

Week 8 Assignment- Apache Solr

Guruprasad Velikadu Krishnamoorthy

College of Science and Technology, Bellevue University

DSC650-T301: Big Data

Professor. Nasheb Ismaily

January 29, 2024

1. Environment Initialization

```
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/solr$ docker-compose up -d
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/solr$
Creating network "solr_default" with the default driver
Creating solr_solr_1 ... done
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/solr$
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/solr$ docker ps
CONTAINER ID   IMAGE                                COMMAND                  CREATED        STATUS        PORTS                               NAMES
49897ecb2126   solr                                 "docker-entrypoint.s..." 6 seconds ago  Up 5 seconds  0.0.0.0:8983->8983/tcp, :::8983->8983/tcp  solr_solr_1
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/solr$
KRISUNICORN@bigdata-ds650:~/dsc650-infra/bellevue-bigdata/solr$ docker exec -it solr_solr_1 bash
solr@49897ecb2126:/opt/solr-9.4.0$
solr@49897ecb2126:/opt/solr-9.4.0$
```

2. Creating a Solr Collection

```
/opt/solr/bin/solr create -c mynewcollection
```

Verify that 'mynewcollection' has been successfully created.

The screenshot shows the Solr Admin web interface in a browser. The address bar shows the URL `http://localhost:8983/solr/#/mynewcollection/core-overview`. The left sidebar contains a navigation menu with items like Dashboard, Logging, Security, Core Admin, Java Properties, Thread Dump, and a dropdown menu for 'mynewcollection'. The 'mynewcollection' dropdown is highlighted. The main content area displays the 'Overview' tab for the 'mynewcollection' core. It includes a 'Statistics' section with fields like Last Modified, Num Docs, Max Doc, Deleted Docs, Version, Segment Count, and Current. Below this is a 'Replication (Leader)' table showing the status of the leader (Searching) and replicable. The 'Instance' section on the right shows the CWD, Instance, Data, Index, and Impl paths. The 'Healthcheck' section indicates that the ping request handler is not configured with a healthcheck file. The bottom of the browser window shows the Windows taskbar with the system clock displaying 10:25 AM on 12/24/2023.

3. Generating and Adding Data to the Collection

```
echo '[
  {"id":"1", "name":"Product A", "category":"Electronics",
  "price":"100"},
  {"id":"2", "name":"Product B", "category":"Books",
  "price":"20"}
]' > /tmp/products.json
```

```
/opt/solr/bin/post -c mynewcollection /tmp/products.json
```

[illegible]

```
solr@43b946b792:/opt/solr-9.4.0 cd /tmp
solr@43b946b792:/tmp$ cd -
/opt/solr-9.4.0
solr@43b946b792:/opt/solr-9.4.0$ cat /tmp/products.json
{
  [{"id":1,"name":"Product A","category":["Electronics","price":100],
  ("id":2,"name":"Product B","category":["Books","price":20]}
]
solr@43b946b792:/opt/solr-9.4.0$ curl "http://localhost:8983/solr/mynewcollection/select?q=*:*"
{"responseHeader":{"status":0,
"qtime":0,
"params":{"q":"*:*"}
},
"numFound":10,
"start":0,
"numFoundExact":true,
"docs":[{"id":1,
"name":["Product A"],
"category":["Electronics"],
"price":100,
"version_":"17961814892292000"},
{"id":2,
"name":["Product B"],
"category":["Books"],
"price":20,
"version_":"17961814892292000"}
]}
solr@43b946b792:/opt/solr-9.4.0$ curl "http://localhost:8983/solr/mynewcollection/select?q=category:Electronics"
{"responseHeader":{"status":0,
"qtime":16,
"params":{"q":"category:Electronics"}
},
"numFound":1,
"start":0,
"numFoundExact":true,
"docs":[{"id":1,
"name":["Product A"],
"category":["Electronics"],
"price":100,
"version_":"17961814892292000"}
]}
solr@43b946b792:/opt/solr-9.4.0$
```

Query 'mynewcollection' in Solr Web Interface

The screenshot shows the Solr Admin web interface in a browser. The address bar displays the URL: `http://localhost:8983/solr/mynewcollection/query?q=*:*&q.op=OR&indent=true&useParams=`. The interface is divided into a left sidebar with navigation links (Dashboard, Logging, Security, Core Admin, Java Properties, Thread Dump, mynewcollection, Overview, Analysis, Documents, Parameters, Files, Ping, Plugins / State, Replication, Schema, Segments info) and a main content area. The main content area is titled 'Request-Handler (/solr)' and shows the query `/select`. The 'common' section has 'q' set to `*:*` and 'sort' set to `start,rows`. The 'start,rows' section has 'start' set to `0` and 'rows' set to `10`. The 'params' section has 'wt' set to `json` and 'indent' checked. The 'JSON Query' section contains the query `*:*`. The 'Execute Query' button is visible. On the right, the JSON response is displayed, showing the search results for the query `*:*`. The response includes the status, qtime, params, numFound, start, numFoundExact, and a list of documents. The documents are Product A and Product B. The response is highlighted with red boxes around the `numFound` and `start` fields.

6. Try a faceted search to count the number of products in each category.

The screenshot shows the Solr Admin web interface. On the left, the 'Query' tab is selected, and the 'facet' checkbox is checked. The 'facet.field' is set to 'category'. The URL in the browser is `http://localhost:8983/solr/#/mynewcollection/select?fq=*:*&q.op=OR&indent=true&facet=true&facet.field=category&useParams=`. The right pane displays the JSON response, which includes a `facet_counts` section with `facet_fields` and `facet_counts` arrays. The `facet_fields` array contains `category` and `price` fields. The `facet_counts` array contains `product` and `price` fields. The `product` field is highlighted with a red box, and the `price` field is highlighted with a red box.

6.2. Faceting based on Price:

The screenshot shows a terminal window with a `curl` command and its output. The command is `curl -X GET "http://localhost:8983/solr/#/mynewcollection/select?q=*:*&facet=true&facet.field=price"`. The output is a JSON response that includes a `facet_counts` section with `facet_fields` and `facet_counts` arrays. The `facet_fields` array contains `price` and `product` fields. The `facet_counts` array contains `product` and `price` fields. The `price` field is highlighted with a red box, and the `product` field is highlighted with a red box.

7. Use the Web Interface to experiment further with various query parameters and filters.

7.1.Sorting the data based on the “id” column in descending order:

The screenshot shows the Solr Admin web interface. The URL in the browser is `http://localhost:8983/solr/#/mynewcollection/query?q=*:*&q.op=OR&indent=true&sort=id%20desc&useParams=`. The 'Request-Handler (st)' section is set to `/select`. The 'common' section has `q` set to `*:*`. The 'sort' section is set to `id desc`. The 'start, rows' section has `start` at 0 and `rows` at 10. The 'df' section is empty. The 'paramset(s)' section is empty. The 'wt' section is set to `json`. The 'indent on' checkbox is checked. The 'debugQuery' checkbox is unchecked. The 'defType' section is set to `lucene`. The 'Raw Query Parameters' section is empty. The 'JSON Query' section is empty. The 'Execute Query' button is visible. The response on the right is a JSON object with the following structure:

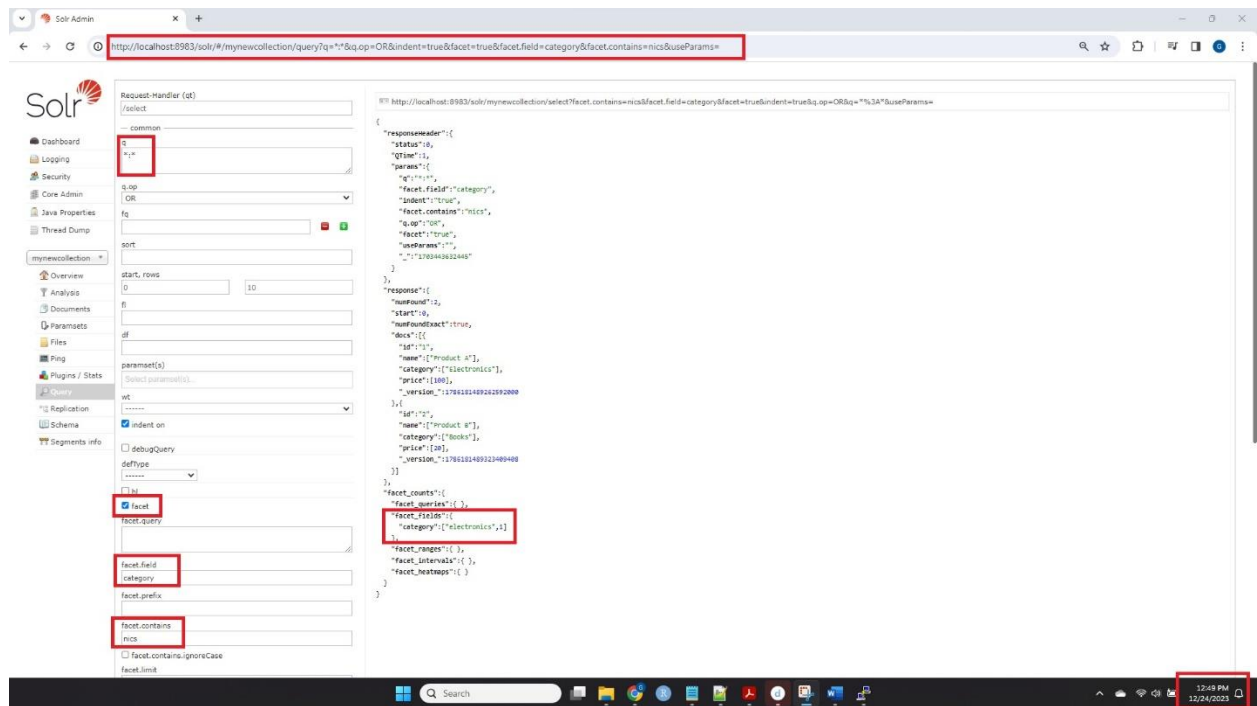
```
{
  "responseHeader": {
    "status": 0,
    "qtzime": 1,
    "params": {
      "q": "*:*",
      "indent": "true",
      "q.op": "OR",
      "sort": "id desc",
      "useParams": "true",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 2,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "id": "1706131409212409480",
        "category": ["Books"],
        "price": 100,
        "version": "1.0"
      },
      {
        "id": "1706131409212409480",
        "category": ["Electronics"],
        "price": 100,
        "version": "1.0"
      }
    ]
  }
}
```

7.2. Using “Price” in the field list filter:

The screenshot shows the Solr Admin web interface. The URL in the browser is `http://localhost:8983/solr/#/mynewcollection/query?q=*:*&q.op=OR&indent=true&fl=price&useParams=`. The 'Request-Handler (st)' section is set to `/select`. The 'common' section has `q` set to `*:*`. The 'sort' section is empty. The 'start, rows' section has `start` at 0 and `rows` at 10. The 'df' section is empty. The 'paramset(s)' section is empty. The 'wt' section is set to `json`. The 'indent on' checkbox is checked. The 'debugQuery' checkbox is unchecked. The 'defType' section is set to `lucene`. The 'Raw Query Parameters' section is empty. The 'JSON Query' section is empty. The 'Execute Query' button is visible. The response on the right is a JSON object with the following structure:

```
{
  "responseHeader": {
    "status": 0,
    "qtzime": 1,
    "params": {
      "q": "*:*",
      "indent": "true",
      "q.op": "OR",
      "fl": "price",
      "useParams": "true",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 2,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "price": 100
      },
      {
        "price": 100
      }
    ]
  }
}
```

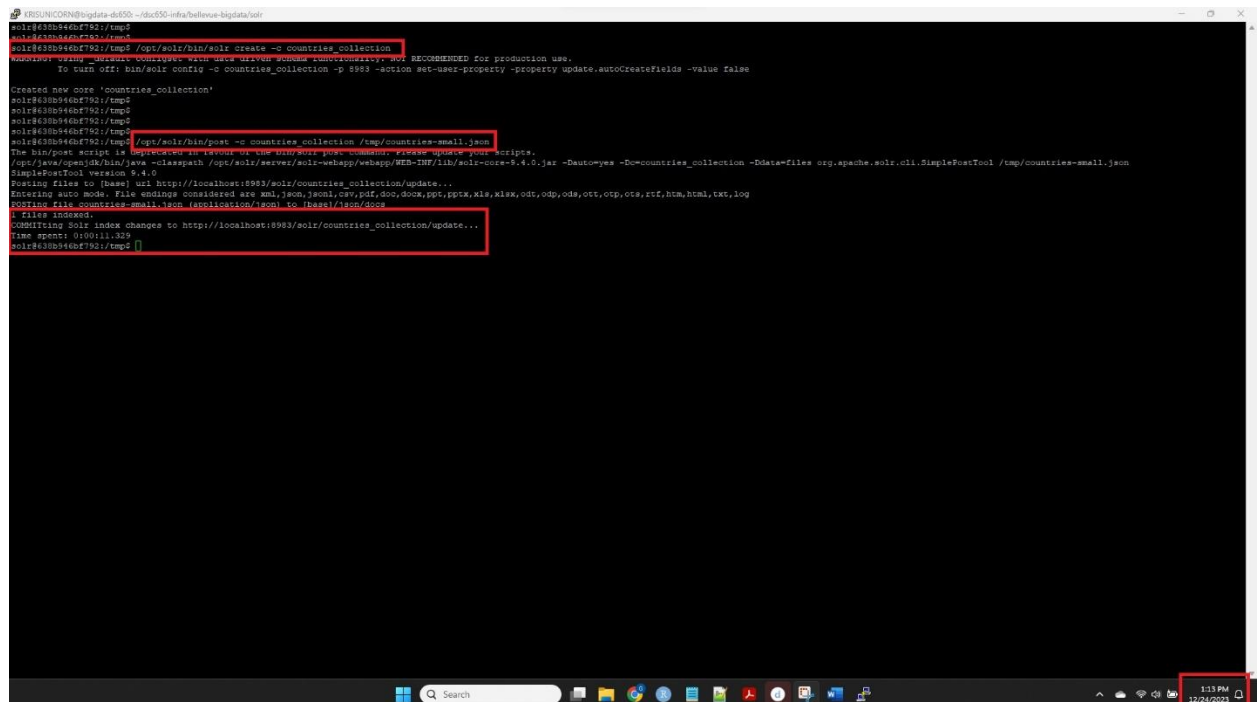
7.3. Using facet on “category” and facet.contains=“nics”



8. Custom JSON Data

8.1. Create a new collection in Solr

Creating a new collection named “**countries_collection**” to search the countries file.



8.2. Querying the countries collection that returns 248 items

The screenshot shows the Solr Admin interface. The URL bar contains the query: `http://localhost:8963/solr/#/countries_collection/query?q=*:*&q.op=OR&indent=true&useParams=`. The left sidebar shows the 'countries_collection' selected. The main panel displays the query results in JSON format, showing a response with 248 items. The response structure includes a 'responseHeader' and a 'response' array. The first item in the array is a country object with fields like 'name', 'iso3', 'iso2', 'lat', 'lon', 'area', 'population', 'capital', 'currency', 'languages', 'translations', etc.

8.3. Faceting on the “currency” field in the collection

The screenshot shows the Solr Admin interface with a query that includes faceting: `http://localhost:8963/solr/#/countries_collection/query?q=*:*&q.op=OR&indent=true&facet=true&facet.field=currency&useParams=`. The left sidebar shows the 'countries_collection' selected. The main panel displays the query results in JSON format, showing a response with faceting information. The response structure includes a 'responseHeader', a 'response' array, and a 'facet_counts' object. The 'facet_counts' object contains a 'facet_fields' section with a 'currency' field, listing various currencies and their counts.

8.4. Faceting on the "regions" field in the collection:

The screenshot shows the Solr Admin web interface. The browser address bar contains the URL: `http://localhost:8983/solr/#/countries_collection/query?q=*:*&q.op=OR&indent=true&facet=true&facet.field=region&useParams=true`. The main content area displays a JSON response from the Solr query. The response includes a large array of document objects, each representing a country with various fields like `name.common`, `name.official`, `region`, and `translations`. At the bottom of the JSON response, there is a `facet_counts` object. This object contains `facet_queries`, `facet_fields`, `facet_ranks`, `facet_intervals`, and `facet_histograms`. The `facet_fields` section shows the results of the faceting on the `region` field, listing regions like `afrika`, `america`, `europa`, `asia`, and `oceania` with their respective counts.

8.5. Filtering for the region that contains "Americas"

The screenshot shows the Solr Admin web interface. The browser address bar contains the URL: `http://localhost:8983/solr/#/countries_collection/query?q=region:Americas&q.op=OR&indent=true&useParams=true`. The left sidebar shows the Solr Admin navigation menu. The main content area displays a JSON response from the Solr query. The response includes a `responseHeader` object with `status` and `params`. The `params` object contains `q` (the query), `q.op` (the operator), and `useParams`. The `response` object contains `numFound` (the number of documents found) and `docs` (an array of document objects). The `docs` array contains a single document object representing a country in the Americas, with fields like `name.common`, `name.official`, `region`, and `translations`.

8.6. Applying Multiple filters

In this example, the results are sorted by the area in **descending order**. Only the top 5 rows from the results are extracted using the **start,rows** option. The results only returns selected fields mentioned in the “**field list**” box. The subregion is set as the “**default field**” and the default value of “America” is included in the “**query**” box.

The screenshot displays the Solr Admin interface for the 'countries_collection' index. The query configuration is as follows:

- Request-Handler:** (q)
- select:** (select)
- common:** America
- q.op:** AND
- fq:** (empty)
- sort:** area desc
- start, rows:** 0, 5
- df:** name,common,altSpellings,area,capital,currency,region,subregion
- subregion:** (empty)
- parameters:** (empty)
- debugQuery:** (unchecked)
- type:** (empty)
- format:** json
- indent on:** (checked)

The raw query parameters are: `q=America&q.op=AND&df=name,common,altSpellings,area,capital,currency,region,subregion&df=subregion&sort=area desc&start=0&rows=5&wt=json&f=name,common,altSpellings,area,capital,currency,region,subregion&indent=true&q.op=AND&format=json`

The response is a JSON array of 5 objects, each representing a country. The fields included are: `altSpellings`, `area`, `capital`, `currency`, `region`, and `subregion`.

```
[{"altSpellings": ["CA", "Canada"], "area": 9984670, "capital": "Ottawa", "currency": "CAD", "name": "Canada", "region": "Americas", "subregion": "Northern America"}, {"altSpellings": ["US", "USA", "United States of America"], "area": 9833517, "capital": "Washington D.C.", "currency": "USD", "name": "United States", "region": "Americas", "subregion": "Northern America"}, {"altSpellings": ["BR", "Brazil", "Federative Republic of Brazil", "Rep\u00fablica Federativa do Brasil"], "area": 8511963, "capital": "Bras\u00edlia", "currency": "BRL", "name": "Brazil", "region": "Americas", "subregion": "South America"}, {"altSpellings": ["AR", "Argentine Republic", "Rep\u00fablica Argentina"], "area": 2780400, "capital": "Buenos Aires", "currency": "ARS", "name": "Argentina", "region": "Americas", "subregion": "South America"}, {"altSpellings": ["CL", "Chile", "Republic of Chile", "Rep\u00fablica de Chile"], "area": 756102, "capital": "Santiago", "currency": "CLP", "name": "Chile", "region": "Americas", "subregion": "South America"}]
```