

# 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
```

Apache Solr

Dashboard

Logging

Cloud

Core Admin

Java Properties

Thread Dump

film\_shard1\_re...

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema Browser

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

010

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)

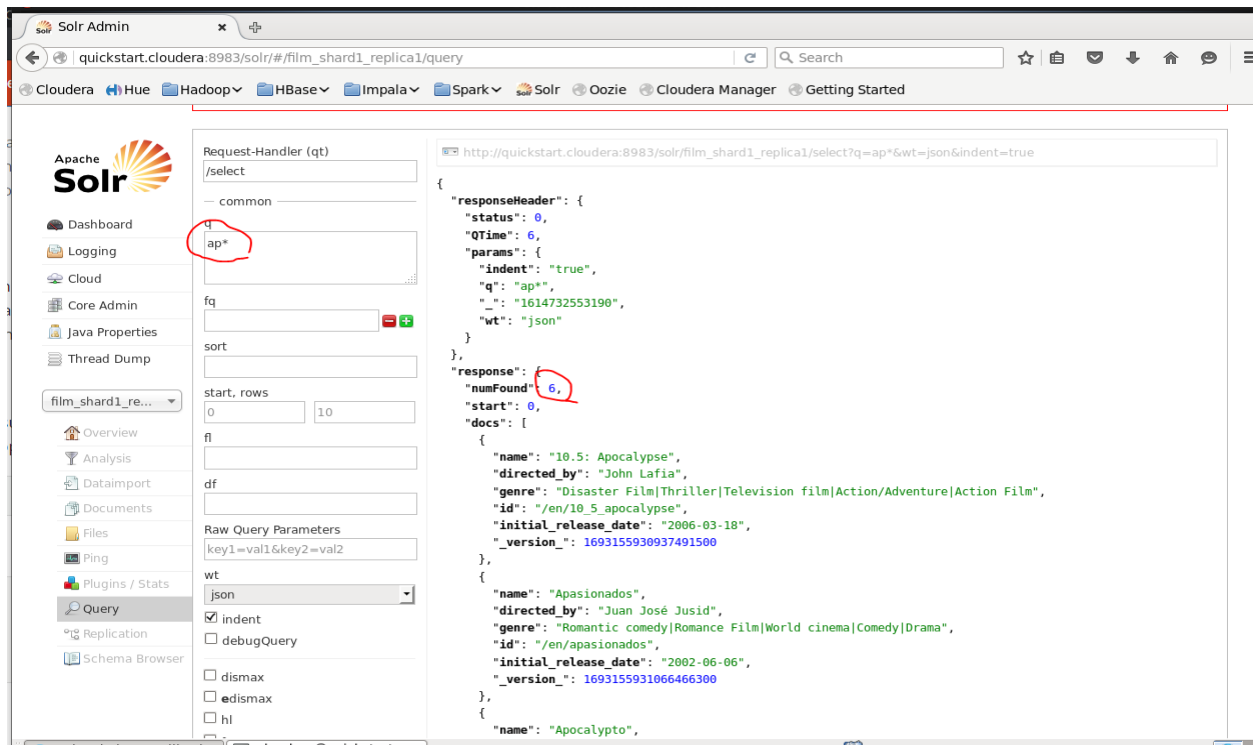
The screenshot displays the Apache Solr Admin interface. On the left, the 'Query' tab is selected. The 'Request-Handler (qt)' section shows the path '/select'. The 'q' field contains the value '300', which is circled in red. The 'wt' field is set to 'json'. The 'response' field is set to 'numFound'. The 'start' field is set to '0' and the 'rows' field is set to '10'. The 'indent' checkbox is checked. The 'Raw Query Parameters' section shows 'key1=val1&key2=val2'. The 'wt' dropdown is set to 'json'. The 'indent' checkbox is checked. The 'dehighlight' checkbox is unchecked. The right pane shows the JSON response for the query. The 'numFound' field is circled in red and has the value '1'. The 'docs' array contains one document with the 'name' field circled in red and the value '300'. The document also includes fields for 'directed\_by', 'genre', 'id', 'initial\_release\_date', and '\_version\_'.

```
http://quickstart.cloudera:8983/solr/film_shard1_replica1/select?q=300&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\*)



The screenshot shows the Apache Solr Admin web interface. On the left, the 'Query' tab is selected in the sidebar. The main panel displays the 'Request-Handler (qt)' configuration for '/select'. The 'q' field contains the query 'ap\*', which is circled in red. The 'wt' field is set to 'json'. The 'numFound' field in the response is also circled in red, showing the value '6'. The response body is a JSON array of document objects, each representing a movie. The first document is for '10.5: Apocalypse' and the second is for 'Apasionados'.

```
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": "Apocalpyto",

```

=> 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)

The screenshot shows the Apache Solr Admin UI. On the left is a sidebar with navigation links: Dashboard, Logging, Cloud, Core Admin, Java Properties, Thread Dump, and a dropdown menu for 'film\_shard1\_re...'. The 'Query' link is selected. The main panel is titled 'Request-Handler (qt)' and shows a search query: `/select`. Below this, the query is defined as `q=name:"24 People"~6`, which is circled in red. The 'fq' field is empty, 'sort' is empty, and 'start, rows' is set to 0 to 10. The 'wt' is set to 'json'. The 'Raw Query Parameters' section shows `key1=val1&key2=val2`. The 'debugQuery' checkbox is checked. On the right, the JSON response is displayed. The `numFound` field is 1, circled in red. The `docs` array contains one document with the following fields: `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), and `_version_` (1693155930940637200).

The screenshot shows the Apache Solr Admin UI with a red banner at the top indicating a 'SolrCore Initialization Failures'. The message states: `logNew2_shard1_replica1: org.apache.solr.common.cloud.ZooKeeperException:org.apache.solr.common.cloud.ZooKeeperException: Specified config does not exist in ZooKeeper:logs_config`. Below the message, it says 'Please check your logs for more information'. The sidebar and main panel are the same as in the previous screenshot, but the query is now `q=name:"24 People"~6`, which is circled in red. The 'wt' is set to 'json'. The 'Raw Query Parameters' section shows `key1=val1&key2=val2`. The 'debugQuery' checkbox is checked. On the right, the JSON response is displayed. The `numFound` field is 1, circled in red. The `docs` array contains one document with the following fields: `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), and `_version_` (1693155930940637200).

=>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])

The image displays two screenshots of the Apache Solr Admin UI, illustrating a range search query and its results.

**Top Screenshot:** The "Request-Handler (qt)" section shows the query `/select?q=initial_release_date:[2010 TO 2015]`. The "Raw Query Parameters" section shows `key1=val1&key2=val2`. The "wt" dropdown is set to "json". The "indent" checkbox is checked. The "response" section shows the JSON output, including the number of documents found (4) and the details of the four documents.

**Bottom Screenshot:** The "Request-Handler (qt)" section shows the same query. The "wt" dropdown is set to "csv". The "Raw Query Parameters" section shows `key1=val1&key2=val2`. The "response" section shows the CSV output, listing the details of the four documents in a tabular format.

=>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)

The screenshot displays a Solr query interface on the left and its JSON response on the right. In the interface, the query field 'q' contains 'apo~1', which is circled in red. The 'wt' dropdown is set to 'json'. The response on the right is a JSON object with a 'responseHeader' and a 'response' section. The 'numFound' value in the response is 1, also circled in red. The response details a document with the name 'Oram Po'.

```
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



Dashboard

Logging

Cloud

Core Admin

Java Properties

Thread Dump

newcollection\_...

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema Browser

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

Submit Document

Status: **success**

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

## 1)Query1(inStock:false)

The screenshot displays the Apache Solr Admin interface. On the left is a sidebar with navigation links: Dashboard, Logging, Cloud, Core Admin, Java Properties, Thread Dump, and a dropdown menu for 'newcollection\_...'. Below these are links for Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query (highlighted), Replication, and Schema Browser.

The main panel is titled 'Request-Handler (qt)' and shows the configuration for the '/select' handler. The 'q' field is set to 'inStock:false', which is circled in red. Other fields include 'fq', 'sort', 'start, rows' (set to 0 and 10), 'fl', 'df', 'Raw Query Parameters' (key1=val1&key2=val2), 'wt' (set to 'json'), and checkboxes for 'indent' (checked), 'debugQuery', 'dismax', 'edismax', 'hl', and 'facet'.

At the top right, the URL is shown: `http://quickstart.cloudera:8983/solr/newcollection_shard1_replica1/select?q=inStock%3Afalse&wt=json&indent=`


The JSON response is displayed on the right, showing a list of documents. The first document is highlighted with a red line under the 'start' field in the response object.

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "inStock:false",
      "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\*)



Dashboard

Logging

Cloud

Core Admin

Java Properties

Thread Dump

newcollection\_...

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema Browser

Request-Handler (qt)

/select

common

q

inStock:true AND author:George\*

fq

sort

start, rows

010

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

Please check your logs for more information

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])

The screenshot displays the Apache Solr Admin interface. On the left is a sidebar with navigation links: Dashboard, Logging, Cloud, Core Admin, Java Properties, Thread Dump, newcollection\_..., Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query (selected), Replication, and Schema Browser. The main panel is titled 'Request-Handler (qt)' and contains several input fields: a text box with '/select', a dropdown menu set to 'common', a text box with 'q: price:[5 TO 7]' (circled in red), a text box for 'fq', a text box for 'sort', a 'start, rows' section with '0' and '10' in input boxes, a text box for 'fl', a text box for 'df', a 'Raw Query Parameters' section with 'key1=val1&key2=val2', a 'wt' dropdown set to 'json', and checkboxes for 'indent' (checked), 'debugQuery', 'dismax', 'edismax', 'hl', and 'facet'. On the right, a URL bar shows the request: 'http://quickstart.cloudera:8983/solr/newcollection\_shard1\_replica1/select?q=price%3A%5B5+TO+7%5D&wt=json&indent=t'. Below the URL, the JSON response is displayed, showing a 'responseHeader' with status 0, QTime 6, and params including indent: true, q: 'price:[5 TO 7]', \_: '1614754874037', and wt: 'json'. The 'response' object contains 'numFound': 5, 'start': 0, and 'docs' with two document entries. The first document has 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', and \_version\_ 1693177759400984600. The second document has id '0812550706' and 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\*)

The screenshot displays the Apache Solr Admin interface. On the left is a sidebar with navigation links: Dashboard, Logging, Cloud, Core Admin, Java Properties, Thread Dump, and a dropdown menu for 'newcollection\_...'. Below these are links for Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query (highlighted), Replication, and Schema Browser.

The main panel is divided into two sections. The top section, 'Request-Handler (qt)', shows the path '/select'. Below it, the 'q' field contains the query 'price:[5 TO 7] OR id:08\*', which is circled in red. Other fields like 'fq', 'sort', 'start, rows' (set to 0 to 10), 'fl', 'df', and 'Raw Query Parameters' (set to 'key1=val1&key2=val2') are visible. The 'wt' dropdown is set to 'json', and the 'indent' checkbox is checked.

The bottom section displays the JSON response from the Solr server. The URL in the address bar is 'http://quickstart.cloudera:8983/solr/newcollection\_shard1\_replica1/select?q=price%3A%5B5%20TO%207%5D+OR+id%3A08\*&wt=json'. The JSON output includes a 'responseHeader' with status 0, QTime 0, and parameters for the query and weight. The 'response' section shows 'numFound': 5, 'start': 0, and a list of documents. The first document is for 'The Black Company' with a price of 6.99 and ID 0812521390. The second document is for 'The Chronicles of The Black Company' with a price of 1693177759400984600 and ID 0812550706.

```
{
  "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)

The screenshot displays the Elasticsearch Quickstart web interface. On the left, the 'Request-Handler (qt)' panel shows a query `series_t:The* AND sequence_i:1` entered in the 'q' field, which is circled in red. Below the query field are options for 'fq', 'sort', 'start, rows' (set to 0 to 10), 'fl', 'df', 'Raw Query Parameters' (set to `key1=val1&key2=val2`), 'wt' (set to 'json'), and various checkboxes for 'indent', 'debugQuery', 'dismax', 'edismax', 'hl', 'facet', 'spatial', and 'snellcheck'. On the right, the JSON response is shown. The 'responseHeader' section includes 'status': 0, 'QTime': 0, 'params': {'indent': 'true', 'q': 'series\_t:The\* AND sequence\_i:1', '\_': '1614755236159', 'wt': 'json'}, and 'response' section shows 'numFound': 3 (underlined in red), 'start': 0, and a list of documents. The first document is for 'The Black Company' with 'sequence\_i': 1. The second document is partially visible.

```
{
  "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_": 1693177759408984600
      },
      {
        "id": "0380014300",
        "cat": [
          "book"
        ]
      }
    ]
  }
}
```

=> Only displays the data that series\_t field value includes The And sequence\_i's field value is 1.

---