在UUSee的应用

演讲人：王睿

EasyHadoop 第六次技术聚会分享

概要

- What
  - 用hadoop做什么
  - hadoop集群情况
  - 统计分析框架与流程
- Why
  - 为什么要用cdh版本
  - 为什么要用flume
  - 为什么要用java写MR
  - 为什么需要hive
  - 为什么准备用ooize
  - 为什么需要sqoop
- How
  - flume最佳实践
  - map-reduce最佳实践
  - sqoop最佳实践
  - 没有ooize的日子

WHAT

用hadoop做什么

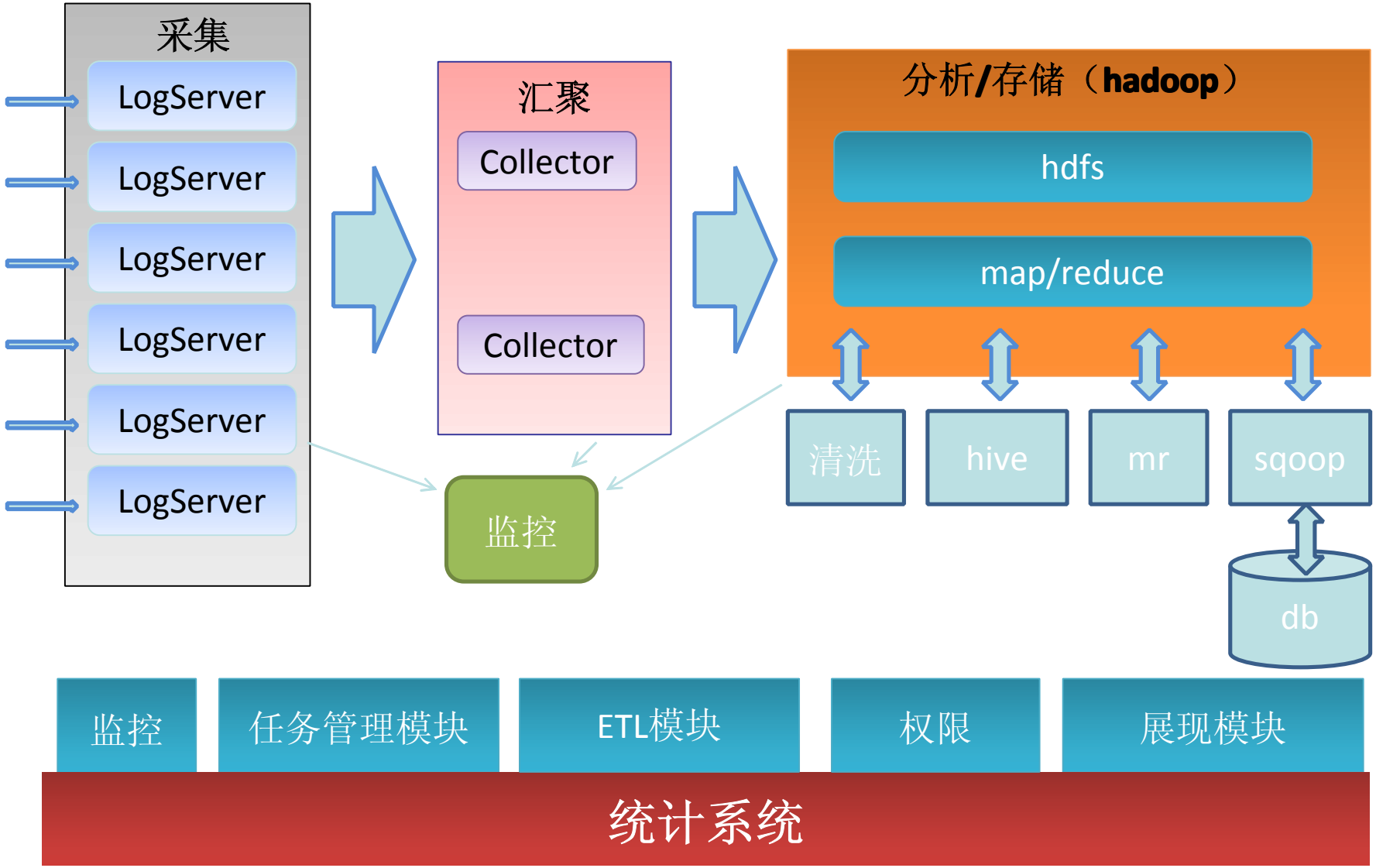
- P2P直播分发、存储
- 日志处理
- 归档存储

hadoop可以做什么？
- 实时搜索
- 日志处理
- 推荐系统
- BI/数据仓库
- 视频和图像分析
- 广告
- 归档

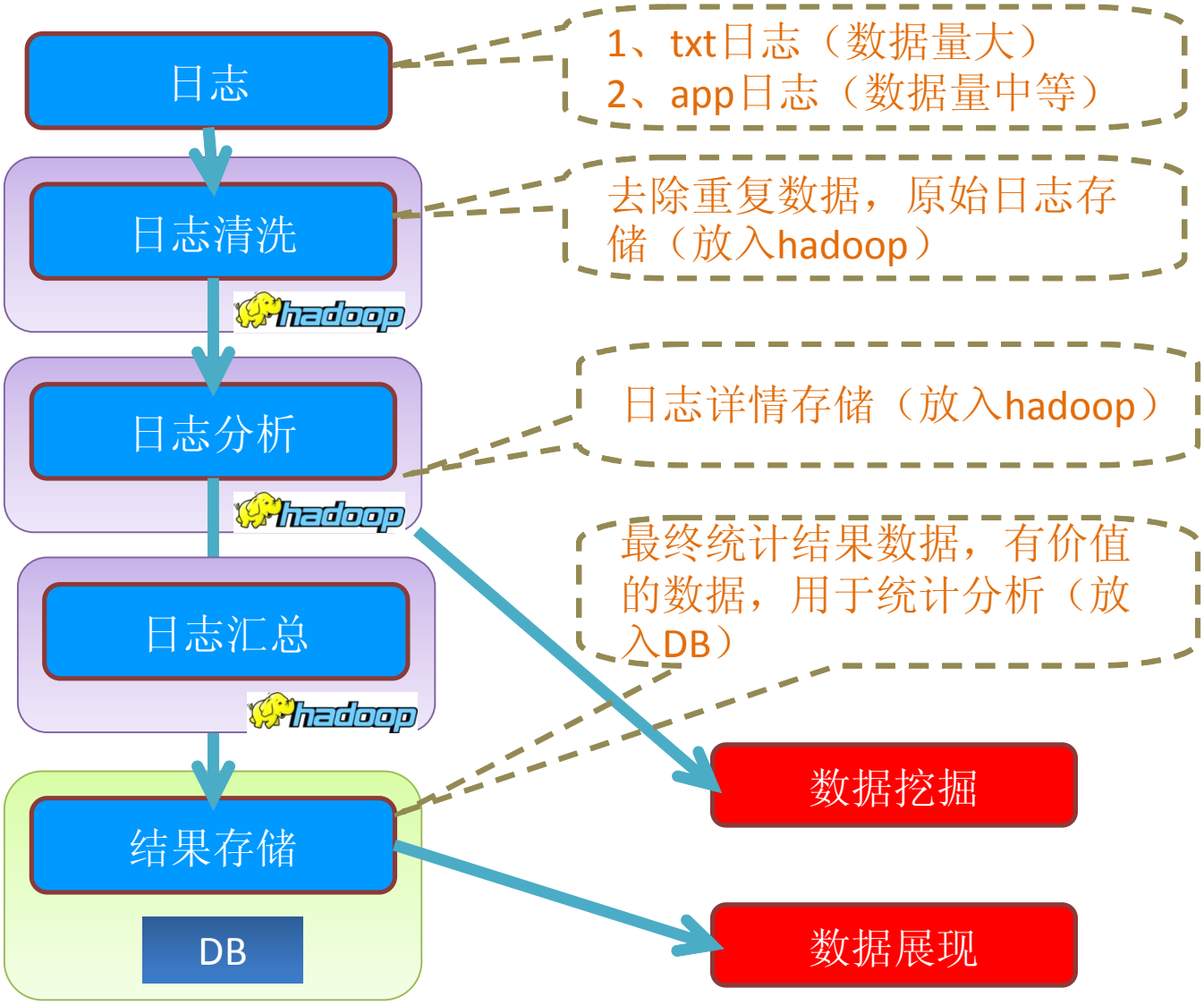
hadoop集群情况

- hadoop版本:
  - cdh3u3
- 使用的组件：
  - flume、hadoop、sqoop
  - ooize（准备用）
  - hive（肯定需要）
- 集群大小：
  - 1台namenode、6台datanode
  - 2台flume collector、n台flume agent
- 日志种类：
  - 十几种
- 数据量：
  - 每天GB级
- 时效性：
  - 有按小时统计需求

统计分析框架


采集
LogServer
LogServer
LogServer
LogServer
LogServer
LogServer
汇聚
Collector
Collector
监控
分析/存储（hadoop）
hdfs
map/reduce
清洗
hive
mr
sqoop
db
监控
任务管理模块
ETL模块
权限
展现模块
统计系统

统计分析流程


日志
1、txt日志（数据量大）
2、app日志（数据量中等）
日志清洗
去除重复数据，原始日志存储（放入hadoop）
hadoop
日志分析
日志详情存储（放入hadoop）
hadoop
日志汇总
最终统计结果数据，有价值的数据，用于统计分析（放入DB）
hadoop
结果存储
DB
数据挖掘
数据展现

WHY

为什么要用CDH版本

- 断定apache版本需要2次开发
- CDH帮我们做了该做的优化
- CDH包括一整套好用的开源组件
- CDH可以紧跟hadoop最新版
- CDH文档丰富
- 用的人多

centos vs redhat ？ vs CDH

为什么要用flume

- 需求
  - 利用现有的UDP日志服务器和HTTP日志服务器
  - LOG服务器与hadoop集群位于不同机房（分布式收集）
  - 实时收集汇聚
  - 可靠
  - 可扩展

为什么要用flume

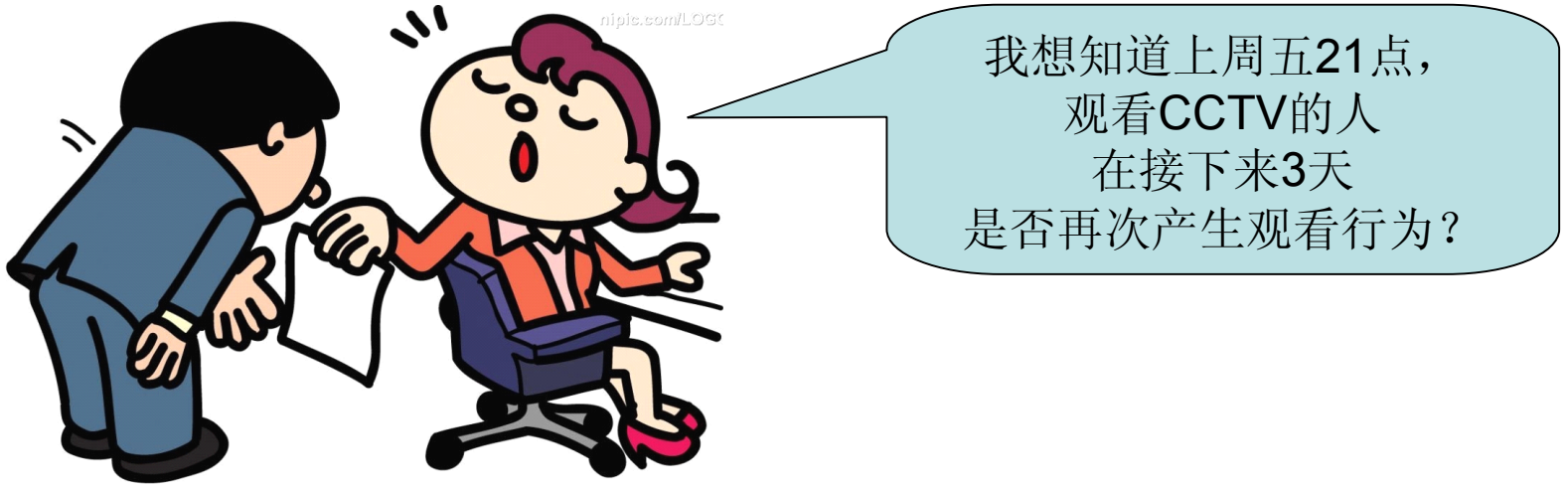
- Flume？
  - 支持多种来源的日志收集
  - 为跨广域网传输进行设计
  - 实时收集
  - 可靠（E2E、DFO、BE）
  - 可扩展
  - 代码量少可控

为什么要用java写MR

- 我们更擅长Java
- Java也高效
- 学习与业务开发并行
- 需要用Java使关键MR更高效

为什么需要HIVE

- 以前的DB解放了，详情数据在HDFS中
- 解决临时需求、一次性需求
- 需要一个HIVE查询直接导excel的工具



为什么准备用OOIZE

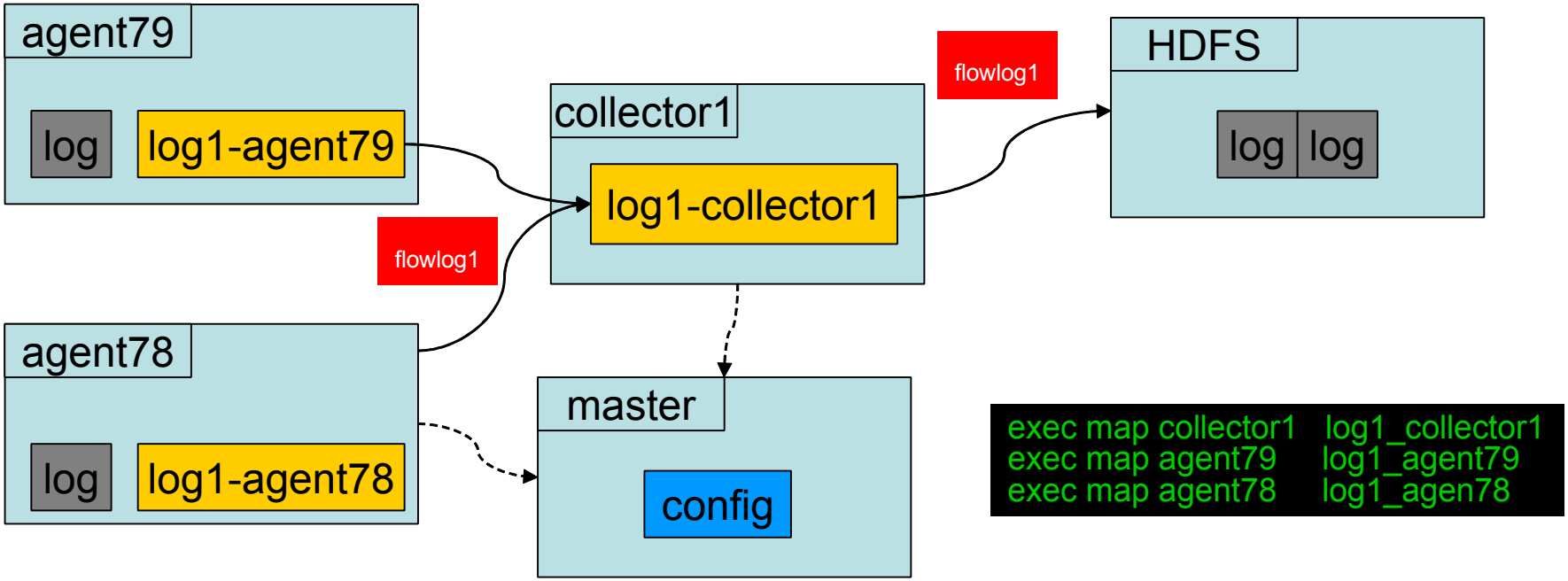
- 最后回答

为什么需要Sqoop

- 从DB到HDFS
  - 老数据
  - 业务数据
  - 自检表
- 从HDFS到DB
  - 最终展现

HOW

Flume最佳实践

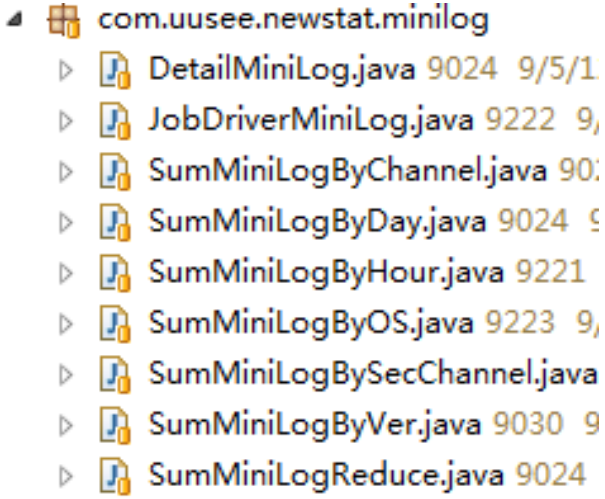



```
exec config log_agent79  flowlog1  'tailDir( "/data0/log1", "log1\\-\\d+\\-\\d+", true )'   'autoDFOChain'
exec config log_agent78  flowlog1  'tailDir( "/data0/log1", "log1\\-\\d+\\-\\d+", true ) '  'autoDFOChain'
exec config log_collector1 flowlog1 autoCollectorSource  'roll(3600000) {unbatch
        escapedCustomDfs("hdfs://xxx:8020/user/flume/original/log1/%Y/%m/%d","log1-%{rolltag}.log",raw())}'
```

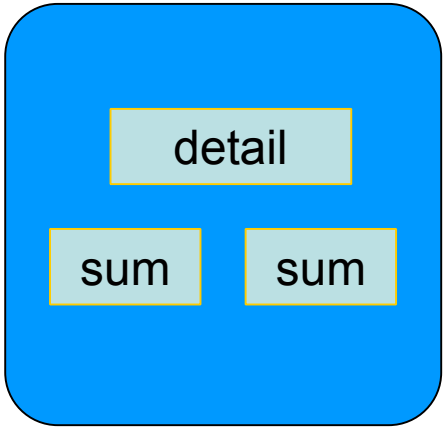
Flume最佳实践

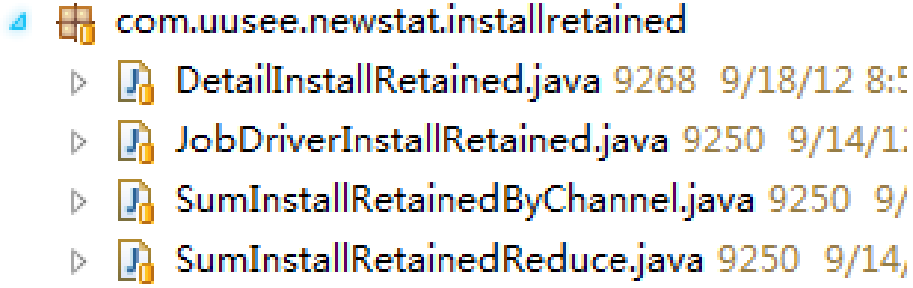
- 可靠性：DFO
- agent参数：flume.agent.logdir.maxage
- 调整linux的open files
- 2个问题：
  - HDFS中出现2个blocksize，64M和128M?
  - HDFS中日志大小为0？

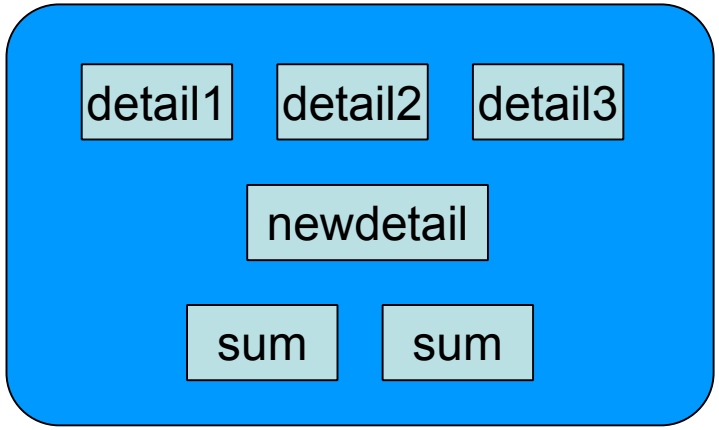
MR最佳实践


com.uusee.newstat.minilog
DetailMiniLog.java 9024 9/5/1
JobDriverMiniLog.java 9222 9/
SumMiniLogByChannel.java 90
SumMiniLogByDay.java 9024 9
SumMiniLogByHour.java 9221
SumMiniLogByOS.java 9223 9/
SumMiniLogBySecChannel.java
SumMiniLogByVer.java 9030 9
SumMiniLogReduce.java 9024


客户端MINI页
按小时
按天
按渠道
按OS版本
按版本


detail
sum
sum


com.uusee.newstat.installretained
DetailInstallRetained.java 9268 9/18/12 8:
JobDriverInstallRetained.java 9250 9/14/1
SumInstallRetainedByChannel.java 9250 9/
SumInstallRetainedReduce.java 9250 9/14/


detail1
detail2
detail3
newdetail
sum
sum

Sqoop最佳实践

- HDFS到DB

```
sqoop export
    --connect jdbc:oracle:thin:@XXX:1521:XXX
    --username XXXX
    --password XXXXX
    --export-dir /user/flume/part-r-00000
    -m 1
    --table NS_SLLOG_CLIENT_SUM_OS
    --update-key ACTIVITYTIME,CLIENT_VER,OS
    --update-mode  allowinsert
    --input-fields-terminated-by '\t'
    --outdir jobs
    --package-name com.uusee.sqoop
```

没有OOIZE的日子

```
#!/bin/bash

#变量定义
INPUT_DATE=$1
LOG_NAME="oldupdatelog"
ARCHIVES_DIR="/user/hadoop/archives/${LOG_PATH}/${INPUT_DATE}"
INSTALL_DETAIL_DIR="/user/hadoop/detail/${LOG_PATH}/install/${INPUT_DATE}"
SUM_DIR="/user/hadoop/sum/${LOG_PATH}/${INPUT_DATE}"
HADOOP_CMD_JAR="/opt/hadoop/bin/hadoop jar"
JOB_JAR="${LOG_NAME}-newstat-uusee.jar"
HADOOP_CMD_RMR="/opt/hadoop/bin/hadoop fs -rmr"
LZO_INDEXER_JAR_MAINCLASS="/opt/hadoop/lib/hadoop-lzo-0.4.15.jar com.hadoop.compression.lzo.DistributedLzoIndexer"
JOBS_DIR="/opt/hadoop/jobs"
JOBSHELL_DIR="/opt/hadoop/jobshell"

#清洗
echo "washing....................."
/bin/sh ${JOBSHELL_DIR}/wash.sh ${LOG_PATH} ${INPUT_DATE}
echo ".....................wash end!"

#i详情
echo "install detail running....................."
/bin/sh ${JOBSHELL_DIR}/detail.sh ${LOG_NAME}  ${ARCHIVES_DIR} ${INSTALL_DETAIL_DIR} ${INPUT_DATE} || exit -1
${HADOOP_CMD_JAR} ${LZO_INDEXER_JAR_MAINCLASS} ${INSTALL_DETAIL_DIR}
echo ".................install detail end!"

#i汇总并导入DB
echo "install sum and export running....................."
${HADOOP_CMD_RMR} ${SUM_DIR}
${HADOOP_CMD_JAR} ${JOBS_DIR}/${JOB_JAR} sumupdateloginstallbyday ${INSTALL_DETAIL_DIR} ${SUM_DIR}
/bin/sh ${JOBSHELL_DIR}/sqoopexport.sh ${SUM_DIR}/part-r-00000 NS_OLDINSTALL_CLIENT_SUM_DAY ACTIVITYTIME
echo ".................install sum and export end!"
```

为什么需要OOIZE

- 大多数统计分析任务不是由一个map-reduce完成的，需要多个map-reduce组合完成，或者还需要其他程序配合。
- 各map-reduce或其他任务间有依赖关系
- 需要组合不同任务：MR、HADOOP命令、SSH、JAVA、Sqoop、MAIL

OOIZE是hadoop的工作流软件，就是为了这个需求产生的

Q&A

感谢大家！

感谢**easyhadoop**！