VirginiaTech
Invent the Future

WIKIPEDIA HADOOP

STEVEN STULGA

SPRING 2016

CS 4624 MULTIMEDIA/HYPERTEXT/INFORMATION ACCESS

PROFESSOR EDWARD FOX

CLIENT: SHIVAM MAHARSHI

BACKGROUND

- Goal: Import the entire English Wikipedia with full revision history into Apache Hadoop and configure it to be searchable from Apache Solr
- A prototype outlining the process
  - importing the data
  - converting the data
  - configuring the system

IMPLEMENTATION

- Creating a prototype
  - Downloading small segment of dataset
  - Uncompressing, extracting, converting this data
  - Import it into local HDFS
  - Configure Solr to index data
- Follow same steps for full data on cluster
  - MUCH larger set of data
  - Working on true cluster

DATA COMPARISON

	Prototype Test Data	English Wikipedia Current Version	English Wikipedia with Full Revision History
Size Compressed	365 KB	23.2 GB	107 GB
Size Decompressed	46 MB	51 GB	10 TB
Number of Articles	21	5,137,000	5,137,000

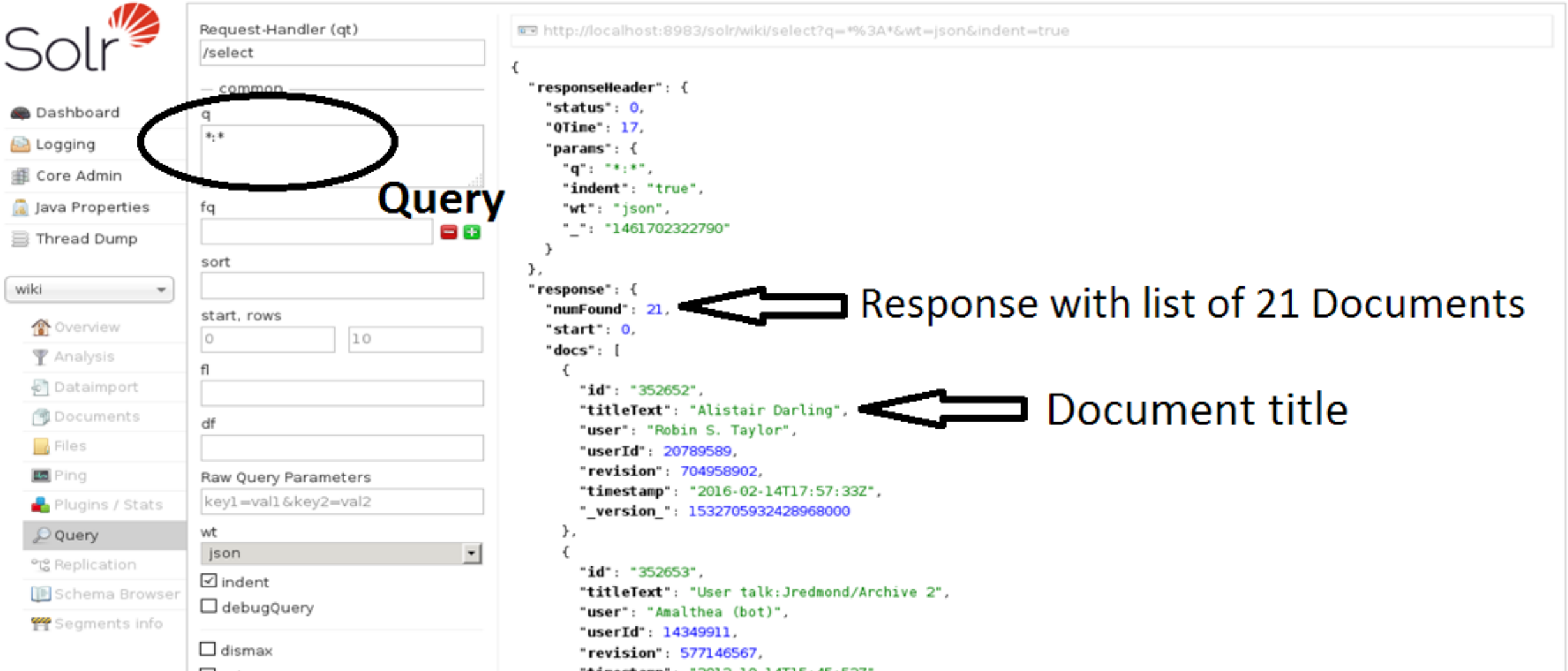
SUCCESSES

```
[cs4624s16_wiki@node1 ~]$ ls -S -lh enwiki/
total 107G
-rw-rw-r-- 1 cs46                                                  5-pages-meta-history27.xml-p042663462p043381320.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  1.2G Mar 15 05:42 enwiki-20160305-pages-meta-history26.xml-p038067203p038736295.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  1.2G Mar 15 10:56 enwiki-20160305-pages-meta-history24.xml-p031451464p031839301.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  1.1G Mar 15 18:08 enwiki-20160305-pages-meta-history17.xml-p013118835p013279304.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki 1005M Mar 15 07:18 enwiki-20160305-pages-meta-history14.xml-p006414990p006679125.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki 1001M Mar 15 13:27 enwiki-20160305-pages-meta-history27.xml-p045050063p046488840.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  946M Mar 15 13:18 enwiki-20160305-pages-meta-history26.xml-p040059433p040860515.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  945M Mar 15 15:21 enwiki-20160305-pages-meta-history27.xml-p046488842p047428985.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  935M Mar 15 17:50 enwiki-20160305-pages-meta-history13.xml-p005040438p005137507.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  890M Mar 15 17:14 enwiki-20160305-pages-meta-history27.xml-p047428986p048617820.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  886M Mar 15 08:15 enwiki-20160305-pages-meta-history23.xml-p027380125p027748345.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  881M Mar 15 15:59 enwiki-20160305-pages-meta-history26.xml-p040860516p041575480.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  832M Mar 15 11:25 enwiki-20160305-pages-meta-history27.xml-p044115376p045050062.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  802M Mar 15 05:17 enwiki-20160305-pages-meta-history19.xml-p016120543p016623894.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  797M Mar 15 11:01 enwiki-20160305-pages-meta-history25.xml-p035147331p035842433.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  792M Mar 15 07:57 enwiki-20160305-pages-meta-history22.xml-p024365252p024822158.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  780M Mar 15 13:28 enwiki-20160305-pages-meta-history23.xml-p028287142p028987754.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  751M Mar 15 15:32 enwiki-20160305-pages-meta-history22.xml-p025684765p026068954.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  751M Mar 15 10:03 enwiki-20160305-pages-meta-history16.xml-p010101153p010386957.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  737M Mar 15 07:40 enwiki-20160305-pages-meta-history16.xml-p009840163p010101151.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  733M Mar 15 20:21 enwiki-20160305-pages-meta-history21.xml-p023538756p023927983.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  728M Mar 15 13:13 enwiki-20160305-pages-meta-history21.xml-p022464639p022860800.7z
-                                                   -20160305-pages-meta-history26.xml-p038736296p039406669.7z
-                                                   -20160305-pages-meta-history24.xml-p033221658p033825989.7z
-rw-rw-r-- 1 cs4624s16_wiki cs4624s16_wiki  710M Mar 15 18:12 enwiki-20160305-pages-meta-history26.xml-p041575481p042289705.7z
```

107GB of compressed XML data

This dataset is ~10TB decompressed

SUCCESSES


Solr
Dashboard
Logging
Core Admin
Java Properties
Thread Dump
wiki
Overview
Analysis
Dataimport
Documents
Files
Ping
Plugins / Stats
Query
Replication
Schema Browser
Segments info
Request-Handler (qt)
/select
common
q
:
Query
fq
sort
start, rows
0
10
fl
df
Raw Query Parameters
key1=val1&key2=val2
wt
json
indent
debugQuery
dismax
http://localhost:8983/solr/wiki/select?q=*%3A*&wt=json&indent=true
{
"responseHeader": {
"status": 0,
"QTime": 17,
"params": {
"q": "*:*",
"indent": "true",
"wt": "json",
"_": "1461702322790"
}
},
"response": {
"numFound": 21,
Response with list of 21 Documents
"start": 0,
"docs": [
{
"id": "352652",
"titleText": "Alistair Darling",
Document title
"user": "Robin S. Taylor",
"userId": 20789589,
"revision": 704958902,
"timestamp": "2016-02-14T17:57:33Z",
"_version_": 1532705932428968000
},
{
"id": "352653",
"titleText": "User talk:Jredmond/Archive 2",
"user": "Amalthea (bot)",
"userId": 14349911,
"revision": 577146567,

LESSONS LEARNED

- Work with data in small pieces
- There isn't always the "how to" guide you need
- Implement full prototype and outline all steps
- Effective learning requires more time and research than you expect
- Communication can be slow and hinder progress

ACKNOWLEDGEMENTS

- Shivam Maharshi, the client, for presenting the project and giving guidance about how to start and tackle challenges
- Sunshin Lee, who provided access to the remote machine and answered questions about execution on the machine
- Dr. Fox, for great feedback and help with the presentation and documentation

THANK YOU!

- Questions?