Flink在云帐房的使用

部门：数据智能部
日期：2021-12-14
分享人：杨成凯

YunZhangFang.com

坚韧、正直、
合作、敬业

目 录

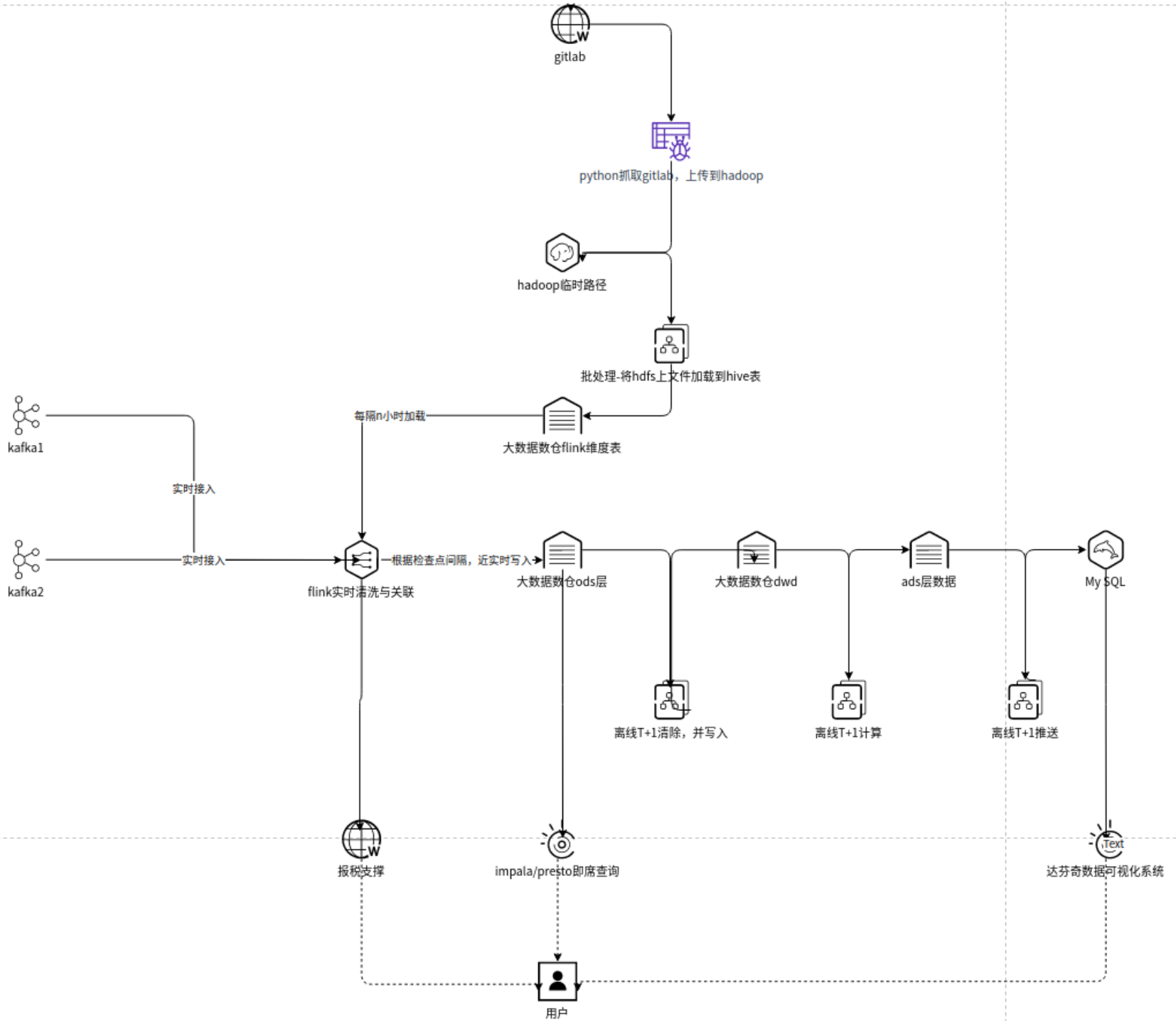
CONTENTS

云帐房
YunZhangFang.com

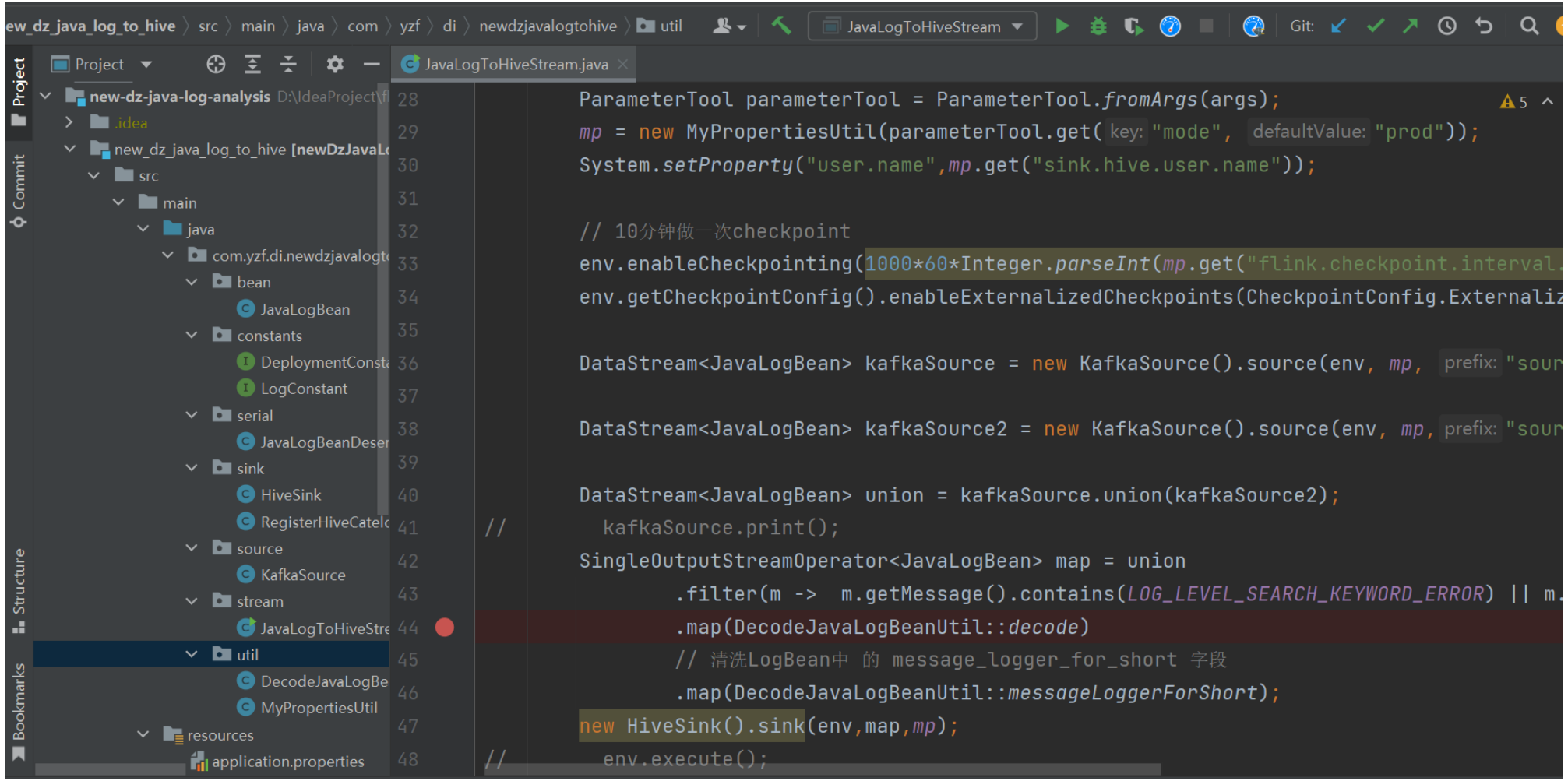
01 新代帐日志追踪

需求说明

- 从多个kafka的broker、topic中消费日志数据。过滤出WARN、ERROR级别的日志，并将每条日志与gitlab上相关信息关联，获取开发人员邮箱。将结果存储进行存储，并根据业务需求进行预警。


gitlab
python抓取gitlab，上传到hadoop
hadoop临时路径
批处理-将hdfs上文件加载到hive表
每隔n小时加载
大数据数仓flink维度表
kafka1
实时接入
kafka2
实时接入
flink实时清洗与关联
根据检查点间隔，近实时写入
大数据数仓ods层
大数据数仓dwd
ads层数据
My SQL
离线T+1清除，并写入
离线T+1计算
离线T+1推送
报税支撑
impala/presto即席查询
达芬奇数据可视化系统
用户

代码走读


ew_dz_java_log_to_hive 〉 src 〉 main 〉 java 〉 com 〉 yzf 〉 di 〉 newdzjavalogtohive 〉 util
JavaLogToHiveStream
Git:
Project
JavaLogToHiveStream.java
new-dz-java-log-analysis D:\IdeaProject\f
.idea
new_dz_java_log_to_hive [newDzJavaL
src
main
java
com.yzf.di.newdzjavalogt
bean
JavaLogBean
constants
DeploymentConsta
LogConstant
serial
JavaLogBeanDeser
sink
HiveSink
RegisterHiveCatelc
source
KafkaSource
stream
JavaLogToHiveStre
util
DecodeJavaLogBe
MyPropertiesUtil
resources
application.properties
Project
Commit
Structure
Bookmarks
28 ParameterTool parameterTool = ParameterTool.fromArgs(args);
29 mp = new MyPropertiesUtil(parameterTool.get(key: "mode", defaultValue: "prod"));
30 System.setProperty("user.name",mp.get("sink.hive.user.name"));
31
32 // 10分钟做一次checkpoint
33 env.enableCheckpointing(1000*60*Integer.parseInt(mp.get("flink.checkpoint.interval
34 env.getCheckpointConfig().enableExternalizedCheckpoints(CheckpointConfig.Externaliz
35
36 DataStream<JavaLogBean> kafkaSource = new KafkaSource().source(env, mp, prefix: "sour
37
38 DataStream<JavaLogBean> kafkaSource2 = new KafkaSource().source(env, mp, prefix: "sour
39
40 DataStream<JavaLogBean> union = kafkaSource.union(kafkaSource2);
41 // kafkaSource.print();
42 SingleOutputStreamOperator<JavaLogBean> map = union
43 .filter(m -> m.getMessage().contains(LOG_LEVEL_SEARCH_KEYWORD_ERROR) || m.
44 .map(DecodeJavaLogBeanUtil::decode)
45 // 清洗LogBean中 的 message_logger_for_short 字段
46 .map(DecodeJavaLogBeanUtil::messageLoggerForShort);
47 new HiveSink().sink(env,map,mp);
48 // env.execute();
5

YunZhangFang.com

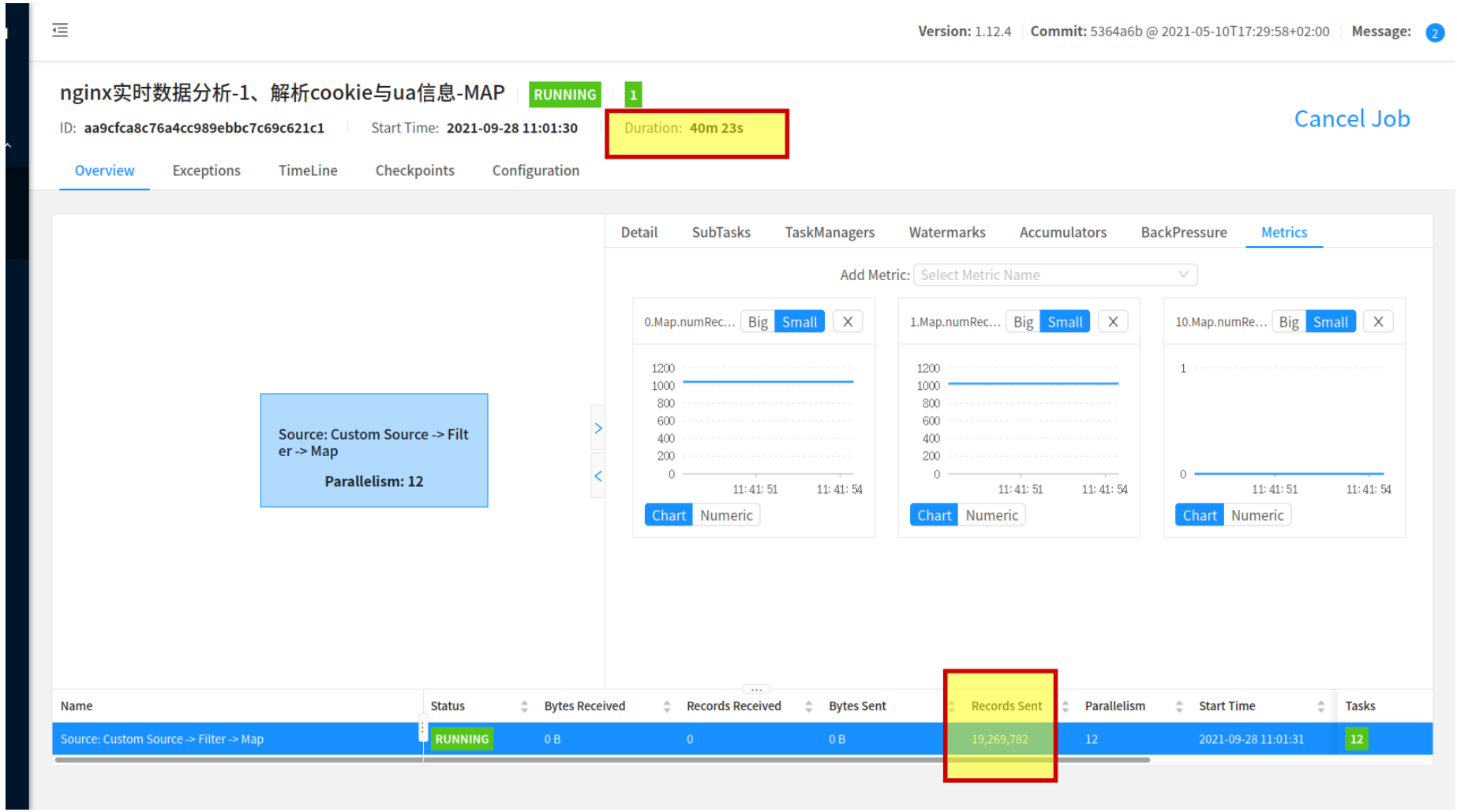
02

nginx日志实时解析

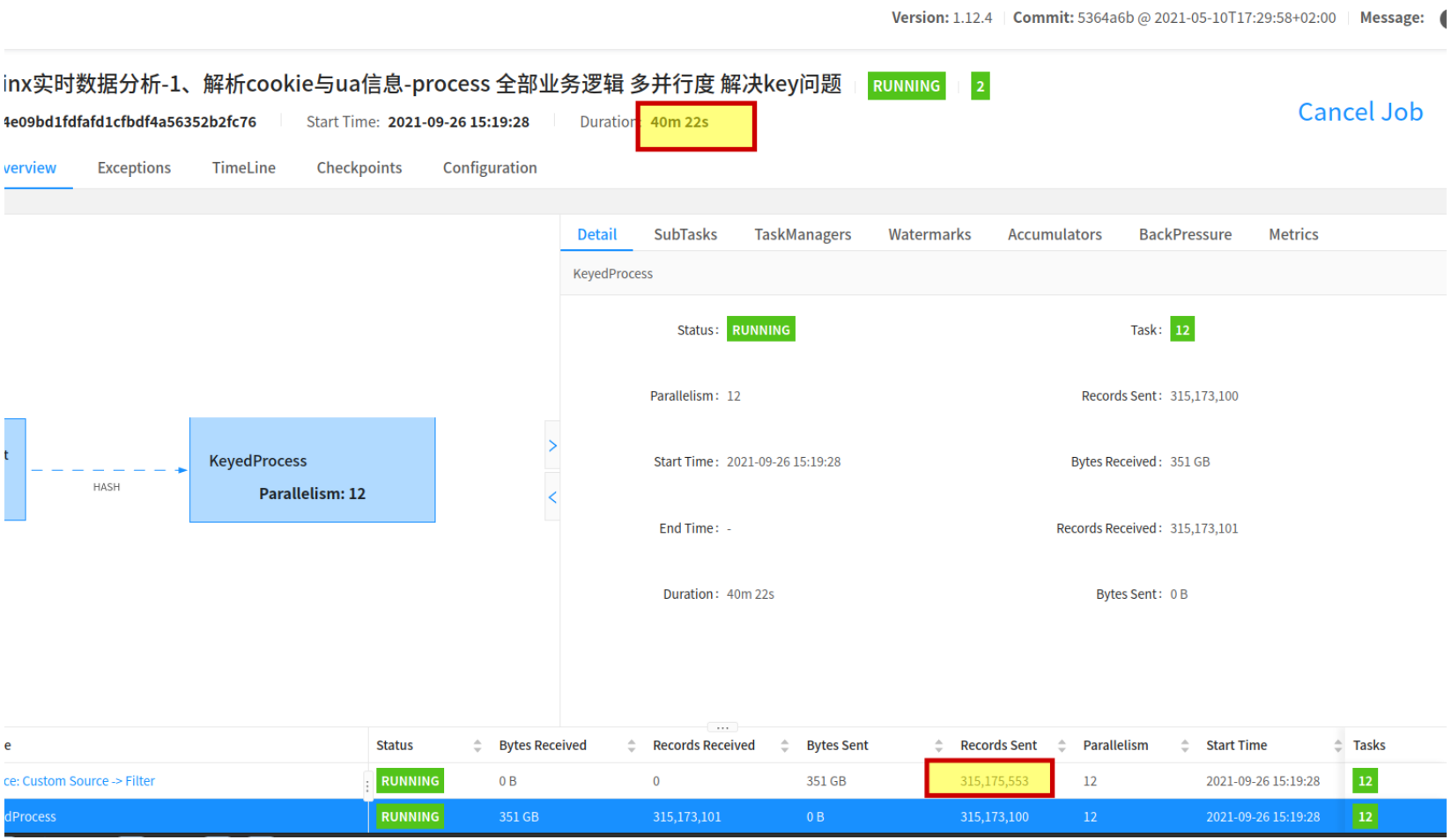
需求说明

- 从nginx中的cookie信息中，提取出用户名，所属代帐公司，进而后续统计相关模块、区域流量。

思路1 使用核心API


Version: 1.12.4 | Commit: 5364a6b @ 2021-05-10T17:29:58+02:00 | Message: 2
nginx实时数据分析-1、解析cookie与ua信息-MAP RUNNING 1
ID: aa9cfca8c76a4cc989ebbc7c69c621c1 Start Time: 2021-09-28 11:01:30 Duration: 40m 23s
Cancel Job
Overview Exceptions TimeLine Checkpoints Configuration
Detail SubTasks TaskManagers Watermarks Accumulators BackPressure Metrics
Add Metric: Select Metric Name
0.Map.numRec... Big Small
1.Map.numRec... Big Small
10.Map.numRe... Big Small
Chart Numeric
Source: Custom Source -> Filter -> Map
Parallelism: 12
Name Status Bytes Received Records Received Bytes Sent Records Sent Parallelism Start Time Tasks
Source: Custom Source -> Filter -> Map RUNNING 0 B 0 0 B 19,269,782 12 2021-09-28 11:01:31 12

思路2 使用底层API


Version: 1.12.4 | Commit: 5364a6b @ 2021-05-10T17:29:58+02:00 | Message:
inx实时数据分析-1、解析cookie与ua信息-process 全部业务逻辑 多并行度 解决key问题
RUNNING
2
4e09bd1fdfafd1cfbdf4a56352b2fc76
Start Time: 2021-09-26 15:19:28
Duration: 40m 22s
Cancel Job
verview
Exceptions
TimeLine
Checkpoints
Configuration
Detail
SubTasks
TaskManagers
Watermarks
Accumulators
BackPressure
Metrics
KeyedProcess
Status: RUNNING
Task: 12
Parallelism: 12
Records Sent: 315,173,100
Start Time: 2021-09-26 15:19:28
Bytes Received: 351 GB
End Time: -
Records Received: 315,173,101
Duration: 40m 22s
Bytes Sent: 0 B
HASH
KeyedProcess
Parallelism: 12
e
Status
Bytes Received
Records Received
Bytes Sent
Records Sent
Parallelism
Start Time
Tasks
ce: Custom Source -> Filter
RUNNING
0 B
0
351 GB
315,175,553
12
2021-09-26 15:19:28
12
dProcess
RUNNING
351 GB
315,173,101
0 B
315,173,100
12
2021-09-26 15:19:28
12

结论

- 使用核心API：40分钟消费了1千万9百万左右的数据，TPS约为8000。
- 使用底层API：40分钟内消费了3亿1千5百万条数据，约351GB数据，TPS约为131000。

参考资料

- svn:数据智能部/01-文档/01-公用/06-技术资料/05-flink/进阶2-性能提升16倍！使用Flink状态后端实时解析TB级别nginx日志.md
- svn:/数据智能部/01-文档/02-项目/04-系统运维部/01-日志分析/03-新日志追踪/设计/整体架构设计.md