## 美股抓取计划

项目介绍：目前我们想要抓取的数据源“Market Watch”使用的财务数据和我们自己的美股数据服务商都是“Factset”。为了能够减轻对于Factset的数据依赖，我们考虑从MW抓取基本的财务数据，并整理他们的规则。

（以下所涉及的抓取流程都用美股“FISV”作为例子进行介绍：）

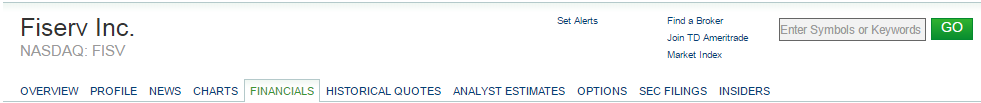
以下是需要抓的URL类表：

|  |  |
| --- | --- |
| 年度IS表 | http://www.marketwatch.com/investing/Stock/FISV/financials |
| 季度IS表 | <http://www.marketwatch.com/investing/Stock/FISV/financials/income/quarter> |
| 年度BS表 | <http://www.marketwatch.com/investing/Stock/FISV/financials/balance-sheet> |
| 季度BS表 | <http://www.marketwatch.com/investing/Stock/FISV/financials/balance-sheet/quarter> |
| 年度CF表 | <http://www.marketwatch.com/investing/Stock/FISV/financials/cash-flow> |
| 季度CF表 | <http://www.marketwatch.com/investing/Stock/FISV/financials/cash-flow/quarter> |

注：进行抓取的时候需要注意根据GAF提供的美股代码对URL里面的代码字段进行替换，并抓取；

具体抓取步骤

1. 公司的代码和公司的名字，同时在存储的时候需要有三个字段：
   1. Name: Fiserv Inc.
   2. Market: NASDAQ
   3. Ticker: FISV

 公司在正常存续区间通过每一个代码所代表的公司名称和财务信息是一致的，但是美股有很多公司经常发现更改代码或者发生转板的情形，所以需要在将数据存入的时候进行判断：

1. 如果在同一个URL的页面抓取下面，“Market”发生了变化，即公司发生了转板的情形，那么需要在数据库里面重新有一个地方存储市场的字段，比如Market\_2同时需要存储抓取到发生变更的时间“年-月-日”；（这里只是一种解决方案，其他方式也请技术支持）
2. 如果在同一个URL的页面抓取下面，公司的Name发生了变化，我们需要有一个新增字段Name\_2来记录新的名字。（这里只是一种解决方案，其他方式也请技术支持）
3. 如果在同一个URL的页面下已经找不到相关公司，说明这家公司发生了代码变更或者公司不再正常交易了。这时候，我们要做的是及时找到发生这种情形的原因，并将更正的信息及时给到技术.
4. 以上三种情形不管是何时发生，都需要有一个报警机制告诉相关的业务人员去核实触发三种情形的原因；
5. 业务这边的同事会定期将最新的代码提供给技术进行抓取，这样能够及时保证新增公司的数据覆盖，以及发生了情形3的公司的补足；
6. 年度和季度数据抓取（IS）
   1. 一般样式抓取：请通过年度IS表提供的链接对于美股的基本样式进行抓取，其中主要包括非财务类科目和非数值类科目的字段.这类数据和财务数据是同时存在于一张表“http://www.marketwatch.com/investing/Stock/FISV/financials”



|  |  |
| --- | --- |
| 字段名称 | 字段内容 |
| Currency | USD |
| Unit | Millions |
| FY | January-December |

* 1. 科目样式抓取：

1. 样式表需要另外存储一张表，和数据表不要放在一起；
2. 科目内容和样式只需要抓取年度数据，即利润表，资产负债表和现金流量表的年度科目样式；
3. 由于获取不到历史期的科目样式，我们目前只是按照当下最新期的科目进行抓取，所以样式表的主键中不需要有财务期；
4. 由于我们目前没有办法对于Market Watch的样式有一个很好的归纳，所以需要技术这边将每一家公司每一个科目都抓取下来，等日后容错和规则完善之后，就不再需要存储全部公司的的样式了。

|  |  |
| --- | --- |
| 字段名称 | 字段内容 |
| 科目序列  Serie | 1,2,3,4,等序列  具体规则：是从“Fiscal year”所在行业的下面一个科目开始。由于很多科目是有加号我们以加好开始的位置进行判断. |
| 科目的层级  Level | L1，L2，L3，L4  具体规则：以抓取的第一个科目作为参照物，如果其他科目的位置是相对这个有1个空格的，我们暂定是L2，有两个空格是L3，以此类推。  如果同级的就是L0。另外，很多科目是有加号我们以加好开始的位置进行判断. |
| 科目的名称  Item | 财务科目名称  具体规则：保证科目抓取完整 |

* 1. 财务数据抓取

1. 根据样式表抓取完成以后，我们对于每一个独一无二的科目都给上一个code，这样对于财务数据表存储数据会带来很多好处；
2. 由于页面上抓取的是科目的名称，需要将其转换成我们自己定义的code，然后将每一期所对应的数据存储到财务数据表；
3. 财务数据存储表是根据公司的代码，市场和财务期所作为主键来构造，同时要将财务日期抓换成“年-月-日”这样的格式；
4. 表结构字段中，需要有一个描述数据频率的字段，主要是年度和季度数据；
5. Market Watch一般会披露5年的数据，对应科目的名称将每一期的数据都存储下来；
6. 对于不同频率的财务数据是否要分表存储，也需要技术评估