Assignment-2

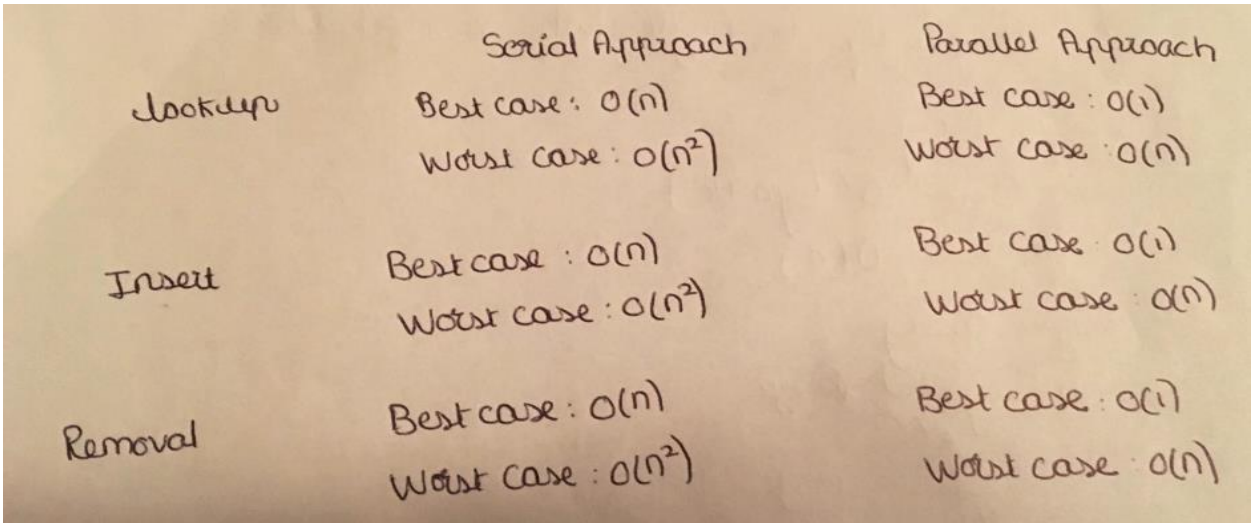
Team:4

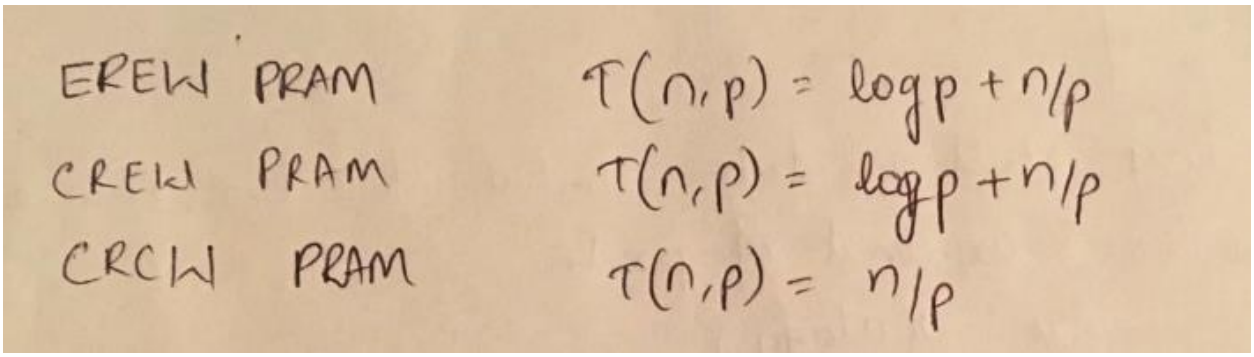
Bhulakshmi Makkena(12)

Naga Mounica(5)

Harshi Priya Yarragonda(25)

1) **linear hashing**





Bloom Filter:

Probability of getting false positive

Pr is $p' = (1-1/m)^{kn} \approx e^{-kn/m} \equiv p$

If p is fraction of 0 bits in the filter then false positive probability is $(1-p)^k \approx (1-p')^k \approx (1-p)^k = (1-e^{-k/c})^k$

Approximations valid as p is concentrated around $E[p]$

Martingale argument suffices.

find optimal at $k = (\ln 2)m/n$ by calculus.

So optimal fpp is about $(0.6185)^{m/n}$

Parallel Searching Techniques

- Range searching:

It generally be single shot vs repetitive

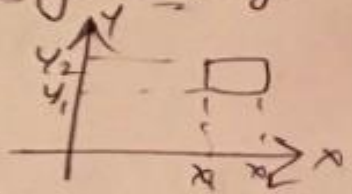
1D Range Query:

Time: $O(\log n + k)$

Space: $O(n)$

2D Range Query:

Orthogonal Range:



Range Trees: Multilevel data structure Main tree is T on the x-coordinate of points.

Size: $O(n \log n)$

Rectangle Data Structure: R-Tree

- Rectangle in leaves
- Internal nodes are minimum bounding rect.

2) The dataset we used contains the information about student like studentID, Fee paid etc.

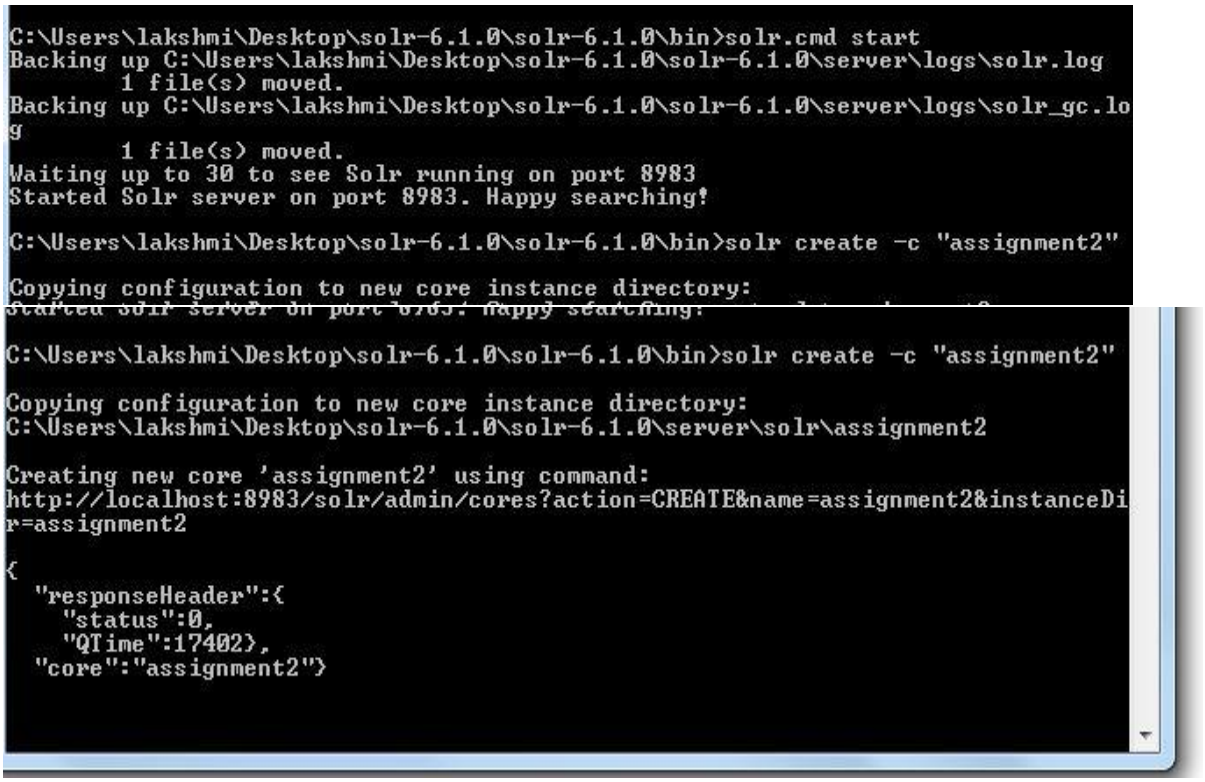
	A	B	C	D	E	F	G
1	StudentID	statecode	county	Fee Paid	Fee Rema	fl_site_lin	fr_site_lir
2	119736	FL	CLAY COU	498960	498960	498960	498960
3	448094	FL	CLAY COU	1322376	1322376	1322376	1322376
4	206893	FL	CLAY COU	190724.4	190724.4	190724.4	190724.4
5	333743	FL	CLAY COU	0	79520.76	0	0
6	172534	FL	CLAY COU	0	254281.5	0	254281.5
7	785275	FL	CLAY COU	0	515035.6	0	0
8	995932	FL	CLAY COU	0	19260000	0	0
9	223488	FL	CLAY COU	328500	328500	328500	328500
10	433512	FL	CLAY COU	315000	315000	315000	315000
11	142071	FL	CLAY COU	705600	705600	705600	705600
12	253816	FL	CLAY COU	831498.3	831498.3	831498.3	831498.3

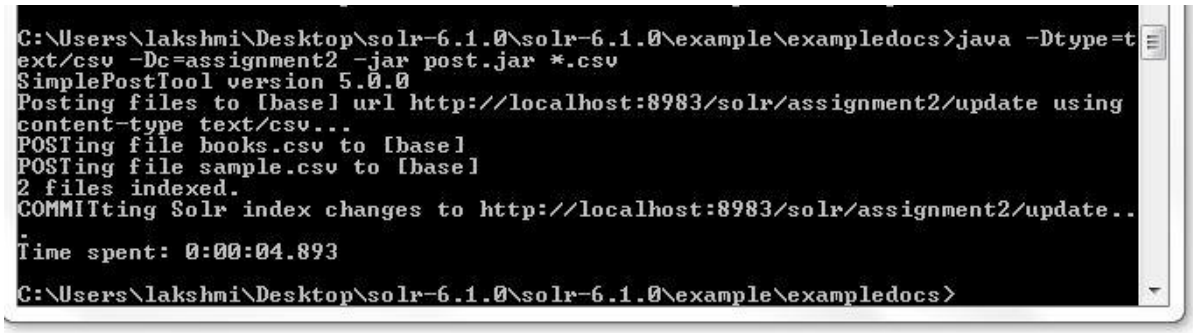
Searching using parallel indexing. Designed a parallel indexing and searching technique and implement it in spark. Submitted the code using github.

3) **Steps to index and querying using Apache solr**

Staring of solr using start command
Using solr.cmd start
Then create an collection using create command
Open http://localhost:8983/solr/admin
Place the dataset sample.csv in example/exampledocs
And run the command to index
The data set we selected is book store which has following columns.

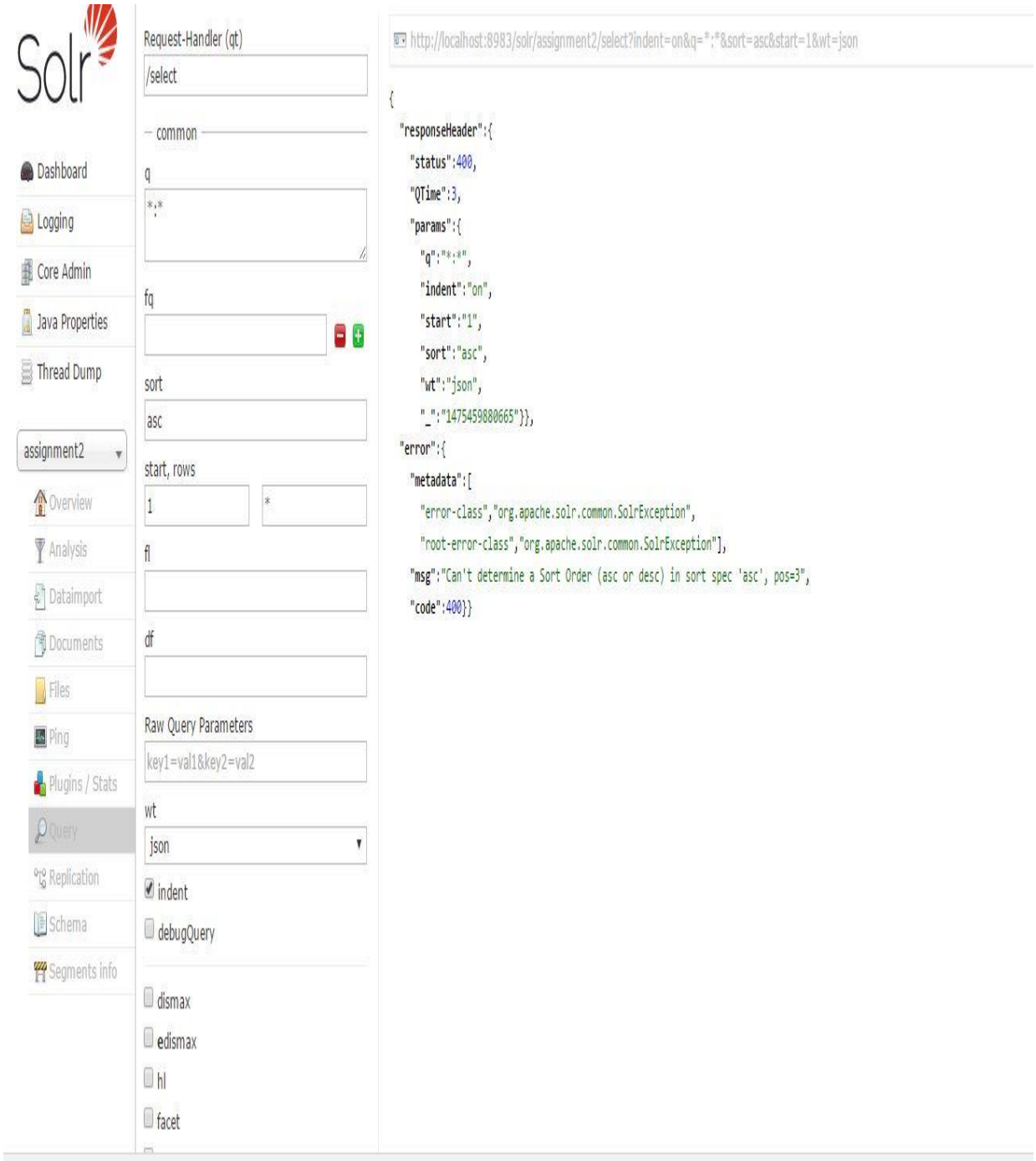
	A	B	C	D	E	F	G	H	I
1	id	cat	name	price	inStock	author	series_t	sequence	genre_s
2	5.54E+08	book	A Game of Thrones	7.99	TRUE	George R.	A Song of	1	fantasy
3	5.54E+08	book	A Clash of Kings	7.99	TRUE	George R.	A Song of	2	fantasy
4	055357342	book	A Storm of Swords	7.99	TRUE	George R.	A Song of	3	fantasy
5	5.53E+08	book	Foundation	7.99	TRUE	Isaac Asir	Foundatic	1	scifi
6	8.13E+08	book	The Black Company	6.99	FALSE	Glen Cook	The Chron	1	fantasy
7	8.13E+08	book	Ender's Game	6.99	TRUE	Orson Sco	Ender	1	scifi
8	4.41E+08	book	Jhereg	7.95	FALSE	Steven Br	Vlad Talto	1	fantasy
9	3.8E+08	book	Nine Princes In Ambe	6.99	TRUE	Roger Zel	the Chron	1	fantasy
10	8.05E+08	book	The Book of Three	5.99	TRUE	Lloyd Alex	The Chron	1	fantasy
11	080508049	book	The Black Cauldron	5.99	TRUE	Lloyd Alex	The Chron	2	fantasy

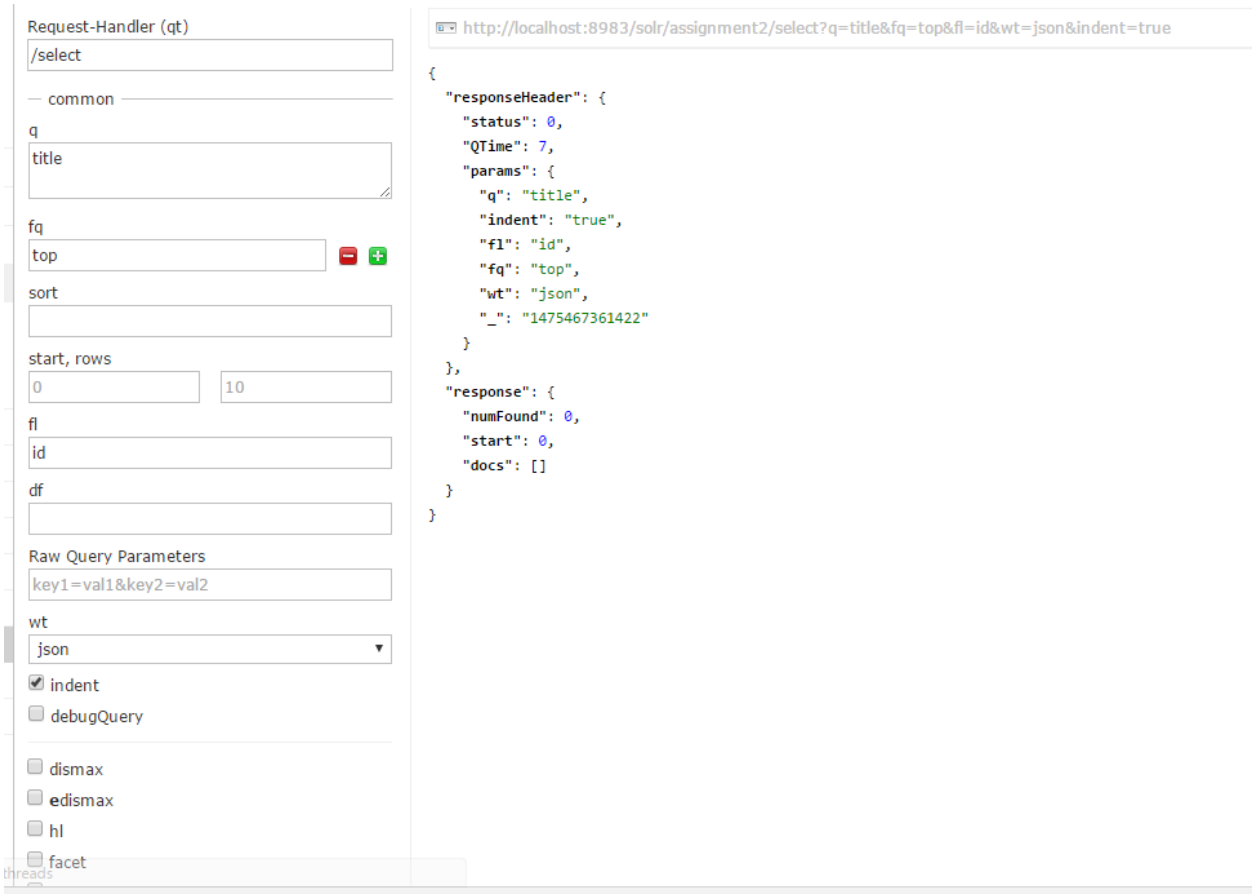

C:\Users\lakshmi\Desktop\solr-6.1.0\solr-6.1.0\bin>solr.cmd start
Backing up C:\Users\lakshmi\Desktop\solr-6.1.0\solr-6.1.0\server\logs\solr.log
1 file(s) moved.
Backing up C:\Users\lakshmi\Desktop\solr-6.1.0\solr-6.1.0\server\logs\solr_gc.log
1 file(s) moved.
Waiting up to 30 to see Solr running on port 8983
Started Solr server on port 8983. Happy searching!
C:\Users\lakshmi\Desktop\solr-6.1.0\solr-6.1.0\bin>solr create -c "assignment2"
Copying configuration to new core instance directory:
C:\Users\lakshmi\Desktop\solr-6.1.0\solr-6.1.0\bin>solr create -c "assignment2"
Copying configuration to new core instance directory:
C:\Users\lakshmi\Desktop\solr-6.1.0\solr-6.1.0\server\solr\assignment2
Creating new core 'assignment2' using command:
http://localhost:8983/solr/admin/cores?action=CREATE&name=assignment2&instanceDir=assignment2
{
"responseHeader":{
"status":0,
"QTime":17402},
"core":"assignment2"}



http://localhost:8983/solr/assignment2/select?q=title&wt=json&indent=true


```
{
  "responseHeader": {
    "status": 0,
    "QTime": 3,
    "params": {
      "q": "title",
      "indent": "true",
      "wt": "json",
      "_": "1475463274418"
    }
  },
  "response": {
    "numFound": 0,
    "start": 0,
    "docs": []
  }
}
```


Solr
Dashboard
Logging
Core Admin
Java Properties
Thread Dump
assignment2
Overview
Analysis
Dataimport
Documents
Files
Ping
Plugins / Stats
Query
Replication
Schema
Segments info
Request-Handler (qt)
/select
common
q
:
fq
sort
asc
start, rows
1
*
fl
df
Raw Query Parameters
key1=val1&key2=val2
wt
json
indent
debugQuery
dismax
edismax
hl
facet
http://localhost:8983/solr/assignment2/select?indent=on&q=*:*&sort=asc&start=1&wt=json
{
"responseHeader":{
"status":400,
"QTime":3,
"params":{
"q":"*:*",
"indent":"on",
"start":"1",
"sort":"asc",
"wt":"json",
"_":"1475459880665"}},
"error":{
"metadata":[
"error-class","org.apache.solr.common.SolrException",
"root-error-class","org.apache.solr.common.SolrException"],
"msg":"Can't determine a Sort Order (asc or desc) in sort spec 'asc', pos=3",
"code":400}}


Request-Handler (qt)
/select
common
q
title
fq
top
sort
start, rows
0
10
fl
id
df
Raw Query Parameters
key1=val1&key2=val2
wt
json
indent
debugQuery
dismax
edismax
hl
facet
http://localhost:8983/solr/assignment2/select?q=title&fq=top&fl=id&wt=json&indent=true
{
"responseHeader": {
"status": 0,
"QTime": 7,
"params": {
"q": "title",
"indent": "true",
"fl": "id",
"fq": "top",
"wt": "json",
"_": "1475467361422"
}
},
"response": {
"numFound": 0,
"start": 0,
"docs": []
}
}

Request-Handler (qt)

/select

common

q

title

fq

top

sort

start, rows

0 10

fl

id

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

indent

debugQuery

dismax

edismax

hl

facet

http://localhost:8983/solr/assignment2/selec


```
{
  "responseHeader": {
    "status": 0,
    "QTime": 7,
    "params": {
      "q": "title",
      "indent": "true",
      "fl": "id",
      "fq": "top",
      "wt": "json",
      "_": "1475467361422"
    }
  },
  "response": {
    "numFound": 0,
    "start": 0,
    "docs": []
  }
}
```