

## 翟士丹 (mail@stanzhai.site)

---

自我介绍	<p><b>7年+工作经验，现居北京</b></p> <p>专注大数据领域，Apache Spark Contributor，有丰富的大数据BI系统及中台构建经验；熟悉SQL、Web技术栈及Linux；工具控；擅长系统架构设计，热爱编码，喜欢探索未知领域；11年开发经验，4年研发团队管理经验（目前带队30多人）。</p>
教育经历	<p><b>曲阜师范大学</b> 软件工程，2008-2012，本科</p>
职业相关	<p><b>方向：</b> 大数据平台、分布式存储/计算；后端开发；基础平台</p> <p><b>技术栈：</b> Spark、HBase、Hadoop、Docker、Nodejs、OpenResty</p> <p><b>编程语言：</b> Scala、Python、Go ...</p> <p><b>编辑器：</b> Vim</p>
工作经历	<p><b>海致：</b> 架构师，技术副总监 2014.12，SaaS BI(BDP)，中台建设</p> <p><b>搜狐·畅游：</b> 研发组长，2012.07，通用数据分析平台；爬虫、舆情分析系统；信息安全审计系统</p> <p><b>浪潮：</b> 实习，2011.12，空军装备研究院，作战指挥系统（C++）</p>
项目经验	<p><b>智能中台(数据+计算+业务)服务公安、金融</b> Spark,ES,JanusGraph,GeoMesa 2018-至今</p> <ul style="list-style-type: none"><li>- 基于标准SQL抽象了算子体系(关联、聚合、追加等)，实现了高度灵活的可视化数据建模系统</li><li>- 基于Structured Streaming实现通用流式计算平台（支持Kafka、可视化拖拽式建模）</li><li>- 实现主题库（对应数仓的DM层）管理模块，支持行业业务抽象的封装</li><li>- 实现通用标签计算引擎，为特定对象自定义标签规则计算标签，支持知识图谱应用</li><li>- 关系库设计实现</li><li>- 异构任务调度平台构建</li><li>- 业务算子系统开发，数据发布接口实现</li><li>- 机器学习平台构建</li><li>- 知识图谱分析平台实现</li><li>- GIS时空分析平台实现</li></ul> <p><b>BDP商业数据分析平台</b> Spark, BI 2014-2018 <a href="https://www.bdp.cn">https://www.bdp.cn</a></p> <ul style="list-style-type: none"><li>- 基于Spark实现高性能、任意多维度分析的OLAP查询引擎(每日20W查询任务，平均2s)</li><li>- Spark性能优化、稳定性改进及Bug修复，任务超时、过载保护及自愈机制（Spark 0.9 -&gt; 2.3)</li><li>- 基于分区表实现MaterialView，支持增量更新，改进了超大表建模及查询速度</li><li>- 扩展了SQL语法，SQL层面支持机器学习处理流程、DataFrame源码层面的任意数据处理</li><li>- 重构了Hive UDF在Spark的注册机制，实现UDF随应用启动自注册，极大简化了开发上线流程</li><li>- 自研Patch算法，基于追加写实现表数据的Insert、Update、Delete的数据接入服务</li><li>- 实现基于版本的并行DAG数据建模任务调度系统（日均调度任务50W）</li><li>- 基于Go实现基于优先级及Quota机制的查询调度模块（高性能请求转发、低内存占用）</li><li>- 基于PhantomJS实现高质量图表、仪表盘后端渲染导出服务</li><li>- 构建服务端Sandbox环境，实现用户自定义图表JS代码的异常检测（死循环代码、超内存消耗）</li><li>- 基于Redis实现多级查询缓存模块，改进了查询缓存命中率，优化图表查询体验</li><li>- 实现了BDP数据平台底层无缝切换至阿里云平台（HDFS+Spark -&gt; ODPS+ADS）</li><li>- 基于Node.js+Knockout实现运维管理平台（数据血缘，SQL IDE，任务监控等）</li></ul> <p><b>MOPlatform数据分析监控平台</b> ASP.NET MVC4, RESTful, MVVM 2013/10-2014/12</p> <ul style="list-style-type: none"><li>- 建立高度自定义的报表模块，用于数据可视化，通过灵活的报表配置，自定义报表内容</li><li>- 基于正则，实现通用的日志解析工具，将日志文件处理为带Schema的结构化数据</li><li>- 实现专用ETL模块，抽象日志元数据生成可以直接分析入库的数据集，统一数据源管理</li><li>- 构建数据分析模块，实现常用的阈值算法（极值，四分位等）对异常数据进行分析报警</li><li>- 实现基于Phantomjs的服务端图表生成模块，保证最佳的图表生成质量</li></ul>

### **舆情监控系统-天兵** C#, Node.js, Phantomjs 2012/07-2014/12

- 实现通用爬虫系统（队列，排重，正文提取，存储，分布式，任务管理及任务指派等）
- 设计并实现舆论分析系统（贝叶斯分类，实时图表，舆情数据可视化，语音及邮件报警等）
- 开发Web信息采集器，带领团队完成特定信息的采集和分析，为竞品分析提供基础数据源
- 实现通用网页正文提取算法：<https://github.com/stanzhai/Html2Article> 被一些公司采用

### **济宁市基层党组织信息管理系统** 2011/10-2011/12

本系统是结合了GIS的信息管理系统，地图部分采用ArcGIS开发，采用WebService与网站通信获取党组织信息。实现了信息录入，信息管理，复杂的信息动态查询，查询结果导出Excel，管理员权限管理及分配等核心模块。  
技术实现：ASP.NET MVC3, Sql Server, NHibernate, ActiveRecord, Spring.NET, Node.js, Lucene.net, Log4net, Nginx

### **Android五棍棋游戏** 2010/10-2010/11

五棍棋是一款深具中国传统文化特色的棋类游戏；  
独立负责Android版整个项目的开发（1W+代码量）；  
采用AndEngine游戏引擎实现，基于SOLID设计原则，分离了走棋算法、游戏交互、UI渲染模块；  
通过剪枝优化了AI核心算法，实现了蓝牙对战、人机对战；  
参加移动MM百万青年创业大赛，全国入围(比赛延期出了幺蛾子)

### **通用浏览器插件框架** 2009/06-2009/09

目前(2009年)通用浏览器的引擎基本分为三种类型：WebKit、Gecko以及Trident等。这些不同的浏览器都有不同的插件标准，开发人员不得不为各种不同的浏览器开发同一种类型的插件。采用了面向对象设计思想，使用C++语言定义一套独立的插件标准，为不同引擎的浏览器提供了相应的适配器。通过使用本浏览器插件框架，开发插件的用户利用我们的标准只需开发一次插件，就能实现在多个浏览器上使用。降低浏览器插件开发成本，提高了开发效率和易维护性。

- 基于Win平台C++动态库、多线程技术，抽象了接口规范（load, exec, unload核心函数）
- 针对IE浏览器插件、mozilla扩展开发库规范，实现了不同平台的框架容器
- 基于接口规范实现了俄罗斯方块游戏demo，作为插件直接运行于不同浏览器
- 本作品参加了山东省齐鲁软件大赛，并获一等奖。

其他

**GitHub:** <https://github.com/stanzhai>

**专栏:** <https://zhuanlan.zhihu.com/bigdata-spark>

**Blog:** <https://stanzhai.site>

**折腾:** <https://www.zhihu.com/question/51314788/answer/144324074>