

潘星星



[这是方便打印的版本, 良好排版请看 <https://paxinla.github.io/my-online-resume/cn/>]

☎ 137-9448-1350 | ✉ aksura@email.cn | 🌐 <https://paxinla.github.io>

期望职位: 数据工程 | 期望薪资: 面议 | 期望城市: 深圳或远程

多年专业数据开发经验, 熟悉数据处理全流程, 包括数据清洗、数据存储、解决各种数据问题 (数据堆积、数据延迟、数据传输效率等)。主要工作领域为企业级数据仓库的设计实施、海量数据的 ETL 加工处理、数据处理平台的设计建设、联机数据库架构的设计/实施/维护、SQL 性能调优等, 在项目中高效地合作与协调。为大型商业银行设计并实施过完整的数据仓库项目, 也为互联网企业落地过大数据平台方案。

曾服务过的行业包括传统IT、互联网、AI 公司等, 喜欢严谨务实的团队, 期望能加入业务有长远发展谋划、致力于持续为社会创造价值的企业。

技术栈

🚀 主要技术栈

- ➔ Python/Scala : ★★★★★
- ➔ OLAP 数据仓库/数据治理/建模集成/ETL : ★★★★★
- ➔ OLTP 数据库方案设计优化 : ★★★★★☆
- ➔ 离线海量数据处理方案设计落地 : ★★★★★☆

🔍 涉猎

- ➔ 其他常用语言 : SQL、Bash、Clojure
- ➔ 其他常用工具 : PostgreSQL、Hadoop/Spark、RabbitMQ、MongoDB、Kettle

> 资格证书

- 🏆 PMP | Project Management Professional
- 🏆 OCP 11g | Oracle Certified Professional 11g
- 🏆 软件设计师 | 计算机技术与软件专业技术资格(水平)

> 教育经历

- 🎓 2008年~2012年7月 重庆邮电大学 计算机科学与技术 - 本科
- 🎓 2008年~2012年7月 重庆邮电大学 通信工程 - 本科二专业

> 工作经历

1. 深圳市星航科技有限公司 -- Data Engineer, 3年, 自 2021 起
2. 八斗智能科技 (广州) 有限公司 -- Data Engineer, 1年, 自 2019 起
3. 深圳爱拼信息科技有限公司 -- Data Engineer, 3年, 自 2015 起
4. 深圳四方精创资讯股份有限公司 -- 研发级DBA, 3年, 自 2012 起

项目经历

【2021.10 - 2025.01】 云原生数据湖仓 项目

- Lakehouse | Spark/Databricks | Delta

该项目建立并持续增强一个大规模的湖仓一体系统，为公司研究团队、各 Web3 产品提供了强大的数据支撑。

我在这个项目中负责的工作有：

1. 结合公司的业务，设计、规划与实施现代化湖仓与服务设施。
2. 分析并构建合理的数据模型与 ETL 流程等，持续建设与改进高性能数据服务。
3. 参与业务的数据调研、分析，满足产品的数据需求，支撑部分业务决策。
4. 总结与维护技术文档，推动发展优质产品。

这个由我负责的企业级别的湖仓系统，是基于 Databricks on AWS 的云原生方案而构建的。它完全摆脱了陈旧的 Hadoop 生态工具，在管理和处理数据方面要优越得多。这个方案在并未增加供应商绑定的风险的同时，极大地降低了公司需要投入的物理基础设施成本和运维成本。我使用该湖仓集成并处理了来自区块链上日均约百万行和智能合约的原始数据，高效地为公司的各个 Web3 数据产品的和数据科学家团队提供了高质量的数据。

【2016.08 - 2019.01】 产品联机数据库管理

- PostgreSQL | 阿里云 | 数据库集群

该项目为公司核心产品“X志愿”的联机数据库，部署在阿里云上，可靠地支撑了产品的数据管理需求。

我在这个项目中负责的工作有：

1. 负责开发、测试、生产等所有环境的数据库系统架构的设计、实施、日常操作、配置、监控及故障处理。负责监控备份情况、恢复策略方案的制定与数据备份恢复计划的制定、执行。
2. 负责数据库性能监控和调优，监控并及时发现问题，优化后台对数据库的访问方式、性能较差的 SQL。
3. 负责项目中的数据库逻辑结构设计。分析业务需求，设计审核并创建维护数据库对象。
4. 根据业务需求批量生成/更新产品所需的核心数据集，维护它们的质量。
5. 对数据库相关的故障和性能问题及时定位问题并解决问题。

“X志愿”是公司的核心数据产品之一，它依托全国大专院校历年的招生录取数据及 AI 算法，根据高考考生的自身情况，预测目标学校专业的当年录取情况，给考生个性化的填报志愿的智能建议。公司为百度 App 中全国唯三的志愿填报服务提供商之一。

公司采用的云服务商是阿里云，生产环境都部署在阿里云上。我全权负责该产品的数据库架构的设计与实施。

公司在考察了阿里云 RDS 的性能与稳定性后，综合考虑了价格、“X志愿”产品的数据访问特征、整个产品对数据库的性能及稳定的要求等因素，选择了在 ECS 上自建基于 PostgreSQL 的生产数据库集群。我设计并落地了数据库集群架构方案。集群全年无故障稳定运行。与之前该产品上一代基于 MongoDB 的数据库方案相比，在读写性能、开发灵活性、功能的丰富程度、批量更新数据的效率及服务稳定性上都有了明显地改善。我为该集群实现了备份策略实施自动化，保障了产品数据的安全。业务系统对数据库的访问及业务基础数据的增长在每年高考前后一段时间内，都会剧增；而其他时间段的负载则较为平稳。如果为以能处理这类峰值访问为标准来投入资源随时待命无疑是巨大的浪费。因此，我设计了集群在业务高峰时期前后能平滑地扩容、缩容的方案。集群在业务高峰时，平稳支撑过 TPS 约 8k；在高峰时段外，则以经济的方式稳定支持业务。

【2013.06 - 2015.08】 特色业务装载平台 项目

- 调度 | 海量数据ETL

该项目为某国有大型商业银行深圳分行各特色应用系统实现 ETL 过程及调度的一个平台，它根据下游各业务系统的需要，定制加工并输送来自异构数据源的数据到业务系统的联机库。

我在这个项目中负责的工作有：

1. 负责对该平台的核心调度程序及作业程序进行维护及二次开发。
2. 设计、维护该平台对于总行下传数据文件、特色系统数据库中数据的 ETL 过程。在平台发生故障时及时处理。
3. 根据不同特色应用系统用户提出的新需求，为其定制 ETL 过程，监控、保障每日批量作用的成功完成。
4. 根据特色应用系统用户的要求，协助对其实施历史数据的追取。

这个项目涉及的数据源是多个异构的，属于其他组织的系统，目的地则是许多不同部门的特色业务系统应用层数据集市及联机数据库，对批量作业的执行时效有一定的要求。在我接手该项目后，完成了对大部分已有的历史作业程序的改造，提高了这些作业程序的容错性和平台自动处理错误的能力。

例如，平台的数据源中有一张记录数约为4000万的账户表，是许多下游目标系统的数据来源。它会在每季度结息日数据量剧增，导致其装载作业(merge模式)运行时间过长且容易失败，过去常需人工到现场监控处理。我在接手后，针对该表的 ETL 过程，通过对数据源文件的自动拆分切片、分析增量变更、合并到存量快照表的方式优化。提高了整体数据装载的效率及自动化程度，使得该作业不论在结息日还是平日的运行时间均保持稳定、工作正常。

致谢

非常感谢您花时间阅读我的简历，期待能有机会和您共事。