使用base模板在本机运行Scrapy框架实现数据采集

- 主页网址为：[财经日历-金十数据]
- 目标数据
    - 时间,数据,重要性,前值,预测值,公布值
- 要求
    1. 使用scrapy 数据抓取目标数据 并存入MySQL数据库
    2. 每个爬虫使用一个setting配置的方式

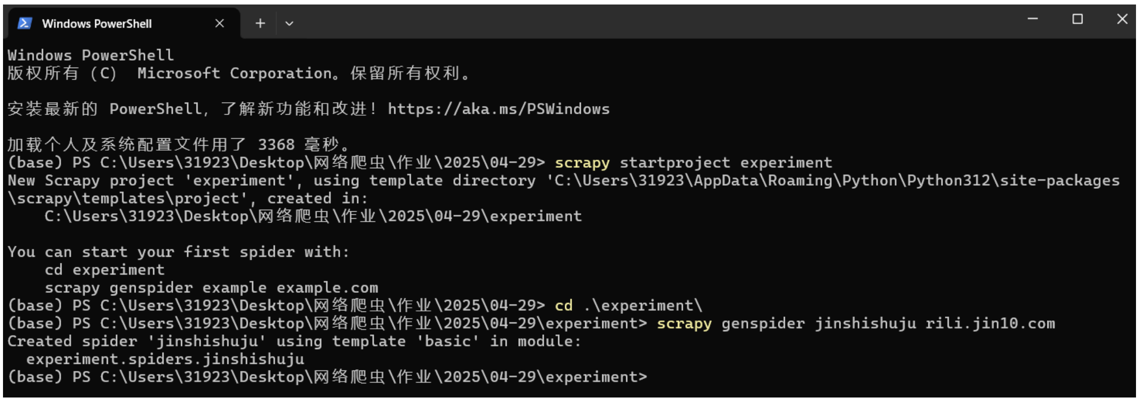
步骤：

1. 创建Scrapy项目

```
scrapy startproject experiment
```

2. 进入项目并创建爬虫

```
cd .\experiment\

scrapy genspider jinshishuju rili.jin10.com
```



```
Windows PowerShell
版权所有 (C)  Microsoft Corporation。保留所有权利。

安装最新的 PowerShell，了解新功能和改进！https://aka.ms/PSWindows

加载个人及系统配置文件用了 3368 毫秒。
(base) PS C:\Users\31923\Desktop\网络爬虫\作业\2025\04-29> scrapy startproject experiment
New Scrapy project 'experiment', using template directory 'C:\Users\31923\AppData\Roaming\Python\Python312\site-packages
\scrapy\templates\project', created in:
    C:\Users\31923\Desktop\网络爬虫\作业\2025\04-29\experiment

You can start your first spider with:
    cd experiment
    scrapy genspider example example.com
(base) PS C:\Users\31923\Desktop\网络爬虫\作业\2025\04-29> cd .\experiment\
(base) PS C:\Users\31923\Desktop\网络爬虫\作业\2025\04-29\experiment> scrapy genspider jinshishuju rili.jin10.com
Created spider 'jinshishuju' using template 'basic' in module:
  experiment.spiders.jinshishuju
(base) PS C:\Users\31923\Desktop\网络爬虫\作业\2025\04-29\experiment>
```

3. 代码实现

    - jinshishuju.py

```
import json
import scrapy
from experiment.items import JinshishujuItem
from datetime import datetime, timedelta

class JinshishujuSpider(scrapy.Spider):
    name = "jinshishuju"
    # 单独设置该爬虫的配置
    custom_settings = {
        'USER_AGENT': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0 Safari/537.36',
        'ROBOTSTXT_OBEY': False,
        'ITEM_PIPELINES': {
            'experiment.pipelines.MySQLPipeline': 300,
        }
    }
    # allowed_domains = ["rili.jin10.com"]
    url =
'https://e0430d16720e4211b5e072c26205c890.z3c.jin10.com/get/data?date='
    headers = {
        "accept": "application/json, text/plain, */*",
        "origin": "https://rili.jin10.com",
        "referer": "https://rili.jin10.com/",
        "user-agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0 Safari/537.36",
        "x-app-id": "sKKYe29sFuJaeOCJ",
        "x-version": "2.0"
    }
    def start_requests(self):
        # 定义起止日期
        start_date = datetime.strptime("2025-05-16", "%Y-%m-%d")
        end_date = datetime.strptime("2025-05-16", "%Y-%m-%d")
        current_date = start_date
        data_list = []
        # 循环生成从起始日期到结束日期之间每天的URL请求
        while current_date <= end_date:
            data_list.append(current_date.strftime('%Y-%m-%d'))
            # 通过yield发送请求，meta中传递当前日期，方便parse中使用
            current_date += timedelta(days=1)  # 日期+1天

        for i in data_list:
            url = self.url + f'{i}&category=cj'
            print(url)
            yield scrapy.Request(url=url,headers=
self.headers,callback=self.parse)

    def parse(self, response, **kwargs):
        data = json.loads(response.text)
        data_dict = data.get('data', [])
        if data_dict is None:
            print("接口没有返回数据")
            return
        for i in data_dict:
            # 时间
            js_time = i.get('actual_time')
            if js_time is None:
                js_time = '时间数据为空'

            # 数据
            country = i.get('country') or ''
            time_period = i.get('time_period') or ''
            indicator_name = i.get('indicator_name') or ''
            js_data = country + time_period + indicator_name
            if not js_data:
                js_data = '数据为空'

            # 重要性
            js_star = i.get('star', None)
            if js_star is not None:
                if js_star == 1:
                    js_star = '很低'
                if js_star == 2:
                    js_star = '低'
                if js_star == 3:
                    js_star = '中'
                if js_star == 4:
                    js_star = '高'
                if js_star == 5:
                    js_star = '很高'
            else:
                js_star = '重要性为空'
            # 前值
            js_previous = i.get('previous', None)
            if js_previous is not None:
                js_previous += '%'
            else:
                js_previous = '前值为空'

            # 预测值
            consensus = i.get('consensus')
            if consensus is not None:
                consensus += '%'
            else:
                consensus = '预测值为空'

            # 公布值
            js_actual = i.get('actual', None)
            if js_actual is not None:
                js_actual += '%'
            else:
                js_actual = '公布值为空'
            item = JinshishujuItem()
            item['time'] = js_time
            item['data'] = js_data
            item['importance'] = js_star
            item['previous'] = js_previous
            item['actual'] = js_actual
            item['consensus'] = consensus
            # print('时间', js_time)
            # print('数据', js_data)
            # print('重要性', js_star)
            # print('前值', js_previous)
            # print('预测值', consensus)
            # print('公布值', js_actual)
            # print('--------------------------------')
            yield item
```

- items.py

```
import scrapy

class JinshishujuItem(scrapy.Item):
    time = scrapy.Field()
    data = scrapy.Field()
    importance = scrapy.Field()
    previous = scrapy.Field()
    consensus = scrapy.Field()
    actual = scrapy.Field()
```

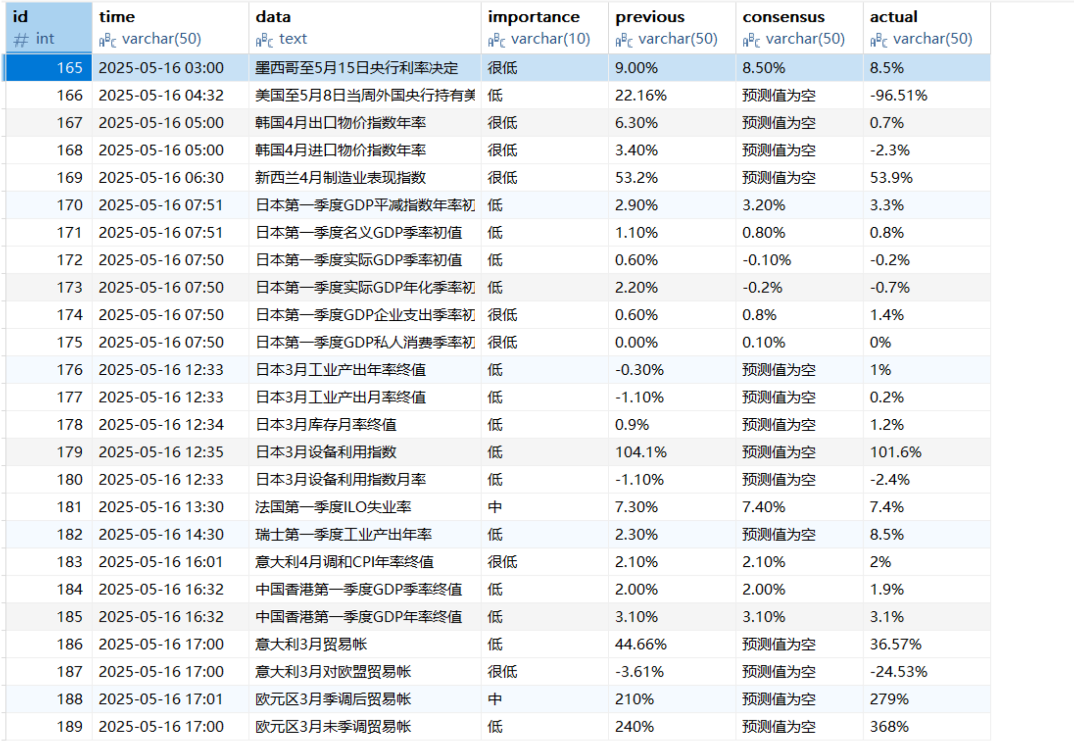
- pipelines.py

```
# Define your item pipelines here
#
# Don't forget to add your pipeline to the ITEM_PIPELINES setting
# See: https://docs.scrapy.org/en/latest/topics/item-pipeline.html


# useful for handling different item types with a single interface
from itemadapter import ItemAdapter


# pipelines.py
import pymysql
from pymysql.err import OperationalError

class MySQLPipeline:
    def open_spider(self, spider):
        self.conn = pymysql.connect(
            host='localhost',
            user='root',
            password='123456',
            database='cj_data',
            charset='utf8mb4',
            cursorclass=pymysql.cursors.DictCursor
        )
        self.cursor = self.conn.cursor()

    def process_item(self, item, spider):
        sql = """
        INSERT INTO jinshishuju (time, data, importance, previous, consensus, actual)
        VALUES (%s, %s, %s, %s, %s, %s)
        """
        try:
            self.cursor.execute(sql, (
                item.get('time'),
                item.get('data'),
                item.get('importance'),
                item.get('previous'),
                item.get('consensus'),
                item.get('actual')
            ))
            self.conn.commit()
        except OperationalError as e:
            spider.logger.error(f"写入数据库错误: {e}")
            self.conn.rollback()
        return item

    def close_spider(self, spider):
        self.cursor.close()
        self.conn.close()
```

- run.py

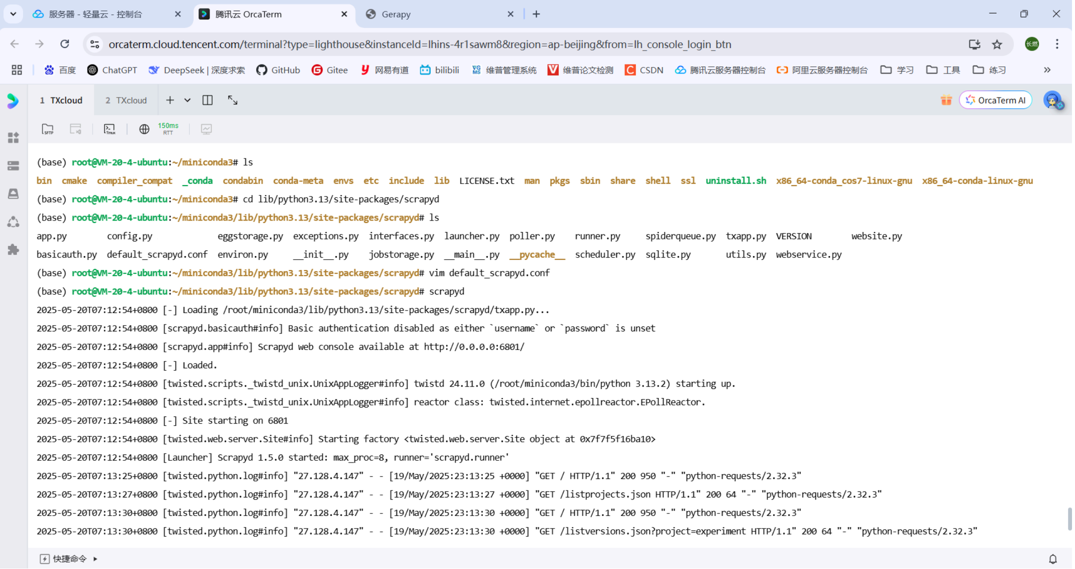
```
from scrapy.cmdline import execute

execute('scrapy crawl jinshishuju'.split())
```

- 运行结果（时间范围可修改，以2025-05-16到2025-05-16为例）



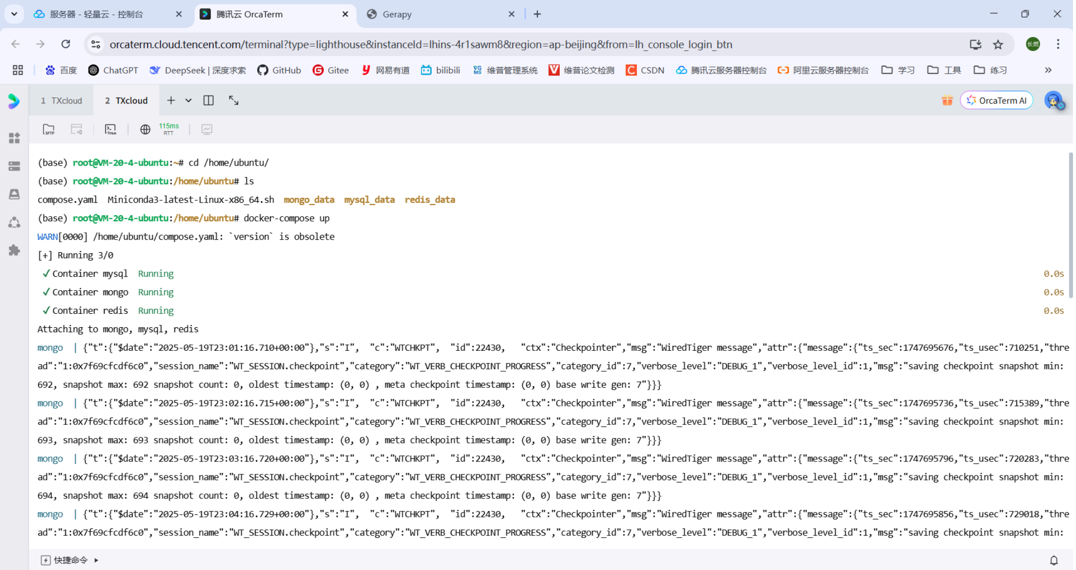
id # int	time varchar(50)	data text	importance varchar(10)	previous varchar(50)	consensus varchar(50)	actual varchar(50)
165	2025-05-16 03:00	墨西哥至5月15日央行利率决定	很低	9.00%	8.50%	8.5%
166	2025-05-16 04:32	美国至5月8日当周外国央行持有美	低	22.16%	预测值为空	-96.51%
167	2025-05-16 05:00	韩国4月出口物价指数年率	很低	6.30%	预测值为空	0.7%
168	2025-05-16 05:00	韩国4月进口物价指数年率	很低	3.40%	预测值为空	-2.3%
169	2025-05-16 06:30	新西兰4月制造业表现指数	很低	53.2%	预测值为空	53.9%
170	2025-05-16 07:51	日本第一季度GDP平减指数年率初	低	2.90%	3.20%	3.3%
171	2025-05-16 07:51	日本第一季度名义GDP季率初值	低	1.10%	0.80%	0.8%
172	2025-05-16 07:50	日本第一季度实际GDP季率初值	低	0.60%	-0.10%	-0.2%
173	2025-05-16 07:50	日本第一季度实际GDP年化季率初	低	2.20%	-0.2%	-0.7%
174	2025-05-16 07:50	日本第一季度GDP企业支出季率初	很低	0.60%	0.8%	1.4%
175	2025-05-16 07:50	日本第一季度GDP私人消费季率初	很低	0.00%	0.10%	0%
176	2025-05-16 12:33	日本3月工业产出年率终值	低	-0.30%	预测值为空	1%
177	2025-05-16 12:33	日本3月工业产出月率终值	低	-1.10%	预测值为空	0.2%
178	2025-05-16 12:34	日本3月库存月率终值	低	0.9%	预测值为空	1.2%
179	2025-05-16 12:35	日本3月设备利用指数	低	104.1%	预测值为空	101.6%
180	2025-05-16 12:33	日本3月设备利用指数月率	低	-1.10%	预测值为空	-2.4%
181	2025-05-16 13:30	法国第一季度ILO失业率	中	7.30%	7.40%	7.4%
182	2025-05-16 14:30	瑞士第一季度工业产出年率	低	2.30%	预测值为空	8.5%
183	2025-05-16 16:01	意大利4月调和CPI年率终值	很低	2.10%	2.10%	2%
184	2025-05-16 16:32	中国香港第一季度GDP季率终值	低	2.00%	2.00%	1.9%
185	2025-05-16 16:32	中国香港第一季度GDP年率终值	低	3.10%	3.10%	3.1%
186	2025-05-16 17:00	意大利3月贸易帐	低	44.66%	预测值为空	36.57%
187	2025-05-16 17:00	意大利3月对欧盟贸易帐	很低	-3.61%	预测值为空	-24.53%
188	2025-05-16 17:01	欧元区3月季调后贸易帐	中	210%	预测值为空	279%
189	2025-05-16 17:00	欧元区3月未季调贸易帐	低	240%	预测值为空	368%

- 项目部署到服务器上执行
  - 启动scrapyd服务



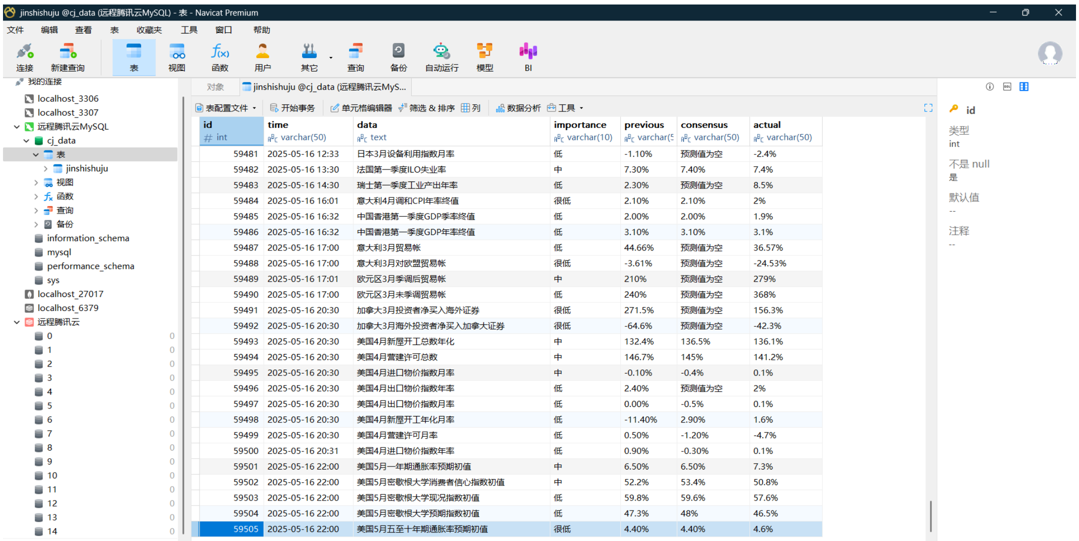
```
(base) root@VM-20-4-ubuntu:~/miniconda3# ls
bin  cmake  compiler_compat  _conda  condabin  conda-meta  envs  etc  include  lib  LICENSE.txt  man  pkgs  sbin  share  shell  ssl  uninstall.sh  x86_64-conda_cos7-linux-gnu  x86_64-conda-linux-gnu
(base) root@VM-20-4-ubuntu:~/miniconda3# cd lib/python3.13/site-packages/scrapyd
(base) root@VM-20-4-ubuntu:~/miniconda3/lib/python3.13/site-packages/scrapyd# ls
app.py        config.py              eggstorage.py  exceptions.py  interfaces.py  launcher.py  poller.py     runner.py     spiderqueue.py  txapp.py  VERSION        website.py
basicauth.py  default_scrapyd.conf  environ.py     __init__.py    jobstorage.py  __main__.py  __pycache__  scheduler.py  sqlite.py       utils.py  webservice.py
(base) root@VM-20-4-ubuntu:~/miniconda3/lib/python3.13/site-packages/scrapyd# vim default_scrapyd.conf
(base) root@VM-20-4-ubuntu:~/miniconda3/lib/python3.13/site-packages/scrapyd# scrapyd
2025-05-20T07:12:54+0800 [-] Loading /root/miniconda3/lib/python3.13/site-packages/scrapyd/txapp.py...
2025-05-20T07:12:54+0800 [scrapyd.basicauth#info] Basic authentication disabled as either `username` or `password` is unset
2025-05-20T07:12:54+0800 [scrapyd.app#info] Scrapyd web console available at http://0.0.0.0:6801/
2025-05-20T07:12:54+0800 [-] Loaded.
2025-05-20T07:12:54+0800 [twisted.scripts._twistd_unix.UnixAppLogger#info] twistd 24.11.0 (/root/miniconda3/bin/python 3.13.2) starting up.
2025-05-20T07:12:54+0800 [twisted.scripts._twistd_unix.UnixAppLogger#info] reactor class: twisted.internet.epollreactor.EPollReactor.
2025-05-20T07:12:54+0800 [-] Site starting on 6801
2025-05-20T07:12:54+0800 [twisted.web.server.Site#info] Starting factory <twisted.web.server.Site object at 0x7f7f5f16ba10>
2025-05-20T07:12:54+0800 [Launcher] Scrapyd 1.5.0 started: max_proc=8, runner='scrapyd.runner'
2025-05-20T07:13:25+0800 [twisted.python.log#info] "27.128.4.147" - - [19/May/2025:23:13:25 +0000] "GET / HTTP/1.1" 200 950 "-" "python-requests/2.32.3"
2025-05-20T07:13:27+0800 [twisted.python.log#info] "27.128.4.147" - - [19/May/2025:23:13:27 +0000] "GET /listprojects.json HTTP/1.1" 200 64 "-" "python-requests/2.32.3"
2025-05-20T07:13:30+0800 [twisted.python.log#info] "27.128.4.147" - - [19/May/2025:23:13:30 +0000] "GET / HTTP/1.1" 200 950 "-" "python-requests/2.32.3"
2025-05-20T07:13:30+0800 [twisted.python.log#info] "27.128.4.147" - - [19/May/2025:23:13:30 +0000] "GET /listversions.json?project=experiment HTTP/1.1" 200 64 "-" "python-requests/2.32.3"
```

  - docker-compose up



```
(base) root@VM-20-4-ubuntu:~# cd /home/ubuntu/
(base) root@VM-20-4-ubuntu:/home/ubuntu# ls
compose.yaml  Miniconda3-latest-Linux-x86_64.sh  mongo_data  mysql_data  redis_data
(base) root@VM-20-4-ubuntu:/home/ubuntu# docker-compose up
WARN[0000] /home/ubuntu/compose.yaml: `version` is obsolete
[+] Running 3/0
 ✓Container mysql  Running                                                                                                              0.0s
 ✓Container mongo  Running                                                                                                              0.0s
 ✓Container redis  Running                                                                                                              0.0s
Attaching to mongo, mysql, redis
mongo  | {"t":{"$date":"2025-05-19T23:01:16.710+00:00"},"s":"I",  "c":"WTCHKPT",  "id":22430,   "ctx":"Checkpointer","msg":"WiredTiger message","attr":{"message":{"ts_sec":1747695676,"ts_usec":710251,"thre
ad":"1:0x7f69cfcdf6c0","session_name":"WT_SESSION.checkpoint","category":"WT_VERB_CHECKPOINT_PROGRESS","category_id":7,"verbose_level":"DEBUG_1","verbose_level_id":1,"msg":"saving checkpoint snapshot min:
692, snapshot max: 692 snapshot count: 0, oldest timestamp: (0, 0) , meta checkpoint timestamp: (0, 0) base write gen: 7"}}}
mongo  | {"t":{"$date":"2025-05-19T23:02:16.715+00:00"},"s":"I",  "c":"WTCHKPT",  "id":22430,   "ctx":"Checkpointer","msg":"WiredTiger message","attr":{"message":{"ts_sec":1747695736,"ts_usec":715389,"thre
ad":"1:0x7f69cfcdf6c0","session_name":"WT_SESSION.checkpoint","category":"WT_VERB_CHECKPOINT_PROGRESS","category_id":7,"verbose_level":"DEBUG_1","verbose_level_id":1,"msg":"saving checkpoint snapshot min:
693, snapshot max: 693 snapshot count: 0, oldest timestamp: (0, 0) , meta checkpoint timestamp: (0, 0) base write gen: 7"}}}
mongo  | {"t":{"$date":"2025-05-19T23:03:16.720+00:00"},"s":"I",  "c":"WTCHKPT",  "id":22430,   "ctx":"Checkpointer","msg":"WiredTiger message","attr":{"message":{"ts_sec":1747695796,"ts_usec":720283,"thre
ad":"1:0x7f69cfcdf6c0","session_name":"WT_SESSION.checkpoint","category":"WT_VERB_CHECKPOINT_PROGRESS","category_id":7,"verbose_level":"DEBUG_1","verbose_level_id":1,"msg":"saving checkpoint snapshot min:
694, snapshot max: 694 snapshot count: 0, oldest timestamp: (0, 0) , meta checkpoint timestamp: (0, 0) base write gen: 7"}}}
mongo  | {"t":{"$date":"2025-05-19T23:04:16.729+00:00"},"s":"I",  "c":"WTCHKPT",  "id":22430,   "ctx":"Checkpointer","msg":"WiredTiger message","attr":{"message":{"ts_sec":1747695856,"ts_usec":729018,"thre
ad":"1:0x7f69cfcdf6c0","session_name":"WT_SESSION.checkpoint","category":"WT_VERB_CHECKPOINT_PROGRESS","category_id":7,"verbose_level":"DEBUG_1","verbose_level_id":1,"msg":"saving checkpoint snapshot min:
```

- 运行结果（时间可修改，这个以2025-01-01到2025-05-19为例）



id int	time varchar(50)	data text	importance varchar(10)	previous varchar(5	consensus varchar(50)	actual varchar(50)
59481	2025-05-16 12:33	日本3月设备利用指数月率	低	-1.10%	预测值为空	-2.4%
59482	2025-05-16 13:30	法国第一季度ILO失业率	中	7.30%	7.40%	7.4%
59483	2025-05-16 14:30	瑞士第一季度工业产出年率	低	2.30%	预测值为空	8.5%
59484	2025-05-16 16:01	意大利4月调和CPI年率终值	很低	2.10%	2.10%	2%
59485	2025-05-16 16:32	中国香港第一季度GDP季率终值	低	2.00%	2.00%	1.9%
59486	2025-05-16 16:32	中国香港第一季度GDP年率终值	低	3.10%	3.10%	3.1%
59487	2025-05-16 17:00	意大利3月贸易帐	低	44.66%	预测值为空	36.57%
59488	2025-05-16 17:00	意大利3月对欧盟贸易帐	很低	-3.61%	预测值为空	-24.53%
59489	2025-05-16 17:01	欧元区3月季调后贸易帐	中	210%	预测值为空	279%
59490	2025-05-16 17:00	欧元区3月未季调贸易帐	低	240%	预测值为空	368%
59491	2025-05-16 20:30	加拿大3月投资者净买入海外证券	很低	271.5%	预测值为空	156.3%
59492	2025-05-16 20:30	加拿大3月海外投资者净买入加拿大证券	很低	-64.6%	预测值为空	-42.3%
59493	2025-05-16 20:30	美国4月新屋开工总数年化	中	132.4%	136.5%	136.1%
59494	2025-05-16 20:30	美国4月营建许可总数	中	146.7%	145%	141.2%
59495	2025-05-16 20:30	美国4月进口物价指数月率	中	-0.10%	-0.4%	0.1%
59496	2025-05-16 20:30	美国4月出口物价指数年率	低	2.40%	预测值为空	2%
59497	2025-05-16 20:30	美国4月出口物价指数月率	低	0.00%	-0.5%	0.1%
59498	2025-05-16 20:30	美国4月新屋开工年化月率	低	-11.40%	2.90%	1.6%
59499	2025-05-16 20:30	美国4月营建许可月率	低	0.50%	-1.20%	-4.7%
59500	2025-05-16 20:31	美国4月进口物价指数年率	低	0.90%	-0.30%	0.1%
59501	2025-05-16 22:00	美国5月一年期通胀率预期初值	中	6.50%	6.50%	7.3%
59502	2025-05-16 22:00	美国5月密歇根大学消费者信心指数初值	中	52.2%	53.4%	50.8%
59503	2025-05-16 22:00	美国5月密歇根大学现况指数初值	低	59.8%	59.6%	57.6%
59504	2025-05-16 22:00	美国5月密歇根大学预期指数初值	低	47.3%	48%	46.5%
59505	2025-05-16 22:00	美国5月五至十年期通胀率预期初值	很低	4.40%	4.40%	4.6%