题目：高频数据量化因子分布式计算

一、题目背景

在股市中，量化因子，或称为因子，是用于分析和预测金融市场的统计指标或变量。它们基于历史数据、市场行为和经济理论，帮助交易策略在复杂的市场环境中识别潜在的投资机会。量化因子的本质是通过数据驱动的方法来捕捉影响股票价格、资产收益的基本特征，从而指导投资决策。相比传统的主观判断，量化因子提供了一种更加系统化、客观化的方式来分析市场和资产，从而降低了情绪和人为偏差对决策的影响。

量化因子的作用体现在多个方面。首先，量化因子帮助投资者识别出市场中的潜在机会。例如，某些因子可以揭示出某一资产或市场的高回报潜力，而另一些因子则能够捕捉市场趋势的变化，帮助投资者调整投资组合。其次，量化因子还能有效地衡量和管理风险。在投资过程中，通过对量化因子的分析，投资者能够更精确地控制风险敞口，避免单一资产或市场暴露带来的过大风险。通过对量化因子的不断优化和调整，投资者可以实现回报最大化的同时，保持合理的风险水平。

在量化因子的应用过程中，IC（信息比率）是衡量因子有效性的一个关键指标。IC 值衡量了因子与资产回报之间的相关性，是判断量化因子预测能力的关键标准。通常，IC 值的范围从-1 到+1，正值表示因子与回报之间具有正向关系，负值则表明二者存在负向关系，而接近零的 IC 值则说明该因子的预测能力较弱。较高的 IC 值意味着该因子能够较好地反映市场的潜在趋势，并帮助投资者作出有效的决策。通过对因子的 IC 值进行分析和筛选，投资者可以评估每个因子的有效性，进而决定其在实际交易中的应用。

二、原始数据：深交所 Level-10 数据（行情快照）

✓ 行情快照数据表：
行情快照数据记录了每个时间现价订单簿的 10 档价量数据，数据为 tradeTime 3 秒频数据，

字段名称	字段英文名	类型	说明
交易日期	tradingDay	Int32	交易日期，如 yyyymmdd

交易时间	tradeTime	Int64	交易时间，如 hmmssffffffff
接收时间	recvtime	Int64	委托接受时间，如 hmmssffffffff
市场识别码	MIC	Object	XSHE,深交所，XSHG,上交所，XBEI，北交所
股票代码	code	Object	
累计成交笔数	cumCut	Int32	当天开盘以来的成交笔数总和
累计成交量	cumVol	Int64	当天总成交股数
累计成交金额	turnover	Int64	当天总成交金额
最新成交价	last	Int64	最新一笔成交价格
开盘价	open	Int64	当天第一笔成交的价格
最高价	high	Int64	当日迄今为止最高成交价
最低价	low	Int64	当日迄今为止最低成交价
全市场买单总量	tBidVol	Int64	当前全部买单的总股数
全市场卖单总量	tAskVol	Int64	当前全部卖单的总股数
加权平均买价	wBidPrc	Int64	以买单数量加权平均的买价
加权平均卖价	wAskPrc	Int64	以卖单数量加权平均的卖价
持仓量	openInterest	Int64	通常用于期货/可转债市场，表示未平仓合约数；对股票而言无意义，因此为 0 或 int64 最小值
买一价	bp1	Int64	买方最优价格
买一量	bv1	Int64	买一价位的委托数量
卖一价	ap1	Int64	卖家最优报价
卖一量	av1	Int64	卖一价位的委托数量
	bp2	Int64	
	bv2	Int64	

	ap2	Int64	
	av2	Int64	
	bp3	Int64	
	bv3	Int64	
	ap3	Int64	
	av3	Int64	
	bp4	Int64	
	bv4	Int64	
	ap4	Int64	
	av4	Int64	
	bp5	Int64	
	bv5	Int64	
	ap5	Int64	
	av5	Int64	
	bp6	Int64	
	bv6	Int64	
	ap6	Int64	
	av6	Int64	
	bp7	Int64	
	bv7	Int64	
	ap7	Int64	

	av7	Int64	
	bp8	Int64	
	bv8	Int64	
	ap8	Int64	
	av8	Int64	
	bp9	Int64	
	bv9	Int64	
	ap9	Int64	
	av9	Int64	
	bp10	Int64	
	bv10	Int64	
	ap10	Int64	
	av10	Int64	

三、LOB 数据 20 个因子表

数据以当前 tradeTime 时刻为 t 时刻。

序号	因子名称	简要描述	公式（以第 t 时刻为准）
1	最优价差	买一卖一价差	$ap1_t - bp1_t$
2	相对价差	价差相对于中间价的	$\frac{ap1_t - bp1_t}{((ap1_t + bp1_t)/2)}$

		比例	
3	中间价	买卖一档均价	$(ap1_t + bp1_t) / 2$
4	买一不平衡	买卖一档挂单不平衡	$\frac{(bv1_t - av1_t)}{(bv1_t + av1_t)}$
5	多档不平衡	前 n 档买卖量不平衡	$\frac{(\Sigma_{\{i=1..n\}} bv(i)_t - \Sigma_{\{i=1..n\}} av(i)_t)}{(\Sigma_{\{i=1..n\}} bv(i)_t + \Sigma_{\{i=1..n\}} av(i)_t)}$
6	买方深度	前 n 档买单总量	$\Sigma_{\{i=1..n\}} bv(i)_t$
7	卖方深度	前 n 档卖单总量	$\Sigma_{\{i=1..n\}} av(i)_t$
8	深度差	买卖深度差	$\Sigma_{\{i=1..n\}} bv(i)_t - \Sigma_{\{i=1..n\}} av(i)_t$
9	深度比	买卖深度比值	$\frac{\Sigma_{\{i=1..n\}} bv(i)_t}{\Sigma_{\{i=1..n\}} av(i)_t}$
10	买卖量平衡指数	全市场买卖总量平衡指标	$\frac{(tBidVol_t - tAskVol_t)}{(tBidVol_t + tAskVol_t)}$
11	买方加权价格 $(VWAPBid(n))$	前 n 档买价按	$\frac{\Sigma_{\{i=1..n\}}(bp(i)_t * bv(i)_t)}{\Sigma_{\{i=1..n\}} bv(i)_t}$

		挂单量加权平均	
12	卖方加权价格 $(VWAPAsk(n))$	前 n 档卖价按挂单量加权平均	$\frac{\Sigma_{\{i=1..n\}}(ap(i)_t * av(i)_t)}{\Sigma_{\{i=1..n\}}av(i)_t}$
13	加权中间价	综合加权中价	$\frac{\Sigma_{\{i=1..n\}}bp(i)_t * bv(i)_t \ + \ \Sigma_{\{i=1..n\}}ap(i)_t * av(i)_t}{(\Sigma_{\{i=1..n\}}bv(i)_t \ + \ \Sigma_{\{i=1..n\}}av(i)_t)}$
14	加权价差(参考因子 11，12)	买卖加权价差	$VWAPAsk\_t(n) \ - \ VWAPBid\_t(n)$
15	买卖密度差	每档平均挂单量差	$(\Sigma_{\{i=1..n\}}bv(i)_t/n) \ - \ (\Sigma_{\{i=1..n\}}av(i)_t/n)$
16	买卖不对称度	按档位衰减加权的不平衡	$\frac{\left(\Sigma_{\{i=1..n\}}\left(\frac{bv(i)_t}{i}\right) - \ \Sigma_{\{i=1..n\}}\left(\frac{av(i)_t}{i}\right)\right)}{(\Sigma_{\{i=1..n\}}(bv(i)_t/i) \ + \ \Sigma_{\{i=1..n\}}(av(i)_t/i))}$
17	最优价变动	最优报价变化幅度	$ap1_t \ - \ ap1_{\{t-\Delta t\}}$
18	中间价变动	中间价的变化	$\left(\frac{ap1_t + bp1_t}{2}\right) - \ (\frac{ap1_{\{t-\Delta t\}} + bp1_{\{t-\Delta t\}}}{2})$
19	深度比变动	买卖深	$\frac{\Sigma_{\{i=1..n\}}\, bv(i)_t}{\Sigma_{\{i=1..n\}}\, av(i)_t} - \frac{\Sigma_{\{i=1..n\}}\, bv(i)_{\{t-\Delta t\}}}{\Sigma_{\{i=1..n\}}\, av(i)_{\{t-\Delta t\}}}$

		度比的变化率	
20	价压指标	价差相对于深度的压力指标	$\frac{(ap1_t - bp1_t)}{(\Sigma_{\{i=1..n\}} bv(i)_t + \Sigma_{\{i=1..n\}} av(i)_t)}$

参数：

n = 5, Δt = 1

四、输出数据

训练阶段（各小组自行编写 MapReduce 阶段）：

给定原始数据集（20240102-20240108 沪深 300 指数全部股票数据），程序需要计算得到第三部分给出的所有因子的因子值序列。对于每一天数据，输出所有因子在 300 只股票上从 9:30:00 开始到 15:00:00 的平均因子值序列。输出格式要求参考标准输出，第一行为列名，分别是 tradeTime 和因子代码(alpha_n)，从第二行开始输出时刻与每个因子在 300 只股票上的平均因子值$\bar{f}_t$。

$f_t(i)$：资产 i 在时间 t 的因子值，在本 project 中资产指的是 CSI300 股票，$i = 1, \dots, N_t$：当期资产池内的资产，N_t=300：

$$\bar{f}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} f_t(i)$$

测试阶段（第 16 周现场测试）：

给定另一天的沪深 300 指数全部股票数据，要求输出所有因子在 300 只股票上从 9:30:00 开始到 15:00:00 的平均因子值序列。输出格式要求与上面一致。

*因子值计算过程可能出现分母为 0 的情况，在分母上添加极小值 1e-7。

五、原始数据

校内网下载：

深沪两市 20240102-20240108 共五天的行情快照数据表,校内坚果云数据下载链接：

csi300.zip - 坚果云 - 南方科技大学

因子值标准答案 - 坚果云 - 南方科技大学

提示：行情快照数据表数据中，不是所有字段都是对于本任务有用的。

六、评分标准

整个 project 占总评的 40%，即满分 40 分。
其中，

- ✓ 技术报告（reports）占 5 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。
- ✓ 展示（presentation）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。
- ✓ 代码（codes）占 30 分，包括准确性测试，速度测试。准确性占 20 分（一题一分），速度占 10 分（按时间排名来算）.

必须使用 HDFS+MapReduce 的方式设计方案和编程实现，必须使用 JAVA 语言。
我们给出因子值作为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。对于每一题，同学计算出的因子平均值与标准因子平均值的误差不超过 1%，误差计算方式为 $(\frac{|标准平均值|-|平均值|}{|标准平均值|})$，则算正确，该题得 1 分；否则该题不得分。

行业通用的评估指标采用 Pearson IC，计算公式如下。本次项目不涉及，仅供自行了解。

- $f_t(i)$：资产 i 在时间 t 的因子值，在本 project 中资产指的是 CSI300 股票，
- $R_{t\to t+h}(i)$：资产 i 从 t 到 $t+h$ 的未来收益，为$(\frac{P_{\{t+h\}}(i)}{P_t(i)}-1)\%$，$h=10$个时间步
- $P_t(i)$：资产 i 在 t 时刻的最新成交价格（last）

- $i = 1, \ldots, N_t$：当期资产池内的资产

- $\bar{f}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} f_t(i)$，$\bar{R}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} R_{t \to t+h}(i)$

Pearson IC(横截面线性相关)

$$\mathrm{IC}_t = \frac{\sum_{i=1}^{N_t} (f_t(i) - \bar{f}_t)\,(R_{t \to t+h}(i) - \bar{R}_t)}{\sqrt{\sum_{i=1}^{N_t} (f_t(i) - \bar{f}_t)^2}\,\sqrt{\sum_{i=1}^{N_t} (R_{t \to t+h}(i) - \bar{R}_t)^2}}$$

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。具体评分机制按照运行时间排序分组，时间排序前 5%的组获得此项所有分数，时间排序后 5%的组获得此项 40%分数，分数按照 5%分组从最快到最慢线性递减。