

Projet scala :

Traitement en temps réel avec spark streaming

Réaliser par :

BENAICH Ayyoub

Nous ajoutons spark dépendances pour l'utiliser comme une librairie dans notre application scala

```
build.sbt x
1  ThisBuild / version := "0.1.0-SNAPSHOT"
2
3  ThisBuild / scalaVersion := "2.12.10"
4
5  lazy val root = (project in file("."))
6  .settings(
7    name := "simple"
8  )
9
10 libraryDependencies ++= Seq(
11   "org.apache.spark" %% "spark-core" % "2.4.6",
12   "org.apache.spark" %% "spark-sql" % "2.4.6",
13   "org.apache.spark" %% "spark-mllib" % "2.4.6",
14   "org.apache.spark" %% "spark-streaming" % "2.4.6" % "provided",
15   "org.scala-sbt" %% "util-logging" % "1.3.0-M2",
16   "org.elasticsearch" %% "elasticsearch-spark-20" % "7.16.2"
17 )
```

Importation des bibliothèques nécessaires pour le projet

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types._
import org.apache.spark.streaming
import org.apache.spark.sql.functions._
import org.apache.log4j._
import org.elasticsearch.spark.sql._
```

Création d'une spark session

```
object Main {

  def main(args: Array[String]): Unit = {

    val spark = SparkSession.builder().master("local[*]").appName("first")
      .config("spark.es.nodes", "localhost")
      .config("spark.es.port", "9200")
      .getOrCreate()
    Logger.getLogger("org").setLevel(Level.ERROR)
  }
}
```

Création du dataframe

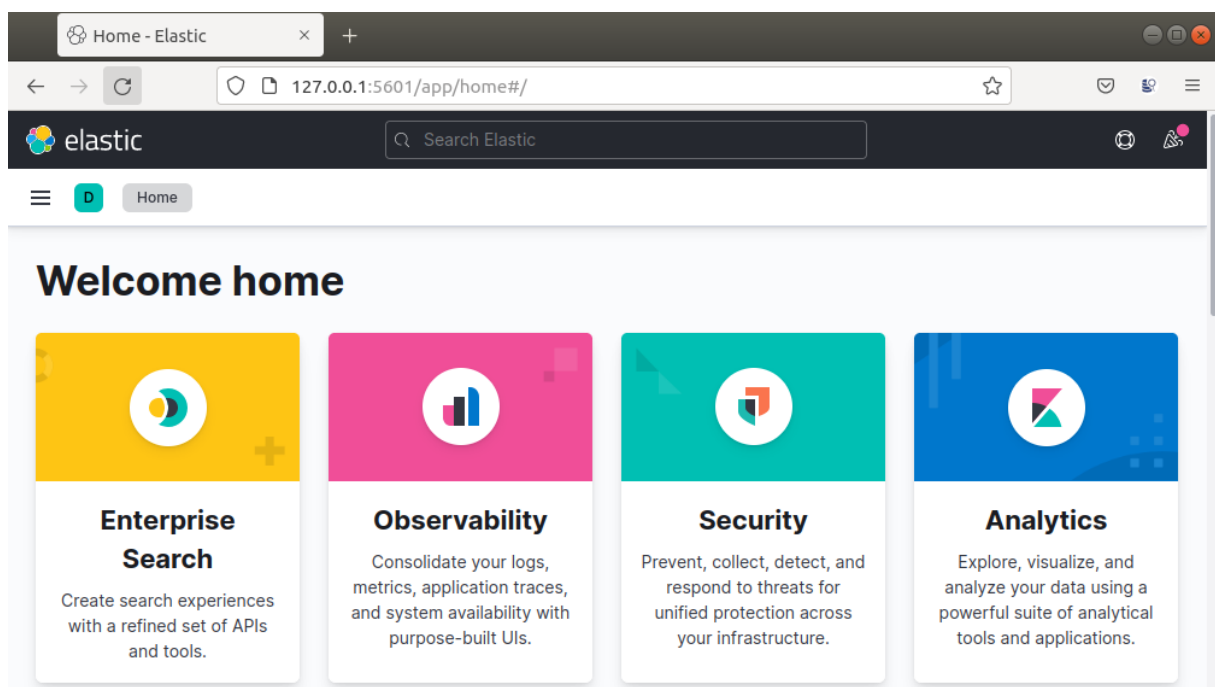
```
//Create DataFrame
val StreamDF = spark
  .readStream.option("delimiter", " ")
  .schema(schema)
  .csv(path = "/home/benaich/Documents/simple/logs")
StreamDF.createOrReplaceTempView("SDF")
val outDF = spark.sql(sqlText = "select * from SDF")
```

Transfère du dataframe vers elasticSearch en mode streaming

```
//write DF to elasticSearch
var query = outDF.writeStream
  .outputMode(outputMode = "append")
  .queryName(queryName = "writing_to_es")
  .format(source = "org.elasticsearch.spark.sql")
  .option("checkpointLocation", "/tmp/")
  .option("es.resource", "logs/doc")
  .option("es.nodes", "localhost")
  .start()

query.awaitTermination()
```

elasticSearch visualisation

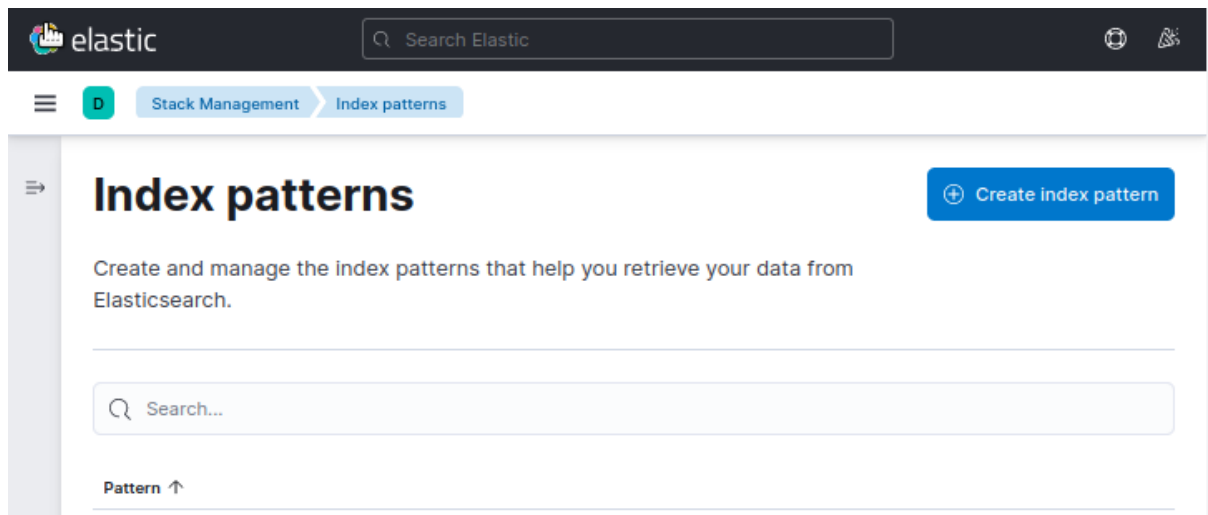


Lancement du log

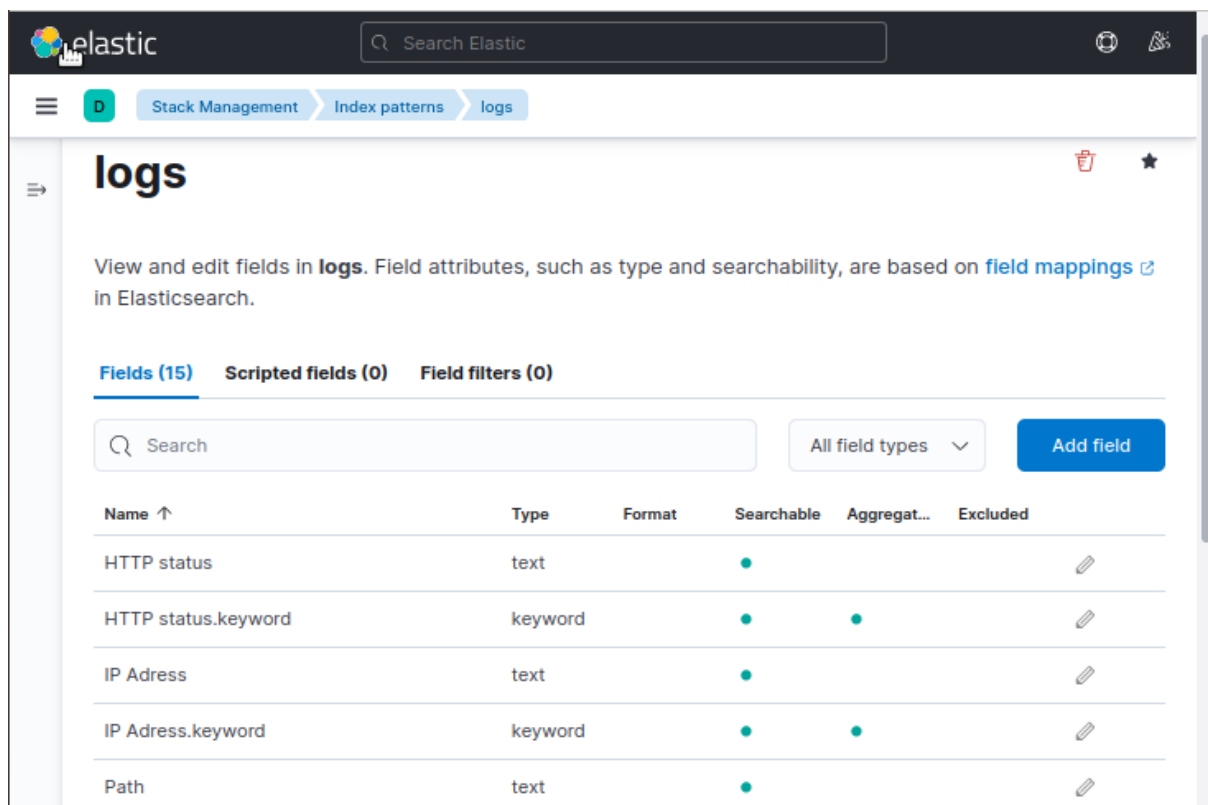
```
ben@tmp:~/Documents/simple$ python3 log-generator.py
```

Puis en transfère les log vers le répertoire

```
ben@tmp:~/Documents/simple$ tail -f /tmp/log-generator.log > /home/benaich/Documents/simple/logs/mydata.txt
```



Récupération du données par sur kibana



On a reçu 6096 enregistrements ,on va faire la visualisation de ces enregistrements

The screenshot shows the Elasticsearch Discover interface. At the top, there's a search bar with 'Search Elastic' and a 'Discover' button. Below the search bar, there's a 'Search' input field and a 'KQL' button. A '+ Add filter' button is also visible. The main area displays 'logs*' as the selected index, with '6,096 hits' shown. On the left, there's a 'Filter by type' section showing '0' filters. Below that, a list of 'Available fields' is shown, including '_id', '_index', '_score', '_type', 'HTTP status', 'IP Address', 'Path', and 'Protocol'. The main content area shows a list of documents, each with fields like 'HTTP status', 'IP Address', 'Path', 'Protocol', 'URL', and '_id'. The documents are sorted by '_score' in descending order.

elastic Search Elastic

Discover

Options New Open Share Inspect Save

Search KQL Refresh

+ Add filter

logs* 6,096 hits

Search field names

Filter by type 0

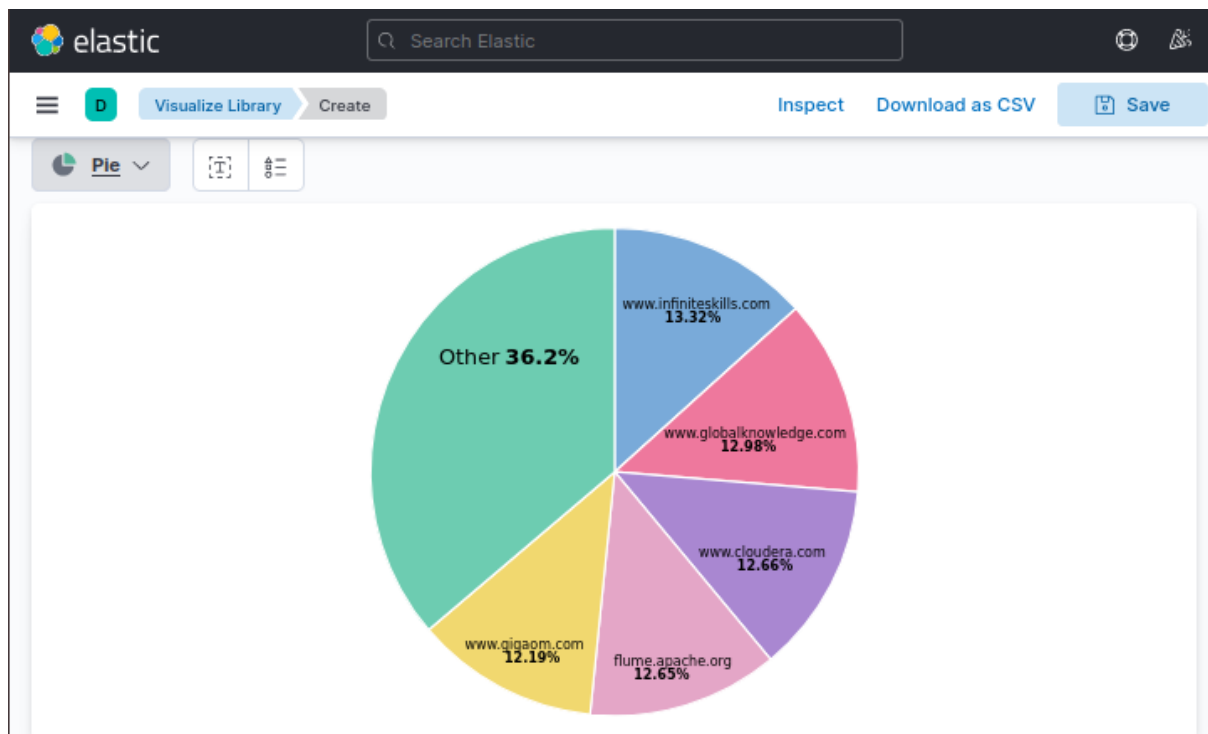
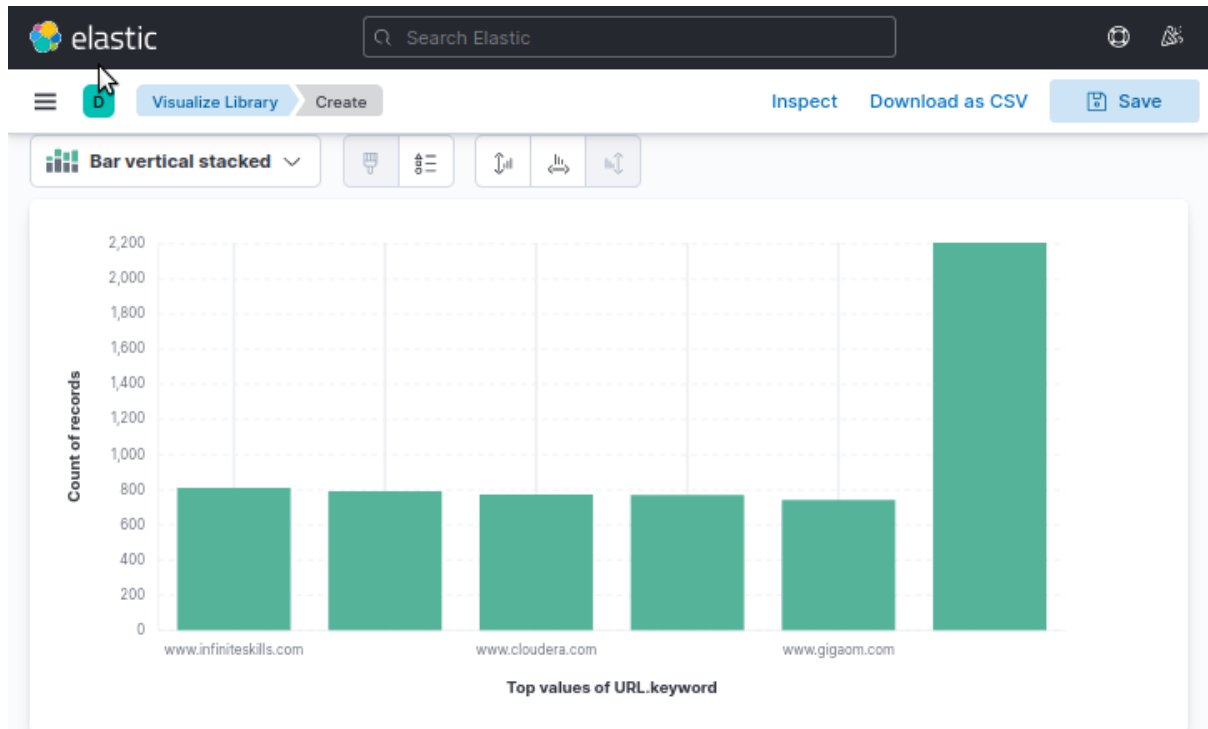
Available fields 9

- _id
- _index
- _score
- _type
- HTTP status
- IP Address
- Path
- Protocol

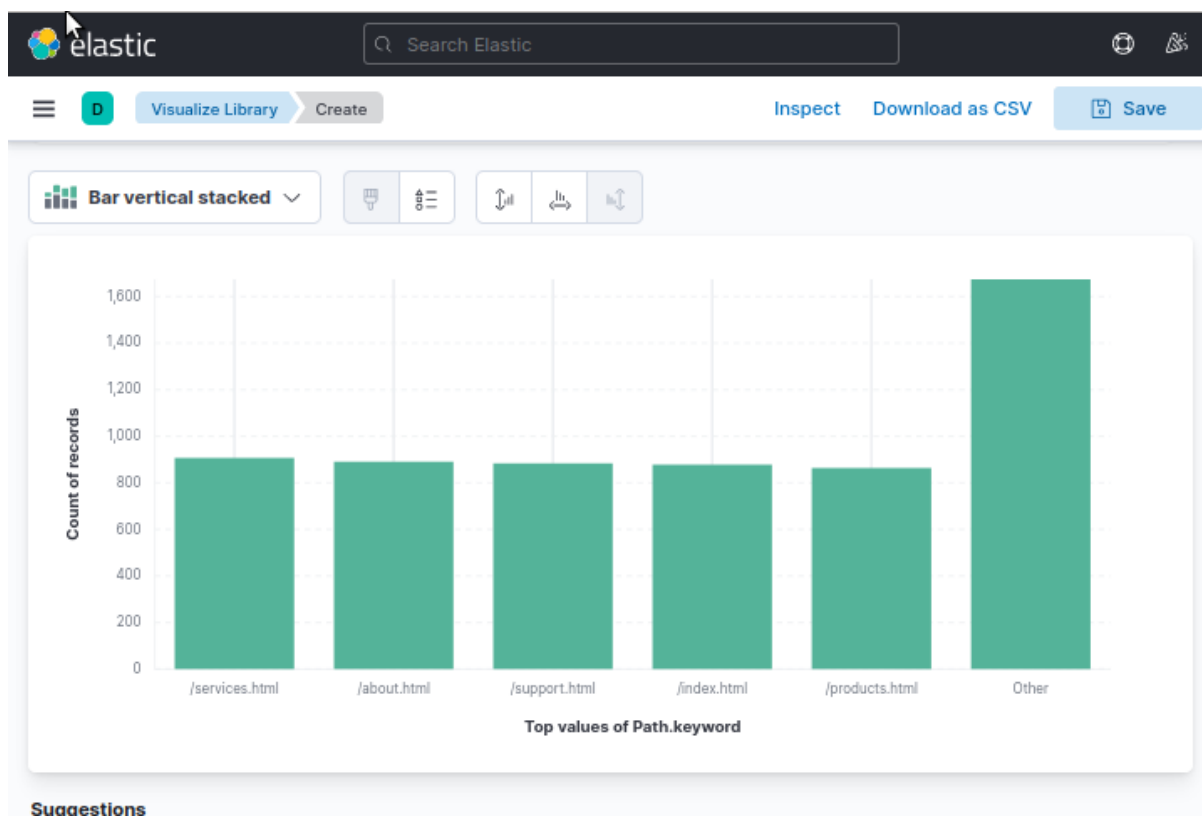
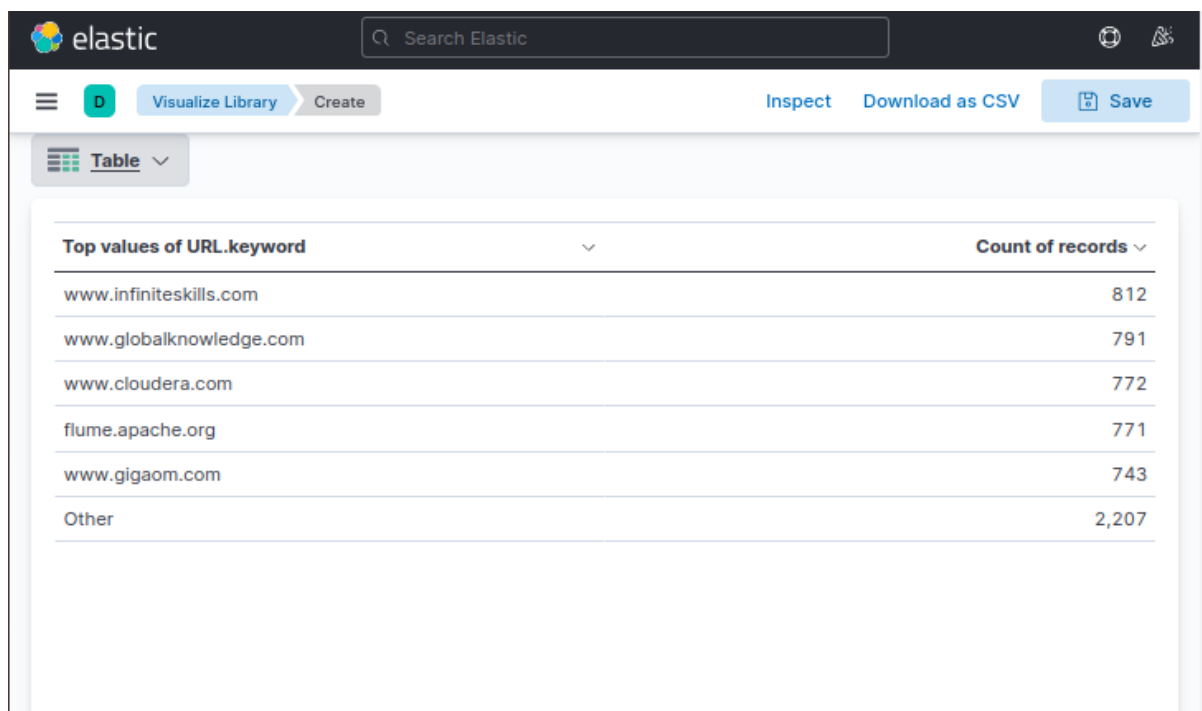
Document

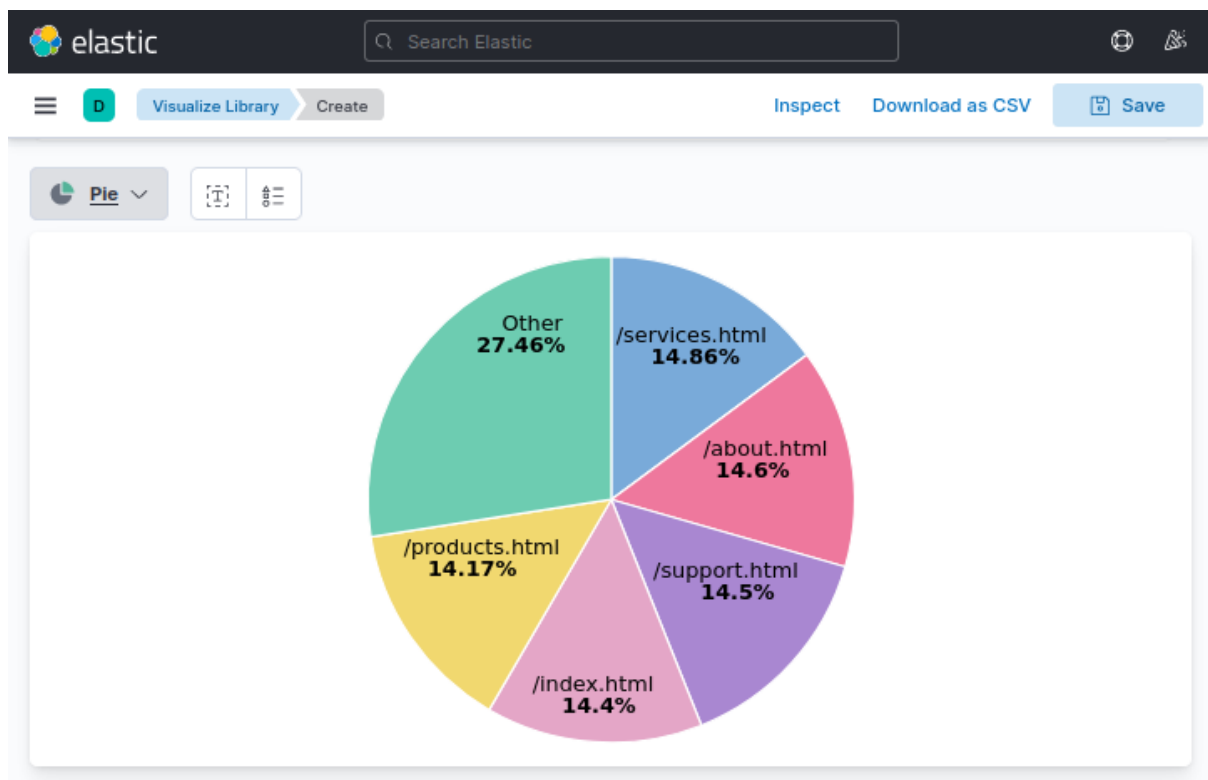
- > HTTP status: 411 IP Address: 43.219.164.83 Path: /contact/submit.html
Protocol: HTTP URL: www.cloudera.com _id: a4oXWX4BtknNtBCe18Ys
_index: logs _score: 1 _type: doc
- > HTTP status: 204 IP Address: 119.186.68.95 Path: /index.html
Protocol: HTTP URL: www.cloudera.com _id: bIoXWX4BtknNtBCe18Yt
_index: logs _score: 1 _type: doc
- > HTTP status: 307 IP Address: 188.215.90.244 Path: /services.html
Protocol: HTTP URL: flume.apache.org _id: bYoXWX4BtknNtBCe18Yt
_index: logs _score: 1 _type: doc
- > HTTP status: 300 IP Address: 79.108.236.241 Path: /contact/submit.html

Visualisation :



Nombres des enregistrement





elastic Search Elastic

Visualize Library Create Inspect Download as CSV Save

Table

Top values of Path.keyword	Count of records
/services.html	906
/about.html	890
/support.html	884
/index.html	878
/products.html	864
Other	1,674

