Scientific Journal Impact Factor (SJIF) : 5.278
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

ELASTICSEARCH

Krishan Kumar Singh[1], Mohit Kumar[2], Mayank Singhal[3] and Aashrita Dubey[4]
[1,2,3,4]*IIMT College of Engineering, Greater Noida*

Abstract- In today's era, it is inconceivable to use traditional techniques / RDBMS to analyze the data as it is growing very quickly. Big data offers the solution for analyzing large amount of data. Using technique of Elasticsearch, access to data can be made faster. Elasticsearch is a search engine based on Lucene. It is near real time search platform. Elasticsearch uses the concept of indexing to make the search faster.

Keywords- Elasticsearch, Lucene.

I. INTRODUCTION

- As name suggests, Elasticsearch is full text search engine library.
- It's an open source search engine built on top of Apache Lucene.
- Elasticsearch is written in java and uses Lucene library internally for indexing and searching.
- Elasticsearch is much more than Lucene, because it aims to make full text search much simpler by hiding the complexities of Lucene behind a simple, coherent, restful API.
- What Elasticsearch is doing?
- Distributed real time document store where every field is indexed and searchable.
- Distributed search engine with real time analytics.
- Capable of scaling to hundreds of servers and petabytes of structured and unstructured data.
- Its near real time as it makes document searchable immediately after indexing the document.

II. BASIC CONCEPTS

Cluster:

- Cluster is a collection of one or more nodes that together holds entire data and provides indexing and searching capabilities.
- A cluster is identified by the unique name ("elasticsearch" by default).
- Name is important because node can only be the part of cluster if the node is set up to join the cluster by its name.
- For instance, we can have logging_dev,logging_stage,logging_prod
  Cluster names for development, stage and production use cluster.
- We can assign cluster name by configuring the elasticsearch.yml file:
  Cluster.name: elasticseach.prod

Note: It is wise to rename your cluster to prevent accidents where by someone's laptop joins the cluster

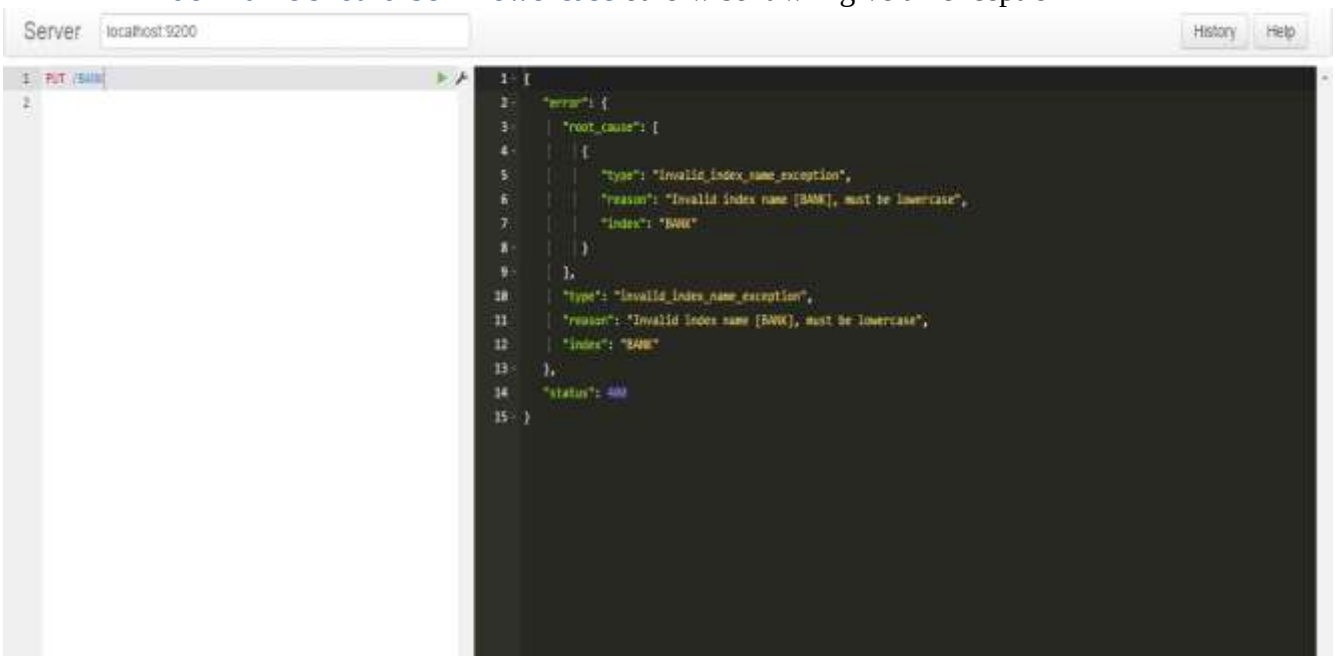
Node:

- Node is a single server that is part of the cluster, stores the data and participate in indexing and searching capabilities.
- Just like cluster, node is also identified by its name. Node name is used for administrative purpose.

- A node can be configured to join the specific cluster by cluster name.(By default node is joined to cluster named " elasticsearch ")
- In single cluster we can have as many as nodes. Even a single cluster can create a cluster.
- Assign Node name in elasticsearch.yml
  Node.name: node name

Index:

- Index is collection of documents that have similar characteristics.
- It is logical namespace.
- Like if we are collecting the documents related to bank so we can name the index as "bank"
  Index name should be in lowercase otherwise it will give an exception



- Index name is used to refer while indexing, searching or analyzing the data.
- In a single cluster we can have many indices.

Type:

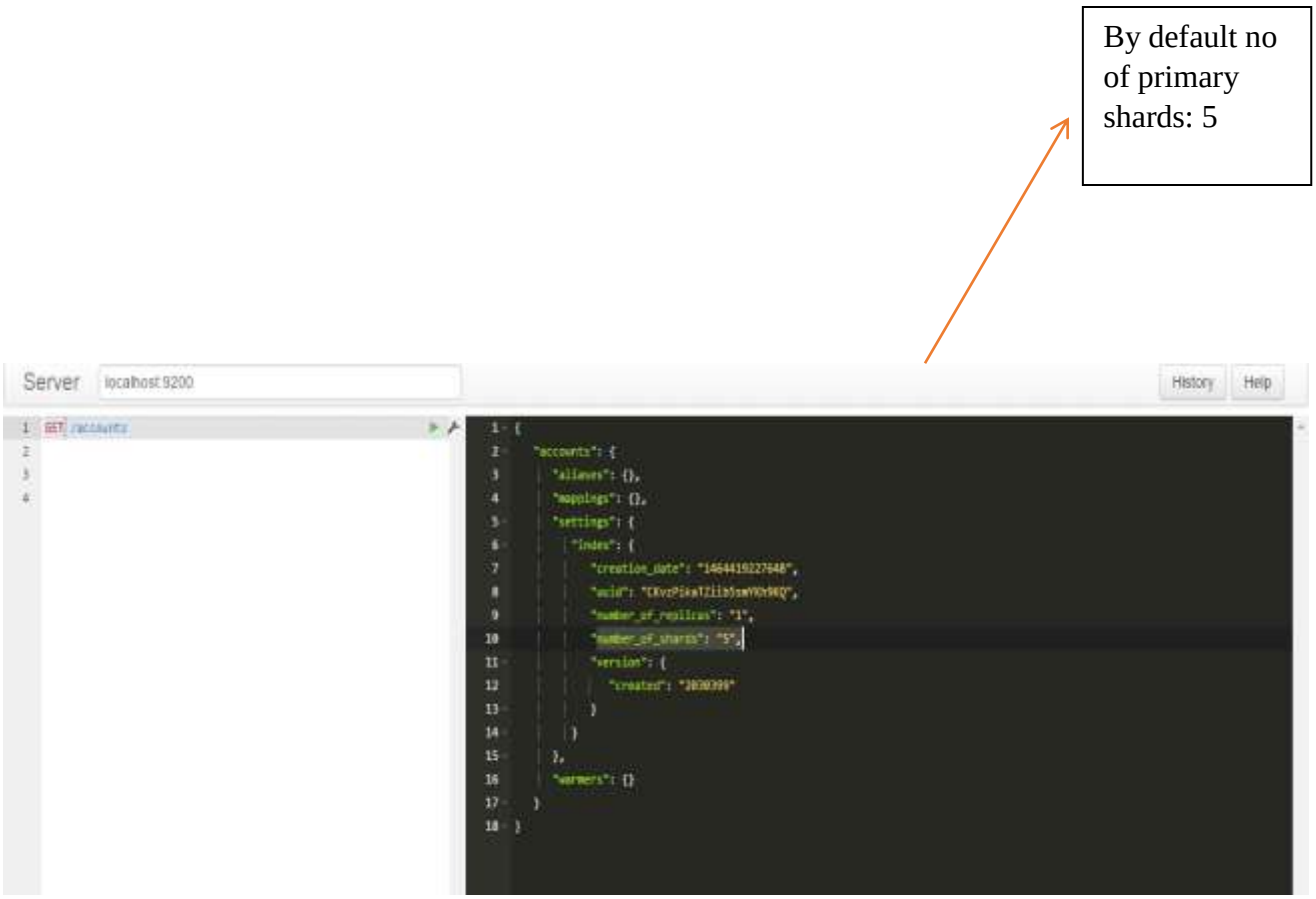
- Within an Index we can have many types. Type contain the documents having common fields.
- Let's say inside Bank Index we can have type "saving" for saving account holders or type "loan" for loan account holders.

Document:

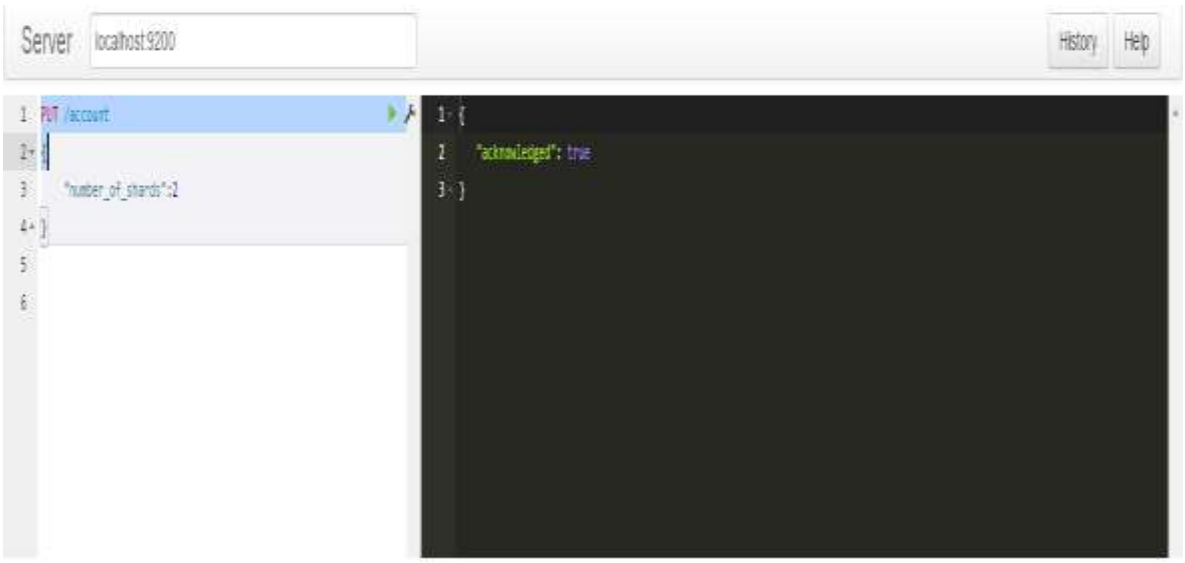
- Document is collection of data or basic unit of information that can be indexed.
- It is expressed in JSON format.
- JSON format is the only format which Elasticsearch supports. (**Limitation of Elasticsearch**).

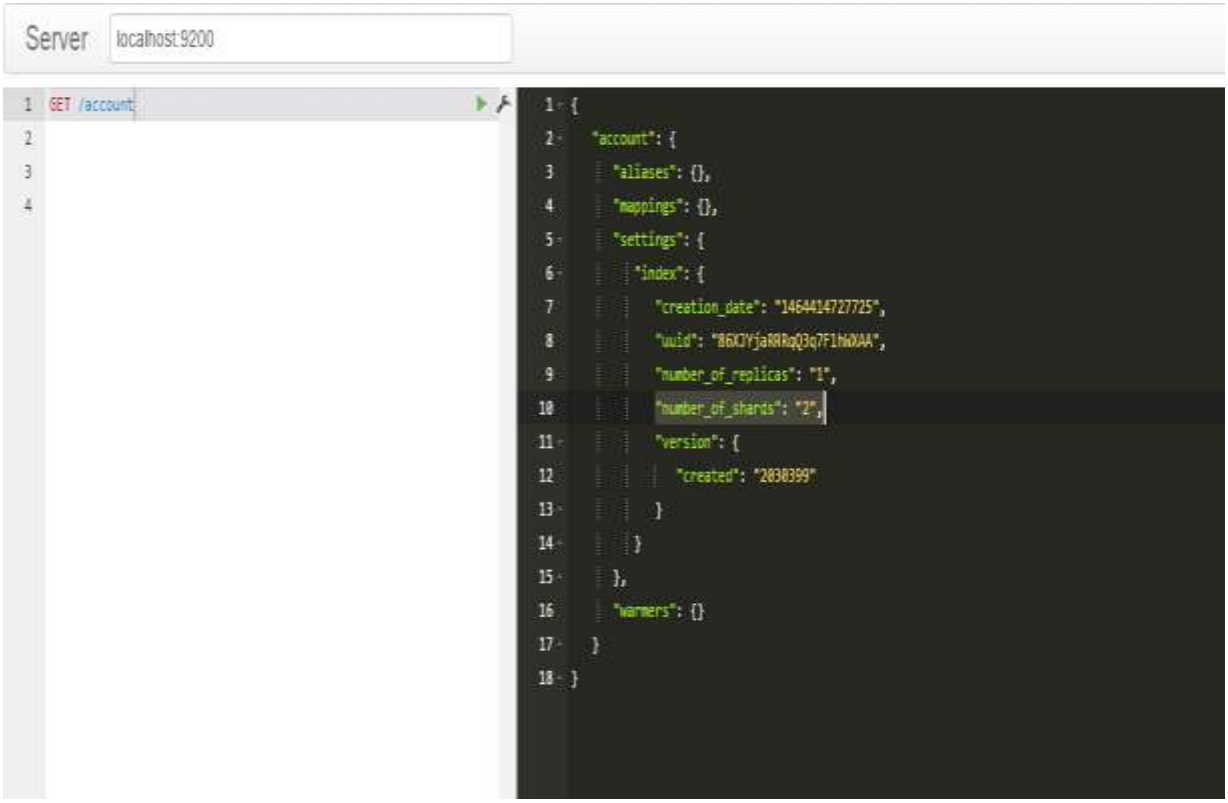
Shards & Replicas:

- As we are indexing a huge amount of data that can exceeds the hardware limit of single node.
- So to handle the huge data ES provides the feature o subdivide the index in multiple parts called **SHARDS.**
- **By default no of shards provided by ES are 5.**



While creating index we can define the number of shards an index can have. After that, Shards will be fully functional index independently.





Shard is important for 2 primary reasons:

- It allows you to horizontally split/scale the data volume
- It allows you to distribute and parallelize operation across the shards(increasing the performance)

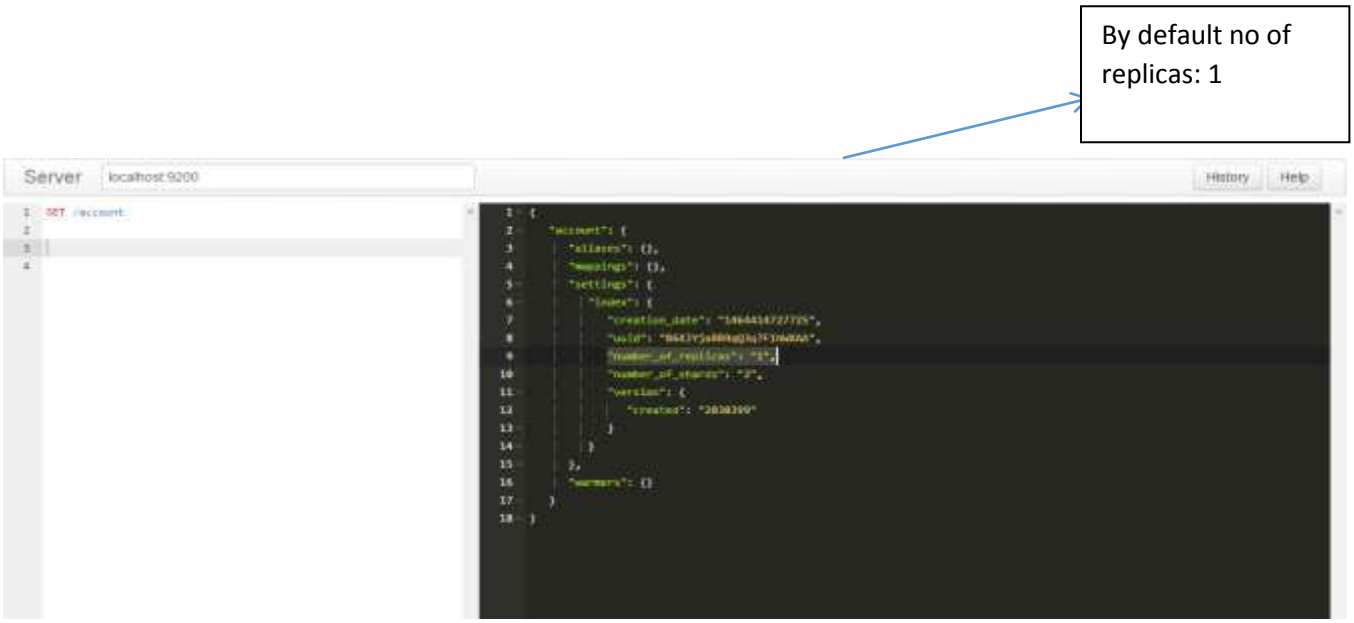
As the data is in huge amount failure in nodes is expected anytime. So it is very useful to handle this failure by having **failover mechanism** in case of any node/shard somehow goes offline or disappears.

To handle this situation ES allows us to have copies of shards to what we call **REPLICA SHARDS.** Primary reasons of having REPLICA SHARDS:

- It provides high availability of data/content in case of node/shard failure. For this reason Replica Shard should not be on the same node on which shard is resided from that it was copied from.
- It provides the scale out your search volume/throughput because search can be executed on replica shards as well in parallel.

By default Replica Shards provided by ES is 1 per PRIMARY SHARD.

But while creating index we can define number of replicas also. It can be 0 or more.



The number of shards (primary + replicas) we will be allocated for at least 2 nodes

By default, each index in Elasticsearch is allocated 5 primary shards and 1 replica which means that if you have at least two nodes in your cluster, your index will have 5 primary shards and another 5 replica shards (1 complete replica) for a total of 10 shards per index.

III. CONCLUSION

The following area of the solution needs to be closely examined by Architects and Developers to ensure component/service reusability across all the Big Data projects:

- **Error Handling**: Procedures and services to handle all types of application-level errors (system, functional, data, and warning) in a consistent way.
- **Alert Notification** – A common mechanism in accordance with the enterprise standards to notify application errors or business events critical for timely actions.
- **Message Logging**: A common mechanism to log all types of messages (error, warning, information, debug) either de-centralized or centralized location as per the enterprise standard. Ability to configure log level and log location is also important.
- **Auditing including business event auditing**: A common mechanism to register and notify business events to appropriate business users – either for auditing, tracing, timely actions/escalations or future analysis.
- **Document Delivery services to Partner using different channels**: Common services to support delivery of outbound documents to the partner for a set of agreed upon protocols and data exchange formats. Inclusive of all security related functionality.
- **Document Receipt services from Partner using different channels**: Common services to support delivery of outbound documents to the partner for a set of agreed upon protocols and data exchange formats. Inclusive of all security related functionality.

REFERENCES

[1] https://www.elastic.co/
[2] https://qbox.io/blog/qbox-elasticsearch-tutorial-1
[3] https://www.digitalocean.com/community/tutorials/how-to-install-and-configure-elasticsearch-on-ubuntu-14-04
[4] https://www.elastic.co/guide/en/elasticsearch/hadoop/current/mapreduce.html#writing-json-new-api

