Isaac Araya Solano | Carnet: 2018151703

Resumen #1

What is Elasticsearch?

Elasticsearch is a distributed search and analytics engine at the heart of the Elastic Stack. It provides real-time search and analytics for all types of data. You can go from simple data retrieval to discovering trends and patterns in your data.

These are some things you can do with Elasticsearch:

- Add a search box to an app or website
- Store and analyze logs, metrics, and security event data
- Use machine learning to automatically model the behavior of your data in real time
- Automate business workflows using Elasticsearch as a storage engine
- Manage, integrate, and analyze spatial information using Elasticsearch as a geographic information system (GIS)
- Store and process genetic data using Elasticsearch as a bioinformatics research tool

Data in: documents and indices

Elasticsearch is a distributed document store that stores the information in complex data structures serelialized as JSON Documents. Stored documents can be accessed immediately from any node of a cluster. All documents stored are indexed and can be accessed within 1 second. The data structure used aids performing full-text searches quickly.

- Index: optimized colletion of documents.
- Document: collection of fields.

Elasticsearch indexes all data in every field and each indexed field has an optimized structure. This management of the information is what makes elasticsearch so fast.

Elasticsearch can also be schema-less, which means that documents can be indexed without further specification. There is also an option of dynamic mapping that detects and adds fields to the index, making easy to index and explore data. You can even define rules to control dinamic mapping yourself.

Information out: search and analyze

Even though Elasticsearch can be used as a document store, its real potential shows when you use the full suite of search capabilities. It provides a simple and coherent REST API fro managing your cluster abnd indexing and searching data.

Searching your data

The Elasticsearch REST APIs support structured queries, full text queries and the combination of both of them. They are similar to the ones you can contruct in SQL. With the full-text queries the system will show the documents that match the query string sorted by relevance.

Analyzing your data

Elasticsearch aggregations help building complex summaries and have access to metrics, patterns and trends. This tools help you get way more info than you could without them.

These aggregations are also really fast and operate alongside search requests. You can perform searches, filter the results and analyze the date at the same time in a single request. And there is even more since you can apply machine learning features to automate the analysis of data and finding patterns, problems and other characteristics of your information.

Scalability and resilience: clusters, nodes, and shards

Elasticsearch is designed to be always available and to scale in relation to your needs. You can add a new server and elasticsearch distributes the information and queries automatically.

There are two types of shards: primaries and replicas. Each document indexed belongs to a primary shard and the replicas are copies of them. The replicas provide the redundant copies to protect the information. The number of primary shards in an index is fixed while the replicas can be modified any time.

It depends...

It is important to know that the number of shards can affect performance. The more shards, the more demanding is the system and the longer it takes to move shards when rebalancing a cluster. It can be faster to query lots of small shards since it can be faster per shard but however querying smaller number of larger shards might be faster. It depends.

In case of disaster

The nodes need reliable connections to each other. To provide better connections it better to have the nodes in the same data center or nearby data centers but it can be troublesome since you need to maintain high availability. The answer is Cross-cluster replication(CCR). CCR can automatically synchronize indices to a secondary remote cluster to serve as a backup.