分布式网络爬虫架构实现

【项目愿景】本系统基于智能爬虫方向对数据由原来的被动整理到未来的主动进攻的转变的背景下，将赋予”爬虫”自我认知能力，去主动寻找”进攻”目标。取代人工复杂而又单调的重复性工作。能够实现在人工智能领域的某一方向上独当一面的作用。

【项目说明】为了能够更好理解优秀框架的实现原理，本项目尽量屏蔽优秀开源第三方jar包实现，自定义实现后再去择优而食，形成对比。

【设计原则】依据形式跟进新技术，不断积累在不同业务场景下对各个开源技术的驾驭能力

【架构及核心模块实现】

1. 爬虫采用心跳机制——便于对任务状态进行监控和减轻任务调度器负担。
2. 爬虫获取资源后统一格式后传输给处理模块——更加有效利用计算机资源。
3. 依据任务数据特征不同动态选择数据存储技术——为框架可扩展性留出空间。
4. 利用对Map的封装实现PO容器(单例模式)——存放系统中具有唯一性的键值对参数，方便调用。
5. 数据源和线程管理上采用连接池技术——以便资源的合理利用。

【分布式爬虫架构原理图】

