Apache Solr

DAPI . Information Description, Storage and Retrieval Course
MIEIC, 2020/21 Edition

Sérgio Nunes
DEI, FEUP, U.Porto

Work in progress

Plan for Today

- ➔ Questions?
- ➔ Groups Presentations (~90 min)
- ➔ Break
- ➔ Milestone #2 Overview
- ➔ Solr overview

Milestone #2 — Information Retrieval

Milestone #2

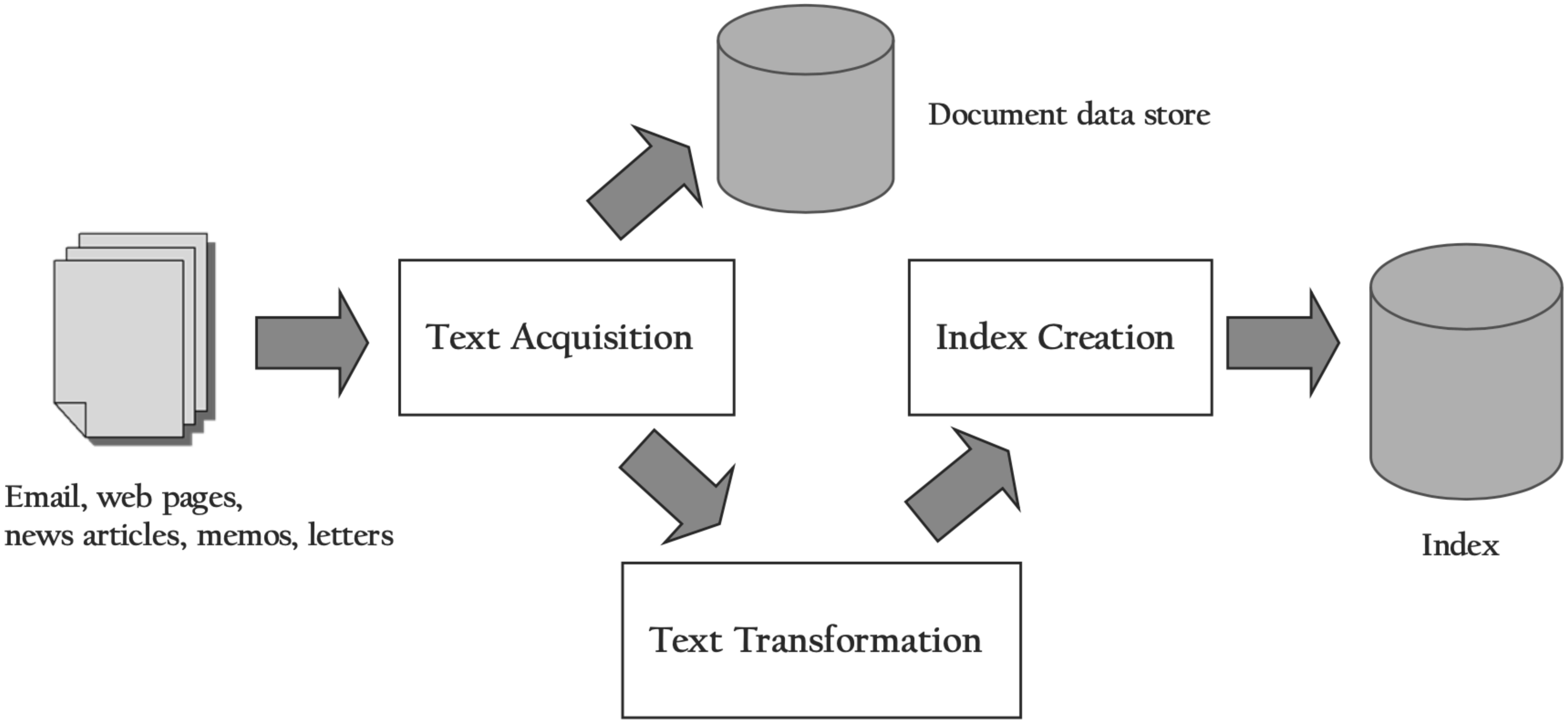
- ➔ Goal: index dataset to support querying using free-text
- ➔ Use open-source tools (i.e. Solr); decide on the document granularity; decide on the search filters.
- ➔ Expected actions:
  - ➔ Choose the information retrieval tool (Solr, Lucene, Terrier, Elasticsearch, …);
  - ➔ Analyze the documents and identify their indexable components;
  - ➔ Identify search parameters that will be offered to the users;
  - ➔ Use the tool API to generate indexes;
  - ➔ Use the tool API to configure the answer to queries;
  - ➔ Demonstrate the indexing and retrieval processes;
  - ➔ Evaluate the results, (ideally) comparing different ranking formulas.
- ➔ More information at https://web.fe.up.pt/~ssn/dokuwiki/teach/dapi/202021/delivery2/index

Search Engine Overview

Architecture of a Search Engine

- ➔ Two primary goals of a search engine:
  - ➔ *effectiveness* (quality) — retrieve the most relevant set of documents;
  - ➔ *efficiency* (speed) — present the results as quickly as possible;
- ➔ Search engines are architected to two support two major functions:
  - ➔ *indexing process* — build the structures to enable search;
  - ➔ *querying process* —use the structures to produce a ranking;

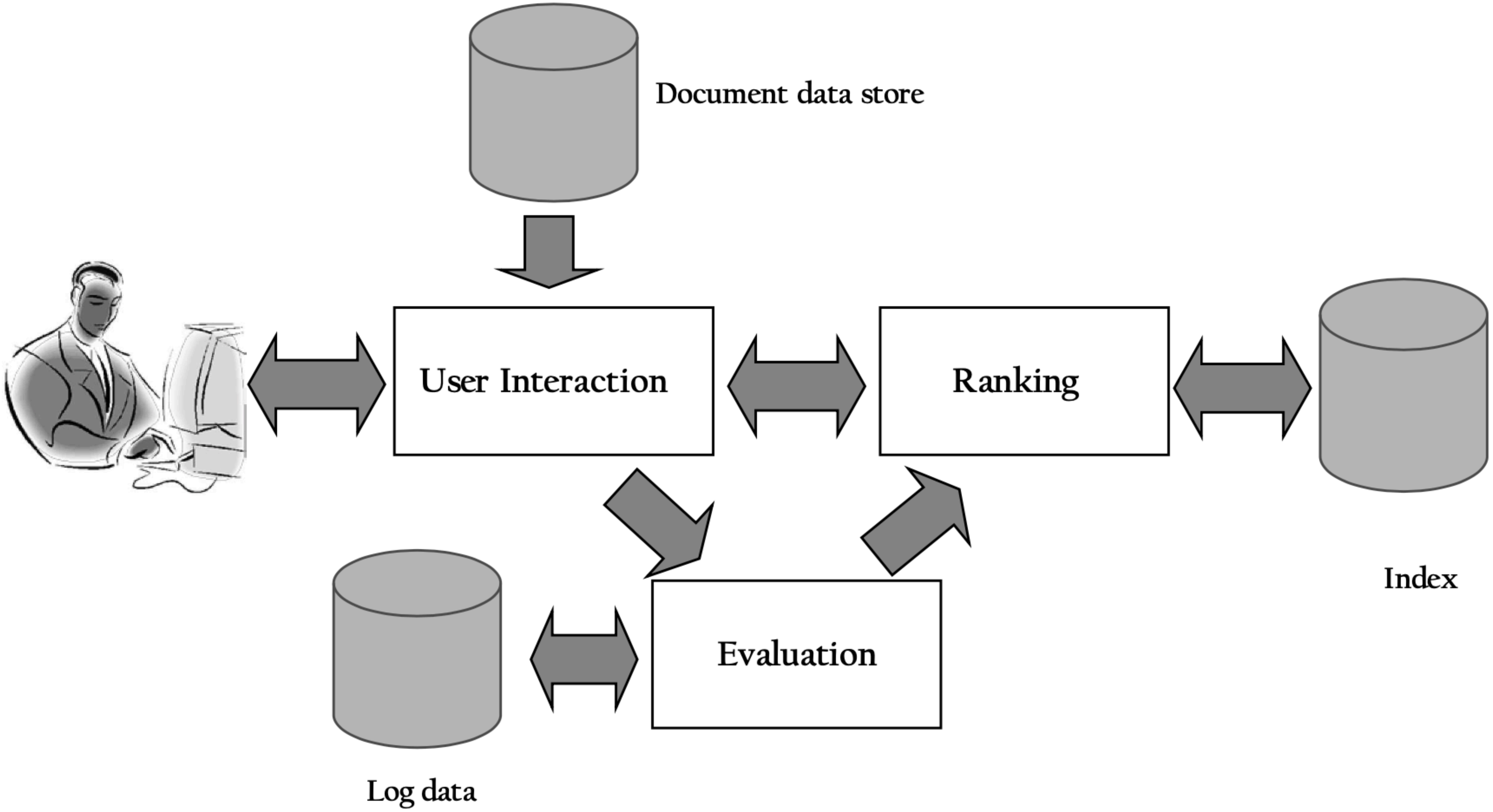
The Indexing Process



Blocks of the Indexing Process

- ➔ Text Acquisition
  - ➔ crawler; conversion; document data store.
- ➔ Text Transformation
  - ➔ parser; stopping; stemming; link extraction; information extraction; classifier.
- ➔ Index Creation
  - ➔ document statistics; weighting; inversion; index distribution;

The Querying Process



Blocks of the Querying Process

- User Interaction
  - query input; query transformation; results output.
- Ranking
  - scoring; performance optimization; distribution.
- Evaluation
  - logging; ranking analysis; performance analysis.

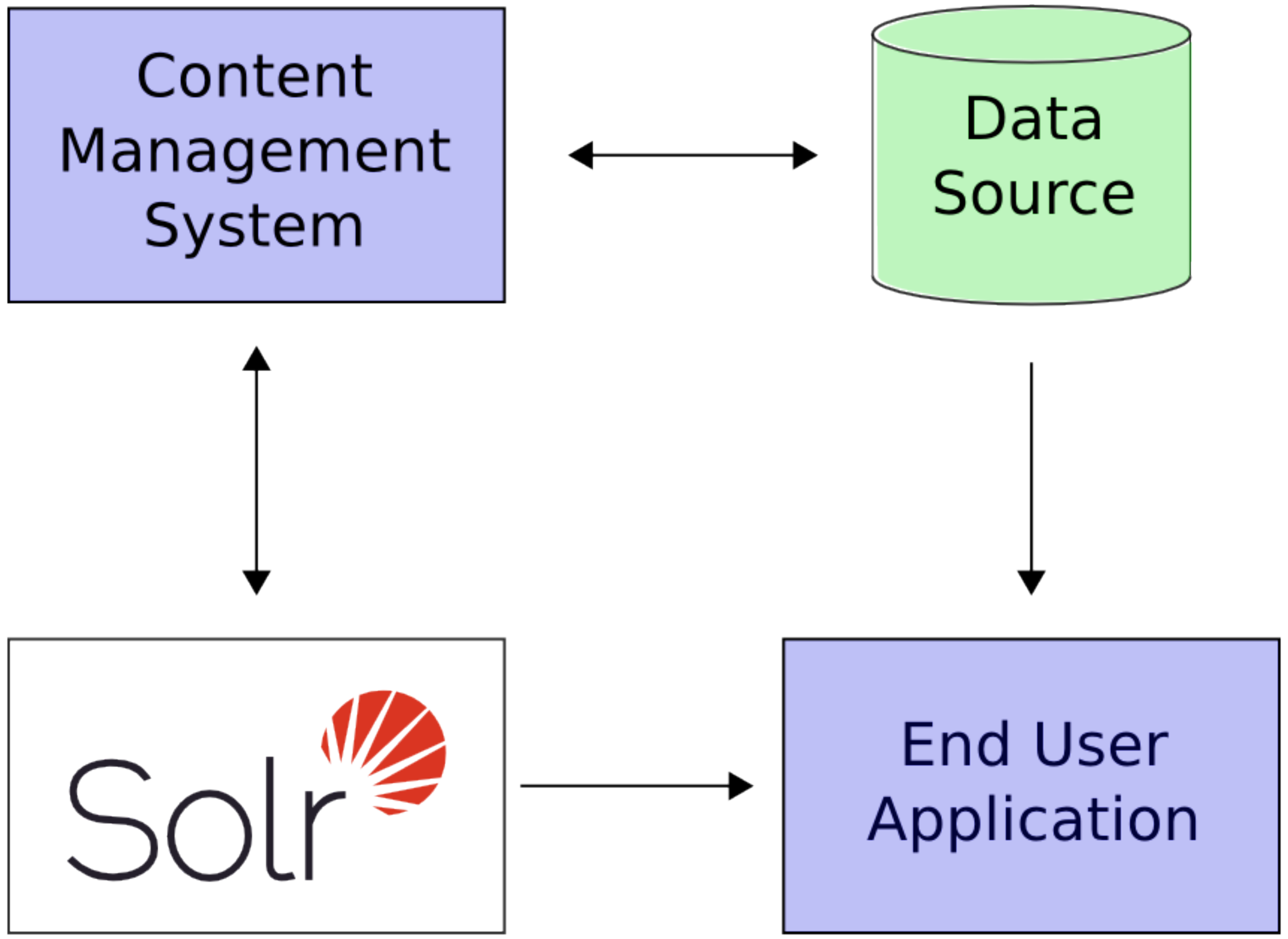
Apache Solr

Apache Solr

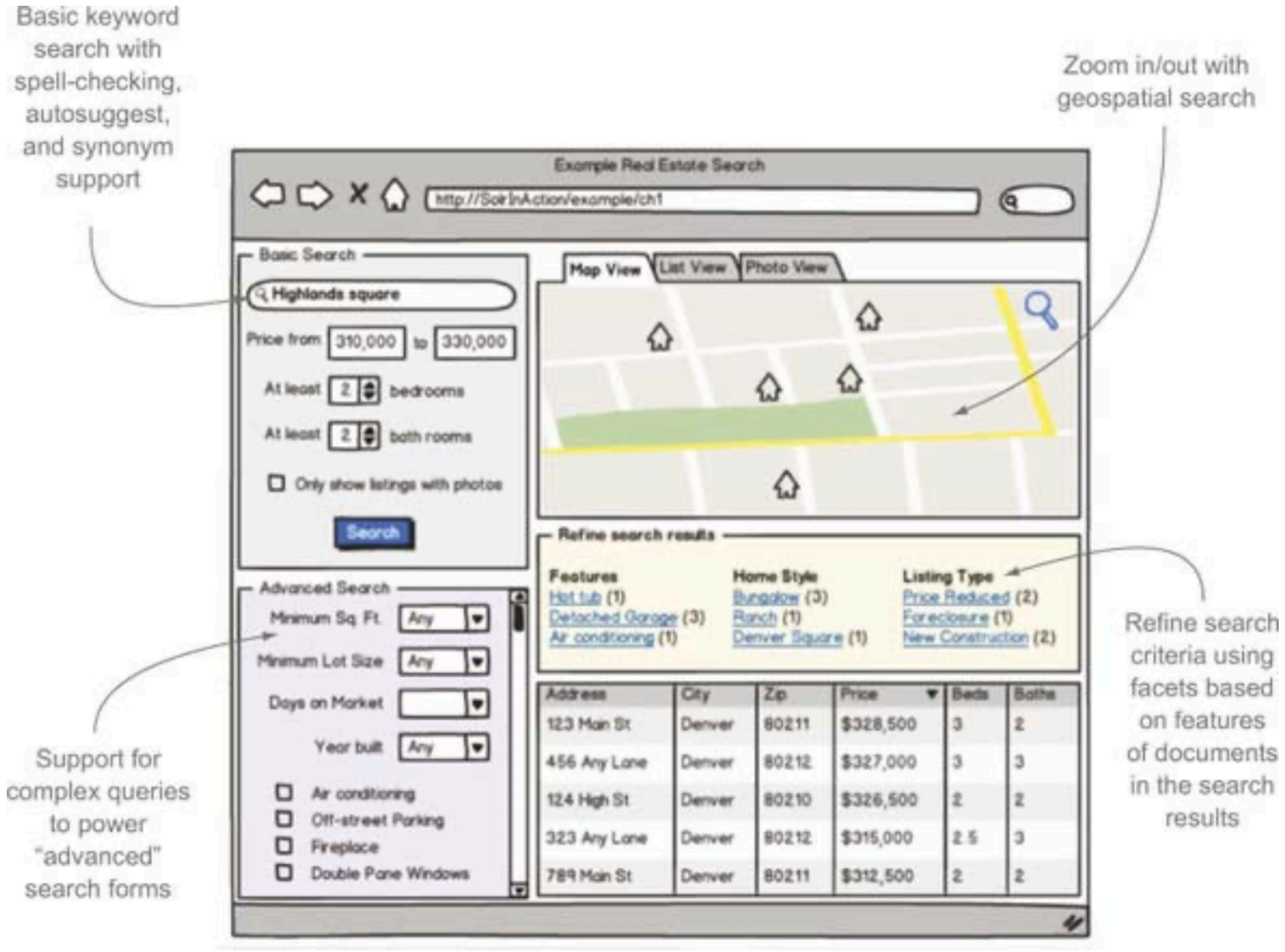
- ➔ Solr is a search server built on top of Apache Lucene, an open source, Java-based, information retrieval library. Standard steps:
  - ➔ Define the schema, to tell Solr about the contents of documents it will be indexing;
  - ➔ Feed Solr documents for which your users will search;
  - ➔ Expose search functionality in your application.
- ➔ Solr offers support for the simplest keyword searching through to complex queries on multiple fields and faceted search results.
- ➔ Because Solr is based on open standards, it is highly extensible. Solr queries are simple HTTP request URLs and the response is a structured document: mainly JSON, but it could also be XML, CSV, or other formats.

From: *Apache Solr Reference Guide*. https://lucene.apache.org/solr/guide

Example Solr Integration



Solr Features



Basic keyword search with spell-checking, autosuggest, and synonym support

Zoom in/out with geospatial search

Refine search criteria using facets based on features of documents in the search results

Support for complex queries to power "advanced" search forms

Highlands square

Price from 310,000 to 330,000

At least 2 bedrooms

At least 2 bath rooms

□ Only show listings with photos

Features
Hot tub (1)
Detached Garage (3)
Air conditioning (1)

Home Style
Bungalow (3)
Ranch (1)
Denver Square (1)

Listing Type
Price Reduced (2)
Foreclosure (1)
New Construction (2)

Minimum Sq. Ft. Any

Minimum Lot Size Any

Days on Market

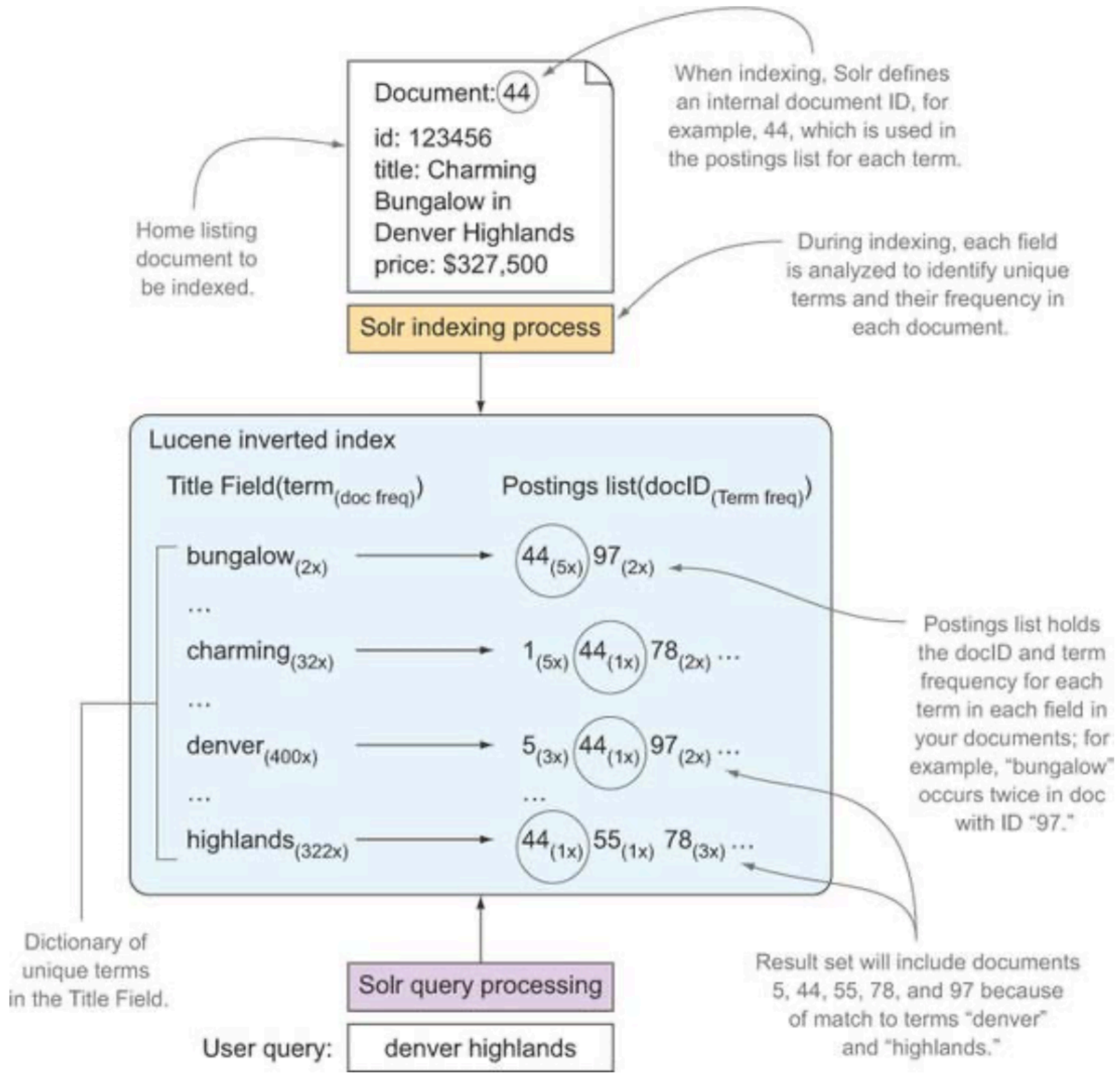
Year built Any

□ Air conditioning
□ Off-street Parking
□ Fireplace
□ Double Pane Windows

Address	City	Zip	Price ▼	Beds	Baths
123 Main St	Denver	80211	$328,500	3	2
456 Any Lane	Denver	80212	$327,000	3	3
124 High St	Denver	80210	$326,500	2	2
323 Any Lane	Denver	80212	$315,000	2.5	3
789 Main St	Denver	80211	$312,500	2	2

From: *Trey Grainger and Timothy Potter. Solr in Action, Manning Publications, 2014.*

The Inverted Index (again)



When indexing, Solr defines an internal document ID, for example, 44, which is used in the postings list for each term.

Home listing document to be indexed.

During indexing, each field is analyzed to identify unique terms and their frequency in each document.

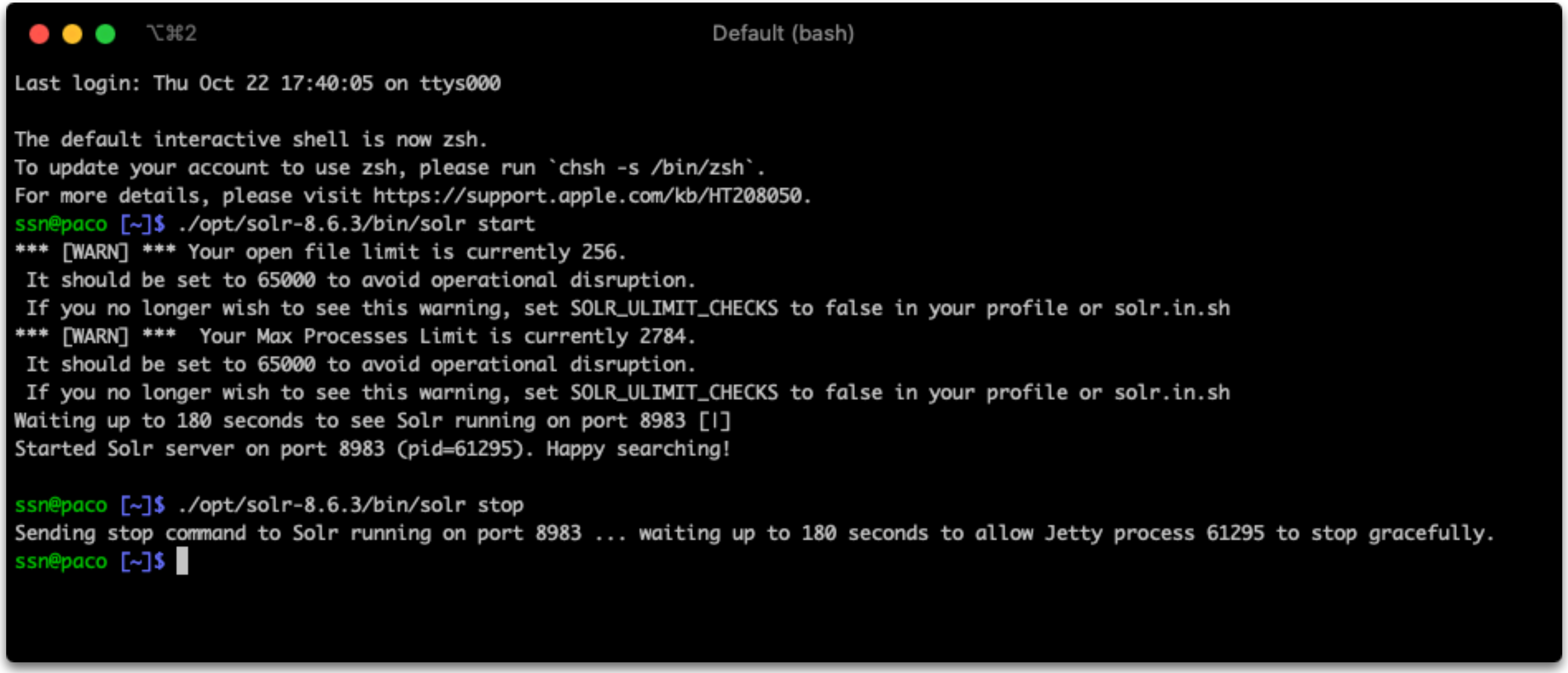
Postings list holds the docID and term frequency for each term in each field in your documents; for example, "bungalow" occurs twice in doc with ID "97."

Dictionary of unique terms in the Title Field.

Result set will include documents 5, 44, 55, 78, and 97 because of match to terms "denver" and "highlands."

From: *Trey Grainger and Timothy Potter. Solr in Action, Manning Publications, 2014.*

Solr Command Line Tool

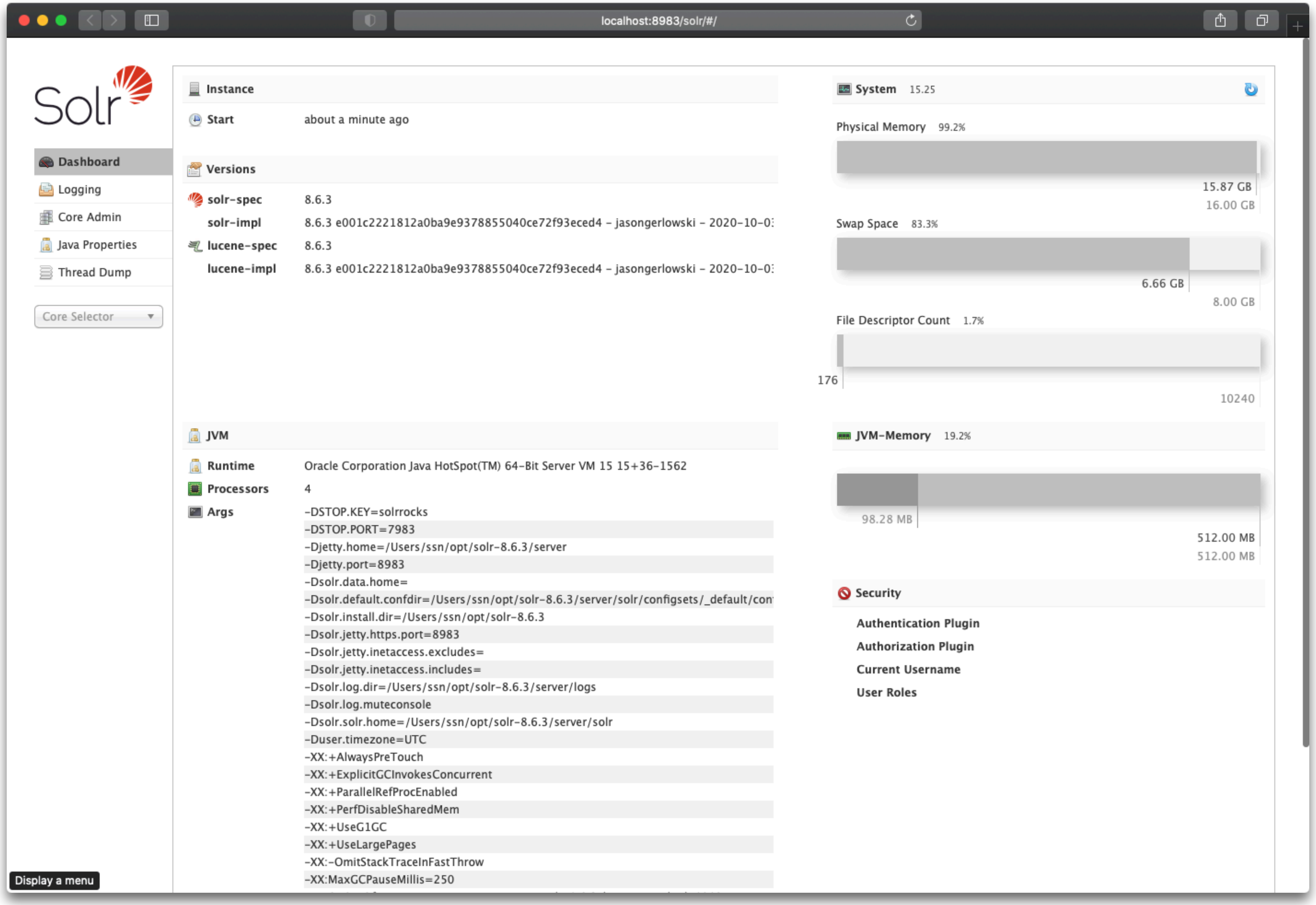


```
Last login: Thu Oct 22 17:40:05 on ttys000

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
ssn@paco [~]$ ./opt/solr-8.6.3/bin/solr start
*** [WARN] *** Your open file limit is currently 256.
 It should be set to 65000 to avoid operational disruption.
 If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in your profile or solr.in.sh
*** [WARN] ***  Your Max Processes Limit is currently 2784.
 It should be set to 65000 to avoid operational disruption.
 If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in your profile or solr.in.sh
Waiting up to 180 seconds to see Solr running on port 8983 [|]
Started Solr server on port 8983 (pid=61295). Happy searching!

ssn@paco [~]$ ./opt/solr-8.6.3/bin/solr stop
Sending stop command to Solr running on port 8983 ... waiting up to 180 seconds to allow Jetty process 61295 to stop gracefully.
ssn@paco [~]$
```

Solr Admin Console



Instance

Start about a minute ago

Versions

solr-spec 8.6.3

solr-impl 8.6.3 e001c2221812a0ba9e9378855040ce72f93eced4 - jasongerlowski - 2020-10-0:

lucene-spec 8.6.3

lucene-impl 8.6.3 e001c2221812a0ba9e9378855040ce72f93eced4 - jasongerlowski - 2020-10-0:

JVM

Runtime Oracle Corporation Java HotSpot(TM) 64-Bit Server VM 15 15+36-1562

Processors 4

Args

```
-DSTOP.KEY=solrrocks
-DSTOP.PORT=7983
-Djetty.home=/Users/ssn/opt/solr-8.6.3/server
-Djetty.port=8983
-Dsolr.data.home=
-Dsolr.default.confdir=/Users/ssn/opt/solr-8.6.3/server/solr/configsets/_default/con
-Dsolr.install.dir=/Users/ssn/opt/solr-8.6.3
-Dsolr.jetty.https.port=8983
-Dsolr.jetty.inetaccess.excludes=
-Dsolr.jetty.inetaccess.includes=
-Dsolr.log.dir=/Users/ssn/opt/solr-8.6.3/server/logs
-Dsolr.log.muteconsole
-Dsolr.solr.home=/Users/ssn/opt/solr-8.6.3/server/solr
-Duser.timezone=UTC
-XX:+AlwaysPreTouch
-XX:+ExplicitGCInvokesConcurrent
-XX:+ParallelRefProcEnabled
-XX:+PerfDisableSharedMem
-XX:+UseG1GC
-XX:+UseLargePages
-XX:-OmitStackTraceInFastThrow
-XX:MaxGCPauseMillis=250
```

System 15.25

Physical Memory 99.2%

15.87 GB
16.00 GB

Swap Space 83.3%

6.66 GB
8.00 GB

File Descriptor Count 1.7%

176
10240

JVM-Memory 19.2%

98.28 MB
512.00 MB
512.00 MB

Security

Authentication Plugin

Authorization Plugin

Current Username

User Roles

Tasks

- ➔ Finish and submit Milestone #1 report.
- ➔ Review goals and organize work for Milestone #2.
- ➔ Experiment with full-text indexing tools.
    - ➔ Apache Solr Tutorial — https://lucene.apache.org/solr/guide/solr-tutorial.html
    - ➔ Experiment with other collections (e.g. project, personal documents, etc).
- ➔ Anticipate indexing and search tasks on the working dataset.

- ➔ **Next week**: finish and submit Milestone #1 report.

References

- ➔ Apache. Solr Tutorial. https://lucene.apache.org/solr/guide/solr-tutorial.html
- ➔ Apache Solr Reference Guide. https://lucene.apache.org/solr/guide/
- ➔ Trey Grainger and Timothy Potter. Solr in Action, Manning Publications, 2014.
- ➔ W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2009. http://ciir.cs.umass.edu/downloads/SEIRiP.pdf