目标：

处理和分析股票市场数据，主要考察如何使用 Hadoop 工具来处理大规模数据。

数据介绍

本 project 数据涉及深圳交易所某一天内所有股票的相关原始数据，共 5.05 GB，包括

- ✓ 逐笔委托数据（文件名：am_hq_order_spot.txt，大小：1.66GB、文件名：pm_hq_order_spot.txt，大小：1.06GB），
- ✓ 逐笔成交数据（文件名：am_hq_trade_spot.txt，大小：1.34GB、文件名：pm_hq_trade_spot.txt，大小：986.6 MB）。

数据已上传至共享文件夹：/shareddata

逐笔委托、逐笔成交是在股票市场和金融领域中常见的数据类型，它们用于跟踪和分析股票市场的交易活动。以下是对这些数据的解释：

1. 逐笔委托数据：
   - 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。这包括投资者的帐户、订单价格、订单数量以及订单时刻等信息。
   - 逐笔委托数据通常是实时的，它们展示投资者正在以什么价格和数量来买入或卖出股票。这对于短期交易者和算法交易非常重要，因为它提供了市场上实际订单的细节，帮助他们做出决策。

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		
联系人	Contactor	C20		
联系方式	ContactInfo	C50		
期限	ExpirationDays	N8		
期限类型	ExpirationType	N8		

2. 逐笔成交数据：

- 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。
- 这些数据可以得到每笔交易的具体细节，例如某股票何时以何种价格成交，这对于交易策略的执行和监控非常有用。

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

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

3. 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.

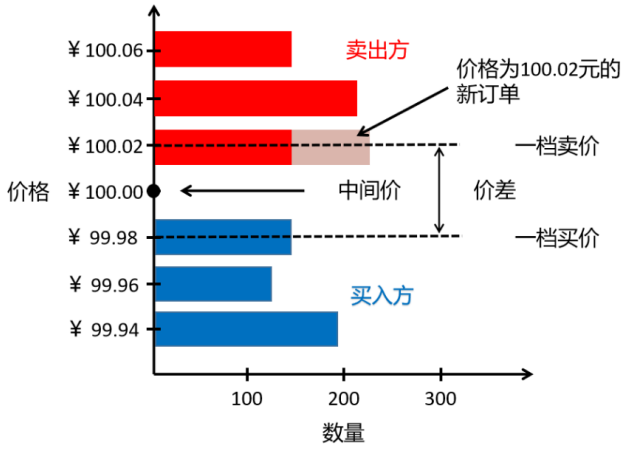


图 1　限价订单簿示意图

4.　市价单的提交类型

（一）对手方最优价格申报；（OrderType=1）

（二）本方最优价格申报；（OrderType=U）

（三）最优五档即时成交剩余撤销申报；（OrderType=1）

（四）即时成交剩余撤销申报；（OrderType=1）

（五）全额成交或撤销申报；（OrderType=1）

（六）本所规定的其他类型。

✓　对手方最优价格申报，以申报进入交易主机时集中申报簿中对手方队列的最优价格为其申报价格。

✓　本方最优价格申报，以申报进入交易主机时集中申报簿中本方队列的最优价格为其申报价格。

✓　最优五档即时成交剩余撤销申报，以对手方价格为成交价，与申报进入交易主机时集中申报簿中对手方最优五个价位的申报队列依次成交，未成交部分自动撤销。

✓　即时成交并撤销申报，以对手方价格为成交价，与申报进入交易主机时集中申报簿中对手方所有申报队列依次成交，未成交部分自动撤销。

- ✓ 全额成交或撤销申报，以对手方价格为成交价，如与申报进入交易主机时集中申报簿中对手方所有申报队列依次成交能够使其完全成交的，则依次成交，否则申报全部自动撤销。

5. 撤单

在逐笔成交数据中，ExecType=4 的订单是撤单。有两种撤单来源，一种是交易者主动撤销未成交的限价单，另一种是市价单中规定的“撤销申报”（如市价类型三四五）。

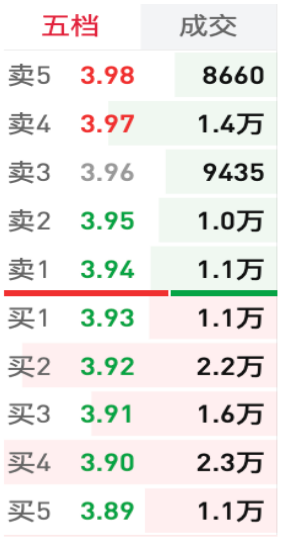
6. 订单索引

在逐笔委托数据中的每一个订单都有一个唯一编号 ApplSeqNum；在逐笔成交数据中，每一笔成交单的 BidApplSeqNum（买方索引）、OfferApplSeqNum（卖方索引）的值即为其在逐笔委托数据中的 ApplSeqNum（取决于委托单是买还是卖）。因此，可以通过这三个索引将逐笔成交数据和逐笔委托数据进行对应。

Project 任务：市价单的价格档位确定

真实的市价单均为 OrderType=1，缺乏更细粒度的区分字段。我们希望确定每笔市价单的最低成交档位，并对这些市价单新增一个字段来记录其最低成交档位。**我们特别考虑连续竞价阶段的数据。**

以如下限价订单簿为例，一个 40000 手的市价卖单，其成交价格为 3.93 元卖 11000 手，3.92 元卖 22000 手，3.91 元卖 7000 手，则最低成交档位为第 3 档（对应 3.91 元），方向为卖。又如，一个 40000 手的市价买单，其成交价格为 3.94 元买 11000 手，3.95 元买 10000 手，3.96 元买 9435 手，3.97 元买 9565 手，则最低成交档位是第四档（对应 3.97 元），方向为买.



五档		成交
卖5	3.98	8660
卖4	3.97	1.4万
卖3	3.96	9435
卖2	3.95	1.0万
卖1	3.94	1.1万
买1	3.93	1.1万
买2	3.92	2.2万
买3	3.91	1.6万
买4	3.90	2.3万
买5	3.89	1.1万

以下是一个处理该问题的一种方法的流程，可以参考实现。

1）从逐笔委托和逐笔成交中选择我们所关心的股票代码对应的数据，并抽取连续竞价阶段的数据，分别记为 Order 和 Trade.

2）从 Trade 数据中依据 ExecType 分离出成交数据 Traded 和撤单数据 Cancel

3）从 Order 数据中依据 OrderType 分离出市价订单数据 MarketOrder 和限价订单数据 LimitedOrder，以及本方最优 SpecOrder

4）对 MarketOrder 中的每一笔订单，新增一个字段 MARKET_ORDER_TYPE，记录其最低成交档位

a) 通过其 ApplSeqNum 索引在 Traded 的 BidApplSeqNum 和 OfferApplSeqNum 索引中找其对应的成交记录

b) 如果找到的成交数据存在 K 种不同成交价格，该市价单的 MARKET_ORDER_TYPE 记为 K。K=0，1，2,....

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。

评分约束：

1）正确性验证只会抽取平安银行的代码：000001

2）必须使用 hdfs+mapreduce 的方式设计方案和编程实现

3）整个程序要求输出一个文件 Output，该文件包含四种订单数据：已标记档位的 MarketOrder 数据，LimitedOrder 数据，SpecOrder 数据和 Cancel 撤单数据。所有订单按照时间从小到大的顺序排列。其中，MarketOrder，LimitedOrder 和 SpecOrder 数据以“委托时间”为准，而 Cancel 数据以“成交时间”为准。该 Output 文件以如下格式组织：

TIMESTAMP, PRICE, SIZE, BUY_SELL_FLAG, ORDER_TYPE, ORDER_ID, MARKET_ORDER_TYPE，CANCEL_TYPE

输出全部委托单，最终输出一个 txt 文本。---------Output.txt

字段	解释
TIMESTAMP	委托时间 / 成交时间（撤单）
PRICE	价格（限价单填写，其余订单类型为空）
BUY_SELL_FLAG	1 = 买，2 = 卖
ORDER_TYPE	委托单类型 1=市价，2 = 限价，U = 本方最优
ORDER_ID	委托单 id （ApplSeqNum）
MARKET_ORDER_TYPE	K （市价单填写，其余订单类型为空）
CANCEL_TYPE	1=是撤单，2=非撤单

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 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。
- ✓ 展示（presentation）占 5 分，包含对任务的理解，整体技术方案，代码的模块化设计思路等。主要考察口头报告的清晰程度。
- ✓ 代码（codes）占 25 分，包括准确性测试和速度测试。准确性占 20 分，速度占 5 分.

准确性测试

按最终输出的 Output 数据的正确性进行评分。具体评分机制稍后给出。

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。具体评分机制稍后给出。

附加分：3 分

未采用前述参考流程，自行设计了新的方案。视新方案与参考方案差异给分。