思考：
1. 数据预处理
2. 如何分配并行任务

第16周理论课+实验课：做pre

题目：股票市场的主力资金流向计算

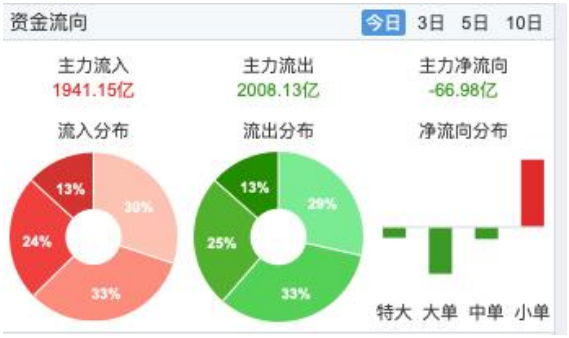
一个完整的大数据分析项目
输入输出都做出来了，所以就做一个并行的加速

一、题目背景

在股市中，主力资金意指集中了大量资金的交易者（机构），具有主导股票涨跌的能力。直观上，跟主力资金买卖，赢面最大。

但是，交易所并没有公布实时及盘后的个股资金流向数据，也没有公布逐笔订单的交易者 ID。因此，我们不能直接从交易所原始数据中获得主力资金信息。

如何识别主力资金？这是个很难的问题。目前的各股票行情软件提供了主力资金流向展示（如下面两图），虽然离准确识别主力资金仍有一定距离，但至少可以一定程度上刻画个股大资金流入流出情况。



序	代码	名称	最新	涨幅	主力净...	集合竞价	实时超大单数据				实时大单数据				实时中单数据				实时小单数据			
							流入	流出	净额	净占比	流入	流出	净额	净占比	流入	流出	净额	净占比	流入	流出	净额	净占比
6	510300	沪深300...	3.906	-0.64%	+4.13亿	1.36亿	45.3亿	37.6亿	+7.66亿	+9.06%	17.8亿	21.4亿	-3.53亿	-4.18%	11.0亿	15.2亿	-4.19亿	-4.96%	9.08亿	9.02亿	+647万	+0.08%
5	AAPL	苹果	233.85	+1.10%	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
4	00700	腾讯控股	415.800	-0.29%	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
3	600519	贵州茅台	1522.50	-2.04%	-4.79亿	1.52亿	18.0亿	22.9亿	-4.90亿	-6.86%	21.4亿	21.2亿	+1096万	+0.15%	30.5亿	25.7亿	+4.79亿	+6.71%	--	14.4万	-14.4万	-0.00%
2	300059	东方财富	20.38	+1.09%	-8.07亿	5.81亿	58.5亿	69.3亿	-10.8亿	-4.71%	65.3亿	62.5亿	+2.71亿	+1.19%	62.3亿	60.6亿	+1.71亿	+0.75%	37.0亿	30.6亿	+6.37亿	+2.78%
1	000001	上证指数	3202.95	+0.05%	-67.0亿	122亿	690亿	702亿	-11.9亿	-0.22%	1251亿	1306亿	-55.1亿	-1.03%	1696亿	1709亿	-13.4亿	-0.25%	1564亿	1483亿	+80.4亿	+1.51%

如何计算主力流入、主力流出和主力净流入？

目前大部分行情软件（如同花顺、东方财富、通达信、国信金太阳、深证云等）普遍采用如下计算方式。即

- 主力流入=所有超大买单金额+所有大买单金额；
- 主力流出=所有超大卖单金额+所有大卖单金额；
- 主力净流入=主力流入-主力流出。

超大单+大单

他们在一些细节处存在区别，比如如何界定小单、中单、大单、超大单？如下图所示。

通达信，同花顺，东方财富，大智慧大单定义的区分		
通达信 L2 大单标准		
类型	股	金额
特大单	>=300000 股（30 万股以上）	>=60 万
大单	>=100000 股（10 万股到 30 万股）	>=20 万 --- 60 万
中单	>=20000 股（2 万股到 10 万股）	>=4 万 --- 20 万
小单	2 万股以下	4 万以下
同花顺上证 A 股、深证主板大单标准		
类型	股	金额
特大单	>=200000 股（20 万股以上）	>=100 万
大单	>=60000 股（6 万股到 20 万股）	>=30 万 --- 100 万（或者占流通盘 0.1%）
中单	>=20000 股（2 万股到 6 万股）	>=5 万 --- 30 万
小单	1 万股以下	5 万以下
同花顺中小板、创业板大单标准		
类型	股	金额
特大单		>=50 万
大单		>=20 万 --- 50 万
中单		>=5 万 --- 20 万
小单		5 万以下
东方财富大单标准		
类型	股	金额
特大单	>=500000 股（50 万股以上）	>=100 万
大单	>=100000 股（10 万股到 50 万股）	>=20 万 --- 100 万
中单	>=20000 股（2 万股到 10 万股）	>=4 万 --- 20 万
小单	2 万股以下	4 万以下
大智慧 L2 大单标准		
类型	股	金额
特大单	>=500000 股（50 万股以上）	>=100 万
大单	>=100000 股（10 万股到 50 万股）	>=20 万 --- 100 万
中单	>=20000 股（2 万股到 10 万股）	>=4 万 --- 20 万
小单	2 万股以下	4 万以下

为简单起见，我们采取华泰金工 insight 的判断标准（本项目仅考虑主板），即对任何一个成交单，其成交量、成交额或流通盘占比落入的区间的较高标准即是该单的类型，均可以认为对应的单型。详细描述请见：

三者取最高标准

判断标准	成交量		成交额		流通盘占比	
	中小创	主板	中小创	主板	中小创	主板
超大单	≥10万股	≥20万股	≥50万元	≥100万元	≥0.3%	
大单	6–10万股	6–20万股	20–50万元	30–100万元	0.1%–0.3%	
中单	1–6万股		5–20万元	5–30万元	0.017%–0.1%	
小单	≤1万股		≤5万元		≤0.017%	

https://findata-insight.htsc.com:9151/help/exquisiteArticle/LevelTwo/#2 。

举例：如果一个成交单的成交量是 15 万股，成交额是 120 万元，流通盘占比 0.2%，则记该单为超大单。

成交额满足超大单

需要自行进行数据清洗

二、原始数据：深交所 Level-2 数据（逐笔委托、逐笔成交）

✓ 逐笔委托数据表： order

逐笔委托数据记录了所有投资者在股票市场上委托的买入和卖出订单的详细信息。

中文名	英文名	数据类型	主键	注释
交易日期	tradedate	N8		
数据生成时间	OrigTime	Int64		交易所数据生成时间
发送时间	SendTime	Int64		
接收时间	recvtime	Int64		
入库时间	dbtime	Int64		
频道代码	ChannelNo	uInt16	PK	证券集代号。
行情类别	MDStreamID	C3		
委托索引	ApplSeqNum	Int64	PK	消息 ID
证券代码	SecurityID	C8		证券代码
证券代码源	SecurityIDSource	C4		102 = 深圳证券交易所
委托价格	Price	N(9,3)		委托价格 3 位小数
委托数量	OrderQty	N(9)		委托数量
委托时间	TransactTime	N(20)		委托时间
买卖方向	Side	C2		1 = 买，2 = 卖 G= 借入，F = 借出
委托类别	OrderType	C2		1=市价，2 = 限价，U = 本方最优
定价行情约定号	ConfirmID	C20		0.0
联系人	Contactor	C20		NULL
联系方式	ContactInfo	C50		NULL
期限	ExpirationDays	N8		0
期限类型	ExpirationType	N8		0

✓ 逐笔成交数据表： trade

逐笔成交数据记录了每一笔订单成交的详细信息，包括交易价格、交易数量、买卖方

的帐户信息以及交易时间等等。

中文名	英文名	数据类型	主键	注释
交易日期	tradedate	N(8)		
数据生成时间	OrigTime	Int64		
发送时间	SendTime	Int64		
接收时间	recvtime	Int64		
入库时间	dbtime	Int64		
频道代码	ChannelNo	uInt16	PK	证券集代号。
行情类别	MDStreamID	C3		
成交索引	ApplSeqNum	Int64	PK	消息 ID
证券代码	SecurityID	C8		证券代码
证券代码源	SecurityIDSource	C4		102 = 深圳证券交易所
买方委托索引	BidApplSeqNum	Int64		买方委托索引 从 1 开始计数，0 表示无对应委托
卖方委托索引	OfferApplSeqNum	Int64		卖方委托索引 从 1 开始计数，0 表示无对应委托
成交价格	Price	N(9,3)		成交价格 3 位小数 6位小数
成交数量	TradeQty	N(9)		成交数量
成交类别	ExecType	C2		成交类别 4 = 撤消 F=成交
成交时间	tradetime	N20		成交时间

✓ 订单撮合成交

开市期间分为集合竞价和连续竞价两种阶段. 以中国 A 股市场为例，9:15 至 9:25 为开盘集合竞价时间，9:30 至 11:30、13:00 至 14:57 为连续竞价时间，14:57 至 15:00 为收盘集合竞价时间. 连续竞价时间阶段，市场中的每位交易者可向交易所连续提交两种订单:限价单和市价单. 限价单即限定价格的订单（OrderType=2），需指定买卖方向、价格和量. 市价单是由市场自动确定价格的订单（OrderType=1），需指定买卖方向和量. 交易所接收所有交易者提交的订单，并维护一个限价订单簿(limited order book,

LOB)，用于记录所有已提交且未成交的(限价)订单．如图 1 所示，买卖两个方向最容易成交的价格称为一档买价/一档卖价．一档买价和一档卖价的均值称为中间价(mid-price)．其余档位价格从中间价依次向两边排序.

当有交易者提交新订单时，世界主流交易所均采取连续双向拍卖（continuous double auction, CDA）机制对新订单进行逐笔撮合成交．下面以买单为例进行介绍，卖单遵循相同原则但方向相反．如果是限价单, CDA 将新买单按照“价格优先-时间优先”的顺序在 LOB 的卖出方向从低到高匹配价格．如果匹配成功，则按照订单进入 LOB 的先后顺序依次消掉该档位上的卖单，直到新买单的量全部匹配．如果匹配不成功，则意味着该买单价格没有超过一档买价， 则按照其价格插入 LOB 中买方对应档位.

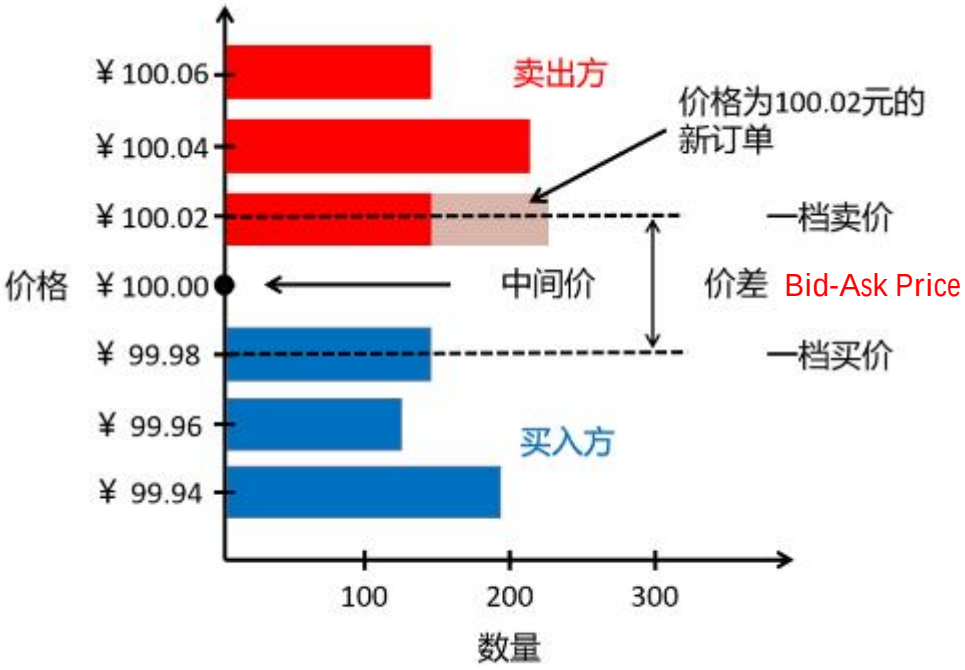


图 1　限价订单簿示意图

三、大体计算过程（可进一步优化）

需要注意，因为存在撤单等原因，委托单中包含的量不一定会全部成交。比如，A 下了一笔 10 万股的卖单，B 以可成交的价格下了一笔 8 万股的买单，因此 A 的这笔委托卖单成交了 8 万股，还剩 2 万股，可能一直不会成交，也可能过一段时间成交，或者被 A 撤

单。同时注意一笔委托单可能对应了多笔成交单（不同交易者的对手方订单）。比如，A 下了一笔 10 万股的卖单，另外有 4 个人分别以可成交的价格下了 2 万、1.5 万、3.5 万、3 万的买单，则该笔委托卖单实际会在逐笔成交表中存在 4 笔记录，这 4 笔记录的卖方委托索引 OfferApplSeqNum 相同（OfferApplSeqNum 为该 10 万股的卖单在逐笔委托表里的委托索引 ApplSeqNum 值），买方委托索引 BidApplSeqNum 则为四个不同的索引值，对应逐笔委托表中的 4 个委托买单（可根据 BidApplSeqNum 值在逐笔委托表里按委托索引 ApplSeqNum 查找）。

因此，存在所谓主动成交和被动成交的区别。主动成交是一种交易行为。具体而言，当一个委托单被下达后，如果能够立即与订单簿中已有的对手方委托单达成交易并成交，这种情况就被称为主动成交。例如，在证券交易中，投资者下达一个买入委托单，而此时订单簿中正好有符合其买入价格等条件的卖出委托单，两者匹配成功并立即成交，这就是主动成交的一种表现。相应地，被成交的已有委托单的这笔成交记录即是该委托单的被动成交。任何一笔逐笔成交记录都包含了涉及的买卖双方委托单索引（BidApplSeqNum 和 OfferApplSeqNum），这个索引对应了逐笔委托中的 ApplSeqNum。不难理解，买卖双方委托单下单较晚的那个单即为主动成交方。下单时间即是逐笔委托数据中的 TransactTime。我们所有的成交量和成交额计算均只考虑主动成交记录。

为应对上述过程，我们建议可以参考如下过程实现：

可以按照此流程图实现，也可以自己规划计算步骤

每股筛选成交数据并判断主动/被动

1）对每一支股票（SecurityID）的逐笔成交数据，筛选为成交的记录（ExecType=F）。并分别依据其 BidApplSeqNum 和 OfferApplSeqNum 索引在逐笔委托中找到对应的 TransactTime，判断该笔成交是主动买还是主动卖。主动买和主动卖可以通过比较买卖双方的 TransactTime 来判断，TransactTime 在后的即为主动单。具体来说可以对比买卖方委托单中 TransactTime，若买方大于卖方，则为主动买单，否则为主动卖单。

去重合并一定时间范围内的总主动单

2）把给定时间范围内的所有 BidApplSeqNum 或 OfferApplSeqNum 相同的主动单进行去重合并，合并方式为将 BidApplSeqNum 或 OfferApplSeqNum 相同的主动单成交量和成交额加和计算。注意，从下一章数据数据的要求来看，这里计算成交量或者成交额的时间范围是一个可变参数。

合并成总成交单，计算三者，判断其类型

3）对合并后的成交量、成交额以及流通盘，按照前述规则判断是否为大单或超大单。

成交量：如果同一个索引（BidApplSeqNum 或 OfferApplSeqNum）对应多笔成交

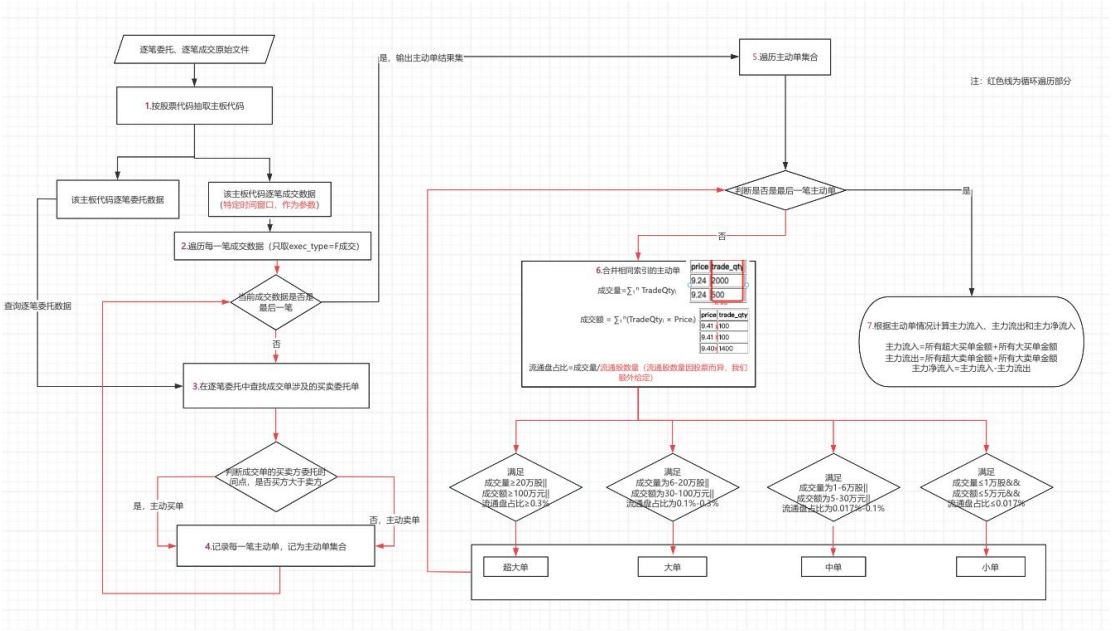
算完最后一笔，最后计算主力净流入

单，则对其 TradeQty 进行求和。

成交额：如果同一个索引（BidApplSeqNum 或 OfferApplSeqNum）对应多笔成交单，其中不成交单可能以不同的 Price 成交，则对其多笔成交的 TradeQty*Price 求和。

流通盘占比：单笔委托的成交量/流通股数量。流通股数量因股票而异，我们额外给定。

4）大体计算过程的流程图如下：



四、输出数据

5.05g量级：用numpy，sklearn可能比MapReduce算得更快
但是本节课要用（数据多了优势就体现出来了）

给定原始数据集（一个交易日的多支股票数据），程序需要按时间范围配置参数 T_window（该时间以 tradetime 字段为依据）。当用户输入时间范围 T_window 的值后，输出如下数据，包括每 T_window 时间内的：

需要一个可以运行的程序
最后要根据test所需参数来设定参数

主力净流入，主力流入，主力流出，超大买单成交量，超大买单成交额，超大卖单成交量，超大卖单成交额，大买单成交量，大买单成交额，大卖单成交量，大卖单成交额，中买单成交量，中买单成交额，中卖单成交量，中卖单成交额，小买单成交量，小买单成交额，小卖单成交量，小卖单成交额。

五、原始数据

校内网下载：http://172.18.30.155/p/DWa7pnYQNRg5IAA

深市 20190102 的全量逐笔委托和逐笔成交，共 5.05GB

提示：逐笔成交数据和逐笔委托数据中，不是所有字段都是对于本任务有用的。

过滤掉一些不用的——考察对题目信息的理解
不考察算法设计，考察分布式程序设计能力

六、评分标准

整个 project 占总评的 40%，即满分 40 分。

其中，

模块化设计：不能all function在一个class里

- ✓ 技术报告（reports）占 10 分，包含问题描述，任务理解，难点分析，整体技术方案（图和文字详细描述），代码的模块化设计思路等。代码必须有详细注释，与有效代码行数相比，至少达到 1:1 比例。主要考察文档撰写的清晰程度。 需要口头ppt+网上可视化展示
- ✓ 展示（presentation）占 10 分。其中口头报告的清晰程度占 5 分，包括对任务的理解，整体技术方案，代码的模块化设计思路等。输出数据的可视化水平占 5 分，建议采用前端技术进行网页展示，网上有许多图形组件供使用。具体地，如采用实时更新每分钟/十分钟上述输出结果的可视化方案，最高可到 5 分。如果只有全天的资金流入流出统计量可视化，按最高 4 分评价。
- ✓ 代码（codes）占 20 分，包括准确性测试，速度测试。准确性占 15 分，速度占 5 分.

必须使用 hdfs+mapreduce 的方式设计方案和编程实现

我们给出若干股票作为示例数据及标准答案供大家完成 project。测试时将采用另外的数据进行测试。

准确性测试

按最终输出的数据的正确性进行评分。我们给出若干股票作为示例数据及标准答案供大家完成 project。测试时将采用另外的数据进行测试并基于新数据进行准确性评分。具体给分方式稍后给出。

去年：分级评分（一定时间内多少分）

速度测试：

按整体程序在课程分配的 docker 中的运行时间来评分（不含可视化部份时间）。具体评分机制稍后给出。

去年有两组市价单的组，跳出了老师给的流程图，速度达到26s（最慢的3分钟，30s是优秀平均水平）给深交所了的程序