|  |  |  |
| --- | --- | --- |
| 项目名称 Project Name | | 密级  Confidentiallity Level |
| Spark数仓（Elasticsearch接入方案） | | 仅供收件查询 |
| 项目编号 Project ID | 版本 Version | 文档编号 Document Code |
| data-etl-bridge | V1.1 | data-etl-bridge-doc-v1.1 |

Spark data warehouse（Elasticsearch）Software Requirements specification

Spark数仓（Elasticsearch接入）系统

需求规格说明书

|  |  |  |  |
| --- | --- | --- | --- |
| Prepared by  拟制 | 王小雷 | Date  日期 | 2019-5-20 |
| Reviewed 评审人 | 李龙 | Date  日期 |  |
| Approved by  批准 |  | Date  日期 |  |

Revision Record

修订记录

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date  日期 | Revision  Version  版本修订 | CR ID/Defect ID  CR/Defect号 | Sec No.  修改章节 | Change description  修改描述 | Author 作者 |
| 2019-5-29 | V0.0.2 |  | 3.4.4 | 网络传输不对称解决方案及调整 | 王小雷 |
| 2019-5-30 | V0.0.3 |  | 3.4.4 | 网络传输不对称解决方案及调整 | 王小雷 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

目录

[1 Introduction 简介 2](#_Toc10143606)

[1.1 Purpose 目的 2](#_Toc10143607)

[2 General description 总体描述 3](#_Toc10143608)

[2.1 Software perspective 系统功能描述 3](#_Toc10143609)

[3 Specific Requirement 具体需求 3](#_Toc10143610)

[3.1 系统用例 3](#_Toc10143611)

[3.2 顶层设计 3](#_Toc10143612)

[3.3 数据流 4](#_Toc10143613)

[3.4 处理模块 4](#_Toc10143614)

[3.4.1 新接入Elasticsearch库 5](#_Toc10143615)

[3.4.2 同步（增量/全量）Elasticsearch库 5](#_Toc10143616)

[3.4.3 失败重试模块 7](#_Toc10143617)

[3.4.4 网络传输不对称解决方案及调整 8](#_Toc10143618)

[3.5 数据字典 9](#_Toc10143619)

# 1 Introduction 简介

## Purpose 目的

该需求规格说明书是关于Spark数仓（Elasticsearch接入）系统提供的功能和性能的描述，该说明书的预期读者为：

项目管理人员；

测试人员；

设计人员；

开发人员。

这份软件需求文档说明书重点描述顶层设计、数据流图、元字典设计，明确系统具体功能，是系统分析人员与开发人员清楚地了解用户需求、实现原理。

# 2 General description 总体描述

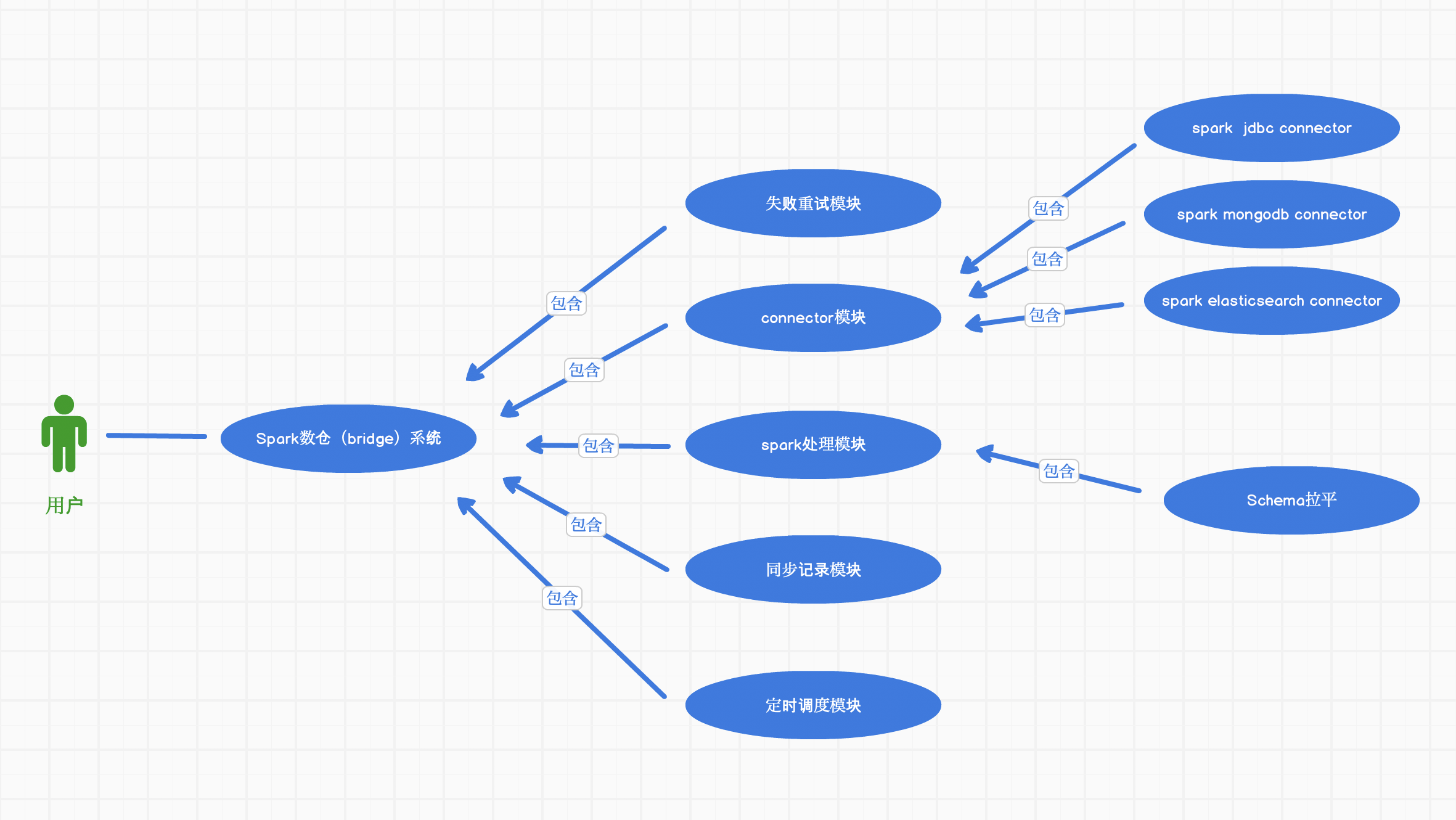
## 2.1 Software perspective 系统功能描述

Spark数仓（Elasticsearch接入）系统，通过spark对Elasticsearch数据进行接入。整个过程通过元字典、调动做到自动化。

# 3 Specific Requirement 具体需求

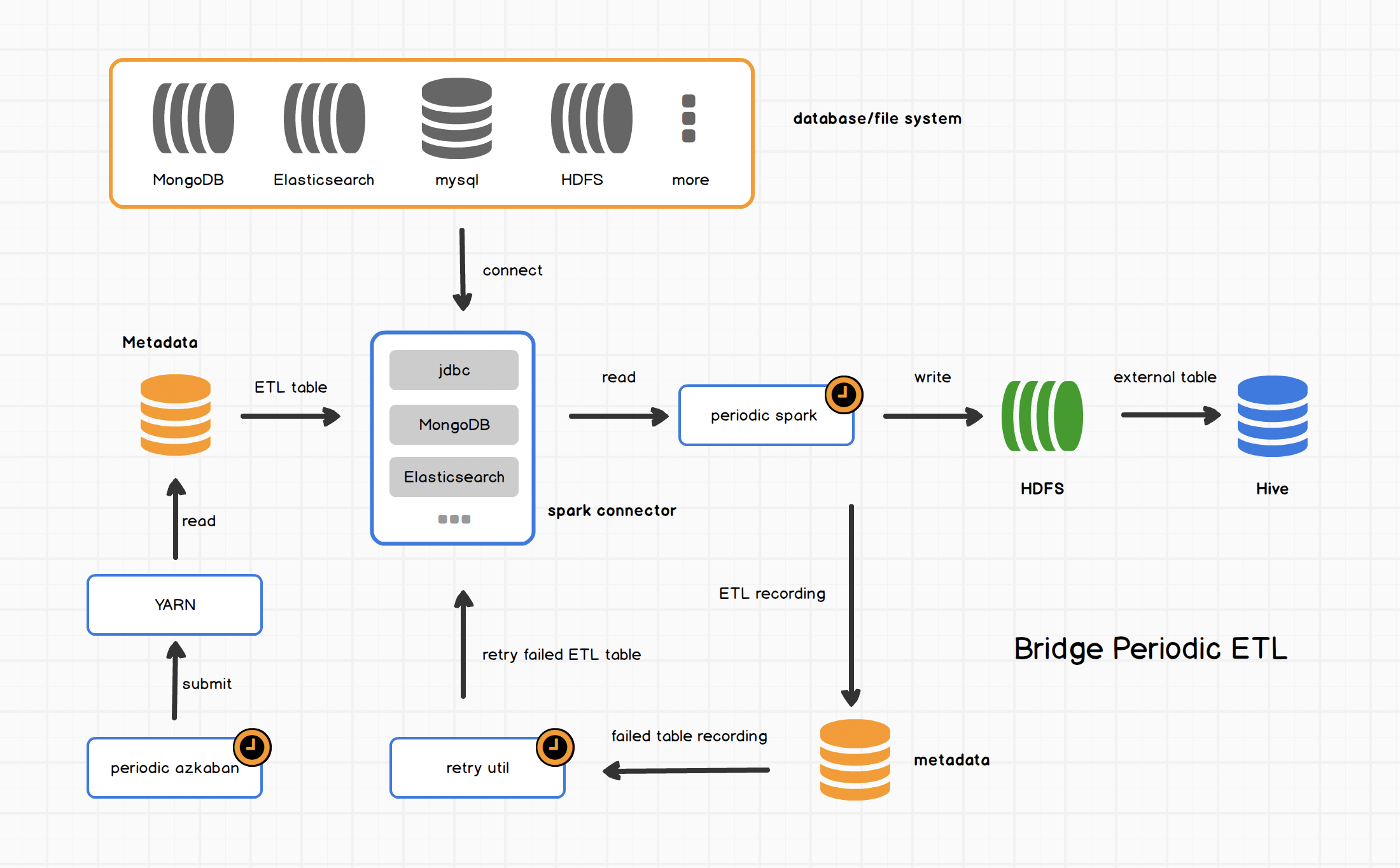
## 3.1 系统用例

系统用例图如下：



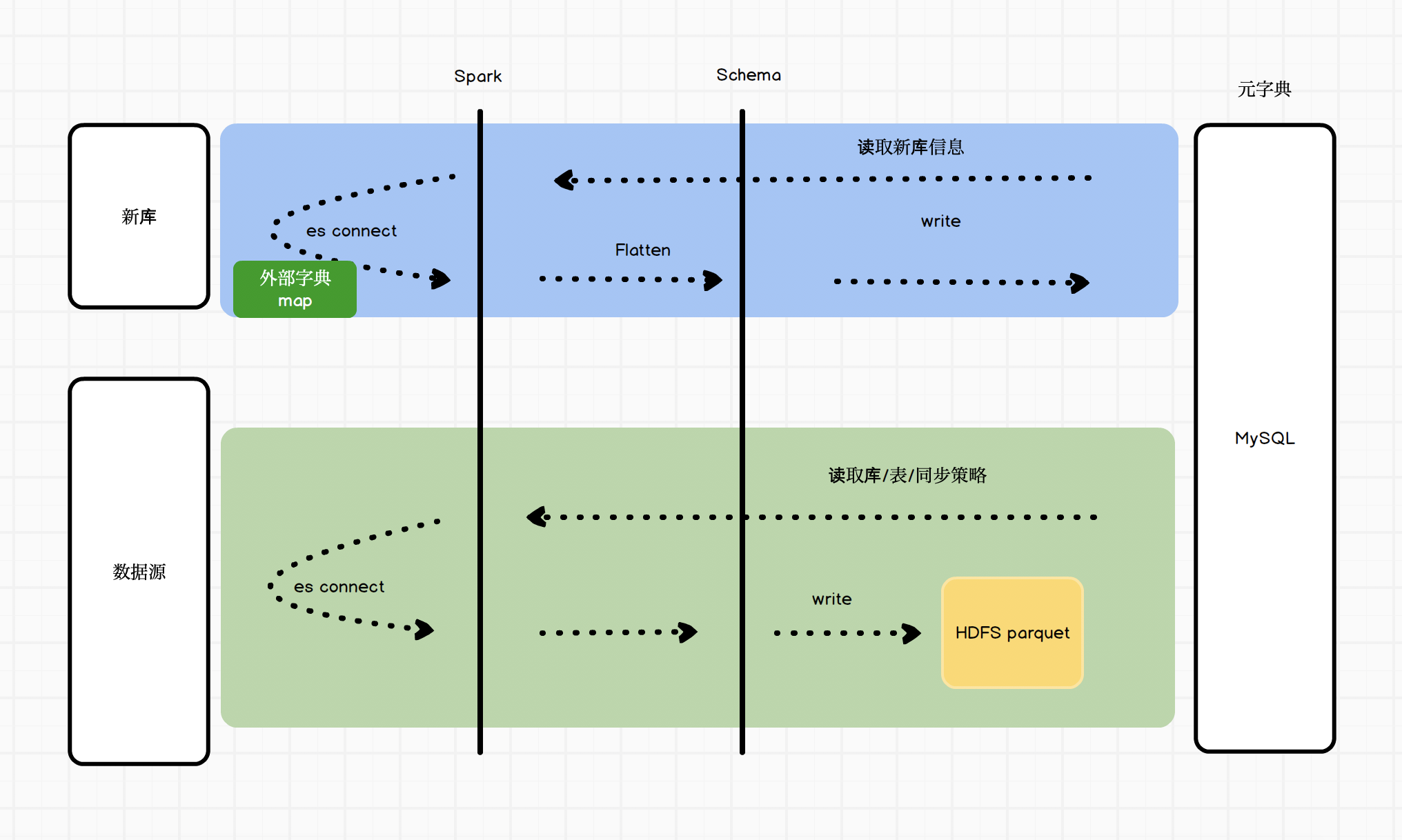
## 3.2 顶层设计

顶层设计图如下：



## 3.3 数据流

Elasticsearch数据流图如下：



## 3.4 处理模块

### 3.4.1 新接入Elasticsearch库

1 介绍

接入新Elasticsearched数据，需要获取数据库的连接信息，之后通过spark对数据库表拉平入元字典。

2 输入

数据库用户、密码、连接地址

3 处理

Elasticsearch连接信息 🡪 元字典 库级别  
 Elasticsearch index 🡪 元字典 表

Elasticsearch Field 🡪 元字典 field

4 输出

Elasticsearch field信息入元字典。

## 3.4.2 同步（增量/全量）Elasticsearch库

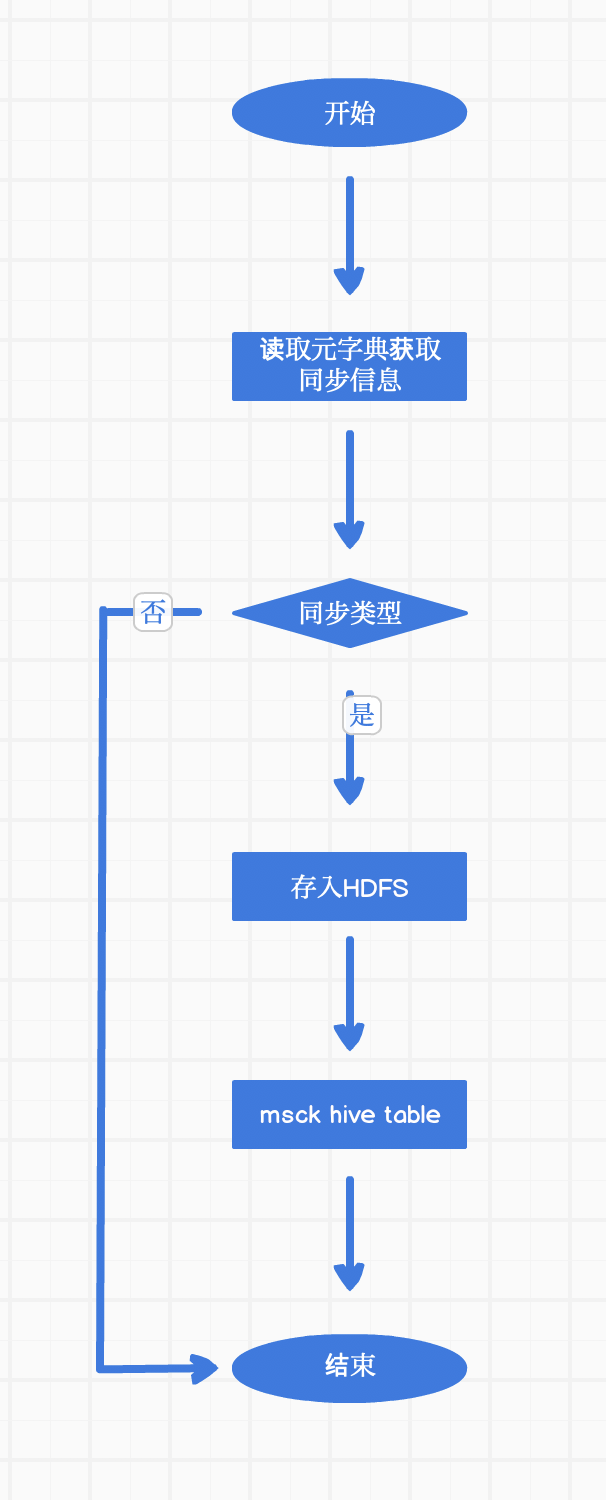
1 介绍

需要同步的Elasticsearch数据库，每日增量同步或全量同步，流水表或者拉链表。

2 输入

元字典同步信息，包含数据库信息、表信息、同步类型

3 处理



4 输出

每日增量/全量存入以parquet形式存入HDFS，并成功刷新Hive表。

## 3.4.3 失败重试模块

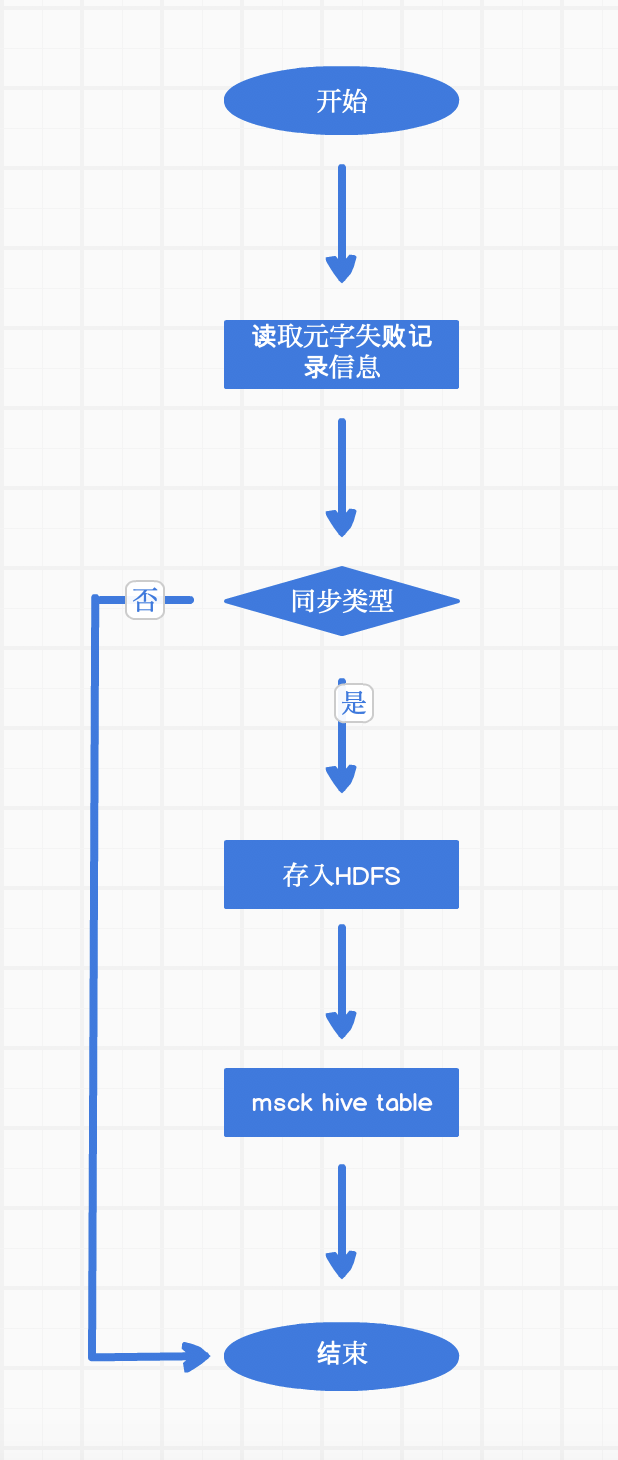
1 介绍

同步记录粒度到表级别，如果表同步失败，需要重新同步该失败的表。

2 输入

同步失败表

3 处理

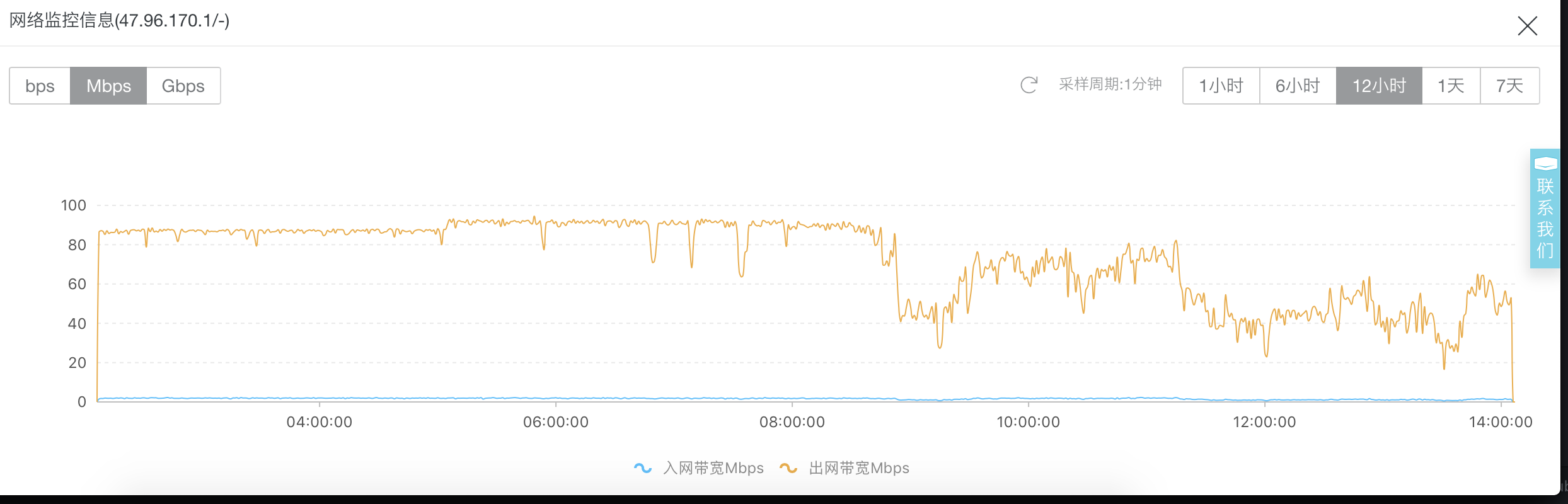


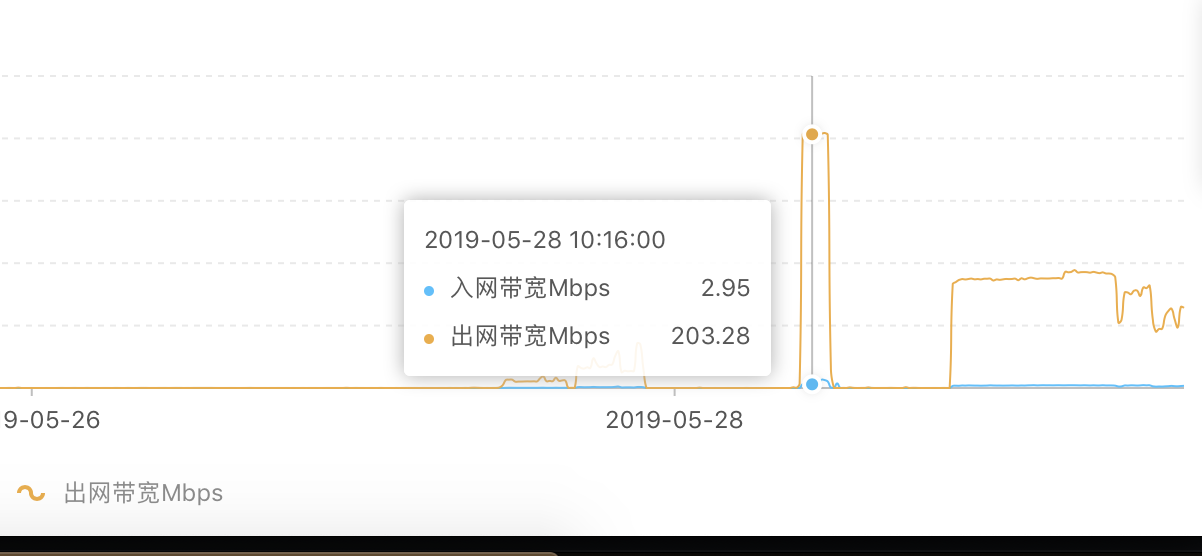
4 输出

失败的数据表重新同步，并记录同步状态。

## 3.4.4 网络传输不对称解决方案及调整

Elasticsearch数据导入时内网传输，数据读取时是外网传输。





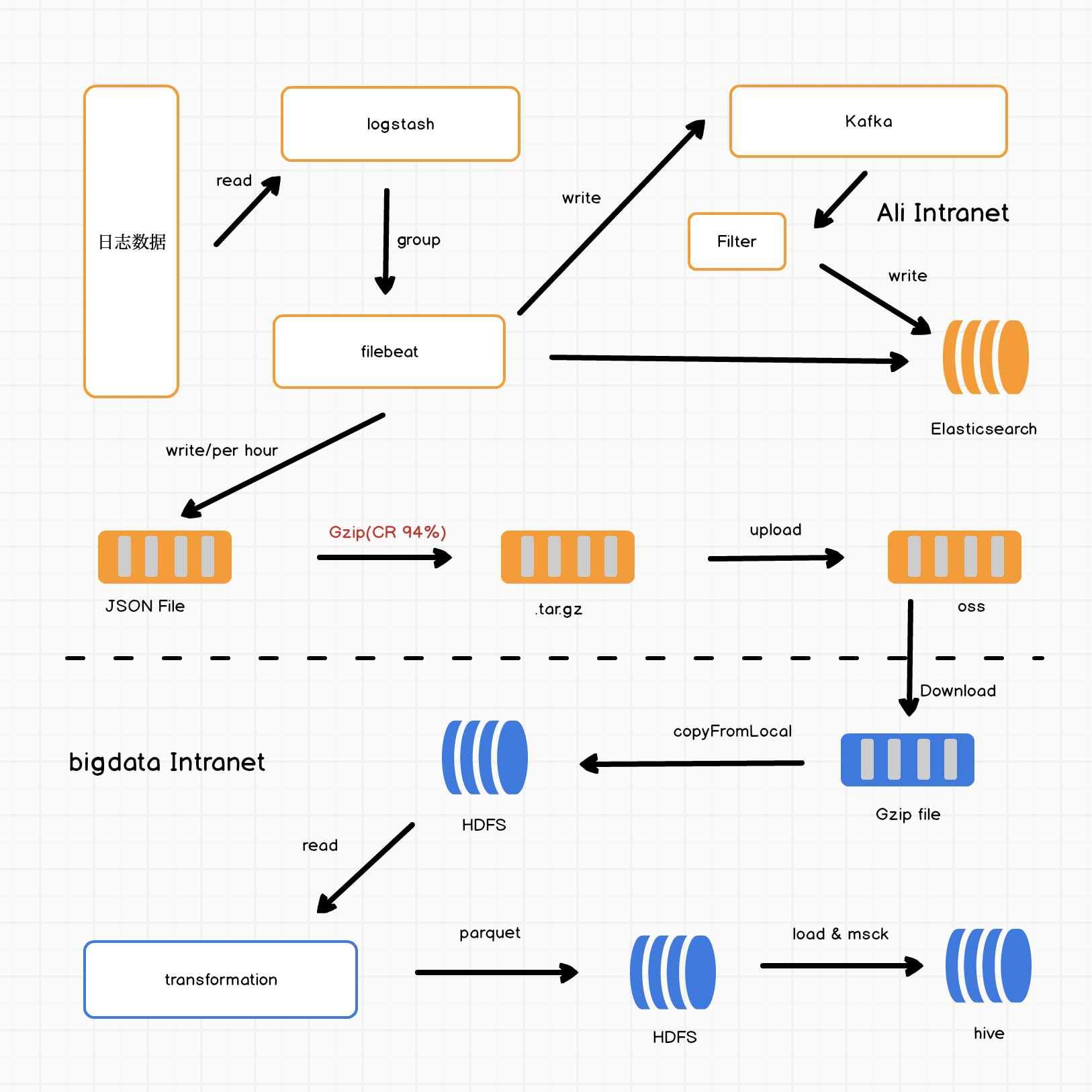
历史数据仍然根据spark-es导入。今日数据调整处理方式。

全链路会受网络抖动影响，不能保证最大带宽。数据批量传输成为瓶颈。

1T数据200Mbps 约25M/s需要11.65小时。

750G，平均 60Mbps即7.5M/s需 30小时。

方案调整如下：



关键操作：  
1、在阿里内网多sink一份数据到本地文件，通过GZIP压缩，原约1350M数据压缩后为约80M，压缩率高达94%；

2、已压缩的数据按小时分批上传至阿里OSS；

3、凌晨开始下载OSS数据并上传至本地大数据集群。

## 3.5 数据字典

元字典图如下：

