10.27 DataRefinement代码实现 副本

Log Refinement

1.根据异常描述（“please analyze the abnormal event between 2025-06-05T17:10:04Z and 2025-06-05T17:33:04Z and provide the root cause.”）找到对应 log 文件，提取出异常发生时段内的 logs 条目。

2.Error 关键字过滤，过滤出“message”字段中含“error”的 logs 条目。

3.去除无关列，保留有用列，"service_name"，"pod_name"，"node_name"，"message"。

4.Drain3 模板提取，使用 Drain3 提取出“message”的模板，并添加新列“template”。

5.log 去重，按 pod-node-template 分组，保留每组第一条日志，并添加计数列“occurrence_count”。

6.按“occurrence_count”降序排列，并保留列"service_name", "pod_name", "node_name", "message", "occurrence_count"。

7.调用 log_agent 提炼出对故障诊断最关键、最有价值的日志。

Trace Refinement

1.根据异常描述找到对应 trace 文件，根据异常发生时段分成正常时段 traces 条目和 异常时段 traces 条目。

2.正常时段 traces 条目用于训练 Isolation Forest。按"parent_pod', "child_pod", "node_name", "operation_name"分组，异常检测器通过拟合每组 trace 条目中的“duration”字段，得到一个异常检测器。

3.异常时段 traces 条目同样按"parent_pod', "child_pod", "node_name", "operation_name"分组，用对应的异常检测器检测该时段内 “duration”异常的条目，提取每一个被检测为异常的条目。

4.traces 去重，每一组用一条 trace 代替，其中"duration"字段用每组平均值"anomaly_avg_duration"代替，并添加计数列"anomaly_count"。

5.按"anomaly_count"降序排列，并保留列"service_name", "node_name", "parent_pod", "child_pod", "operation_name", " normal_avg_duration", "anomaly_avg_duration", "anomaly_count"，其中"normal_avg_duration"，（p95...）是正常时间段内该组“duration”的平均值。

6.调用 trace_agent 提炼出对故障诊断最关键、最有价值的 trace。

Metric Refinement

实现中