Lessons Learned with Spark at the US Patent & Trademark Office

Christopher Bradford

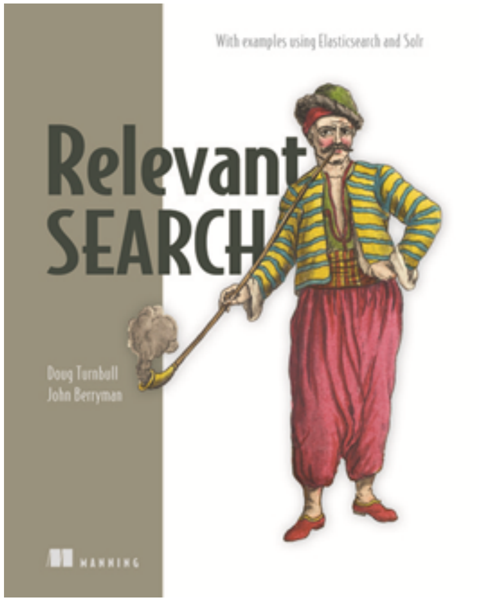
Big Data Architect at OpenSource Connections

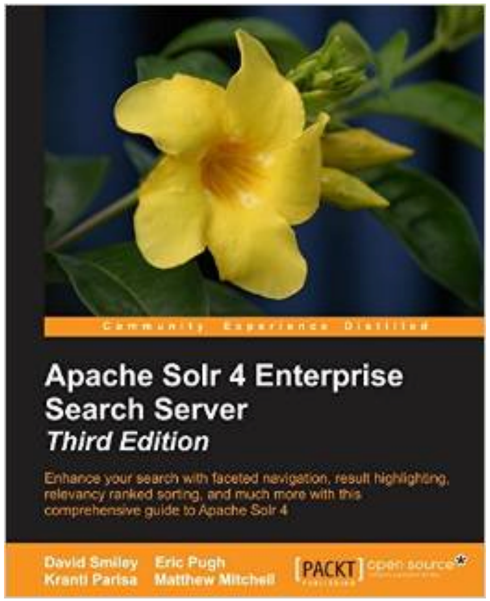
Christopher Bradford

Twitter: @bradfordcp

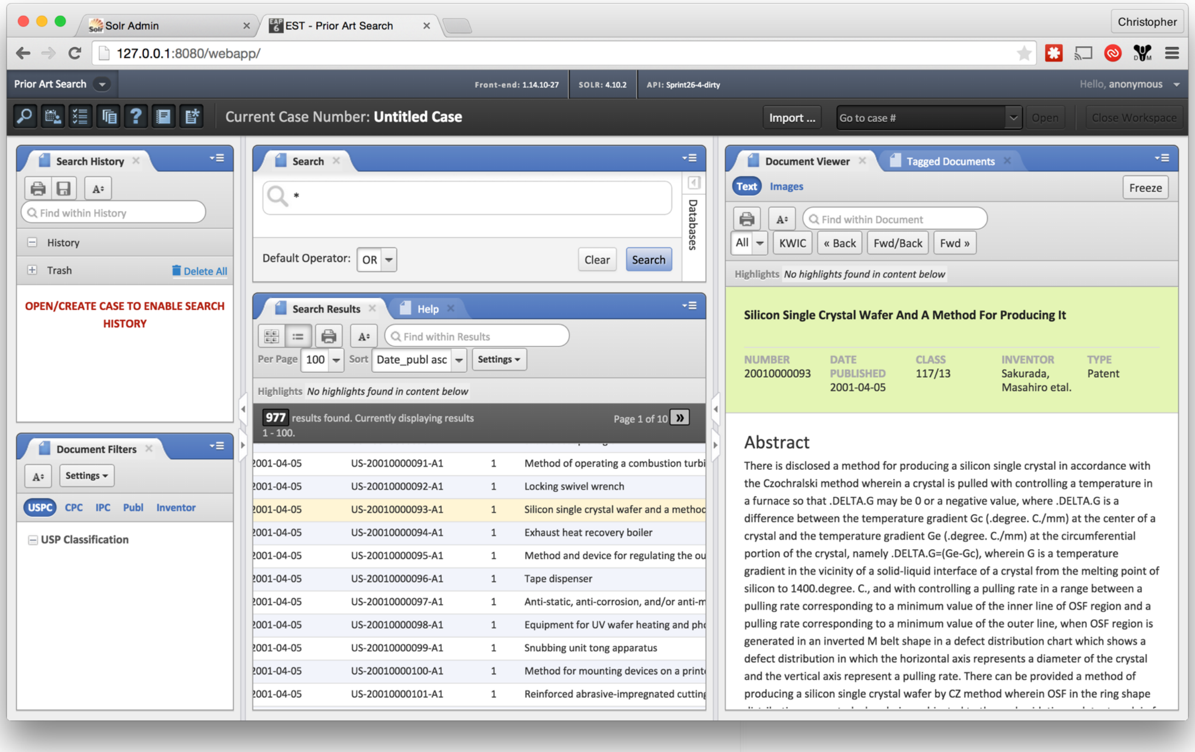
GitHub: bradfordcp

OpenSource Connections




Community Experience Distilled
Apache Solr 4 Enterprise Search Server
Third Edition
Enhance your search with faceted navigation, result highlighting, relevancy ranked sorting, and much more with this comprehensive guide to Apache Solr 4
David Smiley
Kranti Parisa
Eric Pugh
Matthew Mitchell
[PACKT] open source

Exploring Search Technologies - EST



Current Case Number: **Untitled Case**

OPEN/CREATE CASE TO ENABLE SEARCH HISTORY

Search: *

Default Operator: OR

Per Page 100 Sort Date_publ asc

Highlights *No highlights found in content below*

977 results found. Currently displaying results 1 - 100. Page 1 of 10

2001-04-05	US-20010000091-A1	1	Method of operating a combustion turbi
2001-04-05	US-20010000092-A1	1	Locking swivel wrench
2001-04-05	US-20010000093-A1	1	Silicon single crystal wafer and a methoc
2001-04-05	US-20010000094-A1	1	Exhaust heat recovery boiler
2001-04-05	US-20010000095-A1	1	Method and device for regulating the ou
2001-04-05	US-20010000096-A1	1	Tape dispenser
2001-04-05	US-20010000097-A1	1	Anti-static, anti-corrosion, and/or anti-m
2001-04-05	US-20010000098-A1	1	Equipment for UV wafer heating and pho
2001-04-05	US-20010000099-A1	1	Snubbing unit tong apparatus
2001-04-05	US-20010000100-A1	1	Method for mounting devices on a printe
2001-04-05	US-20010000101-A1	1	Reinforced abrasive-impregnated cutting

Highlights *No highlights found in content below*

Silicon Single Crystal Wafer And A Method For Producing It

NUMBER	DATE PUBLISHED	CLASS	INVENTOR	TYPE
20010000093	2001-04-05	117/13	Sakurada, Masahiro etal.	Patent

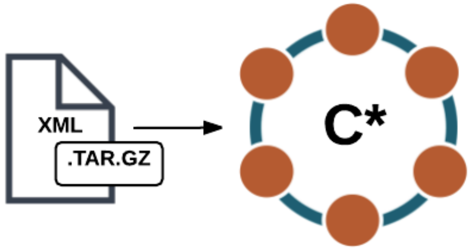
Abstract

There is disclosed a method for producing a silicon single crystal in accordance with the Czochralski method wherein a crystal is pulled with controlling a temperature in a furnace so that .DELTA.G may be 0 or a negative value, where .DELTA.G is a difference between the temperature gradient Gc (.degree. C./mm) at the center of a crystal and the temperature gradient Ge (.degree. C./mm) at the circumferential portion of the crystal, namely .DELTA.G=(Ge-Gc), wherein G is a temperature gradient in the vicinity of a solid-liquid interface of a crystal from the melting point of silicon to 1400.degree. C., and with controlling a pulling rate in a range between a pulling rate corresponding to a minimum value of the inner line of OSF region and a pulling rate corresponding to a minimum value of the outer line, when OSF region is generated in an inverted M belt shape in a defect distribution chart which shows a defect distribution in which the horizontal axis represents a diameter of the crystal and the vertical axis represent a pulling rate. There can be provided a method of producing a silicon single crystal wafer by CZ method wherein OSF in the ring shape

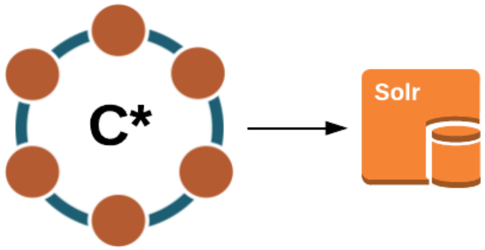
EST – Technology Stack

EST – Data Loading

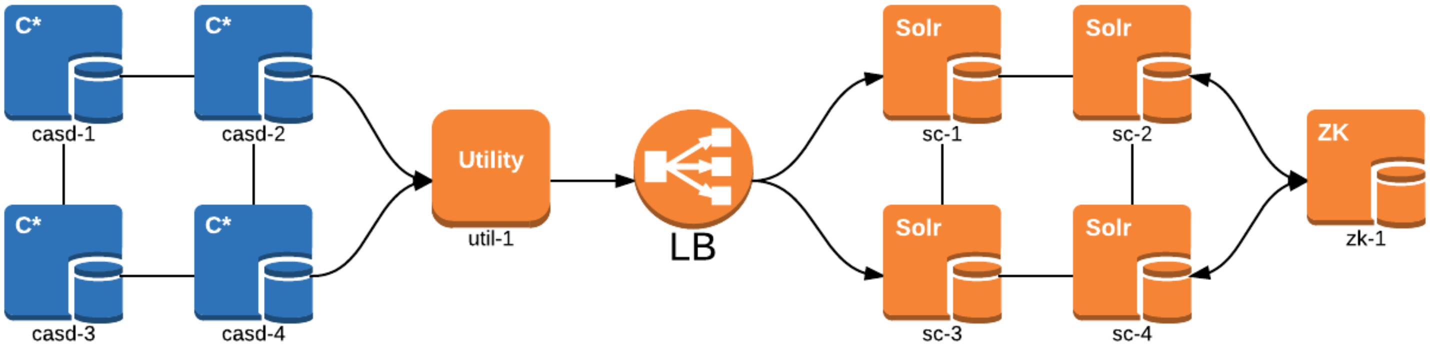
CSS Ingestion (CSS2C)



Solr Ingestion (C2S)

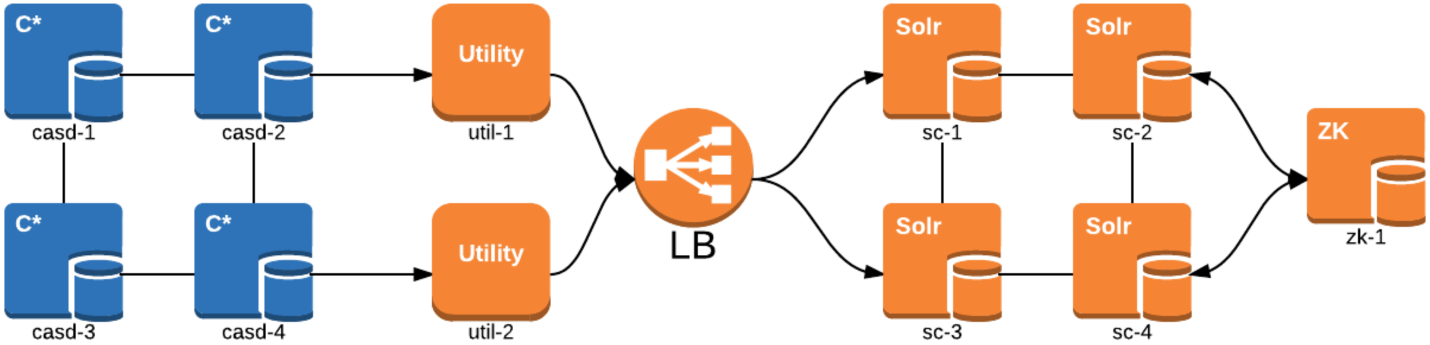


EST – C2S Process



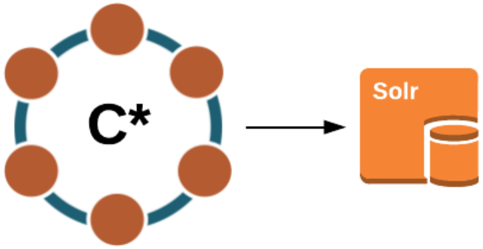
Note: some connections are omitted for clarity

EST – C2S Process (Scaled Out)



Note: some connections are omitted for clarity

EST – C2S Review



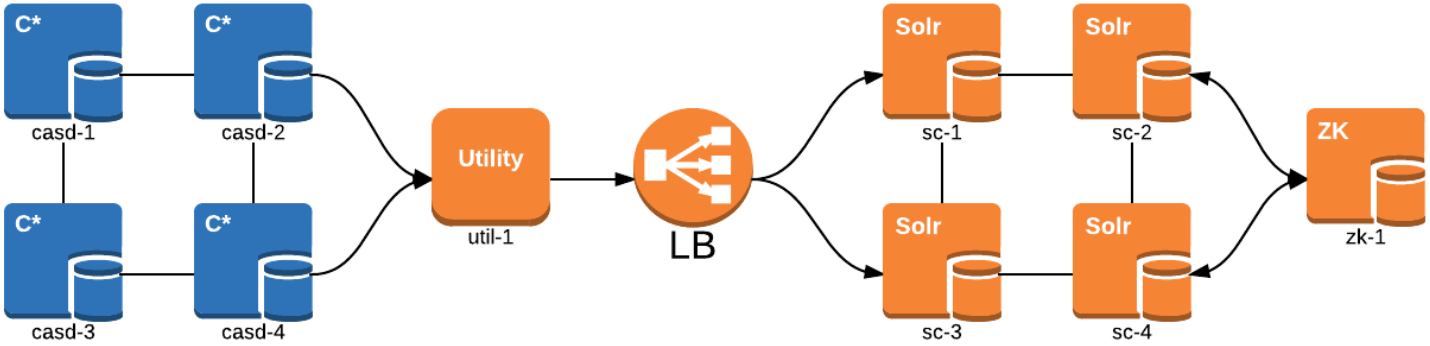
Did it work?

Why change it?

How could we make it better?

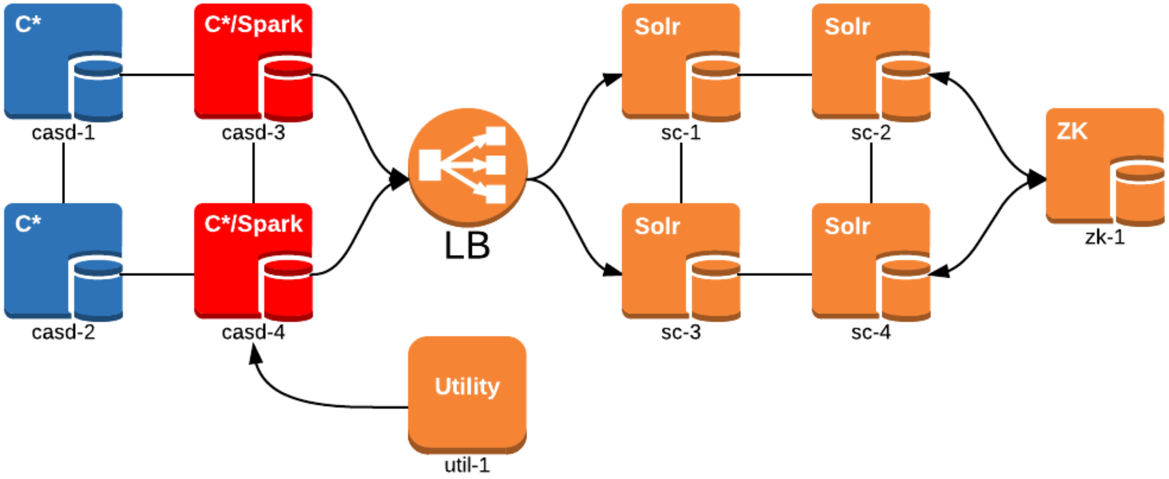
Spark

EST – Old C2S Process



Note: some connections are omitted for clarity

EST – Spark C2S Process



Note: some connections are omitted for clarity

How did this work out?

Poorly

Poor Performance

```
joinedRDD = …
joinedRDD.foreach()
  document = … // build document
  sc = new SolrConnection()
  sc.push(document)
  sc.disconnect()
// Job is done
```

Poor Performance

```
sc = new SolrConnection()
sc.push(document)
sc.disconnect()
```

Optimum Performance

```
joinedRDD = …
sc = new SolrConnection()
joinedRDD.foreach()
  document = … // build document
  sc.push(document)
sc.disconnect()
// Job is done
```

```
joinedRDD = …
joinedRDD.foreachPartition()
  sc = new SolrConnection()
  partition.foreach()
    document = … // build document
    sc.push(document)
  sc.disconnect()
// Job is done
```

Almost

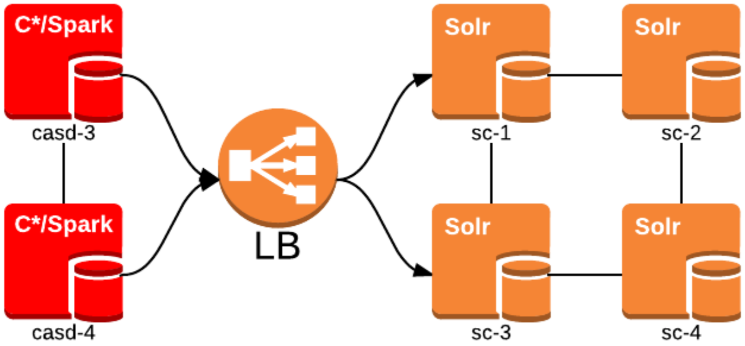
The Solution!

```
joinedRDD = …
joinedRDD.mapPartitions()
  sc = new SolrConnection()
  partition.foreach()
    document = … // build
document
    sc.push(document)
  sc.close()
  return partition.rows
.collect()
```

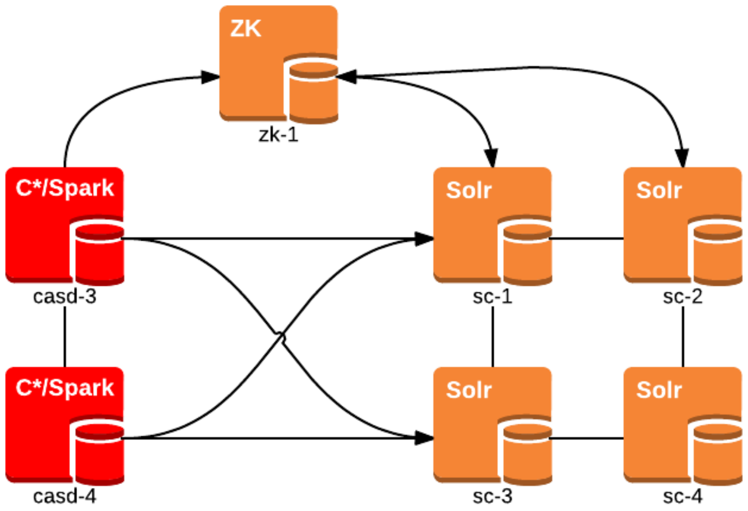
```
joinedRDD = …
joinedRDD.mapPartitions()
  sc = new SolrConnection()
  partition.foreach()
    document = … // build
document
    sc.push(document)
  sc.close()
  return partitions.rows.count
.collect()
```

Results?

Solr Indexing

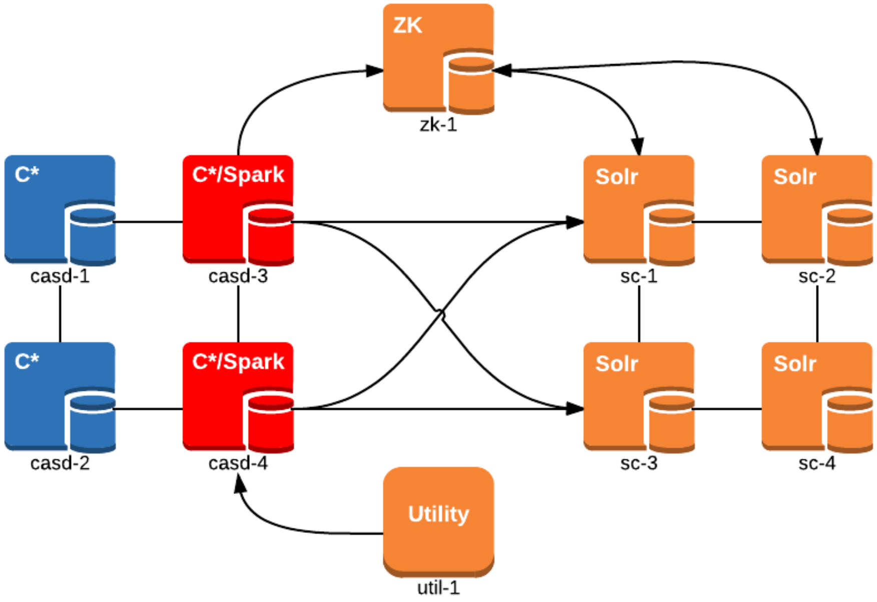


Better Solr Indexing



Note: some connections are omitted for clarity

EST – Spark C2S Process v2



Note: some connections are omitted for clarity

Success?

YUP

5x faster than the original C2S process (with optimizations)

What’s Next?

- Optimization of the C2S Spark job
- More Spark jobs
- Newer version of Spark & DSE
- Scala Spark jobs instead of Java