## Gangtise数据定义文档

源服务器：rm-uf6qosonty87i1729uo.mysql.rds.aliyuncs.com

数据库：dpdata

用户：hifi，密码：\_Huawei123

目标服务器：n5.hifiax.com

数据库：dpdata

用户：hifi，密码：hifi

访问工具：DBeaver

开发服务器：n5.hifiax.com

工作目录：~/pwork/gangtise

1. 数据字典

data\_dic.txt，来自：datapipe\_product\_dic表，共计124个

data\_size.txt，每个表的记录条数、空间大小

data\_dic.xlsx，数据表清单，带类别标签

编写：segment.py，小于100G的数据表，采取dump再导入的方式；大于100G的数据表，只取2025年的数据

list4dump.txt，小表格，90个，1.04G

list4read.txt，大表格，34个

1. 建表语句

编写：get\_schema.py，获取建表语句

124个建表语句存入schema子目录

编写：batch\_create.py，批量建表

1. 批量导入小表格

编写：batch\_dump.sh，从源服务器批量导出小表格，放在dump子目录

编写：batch\_load.sh，在目标服务器批量导入小表格

1. 批量导入大表格

编写：batch\_read.py，从源服务器批量读取大表格，放在y2025子目录，ashare\_main\_product、index\_price表2025年为空记录，往前调半年

每10000条生成一个文件，超过20000条即建子目录

跑3小时，总计2.1G，最大的三个表：

499M ./ashare\_prffrcastdt

527M ./ashare\_prffrcaststat

381M ./ashare\_price\_der

对于大字段文件导入，需要在/etc/mysql/conf.d/mysql.cnf中修改mysql的配置参数，然后重启mysql

[mysqld]

max\_allowed\_packet = 100M

编写：batch\_insert.sh，在目标服务器批量导入大表格，遍历各层子目录

编写：batch\_delete.py，清除大表格数据，顺便查看记录条数

编写：batch\_check.py，查看大表格记录条数

修复工具：

编写：batch\_read\_fix.py，修复表格名错误，已在源头修复

编写：batch\_read\_fixm.py，修复日期未加引号错误

编写：batch\_read\_fixq.py，修复内容中出现引号错误

1. 持续更新

编写：incre\_read.py，从源服务器批量读取所有数据表的增量部分，放在c2025子目录

借用：batch\_insert.sh，在目标服务器批量导入，遍历各层子目录

修复工具：

编写：incre\_read\_fix.py，修复单引号错误，已在源头修复

更新记录：

c20250428、c20250508、c20250620

1. 数据备份

使用华为云OBS存储服务，桶为gangtise

工具目录：~/pwork/obs

备份命令：

python ~/pwork/obs/upload.py c20250508

下载命令：

python ~/pwork/obs/download.py c20250508

1. 常用数据表

ashare\_securityinfo，A股证券基本信息

ashare\_concept\_investmtframe，概念题材投资框架

opinion\_news，观点信息表

1. opinion\_news观点信息表

代码：incre\_opinion.py，同步数据

首先id不是线性的，而是按类型分区，因此要根据update\_time来作同步。

代码：ana\_opinion.py，数据分析

四种类型，只取券商公众号、调研公告。根据相关股票标签，取近2周相关新闻。

建议：每日晚上同步，离线建索引，以方便快速检索。