

INTRODUCTION TO ELASTIC CLOUD



Opensource software by definition earns no money for its creators, since it's free. The primary sponsor then hopes larger companies will sign up for the enterprise product, for which they charge money. That way the user gets offer support, consulting, and it unlocks different features.

For example, Elasticsearch is free, but with the enterprise edition you get machine learning and other features, for a free. Here we show you the Elastic Cloud, hosted version of Elasticsearch.

Opensource companies recently have begun to host clusters themselves. This make this cheaper for its users that buying the product outright. It lets users pay monthly instead of up front.

The CFO loves that, as the company can write off operating costs immediately instead of having to amortize them over time. That's a fancy way of saying they can reduce their taxes sooner than later.

But for technical people like us, it makes the deployment and maintenance easier.

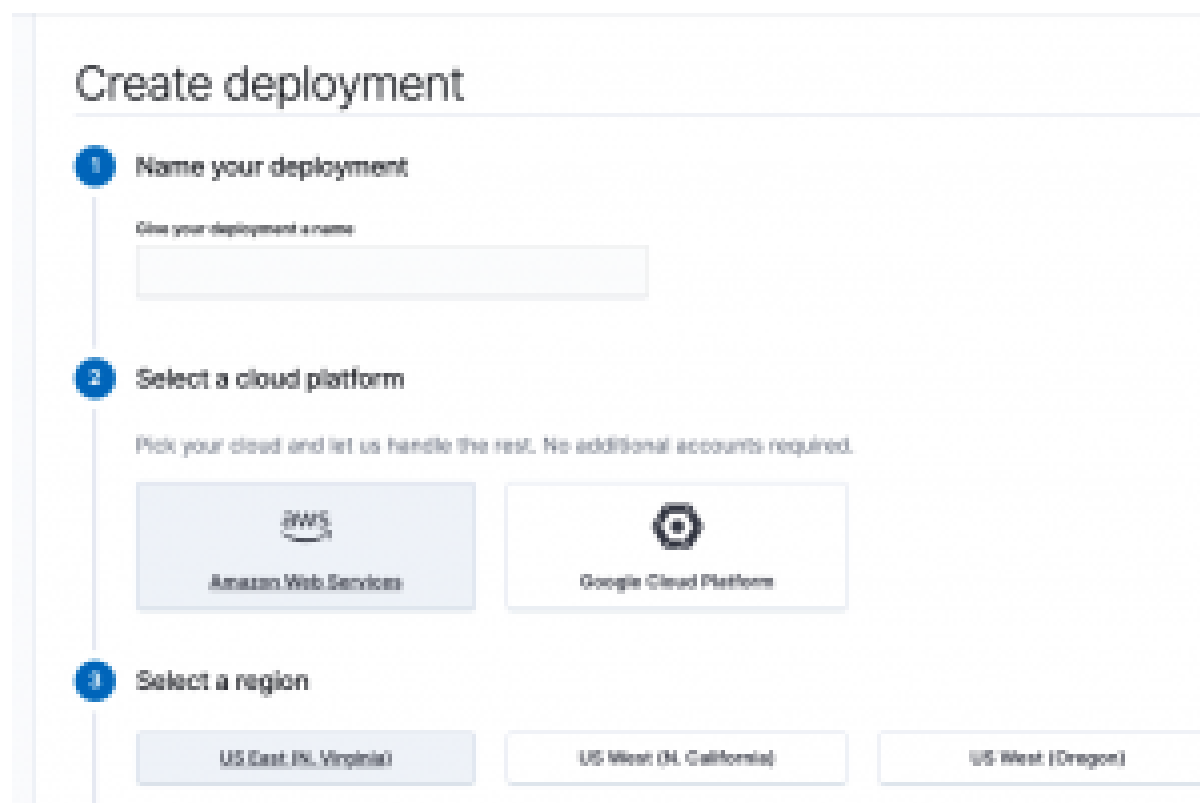
Here we discuss how Elasticsearch has done this with the Elastic Cloud. In short, the takeaway message is that you work with it the same as if you had installed it yourself. The only difference is you don't have ssh access to the instances. But you don't need that since you can use ssh from any other computer to get to those.

Getting Started

You can get a free 14 day Elastic Cloud trial at <https://cloud.elastic.co/>. It's easy to get started, you just create a Deployment and Elasticsearch installs a cluster for you at either Amazon AWS or the Google Cloud as the screen below shows.

The screen look familiar to those of you who have used AWS. It asks you for what region you want

your servers and what size hardware, meaning EC2 template.



The 'Create deployment' screen is divided into three numbered steps. Step 1, 'Name your deployment', features a text input field with the placeholder 'Give your deployment a name'. Step 2, 'Select a cloud platform', includes the instruction 'Pick your cloud and let us handle the rest. No additional accounts required.' and two buttons: 'Amazon Web Services' (with the AWS logo) and 'Google Cloud Platform' (with the GCP logo). Step 3, 'Select a region', shows three buttons: 'US East (N. Virginia)', 'US West (N. California)', and 'US West (Oregon)'.

You can select more memory, a necessity in any production use, storage and nodes on the customization screen. As you can see that is not possible with the trial version.

Trial user?

The trial includes more than enough to get you started with the Elastic Stack. Larger deployments require a credit card to unlock and are not free.

Data 1 configuration

aws-data-highio.i3

DetailsInputMonitor

An i3 optimized Elasticsearch instance running on an AWS i3.

Availability

1 zone2 zones3 zones

RAM per Node

1 GB2 GB4 GB8 GB16 GB24 GB32 GB

Nodes 1

RAM per Zone

4 GB

Summary

4 GB RAM160 GB storage1 node2 zones8 GB RAM240 GB storage

User setting overrides

Machine Learning 1 configuration

aws.ml.m5

Machine Learning

An Elasticsearch machine learning instance running on an AWS m5.

Summary

Version	v7.11
ES data memory	8 GB
ES data storage	240 GB
Total memory	8.5 GB
Total storage	240 GB

Architecture

Zone 1

aws-apt-y4
10 GB RAM

aws-data-h...
8 GB RAM

aws-kibana...
1 GB RAM

Zone 2

aws-data-h...
4 GB RAM

Then it

automatically rolls out a the ElasticSearch servers and the Kibana front end. It gives you an internet domain name so anyone in your organization can access it. It install the SSL certificates and adds a password.

Save your password:

Generated user

You can use the credentials below to login to Elasticsearch or Kibana. Make sure to save the password somewhere as this is the only time we can show it to you.

Username

elastic

Password

XXXXXXXXXXXX

Cloud ID

XXXXXXXXXXXX

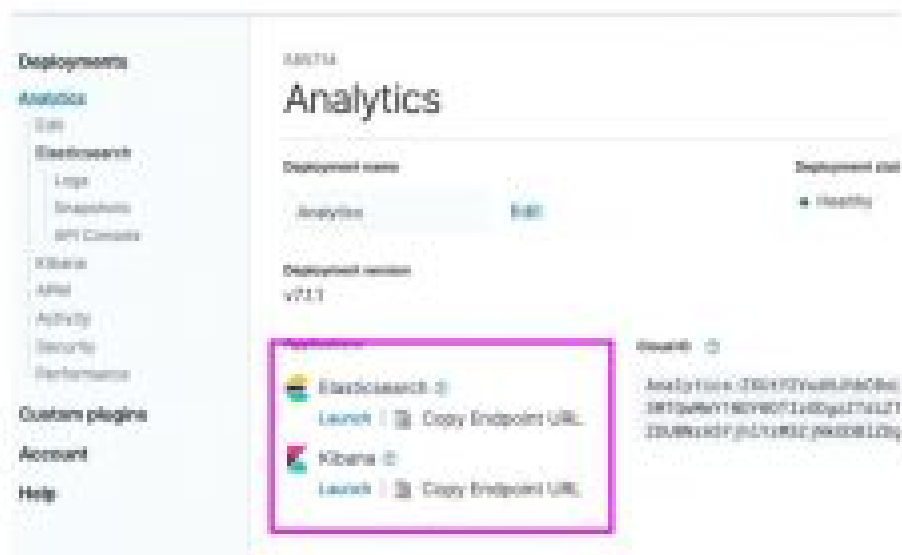
Get started with Beats and Logstash quickly.

The Cloud ID simplifies sending data to your cluster on Elastic Cloud. [Learn more...](#)

APM Server secret token

XXXXXXXXXXXX

From the main screen, click **Copy Endpoint** it will give you the URL for your instance. For example, my URL is <https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243>



Once you get a deployment (aka cluster) you can check its status with curl, if you are used to using the command line, as shown below.

First save `pwd="userid:password"`, used for basic authentication, in an environment variable. Then run the curl:

```
export pwd="elastic:"
```

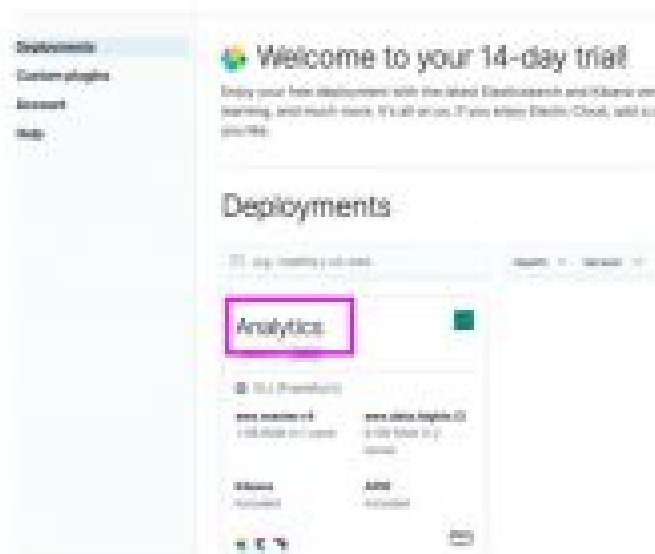
```
curl --user $pwd -H 'Content-Type: application/json' -XGET  
https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243/_cluster/health?pretty
```

Here it shows all is good (green) and I that have 2 data nodes.

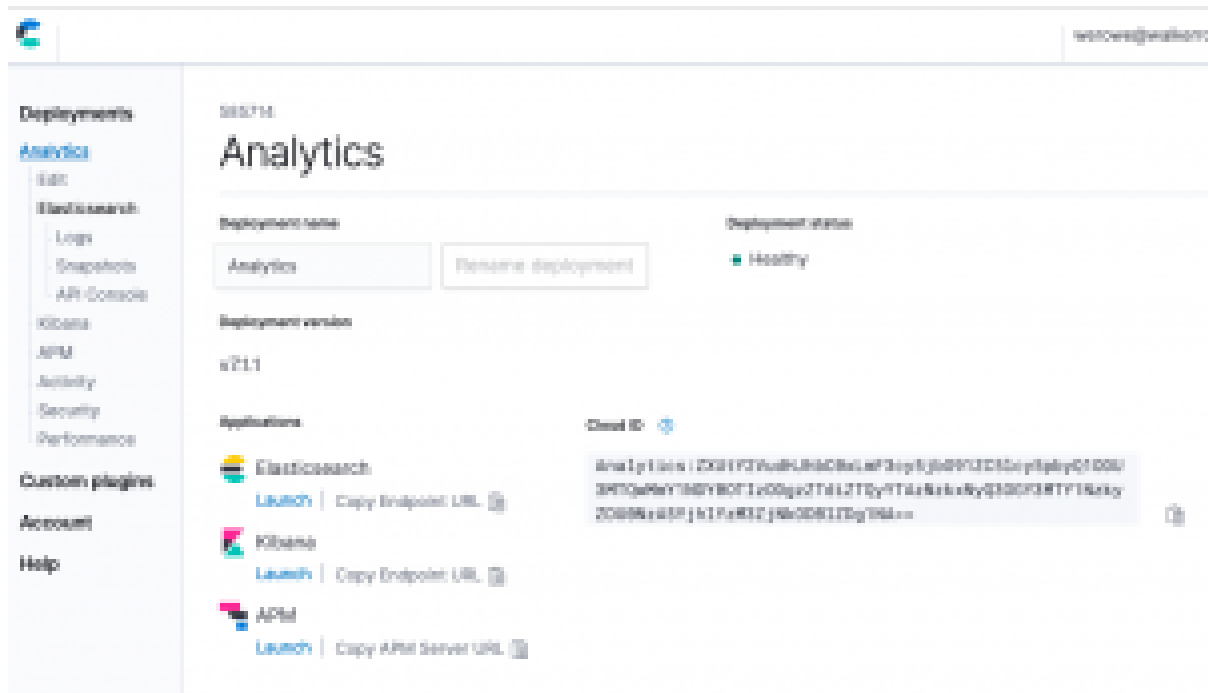
```
{
  "cluster_name" : "58571402f5464923883e7be42a037917",
  "status" : "green",
  "timed_out" : false,
  "number_of_nodes" : 3,
  "number_of_data_nodes" : 2,
  "active_primary_shards" : 6,
  "active_shards" : 12,
  "relocating_shards" : 0,
  "initializing_shards" : 0,
  "unassigned_shards" : 0,
  "delayed_unassigned_shards" : 0,
  "number_of_pending_tasks" : 0,
  "number_of_in_flight_fetch" : 0,
  "task_max_waiting_in_queue_millis" : 0,
  "active_shards_percent_as_number" : 100.0
}
```

And you can see the cluster status from the main screen <https://cloud.elastic.co/deployments>.

Click on the name you gave the deployment. Here I called mine **Analytics**.



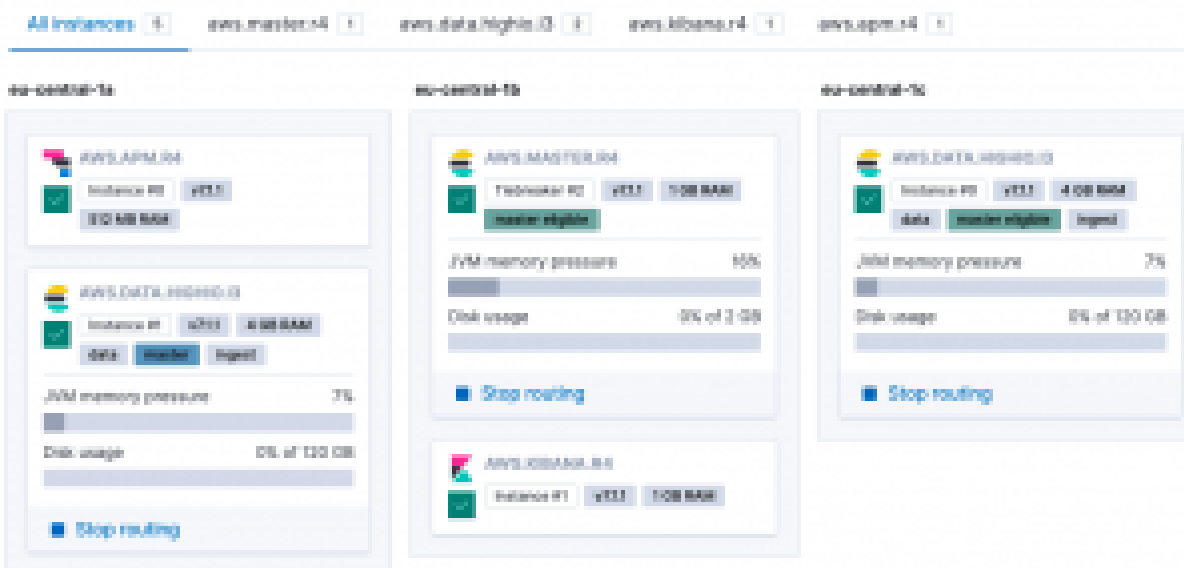
You can see that it has created Elasticsearch, Kibana, and APM, which is their monitoring tool.



Here you can see that I have 5 VMs. Or they could be running in containers. I am not sure.

There's not a lot of memory on these nodes. Which, if you have worked with AWS before, makes sense as the larger the machine the higher the subscription fees.

Instances



If you click Launch ElasticSearch, it simply runs the curl command that gives you information about the instance. But basic authentication is turned on. So the nginx web server that serves as the front end will ask you to login:

Sign in

<https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243>

Username

Password

Cancel

Sign In

Then in the browser it gives you the same data we extracted using curl above.

```
{
  "name" : "instance-0000000001",
  "cluster_name" : "58571402f5464923883e7be42a037917",
  "cluster_uuid" : "x4gHtdPbTiiFlcxWkjDJsQ",
  "version" : {
    "number" : "7.1.1",
    "build_flavor" : "default",
    "build_type" : "tar",
    "build_hash" : "7a013de",
    "build_date" : "2019-05-23T14:04:00.380842Z",
    "build_snapshot" : false,
    "lucene_version" : "8.0.0",
    "minimum_wire_compatibility_version" : "6.8.0",
    "minimum_index_compatibility_version" : "6.0.0-beta1"
  },
  "tagline" : "You Know, for Search"
}
```

Here is that curl again Notice what we have put the **userid:password** into the environment variable **pwd** to avoid having to type that each time.

```
export pwd="elastic:"
```

```
curl --user $pwd
```

```
https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243/
```

Look Some (a lot of Data) Into the Environment

Now, let's stress this environment and show you how to use it by loading it with the extremely large FDA (American Food and Drug Administration) data on drug reactions. Will we use this later to build a machine learning model to try to prective the side effects of a particular drug.

```
wget
```

```
https://download.open.fda.gov/drug/event/all\_other/drug-event-0004-of-0004.js
```

on.zip

unzip drug-event-0004-of-0004.json.zip

There are 4 datasets you can download. As you can see it is quite large > 1 GB when unzipped. And this is just JSON text.

1. -rwxrwxrwx 1 ubuntu ubuntu 33M Jun 4 10:51 drug-event-0001-of-0004.json
2. -rwxrwxrwx 1 ubuntu ubuntu 79M Jun 5 10:51 drug-event-0002-of-0004.json
3. -rwxrwxrwx 1 ubuntu ubuntu 211M Jun 5 10:52 drug-event-0003-of-0004.json
4. -rwxrwxrwx 1 ubuntu ubuntu 637M Jun 4 10:52 drug-event-0004-of-0004.json

We can try to use the bulk loader against this data. But ES complains, saying the JSON file is malformed. Plus files 3 and 4 are larger than 100 MB, which is the maximum size the ES bulk loader will handle:

```
curl --user $pwd -H 'Content-Type: application/x-ndjson' -XPOST
'https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243/_bulk?pretty' --data-binary @drug-event-0004-of-0004.json
```

It will give this error:

```
{
  "error" : {
    "root_cause" : )\n at "
  }
},
```

Load the Data Using Python

So we can load it the slow way by running this python code below. This program will run slow because it connects to the API one time for each JSON record. It will take days to load it into this instance. But we can load enough of it in a few hours to run some analytics against it, which we will illustrate in subsequent posts.

Change the open statement to pick up the name of the file you downloaded:

```
open('drug-event-0004-of-0004.json')
```

And the URL of your instance. Keep **/fda/_doc** on the end, as **fda** is the instance we will use and **/fda/_doc** is the index type.

```
url =
'https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243/fda/_doc/'
```

Here is the code:

```
import json
```

```

import requests
import uuid

def fdict(d):
    ky =

    headers = { 'content-type': 'application/json' }

    url =
"https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243/fda/_doc/"

    data={}
    for i in ky:
        data = d

    response = requests.put(url + str(uuid.uuid4()) , headers=headers,
json=data, auth=('elastic', '

```

You can list some of the data like this:

```

curl --user $pwd -H 'Content-Type: application/json' -XGET
https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243/fda/?pretty

```

And list all indices in this cluster to get the document count:

```

curl --user $pwd -XGET
https://58571402f5464923883e7be42a037917.eu-central-1.aws.cloud.es.io:9243/_cat/indices?v

```

As you can see, the data is large. 47,939 documents is 1.9 GB of storage.

health	status	index	uuid	pri	rep	docs.count
docs.deleted	store.size	pri.store.size				
green	open	fda	DRQnRqnmQsaRcsxHVIT1pg	1	1	47939
1.9gb		1003.6mb				0

Kibana

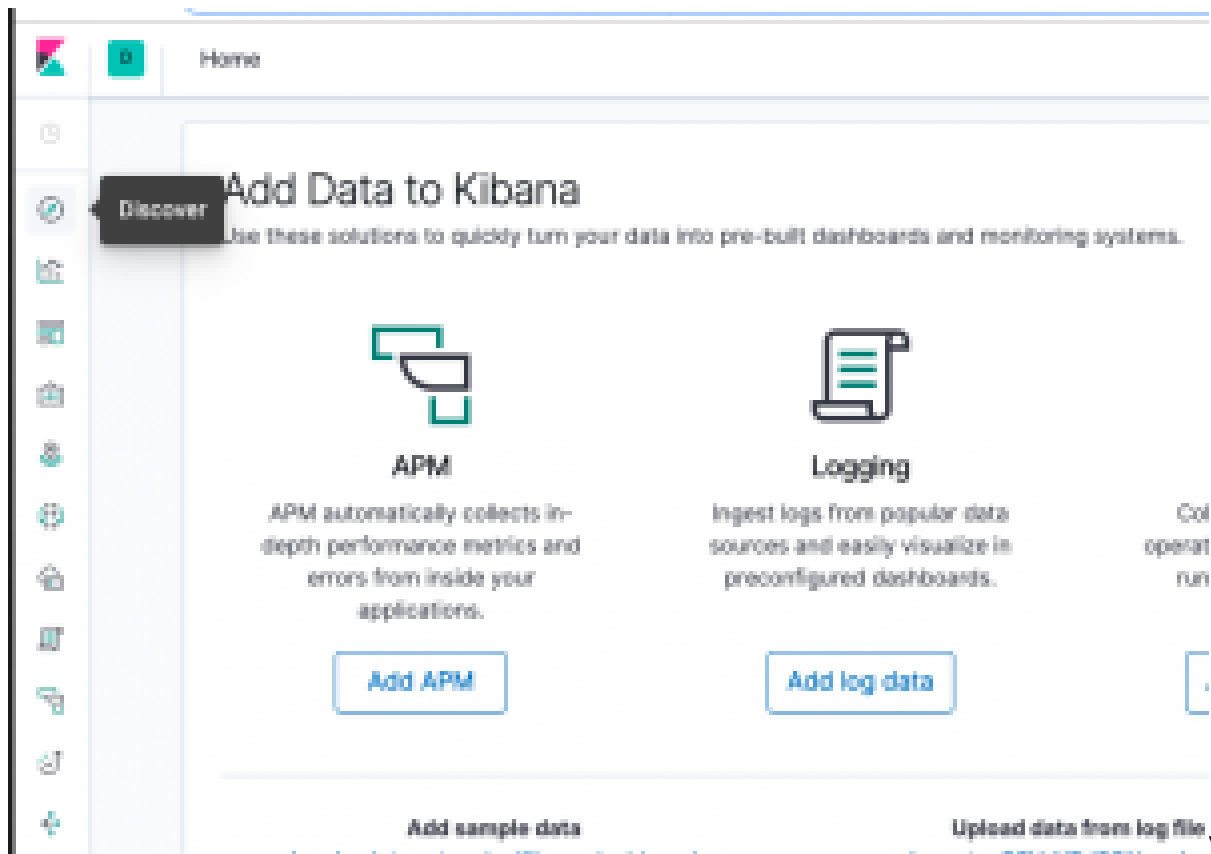
Now let's look at the same data with Kibana. Click on the Kibana link from the main screen. In our case it is

<https://78f7165792d54709b8ec37f3d80ed854.eu-central-1.aws.cloud.es.io:9243/app/kibana>

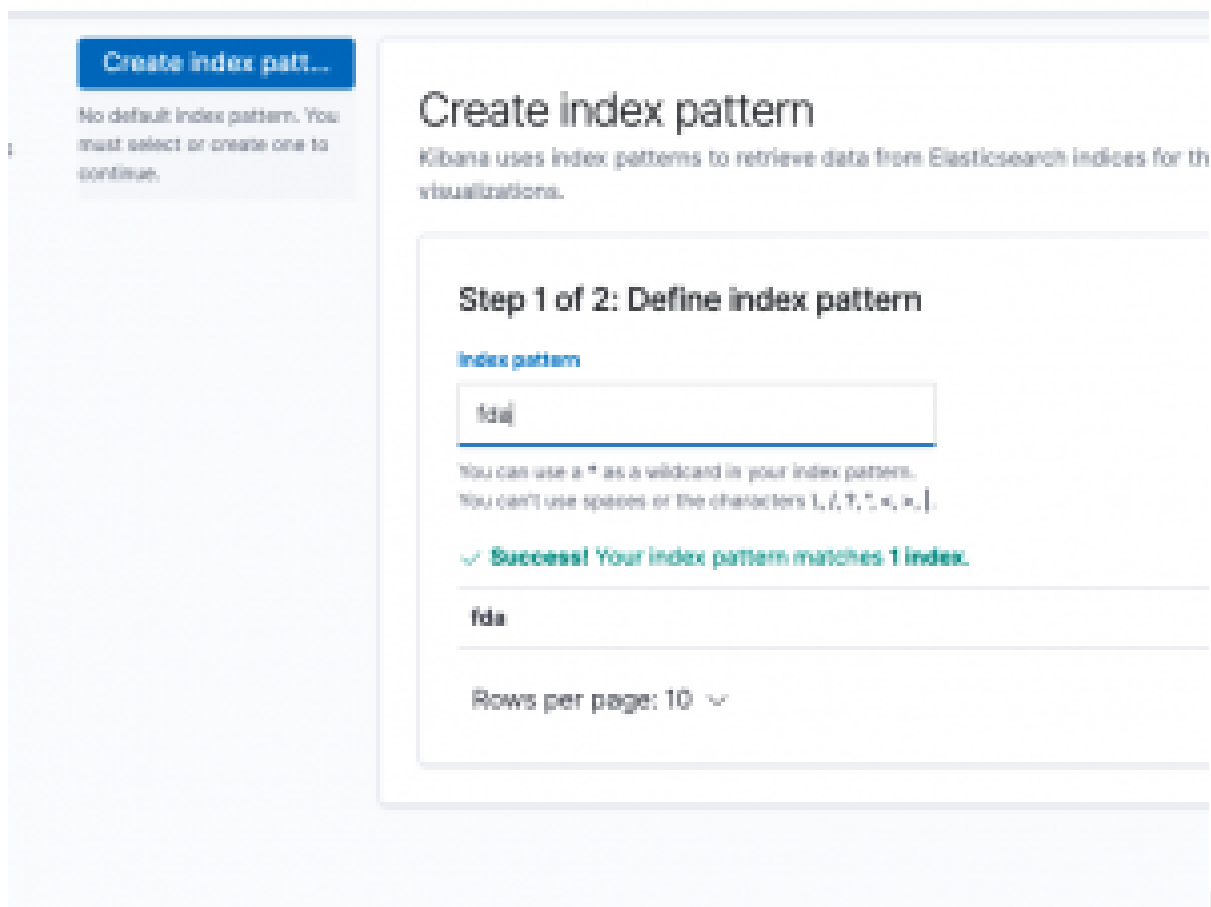
If you're new to Kibana you have to create an **Index Pattern** before you look at any of the data. which you will do using the Discover screen.

The Index Pattern is the name of the index, in this case **fda**, or a wildcard that matches the index name, like **fda***.

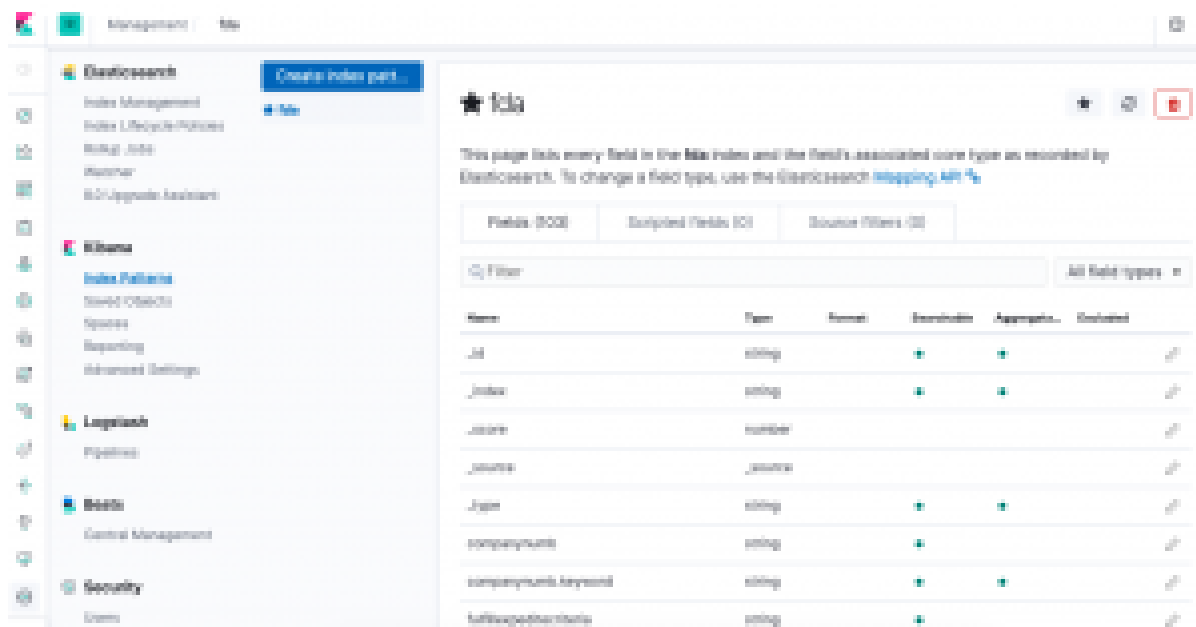
You click the **discover** button in Kibana to see this.



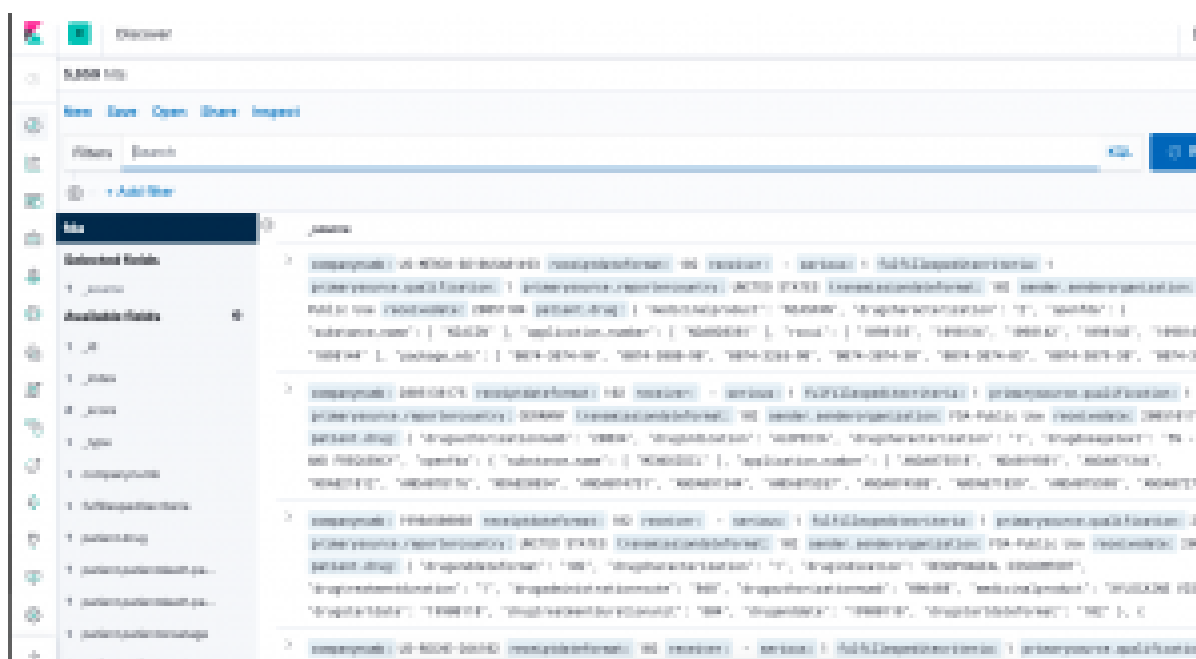
When you first open Kibana you create the first Index Pattern. Elasticsearch will parse the JSON documents it finds there and show you the schema it found. The schema is the Index Pattern.



Kibana will churn for a few seconds and then show you the fields it has auto discovered.



Below it shows the first few records. Below we click on one record to expand it.



Each JSON record from the FDA is enormous. A small part looks like this:

					<pre> { "medicinalproduct": "INSULIN", "drugcharacterization": "1" } </pre>
6	patient.patientonsiteage	63			
6	patient.patientonsiteageunit	801			
6	patient.patientsex	2			
7	patient.reaction	<pre> { "reactionmeddrug": "DRUG TOXICITY" }, { "reactionmeddrug": "GASTRIC VARICES" }, { "reactionmeddrug": "HEPATIC CIRRHOSIS" }, { "reactionmeddrug": "PORTAL HYPERTENSION" }, { "reactionmeddrug": "SYSTEMIC LUPUS ERYTHEMATOSUS" }, { "reactionmeddrug": "VARICES OESOPHAGEAL" } </pre>			
7	primarysource	-			

From here you could connect logstash to it, to ingest some application or hardware logs. We will explain that and how to use analytics in subsequent posts.