



上海立信会计金融学院
SHANGHAI LIXIN UNIVERSITY OF ACCOUNTING AND FINANCE

2025 届本科毕业论文

基于大数据流式引擎构建分析订单簿实时买卖
压力差的趋势量化交易策略（简略）

**Developing a Quantitative Trading Strategy
Based on Big Data Streaming Engines to
Analyze Trends in Real-time Bid-ask Pressure
Differentials From Order Books(Simplified)**

学生姓名：

王灿霖

所在学院：

金融科技学院

专 业：

金融科技

指导老师：

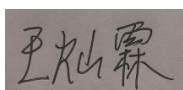
池文涛

2024 年 12 月

声明及论文使用的授权

本人郑重声明：所呈交的论文是本人在导师的指导下取得的科研成果，论文写作严格遵循学术规范。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果。因本毕业论文引起的法律结果完全由本人承担。上海立信会计金融学院享有本毕业论文的研究成果。

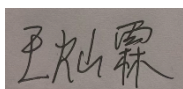
论文作者签名



2025 年 3 月 1 日

本人同意上海立信会计金融学院保留使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以上网公布全部内容，可以采用影印、缩印或其他复制手段保存论文。

论文作者签名



2025 年 3 月 1 日

摘 要

Abstract

一、绪论

（一）研究背景及其意义

（二）文献综述

（三）国内外高频量化交易现状

- 1. 人才资源层面
- 2. 硬件设施层面
- 3. 软件设施层面

国外金融机构广泛采用先进的高频量化交易软件^[32]，如 InfluxDB、KDB、Prometheus 等高性能时序数据库，这些软件由于先发优势目前具有庞大的客户群体和高度的稳定性^[33]。然而近年来，国内金融机构在软件设施方面有后来者居上的趋势。例如 19 年才出现的国产数据库软件如 DolphinDB，在功能和稳定性方面展现出超越 KDB 的后发优势。作为一款多模型数据库，主要专注于高性能的时序数据处理和分析。根据 db-engines 的时间序列（Time Series）DBMS 数据库（表 1）2025 年 3 月的排名中 DolphinDB 在多模型数据库中排名第 8 位。

表 1：全球前十排名时间序列 DBMS 数据库表

Rank		DBMS	Database Model	Final Score
Mar2025	Mar2024			Mar2025
1.	1.	InfluxDB	Time Series, Multi-model	21.50
2.	↑3.	Kdb	Multi-model	7.10
3.	↓2.	Prometheus	Time Series	6.38
4.	↑5.	Graphite	Time Series	4.57
5.	↓4.	TimescaleDB	Time Series, Multi-model	3.48
6.	↑8.	QuestDB	Time Series, Multi-model	3.10

7.	7.	Apache ruid	Multi-model	2.79
8.	↓6.	DolphinDB	Multi-model	2.29
9.	↑11.	GridDB	Time Series, Multi-model	1.98
10.	↓9.	TDengine	Time Series, Multi-model	1.76

如下（图 1）是不同数据库管理系统（DBMS）的排名和评分情况，其中红色箭头所指就是 DolphinDB 数据库软件排名，主要关注的是时序数据库和多模型数据库。部分高性能数据库的排名和评分情况。目前已被国内 TOP10 券商、百亿私募数十家以及公募基金多家采用，并形成了初始的社区生态。

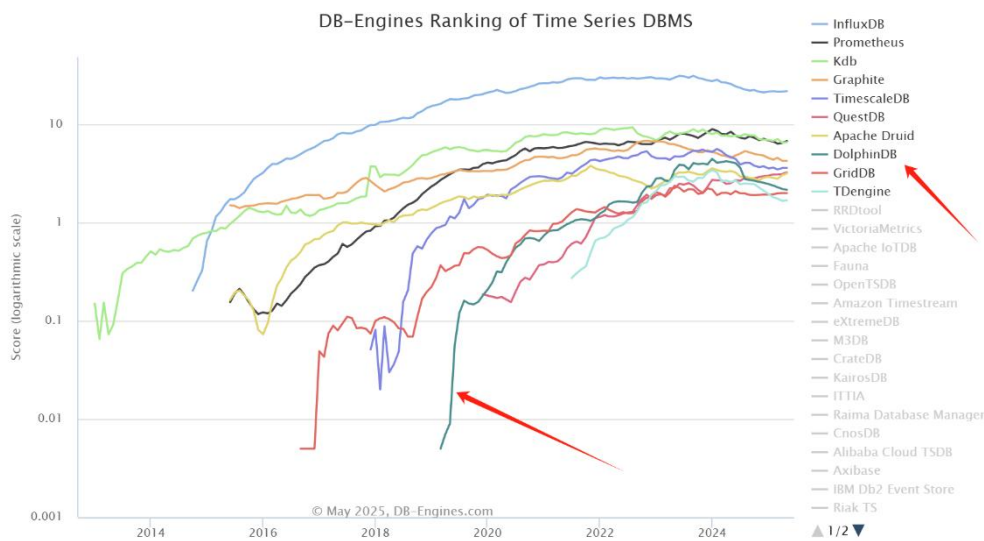


图 1：前十时间序列引擎排名折线图（Dolphin DB 为红色箭头所指）

（四）常见高频因子理论模型方向

二、金融市场流动性与买卖力度评估模型

（一）模型因子解释

1. 加权平均价格（WAP）

WAP（加权平均价格）是一种金融市场中常用的公式，可以通过买卖订单的数量和价格指标，得到当前市场的综合价格。主要应用于高频交易，用来评估市场流动性、价格均衡点以及交易效率。所以 WAP 根据实时数据反映当前市场的交易状态公式如下：

$$WAP_{L1} = \frac{BidPrice_0 * OfferOrderQty_0 + OfferPrice_0 * BidOrderQty_0}{BidOrderQty_0 + OfferOrderQty_0} \quad (1)$$

而将 WAP_{L1} 数据扩充到 WAP_{L5} 数据使用前 5 档买盘和前 5 档卖盘(总共 10 对量价

的数据)的数据进行加权平均，能够更全面地反映订单簿在一定深度内的整体价格均衡，需要将原式更进公式(2)如下：

$$WAP_{L5} = \frac{\sum_{j=0}^4 (BidPrice_j \cdot BidOrderQty_j) + \sum_{j=0}^4 (OfferPrice_j \cdot OfferOrderQty_j)}{\sum_{j=0}^4 BidOrderQty_j + \sum_{j=0}^4 OfferOrderQty_j} \quad (2)$$

2. 深度不平衡公式 (DI_j)

DI_j 是一种衡量市场买卖力量的指标，通过比较买单 (BidOrderQty) 和卖单 (OfferOrderQty) 的数量来计算。主要用于评估市场流动性和买卖力量的平衡情况。当 DI_j 为正时买方力量强市场可能上涨。相反的时候，表明卖单多于买单，卖方力量强可能导致市场的下跌公式如下：

$$DI_j = \frac{BidOrderQty_j - OfferOrderQty_j}{BidOrderQty_j + OfferOrderQty_j}, j = 1, \dots, 5 \quad (3)$$

公式计算出的 DI_j 值位于区间-1 到+1 之间，通过计算不同档位(j=0, 1, ..., 4)的 DI_j，我们可以分析买卖压力在订单簿深度上的分布特征，这为预测价格的短期波动提供了有价值的补充信息。DI_j 常被视为一个领先指标，因为它反映的是尚未成交的委托意愿，可能在价格实际变动之前发出信号。

3. 买卖权重公式 (w_i)

这是一个用于衡量不同价格层级对加权平均价格 (WAP) 影响的重要性指标。它通过当前价格与之前计算的 WAP 因子的差异计算权重。由于里面有之前的公式，所以我们可以称之为 2 阶因子。其含义是当差异越大，则权重越小。该公式广泛应用于高频交易，用于评估价格层级对市场的影响公式如下：

$$w_i = \frac{WAP \div (Price_i - WAP_{L5})}{\sum_{j=1}^5 WAP \div (Price_j - WAP_{L5})} \quad (4)$$

这个公式尝试将权重设定为与价格 Price_i 和 WAP_{L5} 之间差异的倒数相关，并通过分母进行归一化，使得所有档位的权重之和为 1。理论上，当 Price_i-WAP_{L5} 越大时，权重 w_i 越小则差异越大，也就是权重越小的情况。Price_i 代表第 i 档的价格，具体是买价还是卖价取决于计算买压还是卖压的权重。这种基于价格距离的权重计算方法广泛应用于评估不同深度上流动性或压力的重要性。

4. 买卖压力及买卖压力差 (press)

通过计算买压 BidPress 和卖压 AskPress 两个基础的 Press 指标的时候，可以很容易的理解到。当买压 BidPress 大于卖压时，可能预示市场上涨趋势因为买方的量和价格会

更高。反过来卖价和买量组合的 AskPress 也是如此。但是我没要注意的一个点是，卖价一定是低于买价的，这是 Press 指标的先天缺陷。不过也能清晰的算出买方压力和卖方压力的对数差异来衡量市场的整体压力公式如下：

$$\text{BidPress} = \sum_{j=1}^5 \text{BidOrderQty}_j \cdot w_j \quad (5)$$

$$\text{AskPress} = \sum_{j=1}^5 \text{OfferOrderQty}_j \cdot w_j \quad (6)$$

通过比较 BidPress 和 AskPress 的大小，我们可以初步判断市场的整体趋势。当 BidPress 大于 AskPress 时，表明买方在加权后表现出更强的力量，可能预示市场有上涨的倾向；反之，当 AskPress 大于 BidPress 时，则可能预示市场面临下跌压力。为了更清晰、对称地衡量买卖双方压力之间的相对差异，我们通常计算它们的对数比值，称之为 Press 指标，公式(7)表达形式：

$$\text{Press} = \log(\text{BidPress}) - \log(\text{AskPress}) \quad (7)$$

5. 模型因子整 8

将（4）式带入（5）得到（8）式，同理得到 AskPress 的（9）式。

$$\text{BidPress} = \quad (8)$$

$$\text{AskPress} = \quad (9)$$

对于（7）式进行整理后得到（10）式。

$$\text{Press} = \log\left(\frac{\text{BidPress}}{\text{AskPress}}\right) \quad (10)$$

此时将（8）（9）两式代入（10）式得到（11）式结果。

$$\text{Press} = \quad (11)$$

最后将（3）式并入得到最终计算模型 Press_{DI}（12）式。

$$\text{Press}_{DI} = \quad (12)$$

（二）基于金融市场逻辑合理化修正模型

1. 模型初分析

- （1）分母为零风险
- （2）数值敏感性
- （3）非对称性

(4) 结果范围不可控

2. 对于隐性问题改进

(1) 分母为 0 运行错误的隐性问题

$$\text{线性买卖压力权重: } \text{Price}_j - \text{WAP}_{L5} + \epsilon, j = 1, \dots, 5 \quad (13)$$

$$\text{线性绝对买卖压力权重: } |\text{Price}_j - \text{WAP}_{L5}| + \epsilon, j = 1, \dots, 5 \quad (14)$$

(2) 数值敏感性隐性问题

$$\text{非线性绝对买卖压力权重: } \exp(-\lambda|\text{Price}_j - \text{WAP}| + \epsilon), j = 1, \dots, 5 \quad (15)$$

(3) 非对称性问题改进

$$\text{非线性绝对买方压力权重: } \exp(-\lambda|\max(\text{Price}_j, \text{WAP}_{L5}) - \text{WAP}_{L5}| + \epsilon) \quad (16)$$

$$\text{非线性绝对卖方压力权重: } \exp(-\lambda|\min(\text{Price}_j, \text{WAP}_{L5}) - \text{WAP}_{L5}| + \epsilon) \quad (17)$$

(4) 输出的压缩输出范围改进

$$\tanh(X) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (19)$$

最终公式为:

$$\text{press}_{DI_{\tanh}} = \quad (20)$$

(三) 修正模型解释与模型意义

三、数值模拟操作与模拟结果展示

(一) 数据来源与高频计算数据库介绍

本文的数据来源为上海证券交易所旗下的 SimNow 所广播的一档 Tick 数据。高频数据交易软件, 本文章使用的是国内的高性能数据库 Dolphin DB 进行计算。

（二）应用过程概览

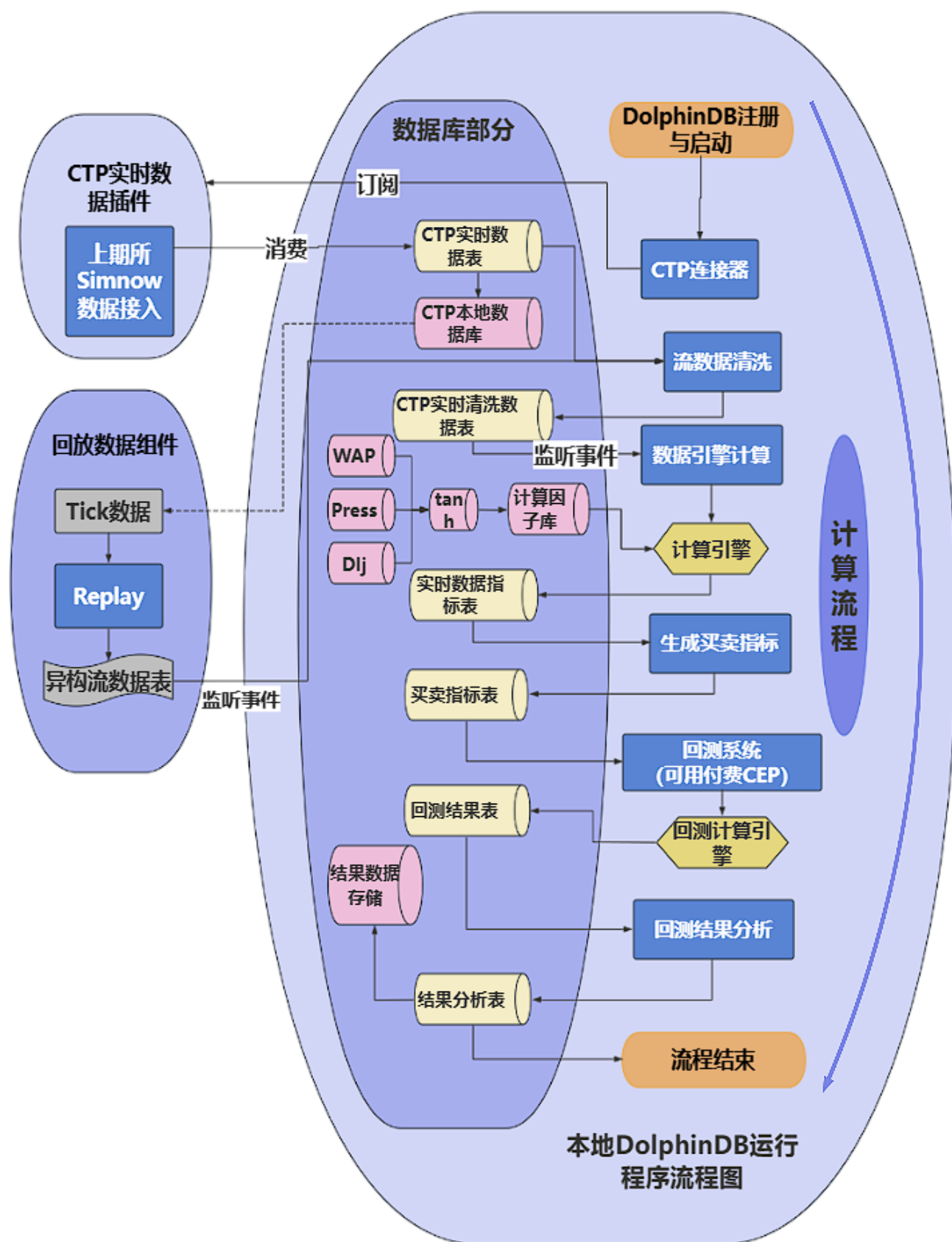


图 2： 本地 Dolphin DB 运行程序流程图

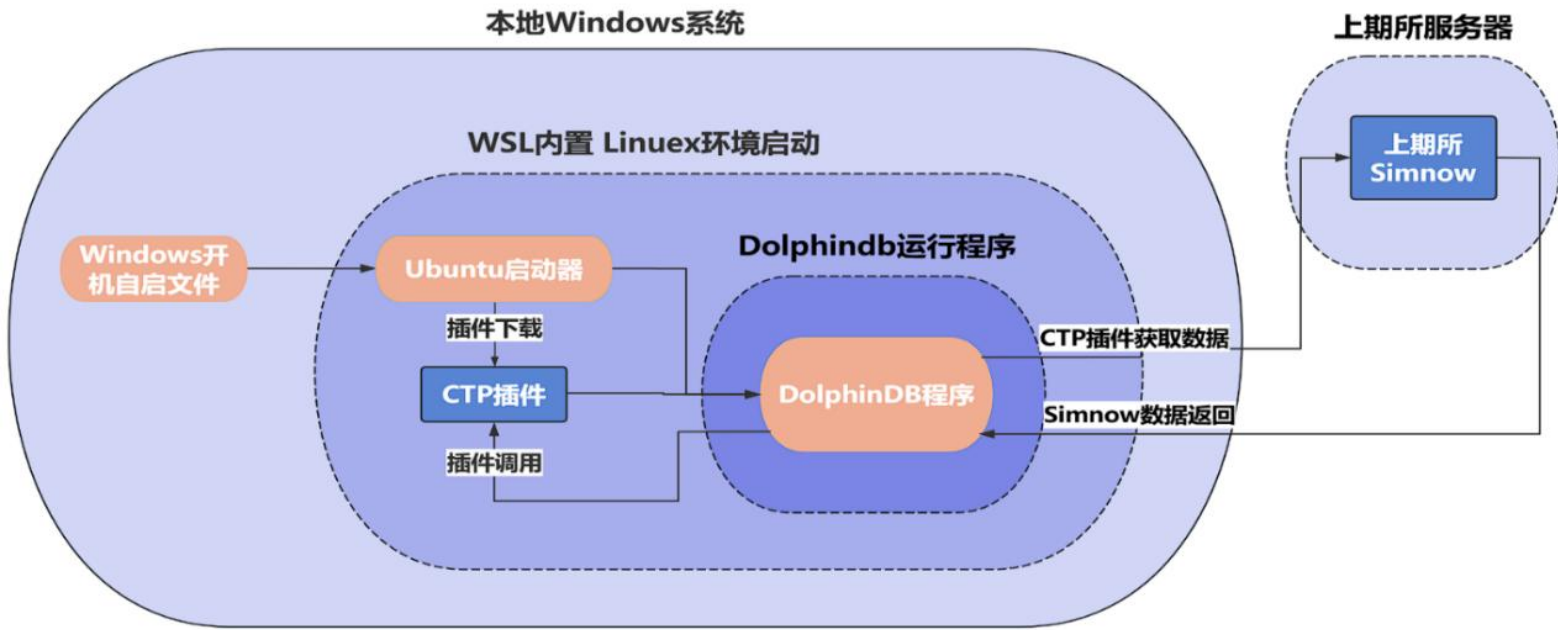


图 3： Dolphin DB 运行环境一览

（三）应用过程

首先使用 CTP 接口连接上期所获得大实时的 tick 数据^[46]，之后现将接受到的数据进行清洗，因为接收全部的数据有期货和期货的混合代码（图 2）。

图 4： 部分在交易期货期权代码获取展示

对数据进行清洗中，由于期权价格并不在本文的讨论范围内，以及某些期货返回数据过少，固本文仅对获得前百返回量的 tick 数据，进行期货流表化实时数据计算与分析如：（图 4、图 5）。

图 5： 筛选后全部期货代码获取展示

之后获取 tick 数据后进行指标计算根据上文的公式 10、公式 16、公式 17 的三种方式进行运算（图 6）和绘制。

	TradingDay	ExchangeID	LastPrice	PreSettlementPrice	PreClosePrice	PreOpenInterest	OpenPrice	HighestPrice	LowestPrice
0	2025.03.17		27	40.5	22	713	22.5	27	22.5
1	2025.03.17		18.5	24.5	12.5	1,805	12.5	20	12
2	2025.03.17		46.480000000000004	48.480000000000004	45.44	108	46.480000000000004	46.480000000000004	46.480000000000004
3	2025.03.17		4,504	4,990	4,570	136	4,466	4,608	4,314
4	2025.03.17		11,906	11,748	11,524	22	12,034	12,034	11,906
5	2025.03.17		3,654	4,702	4,100	357	4,434	4,434	3,580
6	2025.03.17		2,478	3,074	2,754	327	2,716	2,856	2,428
7	2025.03.17		3,384	3,996	3,394	194	3,500	3,592	3,262
8	2025.03.17		270	658	324	718	304	306	248
9	2025.03.17		11,654	12,642	11,450	18	12,240	12,240	11,654

1662394 行 49 列 (633.1 MB) 的表格 ctpMarketDataStream < 1 2 3 4 5 ... 166240 > 10 条/页 跳至 页

图 6： 部分筛选后的期货 tick 数据获取展示（部分）

图 7： 前百种期货的部分实时计算压力值返回（部分）

下图（图 8）是当时段的价格变化。根据基于一天的时间的前半天获得到的价格。

图 8： 当时的一个期货价格变化

通过对于公式 11、公式 17、公式 19 的压力差指标计算。得到如下（图 9、图 10、图 11）的计算结果

图 9： 公式 11 买卖压力差计算结果

图 10： 公式 17 买卖压力差计算结果

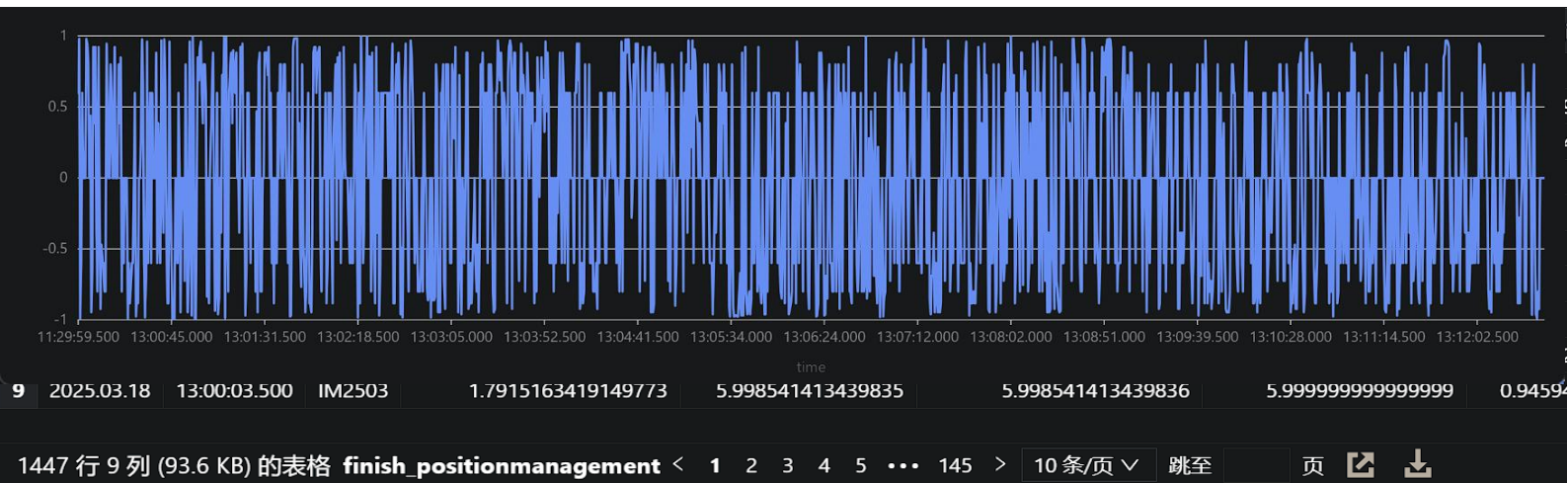


图 11： 公式 19 买卖压力差计算结果（部分）

对公式（19）的指标（图 11）进行分析，指标在 $[-1, +1]$ 区间徘徊，所以以 $+0.5$ 和 -0.5 为界，模拟实际操作数值划分为大于 $+0.5$ ，小于 $+0.5$ 到 -0.5 ，小于 -0.5 的三个区间分别对应 $+1$ 、 0 、 -1 这三个值。对应操作：多头状态、空仓状态、空头状态。并便于接下来数据相关性的理解

（四）盈利原理解析与展示

（五）回测方式与配置

每只入选期货分配资金为：10000000 并进行回测，对于能收集到的高频 97 只期货数据进行全部的计算。其中我们以期货并展示出来，如下表（2）所示。根据图（18）br2505 的回测曲线的结果可以看出，收益从开盘之后上涨趋势明显，说明公式（19）所对应的组合因子收益效果明显。

表 2： br2505 与 ag2509 一日回测的部分指标

期货代码	FG505	ag2509
一日总收益率	0.11201	-0.20267
一日年化收益率	1.74E+14	0
一日最大回撤	0.02945083	0.21616996
夏普比率	0.088202294	0.002027504

FG505 当日总收益达 11.2%，如图（15）。年化收益率更高但是由于回测时间过短，导致数据可能会极度失真。最大回撤也是到达了 2.945%，和回测中的数据可以看得到是可以得到印证的。进而说明按照此方法进行交易是风险极高但是收益极大的策略。但是我们发现夏普比例出奇的低，是因为只有一日的的数据量所计算年化后的波动率并不准确，所以夏普并不能反映出对应的因子状态。

此外我们同时观察到 ag2509 的期货的收益率为-0.2026790（年化到手收益为 0 也是由于时间过短导致的数据失真），与图（16）中的收益曲线一致。说明因子在不同期货的表现不同。

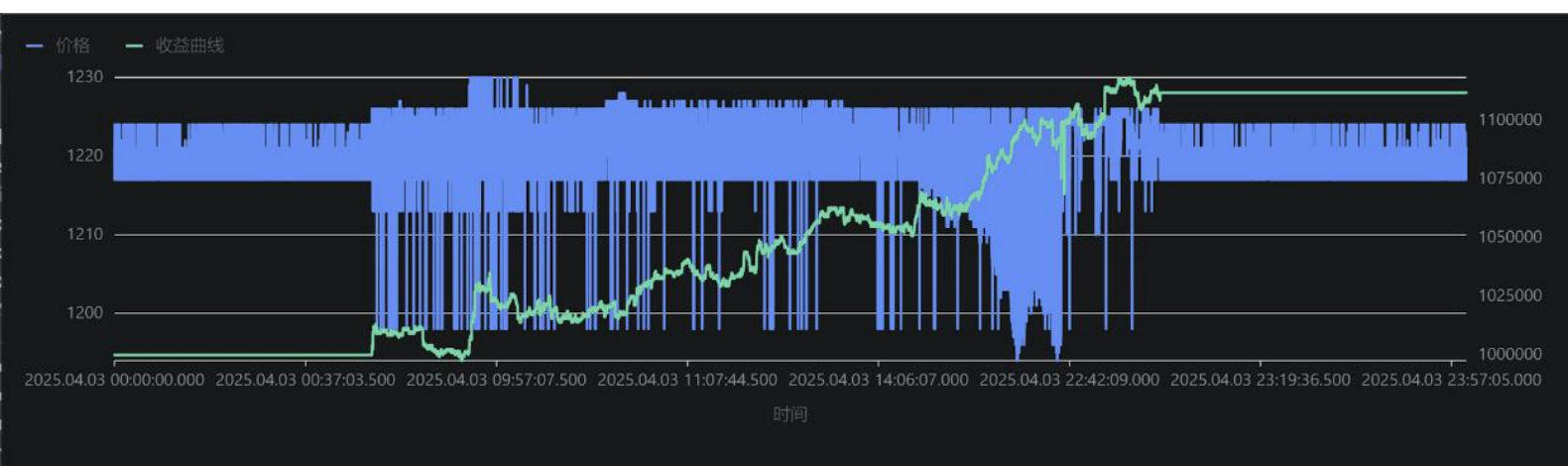


图 15: FG505 期货一日个交易段的回测曲线（上涨案例）

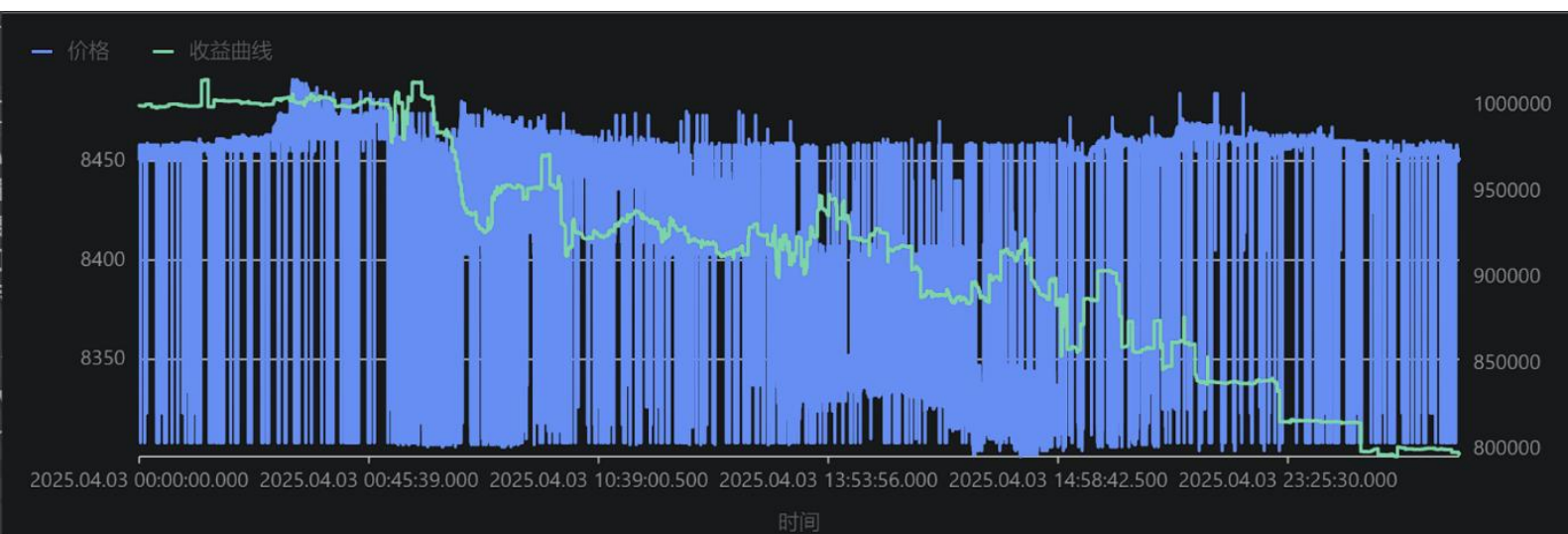


图 16: ag2509 期货一日个交易段的回测曲线（下跌案例）

（六）回测结果分析

对所有计算回测的期货进行正序排名，其中包括：收益率，年化收益率，最大回撤，夏普等指标。图（17）（表（4）是图（20）期货名称的展示）的收益率面积可以很轻易的看出正收益的面积大于负收益。所以整体如果使用均等的资金分配，是有盈利能力的。但是图（18）所示的夏普指标普遍并不高所以还需要长时间的回测数据积累。

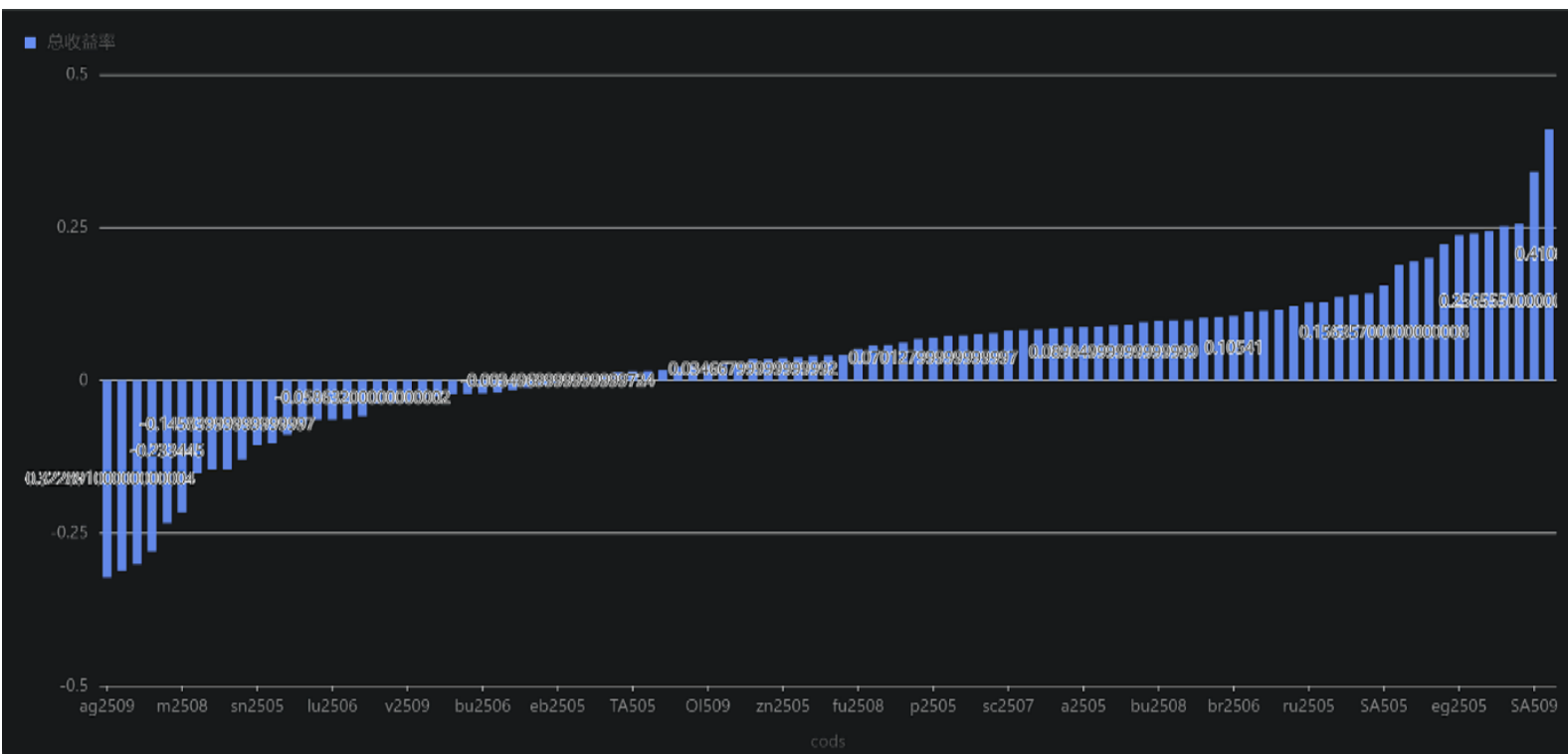


图 17：Python 展示的 97 只期货的收益率正序排名

夏普指标而言很明显有高有低，那么我们可以在之后的改进中减少低夏普的资金分配。用于减少对应的风险。但是像能出现 0.01 以下以及 0.3 以上的极值可能说明了数据量过小导致回测的结果并不正确，还需要长时间的数据获取与计算和回测。

表 4：图 20 期货正收益横轴期货顺序表

1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
ag2509	sn2505	v2509	eb2505	OI509	fu2508	sc2507	bu2508	ru2505	eg2505
ru2507	au2508	ni2508	ag2512	ni2505	SR505	cu2512	m2505	sn2508	i2505
br2507	m2511	MA505	hc2511	au2510	au2512	hc2507	rb2505	jm2505	ag2510
br2512	ao2506	sn2506	b2505	fu2511	sp2507	sc2506	fu2507	SH505	br2508
br2509	cu2508	p2509	ag2505	v2505	ao2505	ao2507	cu2601	sc2505	ru2509

m2508	lu2506	bu2506	TA505	zn2505	p2505	a2505	br2506	SA505	SA509
ru2601	sn2504	sn2507	sc2508	pp2505	cu2505	y2509	FG505	FG509	br2505
bu2507	c2507	lu2507	ao2508	au2505	m2509	y2505	OI505	cu2507	
ni2507	ni2506	fu2510	hc2509	hc2508	sp2511	ag2506	ru2511	fu2505	
lu2508	y2507	au2506	sp2505	hc2510	rb2510	c2505	ag2508	m2601	

四、结论

（一）结论

本文主要探讨了高频量化交易领域中的关键技术和理论。提出了一种基于大数据流式引擎的量化交易策略，目的在于在提高交易的实时性、准确性和因子有效性。强调了高频交易在现代金融市场中的重要性，以及实时数据处理和分析能力在高频交易中的关键作用。提出了基于流式引擎和 Tick 数据的量化交易策略，和与之相关的主流量化公司因子库的现状。

之后对于当前研究现状做了多方面的分析，包括了国内外在高频交易硬件设施、软件设施和人才资源方面的差异。指出国内在硬件和人才方面与欧美发达国家存在差距，但在软件和硬件设施方面，如 DolphinDB 数据库，有后来居上的趋势。另外还说明了高频因子研究现状，综述了当前学界和业界对高频因子的研究情况。指出学界研究相对匮乏，而业界研究在因子构建、风险识别和组合优化等方面具有一定的实践价值。

并通过上述部分引出了文章的实际模型应用，详细介绍了使用的模型计算公式和应用场景。指出了原始模型存在的潜在失真问题与风险，并提出了改进方案，包括添加极小常数、引入非线性的指数衰减函数，以及进行买卖方非对称处理。对于公式的合理化有了提升也提升了可解释性。通过前后公式因子指标计算的对比，说明了改动的合理性。

最后使用 Dolphin DB 软件通过读入 CTP 期货实时数据，对数据清洗得到纯期货数据。然后连接引擎进行并行计算，得出买卖指标进行回测期货数据。

（二）问题与改进

虽说使用（19）式计算的买卖压力差可以及时的观察到价格变化并通过类似股票作 T 的操作。将价格下跌前夕卖出然后价格下跌后买入的操作赚取差价。但是精准度有待提高，特别是有部分时间还是会展示出与价格变化无关的操作知道意见。如图（14、

17) 就有多次的操作变为持有空仓。但是价格并未和预期一致可能会有亏损的产生。

对应的解决方法是，1.数据获取更加全面，其中本文使用的数据是一档数据可以通过消费获得 5 档数据并进行实时的计算，这样更丰富了计算数据会让压力变化的更加平滑。不会展示出极端的上下变化，并且获取时间也要加长这样同一期货的数据就能获得更多计算也会更加完善。2.公式更加完善，本文仅仅是把目前市面上部分主流的因子进行性的组合并通过数据公式使得可读性提高。但是可能原本的因子就已经失效或者效果变差，所以可以通过因子更改的方式进行准确度的提升。3.风控操作的增加。特别是当我算出来每个期货指标的盈利能力的时候，可以调整对应期货的风控指标减少不必要的损失。

总体而言，本文为高频量化交易领域的发展提供了新的视角和实践经验，在高频交易场景下的实时决策提供了可扩展的技术框架。并在最后对于目标期货进行指标的解释说明该组合指标的有效性还是和好的，也提出了 4 条对于改进的措施。

参考文献

代码附录

1. 硬件运行环境介绍：

CPU：AMD Ryzen 9 7940HS w/ Radeon 780M Graphics 4.00 GHz

运行内存：16.0 GB

显卡：英伟达 GeForce RTX 4060 Laptop GPU

网络：1.固定网络 1：500Mbit/s 2.固定网络 2：1000Mbit/s 3.流量网络：未知

（复杂网络获取环境）

制造商：ASUSTeK COMPUTER INC.

2. 软件运行环境介绍：

数据库版本：DolphinDB Linux64 V 3.00.2.3 JIT

CTP 插件版本：etp Linux-X86-3.00.2-3.00.2.7

电脑操作系统：Windows11 家庭中文版（版本更新时间：2025 年 2 月 16 日）

电脑操作子系统：Linux LPTP 5.15.167.4-microsoft-standard-WSL2 #1 SMP Tue Nov 5 00:21:55 UTC 2024 x86_64 x86_64 x86_64 GNU/Linux

linux 操作软件：Ubuntu V 2404.1.68.0

3. 持久化流表路径：

本文的配置：DolphinDB 的 linux 软件内 dolphindb.cfg 持久化路径配置：persistenceDir = /DolphinDB_3.00.2.3/server/streamPersistDir

4. DolphinDB 运行代码：

致 谢

回顾 23 年人生的开启阶段，首先感谢父母的养育之恩，其次是老师的教育之恩，同窗的共进共学之情，谐社会的包容之恩，国家的变强与发展。

之前的中国是一个人情的国家，乡里都是有家族的观念。比如说去农地干点日结的活。当有人去了觉得很好，那么就会一传十的将这等好消息传给同性的同族人。但是近代的改革开放西方社会的模式与传统中国的模式进行了碰撞，特别是资本的形式使得用人的方式发生了巨大的改变，因为资本不看亲情，资本只看能力。所以现在这个社会已经变成了能力社会，所以我特别能同情我父辈年代的人，作为承上启下的作用连接着亲情社会与能力社会，每年还得想着要为同家的人有什么帮助，但到头来还是为无能为力的样子。父辈操心太多了，现在想想理解了很多。

我也能体会到这个社会的变化，特别是我学的金融这个圈子。你多赚必然有人多陪，固然学历很重要，但唯有个人能力万变不离其宗。而能力的高低又和个人的品格息息相关，如果有幸被同窗看此处，希望也能进行勉励先做一个有格局的人。

我的品格可以说是很正的几个了。我清楚记得我的品格塑造是由于小时候去游泳。旁边有水槽子，我的母亲告诉我有脏东西别人会吐进去。但是还没太学会游泳就游一段扒着一段。后来累了干脆一直扒着了。扒着的时候有一个喷壶帽里面的管子被我摸到了，但是没管毕竟是垃圾。后来有两个比我大的小朋友游过来很客气的问我有没有看到喷壶里面的管子，我当时震惊两个比我大的孩子竟然能找我并且客气的询问一个小事情，并答应了帮他们去找，我是从头扒到尾又找了一遍，找到了。然后换给了他们两人，他们也是特别客气的感谢着我。可能助人为乐的种子就此种下了。并基于此，之后的 15 年多我都是坚定的想要帮助社会和国家的人，并铸就了我热衷奉献的品格。

说来教师，首先感谢池文涛老师对上海立信会计金融学院育人贡献，也感谢能有幸接触到池老师这位敢于付出精力在一线教育的企业工作者。其次感谢付一土老师的鼎力支持、没有他的支持我也不会再量化这个方向有坚定的学习信心。江建武老师、王一鸣老师、杨晓诺老师、宋振华老师、周小华老师等等。还要特别谢明的是 Dolphindb 公司、SAIFSA 上高金量化社团等。

其次还要感谢的同窗，名字太多就不一一列举了，希望我们共同进步为祖国和社会添砖加瓦。

致谢人：王灿霖

2025 年 3 月 30 日