联系方式

- 手机（微信）： 18851850600
- 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

尝试留学以及参与 Flink 开源项目（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），但因严格的疫情管控政策与其他因素最终放弃。
- 近期正在系统地学习JVM内存模型与设计模式，欢迎**新石器慧通面试官**点击链接查看我的知识图谱！
- 与阿里Flink committor合作，负责、主导翻译Flink的 3篇官方技术文档，包括：用户自定义connector、Flink Metric Report、Flink SQL的DataType，累计约 4000 行，**成为 Flink 1.15 与 Flink 1.16的 contributor**）

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）

云帐房为高瓴资本领投的E轮财税公司，是财税垂直领域内的 Top2 企业。公司主营业务：为国内近130万中小企业提供算税、报税的 Saas 服务。

我带领大数据部门下的数据服务团队（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）

- 带领团队完成 Flink 的**技术调研与方案落地** 。其中包括：
    - Flink SQL（Hive / Mysql 维表关联、Hive & Kudu 实时写入等主要环节）、Flink DataStream、Flink Stateful API等三层API的实际落地。
    - 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。
- 带领团队基于 Flink 完成了多个项目。其中包括：
    - 1、Nginx 日志实时分析（流量分析、接口分析、反爬虫分析）。
    - 2、Java 日志& GitHub 账户关联、Python 日志处理。
    - 3、Kudu CDC 实时数据同步。
    - 4、用户埋点信息实时分析（反爬虫分析、等）。
    - 5、RPA 机器人程序耗时计算项目。
- **部分Flink 作业调优。**通过对 Flink Stateful 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，超预期完成任务。

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。
- **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，**成为公司最实用的查询工具之一。**
- **元数据系统**：使用 Spring Boot 构建，定时抽取各个应用的mysql源表，构建元数据系统，用于解决源系统频繁分库分表与字段变更的问题。

- **爬虫系统**：指导初级爬虫开发工程师使用 python技术栈（Scrapy、selenium） 抓取企业信息与发票信息、与其他信息。

离线数仓开发

- 负责各个事业部各类应用程序的离线日志分析、用户埋点行为分析。
- 带领团队完成各个事业部的离线数仓开发需求、报表需求。

日常事务类工作

- 在自助查询系统基础上，带领团队完成每月上百个数据工单查询服务。通过定制工单模板、对查询引擎Presto的监控、合理分配资源等手段，提高了用户满意度。
- 完成Apache Atlas 数据血缘监控、Kudu 性能测试、Flink与Hue集成等上级领导分配的任务。

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 Hortonworks 旗下开源工具搭建大数据平台（Apache Storm、Apache Nifi（流处理工具）、Apache Kafka），对各类网络爬虫数据进行实时处理、汇总、入库（HBase）以及预警。
- 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

- 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， 生成风险报告， 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风险关系。

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个部门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）/ 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
- **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

- 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 等方式，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**中获得过一级部门级别的**唯一**优秀员工奖（2018、2021（可提供照片证明）），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。