

|  |
| --- |
| Data Validation using Elasticsearch |
|  |
| November 9  A joint paper by JP Tokyo & Co and MRIIRS  Authored by: Sonakshi Khosla  Reviewed by: Rajesh Nath, Charu Virmani |

Contents

[Title Heading 3](#_Toc527993527)

[Subtitle Text Here 3](#_Toc527993528)

# Elasticsearch

|  |
| --- |
| What is elasticsearch? **Elasticsearch** is a [search engine](https://en.wikipedia.org/wiki/Search_engine_(computing)) based on [Lucene](https://en.wikipedia.org/wiki/Lucene) library. It provides a distributed, [multitenant](https://en.wikipedia.org/wiki/Multitenancy)-capable [full-text search](https://en.wikipedia.org/wiki/Full-text_search) engine with an [HTTP](https://en.wikipedia.org/wiki/HTTP) web interface and schema-free [JSON](https://en.wikipedia.org/wiki/JSON) documents. Elasticsearch is developed in [Java](https://en.wikipedia.org/wiki/Java_(programming_language)) and is released as [open source](https://en.wikipedia.org/wiki/Open_source_software) under the terms of the [Apache License](https://en.wikipedia.org/wiki/Apache_License). Official clients are available in [Java](https://en.wikipedia.org/wiki/Java_(programming_language)), [.NET](https://en.wikipedia.org/wiki/.NET_Framework) ([C#](https://en.wikipedia.org/wiki/C_Sharp_(programming_language))), [PHP](https://en.wikipedia.org/wiki/PHP_(programming_language)), [Python](https://en.wikipedia.org/wiki/Python_(programming_language)), [Apache Groovy](https://en.wikipedia.org/wiki/Apache_Groovy), [Ruby](https://en.wikipedia.org/wiki/Ruby_(programming_language)) and many other languages.[[5]](https://en.wikipedia.org/wiki/Elasticsearch#cite_note-offizsite-5) According to the [DB-Engines ranking](https://en.wikipedia.org/wiki/DB-Engines_ranking), Elasticsearch is the most popular enterprise search engine followed by [Apache Solr](https://en.wikipedia.org/wiki/Apache_Solr), also based on Lucene.[[6]](https://en.wikipedia.org/wiki/Elasticsearch#cite_note-6)  Elasticsearch is developed alongside a data-collection and log-parsing engine called Logstash, and an analytics and visualisation platform called [Kibana](https://en.wikipedia.org/wiki/Kibana). The three products are designed for use as an integrated solution, referred to as the **"Elastic Stack"** (formerly the "ELK stack").  **Kibana** is an open source data visualization plugin for **Elasticsearch**. It provides visualization capabilities on top of the content indexed on an **Elasticsearch** cluster. Users can create bar, line and scatter plots, or pie charts and maps on top of large volumes of data.  Companies using Elasticsearch are: Ebay, IEEE, Citibank, Ottomotors, NYU Library, Shopify, Godaddy, Fitbit, Microsoft, Volkswagen, Accenture, Dell, Facebook, Netflix, Walmart, Slack, BBC, Uber etc. |
|  |

Validating data

Two kinds of validation:

* Structural validation of fields based on a regular expression for  
  example. Perhaps something can be configured in the mapping...
* Integrity validation of document. For example preventing from indexing  
  a document with a field value that already exists.