creazione documento embeddeb

Andiamo a unire le due collezioni in un documento embeddeb, e utilizziamo la funzione lookup di aggregate per creare un documento embeddeb, andando a eliminare il valore Donor ID duplicato.

```
db.donationsv.aggregate([
  {$lookup: {
    from: "donorsv",
    localField: "Donor ID",
    foreignField: "Donor ID",
    as: "donorsv" }},
  { $match: {"donorsv": { $ne: [] }}},
  {$out: "document_finalv" } ])
```

db.getCollection('document\_finalv').find(

New Connection localhost:27017 progettino

db.getCollection('document\_finalv').find({})

document\_finalv 0.003 sec.

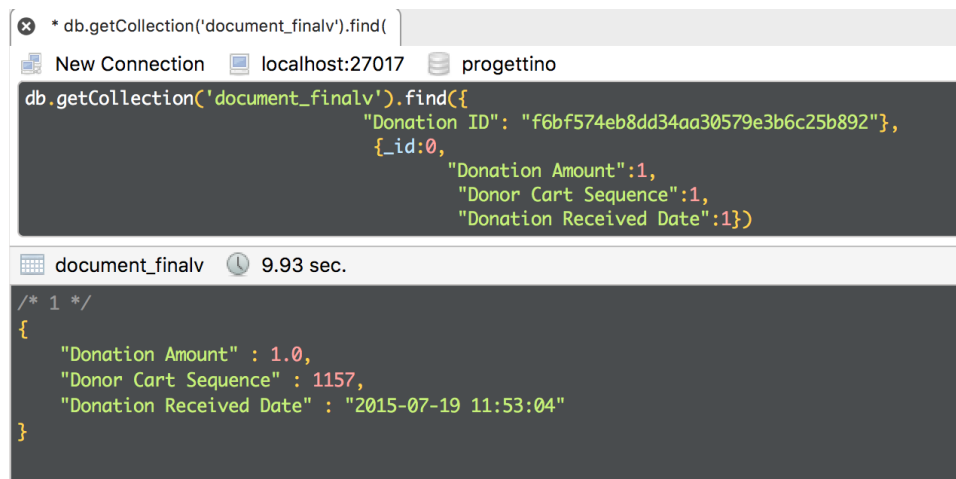
```
{
  "_id" : ObjectId("5ba1646591895f049727df83"),
  "Donation ID" : "18a234b9d1e538c431761d521ea7799d",
  "Donor ID" : "0b0765dc9c759adc48a07688ba25e94e",
  "Donation Included Optional Donation" : "Yes",
  "Donation Amount" : 20.0,
  "Donor Cart Sequence" : 3,
  "Donation Received Date" : "2016-06-06 14:08:46",
  "donorsv" : [
    {
      "_id" : ObjectId("5ba1641391895f049708da5b"),
      "Donor ID" : "0b0765dc9c759adc48a07688ba25e94e",
      "Donor City" : "Riverton",
      "Donor State" : "Utah",
      "Donor Zip" : 840
    }
  ]
}

/* 2 */
{
  "_id" : ObjectId("5ba1646591895f049727df84"),
  "Donation ID" : "688729120858666221208529ee3fc18e",
  "Donor ID" : "1f4b5b6e68445c6c4a0509b3aca93f38",
  "Donation Included Optional Donation" : "No",
  "Donation Amount" : 178.37,
  "Donor Cart Sequence" : 11,
  "Donation Received Date" : "2016-08-23 13:15:57",
  "donorsv" : [
    {
      "_id" : ObjectId("5ba1641591895f04970b6b37"),
      "Donor ID" : "1f4b5b6e68445c6c4a0509b3aca93f38",
      "Donor City" : "West Jordan",
      "Donor State" : "Utah",
      "Donor Zip" : 840
    }
  ]
}
```

## 4 Ottimizzazione Query

### 4.1 Prima Query

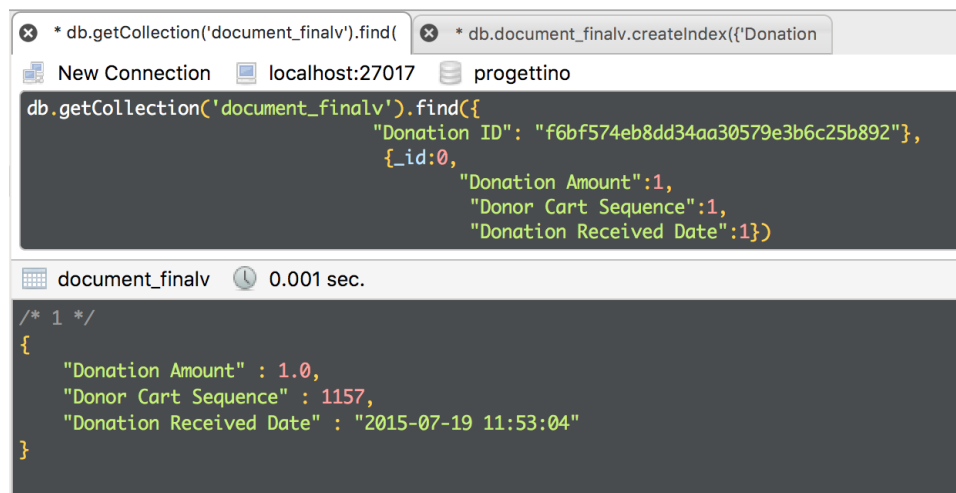
Vogliamo trovare una determinata donazione, quindi con uno specifico Donation ID, e visualizzare l'ammontare della donazione, la data della donazione  
Senza Indice



The screenshot shows the MongoDB Shell interface. The top bar indicates a connection to 'localhost:27017' with the database 'progettino'. The command entered is `db.getCollection('document_finalv').find({"Donation ID": "f6bf574eb8dd34aa30579e3b6c25b892"}, {_id:0, "Donation Amount":1, "Donor Cart Sequence":1, "Donation Received Date":1})`. The execution time is 9.93 seconds. The result is a single document: `{ "Donation Amount" : 1.0, "Donor Cart Sequence" : 1157, "Donation Received Date" : "2015-07-19 11:53:04" }`.

```
* db.getCollection('document_finalv').find(
New Connection localhost:27017 progettino
db.getCollection('document_finalv').find({
  "Donation ID": "f6bf574eb8dd34aa30579e3b6c25b892"},
  {_id:0,
    "Donation Amount":1,
    "Donor Cart Sequence":1,
    "Donation Received Date":1})
document_finalv 9.93 sec.
/* 1 */
{
  "Donation Amount" : 1.0,
  "Donor Cart Sequence" : 1157,
  "Donation Received Date" : "2015-07-19 11:53:04"
}
```

Con indice (Su Donation ID)



The screenshot shows the MongoDB Shell interface with an additional tab for creating an index: `* db.document_finalv.createIndex({'Donation ID':1})`. The command entered is the same as in the previous screenshot. The execution time is significantly faster, at 0.001 seconds. The result is the same document: `{ "Donation Amount" : 1.0, "Donor Cart Sequence" : 1157, "Donation Received Date" : "2015-07-19 11:53:04" }`.

```
* db.getCollection('document_finalv').find(
New Connection localhost:27017 progettino
db.getCollection('document_finalv').find({
  "Donation ID": "f6bf574eb8dd34aa30579e3b6c25b892"},
  {_id:0,
    "Donation Amount":1,
    "Donor Cart Sequence":1,
    "Donation Received Date":1})
document_finalv 0.001 sec.
/* 1 */
{
  "Donation Amount" : 1.0,
  "Donor Cart Sequence" : 1157,
  "Donation Received Date" : "2015-07-19 11:53:04"
}
```

## 4.2 Seconda Query

Vogliamo trovare una determinata donazione, quindi con uno specifico Donation ID, con un Donation Amount maggiore di 15 e Donor Zip superiore a 70.

Senza Indice

New Connection localhost:27017 progettino

```
db.getCollection('document_finalv').find({
  "Donation ID" : "924e80c9340d143cb0adb32637982915",
  "Donation Amount" : {$gt : 15},
  "donorsv.Donor Zip" : {$gt : 70}})
```

document\_finalv 3.31 sec. 0 50

Key	Value	Type
▼ (1) ObjectId("5ba1646791895...)	{ 8 fields }	Object
_id	ObjectId("5ba1646791895f0497292...)	ObjectId
Donation ID	924e80c9340d143cb0adb3263798...	String
Donor ID	00628a2feba15bff07bb982c0a2831...	String
Donation Included Optiona...	Yes	String
Donation Amount	25.0	Double
Donor Cart Sequence	2	Int32
Donation Received Date	2015-12-18 08:03:32	String
▶ donorsv	[ 1 element ]	Array

Con indice (su Donation ID)

New Connection localhost:27017 progettino

```
db.getCollection('document_finalv').find({
  "Donation ID" : "924e80c9340d143cb0adb32637982915",
  "Donation Amount" : {$gt : 15},
  "donorsv.Donor Zip" : {$gt : 70}})
```

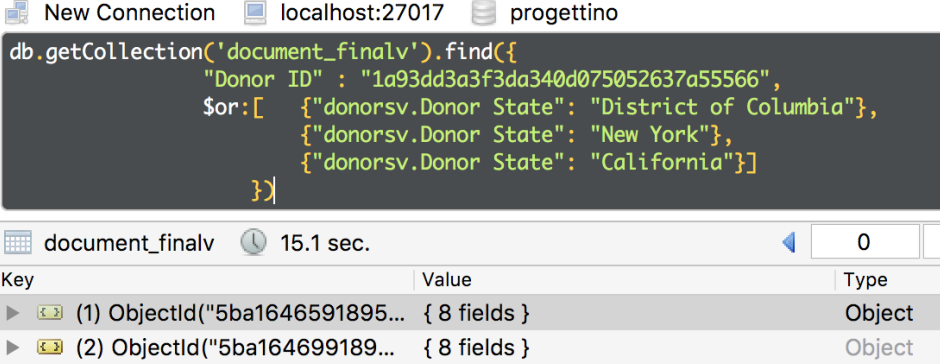
document_finalv 0.006 sec.			0
Key	Value	Type	
▼ (1) ObjectId("5ba1646791895... { 8 fields }		Object	
_id	ObjectId("5ba1646791895f0497292...	ObjectId	
Donation ID	924e80c9340d143cb0adb3263798...	String	
Donor ID	00628a2feba15bff07bb982c0a2831...	String	
Donation Included Optiona...	Yes	String	
Donation Amount	25.0	Double	
Donor Cart Sequence	2	Int32	
Donation Received Date	2015-12-18 08:03:32	String	
donorsv	[ 1 element ]	Array	



### 4.3 Terza Query

Vogliamo trovare tutte le donazioni di un determinato Donor ID, il cui ammontare della donazione sia superiore a 11, il cui donatore faccia parte dello stato di New York o del Distretto della Columbia o della California

Senza Indice



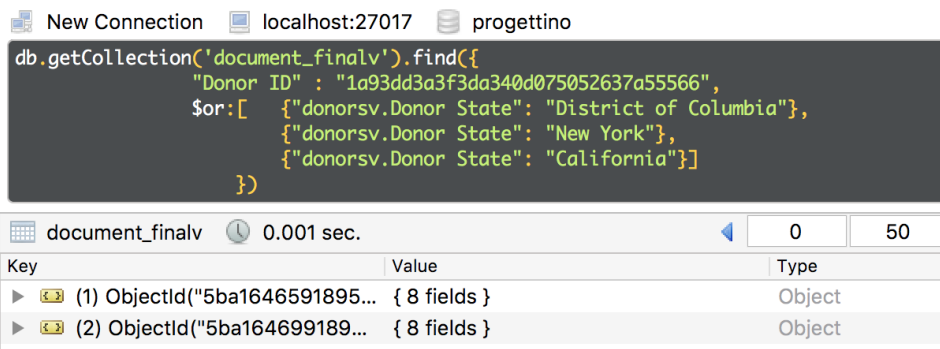
New Connection localhost:27017 progettino

```
db.getCollection('document_finalv').find({
  "Donor ID" : "1a93dd3a3f3da340d075052637a55566",
  $or:[ {"donorsv.Donor State": "District of Columbia"},
        {"donorsv.Donor State": "New York"},
        {"donorsv.Donor State": "California"}]
})
```

document\_finalv 15.1 sec. 0

Key	Value	Type
(1) ObjectId("5ba1646591895...)	{ 8 fields }	Object
(2) ObjectId("5ba164699189...)	{ 8 fields }	Object

Con indice (su Donor ID)



New Connection localhost:27017 progettino

```
db.getCollection('document_finalv').find({
  "Donor ID" : "1a93dd3a3f3da340d075052637a55566",
  $or:[ {"donorsv.Donor State": "District of Columbia"},
        {"donorsv.Donor State": "New York"},
        {"donorsv.Donor State": "California"}]
})
```

document\_finalv 0.001 sec. 0 50

Key	Value	Type
(1) ObjectId("5ba1646591895...)	{ 8 fields }	Object
(2) ObjectId("5ba164699189...)	{ 8 fields }	Object

## 4.4 Quarta Query

Dato un ID di un donatore, (con Donor ID iniziale con 0, contenente 344 e con b alla fine) trovare tutte le donazioni da lui fatte, restituendo l'ID della donazione e il suo ammontare

Senza indice

New Connection localhost:27017 progetto

```
stCollection("document_finalv").find(
  { $and: [ {"donorsv.Donor ID" : /0/}, {"donorsv.Donor ID" : /344/}, {"donorsv.Donor ID" : /b$/} ] },
  { _id:0, "Donation ID":1, "Donation Amount": 1, "donorsv.Donor ID": 1 }
)
```

document\_finalv 6.48 sec.

Key	Value	Type
Donation ID	324b78728d841f7859ffb558092ba0de	String
Donation Amount	1.5	Double
donorsv	[ 1 element ]	Array
(3)	{ 3 fields }	Object
Donation ID	4a129e0bff4f4fb8ffdea5eadcde74f9	String
Donation Amount	25.0	Double
donorsv	[ 1 element ]	Array
[0]	{ 1 field }	Object
Donor ID	0f3c40b8404523e2fdc801bb5b73447b	String
(4)	{ 3 fields }	Object
Donation ID	26f6f84c9c1277b3d6b764b9154017cc	String
Donation Amount	111.0	Double
donorsv	[ 1 element ]	Array
(5)	{ 3 fields }	Object
Donation ID	0d5123130ceb4c15e2eec1abca022a3f	String
Donation Amount	25.0	Double
donorsv	[ 1 element ]	Array
[0]	{ 1 field }	Object
Donor ID	0b79e29344395db2bdd5fffaa8f638b	String
(6)	{ 3 fields }	Object
Donation ID	42c6f1e12410f43ae14338df495cba6	String
Donation Amount	25.0	Double
donorsv	[ 1 element ]	Array

Con indice (su Donor ID)

New Connection localhost:27017 progetto

```
stCollection("document_finalv").find(
  { $and: [ {"donorsv.Donor ID" : /0/}, {"donorsv.Donor ID" : /344/}, {"donorsv.Donor ID" : /b$/} ] },
  { _id:0, "Donation ID":1, "Donation Amount": 1, "donorsv.Donor ID": 1 }
)
```

document\_finalv 3.63 sec.

Key	Value	Type
(1)	{ 3 fields }	Object
(2)	{ 3 fields }	Object
(3)	{ 3 fields }	Object

## 4.5 Quinta Query

Trovare tutti quei Donatori dello stato del Kentucky, (aventi Donor ID che iniziano per 9 e finiscono per cac) Donazione corrispondente avente Donation ID che inizia con 3, e senza donazioni opzionali; con donazione inferiore a 11 o maggiore di 100.

Senza indice

```
db.getCollection("document_finalv").find(
  { $and: [
    {"donorsv.Donor ID" : /^9/},
    {"donorsv.Donor ID" : /cac$/},
    {"Donation ID" : /3$/},
    { $or: [
      {"Donation Amount" : {$lt : 11 }},
      {"Donation Amount" : {$gt : 100 }]}],
    {"Donation Included Optional Donation" : "No"},
    {"donorsv.Donor State" : "Kentucky"}] }
)
```

document\_finalv 4.12 sec.

Key	Value	Type
(1) ObjectId("5ba164a191895f04975ca8a2")	{ 8 fields }	Object

Con indice (su Donor ID)

New Connection

localhost:27017

progetto

```

db.getCollection('document_finalv').find(
  { $and: [
    { "donorsv.Donor ID" :"/49/},
    { "donorsv.Donor ID" :"/cac5/},
    { "Donation ID" :"/35/},
    { $or: [
      { "Donation Amount" : { $lt : 11 }},
      { "Donation Amount" : { $gt : 100 } } ]},
    { "Donation Included Optional Donation" : "No"},
    { "donorsv.Donor State" : "Kentucky" ] }
)

```

document\_finalv

2.8 sec.

Key	Value	Type
(1) ObjectId("5ba164a191895f04975ca8a2")	{ 8 fields }	Object
_id	ObjectId("5ba164a191895f04975ca8a2")	ObjectId
Donation ID	4324d0a547e421e45cab8bfd93c63fb3	String
Donor ID	964ee1dd7408bc3263c04b826bd2cac	String
Donation Included Optional Donation	No	String
Donation Amount	5.0	Double
Donor Cart Sequence	4	Int32
Donation Received Date	2015-01-08 08:23:29	String
donorsv	[ 1 element ]	Array

Con indice (su Donation ID)

document_finalv	7.26 sec.
Key	Value
(1) ObjectId("5ba164a191895f04975ca8a2")	{ 8 fields }

## 4.6 Sesta Query

Trovare tutti quei Donatori che hanno donato piu di una volta, restituendo per ognuno media delle Donazioni e il loro numero.

Senza indice

New Connection localhost:27017 progetto

```
b.getCollection('document_finalv').aggregate([
  {$group: {$_id: "$donorsv.Donor ID",
    "mediaDonazione" : {$avg: "$Donation Amount"},
    "numeroDonazioni" : {$sum: 1}}},
  {$match: {"numeroDonazioni": {$gt: 2}}},
  {$sort: {"mediaDonazione" : -1}},
  {
    allowDiskUse:true,
    cursor:{}
  }
])
```

document\_finalv 40.4 sec.

Key	Value	Type
▼ (1) [1 element]	{ 3 fields }	Object
▶ (1) _id	[1 element]	Array
mediaDonazione	9956.153333333333	Double
numeroDonazioni	3.0	Double
▼ (2) [1 element]	{ 3 fields }	Object
▶ (2) _id	[1 element]	Array
mediaDonazione	8976.286666666667	Double
numeroDonazioni	3.0	Double
▼ (3) [1 element]	{ 3 fields }	Object
▶ (3) _id	[1 element]	Array
mediaDonazione	7838.64	Double
numeroDonazioni	5.0	Double
▶ (4) [1 element]	{ 3 fields }	Object
▶ (5) [1 element]	{ 3 fields }	Object
▶ (6) [1 element]	{ 3 fields }	Object
▶ (7) [1 element]	{ 3 fields }	Object
▶ (8) [1 element]	{ 3 fields }	Object
▶ (9) [1 element]	{ 3 fields }	Object
▶ (10) [1 element]	{ 3 fields }	Object
▶ (11) [1 element]	{ 3 fields }	Object
▶ (12) [1 element]	{ 3 fields }	Object
▶ (13) [1 element]	{ 3 fields }	Object
▶ (14) [1 element]	{ 3 fields }	Object

Con indice (su Donor ID)

New Connection localhost:27017 progetto

```
b.getCollection('document_finalv').aggregate([
  {$group: {$_id: "$donorsv.Donor ID",
    "mediaDonazione" : {$avg: "$Donation Amount"},
    "numeroDonazioni" : {$sum: 1}}},
  {$match: {"numeroDonazioni": {$gt: 2}}},
  {$sort: {"mediaDonazione" : -1}},
  {
    allowDiskUse:true,
    cursor:{}
  }
])
```

document\_finalv 45.9 sec.

Key	Value
▶ (1) [1 element]	{ 3 fields }
▶ (2) [1 element]	{ 3 fields }

## 5 Conclusione

Come dimostrato, per quasitutte le query abbiamo un netto miglioramento;

Ci sono alcune istruzioni, che utilizzano determinati metodi di ricerca, indicati nella documentazione (`$match` e `$sort`), che non traggono beneficio dall'utilizzo dell'indicizzazione, come nella sesta query. Ci sono anche alcuni casi, per esempio la quinta query, dove l'utilizzo errato di un parametro indicizzato ha peggiorato i tempi di ricerca. Su tabelle/collezioni di dataset di dimensioni pi generose, avrebbe comportato un aumento di tempo in molte operazioni di ricerca. Quindi, in conclusione, indicizzare un dataset pu essere molto utile, ma si deve stare attenti a particolari casi di utilizzo dove l'uso non va a migliorare i tempi di risposta.