Slide 1: What I have done so far

Initially, I had completed research on the problem statements (I connected with a couple of developers and did a very basic literature review on this)basically I was trying to answer the following questions.

1. Why Logging is important in codding: It helps to trace the runtime behaviour of the software system. These logging statements play a critical role in monitoring system statuses.
2. Logging conventions: Generally, developers add the logging statements in temporary manner which ended up in fragile logging code (some of them contain insufficient logging information on the other hand others will contain excessing logging information). But anyways the domain knowledge of the developers is something we can leverage and to insight for understanding the existing practices.
3. Importance of logging level. : Basically, they show the urgency of the log (rank the importance of the log statement). It helps to sort the information and control the amount of information to be stored.

Once I understood the problem statement then I started analysing the logging statement of the given project Apache Hadoop

I run some static analysis to get an overall idea of the project

As a primary analysis I looked up the total number of files in the project and a total number of lines of code then programming language distribution of the code over the project.

Then limited my analysis on java only

These are the glimpses of my analysis

```
 Total number of files in the repository -> !git ls-files | wc -l
Total number of java files -> !git ls-files | grep "\.java$" | wc -l
Total number of lines of code in java file -> !git ls-files | grep "\.java$" | xargs cat | wc -l
```

Initially, I run some terminal commands then I thought of running all of then as a shell script, later on, realised that jupyter notebook will help for both visualisation and coding.

Then I jumped into the analysis of the logging statements used in the java code.

As I mentioned above I limited my analysis by choosing standard logging libraries such as **Logback / Log4j / Log4j2 / SLF4J**

And the logging level used in these libraries. Hence calculated the average log level distribution throughout the java codebase.

The analysis showed which all level used over others

Normally each log statement will contain a timestamp, contextual information, message and level. I came to the irrefutable conclusion that messages have a very vital role in the log statements

- I extracted all the logging statements
- And preprocessed the data to separate out the messages in them
- Then I created a word cloud to analyse the most frequent words used in the messages

- As from our previous study - info messages like log debug enabled and statechangelog enabled and encountered exceptions are examples of most frequent words
- And then followed by a similarity testing using spacy similarity (which is calculated by cosine similarity and the word vectors are trained from glove model- unlike the word to vec glove considered the word co-occurrence)
- I understood that log messages are pivotal information sources and we have to investigate more on that

Then I move on to analyse the dependency of log statements to other source code. In order to address the question of where to log, we must have to consider the relative location of the logging statement in the codebase.