

The rise of ElasticSearch

Zoltan Mathe

Wednesday 23rd May, 2018



1 Introduction to ElasticSearch

- Elasticsearch
- Architecture
- Elasticsearch Clients
- Products
- ELK

2 DIRAC and ElasticSearch

- Reminder
- EL integration to DIRAC

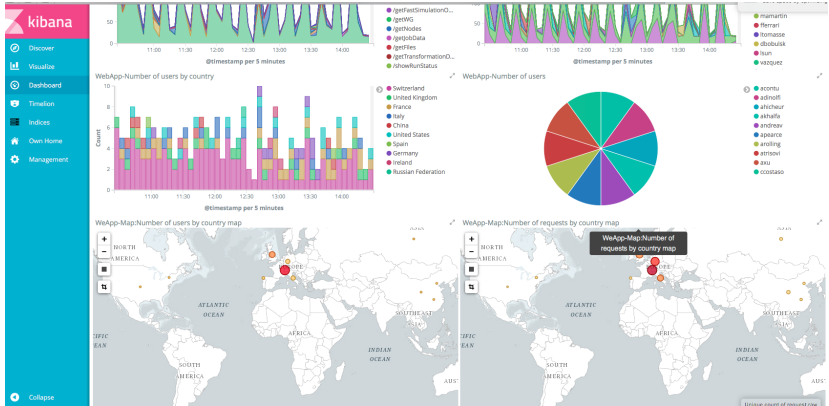
3 Future usage of EL

- distributed search and analytic engine based on Apache Lucene
- developed in Java
- easy to install and to scale to PB of data
- near real time system
- stable well maintained
- open source
<https://github.com/elastic/elasticsearch>

- the EL **cluster** consists of different **nodes**:
 - master node: control the cluster
 - data node: stores the data
- data is stored in schema less JSON **documents**
 - each document belong to a **document type**
 - **mapping** defines how a document and the fields are stored.
- documents are stored in EL **index**
 - index is split into **shards**
 - each shard may be on a different node in a cluster
 - every shard is a self contained Lucerne index
- cluster contains *primary* and *replica* shards for reliability
 - number of primary shards can not be changed later. If it needs to be changed, the data needs to be re-indexed.
 - more replica can be created
- RDBMS: row – > document, table – > document type, database – > index

- multi language, officially supported: Java, javascript, Groovy, .NET, PHP, Perl, Python, Ruby
- REST API to interact with data

- Kibana for visualizing the Elasticsearch data
- Logstash for server side data processing pipeline
- Machine Learning for automatically model the behaviour of Elasticsearch data
- ES-Hadoop for indexing Hadoop data into EL
- and many more



- Accounting system is not for real time monitoring:
 - not efficient for handling time-series data
 - does not scale to hundred of million rows
- Our study based on the following technologies:
 - InfluxDB: distributed time series database
 - OpenTSDB: distributed time series database based on HBase
 - Elasticsearch: distributed search and analytic engine
 - Grafana: metric dashboard and graph editor for InfluxDB, Graphite and OpenTSDB
 - Kibana: flexible analytic and visualization framework for Elasticsearch
- decided to use **Elasticsearch** and develop the DIRAC **Monitoring System**

- Accounting web application for data visualization and EL for storing data
- How?:
- develop the Monitoring system, which is based on:
 - Query DSL (Domain Specific Language) based on JSON to define queries
 - elasticsearch-dsl is python library for writing and running queries against Elasticsearch
 - existing DIRAC libraries
- Monitoring system is available form v6r16 release
- Possible usage of the Monitoring system
 - WMS monitoring
 - Component monitoring (required improvements)

- Job traceability (GSoC 2018 project)
http://hepsoftwarefoundation.org/gsoc/2018/proposal_LHCbJobTraceability.html
- centralized logging:
 - send service and agent logs to EL
 - use Kibana for monitoring
- Component monitoring
 - replace gMonitor
- for your experiment specific use?

Thank you! Questions?