DolphinDB

高频量化的投研与交易

汇报人：谢斐

目录

CONTENT

01

高频数据与常见的高频量化策略

high frequency data & common high frequency quantization strategy

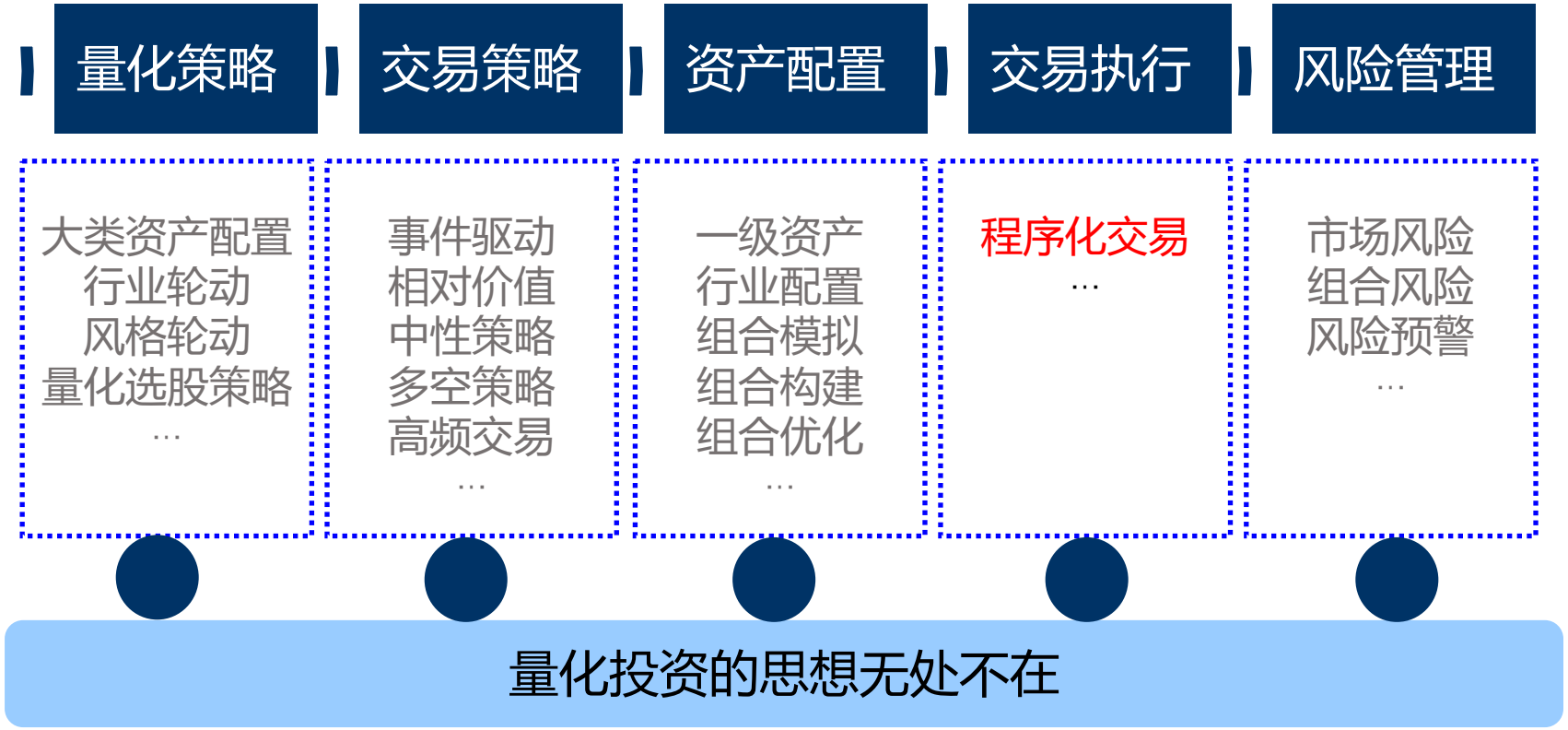
量化投资概念

简单来说，**量化投资**就是利用计算机科技并采用一定的数学、统计模型去实现投资理念、实现投资策略的过程。

量化投资主要依靠**数据和模型**来寻找**投资标的**和**投资策略**。

量化投资的思想

量化投资强调纪律性、系统性和大概率事件



高频量化交易

High frequency quantitative trading

高频交易是指通过预设的计算机算法实现高速度、高频次报单的交易方式，高频交易指令间隔通常小于五毫秒（甚至可达微秒、纳秒级），主要获利方式是通过市场短暂的价格波动进行套利交易。这种交易方式还具有**低隔夜持仓、高报撤单频率、高建仓平仓频率、高换手率**等特点。

高频交易以盘面分时图界面实时变动的客观数据来确定下买入单（或卖出单）的位置和数量，同时根据即时成交信息来决定下一步行动，如买入、卖出或撤单。据观察，目前国内的高频交易有一多半是T+0交易员通过手动下单完成的，还有一部分是通过量化策略模型生成的交易指令由计算机下单完成的。高频量化模型的每一项交易指令都不需要人工决策，将反应时间最小化，以抢占市场先机。

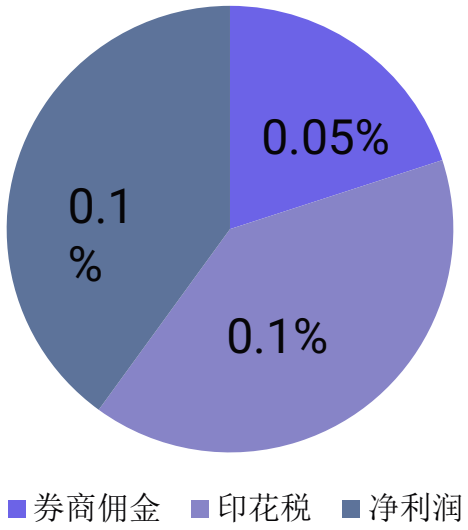
高频交易的特征

国际证监会组织（IOSCO）2011 年发布的《技术变革对市场的影响引发的监管问题》指出了高频交易的一些共同特征：

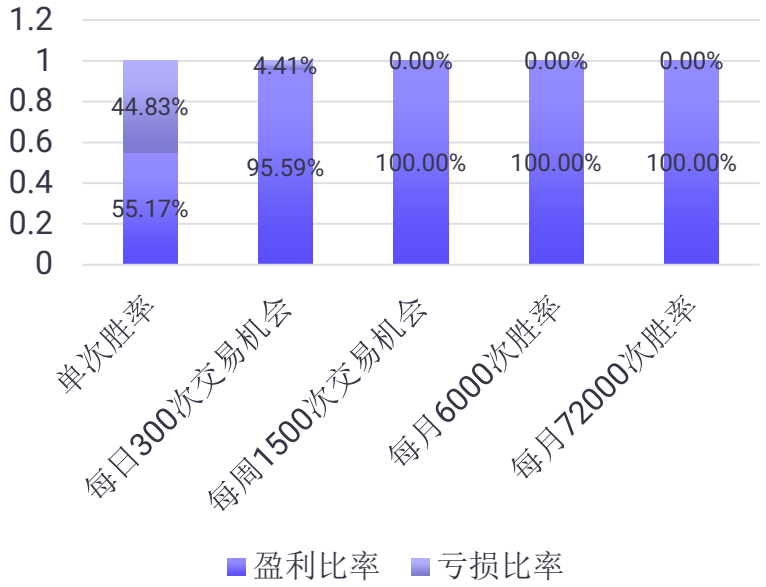
高频视角下的胜率

- 高频交易的单笔胜率
- 单日多笔订单的累计日胜率
- 高频交易下的回撤
- 积小胜而大胜

某策略交易利润分配



交易次数越多，胜率越高



以某策略为例：

按次计算，单次交易胜率55%

按日计算，300次交易总胜率大约95%

按周以上周期计算，日内T+0交易基本不会发生亏损

高频数据

- 高频数据概念
- 数据类型
  - 逐笔委托、逐笔成交、快照数据
- 订单的维度
  - 方向、金额、价格
  - 主动单、被动单
  - 成交、盘口

上交所 Level-2	上交所上市的股票，基金，债券，权证与上交所指数等品种	市场十档行情	2006-12	采样频率为每 3 秒一次的快照数据，提供市场 10 档价量委托行情、最新成交价、成交金额、成交量、成交笔数、市场状态、委买委卖总量、加权平均委买价格、加权平均委卖价格、净值（基金）、到期收益率（债券）、行权总量、涨跌停价（权证）等信息。	
		逐笔成交	2006-12	实时采样（非快照）的超高频数据，完整记录各证券每一笔成交详情成交时间（精确至毫秒）、成交通道、成交序号、成交价格、成交量等	
		指数行情	2006-12	采样频率为 3 秒一次的快照数据，记录指数的最新价、累计成交量、累计成交金额、最高价、最低价、昨收盘、今收盘等信息	
		集合竞价	2006-12	采样频率为 3 秒一次的快照数据，记录集合竞价阶段的虚拟开盘参考价、虚拟匹配量、买方虚拟未匹配量、卖方虚拟未匹配量等信息	
		委托队列	2006-12	采样频率为 3 秒一次的快照数据，记录买一和卖一价位上至多 50 位长度的委托明细。	
深交所 Level-2	深交所上市的股票、基金、债券、权证与上交所指数等品种	证券静态信息	2012-08	提供不同种类证券基本必须的证券静态信息内容，例如流通股数、每股净利润、行权价格、涨幅价格、跌幅价格等等。	由于数据库设计一些问题，造成不同内容信息的开始时间不一样，一共两个起始时间：2010 年 5 月以及 2012 年 8 月
		证券状态	2012-08	实时提供证券实时变动的动态状态信息，包括证券竞价状态、融资（券）买入（出）状态、回购、转股、创设、注销等信息	
		十档行情快照	2010-05	采样频率为每 3 秒一次的快照数据，提供市场 10 档价量委托行情、最新成交价、成交金额、成交量、成交笔数、市场状态、委买委卖总量、加权平均委买价格、加权平均委卖价格、净值（基金）、到期收益率（债券）、行权总量、涨跌停价（权证）、市盈率等信息。	
		最优买卖委托队列	2010-05	采样频率为 3 秒一次的快照数据，记录买一和卖一价位上至多 50 笔的委托明细。	
		指数快照	2010-05	采样频率为 3 秒一次的快照数据，记录指数的最新价、累计成交量、累计成交金额、累计成交笔数、最高价、最低价、昨收盘、今收盘等信息	
		逐笔委托	2012-08	实时采样（非快照）的超高频数据，提供证券的逐笔委托明细数据，包括委托索引、委托价格、委托数量、委托类别委托代码等信息	
		逐笔成交	2010-05	实时采样（非快照）的超高频数据，完整记录各证券每一笔成交详情成交时间（2012-08 开始的数据精确至毫秒）、	

常见的高频量化策略

DolphinDB

分时趋势策略

T0策略

T1动量策略

高频做市策略

分时趋势策略

分时趋势策略是指利用股票日内波动中出现的趋势性运动机会，利用市场具有的动量效应和反转效应，在股价向上突破时买进，或在跌落悬崖之前卖出，以获取日内差价利润的交易策略。A股中的分时趋势策略同样需要之前持有该股票底仓作为完成交易的必备条件。当前A股许多个股日内波动足够大，投资者经常见到一天之内个股出现几波快速上涨或下跌。分时趋势策略包括了突破策略和反转策略，尽管突破和反转对许多交易者来说是不同的，但对于量化模型来说，本质上都是对于趋势的捕捉。**在分时趋势策略量化模型的交易策略中，有些是通过市场信息预判趋势而进行交易的，有些是追踪并跟随趋势下单强化趋势，有些甚至可以制造趋势，触发蝴蝶效应。**

分时趋势策略

突破策略

突破是市场敏感行为，利用突破对市场参与者心里和行为造成的冲击来获利，是分时趋势策略中比较高阶的方法。当前A股交易中有部分机构投资者会采用“按时间加权平均价格订单”（TWAP）对大额买卖单进行拆分，将大额订单在一定时间内分拆为多个等量的小单，以市价进行交易。TWAP订单在盘面分时成交明细中表现出有规律、同方向、连续性强的鲜明特征，很容易被监控市场的量化模型捕捉到。

反转策略

“接飞刀”策略

全市场监控价格的异常性突发变动

低延迟下单和迅速反手平仓

T0策略

T0策略俗称股票日内交易策略，基于对未来短期股价走势的判断，买多或者做空股票，并且在很短的时间里平仓，一般而言，操作周期通常在3-5分钟。优势在于短期预测的高胜率。

策略精髓：基于对未来短期股价走势的判断，买多或做空股票，并且在很短时间内获利平仓。

策略特点：虽然单笔交易利润很薄，但基于高质量的未来短期股价预测，可以实现在每日多次成交的前提下提高胜率。由于持仓时间短，策略本身基本不受宏观因素影响，在比较活跃的市场交易量的情况下，大概率可以做到“稳赚不亏”。

由于A股交易规则是不允许当日卖出股票，所以如果需要完成当日买卖一个周期，那么就要求策略本身持有该股票作为底仓。

常见的高频T0策略

事件驱动策略

通过刻画特定的价量形态和事件来触发交易，这些形态和事件外，策略对于未来的价格走势不做判断。

多因子策略

定义目标价格作为模型的预测对象→寻找有稳定预测能力的因子→线性或非线性拟合因子成最终的“预期收益率”→扣除费用和冲击，作为是否发生买卖的依据

混合型策略

把事件驱动和多因子的研究框架合并在一起，通过定义特定的场景或事件，降低预测未来收益的难度，提高准确度，可更有效的打败交易费用，更有效的降低同类型策略之间的相关性。

高频T0策略的局限

01

容量小

高频T0策略的容量很多时候取决于底仓是否充沛，加之因子在有限的盘口和流动性下，策略的容量也会被大大的限制。

02

策略同质性高

预测到的数据基本就只限于高频的价量数据和少部分实时的另类数据，不同机构的策略底层因子构成的相似度会很高。

03

交易规则限制

不能当日卖出股票、不能净做空股票，对于撤单率的限制，没有做市返佣，印花税偏高等。

T1动量策略

动量交易策略，即预先对股票收益和交易量设定过滤准则，当股票收益或股票收益和交易量同时满足过滤准则就买入或卖出股票的投资策略。

策略思想：当一段时间内买方与卖方力量差别足够大，且对手方新增订单相对够少，则未来继续有一段趋势的可能性非常大!

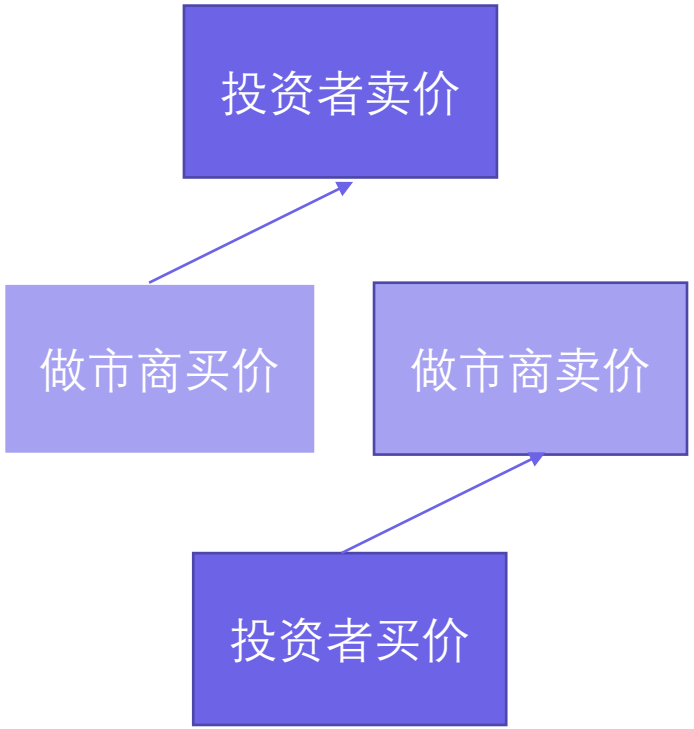
选取上月的活跃品种

选取上月交易策略合适的品种.

TI逐笔成交数据和快照数据的回测

高频做市策略

高频做市策略是指利用盘口买卖盘之间的价差，同时挂出买单和卖单，让其他急于成交的市场参与者与自己的买卖单分别成交，实现利润的策略。A股中的高频做市策略需要之前持有该股票底仓作为挂出卖单的先决条件。高频做市策略需要个股窄幅波动以及市场整体较为平静，在当前市场环境下，个股日内走势大部分时间都在中枢震荡中度过，这就为高频做市策略的入场创造了很好的条件。



高频做市策略

高频做市行为一方面**为市场提供流动性**，另一方面**通过价差收益实现盈利**。由于高频交易单笔获利极小，因此需要**交易量**来保证自己获利，这就会为市场带来大量流动性。

市场其他参与者要想高效的参与股票市场交易机会，就需要一个盘口价差小，成交量足够的市场，高频做市策略**缩小了盘口价差**，为市场其他参与者提供了对手盘，增加了盘面买单和卖单挂单数量，增强了市场整体流动性，在一定程度上降低了其它市场参与者买卖股票的冲击成本。

做市策略要点

合适的品种流动性要求

底仓存货风险控制

单边暴露的风险控制

与套利策略的结合

高频交易的技术基础

02

市场微观结构和算法交易研究

Research on market microstructure and algorithmic trading

程序化交易实盘与策略回测的差异

DolphinDB

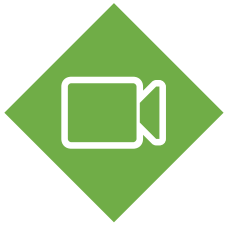
策略的执行细节（算法）

实盘的成交概率

实盘的市场容量

实盘行情、交易的延时

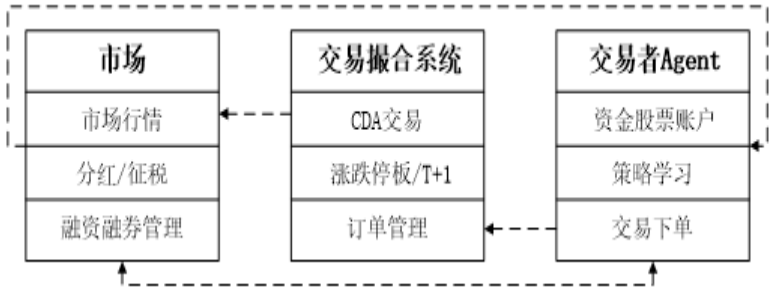
实盘的冲击成本

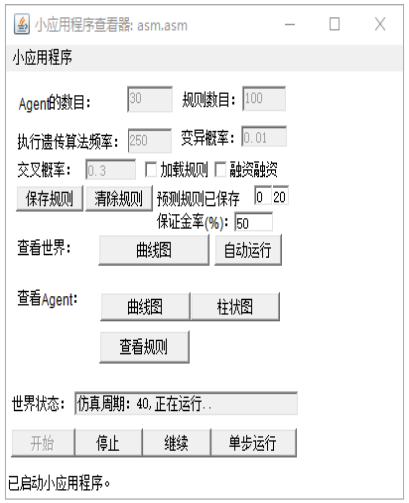


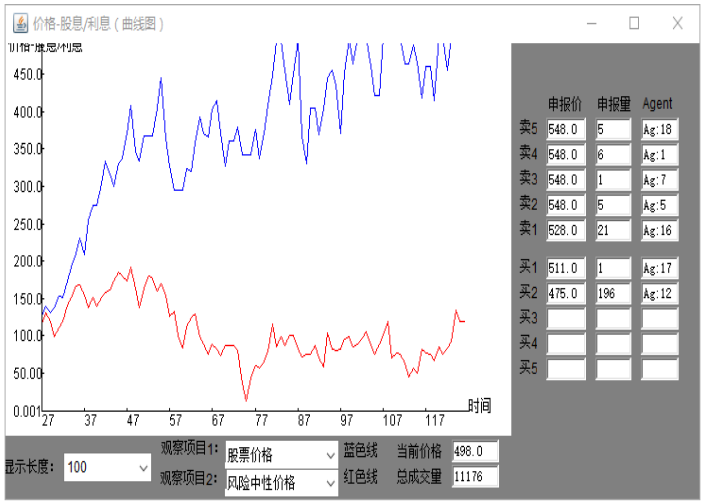
实盘的资金风险控制

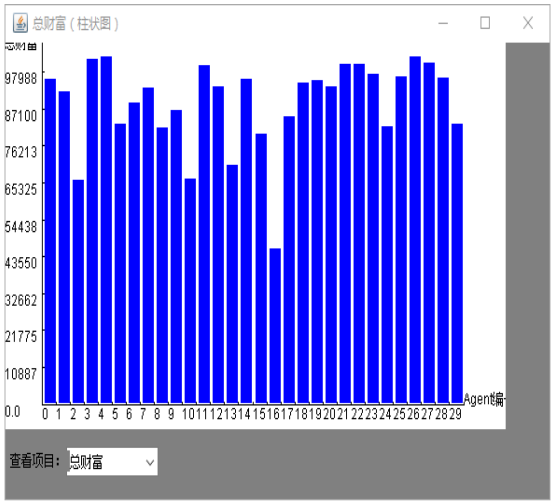
基于连续双向拍卖制的高频交易实验

- 实验模型设计



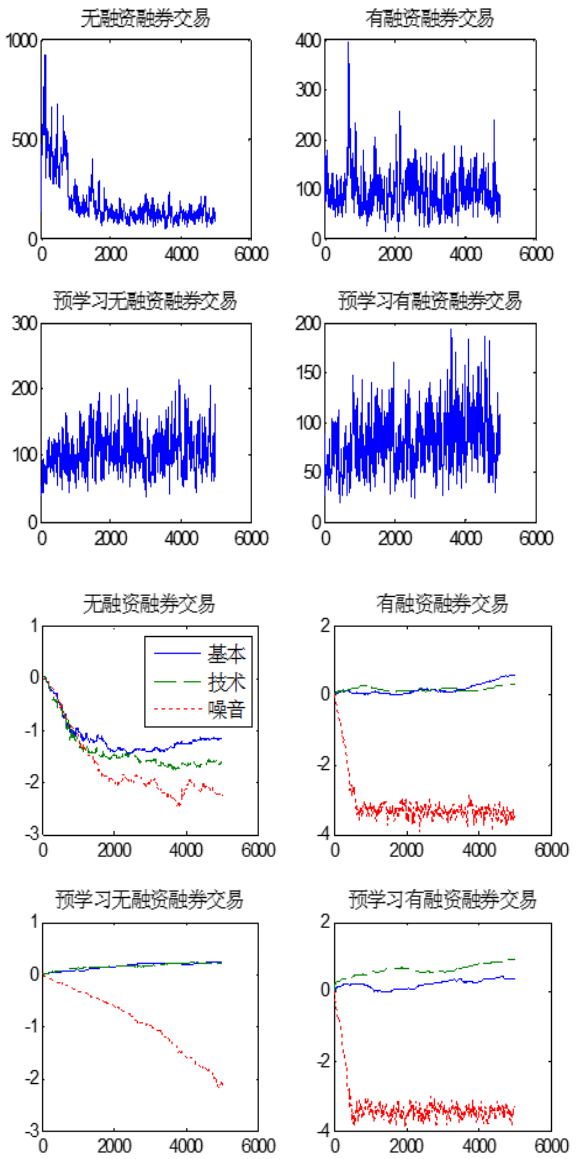






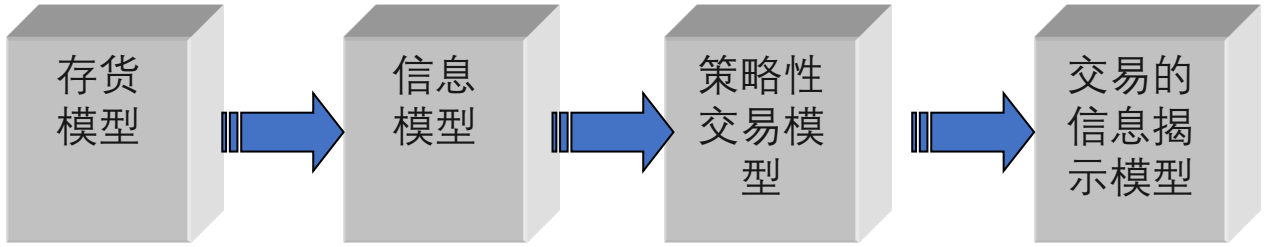
基于连续双向拍卖制的高频交易实验

- 实验结论
  - 日内回转交易（T+0交易）有助于稳定市场
  - 基本面投资可以在高频交易中获取超额收益
  - T+0制度非常有助于高频趋势技术交易
  - 交易指令细节会影响最后的收益



市场微观结构的研究

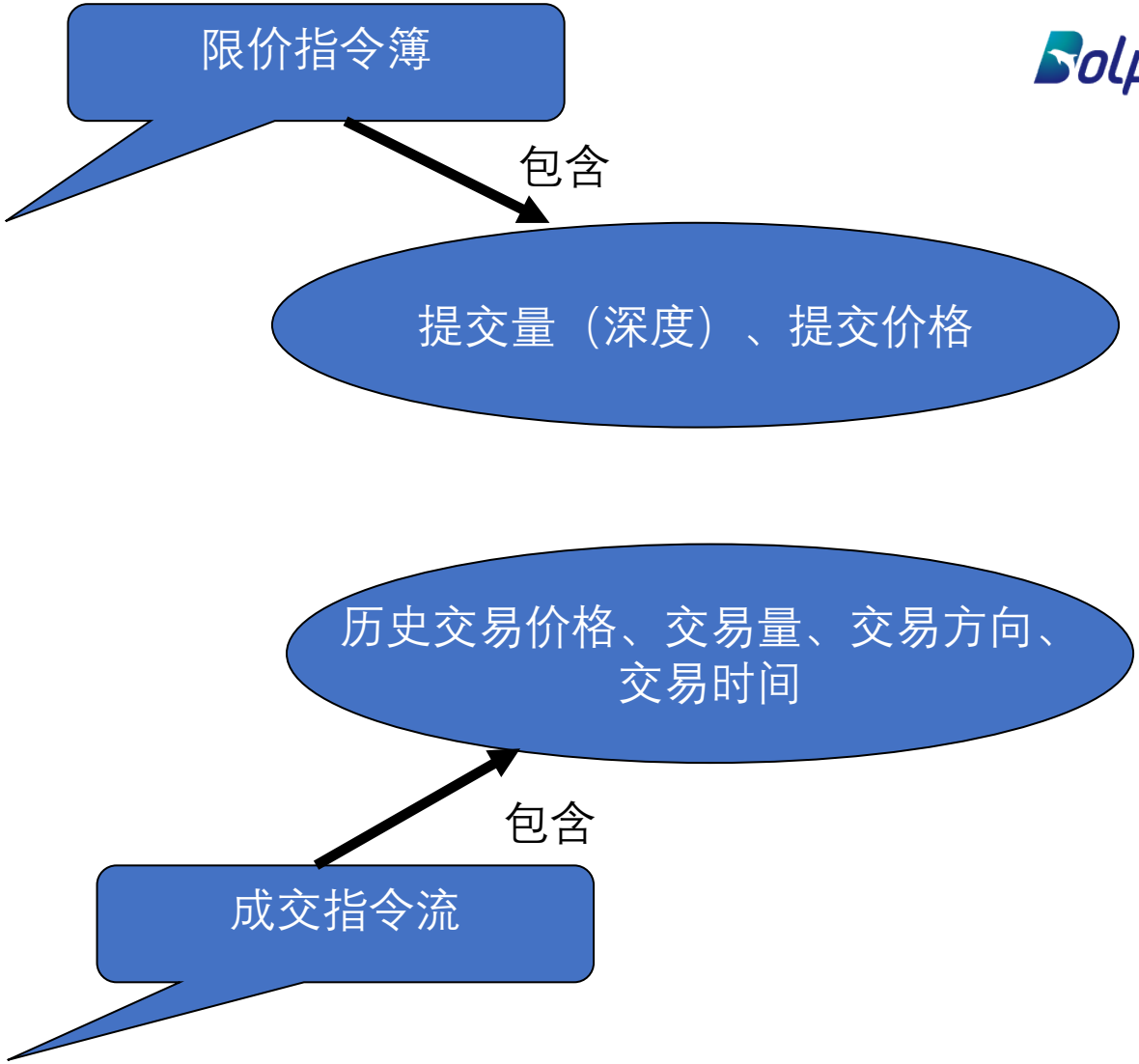
- 根据市场流动性提供者区分
  - 订单驱动市场（Order Driven Market）
    - 连续双向拍卖机制（Continuous Double Auction，简称CDA）
  - 做市商市场（Market Maker）
- 交易是传递信息的主要信号



指令驱动市场

卖⑤	31.69		150
卖④	31.68		15
卖③	31.65		1
卖②	31.60		1000
卖①	31.59		108
买①	31.58		12
买②	31.57		6
买③	31.56		14
买④	31.55		33
买⑤	31.51		144
现价	31.60	今开	32.70
涨跌	-1.13	最高	32.70
涨幅	-3.45%	最低	31.18
总量	27.3万	量比	0.87
外盘	15.0万	内盘	12.3万
市盈	35.2	股本	43.5亿
换手	0.8%	流通	32.6亿
净资	5.90	收益(一)	0.22

14:58	31.60	91 B
14:58	31.60	177 B
14:58	31.60	218 B
14:58	31.60	421 B
14:59	31.60	43 B
14:59	31.60	203 B
14:59	31.59	127 S
14:59	31.60	643 B
14:59	31.60	35 B
14:59	31.60	481 B
14:59	31.51	406 S
14:59	31.60	41 B
14:59	31.60	152 B
14:59	31.59	173 S
14:59	31.60	189 B
14:59	31.59	206 B
15:00	31.59	75 B
15:00	31.59	55



价格笼子（科创板、创业板）

- 在连续竞价交易中设置有效申报价格范围
- 科创板：
  - 在连续竞价阶段，买入价格有上限、无下限，不超买入基准价格的102%；卖出价格有下限、无上限，不低于卖出基准价格的98%。交易价格同时要符合相应的涨跌幅规定要求。
  - 基准价格的确定。监控细则中，科创板的有效申报价格范围，优先以即时揭示的买卖价格为基准，无即时揭示价格的依次以最新成交价、前收盘价为基准。
- 创业板：
  - 超过价格笼子（±2%）的报价将在主机里“缓存”，等即时成交价发生变化，导致这些报价进入有效范围时，这些单子会被自动激活，参与到交易中。

算法交易（Algorithmic Trading）

- 一般认为算法交易是自动化交易的一种子类别。
- **欧盟**于2018 年开始施行的《欧洲金融工具市场指令II》（Markets in Financial Instruments Directive II, MIFID II）对其进行了定义。
- MIFID II 指出，通过计算机算法自动决定金融工具交易订单某个或某些要素（例如是否发起订单、发起订单的时间、订单的价格、订单的成交量等要素），且较少或完全没有人为干预的金融工具交易即为算法交易（Algorithmic Trading）

常见的算法交易策略

01

第一代算法交易

- 成变量加权平均价格算法（VWAP）
- 时间加权平均价格算法（TWAP）
- 跟量算法（TVOL）

02

第二代算法交易

- 执行落差算法（IS）

03

第三代算法交易

- 眼镜蛇（Cobra）
- 夜鹰（Nighthawk）
- 游击战（Guerrilla）
- 狙击手（Sniper）
- 搜寻者（Sniffers）

市场微观结构、高频因子研究

对订单簿（OrderBook)和指令流（MessageBook）的研究

算法交易

流动性的估计

高频策略的回测

逐笔撮合

逐笔数据特殊性

无法按普通时间序列对齐，窗口聚合

03

高频回测数据库构建

High frequency database for backtesting

高频回测数据库的需求

DolphinDB

个性化需求	• 行情厂商标准数据无法提供自定义的Minbar、HourBar，连续合约及自定义合约等拼接规则，导致量化策略无法实现。
历史数据采集现状	• 历史数据数据量大，采集慢访问慢，一年数据可能要采集几天，还不能保证数据没有遗漏。
行情实时处理现状	• 没有高效的内存型数据库无法实现实时的因子计算。
自建系统困难	• 自建数据库采集工具成本高，高频处理系统开发难度高。

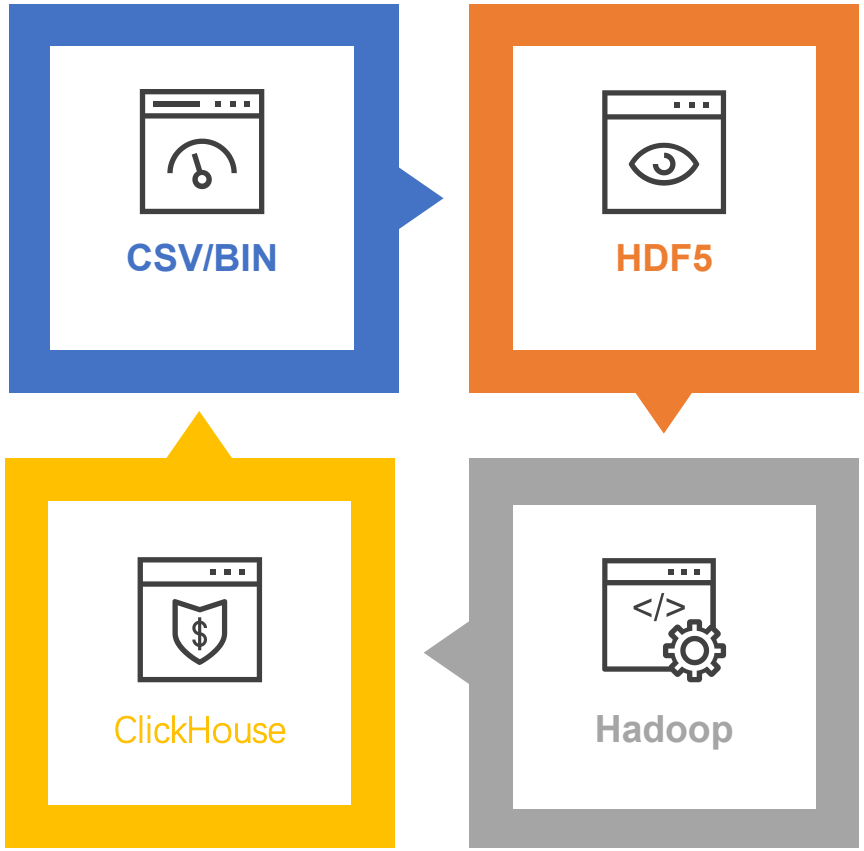
高频数据质量的保障

数据库选型的尝试

DolphinDB

文件式存储

压缩解压操作较为繁琐
数据修正困难



HDF5

存储和读取效率较低

ClickHouse

支持较少、部署维护困难

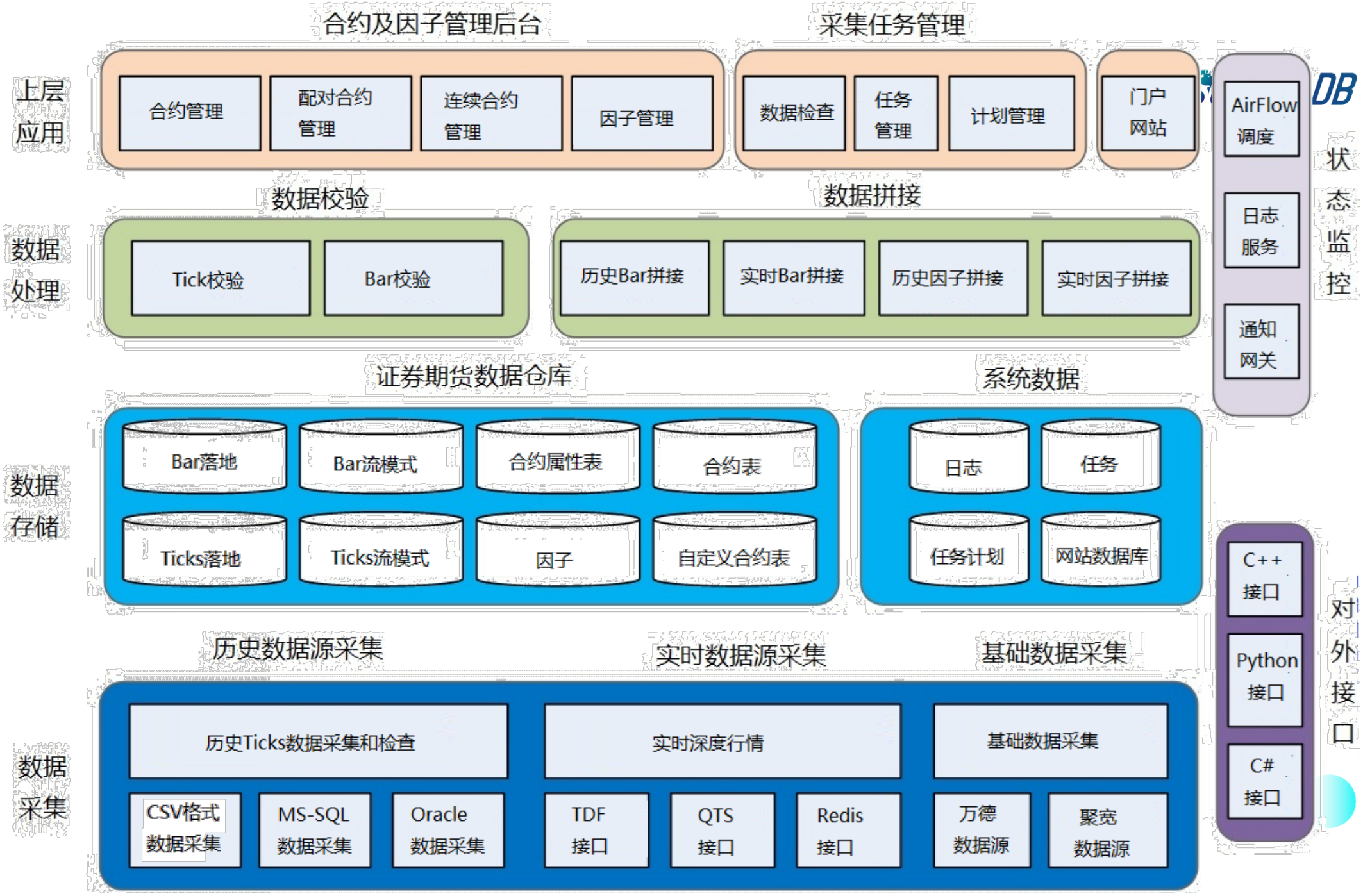
Hadoop

计算脚本编写具有一定门槛

基于DolphinDB的高频数据库设计

- 本系统相比传统数据采集方案，采用高性能可靠性的架构设计，实现高速自动采集证券/期货历史数据。
- 借助Airflow工作流平台全自动调度实现处理金融资产高频历史数据。
- 结合实时行情为多因子量化投资策略等工具提供高效完整的数据计算基础。
- 数据存储使用分布式高性能DolphinDB集群数据库。
- 采用ETL方案统一数据结构，保证了数据质量，前后依赖关系确保数据的一致性和稳定性。
- 扩展性强，支持Python、C++、C#、Java等接口对接第三方系统。

系统架构设计


合约及因子管理后台
采集任务管理
上层应用
合约管理
配对合约管理
连续合约管理
因子管理
数据检查
任务管理
计划管理
门户网站
AirFlow调度
DB
状态监控
日志服务
通知网关
数据校验
数据拼接
数据处理
Tick校验
Bar校验
历史Bar拼接
实时Bar拼接
历史因子拼接
实时因子拼接
证券期货数据仓库
系统数据
数据存储
Bar落地
Bar流模式
合约属性表
合约表
日志
任务
Ticks落地
Ticks流模式
因子
自定义合约表
任务计划
网站数据库
C++接口
对外接口
Python接口
C#接口
历史数据源采集
实时数据源采集
基础数据采集
数据采集
历史Ticks数据采集和检查
实时深度行情
基础数据采集
CSV格式数据采集
MS-SQL数据采集
Oracle数据采集
TDF接口
QTS接口
Redis接口
万德数据源
聚宽数据源

AirFlow调度及高频数据采集流程

1 基于AirFlow调度平台可以实现数据每日自动采集，实时监控采集状态，出错任务自动重新执行确保采集数据准确性

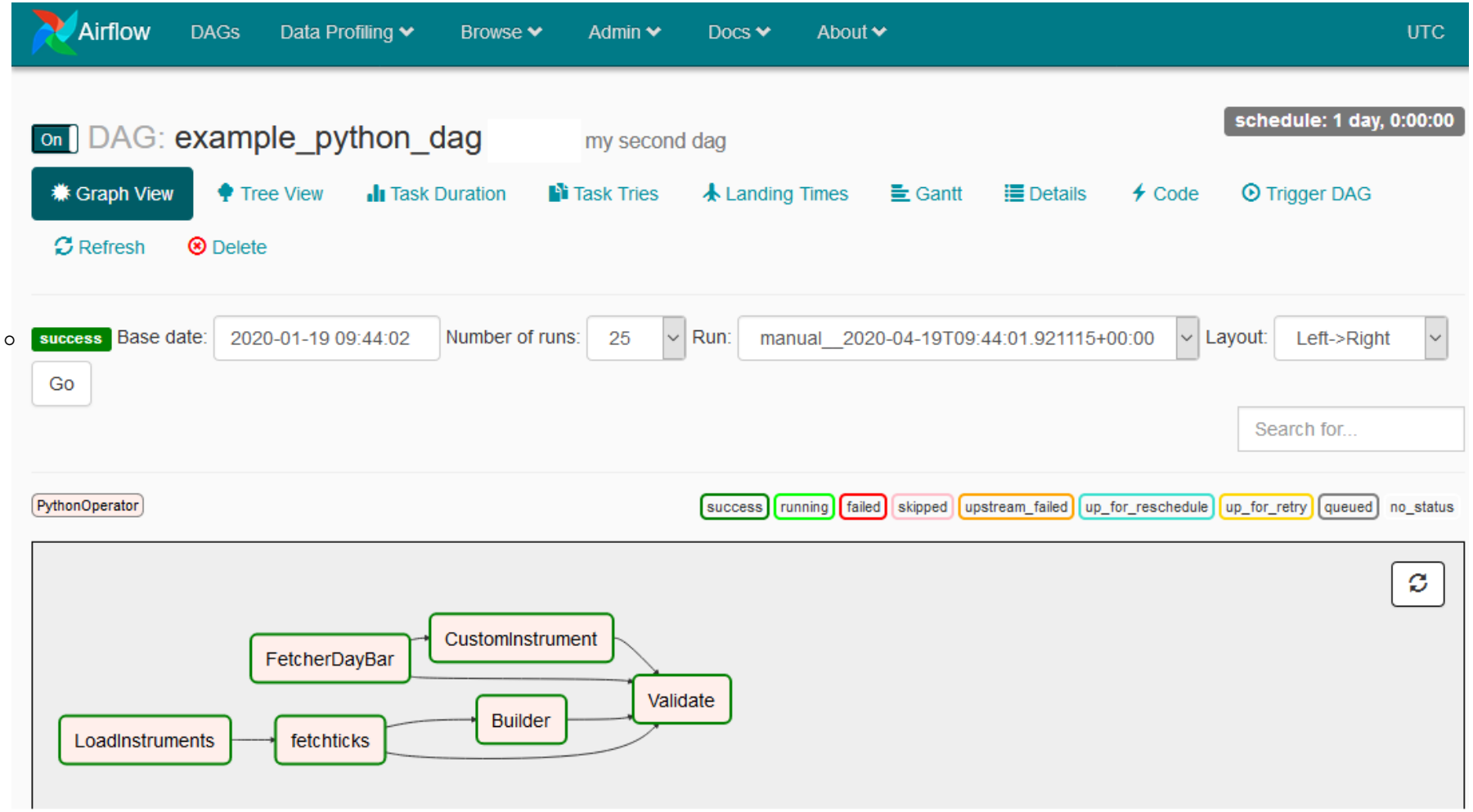
2 快照数据自动运行流程如下：

(1) 载入合约表—读取Ticks—校验。

(2) 载入合约表—读取Ticks—拼接生成Minbar，Hourbar—校验。

(3) 读取合约日线，自定义拼接—校验。

(4) 系统自动监控运行状态，出错后可自动或手动重做任务。



历史数据采集性能

写入性能：
某大型券商SHFE一周Tick数据（712万条）采集入库性能对比（同等硬件条件）。

对比项目	数据存储采用Mongodb（无法发现数据缺失，采用人工抽样或单独编写脚本检查）	本系统存储采用DolphinDB（含多线程采集TDB，并同时写TaskId、JobId到postgres数据库用于记录任务状态）
耗时	1000s	120s

读取性能：
读取深市某股票一年数据

数据类型	数据量	耗时	说明
快照数据	113W*45列	6.36s	113万条记录，每条记录45个字段
逐笔成交数据	1593W×11列	4.8s	每条记录11列字段

DolphinDB

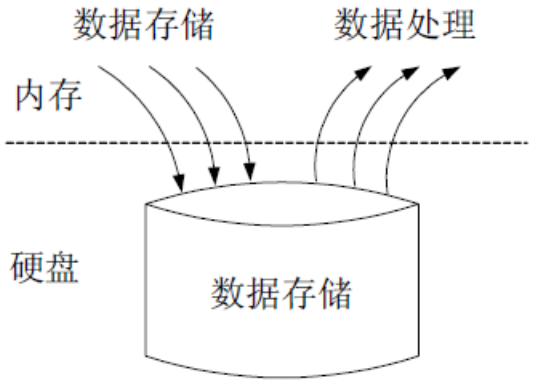
04

高频因子流式计算

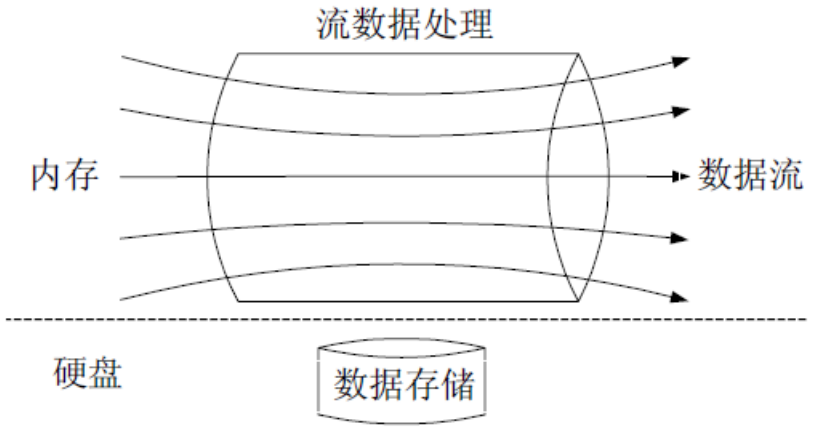
High frequency factor flow calculation

流式计算与批量计算

大数据的计算模式主要分为批量计算(batch computing)、流式计算(stream computing)、交互计算(interactive computing)、图计算(graph computing)等。其中，流式计算和批量计算是两种主要的大数据计算模式，分别适用于不同的大数据应用场景。



大数据批量计算



大数据流式计算

流式计算与批量计算

流式计算和批量计算分别适用于不同的大数据应用场景:对于**先存储后计算,实时性要求不高**,同时,数据的**准确性、全面性更为重要**的应用场景,**批量计算**模式更合适;对于无需先存储,可以**直接进行数据计算,实时性要求很严格**,但数据的**精确度要求稍微宽松**的应用场景,**流式计算**具有明显优势。

性能指标	大数据流式计算	大数据批量计算
计算方式	实时	批量
常驻空间	内存	硬盘
时效性	短	长
有序性	无	有
数据量	无限	有限
数据速率	突发	稳定
是否可重现	难	易
移动对象	数据移动	程序移动
数据精确度	较低	较高

流式计算在金融行业的应用

金融银行系统内部,每时每刻都有大量的往往是结构化的数据在各个系统间流动,并需要实时计算.同时,金融银行系统与其他系统也有着大量的数据流动,这些数据不仅有结构化数据,也会有半结构化和非结构化数据.通过对这些大数据的流式计算,发现隐含于其中的内在特征,可以帮助金融银行系统进行实时决策.

金融银行的实时监控场景中,大数据流式计算往往体现出了自身的优势.如:

- **风险管理**.包括信用卡诈骗、保险诈骗、证券交易诈骗、程序交易等,需要实时跟踪发现;
- **营销管理**.如,根据客户信用卡消费记录,掌握客户的消费习惯和偏好,预测客户未来的消费需求,并为其推荐个性化的金融产品和服务;
- **商业智能**.如,掌握金融银行系统内部各系统的实时数据,实现对全局状态的监控和优化,并提供决策支持.

利用DolphinDB流式计算构建计算服务器

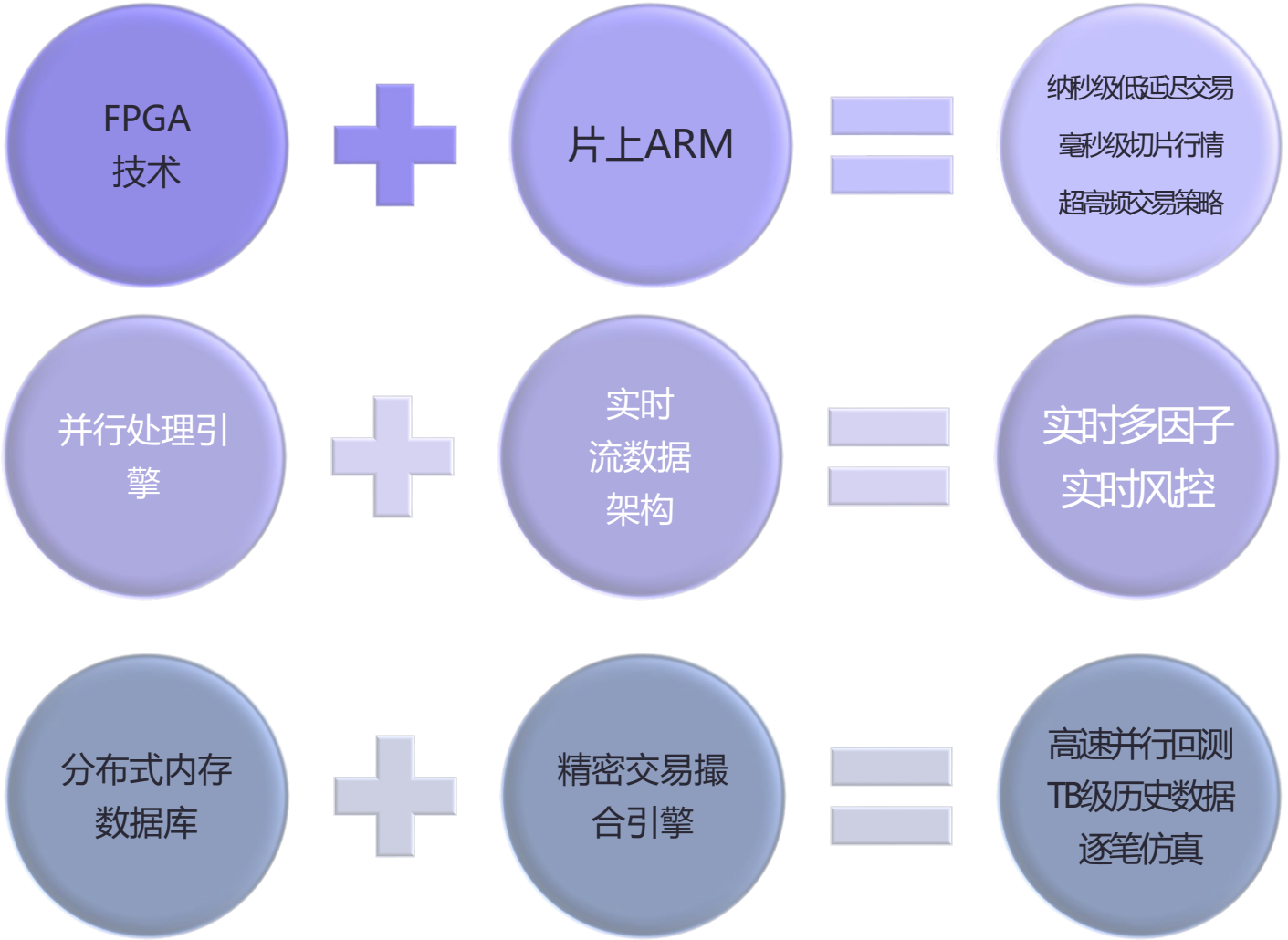
流式数据计算性能

操作项目	耗时	说明
获取行情	1~2毫秒	原始行情写入DolphinDB原始行情表
行情数据转换	1~3毫秒	对原始行情数据加工
计算和存储因子	1~6毫秒	调用因子计算函数，将结果写入因子表

1 借助底层DolphinDB内存表计算性能，提供低至10ms以内的因子计算性能，可按需订阅自己需要的多个合约和因子数据，并行支持上千个因子的实时计算

2 支持C# 或Python等语言订阅因子表：

高频量化交易的新趋势


FPGA
技术
片上ARM
纳秒级低延迟交易
毫秒级切片行情
超高频交易策略
并行处理引擎
实时
流数据
架构
实时多因子
实时风控
分布式内存
数据库
精密交易撮合引擎
高速并行回测
TB级历史数据
逐笔仿真

谢　谢