Log analysis and classification of CERN control systems

Z. Kassabov
CERN EN-ICE
August 19, 2014
WinCC OA / PVSS

- Application for control systems used by:
  - LHC experiments
  - Many other systems at CERN (e.g. electricity).

- Has been successfully used for many years for increasingly complex systems.

- Produces lots of log messages.
  - Too many to be analyzed by humans.
  - Can we use them to understand the system behavior?
WinCC OA / PVSS

- Application for control systems used by:
  - LHC experiments
  - Many other systems at CERN (e.g. electricity).

- Has been successfully used for many years for increasingly complex systems.

- Produces lots of log messages.
  - Too many to be analyzed by humans.
  - Can we use them to understand the system behavior?
Application for control systems used by:
- LHC experiments
- Many other systems at CERN (e.g. electricity).

Has been successfully used for many years for increasingly complex systems.

Produces lots of log messages.
- Too many to be analyzed by humans.
- Can we use them to understand the system behavior?
Z. Kassabov

Log analysis and classification of CERN control systems
Analyzing the logs

- Can we aggregate and visualize the flow of log messages?
- Can we detect anomalies in the system?

We need a log monitoring system.

- ...but mustn’t consume too many resources from production machines.
- We analyze the performance of the *ELK Stack*.
  
  - Logstash: Reads and parses the log files.
  - Elasticsearch: Search server and data store.
  - Kibana: Web interface for Elasticsearch.
Analyzing the logs

- Can we aggregate and visualize the flow of log messages?
- Can we detect anomalies in the system?

We need a log monitoring system.

- ...but mustn’t consume too many resources from production machines.
- We analyze the performance of the ELK Stack.

  - Logstash: Reads and parses the log files.
  - Elasticsearch: Search server and data store.
  - Kibana: Web interface for Elasticsearch.
Analyzing the logs

- Can we aggregate and visualize the flow of log messages?
- Can we detect anomalies in the system?

We need a log monitoring system.

- ...but mustn’t consume too many resources from production machines.
- We analyze the performance of the *ELK Stack*.

  - **Logstash**  Reads and parses the log files.
  - **Elasticsearch**  Search server and data store.
  - **Kibana**  Web interface for Elasticsearch.
Analyzing the logs

Can we aggregate and visualize the flow of log messages?
Can we detect anomalies in the system?

We need a log monitoring system.
...but mustn’t consume too many resources from production machines.
We analyze the performance of the *ELK Stack*.

- **Logstash** Reads and parses the log files.
- **Elasticsearch** Search server and data store.
- **Kibana** Web interface for Elasticsearch.
Expected result: The Kibana interface

Z. Kassabov

Log analysis and classification of CERN control systems
Measuring Logstash performance

- Understand what are the relevant metrics.
- Compare with resources available on production systems.
- Set up a testing system.
- Custom Python script made to monitor performance.
- Another script (based on Fabric and Zdaemon) used to automate everything.
Measuring Logstash performance

- Understand what are the relevant metrics.
- Compare with resources available on production systems.
- Set up a testing system.
- Custom Python script made to monitor performance.
- Another script (based on Fabric and Zdaemon) used to automate everything.
Measuring Logstash performance

- Understand what are the relevant metrics.
- Compare with resources available on production systems.
- Set up a testing system.
- Custom Python script made to monitor performance.
- Another script (based on Fabric and Zdaemon) used to automate everything.
Log analysis and classification of CERN control systems
Log analysis and classification of CERN control systems
Results of the measurement

Results show *high* resource usage on a *production* system:

- 400 − 700 Mb of RAM
- \( \sim 10\% \) of a CPU core (up to 200\%) on bursts of errors.
- Still considered acceptable compared to the alternatives.
  - It’s a small percentage of the freely available resources.
Results show high resource usage on a production system:

- 400 – 700 Mb of RAM
- $\sim 10\%$ of a CPU core (up to 200\%) on bursts of errors.
- Still considered acceptable compared to the alternatives.
  - It’s a small percentage of the freely available resources.
Can we predict whether a log message is important?

Idea

Use the same technique as spam filters (or priority inboxes): Naive Bayes Filtering.

- Easy to understand and deploy.
- Has proven to work in the past.
- ...but can’t pick correlations.
Can we predict whether a log message is important?

Idea

Use the same technique as spam filters (or priority inboxes): Naive Bayes Filtering.

- Easy to understand and deploy.
- Has proven to work in the past.
- ...but can’t pick correlations.
Can we predict whether a log message is important?

Idea

Use the same technique as spam filters (or priority inboxes): Naive Bayes Filtering.

- Easy to understand and deploy.
- Has proven to work in the past.
- ...but can’t pick correlations.
Can we predict whether a log message is important?

Idea

Use the same technique as spam filters (or priority inboxes): *Naive Bayes Filtering*.

- Easy to understand and deploy.
- Has proven to work in the past.
- ...but can’t pick correlations.
Future work

Naive bayesian filters cannot pick correlations between different features, or patterns in events.

- Another team is working on pattern recognition.
- Could it be combined with pattern analysis?
  - Could be implemented by computing higher order features (e.g. Number of similar log messages in the last 5 minutes).
Future work

Naive bayesian filters cannot pick correlations between different features, or patterns in events.

- Another team is working on pattern recognition.
- Could it be combined with pattern analysis?
  - Could be implemented by computing higher order features (e.g. *Number of similar log messages in the last 5 minutes*).