Apache Flink executes a large job on the cluster. But one worker fails to execute. What went wrong inside the cluster? Which worker is affected?

Large job graph is optimized and transformed to execution graph.

A) Impossible to copy whole graph.
B) Analyzing logging data takes too long and it is hard to find the counter part in code.

The idea is to get a small subset from remote cluster and record it to a local machine. With the subset data scientists are able to investigate failures in the local environment.

Recording only specific data is hard. There must be information about a locality. What kind of data is needed to do a local replay?

How is the environment looking like? Can we control the operating systems process also? Which conditions need to be met to do a local replay?

The basic workflow is to serialize the job graph, append additional environment and application variables. Use a Flink runner to do the local replay.

The first attempt is a prototype Flink 1.0 with a modified application runtime. The Flink runtime was modified manually.

The second attempt is a prototype Flink 2.0 with automatic code insertion via bytecode instrumentation and native code post processing in a separate process.