Planet Cassandra Community

DATASTAX

Adventures in Discoverability with C* and Solr

Patricia Gorla, Systems Engineer opensource connections
@pgorla
@o19s

About Me

- Solr
- Cassandra
- Information retrieval

Paul Hostetler - phostetler.com

How Do I Find What I'm Looking For?

Simple

Aristotle's birthplace?

```
select birthPlace
where name = "Aristotle";
```

Coordinates of Stagira?

```
select coord
where name = "Stagira";
```

Complex

All ancient Greek philosophers?

```
create index on tag;

select *
where tag = "Greek philosophy";
```

All cities within 100km of Stagira

```
???
```

How Do I Find What I'm Looking For?

Simple

Aristotle's birthplace?

```
q=Aristotle&fl=birthPlace
```

All ancient Greek philosophers?

```
q=Greek philosophy
```

Coordinates of Stagira?

```
q=Stagira&fl=point
```

All cities within 100km of Stagira

```
q=*:*&fq={!geofilt pt=40.530,
23.752 sfield=point d=100}
```

Approaches to Search

- Google Site Search
- MySQL ‘like’ statements

Seth Casteel - littlefriendsphoto.com

Approaches to Search

- Full-text search
- Ranking (Scoring)
- Tokenization
- Stemming
- Faceting

Approaches to Search

Apache Solr

- Full-text search
- Ranking (Scoring)
- Tokenization
- Stemming
- Faceting

Aol.

THE WHITE HOUSE WASHINGTON

Zappos.com

and many more!

Inverted Index

[1] Pleasure in the job puts perfection in the work.

[2] Education is the best provision for the journey to old age.

[3] If some animals are good at hunting and others are suitable for hunting, then the gods must clearly smile on hunting.

[4] It is the mark of an educated mind to be able to entertain a thought without absorbing it.

Term	Freq	Documents
education	2	[2] [4]
hunting	3	[3]
perfection	1	[1]

Index-side Analysis

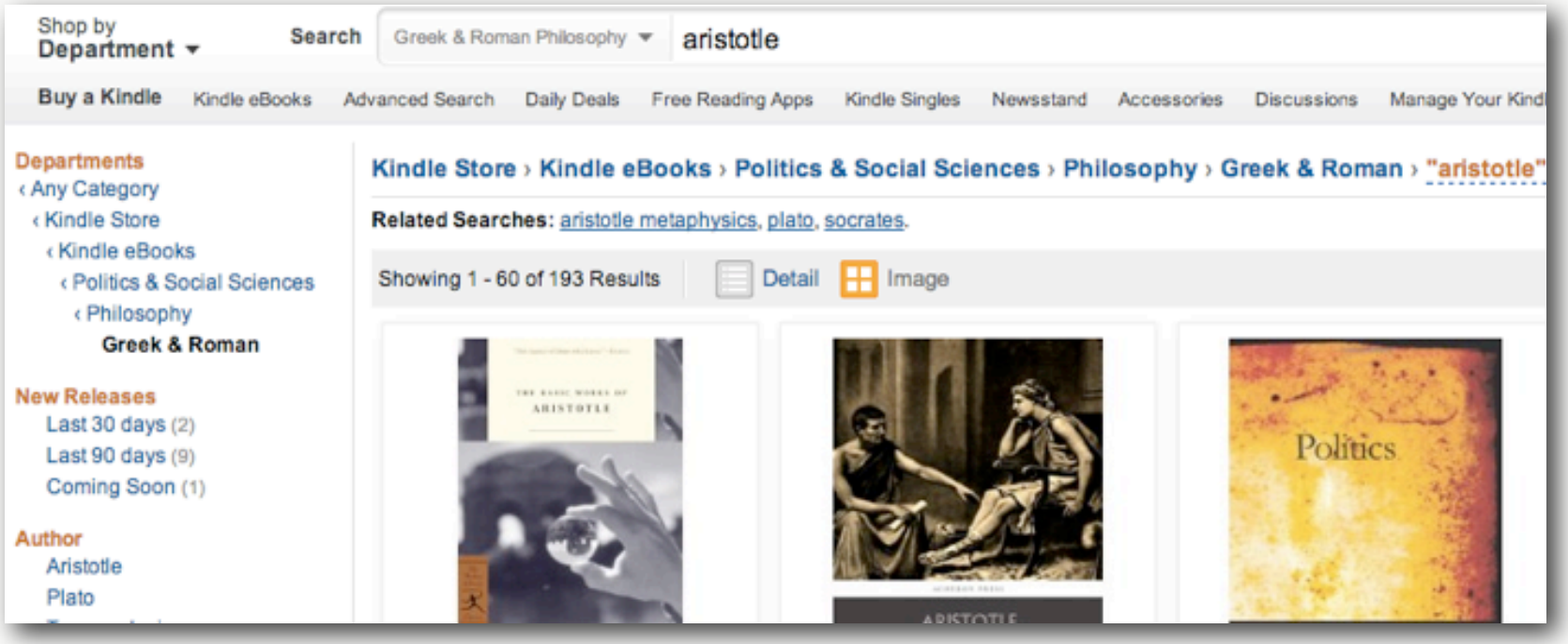
	Hope is a waking dream.
Punctuation	Hope is a waking dream
Stop Words	Hope waking dream
Lowercase	hope waking dream
Stemming	hope wake dream

Query-side Analysis

	Hope is a waking dream.
Punctuation	Hope is a waking dream
Stop Words	Hope waking dream
Lowercase	hope waking dream
Stemming	hope wake dream
Synonyms	hope wake dream desire awake wish

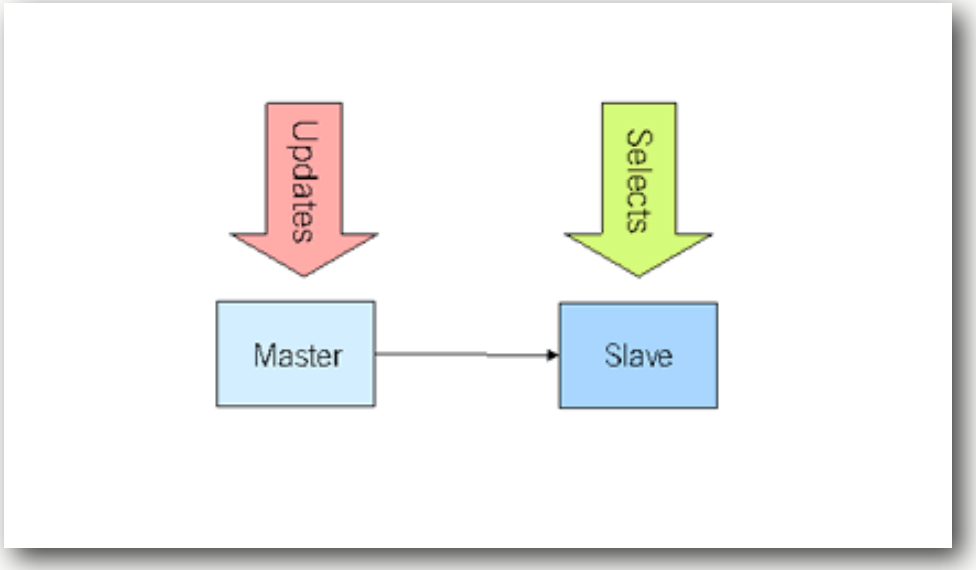
Faceting

```
facet_fields: {
    tags: [hunting: 1, education: 2, work: 2],
    locations: [Stagira: 5, Chalcis: 3]
}
```



Approaches to Distribution

- Pay Google more $$
- MySQL 'like' shards
- Master/Slave replication
- SolrCloud



“Distributed search is hard.”

Solr + Cassandra: Datastax Enterprise

- Full-text search
- Tokenization
- Stemming
- Date ranges
- Aggregation
- High Availability
- Distributed Nature

Examining DBpedia.org Datasets

Person

```
<http://xmlns.com/foaf/0.1/name> "Aristotle"@en .
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> Person .
<http://purl.org/dc/elements/1.1/description> "Greek philosopher"@en .
<http://dbpedia.org/ontology/birthPlace> Stagira
<http://dbpedia.org/ontology/deathPlace> Chalcis .
```

Place

```
<http://dbpedia.org/resource/Stagira> <http://www.opengis.net/gml/_Feature> .
<http://dbpedia.org/resource/Stagira#lat> "40.5916667" .
<http://dbpedia.org/resource/Stagira#long> "23.7947222" .
<http://dbpedia.org/resource/Stagira#point> "40.591667 23.7947222"@en .
```

“Love is a single soul inhabiting two bodies.”

Querying for data

```
curl http://localhost:8983/solr/solr.person/q=Aristotle
```

```
▼ <result name="response" numFound="6" start="0">
    <doc>
      <str name="id"> Aristotle </str>
      <str name="birthPlace"> Stagira (ancient city) </str>
      <str name="deathPlace"> Chalcis </str>
    </doc>
    <doc>
      <str name="id"> Aristotle Onassis </str>
      <date name="birthDate"> 1906-01-15T00:00:00Z </date>
      <str name="birthPlace"> Ottoman Empire </str>
      <date name="deathDate"> 1975-03-15T00:00:00Z </date>
      <str name="deathPlace"> France </str>
    </doc>
    <doc>
```

Filtering by location

```
curl http://localhost:8983/solr/solr.location/select?q=*:*
&spatial=true&fq={!geofilt pt=40.53027,23.7525 sfield=point
d=100}
```

```
<result name="response" numFound="158" start="0">
  <doc>
    <str name="name">Great Lavra</str>
    <str name="point">40.17111111111111,24.38277777777778</str>
  </doc>
  <doc>
    <str name="name">OTE Tower</str>
    <str name="point">40.62620555555556,22.954591666666666</str>
  </doc>
  <doc>
    <str name="name">Lake Volvi</str>
    <str name="point">40.681666666666665,23.467222222222222</str>
  </doc>
  <doc>
```

“There is no great genius without a mixture of madness.”

Schema.xml

Unified Schema

```
<fields>
    <field name="id" type="string" />
    <field name="name" type="text" />
    <dynamicField name="*Date" type="date" />
    <dynamicField name="*Place" type="text" />
    <dynamicField name="*Point" type="location" />
    <dynamicField name="*_tag"  type="text" />
</fields>
```

Upload to DSE, Create Core

```
curl http://localhost:8983/solr/resource/solr.location/schema.xml \
     --data-binary @solr/location_schema.xml \
     -H 'Content-type:text/xml; charset=utf-8 '


curl http://localhost:8983/solr/resource/solr.location/solrconfig.xml \
     --data-binary @solr/location_solrconfig.xml \
     -H 'Content-type:text/xml; charset=utf-8 '

curl http://localhost:8983/solr/admin/cores?action=CREATE&name=solr.location
```

Cassandra Schema

Location Schema

```
cqlsh:solr> DESC COLUMNFAMILY location;

CREATE TABLE location (
  id text PRIMARY KEY,
  "_docBoost" text,
  "_dynFld" text,
  location text,
  name text,
  solr_query text,
  tags text
) WITH COMPACT STORAGE AND
  bloom_filter_fp_chance=0.010000 AND
  caching='KEYS_ONLY' AND
  comment='' AND
  dclocal_read_repair_chance=0.000000 AND
  gc_grace_seconds=864000 AND
  read_repair_chance=0.100000 AND
  replicate_on_write='true' AND
  populate_io_cache_on_flush='false' AND
  compaction={'class':
'SizeTieredCompactionStrategy'} AND
  compression={'sstable_compression':
'SnappyCompressor'};
```

```
CREATE INDEX
solr_location__docBoost_index ON location
("_docBoost");

...

CREATE INDEX
solr_location_solr_query_index ON
location (solr_query);
```

What Changes

Solr

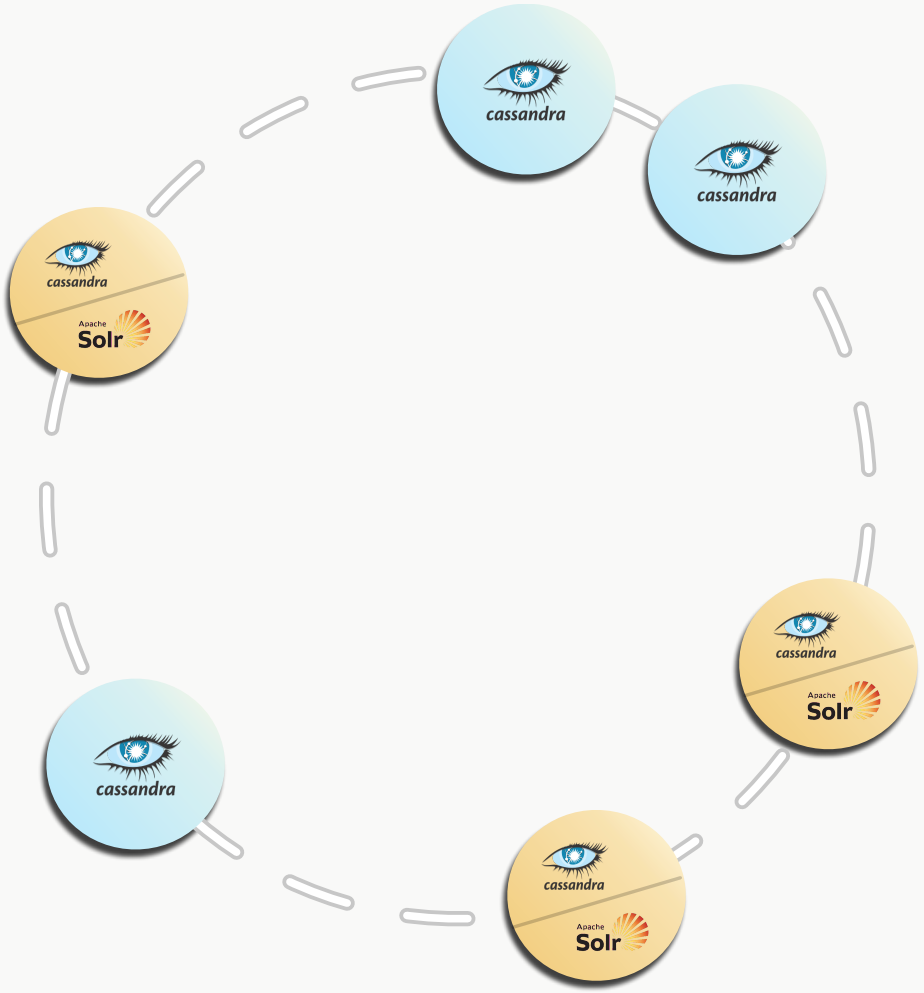
- No multiValued fields
- No JOIN*

Cassandra

- No composite columns
- No counter columns

Bringing it All Together

- Fault tolerant, available search



“Thank you.”

What we discussed today…

- All information on Github, including slides
- http://github.com/pgorla/million-books

THANK YOU

pgorla@opensourceconnections.com
@pgorla
@o19s

All information, including slides, are on `http://github.com/pgorla/million-books`