Logging Statements analysis on Apache Hadoop

Bhagya C

24 Aug 2020

Agenda

- What I have done so far
- Observations and conclusions
- Ideas/Actions Plans

What I have done so far

- Understanding Problem statement
  - Why logging is important in coding
  - Logging conventions
  - Importance of logging levels

What I have done so far

- Static Analysis
    - Total number files
    - Lines of code
    - Code distribution over programming languages

What I have done so far

- Analysis of distribution log libraries and log level usages
- Analysis of messages used in the log level - basic similarity analysis using spacy
- Average length of messages
- Analysis on most frequent words in logging messages

What I have done so far

- Analysis on dependencies of Log statements usages and source code
    - Try
    - Catch
    - If
    - Else
- Analysis on the dependencies of Log level and the source code
- Analysis of usage of multiple logging statements together

Observations and Conclusion

- Challenges while analysing
  - No **conventional system** for using logging statements
  - **Changes** occured in logging statements over time
  - **Inconsistencies** in the usages of logging libraries and logging level
- Logging statements are **inevitable** part of engineering practices
- Domain knowledge and expertise in the domain will affect the method of following logging statements
- Changing logging statements will cause lots of **error** in the production environment

Observations and Conclusion

- There are **many factors affecting** the possibility of **using log statements and log levels in the coding practices**
  - Context of the previous code lines (which type of condition or method)
  - Expertise
  - Type of logging libraries
  - And previous usages of logging statements

Ideas/Action Plans

- Best logging practices will always help in the quality of code as well as finding the issues at the run time
- Logging message suggestion using language generation methods seems promising in this domain (real time suggestions)
- Logging Level predictions are also valid
- Both of the problem statement should incorporate domain expertise in the solution to improve the trustability of the model

Thank you