# 联系⽅式

- ⼿机： 18810276930

- Email： 18810276930@163.com

- Github ID：https://github.com/ChengkaiYang2022

# 个人信息

- 杨成凯/男/1993

- 统招本科/北京理⼯⼤学 计算机科学与技术 2015届

- ⼯作年限：7年

- ⽬前状态：已离职

- 职位： 大数据开发工程师

# ⼯作经历

### 自由职业与尝试留学（2022年4月-2022年12月）

- 辞职后尝试出国留学（准备雅思考试、文书等），但因疫情影响与其他因素最终放弃。

- 系统地学习JVM内存模型、设计模式、大数据经典论文、部分 Flink 代码（connector）。

- 与阿里Flink committor合作，主导翻译Flink开源社区的3篇技术文档包括 Flink Metric Report、用户自定义connector、Flink SQL的DataType，累计约 4000 行，成为 Flink 1.15 与 Flink 1.16的 contributor）

- 自由职业（外包网络爬虫项目）。

### 云帐房网络科技有限公司 大数据开发工程师（2019年5月-2022年1月）

\*公司业务：为代账公司算税、报税的 Saas 服务，5年内完成了5轮投资10亿投资，垂直行业内 Top2 企业。\*

我带领数据团队（5-6人）为其他4个事业部10个团队提供数据服务，包括实时数据开发（Flink）、数据产品开发（Java后端开发）、离线数仓开发、大数据工单服务、部分大数据组件调研与监控等。

### 实时数据开发（Flink）

- \*\*Flink 集群的搭建、监控、运维\*\*。仅使用3台 128G8C 服务器搭建的 Flink Standalone 集群，便支撑了全部实时任务，平台日吞吐量约为5TB。完成 Flink prometheus 监控与钉钉预警。

- \*\*带领团队完成Flink SQL 的技术调研与方案落地\*\* 。其中主要使用了 Hive / Mysql 维表关联、- Flink CEP、Flink SQL、Flink DataStream、Flink Stateful API、Hive & Kudu 实时写入等环节。此技术方案已在部门内部推广，团队基于此方案完成了1、Nginx 日志实时分析、2、用户埋点信息反爬虫、3、爬虫程序耗时计算项目、4、Java 日志& GitHub 账户关联、5、Python 日志、6、Kudu CDC实时数据同步等实时作业开发、7、Flink CDC DDL语句预警项目 将开发时间从周减低为日。

- \*\*Flink 作业调优\*\*。通过 Flink 底层 API 的技术调研，将一个基于Flink Core API 的作业性能提升至16倍，经测试，吞吐量可达到每小时1TB。

### 数据产品开发

- \*\*供应链关系图谱系统\*\*：使用 Spring Boot、Mybatis、Spring Data、Neo4j 完成百万余家企业的供应链图谱、财税报表报告，为代账公司提供获客、信贷评估、税筹数据支持。

- \*\*元数据系统\*\*：使用 Spring Boot 构建+Flink，定时抽取各个应用的mysql源表，解决分库分表与字段变更问题，构建元数据数据字段。

- \*\*自助查询系统\*\*：第一版本使用 Spring Boot 构建，初步实现模板定制、SQL 提交与查询。 第二版本通过对 Hue （Django/Python）二次开发，完成了与Presto的集成、用户大SQL预警，权限控制等，并推广到其他事业部。

- \*\*企业信息爬虫系统\*\*：使用 Scrapy（python）抓取企业信息与发票信息。

### 离线数仓开发

- 负责各个事业部各类应用程序的日志分析，以及用户埋点行为分析。

- 带领团队完成各个事业部的数仓开发需求、报表需求。

### 其他工作

- 带领团队支撑每月上百个工单查询服务，通过定制工单模板、权限校验等功能，提高了用户的查询效率。

- Apache Atlas 数据血缘监控、Kudu 性能测试。

## 中⾦云⾦融（⼤数据） 科技股份有限公司 数据研发⼯程师（2017年2⽉ ~ 2019年5⽉ ）

\*公司业务：面向政府的金融监管科技服务商，辅助各地金融局对当地 P2P 金融公司进行监管。 我带领4-5人团队完成网络爬虫、实时/离线数据处理、舆情数据分析等。

### 实时数据开发与离线开发

- 使⽤Apache Storm、Apache Nifi（流处理工具）、Apache Kafka进⾏对各类网络爬虫数据进行实时处理、汇总以及预警。

- 使⽤ Hortonworks 旗下开源⼯具搭建大数据平台。参与离线数据仓库建设。使⽤Sqoop、 Apache Nifi 等 ETL ⼯具构建ODS数据缓冲层， 收集 MySQL、 MongoDB 数据到 Apache Hive 中。

- 使⽤ HSQL 清洗数据，编写ETL脚本。

### 主导爬⾍系统的开发

\*抓取业务范围包括企业⼯商信息、企业官⽹、各⼤⽹站的招聘信息、舆情信息等多个领域。 抓取的数据主要满⾜政府监管需求，构建企业线上⻛险画像， ⽤于⽀持公司其他内部系统与⻛险报告的数据来源\*

- 从零建⽴企业爬⾍系统， 使⽤requests， scrapy， scrapy cluster， webmagic等⼯具与框架开发30余个网站的爬⾍， 完成了公司的数据采集需求。

- 解决了企查查登陆、 验证、 被封号等问题， 有效采集请求达到15万-20万次/天。支持定时调度与接⼝调度， 另外使⽤kafka缓存爬⾍结果，便于下一步数据收集。

- 使用Redis⾃建了动态ip代理池， 根据爬⾍调度情况⾃动伸缩代理池⼤⼩， 降低成本到原先1/5。

### 其他数据分析相关工作

- 使⽤ python 制作数据分析报告，⽤于数据抽取、 清洗， ⽣成⻛险报告， 降低⼈⼒成本。

- 使⽤ python 的 networkx 与 Neo4j 图数据库解决金融的图相关问题，⽀持复杂关系查询如企业间资⾦内循环等⻛险关系。

## 北京宇信易诚科技有限公司 数仓开发工程师（ 2015年6⽉ - 2017年2⽉ ）

- 华夏银⾏财务类报表项⽬

负责计划财务部、 个⼈业务部等5个部⻔98张报表的设计与开发， 采⽤多维模型的概念， 避免了报表导向开发不够灵活、 数据冗余等问题。

- 华夏银⾏⾮税收⼊收缴项⽬

⾮税收缴收⼊逻辑极为复杂， 优化Oracle存储过程。 抗压能⼒强， 多次2⼩时内解决⽣产数据问题， 得到⾏⽅的肯定。

- 华夏银⾏个⼈业务部报表项⽬ 、 1104⾮现场监管项⽬

个⼈业务贷款存款类报表， 计算利率利息等。

# 技术栈

- \*\*语言\*\*：Java, Python, Scala, SQL

- \*\*实时数据处理\*\*：Flink（Core API/Stateful API/Flink SQL/CDC）/ Storm / Apache Nifi

- \*\*数仓开发（计算引擎）\*\*：Hive / Impala / Presto

- \*\*数据存储\*\*：Hadoop / Kudu / HBase / Kafka / Mysql / Neo4j / Oracle / Redis （MongoDB / Solr ）

- \*\*后端开发\*\*：Spring Boot / Spring Data / Mybatis / Flask / Django。

- \*\*CI/CD\*\*：git / docker

- \*\*计算机基础\*\*：在 JVM 内存模型、设计模式、Java多线程/JUC/NIO 常用库、大数据经典论文。

- \*\*长期专注于数据类工作\*\*。在数据采集（网络爬虫/CDC）、离线数仓、实时数据处理、数据产品开发（后端开发）、大数据组件、数据治理、大数据平台监控等领域有较多经验。

# 沟通与管理能力

- 在近两家公司中有小型团队（4-6人）管理经验。带领团队分阶段完成部门OKR，梳理个人职业规划、拆解任务；帮助团队成员攻克技术、业务难题。

- 具备较好的沟通与协调能力。日常为4个事业部近10个团队提供各类数据服务，通过制定工单规范、优化需求流程、梳理，公示阶段性 OKR 等手段，一定程度上解决了之前需求评审不规范、排期不合理等问题，缓解了部门内、部门间紧张的工作情绪。

- 在\*\*两家公司\*\*获得过一级部门级别优秀员工奖（2018、2021）。在工作中具有主人翁意识，积极了解公司业务背景与痛点，并能分析、识别业务，组织架构中的关键问题，使用合适的方案按阶段解决问题。