OpenSearch

2024 - Mar

Cos’è OpenSearch

OpenSearch è una suite open-source di ricerca e analisi derivata da Elasticsearch 7.10.2 e Kibana 7.10.2, con licenza Apache 2.0

Offre funzionalità di ricerca a larga scala, full-text, distribuita, analitica e in tempo reale

OpenSearch si applica a vari casi d’uso come la ricerca sul web, la ricerca aziendale, l’intelligenza aziendale e l’analisi dei big data

OpenSearch vs. Elasticsearch

- Licenza
  - Apache 2.0 vs. SSPL
- Modello di Governance
  - Guidato da una comunità vs. Elastic
- Funzionalità
  - Divergono nel tempo

Amazon OpenSearch Service

- Servizio Managed
- Processo di Configurazione e Gestione semplificato
  - Backup
  - Patch
  - Scalabilità
- Integrazione con servizi AWS
  - IAM, KMS
- Disponibile anche in modalità serverless

Architettura

- Cluster e Nodi
- Indici e Documenti
- Nodi di coordinamento
- Sharding
- Motore di ricerca e Data Store
- Visualizzazione e UI

Indici e Mappatura dei Campi

Meccanismo di organizzazione dei dati per un loro recupero veloce

- La creazione è automatica quando si inserisce un documento ad un indice che non esiste
- OpenSearch richiede la presenza di un indice univoco di documento
  - Automatico o custom
- La richiesta viene inviata ad uno shard primario
  - Dopo la scrittura viene inviata agli shard di replica
    - È possibile specificare il minimo numero di shard disponibili (resilienza)
- Documenti master/detail
  - Conservati sullo stesso shard
  - Opzioni di routing come ElasticSearch

Creazione di Indici

`PUT <index_name>`

- Restrizioni
  - Lowercase
  - No simboli di punteggiatura
  - Non iniziano con underscore o trattino

Parametri di Query String

- `wait_for_active_shards`
- `cluster_manager_timeout`
- `timeout`

Request body

- `settings`

Creazione di Indici

```
PUT /sample-index1
{
  "settings": {
    "index": {
      "number_of_shards": 2,
      "number_of_replicas": 1
    }
  },
  "mappings": {
    "properties": {
      "age": {
        "type": "integer"
      }
    }
  },
```

Indici

Meccanismo di organizzazione dei dati per un loro recupero veloce

La creazione è automatica quando si inserisce un documento ad un indice che non esiste

Recupero dei Dati

```
GET movies/_doc/1


{
  "_index" : "movies",
  "_type" : "_doc",
  "_id" : "1",
  "_version" : 1,
  "_seq_no" : 0,
  "_primary_term" : 1,
  "found" : true,
  "_source" : {
    "title" : "Spirited Away"
  }
}
```

Recupero «bulk»

```
GET _mget
{
  "docs": [
    {
      "_index": "<index>",
      "_id": "<id>"
    },
    {
      "_index": "<index>",
      "_id": "<id>"
    }
  ]
}
```

Gestione degli Indici

Index template

- Modello da usare per indici che soddisfino ad una determinata condizione

Vantaggi

- Impostazione standard per indici che contengono determinate tipologie di documenti

Index alias

- Indice «virtuale» che punta a uno o più indici

Gestione degli Indici

Manutenzione

- Periodicamente potrebbe essere necessario effettuare operazioni di manutenzione

Index State Management

- Plugin che automatizza le operazioni amministrative attraverso la configurazione di apposite policies

Gestione degli Indici

Policies

- Documenti JSON che definiscono
  - Stato
    - Es. tipologia di accesso
  - Azioni
    - Es. esecuzione di rollover
  - Condizioni

Ingest Pipeline

Sequenza di processori applicati ai documenti man mano che vengono inseriti in un indice

- Si tratta di una **CoR** in cui ogni processore applica una determinata operazione sui dati

Ingest Pipeline

```
PUT _ingest/pipeline/my-pipeline
{
  "description": "This pipeline processes student data",
  "processors": [
    {
      "set": {
        "description": "Sets the graduation year to 2023",
        "field": "grad_year",
        "value": 2023
      }
    },
    {
      "set": {
        "description": "Sets graduated to true",
        "field": "graduated",
        "value": true
      }
    },
    {
      "uppercase": {
        "field": "name"
      }
    }
  ]
```

Queries

Queries

Leaf queries

- Full-text
- Term-level
- Geographic e xy
- Joining
- Span
- Specialized

Queries composte

Queries

Clausole eseguite in contesto

- di filtro
  - Quando si intende ottenere un risultato basato su un confronto di tipo booleano (dentro/fuori)
- di query
  - Quando si intende ottenere un risultato «valutato» su un punteggio di pertinenza
    - numero positivo a virgola mobile registrato nel campo dei metadati per ogni documento nella proprietà _score

Filter Context

Una clausola in un contesto di filtro risponde ad una domanda di tipo booleano che servono per inserire i documenti che soddisfano il filtro nel risultato finale

```
GET students/_search
{
  "query": {
    "bool": {
      "filter": [
        { "term":  { "honors": true }},
        { "range": { "graduation_year": { "gte": 2020, "lte": 2022 }}}
      ]
    }
  }
}
```

Query Context

La richiesta non ha una risposta binaria, ma uno score che rappresenta il punteggio di pertinenza con la domanda

Utile per cercare in contesti full-text parole flesse o sinonimi

Queries

Term-level

- Ricerca di documenti che contengono un termine esatto
- Risultati sulla base della rilevanza
- Con dati di testo sono utilizzabili sono campi mappati come keyword

Term-Level

- `term`
  - Cerca i documenti contenenti un termine esatto in un campo specifico
- `terms`
  - Cerca i documenti contenenti uno o più termini in un campo specifico
- `terms_set`
  - Cerca i documenti che corrispondono a un numero minimo di termini in un campo specifico
- `ids`
  - Cerca i documenti in base all'ID documento
- `range`
  - Cerca i documenti con valori di campo in un intervallo specifico
- `prefix`
  - Cerca i documenti contenenti termini che iniziano con un prefisso specifico
- `exists`
  - Cerca i documenti con qualsiasi valore indicizzato in un campo specifico
- `fuzzy`
  - Cerca i documenti contenenti termini simili al termine di ricerca entro la distanza massima consentita di Levenshtein
- `wildcard`
  - Cerca i documenti contenenti termini che corrispondono a un modello di caratteri jolly
- `regexp`
  - Cerca i documenti contenenti termini che corrispondono a un'espressione regolare

Aggregazioni

Le aggregazioni consentono di attingere al potente motore di analisi di OpenSearch per analizzare i dati ed estrarne statistiche

I casi d'uso delle aggregazioni variano dall'analisi dei dati in tempo reale per intraprendere un'azione all'utilizzo di OpenSearch Dashboards per creare un dashboard di visualizzazione

OpenSearch è in grado di eseguire aggregazioni su set di dati di grandi dimensioni in pochi millisecondi

- Rispetto alle query, le aggregazioni consumano più cicli di CPU e memoria

Aggregazioni

Elemento fondamentale per analisi

Aggregazioni su campi di testo non sono supportate

- Nel caso in cui sia necessario aggregare, solitamente si mantiene un campo copia di tipo keyword

```
GET _search
{
   "size": 0,
   "aggs": {
      "NAME": {
         "AGG_TYPE": {}
      }
   }
}
```

Aggregazioni

Metriche

- Single-value
- Multi-value

Bucket

Pipeline

Aggregazioni

- Aggregazioni metriche
  - Single-value
    - singola metrica

      sum, min, max, avg, cardinality, value_count
  - Multi-value
    - più metriche

      stats, extended_stats, matrix_stats, percentile, percentile_ranks, geo_bound, top_hits, scripted_metric

Aggregazioni

Aggregazioni Metriche

- Average
- Cardinality
- Extended stats
- Geobounds
- Matrix stats
- Maximum
- Minimum
- Percentile ranks
- Percentile
- Scripted metric
- Stats
- Sum
- Top hits
- Value count

Aggregazioni

```
GET opensearch_dashboards_sample_data_ecommerce/_search
{
  "size": 0,
  "aggs": {
    "avg_taxful_total_price": {
      "avg": {
        "field": "taxful_total_price"
      }
    }
  }
}
```


```
{
  "took": 85,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 4675,
      "relation": "eq"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "avg_taxful_total_price": {
      "value": 75.05542864304813
    }
  }
}
```

```
...
"aggregations" : {
  "extended_stats_taxful_total_price" : {
    "count" : 4675,
    "min" : 6.98828125,
    "max" : 2250.0,
    "avg" : 75.05542864304813,
    "sum" : 350884.12890625,
    "sum_of_squares" : 3.9367749294174194E7,
    "variance" : 2787.59157113862,
    "variance_population" : 2787.59157113862,
    "variance_sampling" : 2788.187974983536,
    "std_deviation" : 52.79764740155209,
    "std_deviation_population" : 52.79764740155209,
    "std_deviation_sampling" : 52.80329511482722,
    "std_deviation_bounds" : {
      "upper" : 180.6507234461523,
      "lower" : -30.53986616005605,
      "upper_population" : 180.6507234461523,
      "lower_population" : -30.53986616005605,
      "upper_sampling" : 180.66201887270256,
      "lower_sampling" : -30.551161586606312
    }
  }
 }
}
```

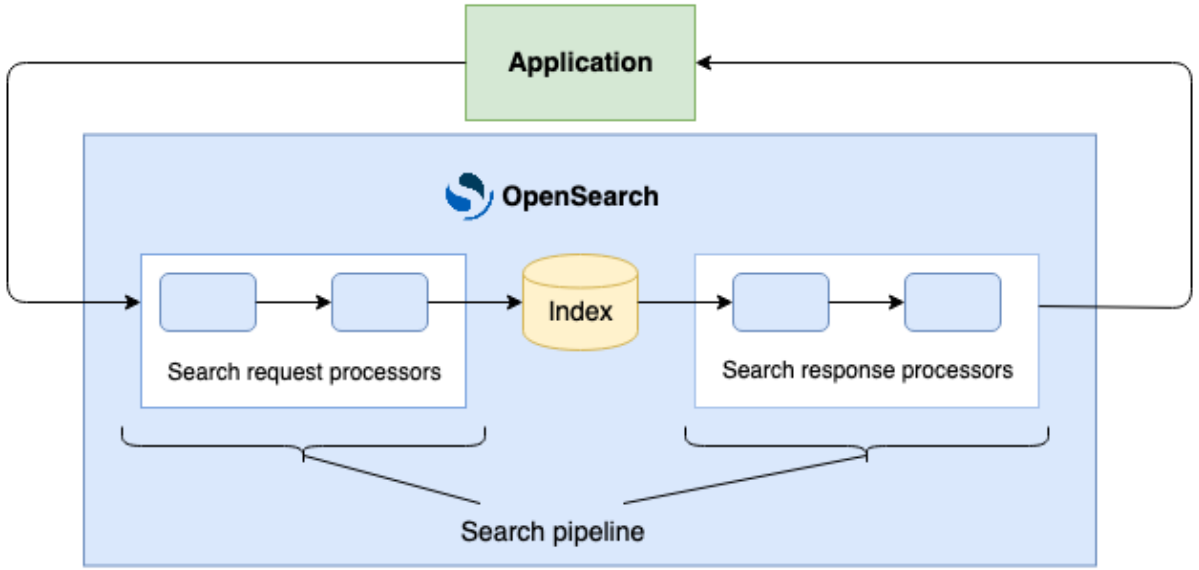
```
GET opensearch_dashboards_sample_data_ecommerce/_search
{
  "size": 0,
  "aggs": {
    "extended_stats_taxful_total_price": {
      "extended_stats": {
        "field": "taxful_total_price"
      }
    }
  }
}
```

Aggregazioni

Pipeline di Ricerca

- Ristrutturazione di query
- Semplificazione dei risultati di ricerca
  - Fanno uso di processori
    - Ognuno dei quali rappresenta un’attività modulare
    - La modularizzazione consente di riorganizzare in maniera semplice la query

Pipeline di Ricerca



Componenti:

- Processore di richiesta
- Processore di risposta
- Processore di risultato
- Processore

Sia l'elaborazione della richiesta che quella della risposta per la pipeline vengono eseguite nel nodo coordinatore, quindi non esiste alcuna elaborazione a livello di shard

Pipeline di Ricerca

Per usare una pipeline con una query occorre specificare il nome della pipeline nel parametro query search_pipeline:

```
GET /my_index/_search?search_pipeline=my_pipeline
```

```
PUT /_search/pipeline/my_pipeline
{
  "request_processors": [
    {
      "filter_query" : {
        "tag" : "tag1",
        "description" : "This processor is going to restrict to publicly visible documents",
        "query" : {
          "term": {
            "visibility": "public"
          }
        }
      }
    }
  ],
  "response_processors": [
    {
      "rename_field": {
        "field": "message",
        "target_field": "notification"
      }
    }
  ]
}
```

Pipeline di Ricerca

Processori di Request

- `filter_query`
  - Aggiunge una query di filtro
- `neural_query_enricher`
  - Imposta un modello predefinito per la ricerca neurale
- `script`
  - Aggiunge uno script che viene eseguito sui documenti indicizzati
- `oversample`
  - Aumenta il valore del parametro size

Pipeline di Ricerca

```
PUT /_search/pipeline/my_pipeline
{
  "request_processors": [
    {
      "filter_query" : {
        "tag" : "tag1",
        "description" : "This processor is going to restrict to publicly visible documents",
        "query" : {
          "term": {
            "visibility": "public"
          }
        }
      }
    }
  ]
}
```

Filter Query Processor

Intercetta una richiesta di ricerca e applica una query aggiuntiva alla richiesta filtrando i risultati

- Utile quando occorre applicare alla query un filtro senza scriverlo direttamente nella query

Pipeline di Ricerca

```
PUT /_search/pipeline/my_pipeline
{
  "request_processors": [
    {
      "oversample" : {
        "tag" : "oversample_1",
        "description" : "This processor will multiply `size` by 1.5.",
        "sample_factor" : 1.5
      }
    }
  ]
}
```

Oversample Processor

Moltiplica il parametro size della richiesta (memorizzato in `original_size`) per un valore specificato

Pipeline di Ricerca

Processori di Response

- `rename_field`
  - Rinomina un campo esistente
- `rerank`
  - Riclassifica i risultati
- `collapse`
  - Raccoglie i risultati
- `truncate_hits`
  - Ignora gli hit di ricerca dopo il raggiungimento di risultati specificato

Pipeline di Ricerca

```
PUT /_search/pipeline/my_pipeline
{
  "response_processors": [
    {
      "rename_field": {
        "field": "message",
        "target_field": "notification"
      }
    }
  ]
}
```

Incremento di Performance in Ricerca

Ricerca Asincrona

La ricerca di grandi volumi di dati può richiedere molto tempo, soprattutto se si esegue la ricerca in nodi caldi o in più cluster remoti

- La ricerca asincrona in OpenSearch consente di inviare richieste di ricerca eseguite in background
- È possibile monitorare lo stato di avanzamento di queste ricerche e ottenere risultati parziali non appena diventano disponibili
- Al termine della ricerca, è possibile salvare i risultati per esaminarli in un secondo momento

Incremento di Performance in Ricerca

Ricerca simultanea di segmenti
Concurrent Segment Search

La ricerca simultanea lavora su segmenti in parallelo durante la fase di query

Utile:

- quando si inviano richieste a esecuzione prolungata
  - ad esempio richieste che contengono aggregazioni o intervalli di grandi dimensioni
  - in alternativa all'unione forzata dei segmenti in un unico segmento al fine di migliorare le prestazioni

SQL e PPL

Oltre al DSL, è possibile usare

- SQL
- PPL

anche in Dashboard

SQL

SQL colma il divario tra i tradizionali concetti di database relazionali e la flessibilità dell'archiviazione dei dati orientata ai documenti di OpenSearch

- Questa integrazione dà la possibilità di utilizzare le conoscenze SQL per interrogare, analizzare ed estrarre informazioni dai dati OpenSearch

SQL

```
FROM index
 WHERE predicates
  GROUP BY expressions
   HAVING predicates
    SELECT expressions
     ORDER BY expressions
      LIMIT size
```

```
SELECT [DISTINCT] (* | expression) [[AS] alias] [, ...]
FROM index_name
[WHERE predicates]
[GROUP BY expression [, ...]
 [HAVING predicates]]
[ORDER BY expression [IS [NOT] NULL] [ASC | DESC] [, ...]]
[LIMIT [offset, ] size]
```

Ordine di Esecuzione

SQL

Join

Sono supportati

- inner join
- cross join
- left outer join

I join hanno una serie di restrizioni

- È possibile mettere in join solo 2 indice
- Occorre usare gli alias per gli indici
- All’interno di una clausola ON, è possibile usare solo condizioni AND
- In un’istruzione WHERE non è possibile combinare alberi che contengono indici multipli
- Non è supportato LIMIT con OFFSET
- Non è possibile usare GROUP BY o ORDER BY per i risultati

SQL

Funzioni

Il linguaggio SQL supporta tutte le funzioni comuni del plug-in SQL

- inclusa la ricerca per pertinenza

ma introduce anche alcuni sinonimi di funzione, che sono disponibili solo in SQL

Funzioni di Aggregazione

Le funzioni di aggregazione operano su sottoinsiemi definiti dalla clausola

- In assenza di una clausola, le funzioni di aggregazione operano su tutti gli elementi del set di risultati

PPL

Piped Processing Language (**PPL**) è un linguaggio di query incentrato sull'elaborazione dei dati in modo sequenziale e dettagliato

Utilizza l'operatore pipe (|) per combinare i comandi per trovare e recuperare i dati

- È il linguaggio principale utilizzato con l'osservabilità in OpenSearch e supporta le query multi-data

PPL

Ogni query PPL inizia con il comando `search`

- che specifica innanzitutto quale indice interrogare

Poiché non esistono, al momento, altri comandi, il comando `search` può essere omesso

```
search source=<index> [boolean-expression]
source=<index> [boolean-expression]
```

Dashboards

Con l’app Dashboards è possibile

- Visualizzare dati diversi in un’unica vista
- Creare visualizzazioni dinamiche
- Creare e condividere report

Dashboards

Panels

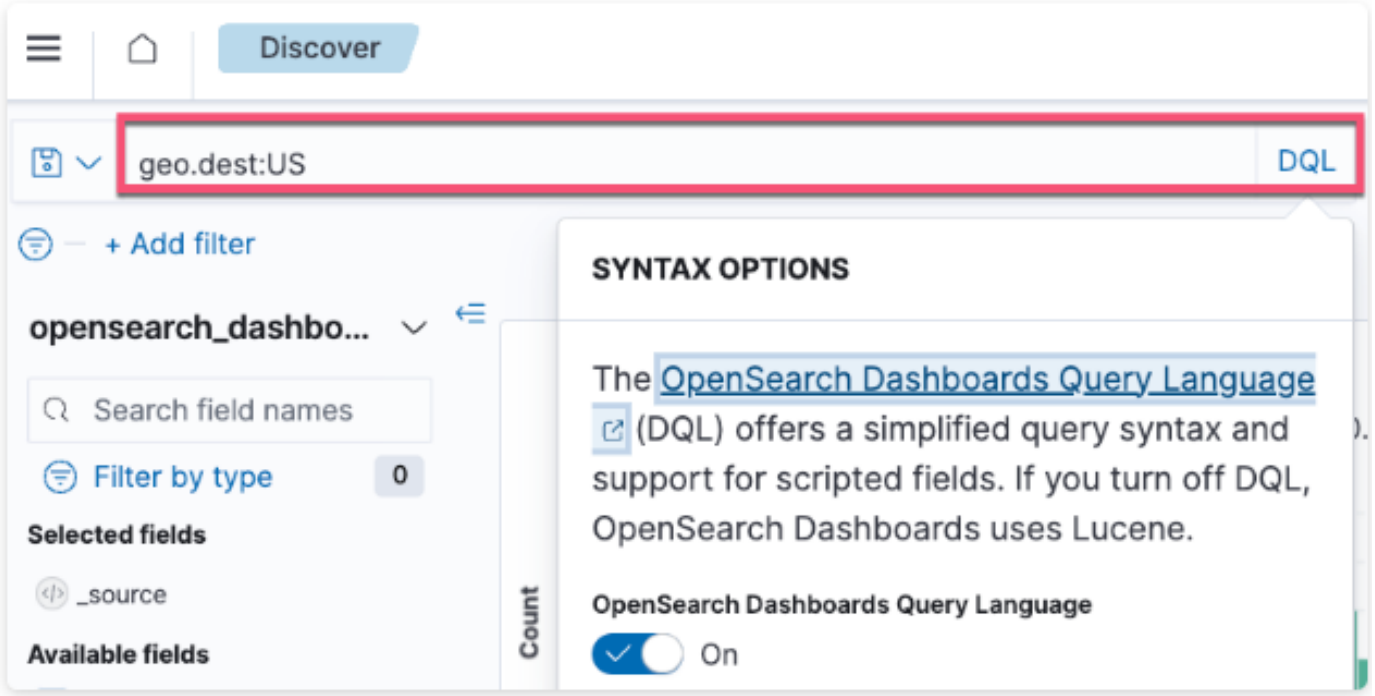
I pannelli sono contenitori di viste

Interazione

Consentono di analizzare i dati in modo più approfondito e filtrarli in diversi modi

Dashboards Query Language

Semplice linguaggio di query basato su testo utilizzato per filtrare i dati in OpenSearch Dashboards



DQL

Ricerca per termini

Ricerca in campi anche con wildcards o ranges

Supporta gli operatori booleani

Supporta il path per le proprietà interne agli oggetti

Data Visualization

products.price	Average products.base_price
0	39.52
300	562.5
900	562.5
	12,100

Query Workbench

Strumento all'interno di OpenSearch Dashboards

Utilizzato per:

- eseguire query SQL e PPL
- tradurre le query nelle chiamate API REST equivalenti
- visualizzare e salvare i risultati in diversi formati di risposta

Non supporta operazioni di `delete` e `update`

L’accesso ai dati è di tipo read-only