Darwin 技术白皮书

V5.0.7

2015 年 06 月

1、产品简介

Darwin 是基于 Hadoop 的企业级应用支撑平台，通过组件化、可视化的方式，帮助用户快速实现大数据的存、查、分析和管理，进而提升企业对市场的响应能力，降低企业大数据应用的实施成本。

Darwin 底层基于 Spark 内存计算技术，性能可提升 10-100 倍。Darwin 易平台的目标是让 Hadoop 和 Spark 的使用更简单，让大数据从汇集到被应用更简单。

2、产品背景

在信息化技术高速发展的今天，企业的信息化，逐渐由内部管理，转向借助互联网和移动互联网技术，适应业务发展的需求。与此同时，企业面临着日益膨胀的各类数据。

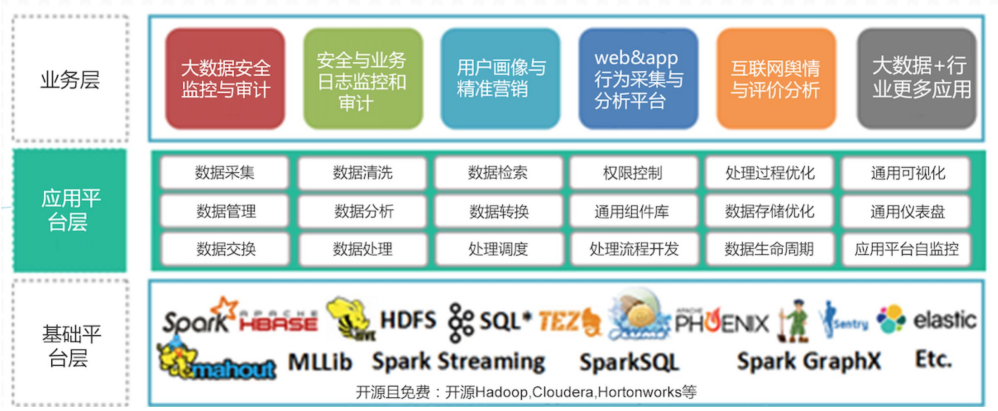
大数据产品的需求应运而生，Hadoop 便是其中的代表型开源产品体系。然而 Hadoop 体系的产品产生于互联网，也是最先在互联网企业得以使用，对企业来说太偏技术，大部分企业不知道如何去运维、使用这样的一个平台，即使是目前流行的商用 Hadoop 产品，也仅仅解决了“运维”这部分。

易平台产品的定位，重点关注客户对 Hadoop 使用中，最为困难的 4 个问题：

1. 基于 Hadoop 的软件研发困难
    a) 海量数据的分析需求快速满足困难
    b) 懂 MapReduce、Spark 技术人才匮乏
2. 数据及数据处理的管理和维护困难
    a) 大数据平台需要满足数据的组织、管理和安全控制
    b) 数据处理需要按需/周期的调度和监控
3. 海量数据处理效率问题
4. Hadoop / Spark 的运维管理问题

3、产品定位

Darwin 平台定位于应用支撑平台，其介于底层 Hadoop 基础平台和业务平台之间，起到承上启下的作用，定位层级如下图所示：



Darwin 平台实现了数据的实时 / 批量的数据汇集、数据清洗、关联、分析处理组件库、数据交换、图形化数据处理流程开发、数据管理、数据服务（海量数据检索，大数据库，存储）和数据可视化等涉及到企业数据全生命周期的功能。

4、技术指标

Hadoop 集群 x86 服务器共 7 台	
硬件配置	CPU：Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz *2，8 core per CPU 内存：128GB　硬盘：1000G SAS * 4
服务器数量	7 台（1 台 Namenode，6 台 Datanode，2 台 Darwin 复用 Datanode）
操作系统	CentOS 6.4
安装软件	CDH 5.3.3 发行版 Darwin 5.0.1 发行版
备注	Darwin 与 Hadoop 集群复用 2 台机器

大数据平台性能指标		
测试数据项		
1T 原始数据，每条数据 9 个数据项，200 多字节，共约 50 亿条		
类别	测试内容	性能指标 （多次测试平均值）
数据加载	1T 数据单机加载到 HDFS	耗时：149 分钟 加载性能：112M/s
	1T 数据导入 BigDB 数据库	耗时：132 分钟 入库性能：126M/s，62 万 EPS
	1T 数据建立全文索引	耗时：347 分钟 索引性能：50.4M/s，24 万 EPS
SQL 查询	任意纬度查询，简单查询（返回结果总数 1700 万条以内）	响应时间：小于 1 秒
	任意带主键多纬度条件查询（800 并发以下）	响应时间：100 毫秒~1 秒
	任意纬度查询，复杂联表聚合查询（扫描数据 1000 万条以内）	响应时间：小于 3 秒
全文模糊查询	全文检索查询，简单模糊查询	响应时间：小于 100 毫秒
	任意 3 个纬度聚合检索	响应时间：小于 3 秒
数据处理	分组统计	耗时：10 分钟 性能：1.68G/s，820 万 EPS
	ETL 清洗转换	耗时：26 分钟 性能：641M/s，320 万 EPS
备注		

5、运行环境

单机版：

软件：

推荐运行于 CentOS 6.4 64bits 系统

硬件：

服务器硬件项	推荐配置	描述
CPU	2*8Cores＋	
内存	64GB＋	
硬盘	2TB＋	5 块 500GB SATA 硬盘，做 raid5
网卡	双万兆网卡	

集群版：

推荐集群：

开源社区 Hadoop 集群 cloudera 集群

Darwin 部署环境：

单机版＊2+

云服务版：

请直接登录产品页面 http://121.42.25.148:8000 ， 联系 joycexu@stonesun-tech.com 申请试用账号后即可体验云服务版。

6、产品优势

- 快速部署安装

  对已有/建设中/无 Hadoop 的企业，可快速安装部署 Darwin，并接入 Hadoop/Spark。

- 零学习成本

  人性化的交互设计，使各类人员能够快速使用 Darwin，完成数据处理的需求。

- 基于 SQL 的分析过程，简单易用

  易平台内置各类丰富的组件，可完成多种数据的汇集、处理、导出。并定期提供（免费 / 付费）打包组件或工作流，可完成更复杂的数据处理过程。屏蔽底层 Hadoop / Spark 的复杂性，通过编写 SQL 轻松完成数据分析与处理，极易操作的任务管理拖拽式完成任务的依赖与自动调度。

- 服务一体化

  多年运维 Hadoop 的技术团队提供专业服务。

7、特别适合场景

场景一：海量数据实时查询

客户诉求：

对于有大量历史明细数据的行业（如交通、电信），希望通过对海量明细数据实时查询，获得最细粒度的数据，用以上层业务系统的支撑。

Darwin 提供的解决方案：

1、实时 / 离线获取明细日志，经过流程处理与清洗后写入搜索引擎。

2、通过灵活的仪表盘配置，创建专用检索仪表盘，实现快速实时查询的需求。

客户价值：

- PB 级数据秒级查询性能
- 业务快速响应能力

场景二：在线交易系统运维监测

客户诉求：

在线交易系统平时交易量平稳，但有市场活动时，业务量会突增，虽然可以通过服务器监控产品，了解服务器的负载，但难以定位是哪个业务环节（负载均衡→web 服务器→WEB 应用→DB 层→LDAP 等）造成的瓶颈，希望能通过对各类业务日志的监控，将日志实时汇集并关联分析，将各个层次的异常一并展现，以便及时发现系统的瓶颈。

Darwin 提供的解决方案：

1、实时获取各类日志，并对被获取日志的服务器零性能影响。

2、通过灵活配置实时检索 / 各类分析（支持 Spark－SQL 与自定义程序）实现多维及关联分析

3、各类检索 / 分析结果通过仪表盘实时呈现，便于问题的分析与定位

4、告警信息及时通过邮件或短信发出，便于问题的及早发现。

客户价值：

- 业务日志关联分析

- 运维问题快速分析

场景三：互联网站点/APP 企业网站指标分析与用户画像

客户诉求：

企业客户逐渐将业务从线下往线上延生，并已逐渐意识到数据对企业的价值，需要将互联网站点或 APP 的用户行为收集到企业内部，作为业务分析使用。企业期望有个自助快速分析平台，可以对站点/APP 的各项指标进行分析。如常规指标：PV, UV, Bounce Rate, 停留时长及其趋势；对关键页面、业务、推广的 Landing page 的 UV、PV、来源等进行分析。除此之外，通过在积累的日志中，建立用户的统一视图，分析用户的兴趣喜好，并辅助企业更好的服务于用户。

Darwin 提供的解决方案：

1、通过 js 插码（智通－易分析插码模块），或 SDK（智通－易感知产品模块）收集互联网站点/APP 的行为日志

2、Darwin 配置数据源，周期/实时将行为汇集

3、Darwin 配置检索 / 分析流程，对数据进行各类分析，无需开发

4、各种分析结果以合适的展现形式，加入到仪表盘 / 告警，完成分析

客户价值：

- 获得原始日志，并获得百度统计 / 友盟同等分析报告
- 积累用户画像，便于精准营销

场景四：DW 大表的迁移和数据分析

客户诉求：

随着数据量的增多，越来越多企业此前建设的 DW 中，大量表数据量已经达到关系型数据库单表的极限（1~10 亿条/表），而此前建设 DW 时，对数据的分析均采用 SQL 及 SQL 连表方式完成，但目前很多客户 DW 忙于各种批处理 SQL 已经无法满足客户需求。

Darwin 提供的解决方案：

1、评估给 DW 带来压力的表，及表操作 SQL，找到关键的大表及分析需求

2、通过 Darwin 配置数据源，从大表导出数据到 Hadoop

3、通过 Darwin 配置检索 / 分析流程，对表进行 SQL 操作

4、分析结果提供 API 方式导出

客户价值：

- 快速搭建大数据平台
- 解决 DW 大表性能瓶颈