日志流收集组件迁移方案说明书

1. 组件描述：

日志流收集组件定期将现网业务日志数据转换为parquet格式，并存储到HDFS中。

* 1. 组件逻辑结构：



* 1. 组件清单

|  |  |  |
| --- | --- | --- |
| 类别 | 文件清单 | 描述 |
| nginx | openrestry程序包 | 统一的openrestry安装包，可由运维统一提供 |
|  | 程序包子目录：nginx/conf/online | 业务相关的配置文件统一放置在该子目录中 |
|  | 程序包子文件：nginx/conf/nginx.conf | nginx配置主文件，nginx启动时加载该配置文件 |
| logCenter | logcenter程序包 | 由开发提供的logcenter程序包 |
|  | 日志输出目录：/var/log/logcenter/current | 日志写入目录，可配置 |
| 日志备份 | WeixinLogUpload.sh  MedusaLogUpload.sh  MedusaPostLogUpload.sh  MoreTVLogUpload.sh | crontab调度的日志备份脚本 |
| 转parquet程序 | 程序包：  MoreTVBISpark.jar  WhaleyBISpark.jar  … | 由开发提供的转parquet程序包 |
| HDFS文件目录 | /… | 日志备份目录 |
|  | /mbi/parquet | parquet文件存储目录【电视猫】 |
|  | /log/medusa/parquet | parquet文件存储目录【电视猫】 |
|  | /log/whaley/parquet | parquet文件存储目录【微鲸】 |

1. 迁移步骤：
   1. 前置条件：
2. 流量复制方案：在整个迁移早期实施，保留迁移期原始日志数据，以便业务组件北京机房部署完毕后做迁移期数据补全，在并行期将日志数据流量同时接入两地机房，为业务组件并行运行提供支持
3. HDFS集群迁移：只需要提供HDFS服务即可，不依赖于HDFS的历史数据
4. Spark集群迁移：依赖于Spark 1.6的standalone集群
   1. 迁移步骤：
5. 北京机房服务部署：

* HDFS文件目录准备：确保依赖的HDFS目录存在，且日志备份脚本和转parquet程序运行用户具有相应读写权限
* 部署logcenter服务：运行用户、程序包安装位置、日志文件位置、监听端口遵循运维规划，现网端口从8080开始往后配置
* 部署nginx服务: 运行用户、程序包安装位置、日志文件位置遵循运维规划；对外服务端口为80（保持和现网一致，避免使用方需要做配置修改）；logcenter地址根据实际情况配置
* 部署日志备份脚本：运行用户、脚本位置遵循运维规范；将备份脚本提交到crontab中，每天凌晨30分调度一次，脚本位置遵循运维规划；现网部署位置/script/bi/weixin/shell
* 部署转parquet程序：运行用户、程序包安装位置遵循运维规划，每天凌晨1点调度一次
* 启动流量复制：将日志流量同时引入到北京和杭州机房，进入并行期，流量复制具体操作参见流量复制方案
* 服务正式切换：在服务验证运行无误，机房服务确认整体切换后，将到杭州机房的流量切断，北京机房服务进入正式服务运行期

1. 检查清单：

|  |  |  |
| --- | --- | --- |
| 阶段 | 检测点 | 检测方法 |
| HDFS目录 | HDFS文件目录存在性检查 | hadoop文件shell命令检查 |
| 用户权限检查 | hadoop文件shell命令检查 |
| logcenter服务 | 进程检查+端口检测 | 遵循运维标准 |
| nginx服务 | 进程检查+端口检测 | 遵循运维标准 |
| 日志备份脚本 | crontab列表中脚本确认 | crontab -l命令检查 |
| 转parquet程序 | crontab列表中脚本确认 | crontab -l命令检查 |
| 流量复制 | 杭州北京两地机房日志文件 | 定时检测两地nginx日志文件，要求大小一致 |