
AWS
re:Invent

ANT403

# Advanced search with Amazon ES


Muthu Pitchaimani
Specialist Solution Architect
AWS

Shivang Shah
Principal Engineer
Intuit Inc.

Akbar Rangara
Staff Software Engineer
Intuit Inc.


AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Elasticsearch: Popular database engine

| Rank | | | DBMS | Database Model | Score | | |
|---|---|---|---|---|---|---|---|
| Oct 2020 | Sep 2020 | Oct 2019 | | | Oct 2020 | Sep 2020 | Oct 2019 |
| 1. | 1. | 1. | Elasticsearch | Search engine, Multi-model | 153.84 | +3.35 | +3.67 |
| 2. | 2. | 2. | Splunk | Search engine | 89.40 | +1.51 | +2.57 |
| 3. | 3. | 3. | Solr | Search engine | 52.48 | +0.86 | -5.09 |
| 4. | 4. | 4. | MarkLogic | Multi-model | 11.73 | -0.21 | -1.33 |
| 5. | 5. | ↑ 8. | Algolia | Search engine | 7.30 | +0.52 | +2.79 |

Source: DB-Engines Ranking - https://db-engines.com/en/ranking/search+engine

Open source

Search and analytics

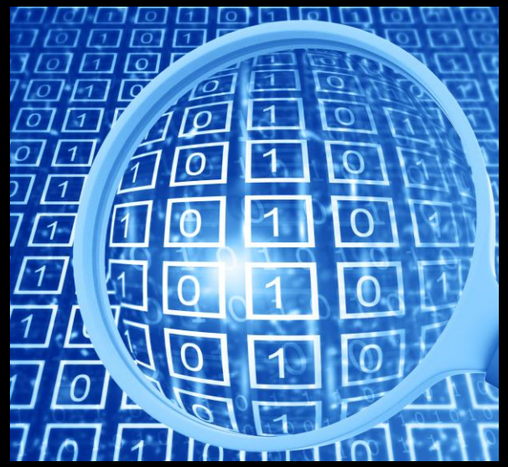
High performance distributed system

# Search and analytics



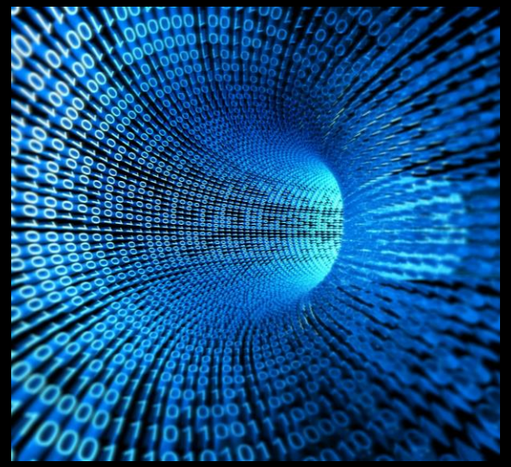

Text Search

Natural language
Scoring
Relevance



Similarity Search

Non-textual data
Vector space



Streaming data

High volume ingest
Near real-time




Analytics

Analysis at scale
Patterns and trends



## Amazon Elasticsearch Service (Amazon ES)

Fully managed service that makes it easy to deploy, manage, and run Elasticsearch cost effectively with industry-leading reliability, scalability, and security.



## Open Distro for Elasticsearch

100% Apache licensed, open-source distribution of Elasticsearch with plugins to deploy a secure, performant Elasticsearch cluster.



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Goals of search

You want to find products you want to buy

You want to find a movie you want to watch

You want to find a restaurant to order food

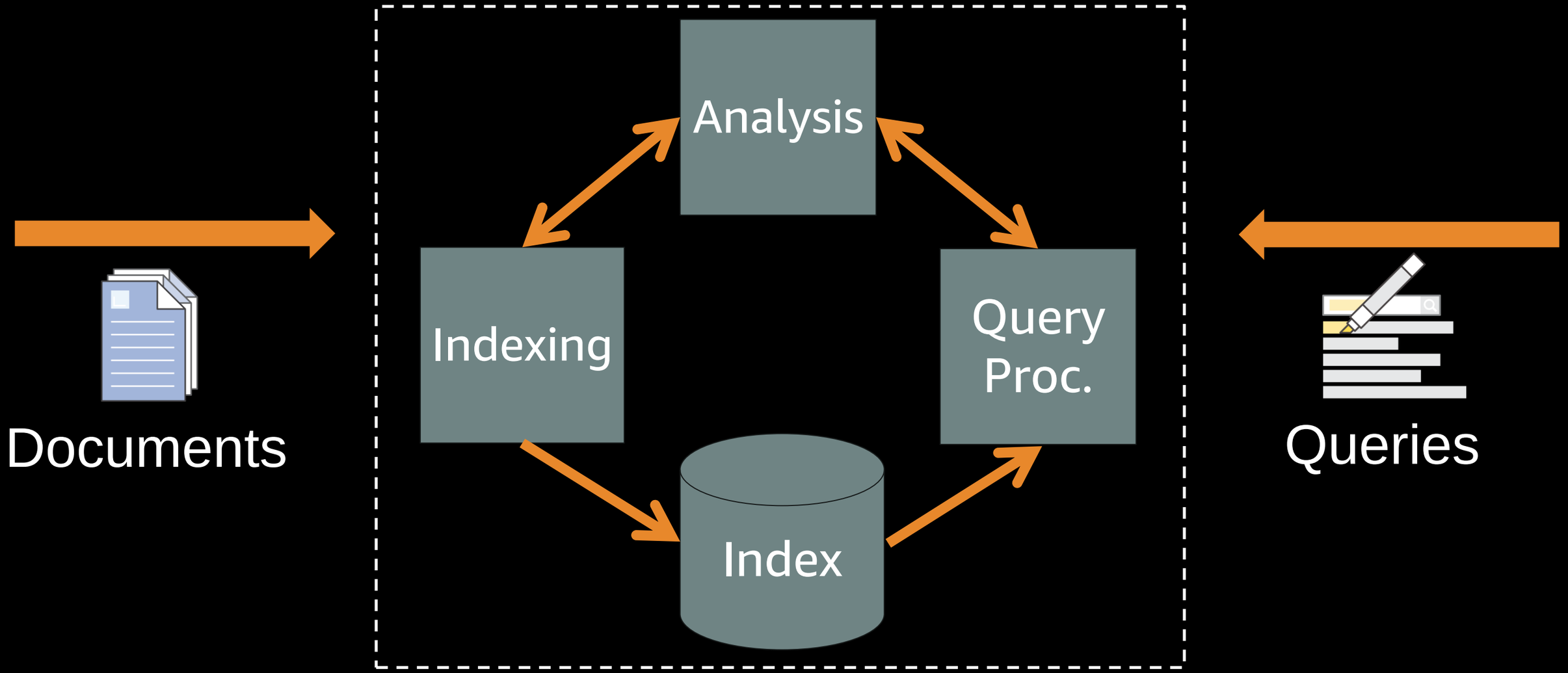
You need to find shared information in your organization

Your developers need to find out what’s wrong with a new feature

Your website is down and you need to diagnose and correct the problem

Search helps find pieces of relevant
information to solve the problem at hand

# Search architecture



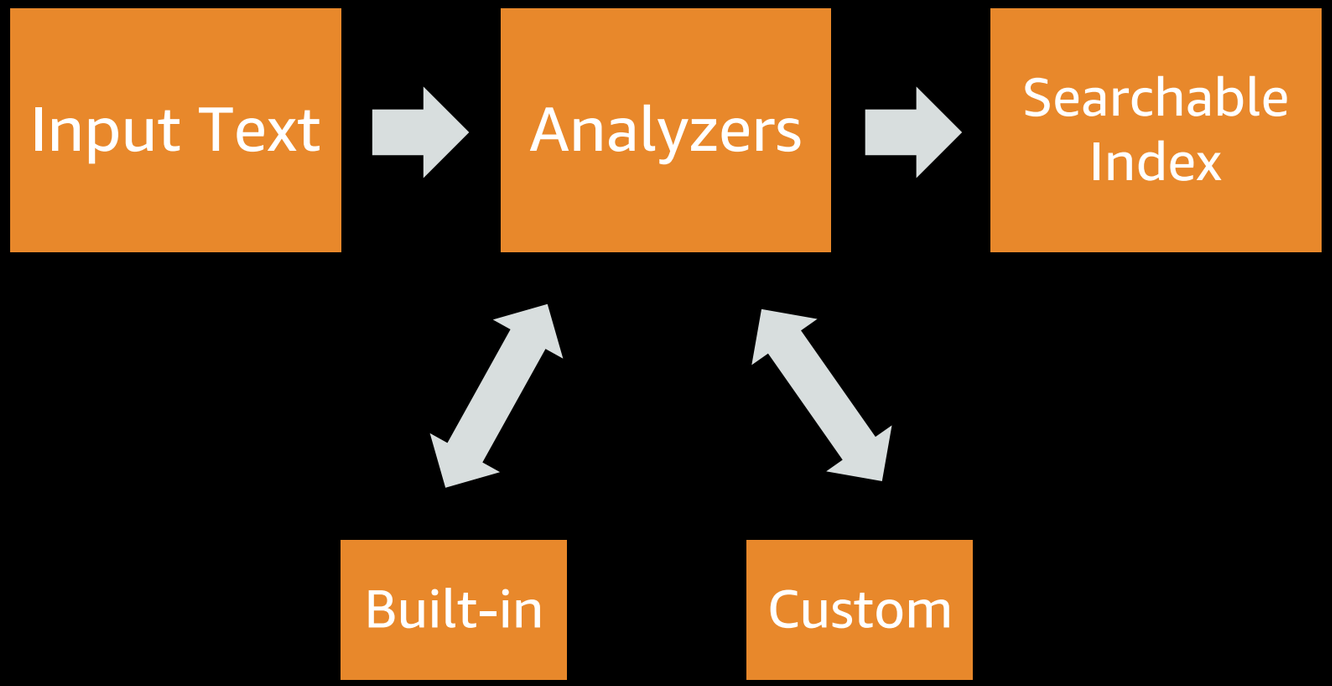
# Text search

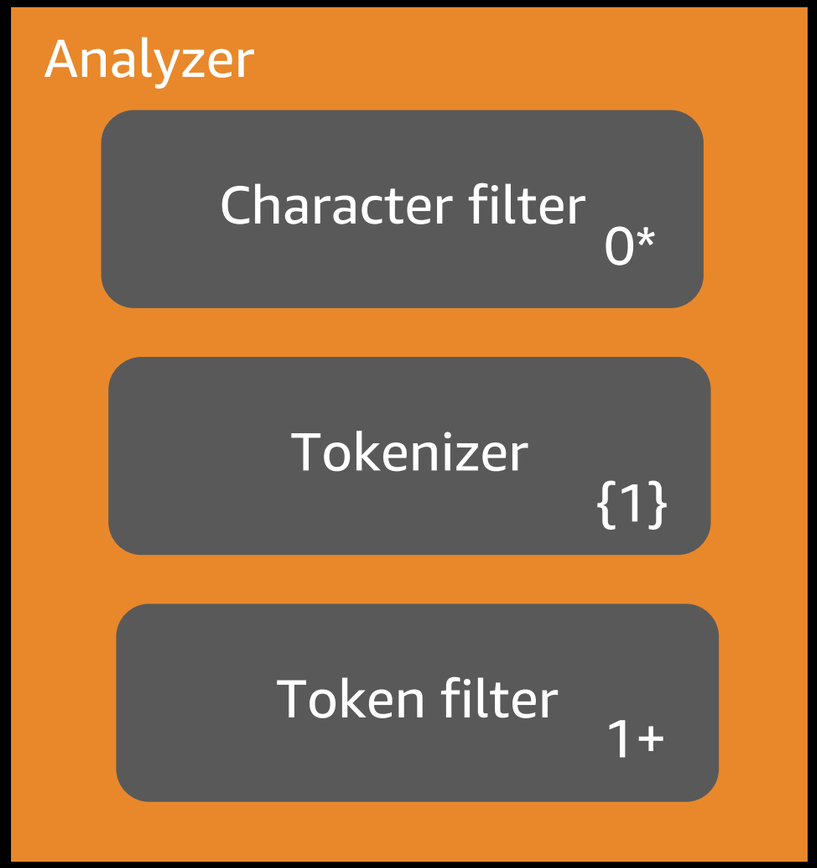




© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Analysis




Amazon Elasticsearch Service is a fully…





# Built-in analyzers

- HTML strip
- Mapping
- Pattern replace

Character filters

- Standard
- Whitespace
- Edge and n-gram
- Keyword
- …

Tokenizers

- Stemmer
- Stop
- Fingerprint
- Whitespace
- …

Token filters

Analyzers : Simple, standard, fingerprint …

# Standard analyzer

```
GET _analyze
{
  "analyzer": "standard",
  "text"    : "Amazon Elasticsearch Service is a fully managed service that
     makes it easy for you to deploy, secure, and run Elasticsearch cost
     effectively at scale. You can build, monitor, and troubleshoot your
     applications using the tools you love, at the scale you need. The service
     provides support for open source Elasticsearch APIs, managed Kibana,
     integration with Logstash and other AWS services, built-in alerting, SQL
     querying, k-NN search, anomaly detection and much more"
}
```

Character filters: , -

Tokenizer: Unicode Text Segmenter

Token filter: lowercase


```
{
  "tokens" : [
    {
      "token" : "amazon",
      "start_offset" : 0,
      "end_offset" : 6,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "elasticsearch",
      "start_offset" : 7,
      "end_offset" : 20,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "service",
      "start_offset" : 21,
      "end_offset" : 28,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
```

# Similarity search



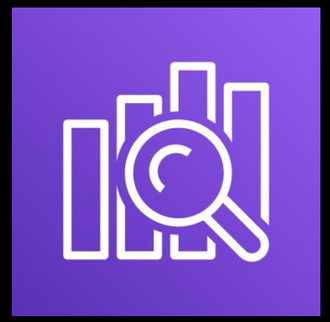


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# K – Nearest Neighbors



?







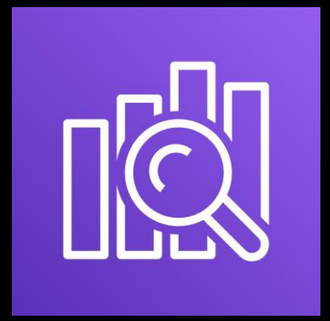
[123.4, 56,…]

[100, 25,…]

[99.8, 12,…]

finding words in a paragraph

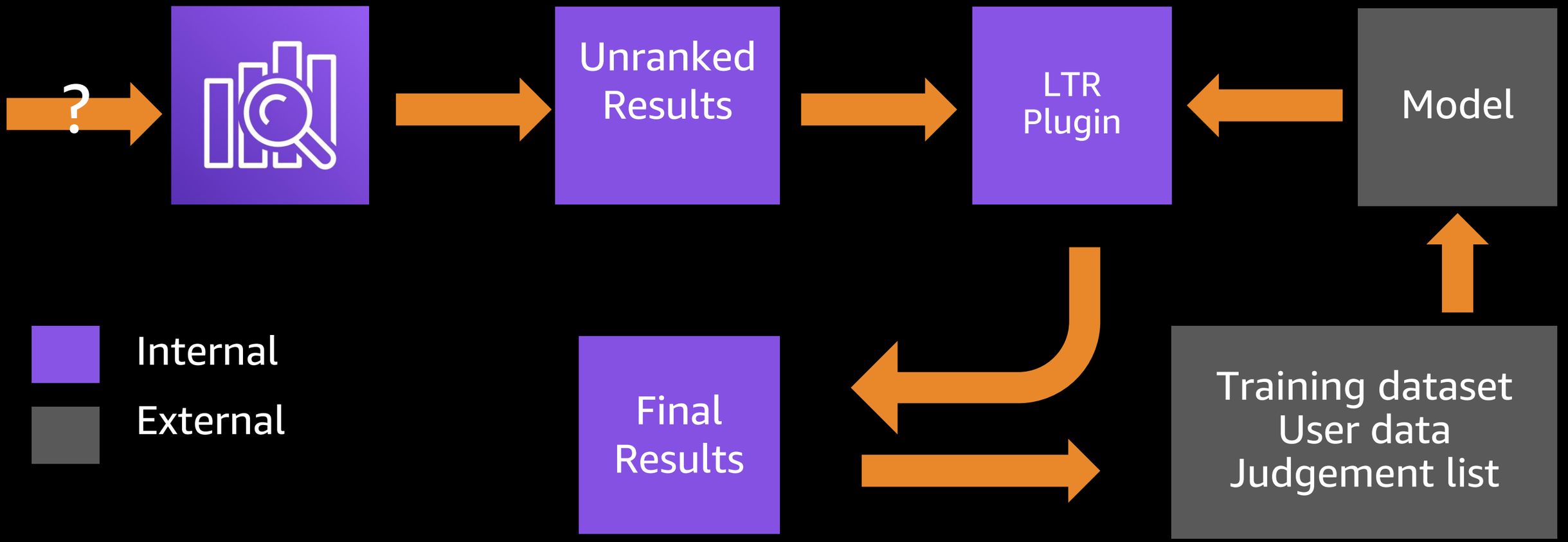
?



search for string in a text

- Data represented as set of numbers (vectors)
- Euclidean distance or cosine similarity
- Product recommendations, Semantic matching, fraud detection

# Relevance and ranking








AWS re:Invent

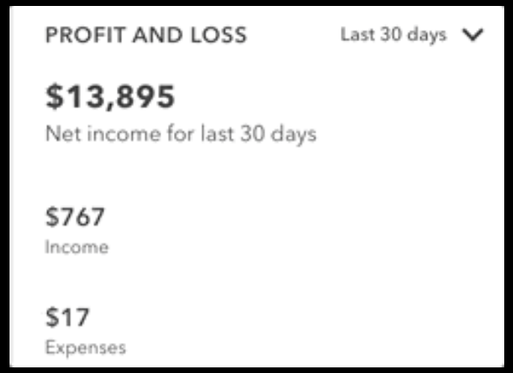
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

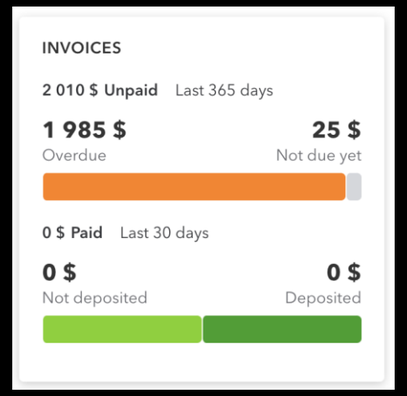

Who we serve:
Consumers
Small businesses
Self-employed
INTUIT
powering prosperity
turbotax
quickbooks
proconnect
mint

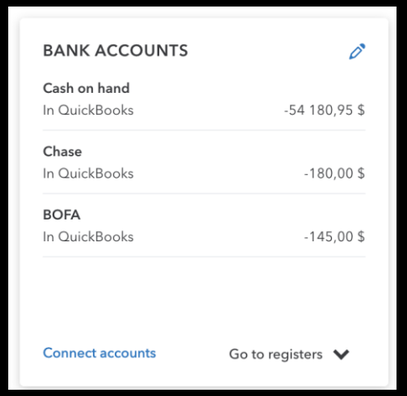
# What is QuickBooks Online?

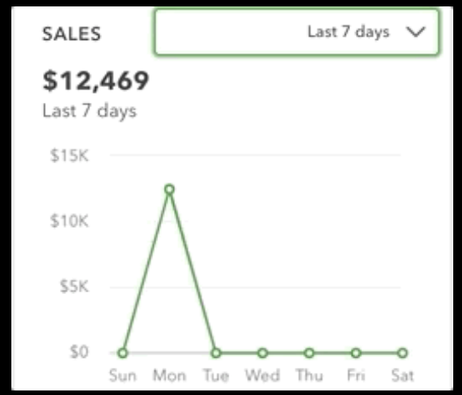
QuickBooks Online is a SaaS accounting offering by Intuit that helps small businesses

- Track their income, expenses & inventory
- Track and pay bills
- Prepare for taxes
- Gain deep business insights
- Automate workflows & integrations
- Connect to experts (accountants) when they need one



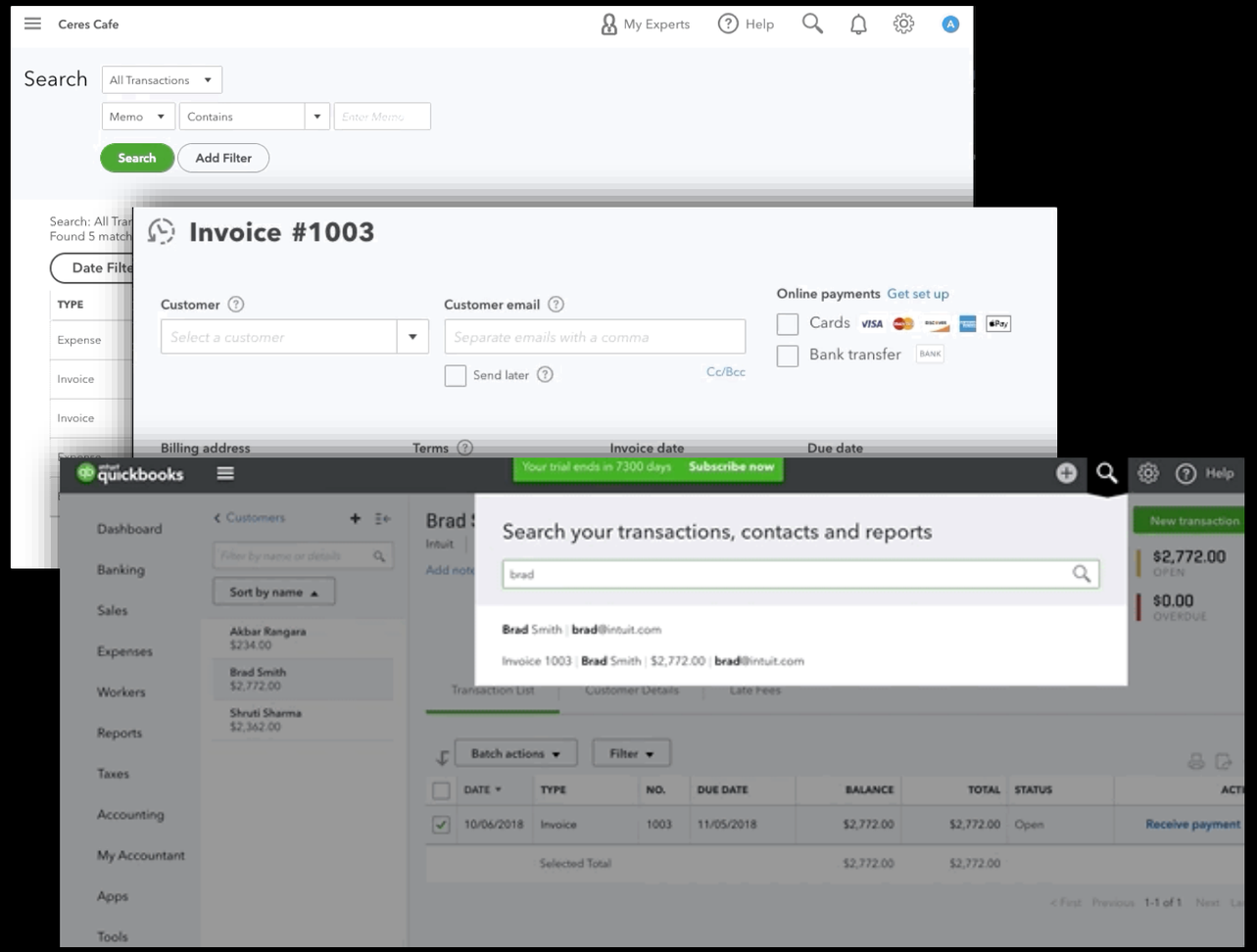




"Leveraging Intuit's AI-Driven Expert Platform to transform QuickBooks Ecosystem from the **source of books** to **source of business**"

# QuickBooks Online: In-product search

- Explicit Filter/Advanced Search
- Type-ahead (Quickfills) Search
- Type-ahead Global Search

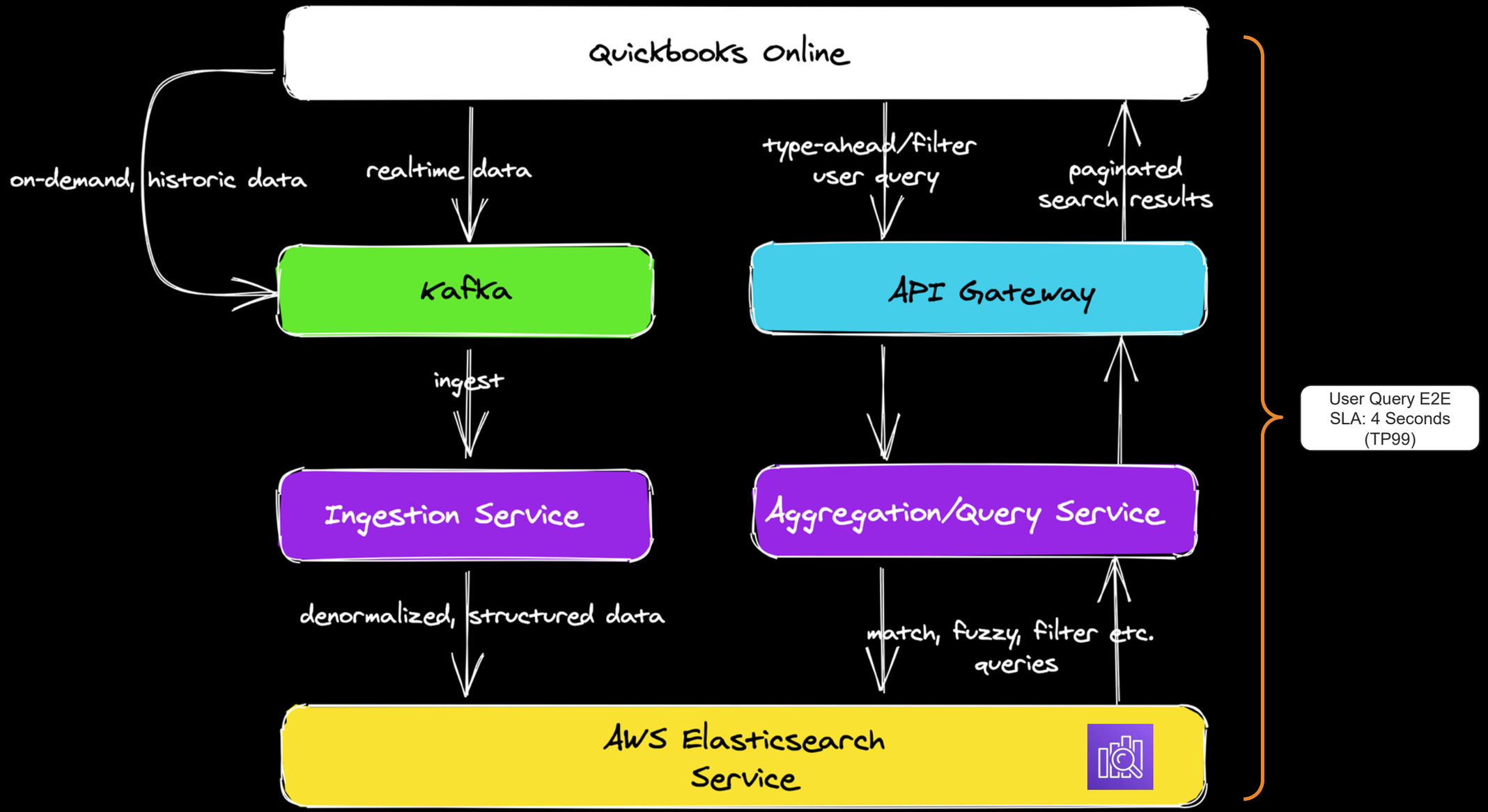


# QuickBooks Online

## SMB GROWTH & INFINITE DATA RETENTION



# High-level architecture



# Elasticsearch deep dive



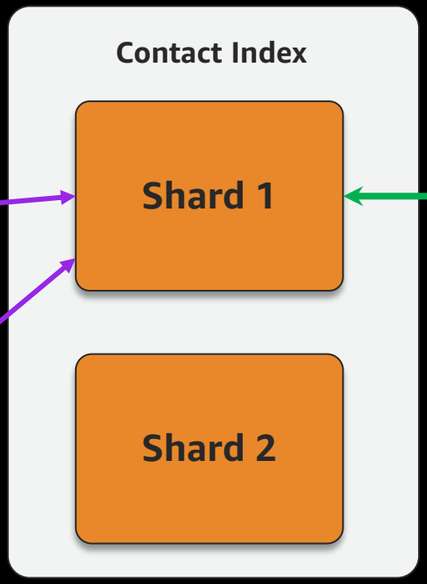
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

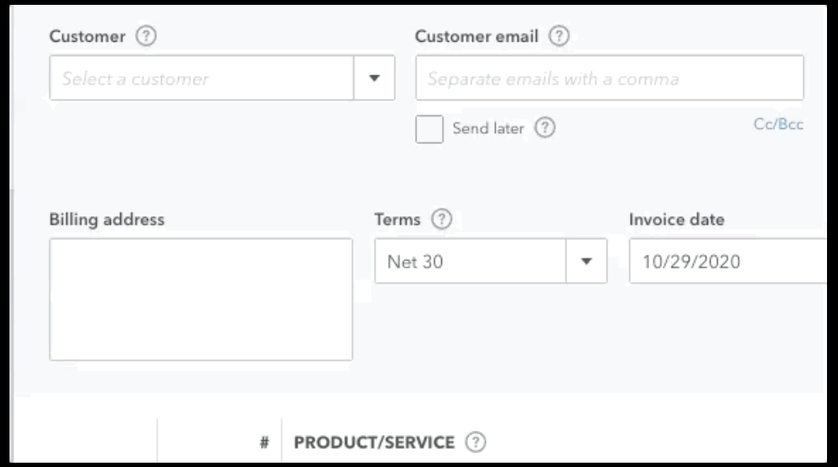
# Routing

```
PUT contact/_doc/1?routing=9130349480871766
{
     "name": "Intuit Inc@mtv20"
     "companyId": "9130349480871766"
}
```

```
PUT contact/_doc/2?routing=9130349480871766
{
  "name": "Intuit Inc@LA"
  "companyId": "9130349480871766"
}
```

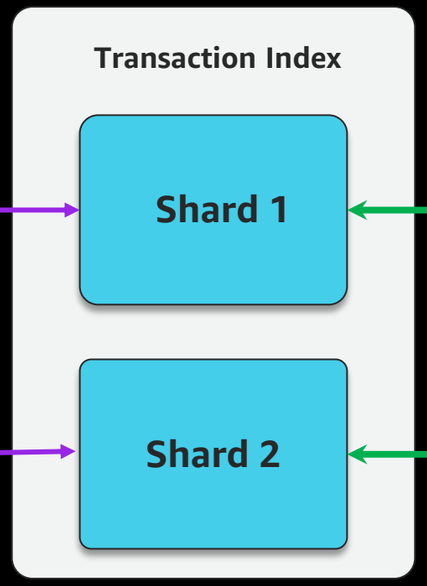



```
POST contact/_search?routing=9130349480871766
```

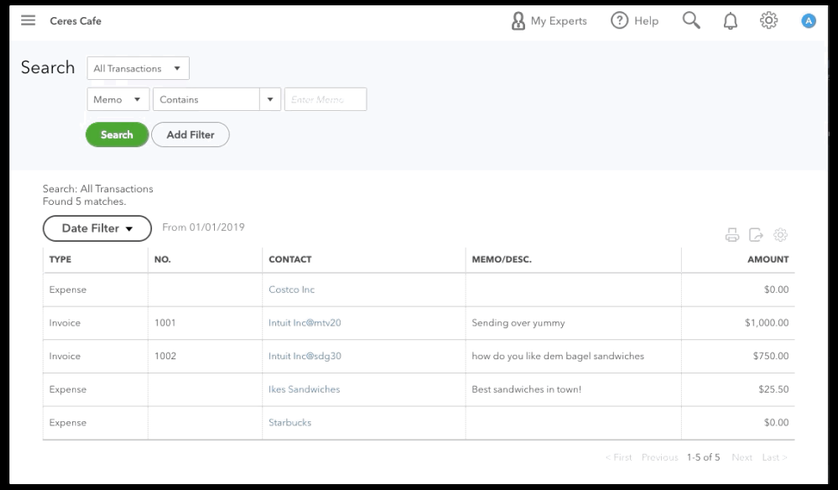



```
PUT transaction/_doc/1
{
     "memo": "Expense for furniture"
     "companyId": "9130349480871766"
}
```

```
PUT transaction/_doc/2
{
     "memo": "Invoice for bagels"
     "companyId": "9130349480871766"
}
```




```
POST transaction/_search
```



# Edge ngrams vs. ngrams

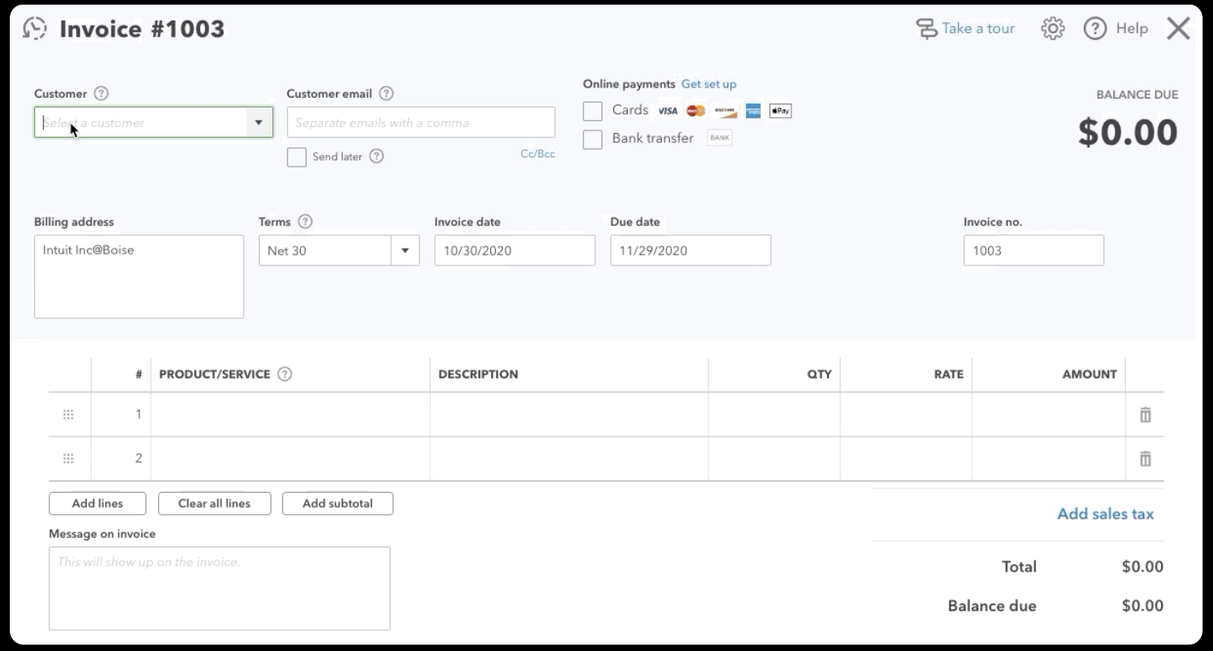
```
"50_edge_ngram": {
    "type": "edge_ngram",
    "min_gram": "1",
    "max_gram": "50",
    "language": "English",
    "token_chars": [
        "letter",
        "digit",
        "symbol",
        "punctuation"
    ]
```

i
in
inc
inc@
inc@m
inc@mt
inc@mtv
inc@mtv2
inc@mtv20

```
"50_ngram": {
    "type": "ngram",
    "min_gram": "1",
    "max_gram": "50",
    "language": "English",
    "token_chars": [
        "letter",
        "digit",
        "symbol",
        "punctuation"
    ]
}
```

i
in
inc
inc@
inc@m
inc@mt
inc@mtv
inc@mtv2
inc@mtv20

n
nc
nc@
nc@m
nc@mt
nc@mtv
nc@mtv2
nc@mtv20
…



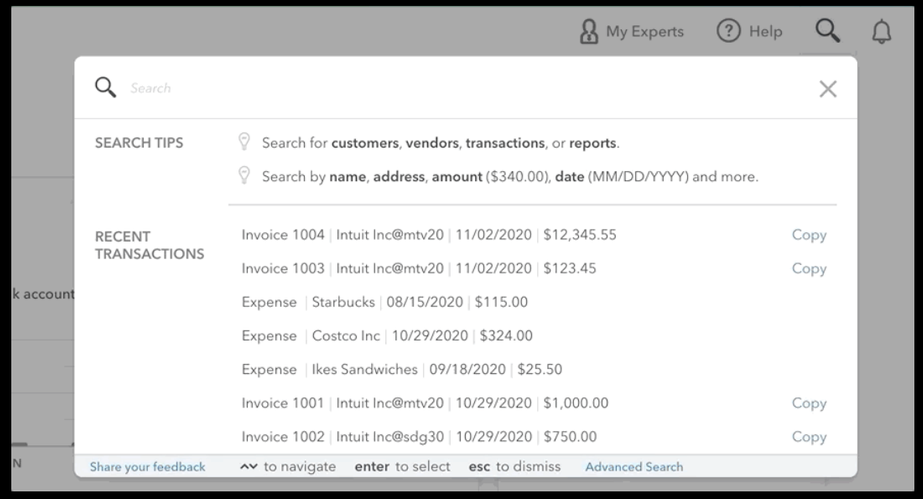
# Custom analyzers

```
"amount_dot_index_analyzer": {
  "type": "custom"
  "tokenizer": "50_edge_ngram",
  "filter": [
    "amount_char_dot_filter",
    "lowercase",
  "char_filter_empty_token_removal",
    "unique"
  ]
}
```

```
"amount_char_dot_filter": {
   "type": "pattern_replace",
   "pattern": "[^0-9\\.]",
   "replacement": ""
}
```

```
"amount_comma_index_analyzer": {
  "type"  "custom",
  "tokenizer": "50_edge_ngram",
  "filter": [
    "amount_char_comma_filter",
    "lowercase",
    "char_filter_empty_token_removal",
    "unique"
  ]
}
```

```
"amount_char_comma_filter": {
   "type": "pattern_replace",
   "pattern": "[^0-9\\,]",
   "replacement": ""
}
```




```
"phone_index_analyzer": {
    "type": "custom",
    "tokenizer": "50_edge_ngram",
    "filter": [
        "phone_char_index_filter",
        "lowercase",

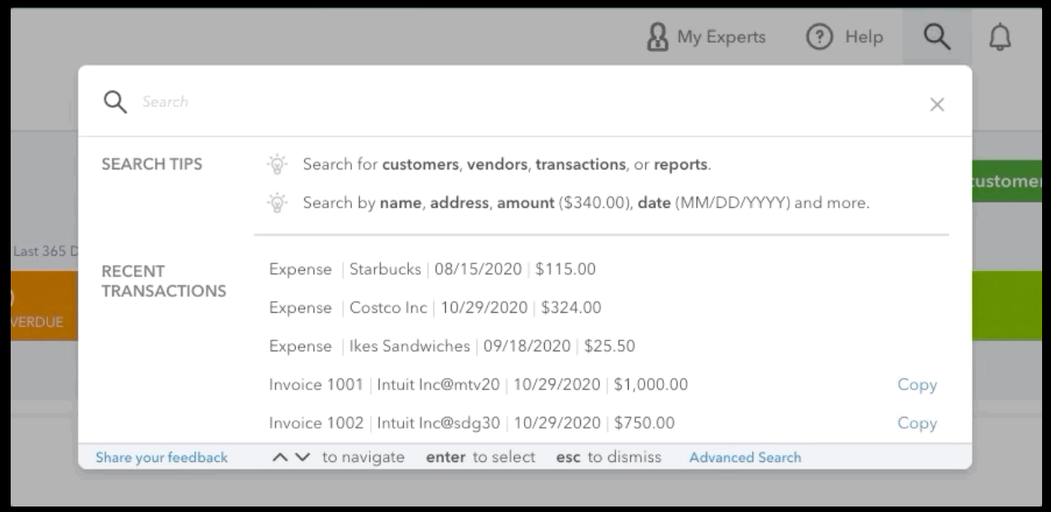
"char_filter_empty_token_removal",
        "unique"
    ]
}
```

```
"phone_char_index_filter": {
    "type":
"pattern_replace",
    "pattern": "[^0-9]",
    "replacement": ""
}
```

```
"phone_search_analyzer": {
    "type": "custom",
    "tokenizer": "whitespace",
    "filter": [
        "phone_char_search_filter",
        "lowercase",

"char_filter_empty_token_removal",
        "unique"
    ]
}
```

```
"phone_char_search_filter": {
    "type": "pattern_replace",
    "pattern": "[\\(\\)\\-\\+]",
    "replacement": ""
}
```



# Optimistic concurrency

Real Time Data

```
PUT transaction/_doc/1?version=22&version_type=external
{
     "title": "Transaction document 1"
     "companyId": "123456789"
}
```

Historic Data

```
PUT transaction/_doc/1?version=22&version_type=external
{
     "title": "Transaction document 2"
     "companyId": "123456789"
}
```



**concurrent update**
which data makes it?



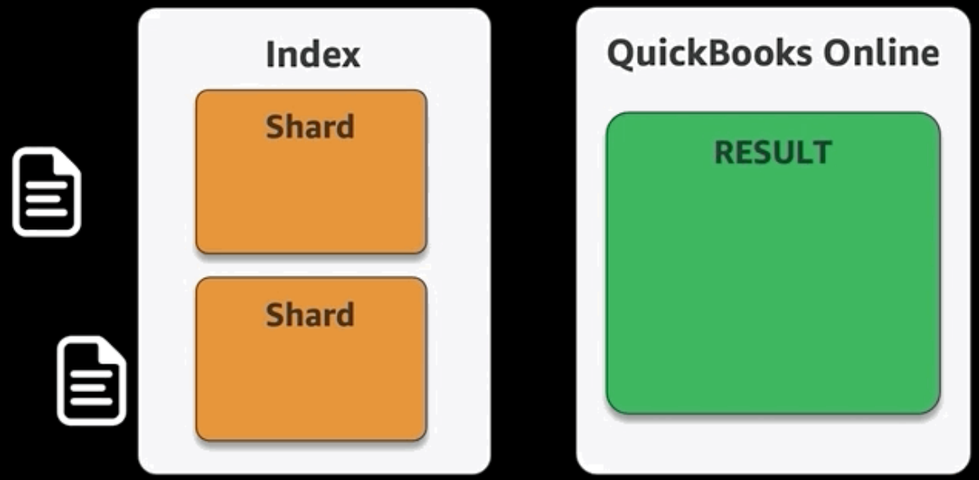
Amazon Elasticsearch Service

**version_type**
(Optional, enum) Specific version type: `internal`, `external`, `external_gte`.

# Takeaways

## Know your data and the use case

- Something as small as a misclassified tokenizer on an attribute can massively impact your storage costs and cause “noise” in your search relevancy (E.g., ngrams vs. edge-ngrams on large strings)
- Avoid multiple re-indexing of data by deeply understanding the use case, and prepare the indexing strategy for the future




## Use “routing” judiciously!

- Strike the right balance - Ingestion throughput vs. Query SLAs
- Choose a “routing key” that guarantees even distribution of data across shards

# Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



# Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.