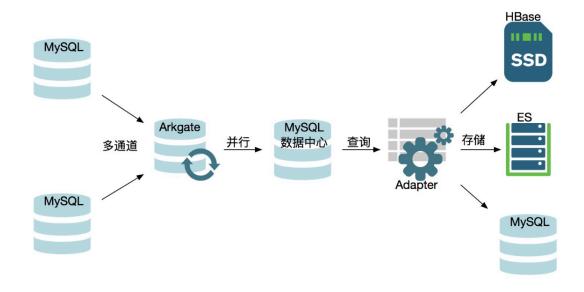
1. Arkcontrol 数据同步功能简介

Arkcontrol 数据同步功能是基于极数云舟自主研发的 Arkgate 的数据异构同步工具。Arkgate 是一款 MySQL 的插件,模拟了 MySQL 从库的 IO Thread,不断地 Dump MySQL 的 Binlog Events,实时地将 MySQL 数据库的增量更新同步到目标数据库中。目标数据库支持 HBase、Elasticsearch 及 MySQL 数据库,以解决不同人群及业务的需求。架构如下图所示:



图中各个节点说明如下:

在线 MySQL

最左边的 MySQL,可以是一个集群,也可以是单节点,都可以通过同一个Arkgate 来做数据传输,一个传输任务称为一个通道,也就是可以有多个通道。要求线上 MySQL 必须记录 Binlog 并且设置 binlog format 为 ROW 模式。

Arkgate

Arkgate 是一个 MySQL 插件,在 MySQL 内部不断运行着传输任务。将 Dump 过来的数据 Binlog Events 解析出来,转换为 JSON,并行地存储到 MySQL 中,这是一个数据的中间状态。这个数据库被称为数据中心。

数据中心

在这个数据库中,一个数据库对应一个通道,一个通道具有唯一的名字,即 Tunnel Name。在这个数据库中,存储了这个通道相关的所有信息,包括复制主从节点、独立配置参数、从节点对应的位置信息、主节点复制的最新位置、白

/黑名单及 JSON 数据。本系统到此为止也是可以使用的,业务可以自己来定制化地查这里面的数据,而如果为了简单方便,可以部署增强系统,即 Adapter 程序。

Adapter 程序

已经做了几种目标数据库的 Adapter 程序,包括 Kafka, HBase、ES 及 MySQL (目前页面支持自动部署的为到 MySQL),这里可以通过配置文件实现表列的对应关系,最终实现实时异构的数据转换。

目前 Arkcontrol 中可以在已经搭建好的 gate 集群中添加创建同步任务,监测任务状态,完成同步任务启停,参数修改、配置等操作。订阅版包含分布式哨兵集群,可以监控同步链路状态,确保高可用。

2. Arkcontrol 同步任务创建

2.1 准备工作

需要手动搭建 Arkgate 集群,用于同步任务

2.1.1 搭建数据库实例

使用\${installDir}/arkcontrol/software/arkgate 下的 ark-mysql57-5.7.18-1.el7.centos.x86_64.rpm,搭建单节点实例(127.0.0.1_3321,用户根据情况调整,这里仅举例说明)

rpm –ivh ark-mysql57-5.7.18-1.el7.centos.x86_64.rpm
/opt/ark/arkdb/scripts/mysql_install_file --port=3321 --target=mysql57 --installdir=/data

2.1.2 拷贝 Arkgate 插件

拷贝 \${installDir}/ArkControl/software/arkgate 下的 arkgate.so 到部署实例的 plugin dir(登陆实例\${installDir}/ArkControl/software/arkgate 查看)下

拷贝 \${installDir}/ArkControl/software/arkgate 下的 arkgate.key 文件到部署 实例的 arkdb/3321/etc 下

2.1.3 安装 Arkgate 插件

mysql -uroot -P3321 -h127.0.0.1 -A

执行如下语句:

install plugin arkgate soname 'arkgate.so';

install plugin arkgate_datacenter_threads soname 'arkgate.so';

```
install plugin arkgate_datacenter_full_list soname 'arkgate.so';
install plugin arkgate_fliter_list soname 'arkgate.so';
install plugin arkgate_slave_list soname 'arkgate.so';
install plugin arkgate_datacenter_list soname 'arkgate.so';
install plugin arkgate_datacenter_options soname 'arkgate.so';
install plugin arkgate_datacenter_tables soname 'arkgate.so';
#查看 arkgate 相关参数确认安装成功
mysql> show global variables like '%arkgate%';
```

2.1.4 修改配置文件

修改 arkdb/3321/etc/my.cnf 添加

```
[mysqld]

arkgate_datacenter_host = 127.0.0.1 ##dc 地址

arkgate_datacenter_port = 3321 ##dc 端, dc 与 gate 公用一个 MySQL 实例

arkgate_datacenter_user = arkgate ##访问 dc 用户

arkgate_datacenter_password = arkgate_test ##访问 dc 密码

arkgate_license_file = /data/ark/arkdb/multi/3321/etc/arkgate.key ##License 验证
```

2.1.5 重启 MySQL 实例

/etc/init.d/mysql.server -P 3321 restart

2.1.6 账号授权

```
mysql -uroot -P3321 -h127.0.0.1 -A
grant all privileges on *.* to 'arkgate'@'%' identified by 'arkgate_test';
```

2.1.7 初始化 ETL 配置库表

```
CREATE DATABASE IF NOT EXISTS 'etl_config';

CREATE TABLE IF NOT EXISTS 'etl_config'.'task_info' (
    'id' int(11) NOT NULL AUTO_INCREMENT,
    'task_name' varchar(128) NOT NULL DEFAULT ",
    'dc_ip' varchar(16) NOT NULL DEFAULT '127.0.0.1',
    'dc_port' int(11) NOT NULL DEFAULT '9991',
    'dc_dbname' varchar(32) NOT NULL DEFAULT ",
```

```
'dc user' varchar(32) NOT NULL DEFAULT ",
    'dc pass' varchar(32) DEFAULT ",
    'dst ip' varchar(16) NOT NULL DEFAULT '127.0.0.1',
    'dst_port' int(11) NOT NULL DEFAULT '3310',
    'dst user' varchar(32) NOT NULL DEFAULT ",
    'dst pass' varchar(32) NOT NULL DEFAULT ",
    'start id' int(11) NOT NULL DEFAULT '1',
    'thread num' int(11) DEFAULT '8',
    'limit_query_rows' int(11) NOT NULL DEFAULT '4000',
    'buffer_size' int(11) NOT NULL DEFAULT '8000',
    'exec retry num' int(11) NOT NULL DEFAULT '3',
   'exec retry interval' int(11) NOT NULL DEFAULT '3',
   'logger file' varchar(128) NOT NULL DEFAULT 'mysql etl.log',
    `double_active` tinyint(1) NOT NULL DEFAULT '0' COMMENT '双活开关, false 为关闭',
    `exec mode` int(11) NOT NULL DEFAULT '0' COMMENT '执行 SQL 报错时的控制参数,默
认值是 0.(0 跳过,1 覆盖执行,2 业务方提供异常处理接口,3 程序退出);',
   'biz conflict api' varchar(256) NOT NULL DEFAULT ",
   'rename flag' tinyint(1) NOT NULL DEFAULT '0' COMMENT '修改表名开关',
   `filter_opt_flag` tinyint(1) NOT NULL DEFAULT '0' COMMENT '操作过滤开关',
   `filter_col_flag` tinyint(1) NOT NULL DEFAULT '0' COMMENT '列过滤开关',
   'create time' timestamp NOT NULL DEFAULT '1970-01-01 09:00:00',
    'update_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT_TIMESTAMP,
   PRIMARY KEY ('id'),
   UNIQUE KEY 'uniq taskname' ('task name')
                          AUTO INCREMENT=1
                                                   DEFAULT
                                                                CHARSET=utf8mb4
      ENGINE=InnoDB
COMMENT='ETL 任务基准表':
 CREATE TABLE 'etl config'.'etl state' (
   'id' int(11) NOT NULL AUTO INCREMENT,
    `task_name` varchar(512) NOT NULL DEFAULT "COMMENT '任务名称,不能重复',
    'is consistent' tinyint(1) NOT NULL DEFAULT '0',
```

```
`delay time` int(11) NOT NULL DEFAULT -1 COMMENT 'etl 相对于 DC 的延迟时间,单
位是秒',
   `qps` int(11) NOT NULL DEFAULT '0' COMMENT '不够精确的 qps',
   'checkpoint' bigint(20) NOT NULL DEFAULT '0',
   'binlog ctime' timestamp NOT NULL DEFAULT '1973-01-01 08:00:00',
   `heart beat` int(11) NOT NULL DEFAULT '0' COMMENT '心跳线程活着时, 每隔一秒种执
行 heart beat++',
   'create_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
   'update_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT_TIMESTAMP,
   PRIMARY KEY ('id'),
   UNIQUE KEY 'uniq taskname' ('task name')
                        AUTO_INCREMENT=1 DEFAULT
                                                             CHARSET=utf8mb4
      ENGINE=InnoDB
 )
COMMENT='ETL 状态表':
 CREATE TABLE IF NOT EXISTS 'etl config'.'rename rule' (
   'id' int(11) NOT NULL AUTO INCREMENT,
   'task name' varchar(128) NOT NULL DEFAULT ",
   'original_db' varchar(64) NOT NULL DEFAULT ",
   'original_table' varchar(64) NOT NULL DEFAULT ",
   'new db' varchar(64) NOT NULL DEFAULT ",
   'new_table' varchar(64) NOT NULL DEFAULT ",
   'create_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
   'update_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT_TIMESTAMP,
   PRIMARY KEY ('id')
 ) ENGINE=InnoDB AUTO_INCREMENT=1 DEFAULT CHARSET=utf8mb4 COMMENT='修
改表名';
 CREATE TABLE IF NOT EXISTS 'etl_config'.'rename_regexp' (
   'id' int(11) NOT NULL AUTO INCREMENT,
```

```
'task name' varchar(128) NOT NULL DEFAULT ",
   'db pattern' varchar(64) NOT NULL DEFAULT ",
   'table pattern' varchar(64) NOT NULL DEFAULT ",
   'new_db' varchar(64) NOT NULL DEFAULT ",
   'new_table' varchar(64) NOT NULL DEFAULT ",
   'create_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
   'update_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT_TIMESTAMP,
   PRIMARY KEY ('id')
 ) ENGINE=InnoDB AUTO_INCREMENT=1 DEFAULT CHARSET=utf8mb4 COMMENT='修
改表名,正则表达式来匹配原表名';
 CREATE TABLE IF NOT EXISTS 'etl_config'.'filter_opt' (
   'id' int(11) NOT NULL AUTO INCREMENT,
   'task name' varchar(128) NOT NULL DEFAULT ",
   'dbname' varchar(64) NOT NULL DEFAULT ",
   'tablename' varchar(64) NOT NULL DEFAULT ",
   'optype' varchar(16) NOT NULL DEFAULT ",
   'action' varchar(128) NOT NULL DEFAULT "COMMENT '多个动作用英文逗号分割,如:
ADD INDEX, DROP INDEX',
   'create_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
   'update_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT_TIMESTAMP,
 PRIMARY KEY ('id')
 ) ENGINE=InnoDB AUTO INCREMENT=7 DEFAULT CHARSET=utf8mb4 COMMENT='操
作过滤表':
 CREATE TABLE IF NOT EXISTS 'etl config'. 'filter col' (
   'id' int(11) NOT NULL AUTO INCREMENT,
   'task name' varchar(128) NOT NULL DEFAULT ",
   'dbname' varchar(64) NOT NULL DEFAULT ",
```

```
'tablename' varchar(64) NOT NULL DEFAULT ",
    `column names` varchar(64) NOT NULL DEFAULT " COMMENT '要过滤多个字段,用英文
逗号分割',
    'create_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
    'update_time' timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT TIMESTAMP,
    PRIMARY KEY ('id')
  ) ENGINE=InnoDB AUTO_INCREMENT=1 DEFAULT CHARSET=utf8mb4 COMMENT='字
段过滤!;
  INSERT INTO 'etl config'. 'task info'(
    id, task name, de ip, de port, de dbname, de user, de pass, dst ip, dst port, dst user, dst pass,
start_id,
    thread num, limit query rows, buffer size, exec retry num, exec retry interval, logger file,
double active, exec mode,
    biz conflict api, rename flag, filter opt flag, filter col flag) VALUES (
    1, 'etl beijing', '118.190.89.67', 9991, 'dc beijing', 'test etl', 'test etl', '118.190.89.67', 3311,
'test etl',
    'test_etl', 1, 8, 2000, 2000, 3, 3, 'mysql_etl.log', 0, 0, ", 1, 1, 1
  );
  #rename_rule 测试数据
  #只改表名
  INSERT INTO 'etl config'. rename rule' (id, task name, original db, original table, new table)
VALUES (1, 'etl beijing', 'da1', 'ta1', 'ta');
  INSERT INTO 'etl config'. rename rule' (id, task name, original db, new table) VALUES (2,
'etl_beijing', 'db', 'tb');
  INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_table, new_table) VALUES (3,
'etl beijing', 'tc1', 'tc');
  #只改库名
  INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_db, original_table, new_db)
VALUES (4, 'etl beijing', 'dd1', 'td1', 'dd');
```

INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_db, new_db) VALUES (5, 'etl beijing', 'de1', 'de1');

INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_table, new_db) VALUES (6, 'etl_beijing', 'tf1', 'df');

#库名表名都修改为指定名称

INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_db, new_db, new_table)
VALUES (7, 'etl beijing', 'dg1', 'dg', 'tg');

INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_table, new_db, new_table)
VALUES (8, 'etl_beijing', 'th1', 'dh', 'th');

INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_db, original_table, new_db, new table) VALUES (9, 'etl_beijing', 'di1', 'ti1', 'di', 'ti');

INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_db, original_table, new_db, new_table) VALUES (10, 'etl_beijing', 'di1', 'ti2', 'di', 'ti');

INSERT INTO 'etl_config'.'rename_rule'(id, task_name, original_db, original_table, new_db, new table) VALUES (11, 'etl beijing', 'di2', 'ti1', 'di', 'ti');

#rename regexp

INSERT INTO 'etl_config'.'rename_regexp'(id, task_name, table_pattern, new_table) VALUES (1, 'etl_beijing', '^rta', 'rta');

INSERT INTO 'etl_config'.'rename_regexp'(id, task_name, db_pattern, new_db) VALUES (2, 'etl_beijing', '^rdb', 'rdb');

INSERT INTO 'etl_config'.'rename_regexp'(id, task_name, db_pattern, table_pattern, new_db, new_table) VALUES (3, 'etl_beijing', '^rdc', '^rtc[a-z]*rtcc\$', 'rdc', 'rtc');

#filter opt 测试数据

INSERT INTO 'etl_config'.'filter_opt'(id, task_name, dbname, tablename, optype, action)
VALUES

(1, 'etl_beijing', 'animal', 'dog', 'ALTERTABLE', 'ADD COLUMN, ADD INDEX'),(2, 'etl beijing', 'animal', 'dog', 'ALTERTABLE', 'DROP COLUMN, DROP INDEX');

INSERT INTO 'etl_config'.filter_opt(id, task_name, dbname, optype) VALUES (3, 'etl_beijing', 'animal1', 'CREATETABLE');

INSERT INTO 'etl_config'.filter_opt(id, task_name, tablename, optype) VALUES (4, 'etl_beijing', 'cat', 'TRUNCATE');

INSERT INTO 'etl_config'.filter_opt(id, task_name, optype) VALUES (5, 'etl_beijing', 'DROPTABLE');

#filter col 测试数据

INSERT INTO 'etl_config'.filter_col(id, task_name, dbname, tablename, column_names)
VALUES (1, 'etl_beijing', 'school', 'teacher', 'hobby,address'); #1.普通字段过滤

INSERT INTO 'etl_config'.filter_col(id, task_name, dbname, tablename, column_names)
VALUES (2, 'etl_beijing', 'school', 'teacher', 'column'); #2.特殊字段过滤;12.多个过滤规则作用在同一张表上的字段过滤

INSERT INTO 'etl_config'.filter_col(id, task_name, dbname, tablename, column_names)
VALUES (3, 'etl_beijing', 'school', 'student', 'math, english'); #4.过滤多个字段

INSERT INTO 'etl_config'.filter_col(id, task_name, dbname, tablename, column_names)
VALUES (4, 'etl_beijing', 'school', 'student', 'idxcol'); #5.要过滤的字段上有索引

INSERT INTO 'etl_config'.filter_col(id, task_name, dbname, tablename, column_names)
VALUES (5, 'etl_beijing', 'school', 'student', 'xingming'); #6.要过滤的字段有发生了 change 字段

INSERT INTO `etl_config`.filter_col(id, task_name, dbname, column_names) VALUES (6, 'etl_beijing', 'school', 'telephone'); #7.过滤一个库下所有表的该字段

INSERT INTO `etl_config`.filter_col(id, task_name, tablename, column_names) VALUES (7, 'etl_beijing', 'all_table', 'col1, col2'); #8.过滤所有表名同为 ta 的一个或多个字段

INSERT INTO 'etl_config'.filter_col(id, task_name, dbname, tablename, column_names)
VALUES (8, 'etl_beijing', 'company', 'all_table', 'col3, col3'); #9.过滤所有表名同为 ta 的一个或
多个字段

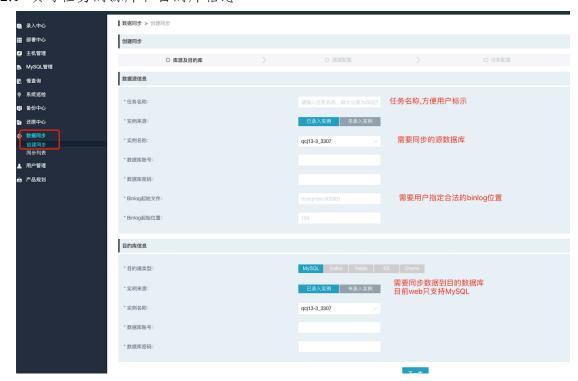
INSERT INTO `etl_config`.filter_col(id, task_name, dbname, tablename, column_names)
VALUES (9, 'etl_beijing', 'company', 'order', 'col100'); #10.如果要过滤的字段不再表中是否有影响

INSERT INTO `etl_config`.filter_col(id, task_name, dbname, tablename) VALUES (10, 'etl_beijing', 'company', 'library'); #13.column_names 为空,不过虑

INSERT INTO 'etl_config'.filter_col(id, task_name, column_names) VALUES (11, 'etl_beijing', 'money');

2.2 创建任务

2.2.1 填写任务的源库和目的库信息



2.2.2 通道配置(读源库配置)



高级配置可以指定修改同步参数,以及黑白名单等

| | 收粒配置 | | | | | |
|---|--|---|--|--|--|--|
| 通道-参数配置 | | ~ | | | | |
| | | | | | | |
| 同步线程: | 5 配置Arkgate的同步并发线程数 | | | | | |
| 工作队列长度: | 10000 | | | | | |
| 源库架构类型: | 例生MySOL 阿里云(ROS) | | | | | |
| 是否包含表结构的全部信息: | 包含 就认包含,用于同步到MySQL场景,不包含适用于HBase、ES等场景 | | | | | |
| 数据保留时长(小时): | 168 Arkgate集群中数据保留时长以及清理策略 | | | | | |
| 删除数据分配大小(行数): | 5000 | | | | | |
| 删除数据线程数: | 5 | | | | | |
| 1 | | | | | | |
| 通道-过滤配置 | | ~ | | | | |
| 白名单:只同步符合要求的表数据。黑名单:忽略符合要求的表数据。两种 | 者只能配置一种 | | | | | |
| 英型: 黒名単 白名単 | 表名称 | | | | | |
| 数据库名称 database name | 农合和 table name 樹脂 | | | | | |
| 添加 | | | | | | |
| | | | | | | |
| 通道-数据源高可用配置 | | ~ | | | | |
| 备注:数据赛高可用配置,用于当同步源数据库出现岩机后,或者别的原因导致不可用的时候,如果配置了高可用,逐速可以自动的切换到可用的从库,继续同步,确保同步正常。 | | | | | | |
| * 实例来源: 已录入实例 未录入实例 | | | | | | |
| * 突例名称: 请选择 | 無印象 | | | | | |
| 语加观从库 | | | | | | |
| | | | | | | |

2.2.3 任务配置(写目的库配置)

仅展示用户设置的通道名称

| 数据同步 > 创建同步 | | | | | | | | |
|-------------|--|--------------|--------------|--------|--|-------|--|--|
| 创建同步 | | | | | | | | |
| 〇 库源及目的库 | | ○ 通道配置 | | ○ 任务配置 | | 〇 預检查 | | |
| 任务基本信息 | | | | | | | | |
| 通道名称: | | arkgate_demo | 仅展示用户配置的通道名称 | | | | | |
| 用4000世 | | | | | | | | |
| | | | 上一步 预检查 | | | | | |

高级设置可以配置写目的库时的并发,表重命名,操作过滤等策略



2.2.4 预检查及提交

配置完成后点击预检查



如果检测不通过,可以点击失败,显示详情



如果检查通过则可以点击提交,完成任务创建

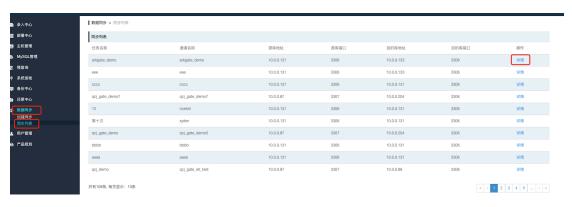


提交后等待几秒钟(创建同步任务比较耗时),会自动跳转到详情页面



3 Arkcontrol 同步任务管理

3.1 进入数据同步, 查看同步列表



- 3.2 点击详情进入详情页面
- 3.3 修改通道信息(读取源库配置)
- 3.3.1 添加黑白名单

| alpha版 | | | | | | |
|---------------------------|------------------------------------|--------|-----------|------|---------|----|
| | | 新增黑白名单 | | × | | |
| 实例名称: sk1_3306 | 数: | | | | □: 3306 | |
| Binlog起始文件: mysql-bin.000 | 023 Bin | 数据库名称: | filter_db | | | |
| | | 表名称: | % | | | |
| 目的库信息 | | 过滤类型: | 白名单 💠 | | | v |
| | | | | | | |
| 实例名称: sk3_3306 | 数主 | | | | □: 3306 | |
| 实例类型: MySQL | | | | | | |
| | | | | 取消添加 | | |
| 通道信息 | | | | | | ~ |
| 同步工作集群 黒白名单 同 | 步参数配置 数据源切换 日志信息 | | | | | |
| 白名单: 只同步符合要求的表数 | | | | | | |
| 黑名单: 忽略符合要求的表数据 | | | | | | |
| 白名单优先于黑名单 | | | | | | |
| *由于新增和修改操作影响比较重 | | | | | | |
| 数据库名称 | 表名称 | | | 过滤类型 | | 操作 |
| 新増 保存 制約 | 添加完成后,点击保存才会生效 可以一次性添加多个,然后保存生效 | | | | | |
| | 可以一次性添加多个,然后保存生效 | | | | | |

3.3.2 数据源切换

数据源切换,可以开始,关闭同步任务



3.4 修改任务-规则信息(写入目的库配置)

