CSEE5590 Big Data Programming

In Class Programming –6 Report
(Jongkook Son)

Project Overview:

Solr is highly reliable, scalable and fault tolerant, providing distributed indexing,
replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest internet sites.

Requirements/Task(s):

1. Film Dataset

- Keyword matching

- Wildcard matching

- Proximity matching

- Range searches

- Fuzzy logic

2. Books Dataset

ICP question: Execute any 5 queries on the given dataset

What I learned in ICP:

I could have learned about Solr and its functionality to index the data. In the ICP I learned how to create a collection in solr and how to edit the schema to match the data fields of a given data set. And how to import given dataset to solr. Also I could understand the various queries in Solr it was easy to grasp the concept. Configuring was a little tricky first, But I could figure out thanks to the reference.

ICP description what was the task you were performing and Screen shots that shows the successful execution of each required step of your code

<Task1> Films

1- Generating a new instance using the below command:

```
$ solrctl instancedir --generate /tmp/film
```

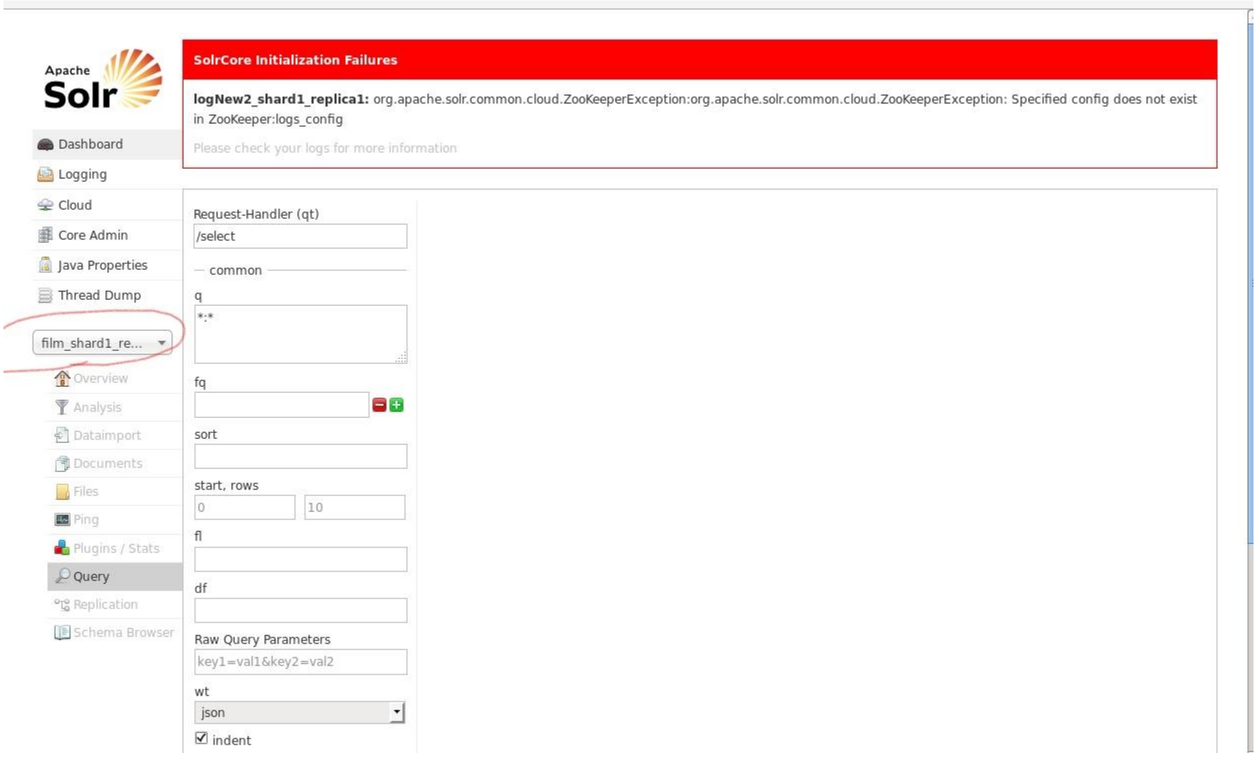
2- To open the schema file, I used ***gedit***, which comes with the system.

```
$ gedit /tmp/film/conf/schema.xml
```

3- Then I added some fields to it and saved it.

4- Finally, created a new instance from this schema and created my collection from it.

```
$ solrctl instancedir --create film /tmp/films
$ solrctl collection --create films
```



SolrCore Initialization Failures

logNew2_shard1_replica1: org.apache.solr.common.cloud.ZooKeeperException:org.apache.solr.common.cloud.ZooKeeperException: Specified config does not exist in ZooKeeper:logs_config

Please check your logs for more information

Request-Handler (qt): /select

common

q: *:*

fq:

sort:

start, rows: 0 10

fl:

df:

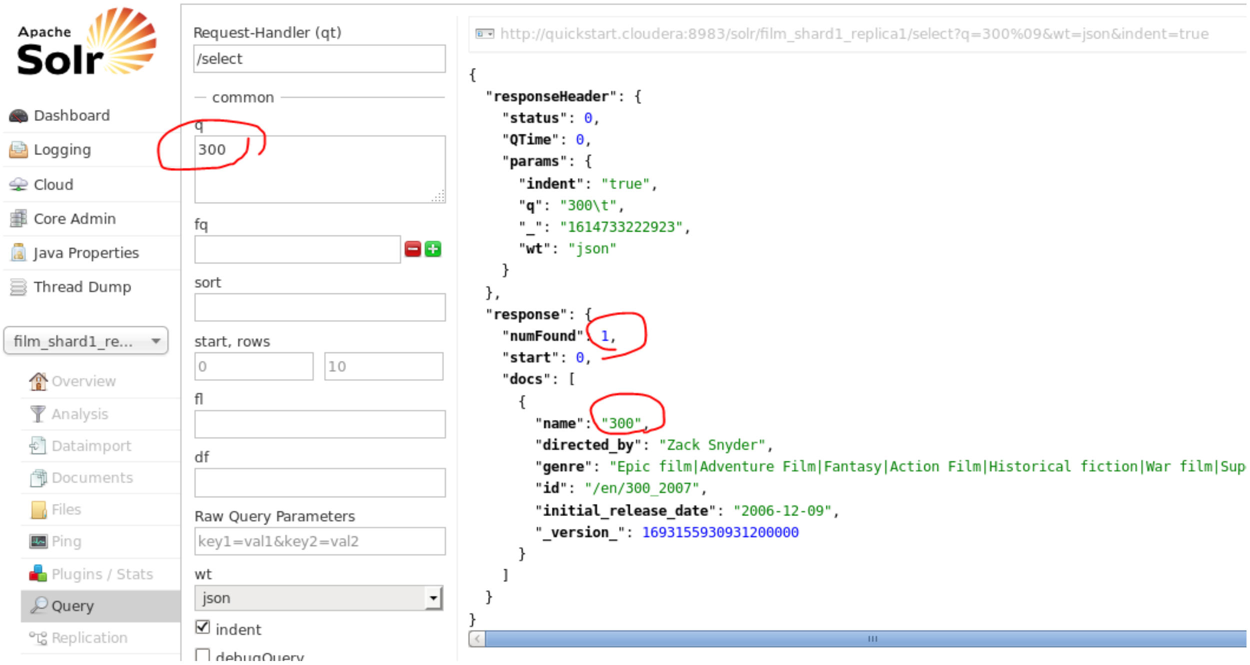
Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

=>One can find out that the collection is well created and displayed in the solr admin page

1)Keyword Matching(300)



Request-Handler (qt): /select

common

q: 300

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

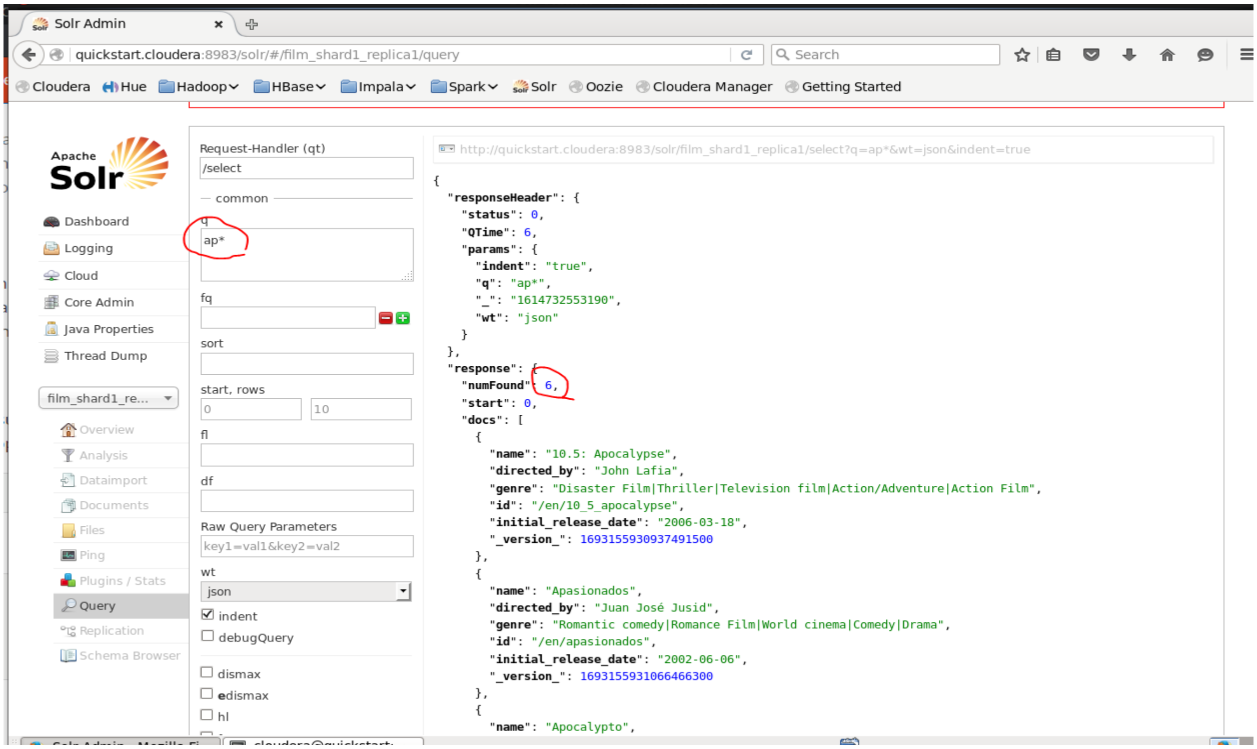
□ debugQuery

http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=300%09&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "indent": "true",
      "q": "300\t",
      "_": "1614733222923",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1,
    "start": 0,
    "docs": [
      {
        "name": "300",
        "directed_by": "Zack Snyder",
        "genre": "Epic film|Adventure Film|Fantasy|Action Film|Historical fiction|War film|Sup
        "id": "/en/300_2007",
        "initial_release_date": "2006-12-09",
        "_version_": 1693155930931200000
      }
    ]
  }
}
```

=>Just display the data which field’s value is 300

2)Wildcard Matching(ap*)



Request-Handler (qt): /select

common

q: ap*

fq

sort

start, rows: 0 10

fl

df

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

□ debugQuery

□ dismax

□ edismax

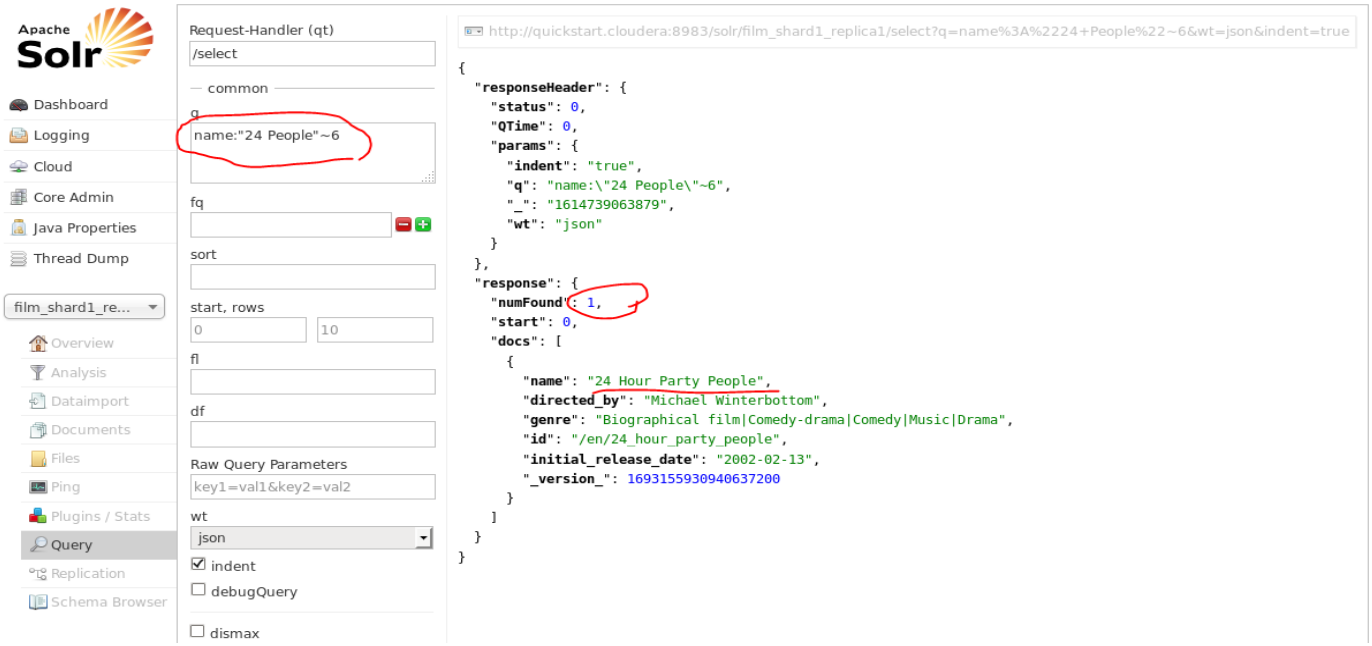
□ hl

http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=ap*&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 6,
    "params": {
      "indent": "true",
      "q": "ap*",
      "_": "1614732553190",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 6,
    "start": 0,
    "docs": [
      {
        "name": "10.5: Apocalypse",
        "directed_by": "John Lafia",
        "genre": "Disaster Film|Thriller|Television film|Action/Adventure|Action Film",
        "id": "/en/10_5_apocalypse",
        "initial_release_date": "2006-03-18",
        "_version_": 1693155930937491500
      },
      {
        "name": "Apasionados",
        "directed_by": "Juan José Jusid",
        "genre": "Romantic comedy|Romance Film|World cinema|Comedy|Drama",
        "id": "/en/apasionados",
        "initial_release_date": "2002-06-06",
        "_version_": 1693155931066466300
      },
      {
        "name": "Apocalypto",
```

=> Solr’s standard query parser supports single and multiple character wildcard searches within single terms. * indicates Multiple characters (matches zero or more sequential characters)

3)Proximity Search(name:”24 People”~6)



Request-Handler (qt): /select

common

q: name:"24 People"~6

fq

sort

start, rows: 0 10

fl

df

Raw Query Parameters: key1=val1&key2=val2

wt: json

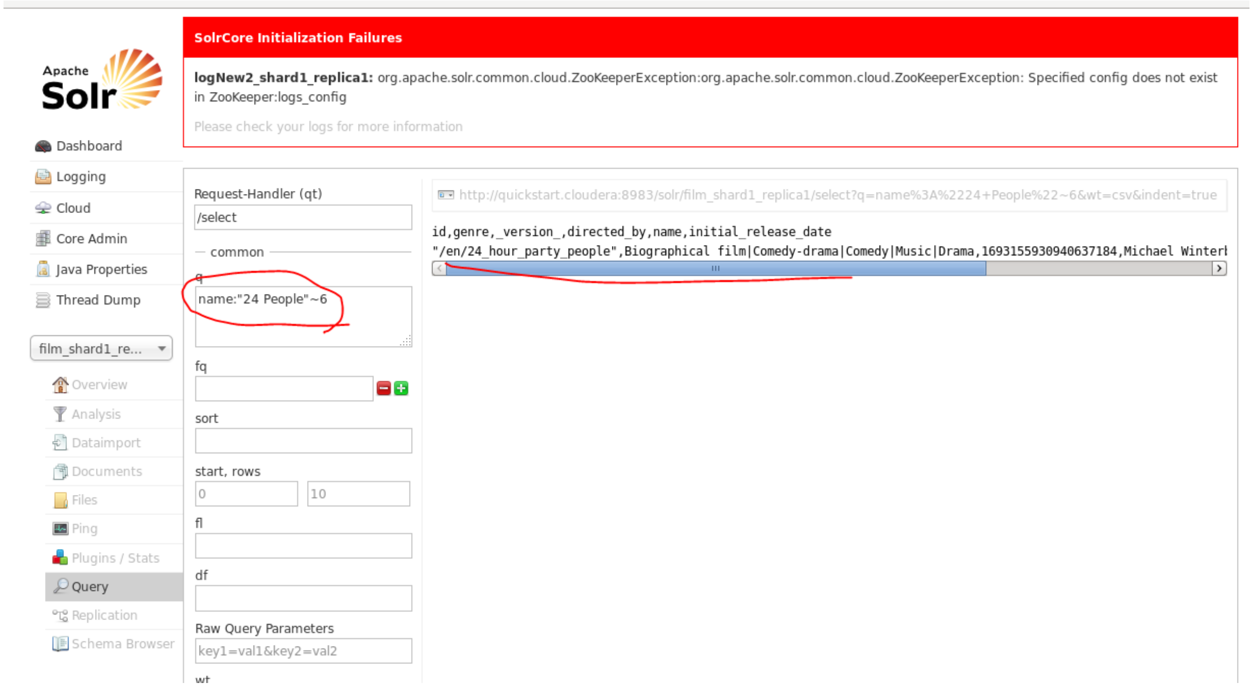
☑ indent

☐ debugQuery

☐ dismax

http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=name%3A%2224+People%22~6&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "indent": "true",
      "q": "name:\"24 People\"~6",
      "_": "1614739063879",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1,
    "start": 0,
    "docs": [
      {
        "name": "24 Hour Party People",
        "directed_by": "Michael Winterbottom",
        "genre": "Biographical film|Comedy-drama|Comedy|Music|Drama",
        "id": "/en/24_hour_party_people",
        "initial_release_date": "2002-02-13",
        "_version_": 1693155930940637200
      }
    ]
  }
}
```



SolrCore Initialization Failures

logNew2_shard1_replica1: org.apache.solr.common.cloud.ZooKeeperException:org.apache.solr.common.cloud.ZooKeeperException: Specified config does not exist in ZooKeeper:logs_config

Please check your logs for more information

Request-Handler (qt): /select

common

q: name:"24 People"~6

fq

sort

start, rows: 0 10

fl

df

Raw Query Parameters: key1=val1&key2=val2

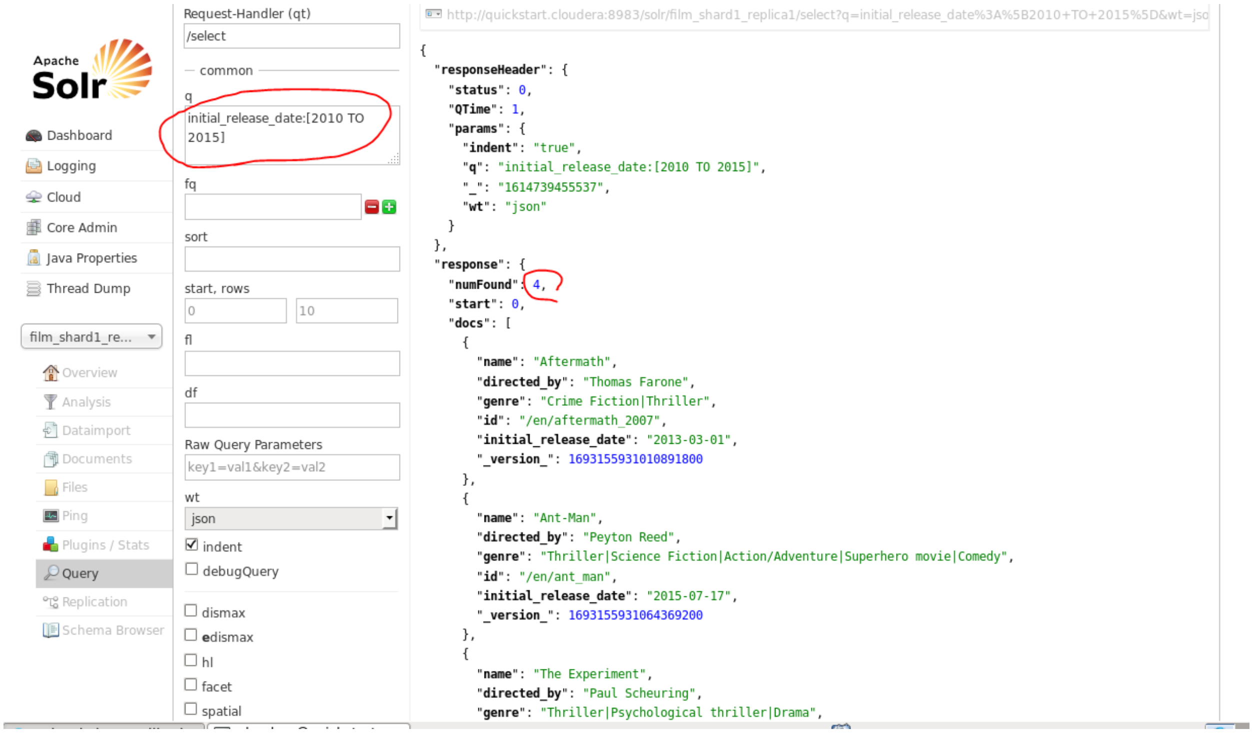
wt

http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=name%3A%2224+People%22~6&wt=csv&indent=true

```
id,genre,_version_,directed_by,name,initial_release_date
"/en/24_hour_party_people",Biographical film|Comedy-drama|Comedy|Music|Drama,1693155930940637184,Michael Winterl
```

=>A proximity search looks for terms that are within a specific distance from one another. In my query, We could find out some matched result which in name field there is 6 word distance from 24 and People. And the name field that matched with this term is “24 hour party people”

4)Range Search(initial_release_date:[2010 TO 2015])



Request-Handler (qt)
/select

common

q
initial_release_date:[2010 TO 2015]

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

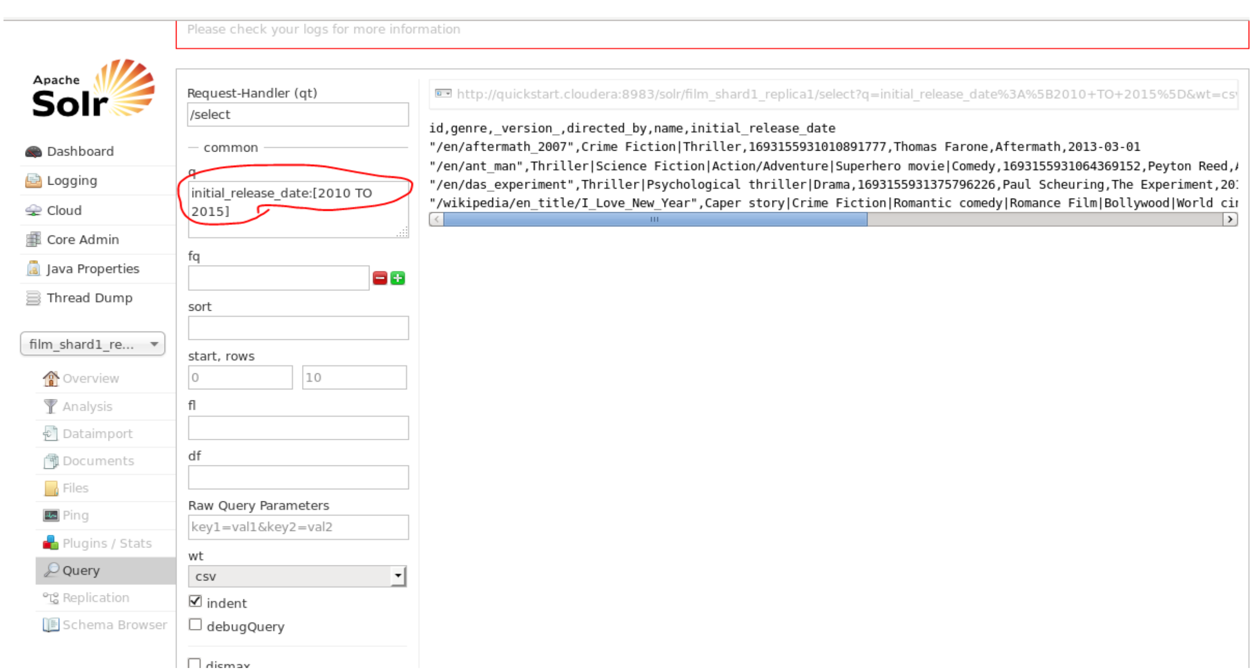
wt
json

☑ indent
□ debugQuery

□ dismax
□ edismax
□ hl
□ facet
□ spatial

http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=initial_release_date%3A%5B2010+TO+2015%5D&wt=jso

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "initial_release_date:[2010 TO 2015]",
      "_": "1614739455537",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 4,
    "start": 0,
    "docs": [
      {
        "name": "Aftermath",
        "directed_by": "Thomas Farone",
        "genre": "Crime Fiction|Thriller",
        "id": "/en/aftermath_2007",
        "initial_release_date": "2013-03-01",
        "_version_": 1693155931010891800
      },
      {
        "name": "Ant-Man",
        "directed_by": "Peyton Reed",
        "genre": "Thriller|Science Fiction|Action/Adventure|Superhero movie|Comedy",
        "id": "/en/ant_man",
        "initial_release_date": "2015-07-17",
        "_version_": 1693155931064369200
      },
      {
        "name": "The Experiment",
        "directed_by": "Paul Scheuring",
        "genre": "Thriller|Psychological thriller|Drama",
```


Please check your logs for more information

Request-Handler (qt)
/select

common

q
initial_release_date:[2010 TO 2015]

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
csv

☑ indent
□ debugQuery

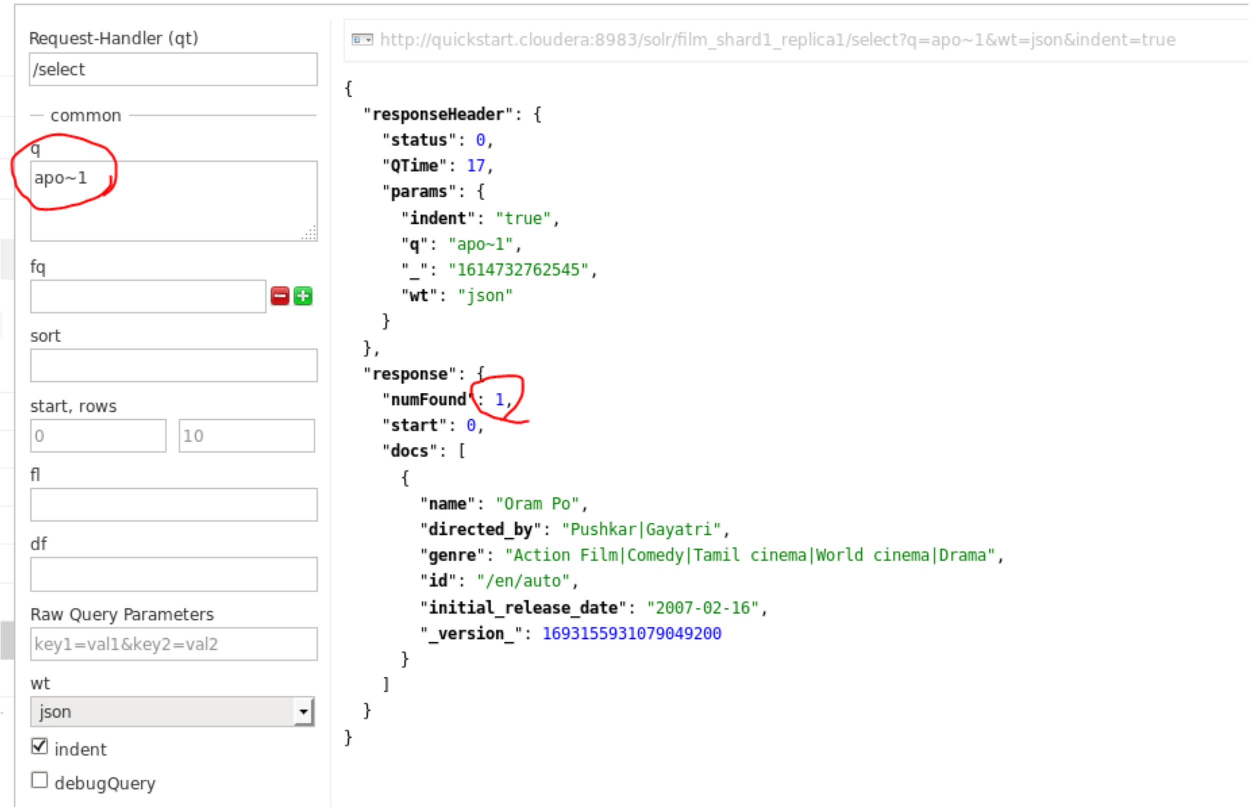
□ dismax

http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=initial_release_date%3A%5B2010+TO+2015%5D&wt=csv

```
id,genre,_version_,directed_by,name,initial_release_date
"/en/aftermath_2007",Crime Fiction|Thriller,1693155931010891777,Thomas Farone,Aftermath,2013-03-01
"/en/ant_man",Thriller|Science Fiction|Action/Adventure|Superhero movie|Comedy,1693155931064369152,Peyton Reed,A
"/en/das_experiment",Thriller|Psychological thriller|Drama,1693155931375796226,Paul Scheuring,The Experiment,20
"/wikipedia/en_title/I_Love_New_Year",Caper story|Crime Fiction|Romantic comedy|Romance Film|Bollywood|World ci
```

=>A range search specifies a range of values for a field (a range with an upper bound and a lower bound). The query matches documents whose values for the specified field or fields fall within the range. In my query, I could extract some results which initial release date's field value is between 2010 to 2015.

5)Fuzzy Search(apo~1)



Request-Handler (qt)

/select

common

q

apo~1

fq

sort

start, rows

0 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☑ indent

☐ debugQuery

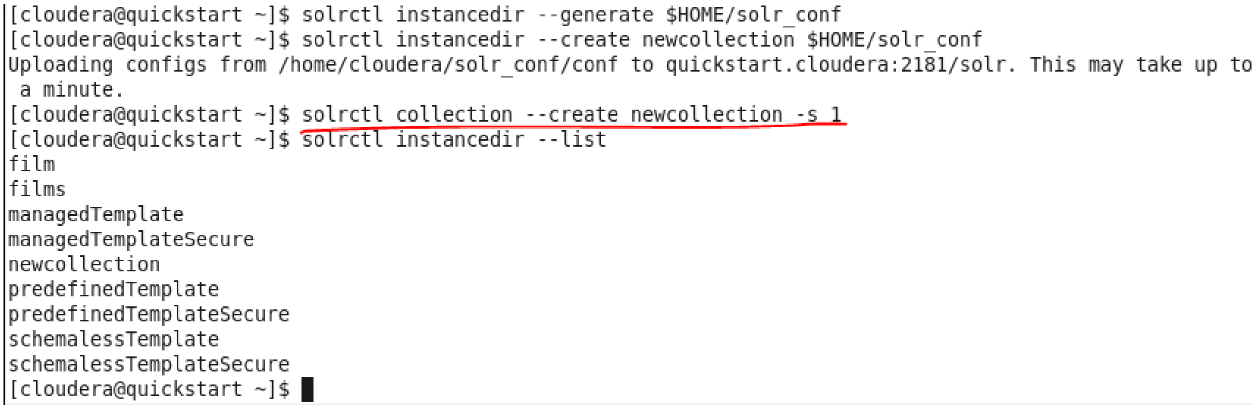
http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=apo~1&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 17,
    "params": {
      "indent": "true",
      "q": "apo~1",
      "_": "1614732762545",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1,
    "start": 0,
    "docs": [
      {
        "name": "Oram Po",
        "directed_by": "Pushkar|Gayatri",
        "genre": "Action Film|Comedy|Tamil cinema|World cinema|Drama",
        "id": "/en/auto",
        "initial_release_date": "2007-02-16",
        "_version_": 1693155931079049200
      }
    ]
  }
}
```

=> Fuzzy searches discover terms that are similar to a specified term without necessarily being an exact match. In my query, some words like “apol” “bpo” will be matched because edit distance is 1 but word like “bpol” will not be matched because edit distance is 2.

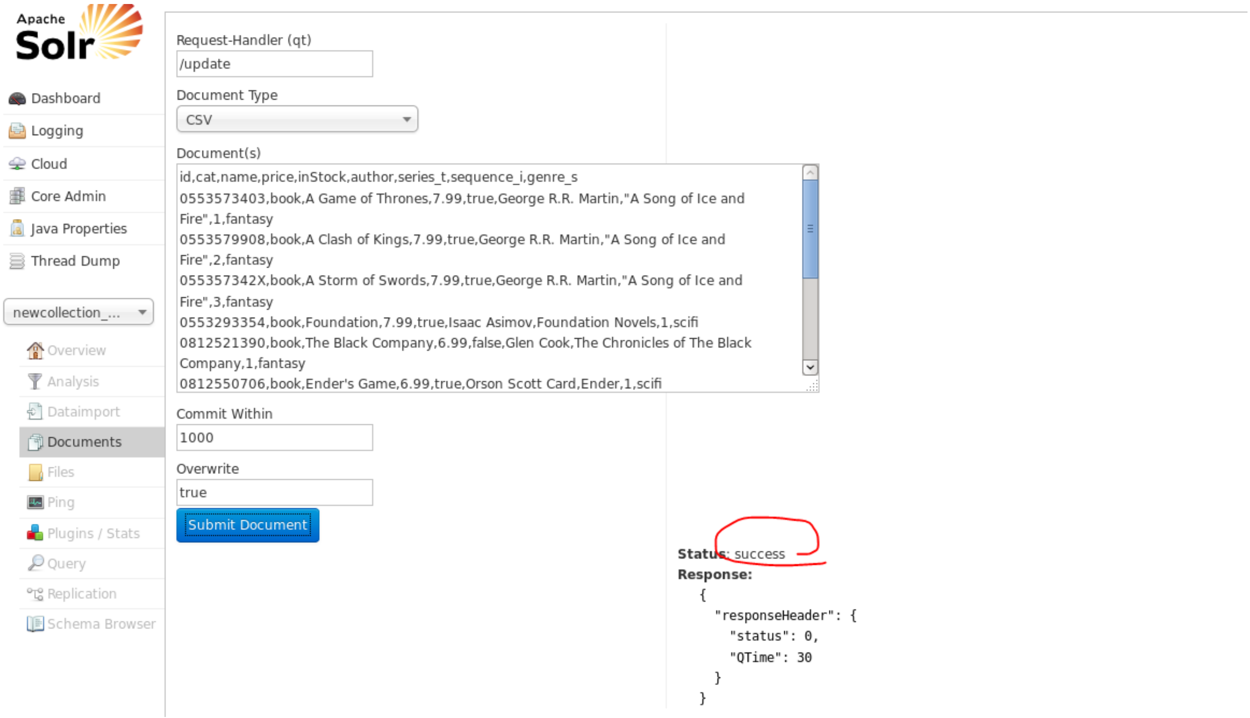
Task2 Books

Creating a collection for the second task.



```
[cloudera@quickstart ~]$ solrctl instancedir --generate $HOME/solr_conf
[cloudera@quickstart ~]$ solrctl instancedir --create newcollection $HOME/solr_conf
Uploading configs from /home/cloudera/solr_conf/conf to quickstart.cloudera:2181/solr. This may take up to
 a minute.
[cloudera@quickstart ~]$ solrctl collection --create newcollection -s 1
[cloudera@quickstart ~]$ solrctl instancedir --list
film
films
managedTemplate
managedTemplateSecure
newcollection
predefinedTemplate
predefinedTemplateSecure
schemalessTemplate
schemalessTemplateSecure
[cloudera@quickstart ~]$ 
```

Upload a document



Request-Handler (qt): /update

Document Type: CSV

Document(s):

```
id,cat,name,price,inStock,author,series_t,sequence_i,genre_s
0553573403,book,A Game of Thrones,7.99,true,George R.R. Martin,"A Song of Ice and
Fire",1,fantasy
0553579908,book,A Clash of Kings,7.99,true,George R.R. Martin,"A Song of Ice and
Fire",2,fantasy
055357342X,book,A Storm of Swords,7.99,true,George R.R. Martin,"A Song of Ice and
Fire",3,fantasy
0553293354,book,Foundation,7.99,true,Isaac Asimov,Foundation Novels,1,scifi
0812521390,book,The Black Company,6.99,false,Glen Cook,The Chronicles of The Black
Company,1,fantasy
0812550706,book,Ender's Game,6.99,true,Orson Scott Card,Ender,1,scifi
```

Commit Within: 1000

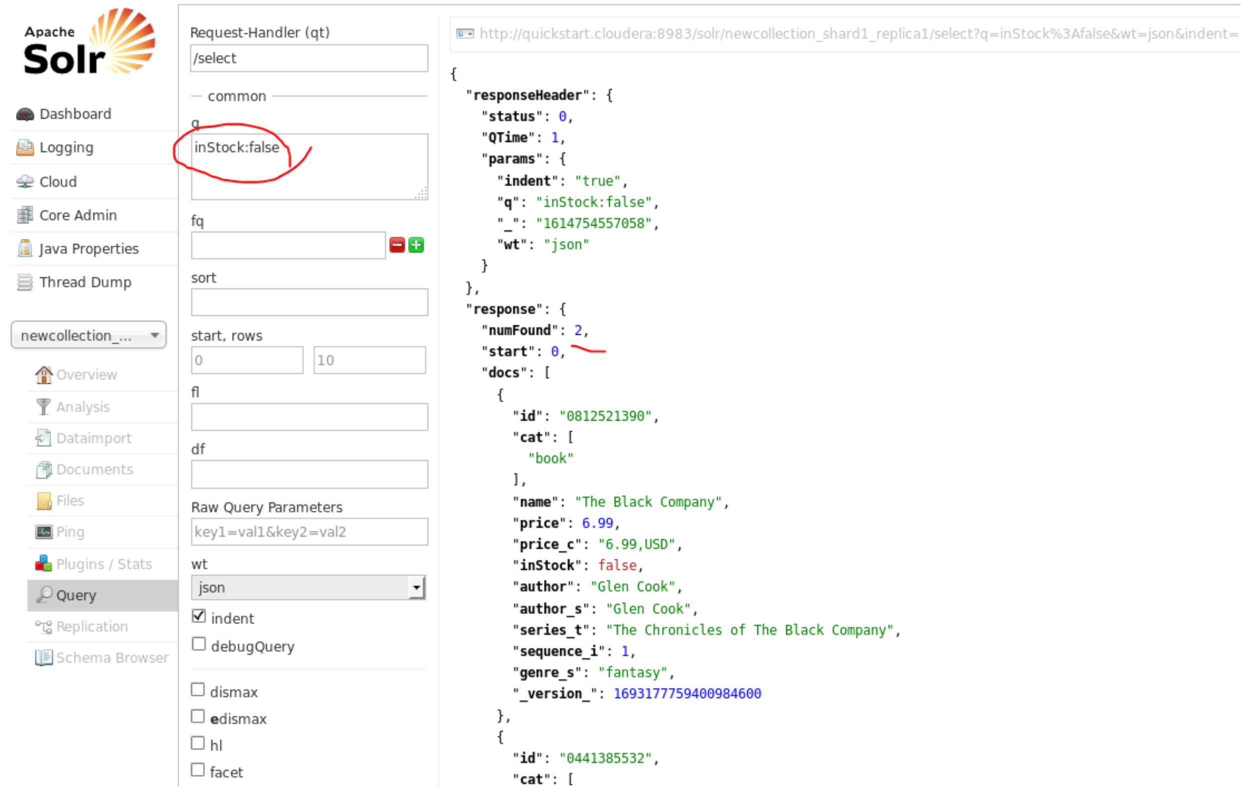
Overwrite: true

Status: success

Response:

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 30
  }
}
```

1)Query1(inStock:false)



Request-Handler (qt): /select

common

q: inStock:false

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

□ debugQuery

□ dismax

□ edismax

□ hl

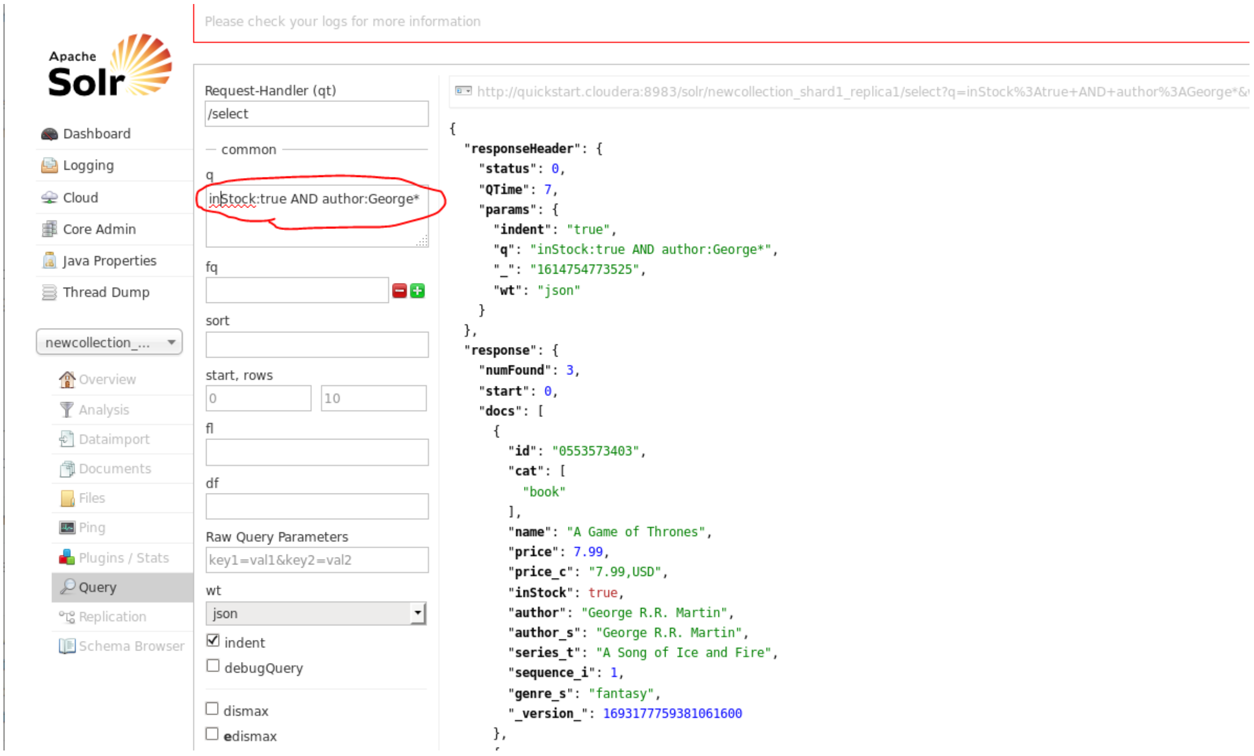
□ facet

http://quickstart.cloudera:8983/solr/newcollection_shard1_replica1/select?q=inStock%3Afalse&wt=json&indent=

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "inStock:false",
      "_": "1614754557058",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 2,
    "start": 0,
    "docs": [
      {
        "id": "0812521390",
        "cat": [
          "book"
        ],
        "name": "The Black Company",
        "price": 6.99,
        "price_c": "6.99,USD",
        "inStock": false,
        "author": "Glen Cook",
        "author_s": "Glen Cook",
        "series_t": "The Chronicles of The Black Company",
        "sequence_i": 1,
        "genre_s": "fantasy",
        "_version_": 1693177759400984600
      },
      {
        "id": "0441385532",
        "cat": [
```

=>Display the data based on the condition that inStock field value is false

2)Query2(inStock:true AND author:George*)



Request-Handler (qt): /select

common

q: inStock:true AND author:George*

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

☐ debugQuery

☐ dismax

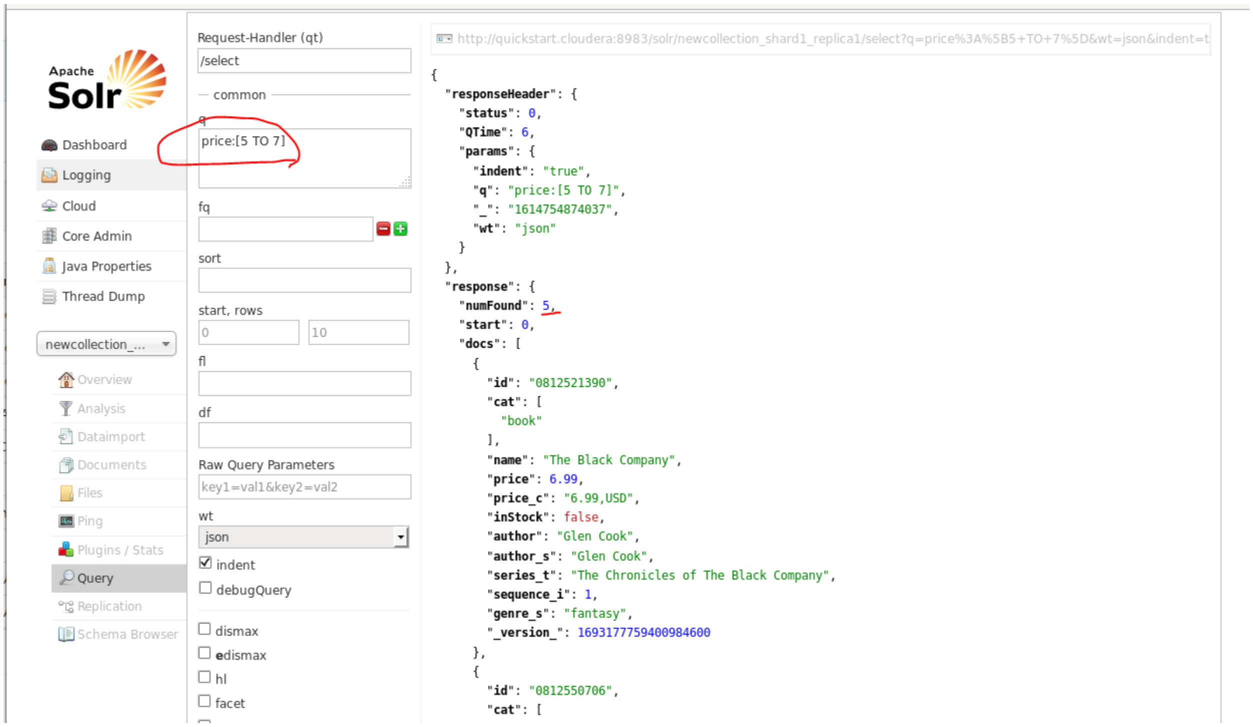
☐ edismax

http://quickstart.cloudera:8983/solr/newcollection_shard1_replica1/select?q=inStock%3Atrue+AND+author%3AGeorge*&

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 7,
    "params": {
      "indent": "true",
      "q": "inStock:true AND author:George*",
      "_": "1614754773525",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 3,
    "start": 0,
    "docs": [
      {
        "id": "0553573403",
        "cat": [
          "book"
        ],
        "name": "A Game of Thrones",
        "price": 7.99,
        "price_c": "7.99,USD",
        "inStock": true,
        "author": "George R.R. Martin",
        "author_s": "George R.R. Martin",
        "series_t": "A Song of Ice and Fire",
        "sequence_i": 1,
        "genre_s": "fantasy",
        "_version_": 1693177759381061600
      },
```

=>Display the data based on the condition that inStock field value is true and author field value is matched with including "George". Wildcard search that Multiple characters (matches zero or more sequential characters) is used.

3) Query3(price:[5 TO 7])



Request-Handler (qt): /select

common

q: price:[5 TO 7]

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent
□ debugQuery

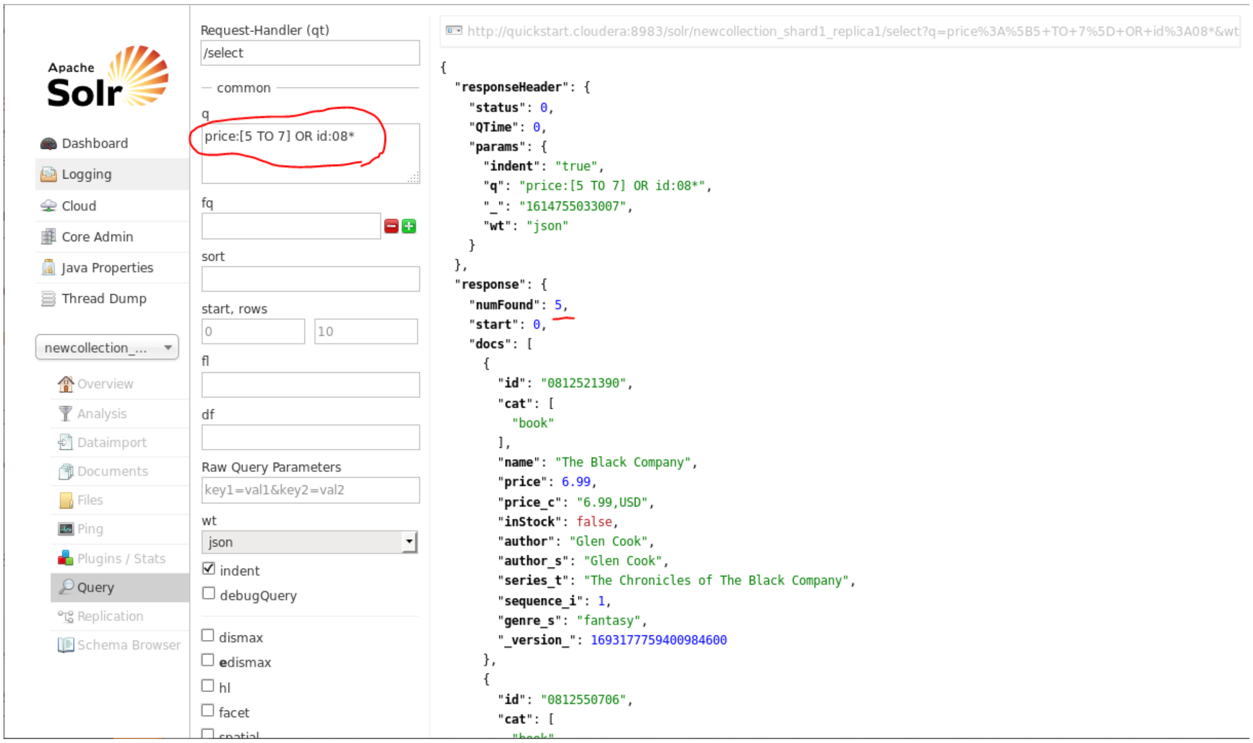
□ dismax
□ edismax
□ hl
□ facet

http://quickstart.cloudera:8983/solr/newcollection_shard1_replica1/select?q=price%3A%5B5+TO+7%5D&wt=json&indent=t

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 6,
    "params": {
      "indent": "true",
      "q": "price:[5 TO 7]",
      "_": "1614754874037",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 5,
    "start": 0,
    "docs": [
      {
        "id": "0812521390",
        "cat": [
          "book"
        ],
        "name": "The Black Company",
        "price": 6.99,
        "price_c": "6.99,USD",
        "inStock": false,
        "author": "Glen Cook",
        "author_s": "Glen Cook",
        "series_t": "The Chronicles of The Black Company",
        "sequence_i": 1,
        "genre_s": "fantasy",
        "_version_": 1693177759400984600
      },
      {
        "id": "0812550706",
        "cat": [
```

=> Range Search is used in Query. Only display the data that price field value is between 5 and 7

4)Query4(price[5 TO 7] OR id:08*)



Request-Handler (qt): /select

common

q: price:[5 TO 7] OR id:08*

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

□ debugQuery

□ dismax

□ edismax

□ hl

□ facet

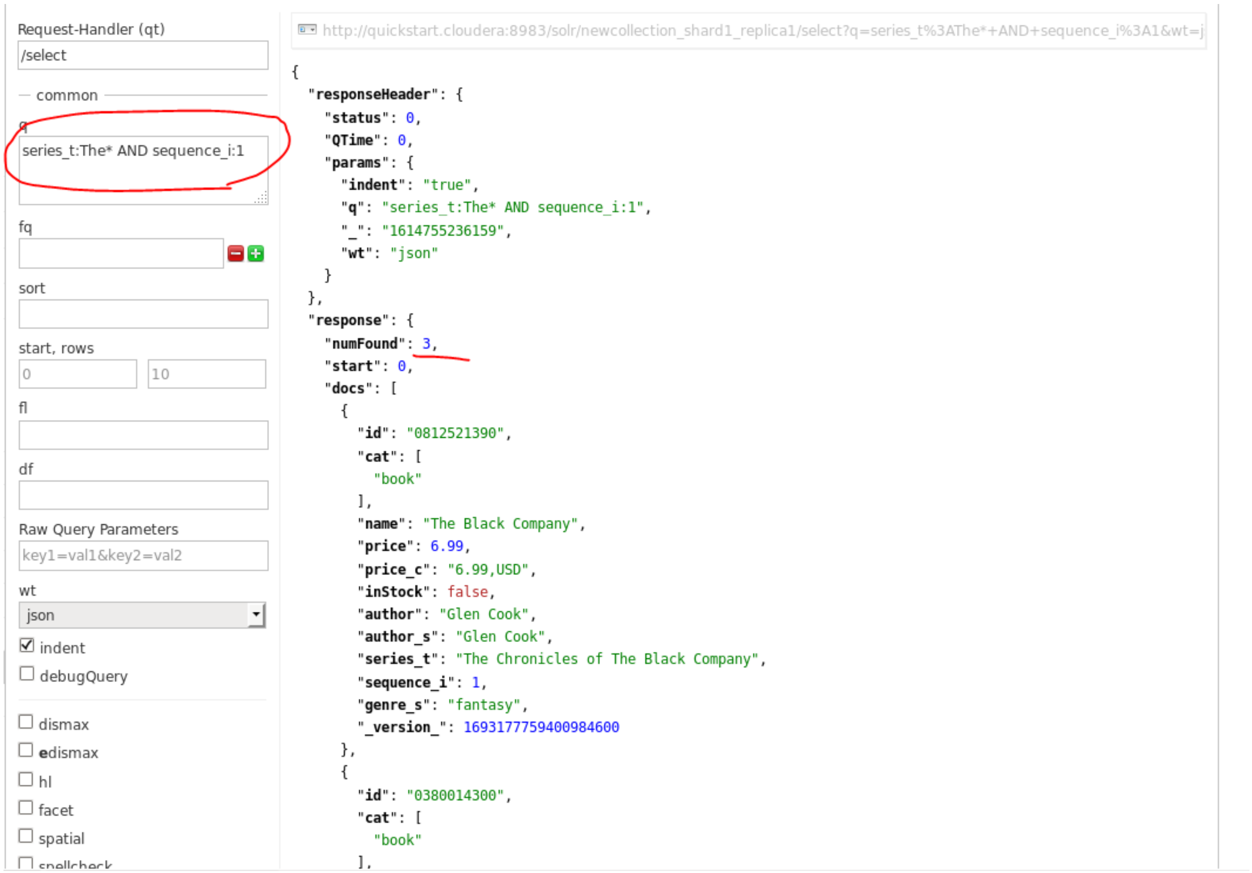
□ spatial

http://quickstart.cloudera:8983/solr/newcollection_shard1_replica1/select?q=price%3A%5B5+TO+7%5D+OR+id%3A08*&wt

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "indent": "true",
      "q": "price:[5 TO 7] OR id:08*",
      "_": "1614755033007",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 5,
    "start": 0,
    "docs": [
      {
        "id": "0812521390",
        "cat": [
          "book"
        ],
        "name": "The Black Company",
        "price": 6.99,
        "price_c": "6.99,USD",
        "inStock": false,
        "author": "Glen Cook",
        "author_s": "Glen Cook",
        "series_t": "The Chronicles of The Black Company",
        "sequence_i": 1,
        "genre_s": "fantasy",
        "_version_": 1693177759400984600
      },
      {
        "id": "0812550706",
        "cat": [
          "book"
```

=>Using Range Search and Wildcard Search, only display the data that price field value is between 5 to 7 OR the id field value includes 08.

5)Query5(series_t:The* AND sequence_i:1)



Request-Handler (qt)
/select

common

q
series_t:The* AND sequence_i:1

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

☑ indent
☐ debugQuery

☐ dismax
☐ edismax
☐ hl
☐ facet
☐ spatial
☐ spellcheck

http://quickstart.cloudera:8983/solr/newcollection_shard1_replica1/select?q=series_t%3AThe*+AND+sequence_i%3A1&wt=j

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "indent": "true",
      "q": "series_t:The* AND sequence_i:1",
      "_": "1614755236159",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 3,
    "start": 0,
    "docs": [
      {
        "id": "0812521390",
        "cat": [
          "book"
        ],
        "name": "The Black Company",
        "price": 6.99,
        "price_c": "6.99,USD",
        "inStock": false,
        "author": "Glen Cook",
        "author_s": "Glen Cook",
        "series_t": "The Chronicles of The Black Company",
        "sequence_i": 1,
        "genre_s": "fantasy",
        "_version_": 1693177759400984600
      },
      {
        "id": "0380014300",
        "cat": [
          "book"
        ],
```

=> Only displays the data that series_t field value includes The And sequence_i's field value is 1.