UBS 实习汇报

汇报人 许震宇 日期 2026.1.29

一、 一个简短的自我介绍

- 许震宇
- 中国科学院大学(北京怀柔),分数线仅次于清北
- 家在深圳、港籍
- 徒步、网球、滑雪...

微信:

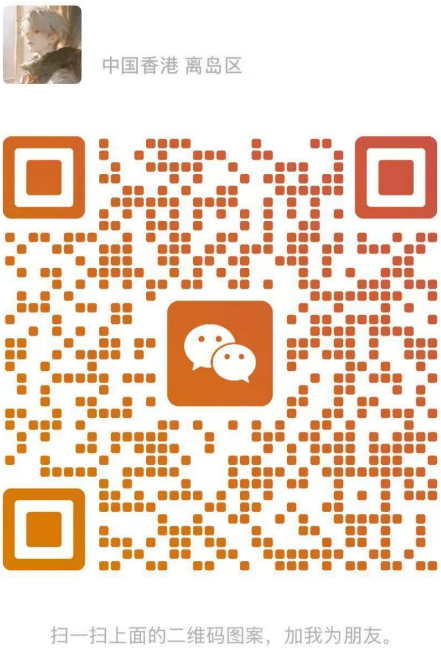

图 1:

二、 数据格式调研

数据格式	适用场景与优点	备注
CSV	通用、可读性强、便于快速查看与交换	体积大、读写慢、类型易丢失
Parquet	列式存储、压缩好、读取快，适合大规模因子/行情特征	生态成熟（Spark/Arrow）
Feather/Arrow	序列化快、零拷贝传输，适合进程间共享与临时缓存	适合内存分析
HDF5 (.h5)	分层结构、可存多表/多维数组、支持元数据与随机访问	适合大体量数值数据
SQLite	轻量数据库、支持 SQL 查询，适合中小规模结构化数据	单文件便携
Pickle	Python 对象持久化，开发效率高	跨语言差、版本兼容风险
JSON	可读、跨语言、适合配置与元数据	体积大、速度慢
NPZ	压缩数组打包，便于分发与存档	适合纯数值数组

表 1: 量化常见数据格式对比

三、 数据调研

- 数个指标从 2010 年到 2024 年的分时数据
- 层级：分时数据项 < 每日数据 < 从 2014 ~2024 的数据 < 3 个指标的从 2014 ~2024 的数据
- 每个分时数据项包含：Close High Low Open Time Date Amount Volume