BEHAVOX

Troubleshooting of distributed system

Behavox platform is a complex distributed system which processes huge amounts of unstructured data. The platform must be available 24/7 for our clients and in case of an issue a support engineer should provide a detailed report with technical context and potential root cause analysis.

The test task consists of three steps:

1. Understanding a complex distributed system. The goal is to evaluate how far the candidate can go without any help in learning a new technology using just publicly available resources.
2. Apply the knowledge gained in step 1 to a real problem. The goal of support engineers is to help Development/DevOps/Delivery teams fix production bugs fast and minimize the impact on client's operations.
3. Usage of troubleshooting tools. In the case of the Behavox platform, troubleshooting is mainly based on an understanding of the system architecture, and on log analysis. Distributed systems produce a lot of logs and it's critical to understand how to analyze them using available tools.

Apache HBase is one of the storage systems that is heavily used in the Behavox platform. Log analysis is implemented using Kibana. The goal is to use Apache HBase as the target testing platform, and to use Kibana for a simple statistical analysis.

1. Please study the HBase architecture guide or the whole specification. The scope is up to you. https://hbase.apache.org/book.html#_architecture
2. Please study the following bug: https://issues.apache.org/jira/browse/HBASE-14498
    a. Please prepare a visual diagram of the problem flow in this bug. The bug is about incorrect behaviour in a distributed system, affecting high availability. The diagram should clearly demonstrate where the root cause of the problem is.
    b. Please provide your version of the description of the bug. Keep in mind that some details may be missing in the initial bug report and were added in the follow-up discussion in the issue tracker.
3. Kibana. Please use the following demo cluster: https://demo.elastic.co and:
    a. Count all unique host names in filebeat-* index. Tip: use the Data table visualization. Use the time range of one year.
    b. Advanced part of the task: list all host names.

Submission deadline: 4 days from the date of receipt