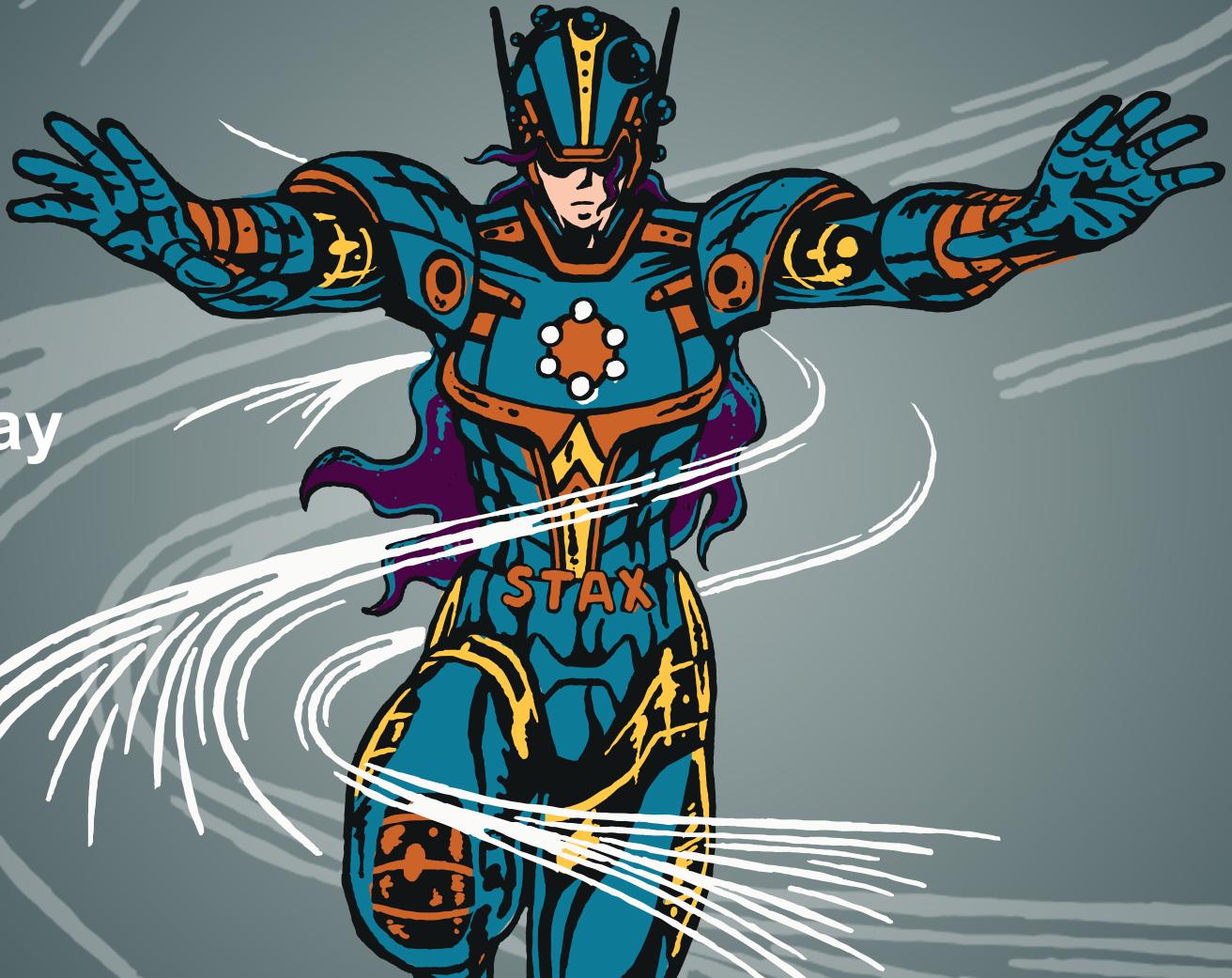


Developer Day DSE Search



What are we doing today?

- Explore the product catalog use case
- Discuss how your use case might be challenging using just Cassandra
- Use DSE Search to perform queries on different columns
- Make changes to our DSE Search schema
- Dive into full text searching

The Product Catalog

What Functionality Do We Need?

- Querying columns
- Search in text
- Sorting through results
- Counting
- Pagination

Searching

Screenshot of the Walmart website showing a search result for Huggies Little Snugglers Diapers, Size 4, 116 Diapers.

The page includes the following elements:

- Breadcrumbs:** Baby > Diapering > Diapers > Disposable Diapers
- Product Title:** HUGGIES Little Snugglers Diapers, Size 4, 116 Diapers
- Rating:** ★★★★☆ 12102 reviews
- By:** HUGGIES Walmart #: 554995154
- Image:** A large image of a box of Huggies Little Snugglers diapers featuring a smiling baby.
- Price:** \$33.87 (\$0.30 / each)
- Shipping:** 2-DAY SHIPPING
- Sold by:** Sold & shipped by Walmart
- Quantity:** 1
- Add to Cart:** Button
- Other Options:** Add to List, Add to Registry
- Shipping Information:** FREE 2-DAY SHIPPING on orders \$35+, Arrives by Thursday, Apr 26. Ship to 94016 See shipping options
- Pickup Information:** FREE PICKUP at San Leandro, 1919 Davis St Available Thu, Apr 26 See more stores
- Highlights:** DryTouch liner absorbs moisture on contact to help keep baby dry

Annotations with arrows highlight the following:

- A double-headed arrow points to the product title "HUGGIES Little Snugglers Diapers, Size 4, 116 Diapers".
- A single-headed arrow points to the large product image of the diaper box.
- A single-headed arrow points to the price "\$33.87 (\$0.30 / each)".

Searching

- Cassandra designed to allow on specific columns – partition key columns
- Some additional features to make querying more flexible
 - Secondary Indexes
 - Materialized Views
 - SASI – SSTable attached secondary indexes



Sorting

Categories ▾ Deals ▾

🔍 baby food

My Account ▾ 0 Items 🛒
Sign in

“baby food”

258 items

related searches

🔍 gerber baby food

🔍 organic baby food

🔍 baby food maker

filter results

Category

baby (258)

food & beverage (6)

shipping & pickup

buy online & pick up

in stores

Deliver to 94112 ▾

all delivery options

free 2-day shipping

In stock at San Francisco South ▾

shipping & pickup ▾

24 per page ▾

sort by relevance ▾

• relevance ✓

Featured

price-low to high

price-high to low

average ratings

best seller

newest



add to cart



add to cart



add to cart



Sorting

- Limited in Cassandra by how data is stored to disk
- Can only sort within a partition
 - Need to search on partition key
- Only clustering columns are ordered
 - Need to know what columns that can order ahead of time
 - Need to re-create table if new ordering requirements
- Clustering columns sorted in groups following the primary key ordering
- Cannot arbitrarily change clustering column order
 - Depends on the order of the proceeding clustering columns

Counting

ebay Shop by category ▾

swaddle blankets

Related: [swaddle me](#) [aden and anais](#) [swaddle wrap](#) [swaddle blankets lot](#) [muslin swaddle blanket](#) [baby receiving blankets](#) [miracle blanket](#) [swaddle designs](#)

Blankets & Throws **Search** Advanced Include description

Categories

All

↳ Baby

↳ Nursery Bedding

Blankets & Throws

Sleeping Bags & Sleepsacks

Sheets & Sets

Quilts & Coverlets

More ▾

Clothing, Shoes & Accessories

Home & Garden

Books

Cameras & Photo

Show more ▾

Gender see all

Boys (996)

Girls (2,588)

Unisex (10,488)

Not Specified (5,916)

Type see all

Blanket Sets (1,248)

Multi-Purpose Blanket (1,259)

Receiving Blanket (1,312)

Swaddling Blanket (15,214)

All Listings Auction Buy It Now

Sort: Best Match View: Group Similar Listings

Guaranteed 3 day delivery

20,396 results for swaddle blankets Save this search

Shop by Gender

Girls Boys Unisex

 **SwaddleMe Muslin Swaddle Blankets, Anchors**
Brand New  8 product ratings
\$12.95
Buy It Now
Free Shipping
119 Sold

 **Baby Boy Layette 5 Swaddling Flannel Receiving Blankets Camel Grey Orange White**
Brand New  2 product ratings
\$10.95

Browse related

 Sports Nursery Swaddling Blankets

 Girls Nursery Swaddling Blankets

 Swaddling Blanket

Counting

- Cassandra does have COUNT, but...
- Need to read through partitions to get the count
- If not restricting to a partition, that means doing a full table scan
- Also no way to count how much each value shows up in a column



Pagination

The screenshot shows the Newegg homepage with a search bar containing "baby monitor". The search results page displays a grid of baby monitors. A sidebar on the left shows categories like "Department" and "Customer Insights". The main content area includes filters for "Search Within", "Sold By", "Sort by", and "View". A navigation bar at the top includes links for "Log in or Register", "Try PREMIER", "0 Items", "Wish List", and "Customer Service".

Search Results: "baby monitor"

"BABY MONITOR" Top Sellers Free Shipping Newegg Premier Eligible

Search Within: GO

Sold By: Newegg All Sellers Sort by: Featured Items

View: 12

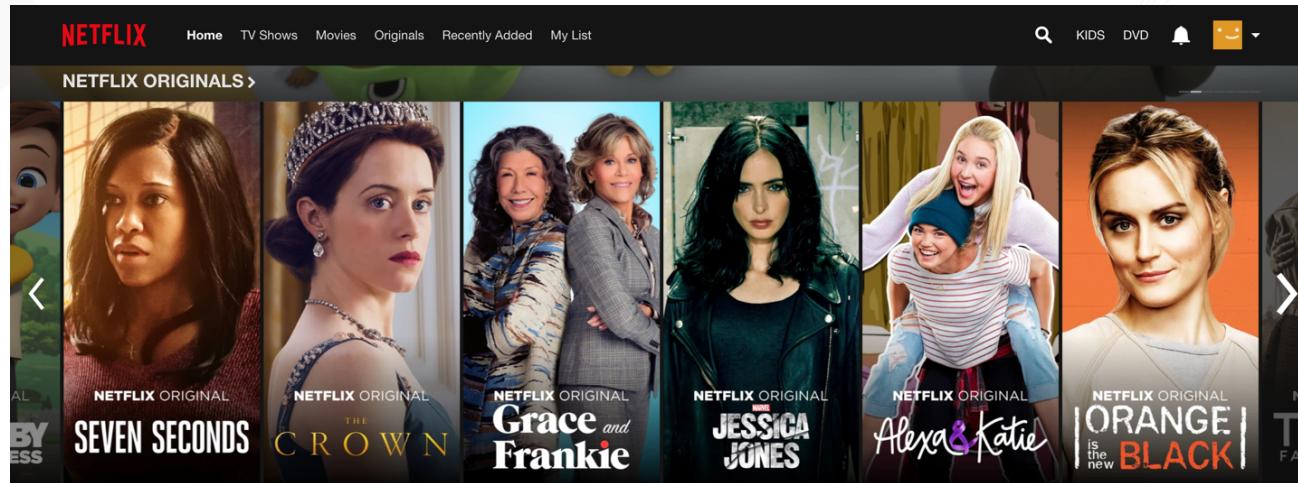
Page 1/29

Advertisement

Image	Product Name	Rating	Details
	Sense-U Breathing & Rollover Baby Movement Monitor: Monitor Your Baby's Breathing, Rollover and	5 stars (9)	<small>Limited Time Only</small>
	Motorola 7" Baby Monitor/Digital Picture Frame with Wireless Camera and Live Video	4 stars	
	Motorola Digital Audio Baby Monitor with 2 Parent Units - MBP16-2	4 stars	

Pagination

- Affected by limitations of counting, cannot efficiently do offset paging
- Cassandra driver can do cursor-based paging through results
- Essentially can only go forward or back from the current page



Introducing DSE Search



Apache Solr™

- Open-source enterprise search platform
- Provides tools and an interface for running search queries



Apache Lucene™

- Text indexing and search engine library
- The core of the indexing and search capabilities available with DSE Search

DSE Search

- Rich functionality not available in Cassandra
- More convenient way to access data
 - Doesn't require complex data models or data duplication
 - Less work needed on the application side to format results
- Accessible from CQL
 - Core functionality using pure CQL syntax
 - More features available using the `solr_query` column
 - Changes to the search index schema or configuration

Getting Started with Search

Creating a search index

- Use CQL to create the search index
 - Command runs on all Search nodes in the datacenter
 - Uses a default search schema and config, which can be altered later
 - Also indexes existing table data
 - New data that you add to the table is automatically indexed

```
// Index all columns in the table  
CREATE SEARCH INDEX ON keyspace.table;
```

```
// Only index certain columns in the table  
CREATE SEARCH INDEX ON keyspace.table WITH COLUMNS column1, column2, ...;
```

CQL Search Query

- Accessing a query using search index can be done through CQL
- Will execute query just using Cassandra, if possible
- Otherwise will use the search index

```
SELECT * FROM keyspace.table WHERE predicate1 AND predicate2 AND ...;
```

```
SELECT column1, column2, ... FROM keyspace.table WHERE predicate1 AND predicate2 AND ...;
```

```
SELECT COUNT(*) FROM keyspace.table WHERE predicate1 AND predicate2 AND ...;
```

```
SELECT * FROM keyspace.table WHERE predicate1 AND predicate2 AND ... LIMIT #rows;
```

CQL Query

Predicates

```
CREATE TABLE killrvideo.users (
    userid UUID,
    created_date TIMESTAMP,
    email TEXT,
    firstname TEXT,
    lastname TEXT,
    phone_number SET<TEXT>
    PRIMARY KEY ((userid))
);
```

- Equality and Inequality

```
SELECT * FROM killrvideo.users WHERE email = 'eboyeri5@aol.com';
SELECT * FROM killrvideo.users WHERE email != 'eboyeri5@aol.com';
```

- Range

```
SELECT * FROM killrvideo.users
WHERE created_date >= '2018-04-01' AND created_date < '2018-05-01';
```

- Contains

```
SELECT * FROM killrvideo.users
WHERE phone_number CONTAINS '650-389-6000';
```

- In

```
SELECT * FROM killrvideo.users
WHERE firstname IN ('Beauregard', 'Muffin');
```

- Like

```
SELECT * FROM killrvideo.users WHERE lastname LIKE 'McD%';
```

CQL Query

Order By

- Sort by any column
- Text can't be sorted by default, but can be changed in the search index schema
- Ascending (default) or descending order

```
SELECT email, added_date, lastname, firstname  
FROM killrvideo.users  
ORDER BY added_date DESC, lastname ASC, firstname ASC;
```

Time for an exercise!

Search Queries



Text Search

Text Search

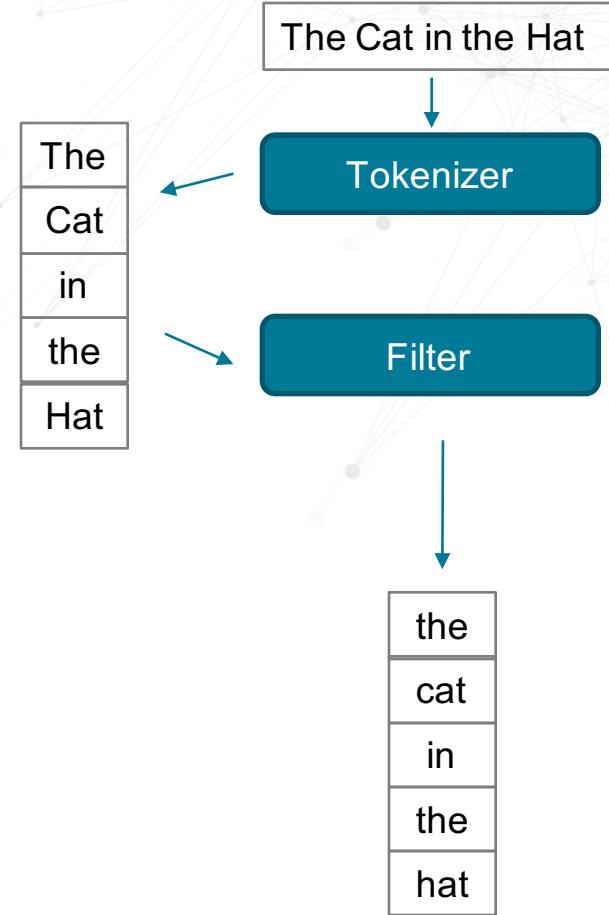
- One of DSE Search's strengths is in full-text search
- Retrieves results based on how well the text matches the search parameters
 - Calculates a relevancy score
 - Only includes the rows with the highest score are included in the search results
- Uses the more expressive Lucene query syntax instead of just CQL

Field Types

- The Search equivalent to data types found in the search index schema
- Cassandra data types map to a corresponding field type
- Some data types may have several compatible field types
 - For the Cassandra TEXT data type, you can use:
 - StrField (default)
 - TextField (text search capabilities)
- Power users can even create fully customizable field types

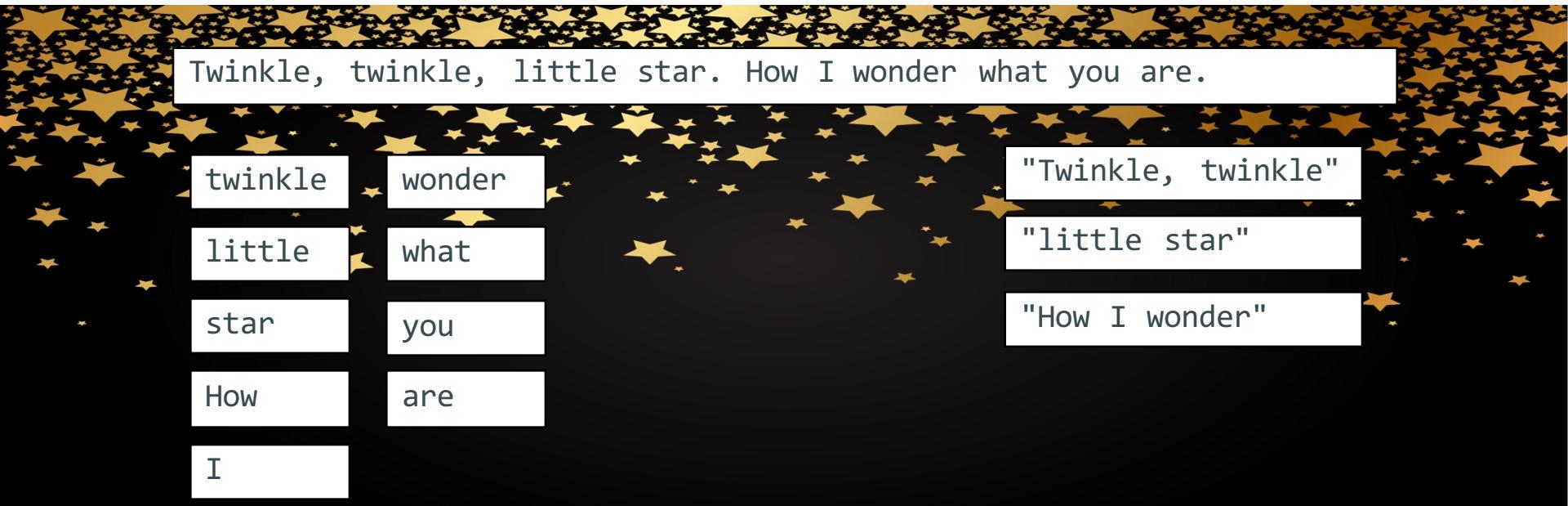
TextField

- TextField data will undergo some processing as it is indexed
 - Analysis Chain
 - Tokenizer – breaks up text into tokens
 - Filter – performs some sort of processing
 - Resulting terms is what is indexed
- Search parameters also goes through analysis chain
 - Compared against the indexed terms
 - Row may be included in the results if matches



Terms and Phrases

- Term – Tokenized data, or words, from text analysis or search input
- Phrase – Terms that are positioned in a certain order



Search index schema

- Written and stored as a XML file
- Can edit using CQL, or by uploading the new schema XML file
- Defines field types; declares columns / fields and its field type

```
DESCRIBE ACTIVE SEARCH INDEX SCHEMA ON keyspace.table; // CQLSH only  
DESCRIBE PENDING SEARCH INDEX SCHEMA ON keyspace.table; // CQLSH only
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>  
<schema name="autoSolrSchema" version="1.5">  
  <types>  
    <fieldType class="org.apache.solr.schema.StrField" name="StrField"/>  
    :  
  </types>  
  <fields>  
    <field indexed="true" multiValued="false" name="title" type="StrField"/>  
    :  
  </fields>  
  <uniqueKey>video_id</uniqueKey>  
</schema>
```

Making changes to the schema

- We need to define our TextField and change the field type for our fields
- Schema with changes that haven't been applied yet is PENDING
- Currently running schema is ACTIVE

```
ALTER SEARCH INDEX SCHEMA ON keyspace.table
ADD fieldType[@name='TextField', @class='solr.TextField']
WITH '{"analyzer": [{"tokenizer": {"class": "solr.StandardTokenizerFactory"}, "filter": {"class": "solr.LowerCaseFilterFactory"}}]}';
```

```
ALTER SEARCH INDEX SCHEMA ON keyspace.table
SET field[@name='name']@type='TextField';
```

Applying changes to the search index

- Reload the search schema to apply the changes to the schema
 - PENDING schema replaces the ACTIVE schema
- Rebuild the search index to reindex the existing data in the table
 - Current indexes would not match the new schema
 - Not needed if changes only made to the search index config



```
RELOAD SEARCH INDEX ON keyspace.table;  
REBUILD SEARCH INDEX ON keyspace.table;
```

Using solr_query in CQL

- solr_query is a pseudo-column created with the search index
- Set a Lucene query string to the solr_query column in the WHERE clause

```
SELECT select-clause FROM keyspace.table WHERE solr_query = 'Lucene-query';
```

- Passed to DSE Search / Solr to be executed
- Search results return to Cassandra, and then retrieves the actual row and column data

Lucene Query syntax in a nutshell

- Search all the things: *:*****
- Term search: *field-name:term*
- Phrase search: *field-name:"term term term"*
must be capitalized!
- Multiple terms: *field-name:(term OR (term AND term))*
- Multiple fields: *field-name:term AND (field-name:term OR field-name:term)*
- Range search: *field-name:(1 TO 0]*

exclusive bound

inclusive bound

must be capitalized!

```
SELECT * FROM killrvideo.videos WHERE solr_query = 'name:cassandra';
SELECT * FROM killrvideo.videos WHERE solr_query = 'name:"Distributed Data Show"';
SELECT * FROM killrvideo.users WHERE solr_query = 'name:(Jack OR Jill)';
SELECT * FROM killrvideo.videos WHERE solr_query = 'name:something AND tags:cats';
SELECT * FROM killrvideo.videos WHERE solr_query = 'year:[2017 TO *]';
```

Levenshtein Distance

- Measure of how similar two strings are
- Based on the number of edits for one string to match the other
- Used by both fuzzy search and proximity search

```
// Distance between the word kitten and sitting is 3
```

```
Edit 1: kitten → sitten (substitution of "s" for "k")
```

```
Edit 2: sitten → sittin (substitution of "i" for "e")
```

```
Edit 3: sittin → sitting (insertion of "g" at the end)
```

Fuzzy Search

- Add ~ at the end of a term
- Degree of similarity to the term is controlled by adding a value after the ~
 - This optional parameter can be 1 or 2
 - Represents the max number of edits to the indexed term

```
SELECT * FROM keyspace.table WHERE solr_query = 'field:term~#';
```

seven~1

The Magnificent Seven
Se7en
The Even Stevens Movie

Proximity Search

- Add ~ at the end of a phrase
- Degree of similarity controlled by adding a value after the ~
- Represents the maximum distance that terms in the phrase can be apart

```
SELECT * FROM keyspace.table WHERE solr_query = 'field:"phrase"~#';
```

"the road"~3'



Kickboxer 2: The Road Back
The Best of Bray Road
The Black Rider: Revelation Road

Time for an exercise!

Text Search



Wrapping Up

In summary...

- The use case for product catalog has certain requirements
 - Can be a challenge to implement with just Cassandra
 - DSE Search makes it much easier
- If you can write CQL, you can search
 - Not absolutely required to learn Lucene search syntax
 - Also for managing your schema and config
 - Use the ***solr_query*** column for more complex search and text search
- DSE Search is great at many things, especially:
 - Text search
 - Counting

Continue Your Learning at DataStax Academy

- [DS310: DataStax Enterprise Search](https://academy.datastax.com)
 - Free self-paced course
 - Everything we discussed today, in more detail!
 - Even more queries!
 - Facet queries
 - Join queries
 - Geo-spatial queries

<https://academy.datastax.com>



Any questions?



Thank You

