Elasticsearch - DIY

Dileep Gadiraju

What is Elasticsearch ?

“The world’s leading free and open search and analytics solution with an emphasis on speed, scale, and relevance. It's transforming how the world uses data”

Its built on Apache Lucene library and provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

What is Kibana ?

Kibana is an free and open frontend application that sits on top of the Elastic Stack, providing search and data visualization capabilities for data indexed in Elasticsearch

Why Elasticsearch ?

Elasticsearch

- Application Search
- Website search
- Enterprise search
- Logging and log analytics
- Infrastructure metrics and container monitoring
- Application performance monitoring
- Geospatial data analysis and visualization
- Security analytics
- Business analytics

Components

Elasticsearch

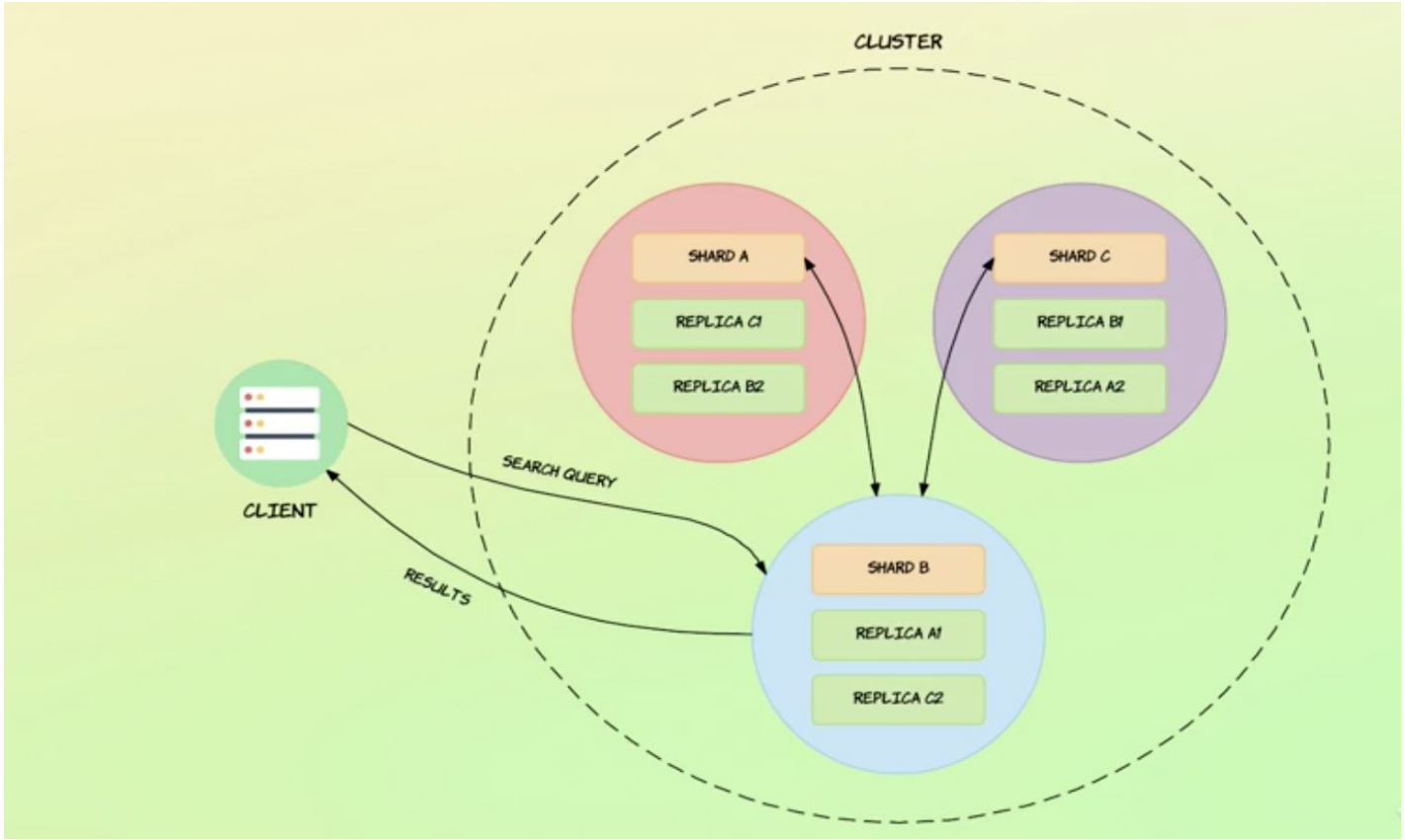
Index	An index is a logical namespace which maps to one or more primary shards and can have zero or more replica shards.
Type	Data types of fields stored within a index. Examples: date,long,keyword,**geo_point**,long etc.
Document	Multiple rows or records stored with in index.
Field	Each document is essentially a JSON structure, which is ultimately considered to be a series of key:value pairs
Mapping	Mapping is similar to database schemas that define the properties of each field in the index.
Node	server which stores a data and part of cluster. A special node called co-ordinating node receives client requests.
Cluster	Collection of Nodes (i.e server) , each node contain a parts of cluster data , being data added to cluster.
KQL	Kibana Query Language (KQL) is a simple syntax for filtering Elasticsearch data using free text search or field-based search.
Document Score	Process to determine the relevance of retrieved documents based on user queries, term frequencies, and other important parameters.

Components

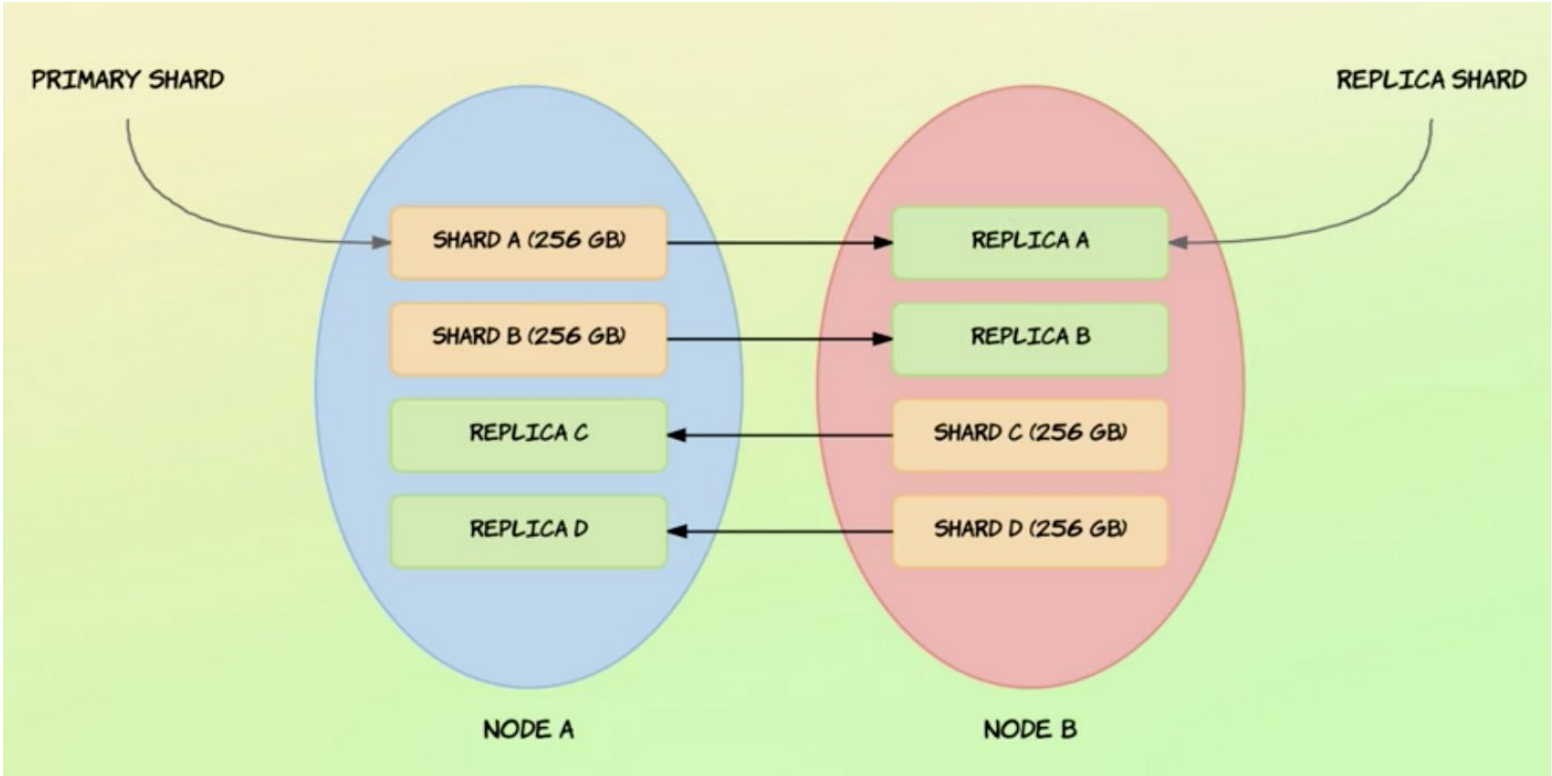
Elasticsearch

Shard	Shards are individual instances of a Lucene index. Each index is comprised of shards across multiple nodes. Types = Primary Shards , Replica Shards.
ILM	Index Lifecycle Management policies to manage indices according to your performance, resiliency, and retention requirements.
Index Template	index template is a way to tell Elasticsearch how to configure an index when it is created

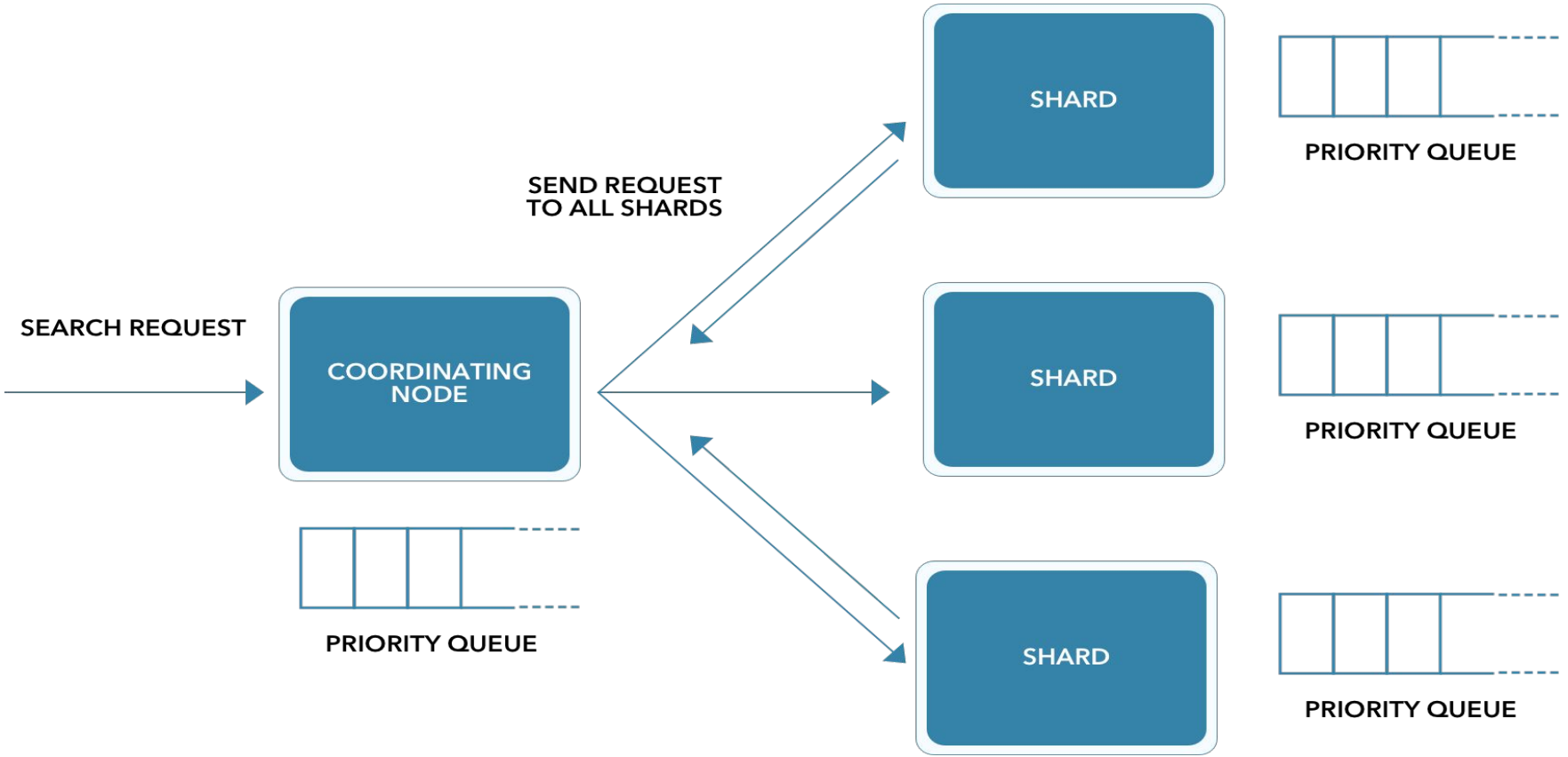
Elasticsearch Architecture



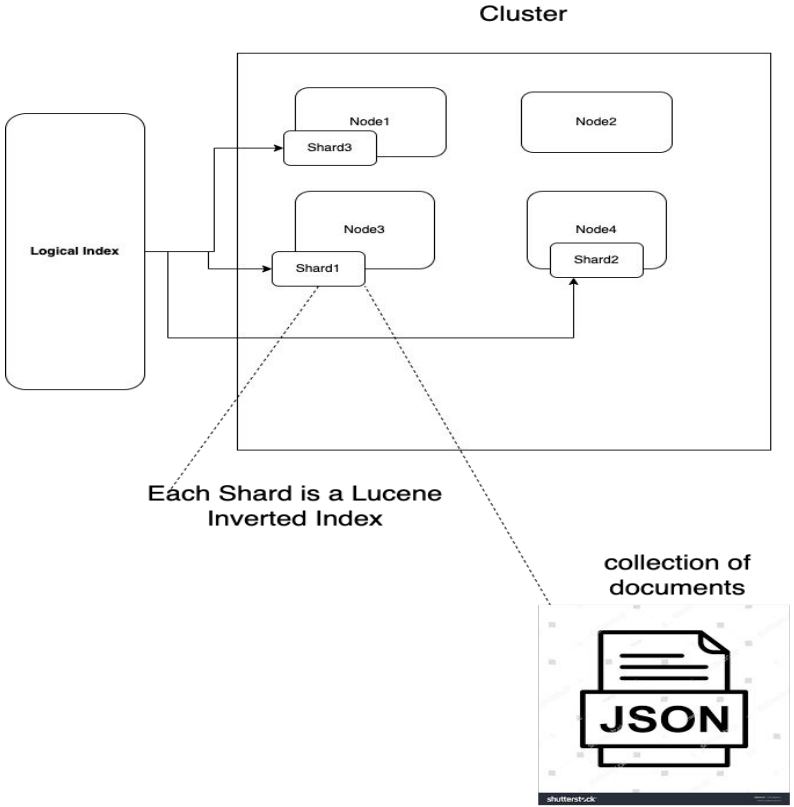
Elasticsearch Sharding



Elasticsearch Architecture



Elasticsearch - Discussion in the Training



Elasticsearch

What is Analyzer ?

Analyzers are the special algorithms that determine how a string field in a document is transformed into terms in an inverted index.

Types of Analyzers

- Standard Analyzer
- Simple Analyzer
- Whitespace Analyzer
- Stop Analyzer
- Keyword Analyzer
- Pattern Analyzer
- Language Analyzer
- Fingerprint Analyzer
- Custom Analyzer

Elasticsearch

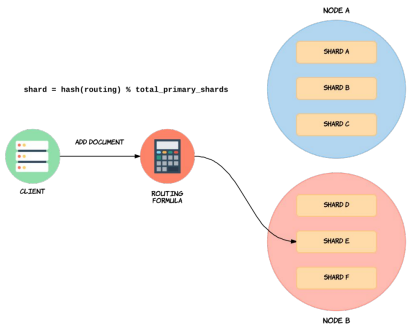
What is Inverted Index ?

Term	Document #1	Document #2
best	X	
carbonara		X
delicious		X
pasta	X	X
pesto	X	
recipe	X	X
the	X	
with	X	

Results of Analyzer’s processing stored as Inverted Index. Inverted index is a mapping between terms and which documents contain those terms. Search queries uses inverted indices to fetch document results.

Elasticsearch

What is Routing ?



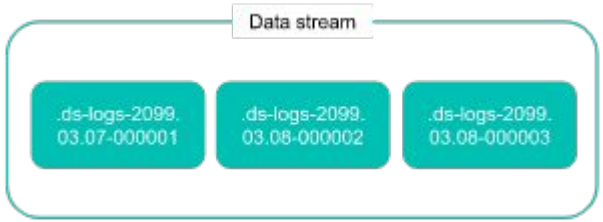
Process of determining which shard a document would reside on.

Default routing done based on document id to evenly distribute the documents.

Custom routing can help speed up the search process. For example use user_id or zip code as routing id.

Elasticsearch

What is Data Stream ?



A Data stream lets you store append-only time series data across multiple indices while giving you a single named resource for requests.

Data streams are well-suited for logs, events, metrics, and other continuously generated data.

Elasticsearch

What is Nested Type ?

Nested is a special type of object that is indexed as a separate document, and a reference to each of these inner documents is stored with the containing document, so we can query the data accordingly.

Settings:

index.mapping.nested_fields.limit

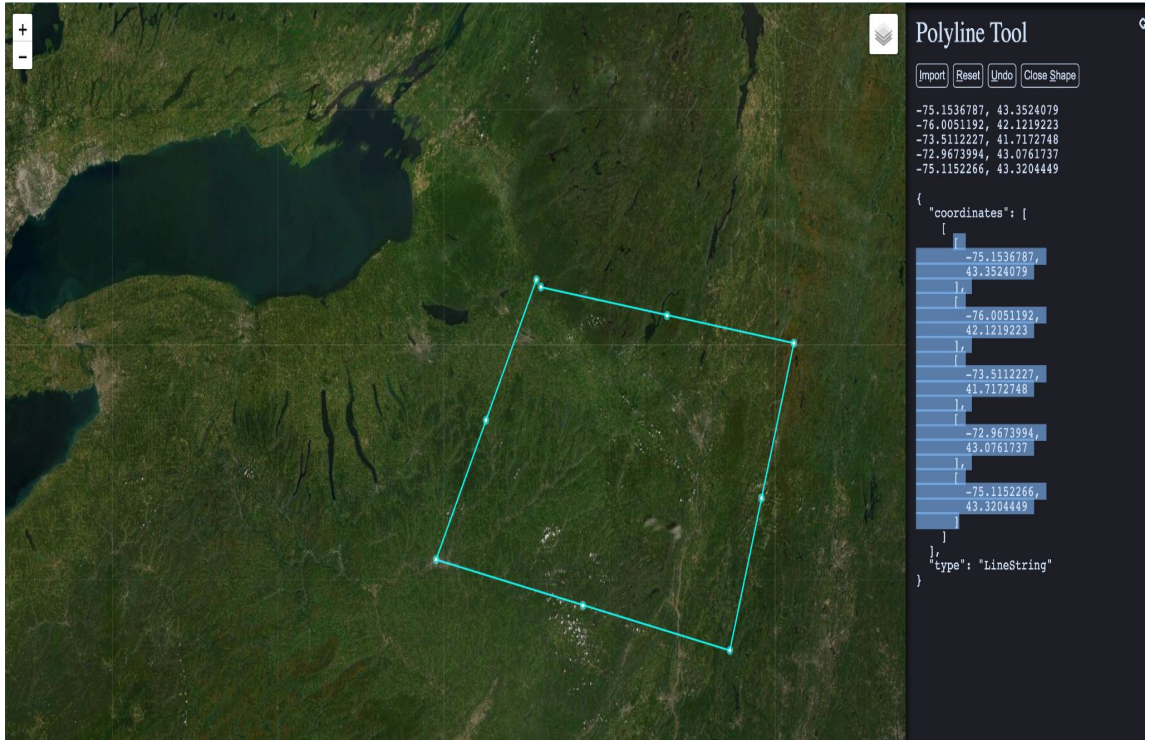
index.mapping.nested_objects.limit

Basic Query Constructs

must	AND condition
should	OR condition
filter	Filter non matching documents without impacting document scores. Frequently used filters will be cached automatically by Elasticsearch, to speed up performance
must_not	NOT condition
match	Returns documents that match a provided text, number, date or boolean value. The provided text is analyzed before matching.
match_phrase	All terms + Order + Without other intervening words. For example match_phrase search for “Hello World” would return document with “**I Just said hello world**”
term	Returns the documents where the value of a field exactly matches the criteria with score.

Geo Queries

- geo_distance
- geo_polygon
- geo_distance_range
- geo_bounding_box
- Geo_shape



Rest APIs

Elasticsearch

- POST /_aliases
- GET /_all/_mapping
- GET _search
- GET /_cat
- GET _cat/indices
- GET /_cat/aliases
- GET /_cat/health
- GET /_cat/nodeattrs
- GET /<index name>/_settings
- POST /<index name>/_doc
- POST _analyze
- POST /<index>/_analyze
- GET /_index_template
- GET /_data_stream
- GET /_ilm/policy
- POST _bulk
- POST /<index>/_update/<_id>
- POST /<index>/_update_by_query

Elasticsearch

Demo

https://github.com/dileep-gadiraju/elastic-search-training

Q&A

Elasticsearch

Thank you!