Lucene’s Faceted Search

Overview and Recent Improvements

Who am I?

- Greg Miller
- Software Engineer
- Apache Lucene Committer / PMC Member
- Working on Amazon’s Product Search Engine
- You can find me here: https://lucene.apache.org/whoweare.html

Who are you?

- Ideally an audience with…
  - A basic understanding of *information retrieval* / *search*
  - Maybe some experience with Apache Lucene
  - Maybe some experience with open source development
- Not necessarily an audience with…
  - *Faceted search* experience

In the next 30 minutes

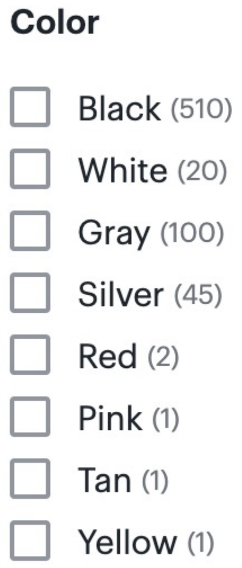
Everyone should walk away from this talk with…

1. A high-level understanding of Lucene's Faceting capabilities
2. A sense of the active development in Lucene's Faceting module

Part I: What is Faceting?

What is faceting?

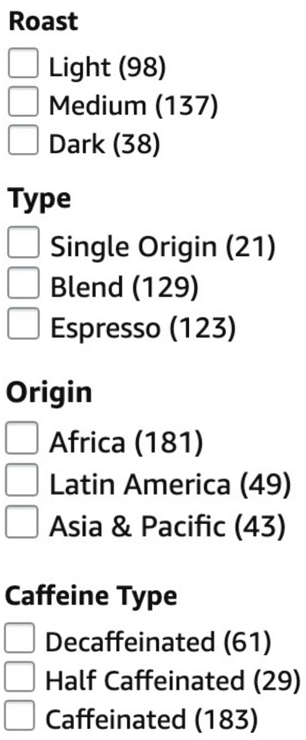
- Commonly supports *filtering* search experiences
    - *Filtering* allows users to modify search results using structured data
    - *Faceting* allows users to “look into the future” to understand how applying a filter will affect search results



Color

□ Black (510)
□ White (20)
□ Gray (100)
□ Silver (45)
□ Red (2)
□ Pink (1)
□ Tan (1)
□ Yellow (1)

Faceting / Filtering: Coffee



Roast

- □ Light (98)
- □ Medium (137)
- □ Dark (38)

Type

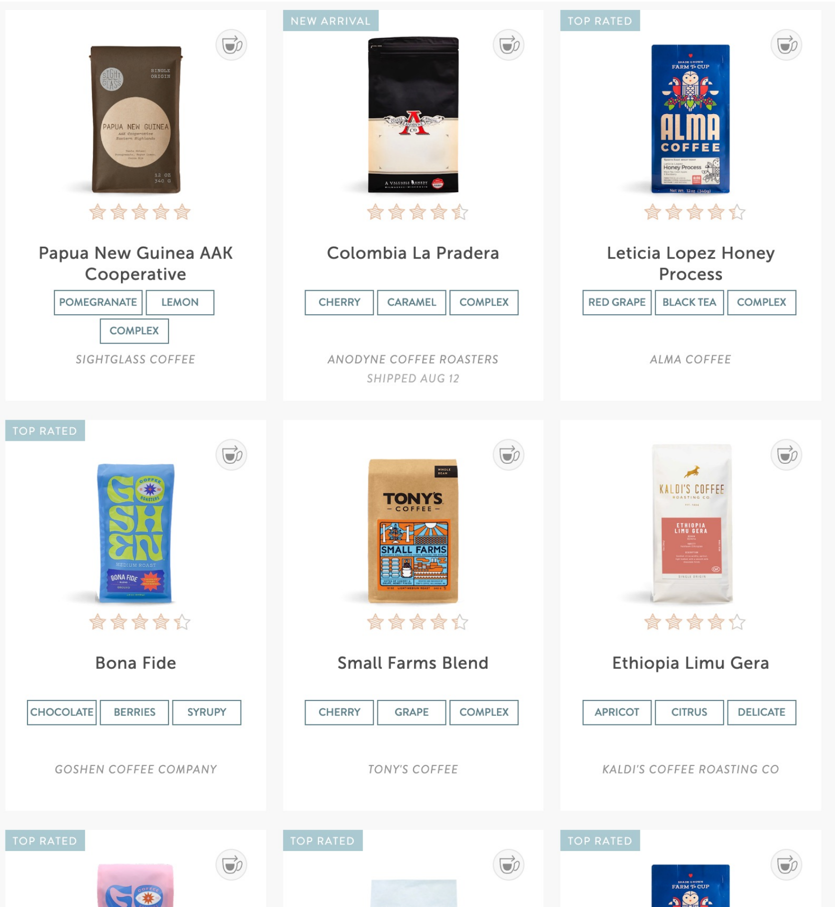
- □ Single Origin (21)
- □ Blend (129)
- □ Espresso (123)

Origin

- □ Africa (181)
- □ Latin America (49)
- □ Asia & Pacific (43)

Caffeine Type

- □ Decaffeinated (61)
- □ Half Caffeinated (29)
- □ Caffeinated (183)



Papua New Guinea AAK Cooperative

POMEGRANATE | LEMON | COMPLEX

SIGHTGLASS COFFEE

NEW ARRIVAL

Colombia La Pradera

CHERRY | CARAMEL | COMPLEX

ANODYNE COFFEE ROASTERS

SHIPPED AUG 12

TOP RATED

Leticia Lopez Honey Process

RED GRAPE | BLACK TEA | COMPLEX

ALMA COFFEE

TOP RATED

Bona Fide

CHOCOLATE | BERRIES | SYRUPY

GOSHEN COFFEE COMPANY

Small Farms Blend

CHERRY | GRAPE | COMPLEX

TONY'S COFFEE

Ethiopia Limu Gera

APRICOT | CITRUS | DELICATE

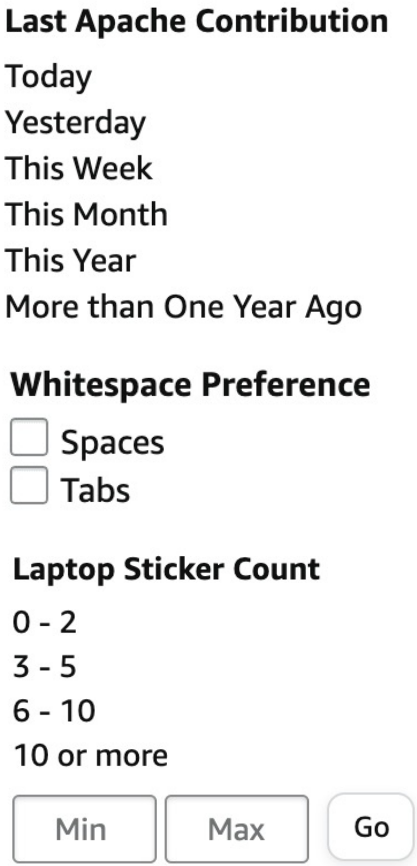
KALDI'S COFFEE ROASTING CO

TOP RATED

TOP RATED

TOP RATED

Faceting / Filtering: ~~Coffee~~ ApacheCon Attendees



Last Apache Contribution

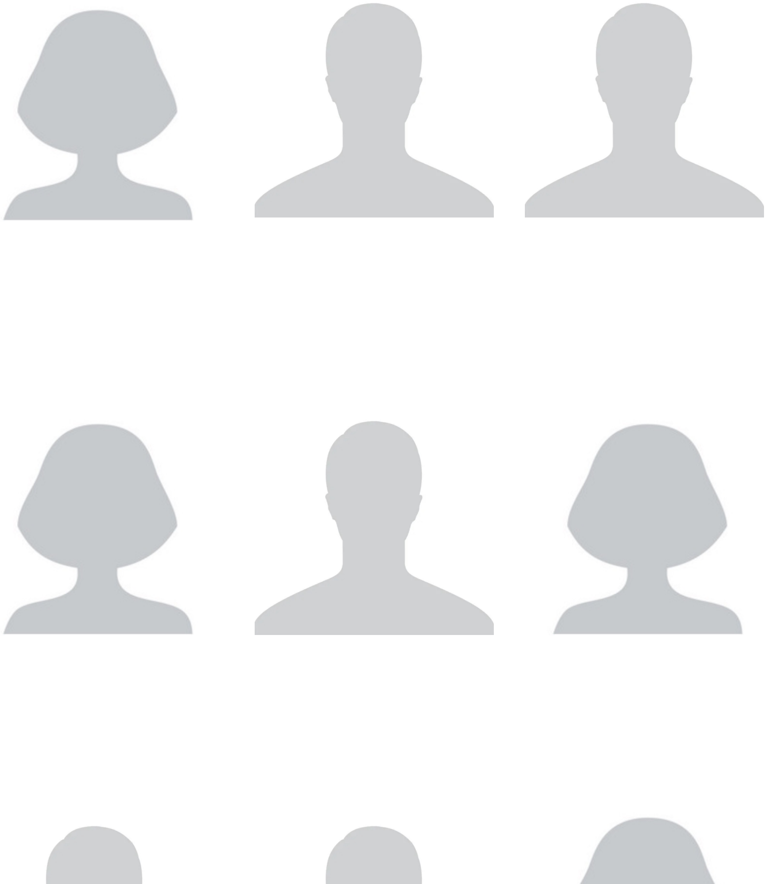
Today
Yesterday
This Week
This Month
This Year
More than One Year Ago

Whitespace Preference

□ Spaces
□ Tabs

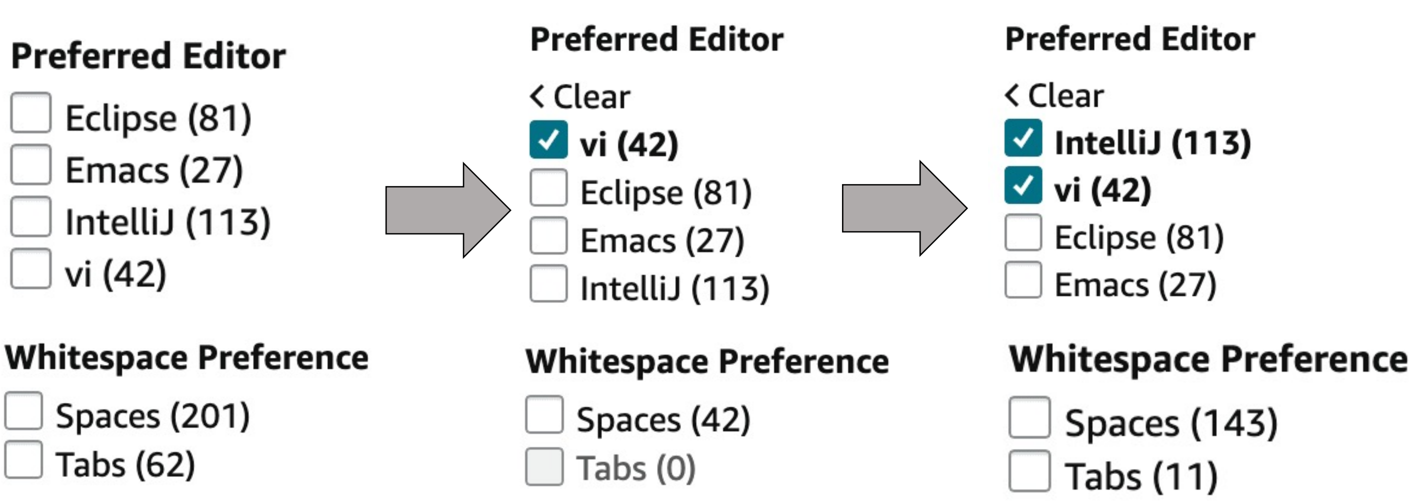
Laptop Sticker Count

0 - 2
3 - 5
6 - 10
10 or more



Faceting / Filtering: ApacheCon Attendees

Faceting can be used to understand a filter's impact before applying it.



Preferred Editor

☐ Eclipse (81)
☐ Emacs (27)
☐ IntelliJ (113)
☐ vi (42)

Whitespace Preference

☐ Spaces (201)
☐ Tabs (62)

Preferred Editor

‹ Clear
☑ **vi (42)**
☐ Eclipse (81)
☐ Emacs (27)
☐ IntelliJ (113)

Whitespace Preference

☐ Spaces (42)
☐ Tabs (0)

Preferred Editor

‹ Clear
☑ **IntelliJ (113)**
☑ **vi (42)**
☐ Eclipse (81)
☐ Emacs (27)

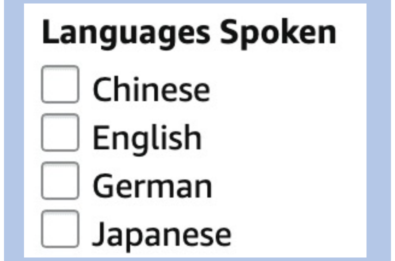
Whitespace Preference

☐ Spaces (143)
☐ Tabs (11)

Faceting / Filtering: ApacheCon Attendees

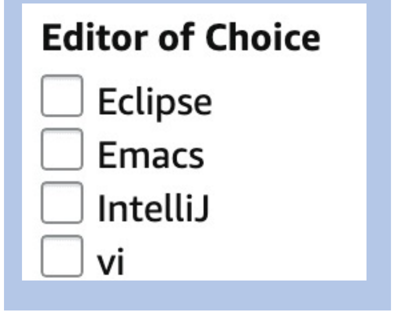
Two high-level types of faceting:

Categorical



Languages Spoken

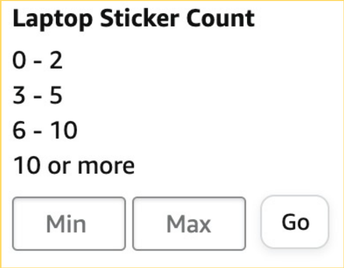
□ Chinese
□ English
□ German
□ Japanese



Editor of Choice

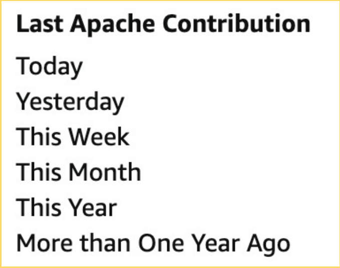
□ Eclipse
□ Emacs
□ IntelliJ
□ vi

Numeric



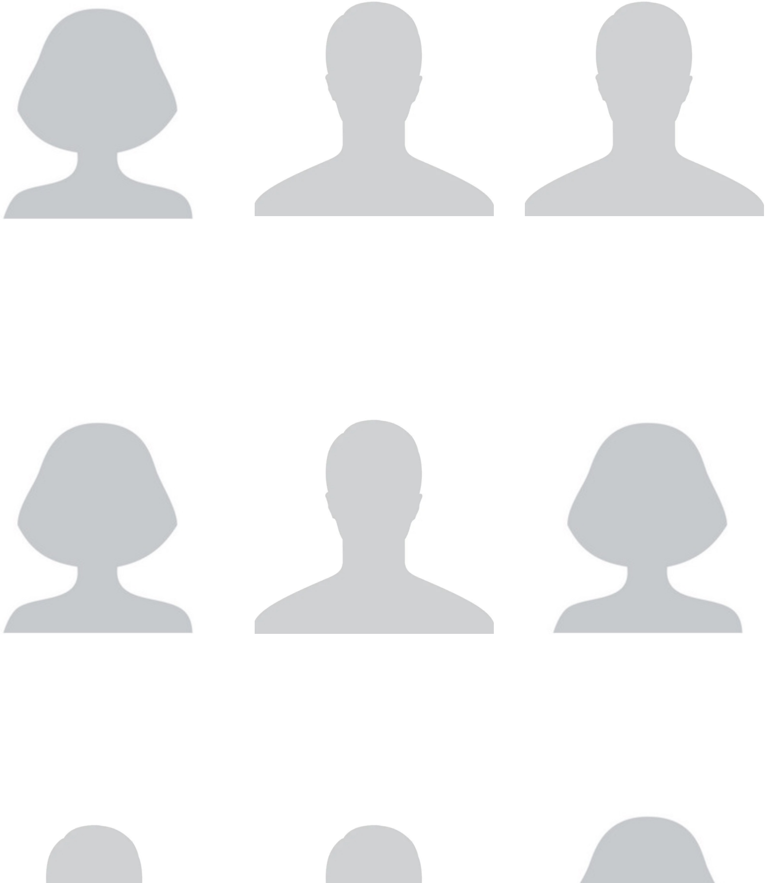
Laptop Sticker Count

0 - 2
3 - 5
6 - 10
10 or more

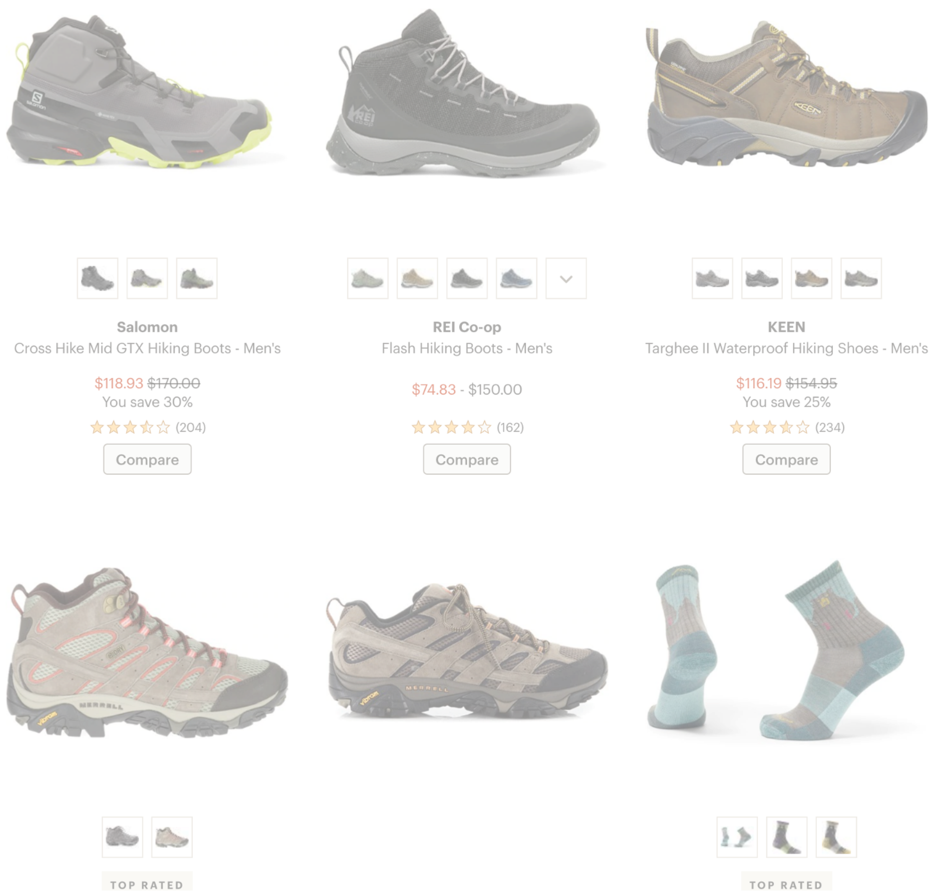


Last Apache Contribution

Today
Yesterday
This Week
This Month
This Year
More than One Year Ago



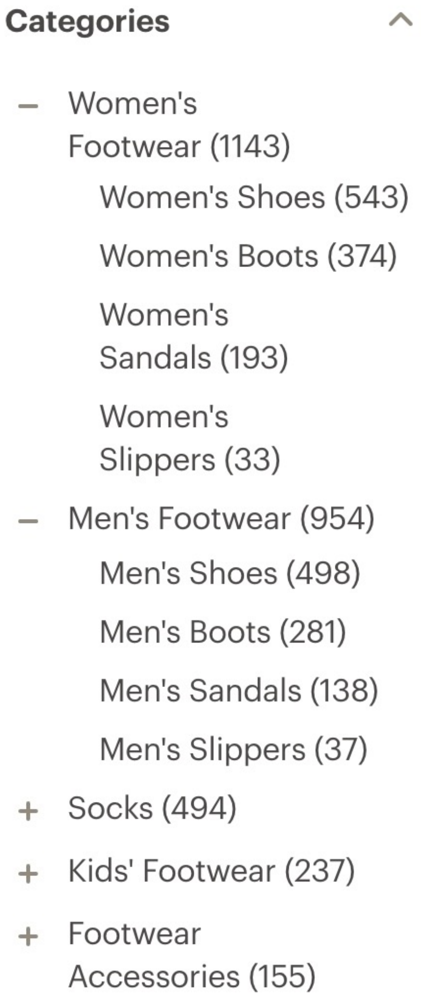
Categorical Faceting + Hierarchies



Salomon
Cross Hike Mid GTX Hiking Boots - Men's
$118.93 ~~$170.00~~
You save 30%
(204)
Compare

REI Co-op
Flash Hiking Boots - Men's
$74.83 - $150.00
(162)
Compare

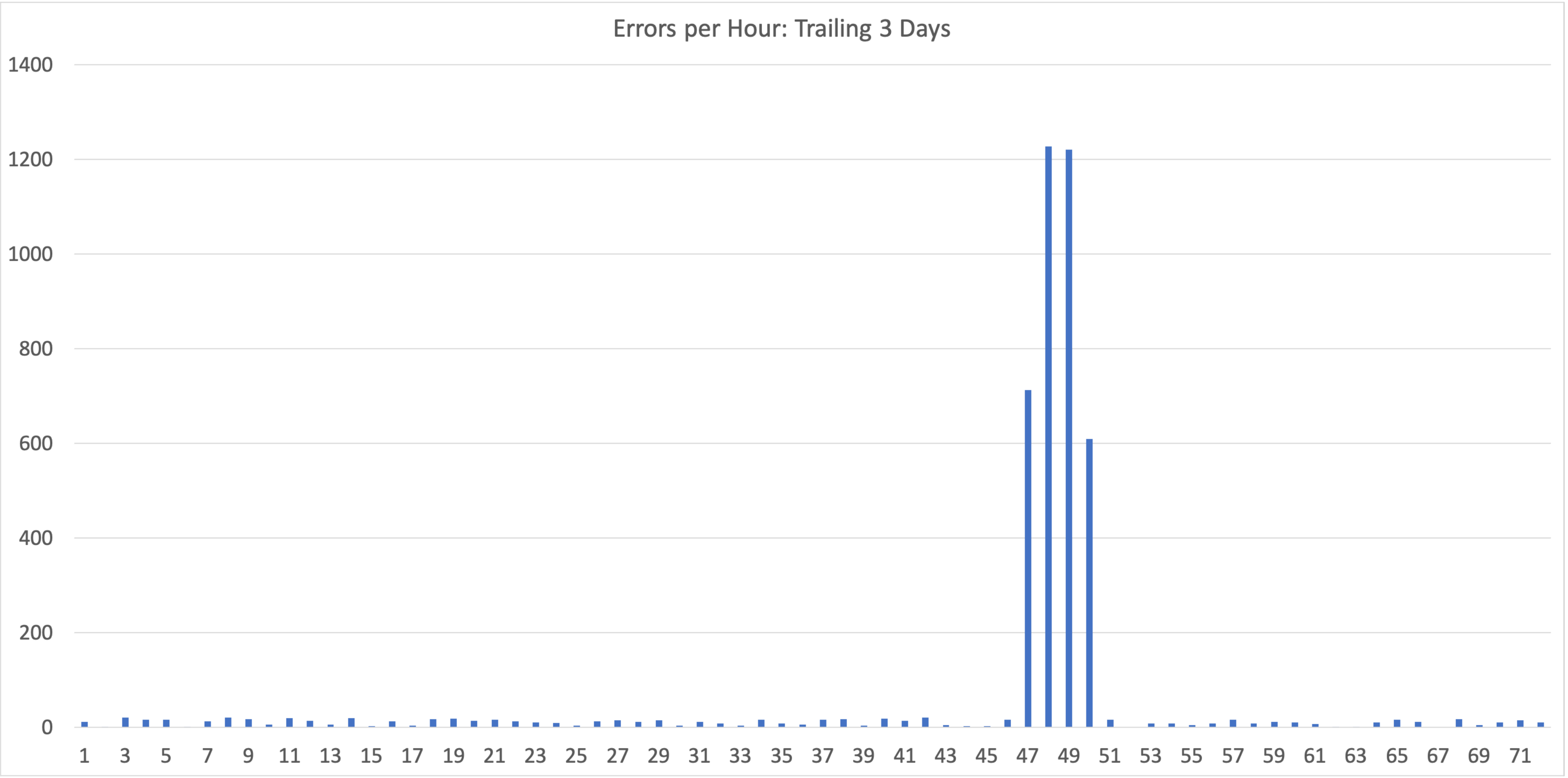
KEEN
Targhee II Waterproof Hiking Shoes - Men's
$116.19 ~~$154.95~~
You save 25%
(234)
Compare



Categories

- – Women's Footwear (1143)
  - Women's Shoes (543)
  - Women's Boots (374)
  - Women's Sandals (193)
  - Women's Slippers (33)
- – Men's Footwear (954)
  - Men's Shoes (498)
  - Men's Boots (281)
  - Men's Sandals (138)
  - Men's Slippers (37)
- + Socks (494)
- + Kids' Footwear (237)
- + Footwear Accessories (155)

More than just filtering



More than just filtering

Documents:

```
{
  msg: “2022-10-03 22:45:56 ERROR NullPointerException …”,
  log_level: “ERROR”,
  timestamp: 1661527803
}
```

Facet on:

```
[
  (1661527800-1661531400),
  (1661531400-1661535000),
  …
]
```

- 72 ranges – one for each hour over past three days

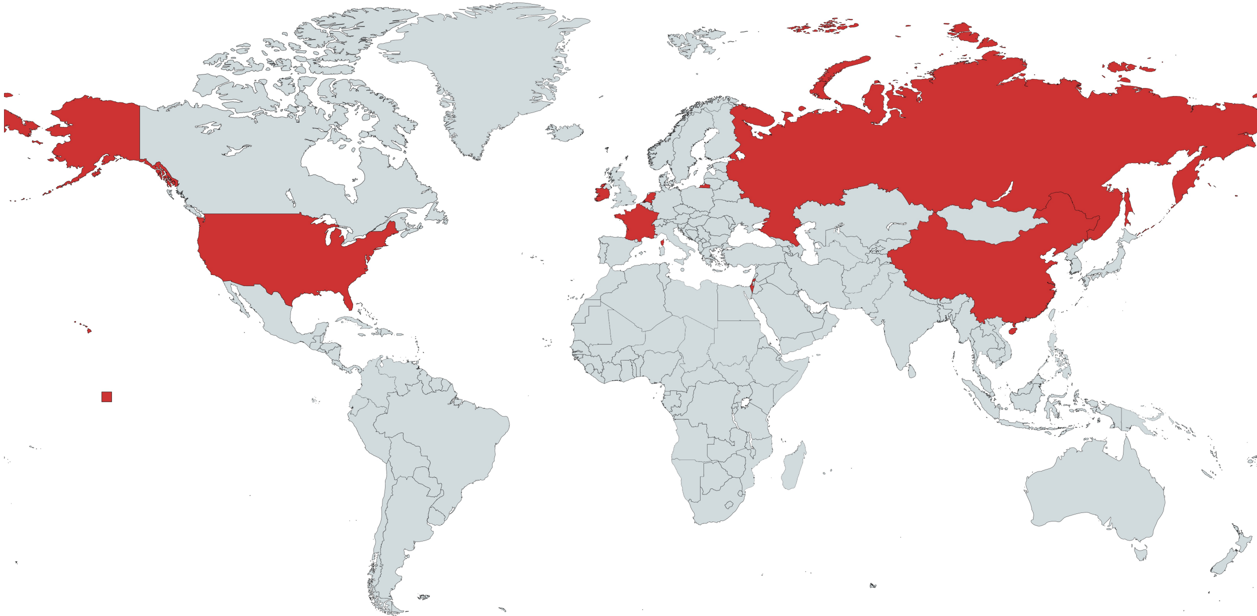
Part II: Community Activity

Development Velocity

- Development tapered off going into 2017; picked back up in 2021
- Faceting issues resolved by year:
  - 2017-2020: 6
  - 2021: 27
  - 2022: 22 (as of April, 28th)
- Nine updates have been made to faceting nightly benchmarks in 2021-2022
- Twenty individuals involved in some capacity over the last two years

Community Participation

- Participating community members in 2021 – 2022 were…
  - Located in nine different countries:
    - China, France, Ireland, Israel, Netherlands, Russia, United States
  - Employed by at least four different companies:
    - Amazon
    - Elastic
    - Headhunter Group
    - MongoDB



Community Participation

- Adrien Grand
- Alexander Lukyanchikov
- Ankur Goel
- Chris Hegarty
- Feng Guo
- Gautam Worah
- Greg Miller
- Grigoriy Troitskiy
- Luca Cavanna
- Marc D’Mello
- Michael McCandless
- Michael Sokolov
- Mike Drob
- Patrick Zhai
- Robert Muir
- Sejal Pawar
- Shai Erera
- Stefan Vodita
- Yuting Gan
- Zachary Chen

Thank You!

Part III: New Features

FacetSets

- New faceting capability added in June, 2022
  - Thanks Marc D’Mello / Shai Erera (and everyone who provided input)
  - Jira: LUCENE-10274
- Enables faceting on value combinations across dimensions
  - For more, see FacetSets.adoc

FacetSets

How many actors or actresses starred in each film genre (e.g., *Action, Adventure, Comedy, Drama*) in the 1980s? How about the 1990s?

```
{
  "name": "Sigourney Weaver",
  "films": [("Comedy", 1988), ("Drama", 2000), …]
}
{
  "name": "Harrison Ford",
  "films": [("Action", 1977), ("Adventure", 1981), …]
}
```

FacetSets

Documents:

```
{
  "name": "Sigourney Weaver",
  "films":
    [
      ("Comedy", 1988),
      ("Drama", 2000)
    ]
}
```

Facet on:

```
[ ("Action", 1980–1989),
  ("Adventure", 1980–1989),
   …
  ("Comedy", 2000–2009),
  ("Drama"), 2000–2009)
]
```

Genre: 1980s

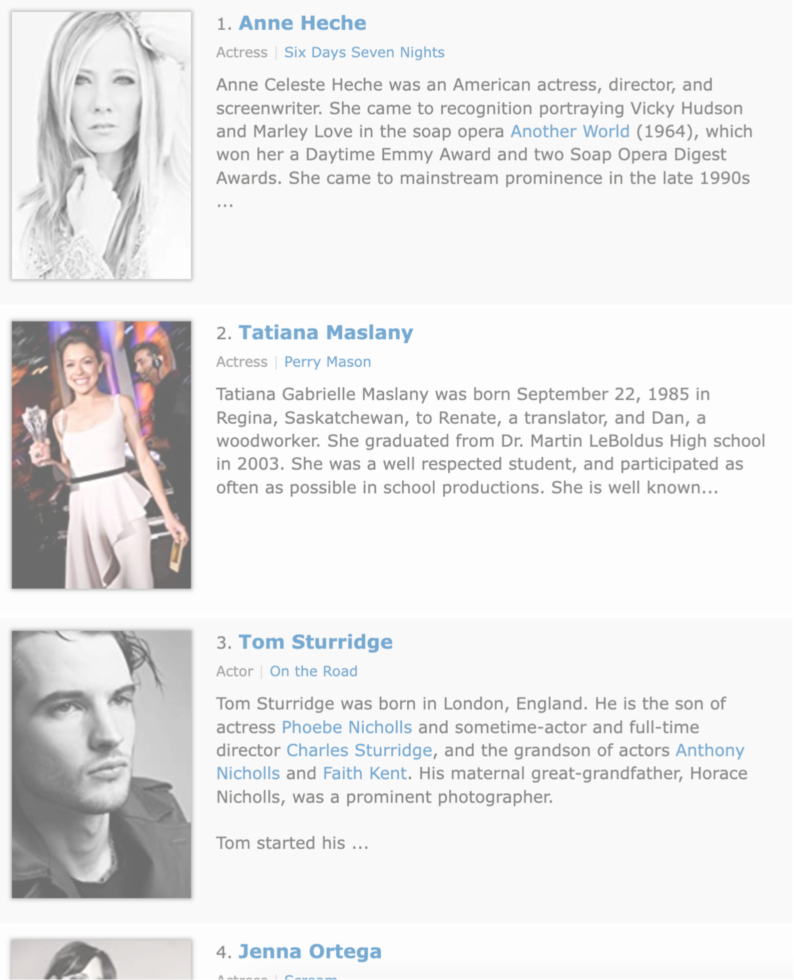
□ Action
□ Adventure
□ Comedy
□ Drama

Genre: 1990s

□ Action
□ Adventure
□ Comedy
□ Drama

Genre: 2000s

□ Action
□ Adventure
□ Comedy
□ Drama



1. **Anne Heche**

Actress | Six Days Seven Nights

Anne Celeste Heche was an American actress, director, and screenwriter. She came to recognition portraying Vicky Hudson and Marley Love in the soap opera Another World (1964), which won her a Daytime Emmy Award and two Soap Opera Digest Awards. She came to mainstream prominence in the late 1990s ...

2. **Tatiana Maslany**

Actress | Perry Mason

Tatiana Gabrielle Maslany was born September 22, 1985 in Regina, Saskatchewan, to Renate, a translator, and Dan, a woodworker. She graduated from Dr. Martin LeBoldus High school in 2003. She was a well respected student, and participated as often as possible in school productions. She is well known...

3. **Tom Sturridge**

Actor | On the Road

Tom Sturridge was born in London, England. He is the son of actress Phoebe Nicholls and sometime-actor and full-time director Charles Sturridge, and the grandson of actors Anthony Nicholls and Faith Kent. His maternal great-grandfather, Horace Nicholls, was a prominent photographer.

Tom started his ...

4. **Jenna Ortega**

Actress | Scream

Feature Parity Between Taxonomy and SSDV

- Lucene has two different implementations of categorical faceting
  - FastTaxonomyFacetCounts
  - SortedSetDocValueFacetCounts
- In January, 2022, “hierarchical faceting” support was added to SSDV faceting, closing the feature gap with taxonomy faceting
  - Thanks Marc D’Mello! (Jira: LUCENE-10250)
  - E.g., “Books/Fiction/Science Fiction”
- Builds taxonomy knowledge from inspecting the field data, requiring no external taxonomy source
  - Cached/reused until index is refreshed/reopened

Feature Parity Between Taxonomy and SSDV

- With feature parity, should we converge to a single implementation?
- Two separate implementations have created healthy competition, pushing improvements to both approaches.
- Remaining differences are now related to usability and performance. Time to incorporate the best of both into a single approach?
- To be continued…
    - Github: #11717

… and more

- Extensible aggregation functions for association faceting
  - Jira: LUCENE-10444
- Strong multi-value field support across all faceting implementations
  - Both on indexed fields as well as “value sources” (i.e., dynamic data sources)
  - Jira: LUCENE-9948, LUCENE-9946, LUCENE-10245
- Faceting on any generic “string” field (i.e., `SORTED` / `SORTED_SET`)
  - Jira: LUCENE-9950

… and more

- New `getTopDims` and `getAllChildren` APIs across all faceting implementations
  - Jira: LUCENE-10325, LUCENE-10550
- DrillSideways now works with concurrent search
  - Jira: LUCENE-9944
- More flexibility when indexing facet fields
  - Jira: LUCENE-9385

Part IV: Performance Improvements

Modernized Ordinal Encoding

- Taxonomy faceting (i.e., `FastTaxonomyFacetCounts`), needs to store category labels assigned to each document.
- Each unique category is assigned an ordinal. The ordinals are stored with the documents.

Map:

```
{
  “color/red”: 0,
  “color/blue”: 1,
  “size/small”: 2,
  “size/medium”: 3
}
```

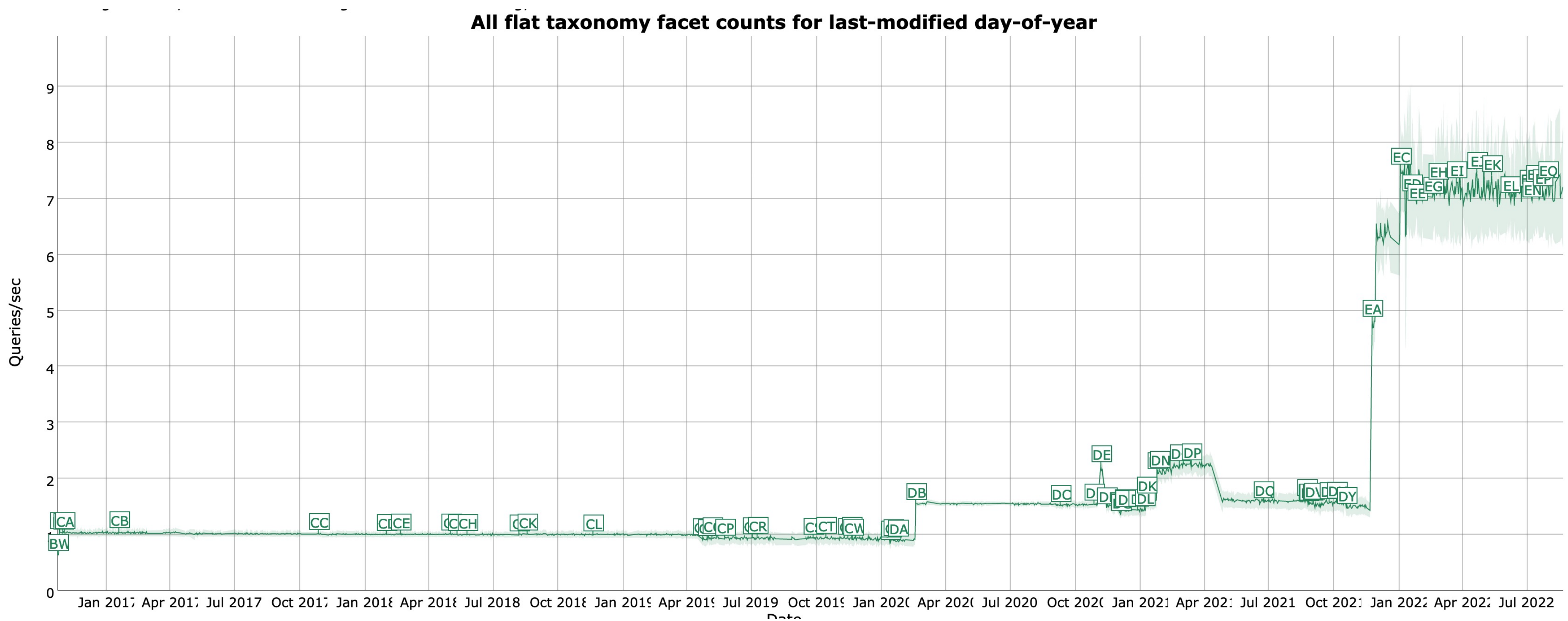
Documents:

```
{ $tfacets: [0, 2] }
{ $tfacets: [1, 3] }
```

Modernized Ordinal Encoding

- Old approach used custom var-int encoding (packed in a `BINARY` field)
- Ordinals are now stored with the same encoding as numeric doc value fields (i.e., `SORTED_NUMERIC`)
  - Uses a common bit-width for each value
- QPS improved by +438% (benchmark measurements)
- No longer need to maintain a custom encoding implementation
- Will continue to benefit from future numeric encoding improvements
- Jira: LUCENE-10062

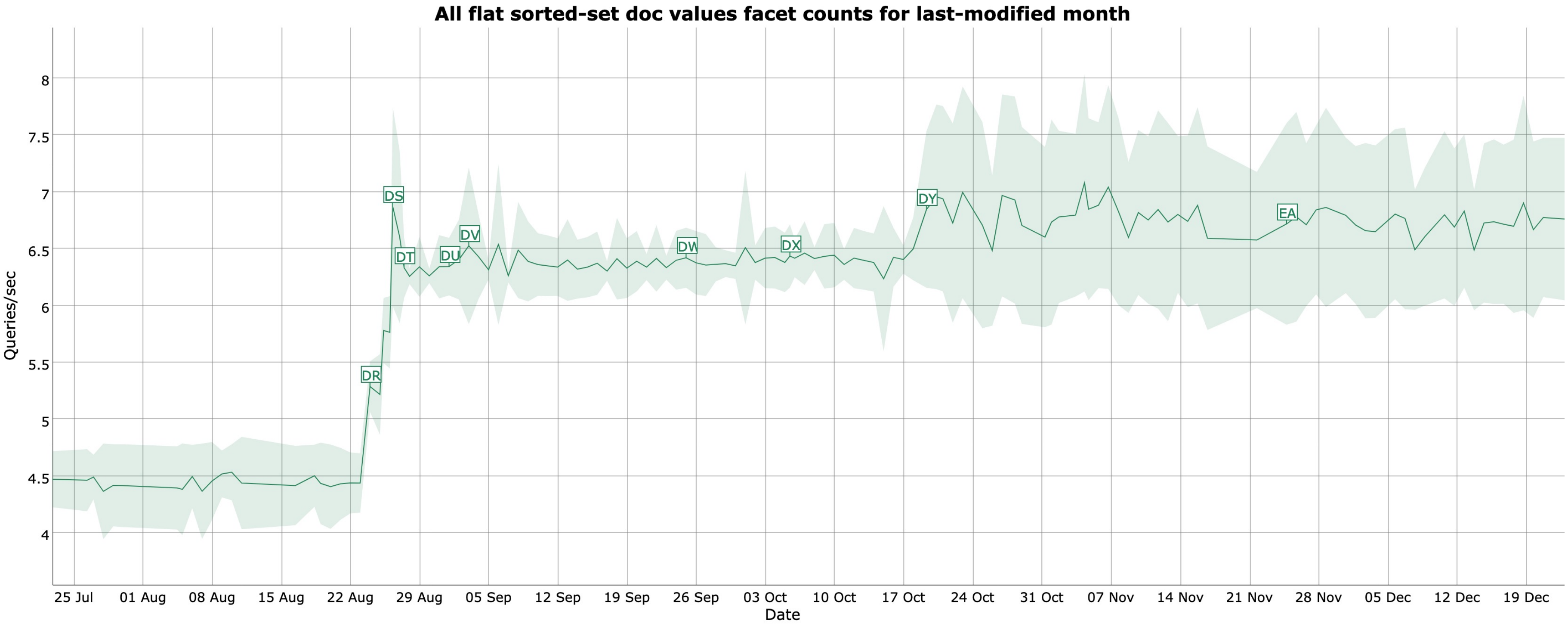
Modernized Ordinal Encoding



Special Handling for Single-Valued Fields

- Common implementation for single- and multi-valued fields
- Adding special handling for the single-valued case
- QPS improved by +45% (benchmark measurements)
- Idea had first been proposed in 2013. Sometimes there’s gold in old issues!
- Jira: LUCENE-5309

Special Handling for Single-Valued Fields



The list goes on…

- Taxonomy index modernization:
  - Migration from stored document fields to modern DocValue fields
    - Jira: LUCENE-9450, LUCENE-9476
  - Parent references migrated from “position” encoding to DocValues
    - Jira: LUCENE-10122
- Optimize away some null checks within a tight counting loop
  - Jira: LUCENE-10350
- Avoid some data structure indirection within a tight counting loop
  - Jira: LUCENE-10379
- FacetsCollector does not need scores in most cases
  - Jira: LUCENE-10481

Part V: So, What’s Next?

So, What’s Next?

- Exciting use-cases for FacetSets?
  - Likely to evolve
  - More optimization potential (e.g., more efficiently identifying matching facet sets)
- Can numeric range faceting automatically discover ranges that “fit” the data?
  - GH: #11028
- Converge on a single implementation for categorical faceting?
  - GH: #11717

Goals (revisited)

Everyone should walk away from this talk with…

1. A high-level understanding of Lucene's Faceting capabilities

2. A sense of the active development in Lucene's Faceting module

Questions?