

HASH AGILE TECHNOLOGIES

TASK- 2


TASK GIVEN :

1. Read about Apache Solr: [Solr Tutorial](#).
2. Install Solr on your local machine.
3. Create a collection in Solr.
4. Index the Employee data from <https://www.kaggle.com/datasets/williamlucas0/employee-sample-data>

STEPS USED :

1) INSTALLATION

Downloaded zip file (Solr.9.7.0.tgz) from Apache Solr Downloader

 [BLOG](#) [NEWS](#) [SECURITY](#) [FEATURES](#) [RESOURCES](#) [COMMUNITY](#) [PROJECT](#) [SOLR OPERATOR](#) [DOWNLOAD](#)

Solr Downloads

Official releases are usually created when the [developers](#) feel there are sufficient changes, improvements and bug fixes to warrant a release. Due to the voluntary nature of Solr, no releases are scheduled in advance.

Solr 9.7.0

Solr 9.7.0 is the most recent Apache Solr release.

- Source release: [solr-9.7.0-src.tgz](#) [PGP] [SHA512]
- Binary releases: [solr-9.7.0.tgz](#) [PGP] [SHA512] , [solr-9.7.0-slim.tgz](#) [PGP] [SHA512]
- OpenAPI specification: [solr-openapi-9.7.0.json](#) [PGP] [SHA512]
- Docker: [solr:9.7.0](#), [solr:9.7.0-slim](#)
- [Change log](#)

Solr 8.11.4

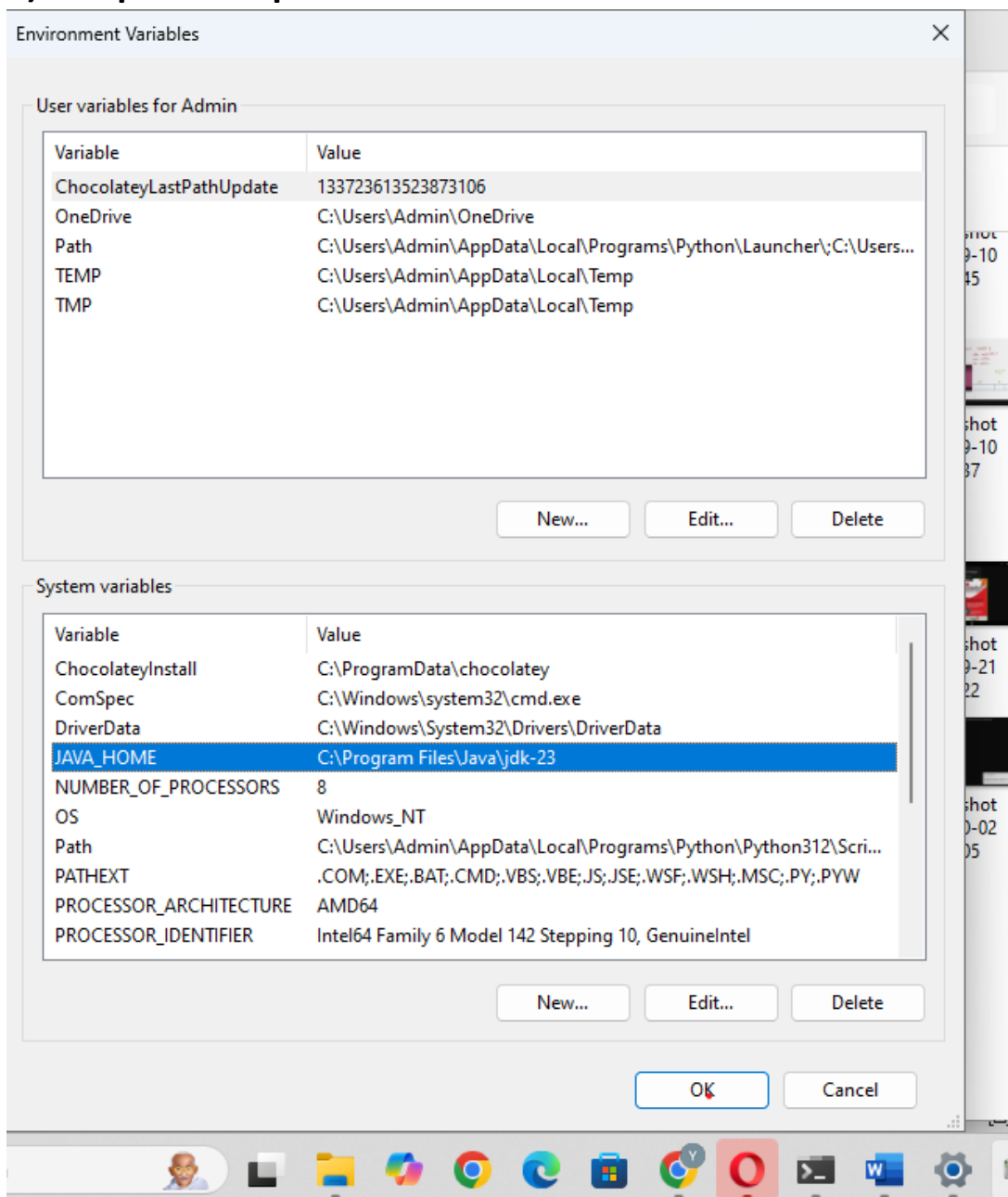
Solr 8.11.4 is the latest release in the 8.x series.

2) Next , Installed Java JDK 23 for establishing connection in the system

```
Microsoft Windows [Version 10.0.22631.4169]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Admin>java --version
java 23 2024-09-17
Java(TM) SE Runtime Environment (build 23+37-2369)
Java HotSpot(TM) 64-Bit Server VM (build 23+37-2369, mixed mode, sharing)
```

3) JDK path setup



Java Environment Variables is created as JAVA_HOME and path has been set up

4) SOLR START

To start the solr the start command should be initialized in command prompt inside the solr/bin directory .

COMMAND :

C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\bin>solr start

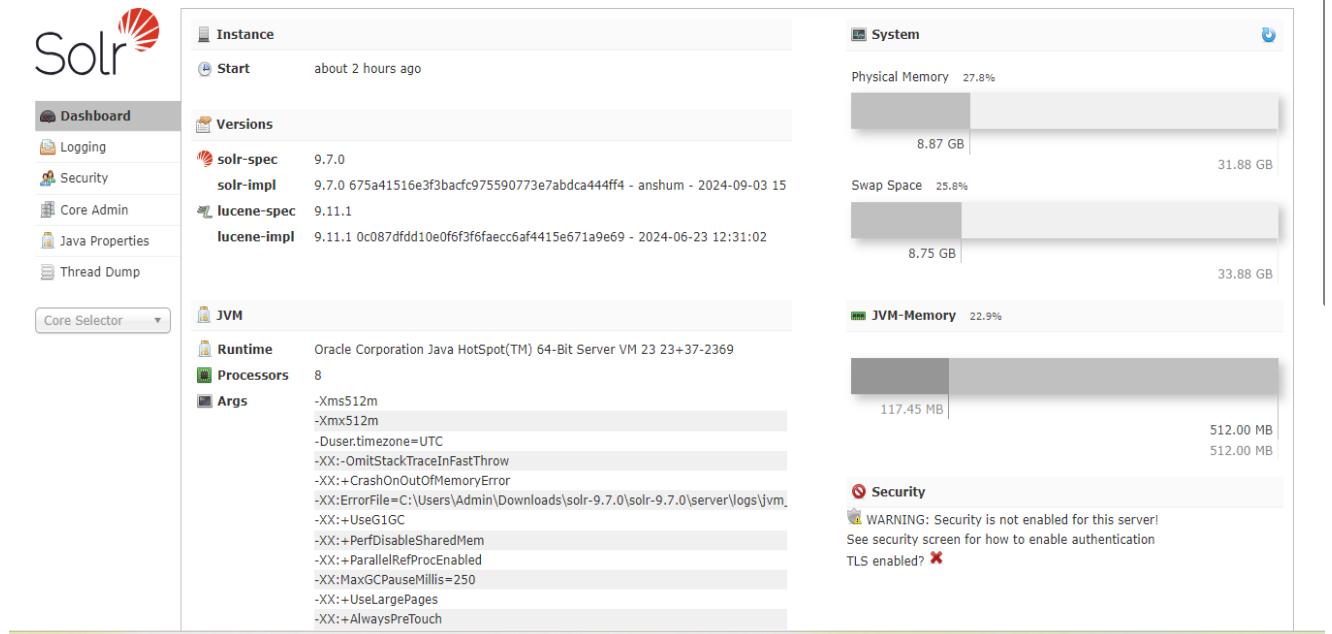
```
C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\bin>solr start
Java 23 detected. Enabled workaround for SOLR-16463
Java HotSpot(TM) 64-Bit Server VM warning: JVM cannot use large page memory because it does not have enough privilege to
lock pages in memory.
WARNING: A command line option has enabled the Security Manager
WARNING: The Security Manager is deprecated and will be removed in a future release
Failed to parse command-line arguments due to: Unrecognized option: --max-wait-secs

usage: bin/solr status [-maxWaitSecs <SECS>] [--solr-url <URL>]

List of options:
  -maxWaitSecs <SECS>  Wait up to the specified number of seconds to see Solr running.
  --solr-url <URL>     Address of the Solr Web application, defaults to: http://localhost:8983.
```

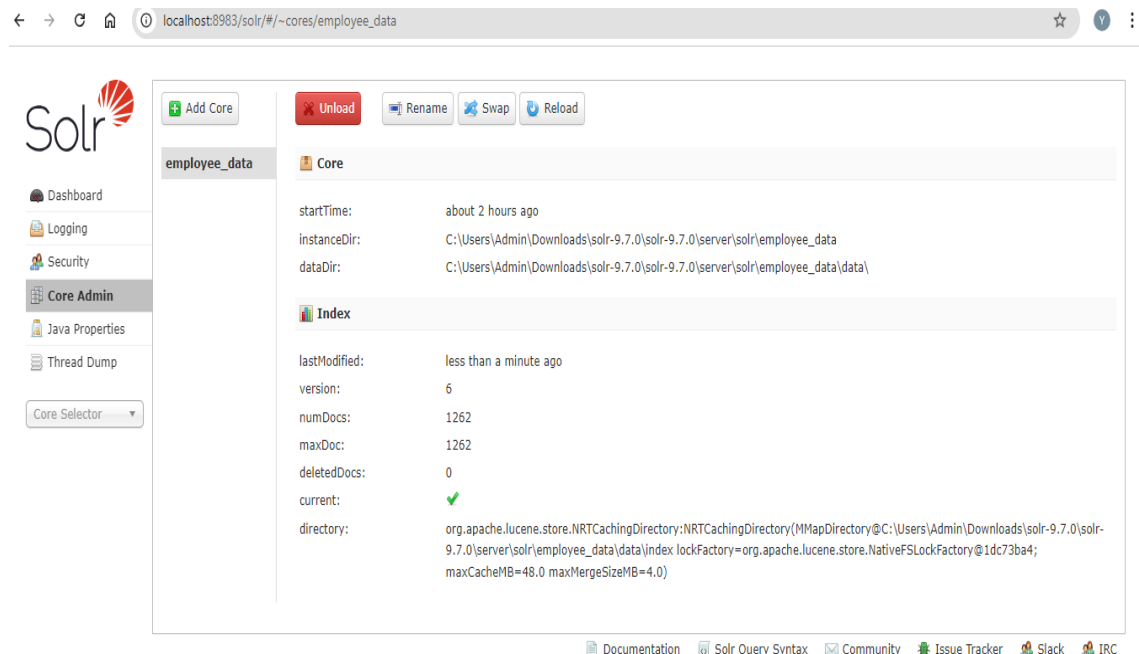
5) LOCAL HOSTING

After the above step , the local machine setup will be initialized will be in the **local host :8393 port** .



6) CREATING COLLECTION

To create a collection , **create-c employee_data**(name of the collection) command is used .



COMMAND

```
C:\Users\Admin\Downloads\solr9.7.0\solr-9.7.0\bin>solr create -c employee_data
```

Here , the employee data collection is created inside the core admin

7) CONVERTING CSV FILE TO JSON FILE

Next step is to convert the csv file to json format , for that I have used python script for automating

CODE :

```
import csv
import json

# Input and output file paths
csv_file = 'Employee Sample Data 1.csv' # Name of the csv file
json_file = 'employee_data.json'

# Read CSV and convert to JSON
data = []

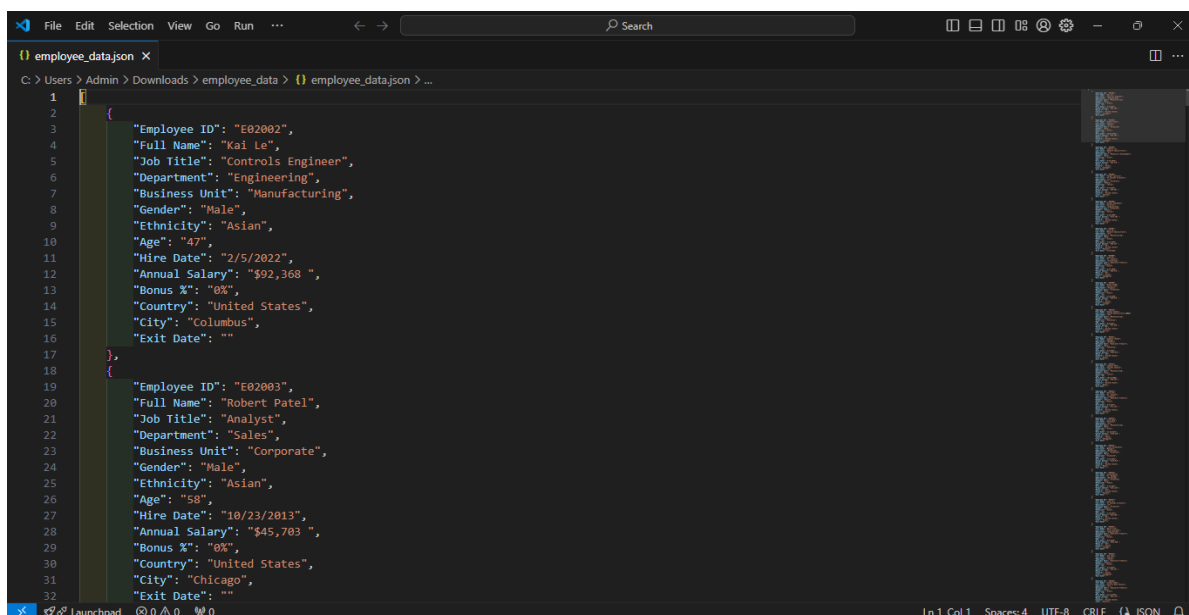
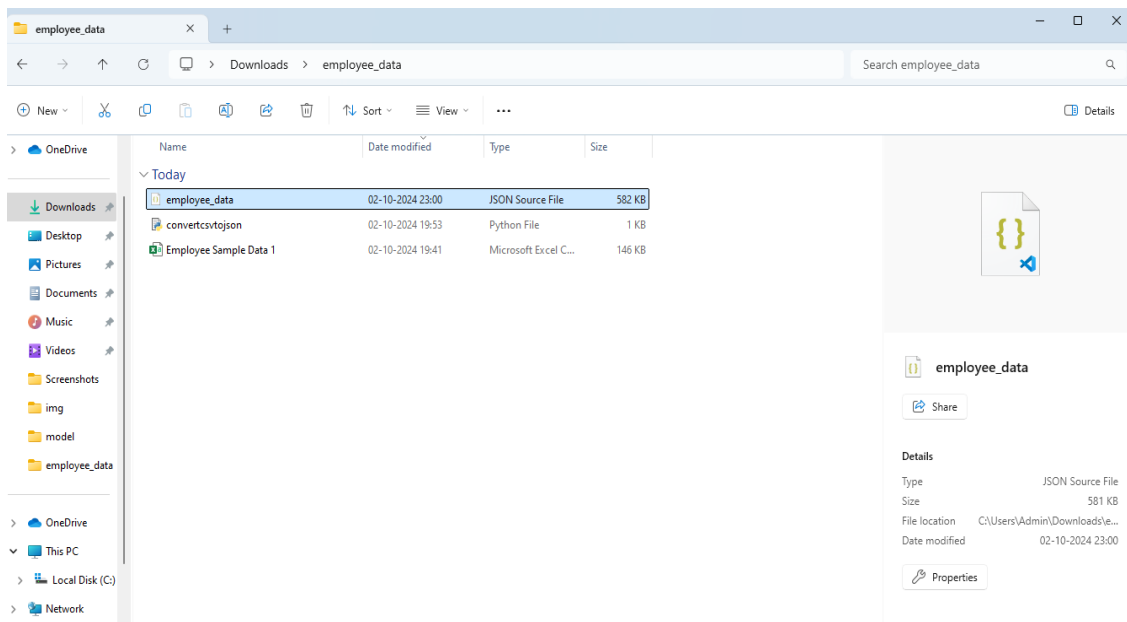
with open(csv_file, newline='', encoding='ISO-8859-1') as csvfile: #
    Change encoding here
    reader = csv.DictReader(csvfile)
    for row in reader:
        data.append(row)

# Write JSON data to file
with open(json_file, 'w', encoding='utf-8') as jsonfile:
    json.dump(data, jsonfile, indent=4)
```

8) EXECUTION

Executed this code as .py file using pythonIDLE ,in the path where the excel file [Employee Test Data 1] is located. Which was downloaded from Kaggle using the provided link.

After executing this script the employee_data.json file will be created

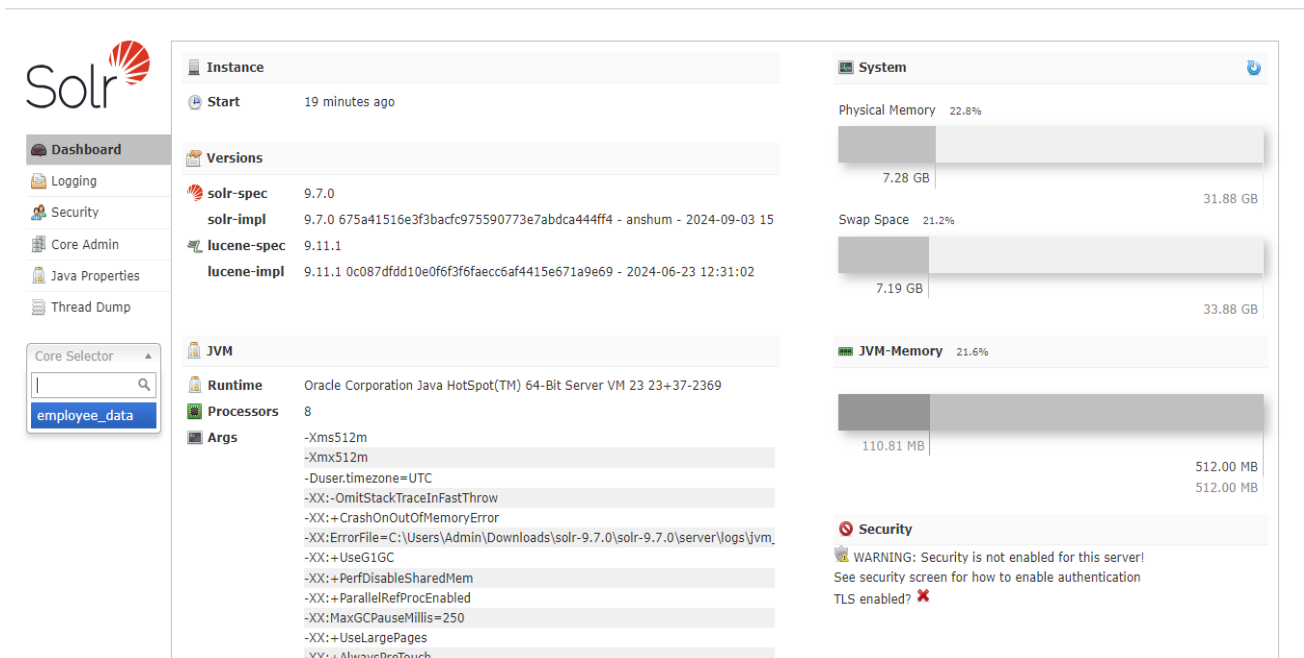


9) POSTING THE JSONFILE TO COLLECTION

Using the **solr.cmd**: to post the JSON file in to the collection

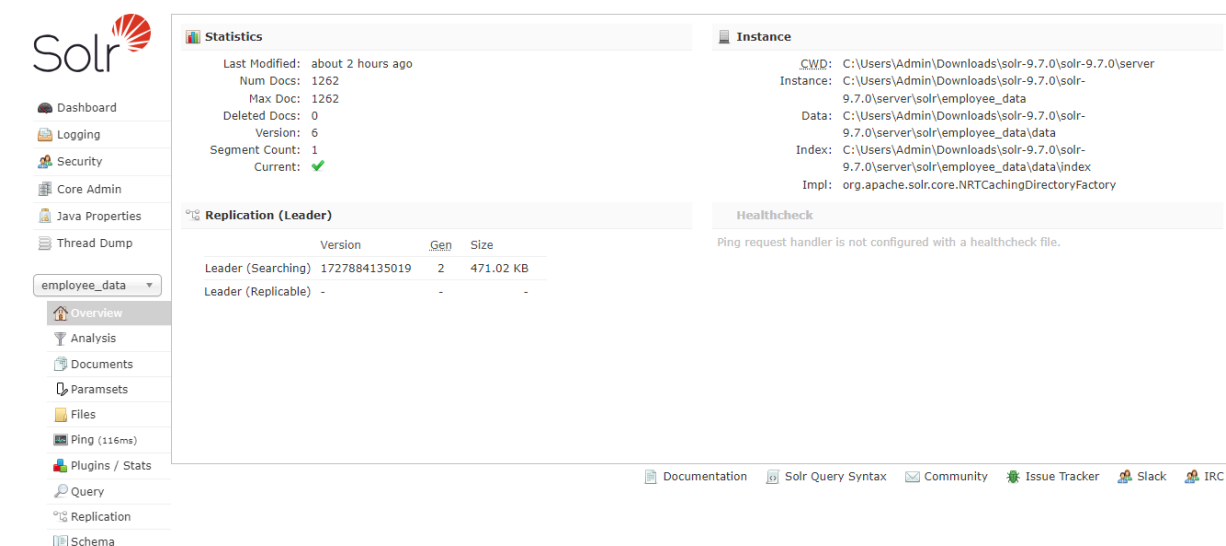
Command :

C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\bin\solr.cmd post -c employee_data C:\Users\Admin\Downloads\employee_data\employee_data.json



The Solr Admin UI Dashboard shows the following information:

- Instance:** Start 19 minutes ago.
- Versions:**
 - solr-spec:** 9.7.0
 - solr-impl:** 9.7.0 675a41516e3f3bacfc975590773e7abdca444ff4 - anshum - 2024-09-03 15
 - lucene-spec:** 9.11.1
 - lucene-impl:** 9.11.1 0c087dfdd10e0f6f3f6faecc6af4415e671a9e69 - 2024-06-23 12:31:02
- JVM:**
 - Runtime:** Oracle Corporation Java HotSpot(TM) 64-Bit Server VM 23.23+37-2369
 - Processors:** 8
 - Args:**
 - Xms512m
 - Xmx512m
 - Duser.timezone=UTC
 - XX:-OmitStackTraceInFastThrow
 - XX:+CrashOnOutOfMemoryError
 - XX:ErrorFile=C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\server\logs\jvm
 - XX:+UseG1GC
 - XX:+PerfDisableSharedMem
 - XX:+ParallelRefProcEnabled
 - XX:MaxGCPauseMillis=250
 - XX:+UseLargePages
 - XX:+AlwaysPreTouch
- System:**
 - Physical Memory:** 22.8% (7.28 GB / 31.88 GB)
 - Swap Space:** 21.2% (7.19 GB / 33.88 GB)
 - JVM-Memory:** 21.6% (110.81 MB / 512.00 MB)
- Security:** WARNING: Security is not enabled for this server! See security screen for how to enable authentication. TLS enabled? ❌



The Solr Admin UI shows the following details for the **employee_data** collection:

- Statistics:**
 - Last Modified: about 2 hours ago
 - Num Docs: 1262
 - Max Doc: 1262
 - Deleted Docs: 0
 - Version: 6
 - Segment Count: 1
 - Current: ✓
- Replication (Leader):**

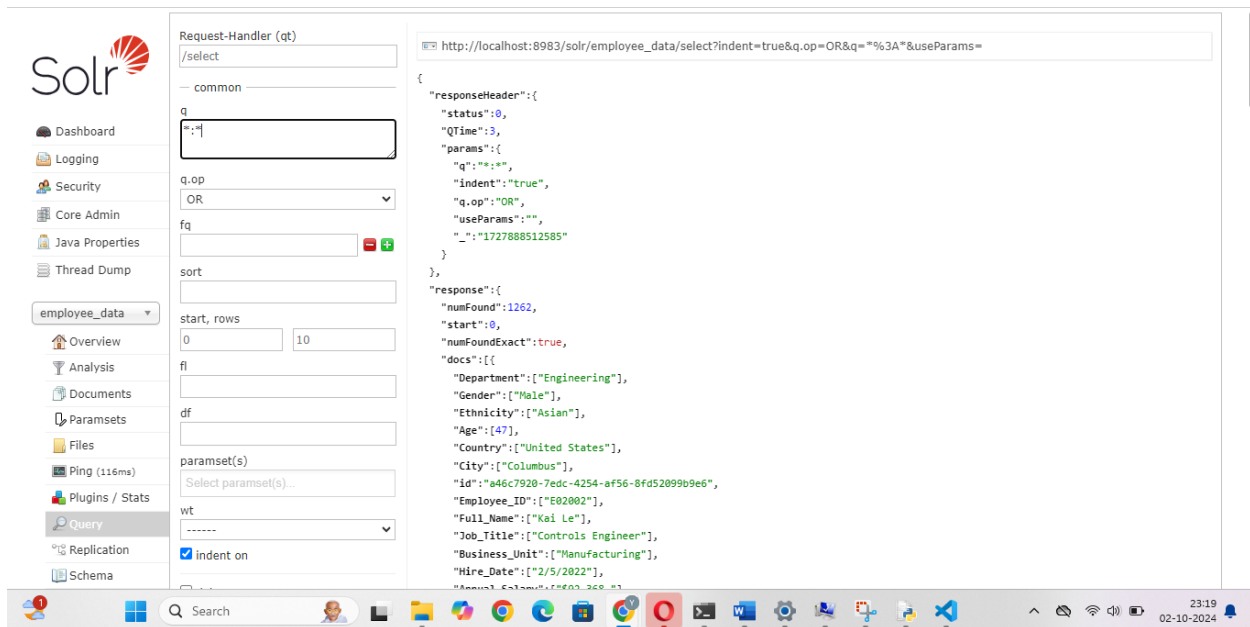
	Version	Gen	Size
Leader (Searching)	1727884135019	2	471.02 KB
Leader (Replicable)	-	-	-
- Instance:**
 - CWD: C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\server
 - Instance: C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\server\solr\employee_data
 - Data: C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\server\solr\employee_data\data
 - Index: C:\Users\Admin\Downloads\solr-9.7.0\solr-9.7.0\server\solr\employee_data\data\index
 - Impl: org.apache.solr.core.NRTCachingDirectoryFactory
- Healthcheck:** Ping request handler is not configured with a healthcheck file.

9) EXECUTING QUERIES

Getting all the values in json format

Query:

.



The screenshot shows the Solr Admin UI with the following details:

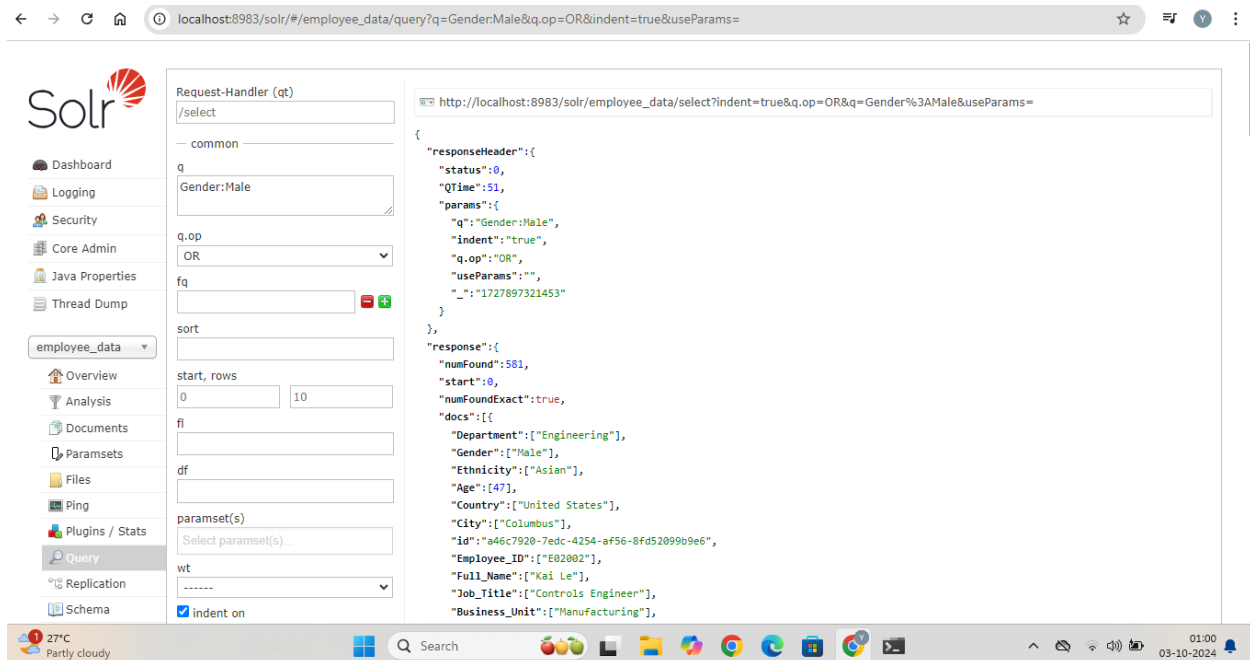
- Request-Handler (qt):** /select
- q:** *.*
- q.op:** OR
- fq:** (empty)
- sort:** (empty)
- start, rows:** 0, 10
- fl:** (empty)
- df:** (empty)
- paramset(s):** Select paramset(s)
- wt:** (empty)
- indent on:** ☒

The URL in the address bar is: `http://localhost:8983/solr/employee_data/select?indent=true&q.op=OR&q=*. *&useParams=`

The response is a JSON object:

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 3,
    "params": {
      "q": " *.*",
      "indent": "true",
      "q.op": "OR",
      "useParams": "",
      "_": "1727888512585"
    }
  },
  "response": {
    "numFound": 1262,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "Department": ["Engineering"],
        "Gender": ["Male"],
        "Ethnicity": ["Asian"],
        "Age": [47],
        "Country": ["United States"],
        "City": ["Columbus"],
        "id": "a46c7920-7edc-4254-af56-8fd52099b9e6",
        "Employee_ID": ["E02002"],
        "Full_Name": ["Kai Le"],
        "Job_Title": ["Controls Engineer"],
        "Business_Unit": ["Manufacturing"],
        "Hire_Date": ["2/5/2022"],
        "Annual_Salary": [55000.00]
      }
    ]
  }
}
```

Query to obtain male employees - **Gender:Male**



The screenshot shows the Solr Admin UI with the following details:

- Request-Handler (qt):** /select
- q:** Gender:Male
- q.op:** OR
- fq:** (empty)
- sort:** (empty)
- start, rows:** 0, 10
- fl:** (empty)
- df:** (empty)
- paramset(s):** Select paramset(s)
- wt:** (empty)
- indent on:** ☒

The URL in the address bar is: `http://localhost:8983/solr/employee_data/select?indent=true&q.op=OR&q=Gender%3AMale&useParams=`

The response is a JSON object:

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 51,
    "params": {
      "q": "Gender:Male",
      "indent": "true",
      "q.op": "OR",
      "useParams": "",
      "_": "1727897321453"
    }
  },
  "response": {
    "numFound": 581,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "Department": ["Engineering"],
        "Gender": ["Male"],
        "Ethnicity": ["Asian"],
        "Age": [47],
        "Country": ["United States"],
        "City": ["Columbus"],
        "id": "a46c7920-7edc-4254-af56-8fd52099b9e6",
        "Employee_ID": ["E02002"],
        "Full_Name": ["Kai Le"],
        "Job_Title": ["Controls Engineer"],
        "Business_Unit": ["Manufacturing"],
        "Hire_Date": ["2/5/2022"],
        "Annual_Salary": [55000.00]
      }
    ]
  }
}
```


Analysis of Age Column

The screenshot shows the Solr Admin interface for the 'employee_data' collection. The 'Analysis' tab is selected, and the 'Age' field is being analyzed. The 'Field Value (Index)' is 'Age'. The 'Field Value (Query)' is empty. The 'Analyze Fieldname / FieldType' is 'Annual_Salary'. The 'Verbose Output' checkbox is checked. The 'Analyze Values' button is visible. The analysis results are displayed in a table with three sections: ST, SF, and LCF.

Section	Field	Type	Value
ST	text	Age	
	raw_bytes		[41 67 65]
	start		0
	end		3
	positionLength		1
	type		<ALPHANUM>
SF	text	Age	
	raw_bytes		[41 67 65]
	start		0
	end		3
	positionLength		1
	type		<ALPHANUM>
LCF	text	age	
	raw_bytes		[61 67 65]
	start		0
	end		3
	positionLength		1
	type		<ALPHANUM>

Query to get data's where age=25

Age:25

The screenshot shows the Solr Admin interface for the 'employee_data' collection. The 'Query' tab is selected. The 'Request-Handler (qt)' is '/select'. The 'q' is 'Age:25'. The 'q.op' is 'OR'. The 'fq' is empty. The 'sort' is empty. The 'start, rows' is '0 10'. The 'df' is empty. The 'paramset(s)' is 'Select paramset(s)'. The 'wt' is 'json'. The 'indent on' checkbox is checked. The 'Response' is displayed in a text area.

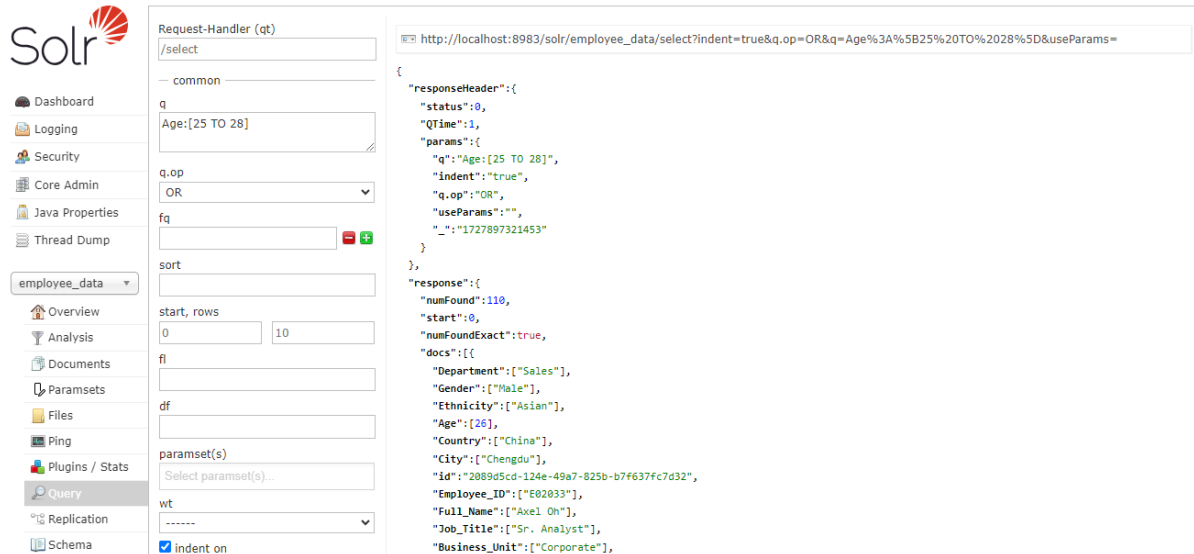
```
http://localhost:8983/solr/employee_data/select?indent=true&q.op=OR&q=Age%3A25&useParams=

{
  "responseHeader": {
    "status": 0,
    "QTime": 14,
    "params": {
      "q": "Age:25",
      "indent": "true",
      "q.op": "OR",
      "useParams": "",
      "_": "1727897321453"
    }
  },
  "response": {
    "numFound": 14,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "Department": ["Marketing"],
        "Gender": ["Female"],
        "Ethnicity": ["Latino"],
        "Age": [25],
        "Country": ["Brazil"],
        "City": ["Sao Paulo"],
        "id": "e6608945-690f-4fb3-847d-a72aae9565e2",
        "Employee_ID": ["E02180"],
        "Full_Name": ["Aubrey Ruiz"],
        "Job_Title": ["Analyst II"],
        "Business_Unit": ["Specialty Products"],

```

Query to obtain data between Age 25 to 28

Age [25 TO 28]



The Solr Admin UI shows the 'Request-Handler (qt)' section with the following configuration:

- q: /select
- q.op: OR
- fq: Age:[25 TO 28]
- sort: (empty)
- start, rows: 0, 10
- df: (empty)
- paramset(s): Select paramset(s)...
- wt: (empty)
- indent on: ☒

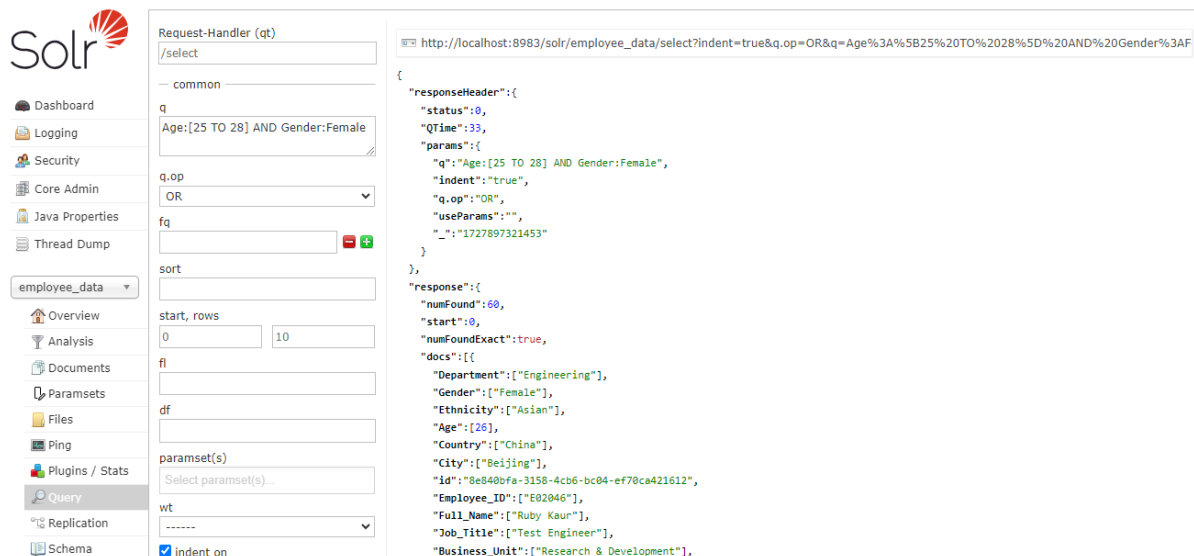
The URL bar shows: `http://localhost:8983/solr/employee_data/select?indent=true&q.op=OR&q=Age%3A%5B25%20TO%2028%5D&useParams=`

The response JSON is as follows:

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "q": "Age: [25 TO 28]",
      "indent": "true",
      "q.op": "OR",
      "useParams": "",
      "_": "1727897321453"
    }
  },
  "response": {
    "numFound": 110,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "Department": ["Sales"],
        "Gender": ["Male"],
        "Ethnicity": ["Asian"],
        "Age": [26],
        "Country": ["China"],
        "City": ["Chengdu"],
        "id": "2889d5cd-124e-49a7-825b-b7f637fc7d32",
        "Employee_ID": ["E02033"],
        "Full_Name": ["Axel Oh"],
        "Job_Title": ["Sr. Analyst"],
        "Business_Unit": ["Corporate"]
      }
    ]
  }
}
```

AND

Age:[25 TO 28] AND Gender:Female



The Solr Admin UI shows the 'Request-Handler (qt)' section with the following configuration:

- q: /select
- q.op: OR
- fq: Age:[25 TO 28] AND Gender:Female
- sort: (empty)
- start, rows: 0, 10
- df: (empty)
- paramset(s): Select paramset(s)...
- wt: (empty)
- indent on: ☒

The URL bar shows: `http://localhost:8983/solr/employee_data/select?indent=true&q.op=OR&q=Age%3A%5B25%20TO%2028%5D%20AND%20Gender%3AF`

The response JSON is as follows:

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 33,
    "params": {
      "q": "Age: [25 TO 28] AND Gender:Female",
      "indent": "true",
      "q.op": "OR",
      "useParams": "",
      "_": "1727897321453"
    }
  },
  "response": {
    "numFound": 60,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "Department": ["Engineering"],
        "Gender": ["Female"],
        "Ethnicity": ["Asian"],
        "Age": [26],
        "Country": ["China"],
        "City": ["Beijing"],
        "id": "8e840bfa-3158-4cb6-bc04-ef70ca421612",
        "Employee_ID": ["E02046"],
        "Full_Name": ["Ruby Kaur"],
        "Job_Title": ["Test Engineer"],
        "Business_Unit": ["Research & Development"]
      }
    ]
  }
}
```

OR

Age:[25 TO 28] OR Department :Engineering



- Dashboard
- Logging
- Security
- Core Admin
- Java Properties
- Thread Dump
- employee_data
- Overview
- Analysis
- Documents
- Paramsets
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema

Request-Handler (qt)

/select

common

q

Age:[25 TO 28] OR Department:Engineering

q.op

OR

fq

sort

Age asc

start, rows

0 10

fl

df

paramset(s)

Select paramset(s)...

wt

☒ indent on

http://localhost:8983/solr/employee_data/select?indent=true&q.op=OR&q=Age%3A%5B25%20TO%2028%5D%20OR%20Department%3AEngineering&sort=Age%20asc

```
{
  "responseHeader":{
    "status":0,
    "Qtime":24,
    "params":{
      "q":"Age:[25 TO 28] OR Department :Engineering",
      "indent":"true",
      "q.op":"OR",
      "sort":"Age asc",
      "useParams":"","
      "_":"1727897321453"
    }
  },
  "response":{
    "numFound":253,
    "start":0,
    "numFoundExact":true,
    "docs":[{
      "Department":["Engineering"],
      "Gender":["Female"],
      "Country":["China"],
      "City":["Shanghai"],
      "id":["d10301e6-85c1-4dd9-b47f-716b8422e105"],
      "Employee_ID":["E02504"],
      "Full_Name":["Victoria Vo"],
      "Job_Title":["Field Engineer"],
      "Business_Unit":["Corporate"],
      "Hire_Date":["12/18/2017"],
    }
  ]
}
```

SORT

Age asc

localhost:8983/solr/#/employee_data/query?q=Age%20%5B20%20TO%2030%5D&q.op=OR&indent=true&sort=Age%20asc&useParams=

Solr

Dashboard

Logging

Security

Core Admin

Java Properties

Thread Dump

employee_data

Overview

Analysis

Documents

Paramsets

Files

Ping

Plugins / Stats

Query

Replication

Schema

q

Age :[20 TO 30]

q.op

OR

fq

sort

Age asc

start, rows

0 10

fl

df

paramset(s)

Select paramset(s)...

wt

☒ indent on

☐ debugQuery

defType

hl

```
{
  "responseHeader":{
    "status":0,
    "Qtime":2,
    "params":{
      "q":"Age :[20 TO 30]",
      "indent":"true",
      "q.op":"OR",
      "sort":"Age asc",
      "useParams":"","
      "_":"1727897321453"
    }
  },
  "response":{
    "numFound":164,
    "start":0,
    "numFoundExact":true,
    "docs":[{
      "Department":["Marketing"],
      "Gender":["Female"],
      "Ethnicity":["Latino"],
      "Age":[25],
      "Country":["Brazil"],
      "City":["Sao Paulo"],
      "id":["e6608945-690f-4fb3-847d-a72aae9565e2"],
      "Employee_ID":["E02180"],
      "Full_Name":["Aubrey Ruiz"],
      "Job_Title":["Analyst II"],
      "Business_Unit":["Specialty Products"],
      "Hire_Date":["1/11/2021"],
      "Annual_Salary":["$70,126 "],
      "Bonus__":["0%"],
      "_version_":1811817833944842262,
      "_root_":["e6608945-690f-4fb3-847d-a72aae9565e2"]
    }
  ]
}
```