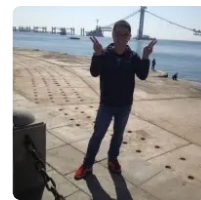


# 杜延龙

男 | 年龄：38岁 | 13552932725

15年工作经验 | 求职意向：数据架构师 | 期望城市：北京



## 个人优势

业务&管理方面：

- 1、具备数据底座（EMR、Clickhouse、StarRocks）从0到1的完整建设经验，能够独立主导技术架构设计与落地实施；
- 2、主导数据开发与治理平台从0到1的研发建设，成功实现平台稳定性提升至99.95%以上，支持高并发、高可用的数据服务场景；
- 3、拥有大规模实时与离线数仓建设经验，支撑实时数据处理规模达TB级以上，离线处理规模达PB级以上，保障数据时效性与准确性；
- 4、有团队管理经验（10人）；
- 5、具备AI Agent开发实践经验，能够结合业务场景实现智能化数据处理与分析功能；
- 6、工作积极主动，责任心强，具备出色的抗压能力与执行力，能够在高强度环境下保持高效输出。

技术方面：

- 1、精通Java、SQL、Scala语言，熟练掌握Spring、Mybatis等主流开发框架，具备扎实的编码能力与系统设计经验；
- 2、熟悉Flink、StarRocks、Spark、Hive、Hadoop、Paimon等计算引擎，具备多引擎选型与性能调优能力；
- 3、熟练使用DataX、Flume、Canal、FlinkX、FlinkCDC、Seatunnel等数据同步工具，对部分工具具备源码级二次开发经验；
- 4、熟悉DolphinScheduler、Airflow源码和二开经验；
- 5、了解Apache Kubbyi、Dinky、Linkis、Zookeeper、Redis、Memcache、Kafka、Dify，看过部分源码；

## 工作经历

### 腾讯音乐娱乐科技（深圳）有限公司

数据架构师

2021.01-至今

围绕企业财经、HR、服务支持等管理域的数字化转型需求，结合业界技术趋势完成整体数据架构规划与落地：

1. 搭建技术底座：确立 Hadoop/Hive/Spark/Datax/StarRocks 离线技术栈、Flink/Kafka/Paimon/StarRocks 实时技术栈，以及 Airflow/DolphinScheduler 调度技术栈，支撑全场景数据处理需求；3000+表、百 TB以上、稳定性 99.97%+；
2. 搭建团队与规范体系：按需完成数据团队招聘与组织架构搭建，制定数据平台、数据仓库使用规范及安全管理规范，保障团队高效合规运作；最高 10 人+；
3. 主导核心平台研发：基于 DolphinScheduler 二次研发优化新一代数据开发平台，自研覆盖元数据采集、资产编目、数据安全、标准与服务的全流程数据资产管理平台；基于数仓构建单源及 Flink 跨源数据质量平台，依托 Alibaba Druid、Nebula Graph 搭建数据血缘平台，同时搭建完善的数据监控运维体系；任务4000+，工作流 300+，日调度万次/日；
4. 技术赋能业务与创新：深入业务一线提供数据应用技术支持，开发轻量化工具保障业务数据的安全性、稳定性与准确性，借助 Data+AI 能力构建问题分析、数据血缘分析机制，实现业务效率显著提升。

### 北京奇虎360科技有限公司

高级数据专家

2017.07-2021.01

- 1、集团各业务线数据计算（离线&实时）需求支持（Hive、Flink、Spark）；
- 2、主导研发分布式数据计算平台（基于 Flink、Spark 提供分布式跨数据源计算的能力），支持 HDFS、Hive、Mysql、Kafka、Hbase、Oracle、Druid、ClickHouse 等跨数据源计算；
- 3、主导研发分布式调度平台（提供高性能、高可用、可控的自研调度平台）；
- 4、开发数据资产系统（提供数据质量计算和数据资产化管理能力）；
- 5、基于业务线需求，建设实时数仓平台，并将平台推广到其他业务线（Flink、Kafka、Druid、Hbase）；
- 6、浏览器业务线数据仓库体系建设；

- 1、参与大数据平台架构的规划和设计；
- 2、基于 Apache Canal 建设 RDBMS 数据接入系统；
- 3、带领团队建设大数据集群和数据监控平台，提高集群稳定性，确保数据产出及时性，建设数仓权限体系确保数据安全性；
- 4、对离线数仓计算体系进行优化，提高数据效率；
- 5、主导开发网约车数据上报同步系统；
- 6、将实时计算作业模板化，降低程序开发时间成本，提高效率；日常实时计算开发工作；

- 1、负责数据平台部 hadoop 集群的升级及集群的日常维护和优化工作；
- 2、搭建公司新一代流量日志收集系统，采用分布式 flume 方式接收数据；
- 3、采用 kafka+storm+hbase 架构开发实时统计报表；
- 4、对公司原有实时集群进行维护并优化；
- 5、对原有 etl 非流量部分数据作业全迁移到 hive上，并完善 hive 的权限机制；
- 6、参与安装和维护 ETL automation 数据仓库工具；
- 7、研究 spark、storm 等新技术。

## 项目经历

### 内容:

数据开发平台是基于管理域业务，为解决数据工程实施过程中开发效率低、维护成本高、底层暴露缺乏产品化、发布管理困难等一系列问题所开发建设的数据产品；该产品基于开源Apache DolphinScheduler Alpha 3.0.0版本进行二次开发；

在任务开发、测试、发布管理、版本管理、维护、安全、展示等功能上做了优化改造，项目团队6人；

### 业绩:

承载 hr 、财经、法务等领域，包括 HR系统员工花名册、编制、人效、激励、干部管理报告、财经管理报告、财经EBS报告、财经收入报表、费控、员工成本等；

任务4000+，工作流 300+，日调度万次以上/日；任务稳定性 99.97%+；

### 内容:

基于集团管理域数字化转型大背景，建设稳定的、高效的、安全的数据存储、传输、计算、作业调度底座支持能力产品；以 Hadoop、Hive、Spark、Datax、StarRocks、Flink作为数据存储与计算，和DolphinScheduler 作为调度的整体技术架构；

### 业绩:

1. 能力分布：构建了具备离线、实时计算、数据传输、调度能力的管理域较为完善的、安全的、稳定的数据底座；
2. 业务规模：承载了管理域百TB以上级数据，5000+表；
3. 架构收益：采用 Apache 体系+Starrocks的架构，相对纯Apache体系数仓，降低了学习、迁移成本，提高了计算效率；
4. 成本管控：通过作业治理管控，在2022至今业务和作业量增长 3.5 倍的前提下，仅扩容 20%计算资源；

### 内容:

随着数据业务不断发展后所面临的管理难、使用难、质量差、以及“近在咫尺而不知何用”等问题而自主研发的数据管理平台；包括数据资产管理、数据质量管理、数据血缘管理（基于 alibaba Druid+ Nebula ）等子系统；项目团队10人；

**业绩：**

产品收益：

- 1.数据质量：建立质量规则 1700+，发现质量问题 1400+次，问题解决率 93%+；
- 2.数据资产、血缘：管理 6000+表，通过任务与表血缘解决任务执行顺序冲突 300+；

项目管理：项目以项目群的形式运作使我在技术架构和管理方面收益良多；

**360—实时数仓**

数据架构师

2019.11-2020.10

**内容：**

基于 Flink 引擎、数据仓库理论，进行业务抽象，以通用平台化形式对外提供实时计算能力，用户可通过平台快速且灵活配置实时计算任务，无需开发代码；针对计算流程建立较为完善监控体系，确保实时数据稳定性；

**业绩：**

前期实时数仓建设技术、功能架构设计和建设；项目开发进度管理和技术难点攻关；  
成功将实时数仓计算平台落地到多个业务线，并建立较为完善的监控体系，实时数据规模百T/日；  
底层通过使用 Flink 计算引擎，数据产出由原来的T+1优化到分钟级，新的实时需求由原来的需要3天+到现在通过配置化3小时- 完成，大大提高工作效率；

**360--数据资产平台**

数据架构师

2018.04-2019.09

**内容：**

数据治理项目的目标是将零散的数据统一收集归纳到系统，对管理的数据通过质量稽核、质量评估、数据清理、安全处理、数据历史回溯、数据血缘分析等对数据进行全方面管理，统一数据出口，对内可用于数据集成、Quicksql 计算对外进行标准数据输出；

**业绩：**

前期项目技术、功能架构设计和建设；项目开发进度管理和技术难点攻关；  
数据质量计算和元数据采集两大模块的开发工作；

**360—分布式数据计算平台**

项目经理

2017.07-2019.09

**内容：**

基于 java 语言开发的分布式数据计算平台，同时支持实时和离线计算，支持结构、半结构、非结构化数据的处理，用户可自由选择拖拽和编写 sql 进行不同的方式配置计算任务，计算平台 ETL 单元对 Flink和 Spark 算子进行封装并进行接口抽象，通过用户配置的业务流程对 ETL 单元进行编排执行；为实时离线数仓建设提供计算能力；支持 HDFS、Hive、Mysql、Kafka、Hbase、Oracle、Druid、ClickHouse、FTP、Redis、ES 等跨数据源计算；

**业绩：**

主要负责平台的技术选型、架构设计和核心功能的开发，调研实时数仓需求，基于平台实现实时计算能力，并接入信息流、浏览器、国际卫士业务建立实时数仓；

**360--分布式调度平台**

数据架构师

2017.07-2019.07

**内容：**

系统架构架构上采用 Master/Slave 分布式结构,由 Admin、Executor和 Seed 三部分组成。Admin、Executor和 Seed 间采用异步通信方式

核心优势：该系统采用 Master/Slave 分布式设计。高容错性：Master/Slave 都为多台机器，提高任务稳定性。高并发：每台 Executor 可以同时最大并发50多个任务(可能根据任务占用 executor 内存、cpu 有关)。任务容错：包括任务状态检查、任务重新调度、机器下线重新调度任务等机制。任务恢复：系统 Executor 升级期间不影响任务执行。类型任务：根据业务需求指定不同的计算模型(支持离线、实时任务)。任务路由：提供任务随机执行和指定机器执行。系统监控：可以监控 Executor 机器内存、cpu、load 、磁盘等指标，并采用相应的任务调度策略。任务拓扑：提供任务依赖拓扑，支持拓扑暂停、拓扑恢复等功能。模型升级：支持模型下发和模型自我检查机制，可以对机器模型进行升级。

**业绩：**

- 1、参与了分布式调度系统需求调研,技术选型,和核心功能开发，包括但不限于新功能节点安全上下线，提高了调度系统的运行稳定性在上下线期间不影响作业正常执行；
- 2、主要负责了调度系统的 excutor 健康度检查、调度策略多维度优化等的开发、测试，使任务能够更合理的进行资源分配；
- 3、负责调度端到端系统的节点监控、任务监控、权限管理、统一登录模块开发及测试，完善端到端系统功能，方便操作管理调度平台和任务；
- 4、负责调度端到端系统任务管理和资源管理模块的功能优化；以及各模块帮助文档完善；

## 教育经历

北京经济技术学院	本科	计算机应用技术	2005-2009
----------	----	---------	-----------