elastic

Open source solutions for big data challenges

Jason Wong
2020-Mar-22, Education Architect, Elastic

“

In the modern world of business, data management has become a critical challenge to overcome - in the “red sea”, companies that invested in data management could dramatically reverse their fate from Zero to Hero

Agenda

1. Differences between data collection and data management
2. What could you do with the data on hand?
3. Cost of data management
4. The process
5. Story telling on the data
6. FAQ

elastic

who’s talking???

```
{
        “name”     : “Jason Wong”,
        “company” : {
                “name”     : “Elastic (SG)”,
                “position” : “Education Architect”,
                “roles”    : [ “run trainings”,
                               “curriculum development”,
                               “prototyping”,
                               “run meetups” ]
        },
        “toolbox” :[ “java”, “react.js”, “vue.js”,
                     “angular.js”, “node”, “golang” ]
}
```

resources

- Github
  - coronavirus demo: https://github.com/quoeamaster/bigdata_blogs/tree/master/coronavirus_COVID_19_usecase
- blog for the demo: https://medium.com/@quoeamaster/use-case-coronovirus-covid-19-investigation-and-visualisation-with-elasticsearch-598df7849e82
- social media
  - https://www.linkedin.com/in/jason-wong-391a7624/
  - https://github.com/quoeamaster
  - https://medium.com/@quoeamaster
  - https://bigdata887.wordpress.com/
  - https://twitter.com/QueMaster3

elastic

1.

where are we now ?
data collection OR
data management ?

elastic

definition

- ## data collection

  the collection of data from vast sources with limited or no processing

- ## data management

  consists of several steps / process including the collection of data, processing of data and data liberation

elastic

brain teasers - data collection or data management?

- collecting logs from different internal systems and treat them as audit logs

- generating trend analysis on granted loans to SME in Singapore

- discovering the relationships between savings and occupation

elastic

brain teasers - data collection or data management?

- collecting logs from different internal systems and treat them as audit logs

data collection

- generating trend analysis on granted loans to SME in Singapore

- discovering the relationships between savings and occupation

elastic

brain teasers - data collection or data management?

- collecting logs from different internal systems and treat them as audit logs

data collection

- generating trend analysis on granted loans to SME in Singapore

data management

- discovering the relationships between savings and occupation

elastic

brain teasers - data collection or data management?

- collecting logs from different internal systems and treat them as audit logs

data collection

- generating trend analysis on granted loans to SME in Singapore

data management

- discovering the relationships between savings and occupation

data management

elastic

2.

what could we do with the data ?

elastic

here are some of them…

- perform analytics
- apply AI analysis (e.g. machine learning on buying behaviour)
- story telling (visualise the data, make them live)
- guide management on decision making (e.g. trend analysis, discovering behaviour and stickiness by regression analysis)

elastic

3.

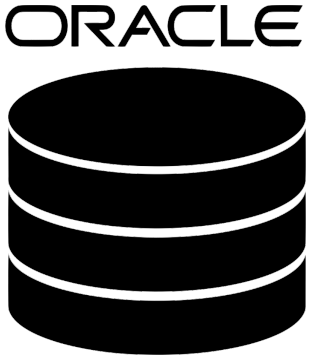
cost of data management

elastic

open source vs commercial products

elastic

open source vs commercial products



myths

- open source is crap
- support is limited on open source products
- a team of 100 vs a collaborations of talents
- the open source company behind is not stable (before and after IPO…)

elastic

costs on the team

- need to setup a team of talents to:
  - integrate various sources of data input (e.g. from databases, web api, csv files etc)
  - process / cleansing of data (e.g. from non-structured data to structured data)
  - apply analytics (e.g. manual or machine-learning)
  - generate visualisation on story telling
  - housekeeping of old data

PS. a good tool stack would help on reducing the above costs

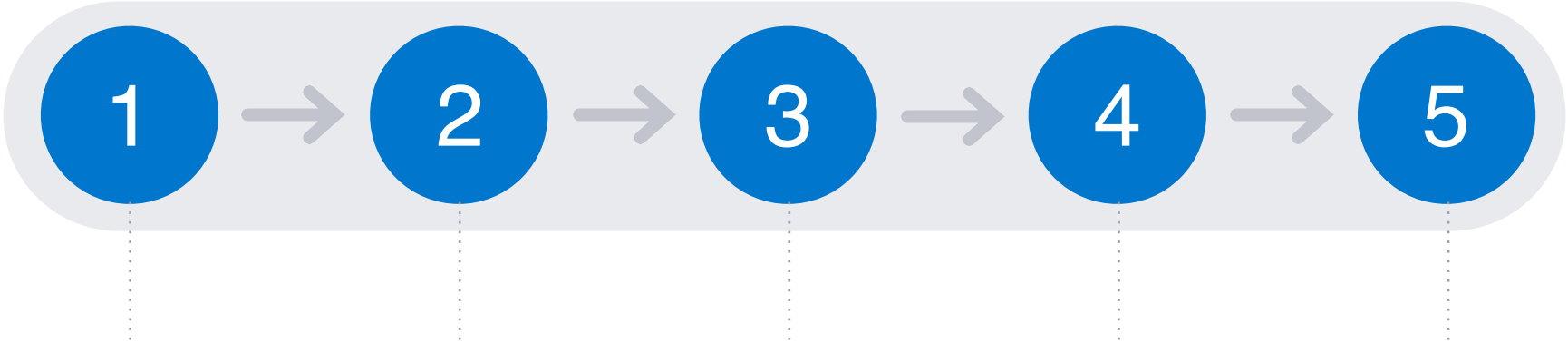
elastic

4.

the process

elastic

process pipeline



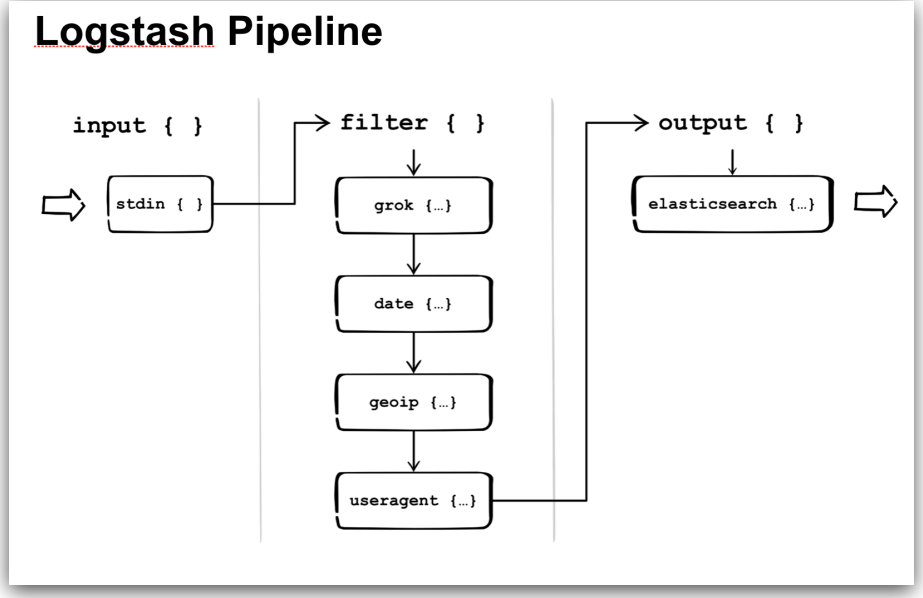
1. integration with data source and collect
2. processing / cleansing of data
3. perform further analysis (e.g. manual or machine learning)
4. generate visualisations and reports
5. housekeeping (e.g. archive old data)

elastic

process 1: integration and collection of data from vast sources

a tool that can be configured to talk with different data sources and extract values out.

For example: logstash



elastic

process 2: preparing and cleansing data

multiple approaches:

- coding level
- tools like logstash
- server side pipelines (e.g. elasticsearch ingest nodes)

```
PUT _ingest/pipeline/access_log
{
  "description" : "Ingest pipeline for Combined Log Format",
  "processors" : [
    {
      "grok": {
        "field": "message",
        "patterns": ["%{IPORHOST:clientip} %{USER:ident} %{USER:auth} \\[%
{HTTPDATE:timestamp}\\] \"%{WORD:verb} %{DATA:request} HTTP/%{NUMBER:httpversion}\" %
{NUMBER:response:int} (?:-|%{NUMBER:bytes:int}) %{QS:referrer} %{QS:agent}"]
      }
    },
    {
      "date": {
        "field": "timestamp",
        "formats": [ "dd/MMM/YYYY:HH:mm:ss Z" ]
      }
    },
    {
      "geoip": {
        "field": "clientip"
      }
    },
    {
      "user_agent": {
        "field": "agent"
      }
    }
  ]
}
```


elastic

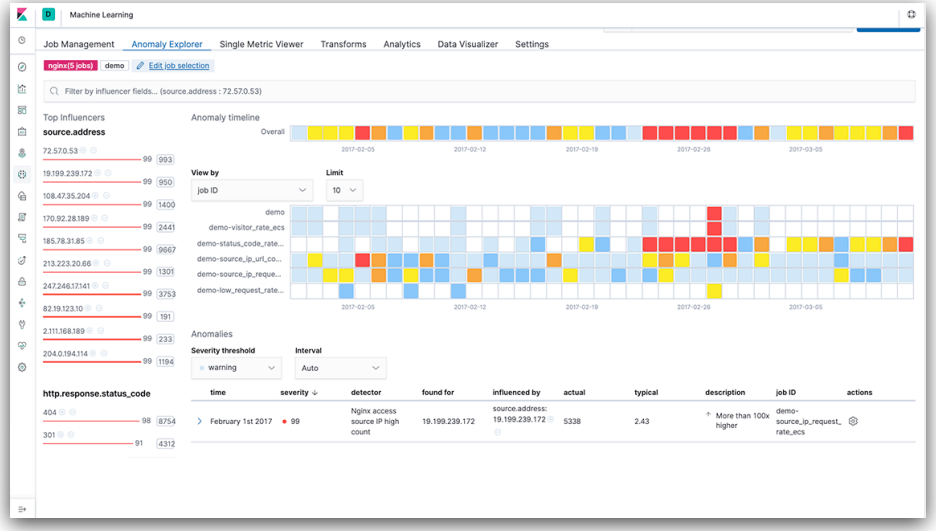
process 3: further processing

manual approach:

- analytic tools required

machine learning approach:

- talents might be required to config a model and pick the right algorithm to start “learning” (e.g. machine learning module in elasticsearch)



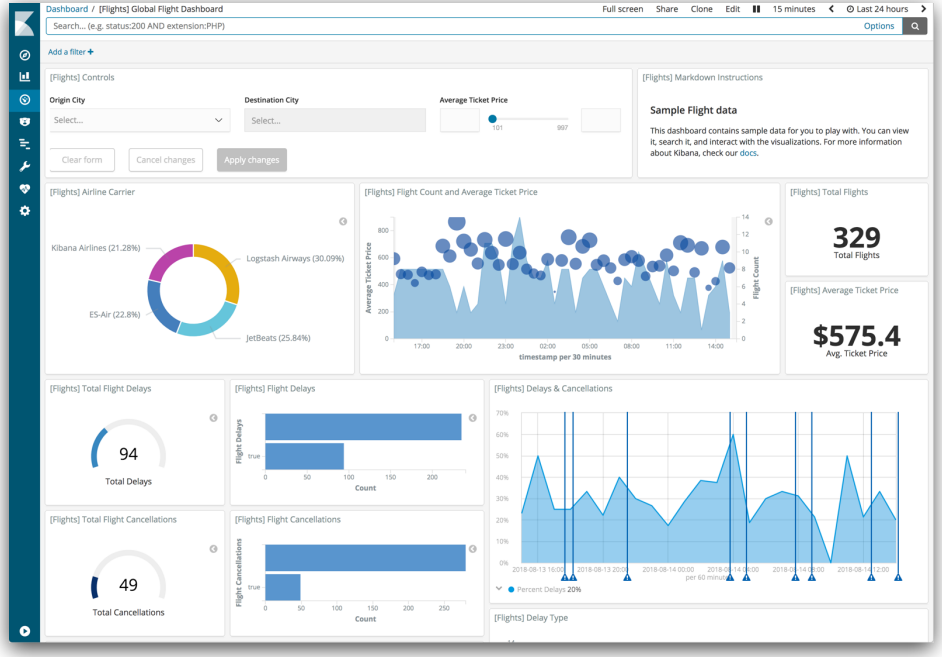
elastic

process 4: data liberation through visualisations

a Business Intelligence tool is required

open source tools are free but usually have limitations

commercial tools usually have more support on visualisation types and able to integrate with different data sources

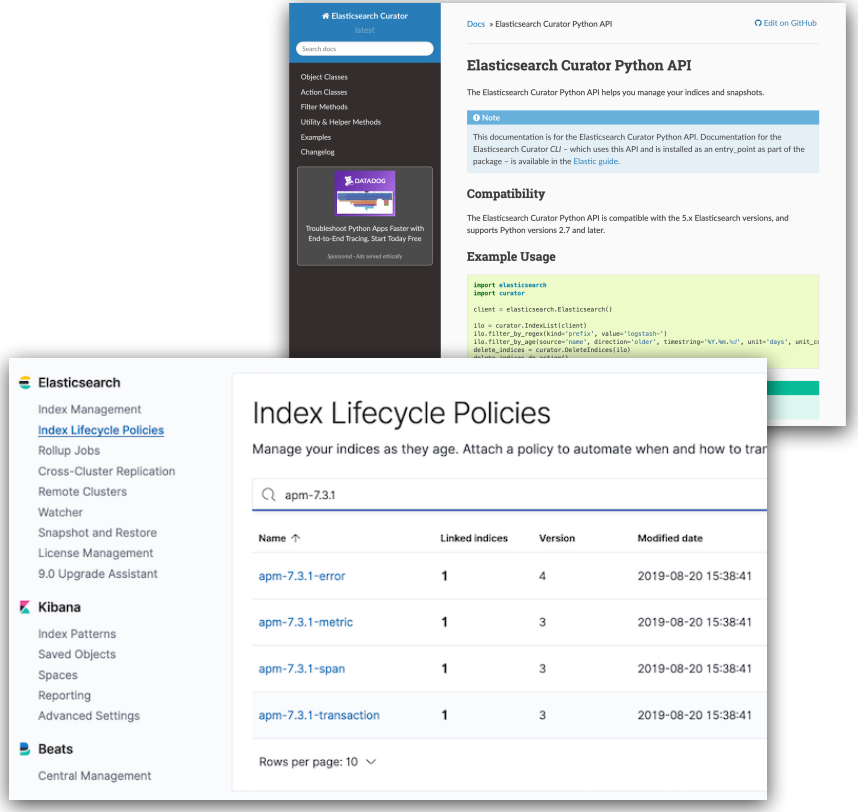


elastic

process 5: housekeep

tools are required to perform housekeep operations:

- archive old data
- re-allocate resources on active data but still have limited support to access archived data
- compacting read-only data to reduce resources consumption
- backup and restore capability of data



5.

story telling

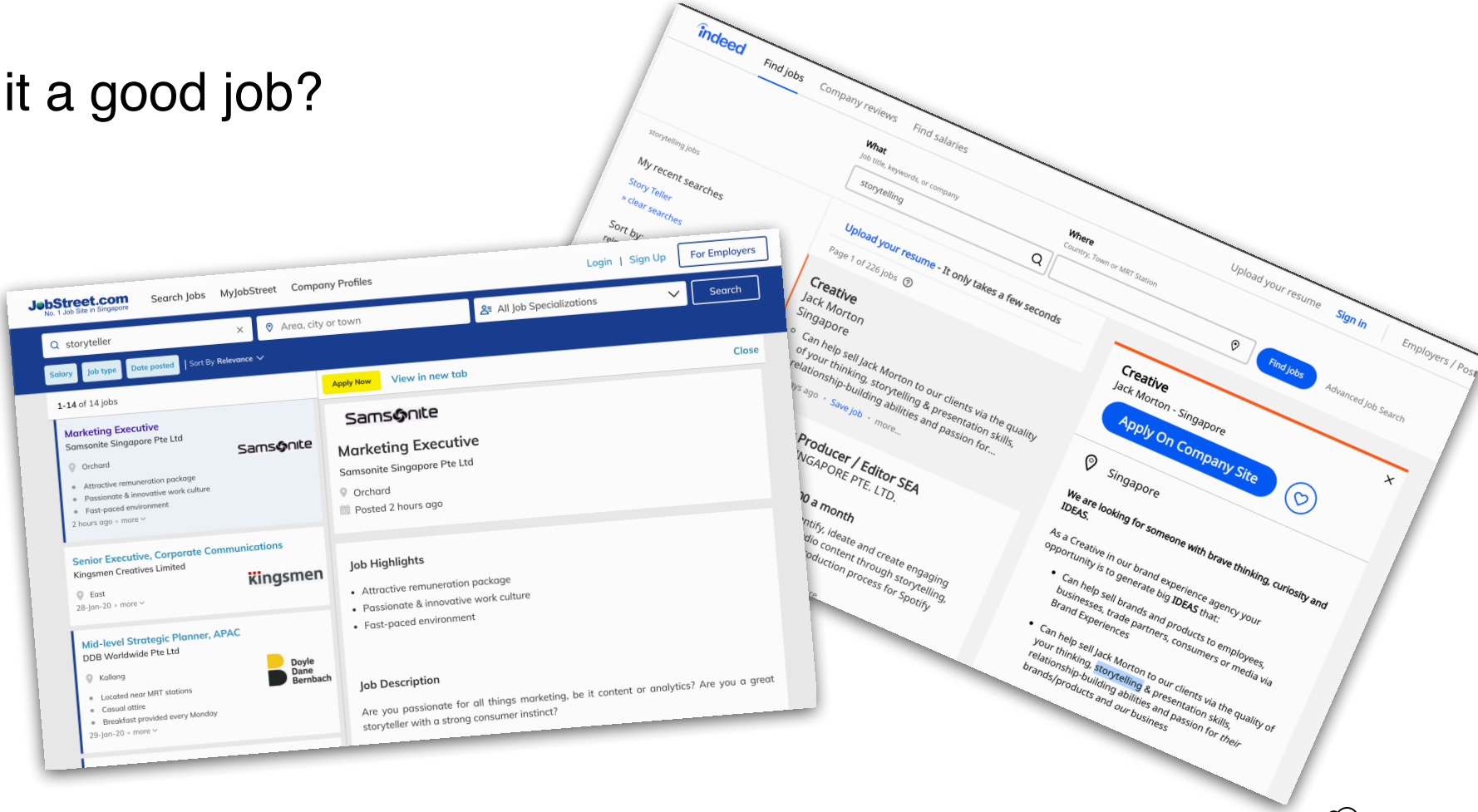
elastic

why is story telling important?

- finding the values hidden in the data
- data collected from customers revealed their behaviour truth (no way to lie) and hence a more accurate way to predict sales OR to tailor make campaigns that sells
- a less boring way to present static data

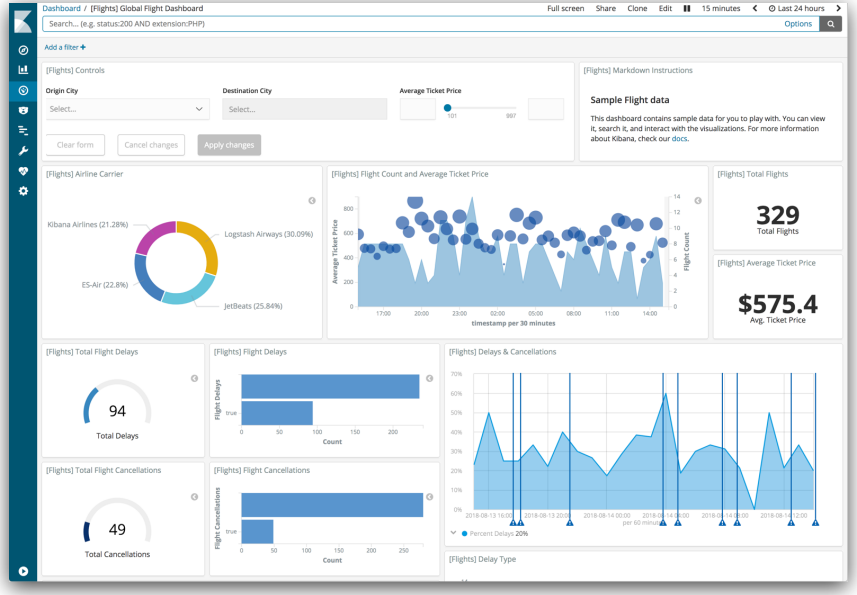
elastic

is it a good job?

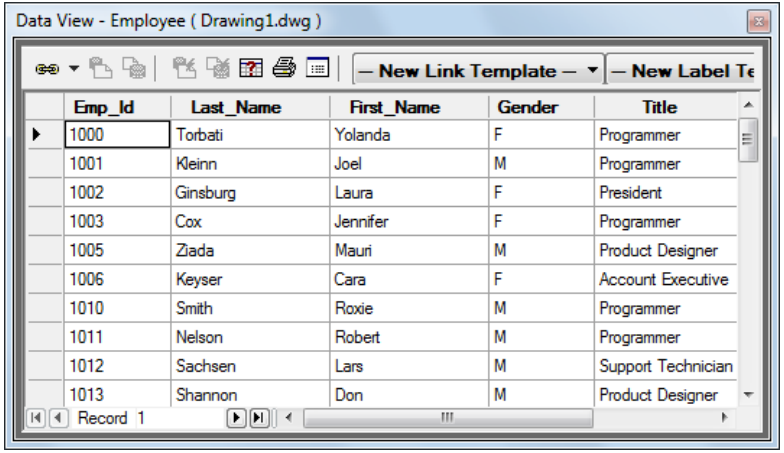


elastic

visualisation helps on storytelling and data liberation



vs



Emp_Id	Last_Name	First_Name	Gender	Title
1000	Torbati	Yolanda	F	Programmer
1001	Kleinn	Joel	M	Programmer
1002	Ginsburg	Laura	F	President
1003	Cox	Jennifer	F	Programmer
1005	Ziada	Mauri	M	Product Designer
1006	Keyser	Cara	F	Account Executive
1010	Smith	Roxie	M	Programmer
1011	Nelson	Robert	M	Programmer
1012	Sachsen	Lars	M	Support Technician
1013	Shannon	Don	M	Product Designer

elastic

Learn more?

elastic

Elastic Stack

SOLUTIONS

Kibana — Visualize & Manage

Elasticsearch — Store, Search, & Analyze

Beats, Logstash — Ingest

SaaS

Elastic cloud

On-Prem

Elastic cloud Enterprise

Elastic cloud On Kubernetes

Standalone

elastic

official trainings

Master the Elastic Stack

The demand for Elasticsearch and Kibana experts increases every day. Our certifications let everyone know that you're just what they're looking for.

Elasticsearch Engineer I

Start the path to certification and learn how to manage deployments and develop solutions.

Learn more

Elasticsearch Engineer II

Develop a deeper understanding of how Elasticsearch works and master advanced deployment techniques.

Learn more

Elastic Certified Engineer

Test your Elasticsearch skills with our performance-based certification exam, and prove you're an expert.

Get certified

DEVELOPMENT AND OPERATIONS

Elastic Certified Engineer

This certification is designed for versatile Elasticsearch experts who can install and manage Elasticsearch clusters, as well as develop solutions for searching and analyzing their indexed data. To prepare for this exam, we recommend taking the **Elasticsearch Engineer I** and **Engineer II** trainings.

Become an Elastic Certified Engineer

ELASTIC CERTIFIED · QUAERERE · INVENIET · ENGINEER

ELASTIC CERTIFIED · IDEST IN · NOTITIA · OPS ANALYST

ANALYSIS AND INSIGHT

Elastic Certified Ops Analyst

Certified ops analysts are experts at using Kibana for both data visualization and analysis. From creating powerful dashboards to analyzing time-series data to developing machine learning jobs, they can do it all. To prepare for this exam, we recommend taking the **Kibana Data and Ops Analyst** training.

Learn about Elastic Certified Ops Analyst exam

Get involved at https://training.elastic.co/

elastic

resources

- Github
  - coronavirus demo: https://github.com/quoeamaster/bigdata_blogs/tree/master/coronavirus_COVID_19_usecase
- blog for the demo: https://medium.com/@quoeamaster/use-case-coronovirus-covid-19-investigation-and-visualisation-with-elasticsearch-598df7849e82
- social media
  - https://www.linkedin.com/in/jason-wong-391a7624/
  - https://github.com/quoeamaster
  - https://medium.com/@quoeamaster
  - https://bigdata887.wordpress.com/
  - https://twitter.com/QueMaster3

elastic

6.

FAQ

elastic

Thank you !

elastic