1. **行业背景**

大数据是信息化发展的新阶段。随着信息技术和人类生产生活交汇融合,互联网快速普及，全球数据呈现爆发增长、海量集聚的特点，对经济发展、社会治理、国家管理、人民生活都产生了重大影响。在过去的2017年里，大数据在政策、技术、产业、应用等多个层面都取得了显著进展。  
  
在政策层面，大数据的重要性进一步得到巩固。党的十九大提出“ 推动互联网、大数据、人工智能和实体经济深度融合”，习近平总书记在政治局集体学习中深刻分析了我国大数据发展的现状和趋势，对我国实施国家大数据战略提出了更高的要求。  
  
在技术层面，以分析类技术、事务处理技术和流通类技术为代表的大数据技术得到了快速的发展。以开源为主导、多种技术和架构并存的大数据技术架构体系已经初步形成。大数 据技术的计算性能进一- 步提升，处理时延不断降低，硬件能力得到充分挖掘，与各种数据库的融合能力继续增强。

1. **产品背景**

数据分析平台的建设流程基本上可以从数据分析流程来理解，比如数据采集、数据整合、数据加工、数据可视化等等，一般的大数据平台都会包括这些流程，叫做一站式大数据平台。

**中琛魔方**是深圳中琛源科技股份有限公司（股票代码839294）旗下大数据服务品牌，企业秉承着“需求决定市场、市场引领技术、技术指导市场”的理念，在企业销售自动化应用和移动营销综合推广领域提出多个先进的专业解决方案，获得“双软”企业认证，AAA级诚信企业，2017年获得“2016年度深圳市软件业务收入前百家企业”。  
  
2017年中琛源开始积极开拓对外的大数据业务，依托自身多年的大数据应用技术和实践，迅速推出了大数据平台类产品以及行业解决方案，为众多企业发展带来了新的增涨点，同时为大数据行业发展提供了实践蓝本！  
  
未来，中琛魔方将在大数据领域持续深耕，力求将最好的技术、最好的产品带向市场，帮助企业运营发展向数据化、智能化转变！

**北京亿信**华辰软件有限责任公司（ESENSOFT，亿信华辰）成立于2006年10月。亿信华辰是中国专业的智能数据产品与服务提供商，深耕商务智能和大数据领域15年，着眼于打造数据全生命周期的智能化产品线，致力于帮助企业和政府解决数据应用难题，实现企业生产力和政府治理能力的数字化转型，让数据驱动进步。  
  
实力雄厚，发展迅速总部位于北京中关村东升科技园，并下设广州、上海、武汉、西安，成都、济南等多个分公司。公司现有800多名员工，技术人员是核心力量，占员工总数的80%。在所有员工的共同努力下，营业收入以31.28%的复合增长率持续增长。  
  
  
专业的团队，用心的服务亿信华辰具有300+大数据分析师和实施顾问、200+研发专家所组成的专业服务团队，服务人员遍布全国各省市。亿信服务团队一直秉承“快速响应，精益求精”的服务理念，以“诚心，专心，用心”为基本服务原则，力求为客户稳定持续的提供“放心，省心，舒心”的服务体验。  
  
产品技术先进，项目成功率高作为国内拥有15年历史的专业BI厂商，亿信华辰秉持“让数据驱动进步”的理念，专注于商务智能领域和大数据相关产品的研究和实践。自成立起，每年以超过销售收入45%以上的研发投入，围绕数据全生命周期打造多款重量级核心智能产品，并以几乎100%的项目成功率，取得了远超同业公司客户满意度。  
  
市场覆盖广亿信华辰基于核心大数据分析产品，服务了8000多合作客户，覆盖100多个细分行业。 50多个国家级部委采用我们的产品完成监管、普查、抽查、统计等应用场景的数据采集与分析；100余家银行采用亿信产品完成统一数据平台的建设；100家以上的全球500强在华企业采用我们的产品实现商务智能系统、完成集团业务的精细化管理；超过1500家以上的ISV、软件集成商牵手亿信在各自优势行业完成数据分析项目的实施和交付。  
  
PRODUCT STRUCTURE亿信产品架构经过15年的技术沉淀，公司围绕数据全生命周期打造了多款重量级核心智能数据产品，其产品线涵盖了数据采集、数据处理、数据存储、数据分析、敏捷分析、数据挖掘等多个领域。亿信旗下所有产品都是基于公共平台进行研发，因此每款产品都可独立应用，同时也可进行产品间的互联互通与无缝集成。

1. **产品介绍**

ABI（ALL in one BI）是亿信华辰历经十五年匠心打造的国产化BI工具，技术自主可控。它打通从数据接入、到数据建模与处理、再到数据分析与挖掘整个数据应用全链路，可满足企业经营中各类复杂的分析需求，帮助企业实现高效数字化转型。



**产品架构**

中琛魔方提供大数据采集、融合、分析、可视化等全套产品服务， 包含“元数据管理、ETL、标签管理、任务管理、网络数据采集、 DMP”等大数据应用的全方位产品。

产品架构图



1. **产品功能对比**

**一、亿信ABI一站式数据分析平台**

可连接多种类型的数据源，包括：关系型数据库，分布式大数据库，文件数据源（Excel、txt、DB）、接口数据源等，支持跨源跨库的分析  
数据处理  
平台内置了数仓实施工具，其中丰富的处理转换组件，通过拖拽式的流程设计，实现了数据抽取、清洗、转换、装载及调度，用于帮助政府和企业构建数据仓库，完成数据融合，提升数据质量，服务数据分析  
指标统一管理  
支持将分散在不同系统的各类指标集中管理，并用统一的标准进行约束，能清晰展现用户指标一览表和指标运算逻辑，帮助企业统一指标口径，构建完善的指标体系  
补录填报与工作流  
平台提供对分析表格进行数据回填设置，对缺失的数据进行补录，也可以设置全新的填报表单用于录入数据。同时，填报可以与工作流结合使用，实现数据填报补录后的业务流程审批  
可视化的报表设计器  
B/S架构部署，纯WEB应用，用户无需安装任何插件，只需通过拖拽的方式即可组合内置组件，轻松实现图表展示及钻取的应用  
中国式复杂报表应用  
采用类似Excel的操作方式，可直接添加斜线表元、多级表头、合并表元等样式，无需依赖于IT人员，也能够快速、高效的设计中国式复杂报表  
精细化的权限控制  
支持资源级别的权限控制，还提供了行列粒度的数据权限控制，可从行列级别去精细保障数据应用的安全性，防止越级访问  
支持移动应用  
支持通过移动端实现数据的采集与分析应用，报表一次制作即可在手机、ipad、大屏等多个终端设别上进行展示







**二、中琛魔方元大数据平台。**

**1.元数据管理子系统**  
元数据(MetaData)狭义的解释是用来描述数据的数据，广义的来看，除了业务逻辑直接读写处理的那些业务数据，所有其它用来维持整个系统运转所需的信息/数据都可以叫作元数据。比如数据表格的Schema信息，任务的血缘关系，用户和脚本/任务的权限映射关系信息等等。  
  
元数据管理系统是具有完全的自主知识产权的产品，从架构到代码均针对元数据数据库及模型进行针对性设计和优化，能适应企业复杂的IT异构环境，可快速部署企业元数据管理系统，提供完善的元数据管理功能，满足超大量数据处理分析需求，支持全中文内核元数据管理界面。本系统能帮助企业在生产运营中更高效的挖掘和使用数据，让平台管理人员能更加有效的做好系统的维护管理工作。



1.1元数据模型  
元数据管理产品的元模型符合 CWM 规范，同时提供内置数据字典、代码管理等技术元  
数据和指标、业务规则等业务元数据，符合企业数据仓库环境的各类元数据管理模型。元模  
型可由用户自定义扩展，满足客户化元数据的需求。有元模型查询、元模型增加、元模型修  
改、元模型删除、添加关系、权限设置等功能

**2.ETL 子系统**

ETL，是英文 Extract-Transform-Load 的缩写，用来描述将数据从来源端经过抽取（extract）、 交互转换（transform）、加载（load）至目的端的过程，为多源、多种类、多格式数据提供灵 活、可靠、高效的海量数据采集、清洗、转换、与加载能力。 中琛魔方 ETL 子系统提供数据标准管理到元数据管理再到标准化 ODS 库的建立的自动 化功能，由用户录入或导入业务标准数据（业务属性），平台自动生成元数据（数据库结构 信息），并自动创建数据库表。从而实现业务标准、元数据、数据库之间的一致性。利用 ETL 工具，建立外部数据和标准 ODS 数据库之间的映射关系，自动比对数据之间的差异，形成 对照表，再利用 ETL 工具的转换功能进行数据转换配置，并执行转换和加载；在标准 ODS 数 据库中，实现各业务数据之间建立关联关系，提供可视化界面，供业务人员按业务逻辑建立 关联关系。



2.1任务中心 任务中心支持管理所有 ETL 子系统中的数据处理任务，分为转换任务与调度任务两大 类，支持分类搜索任务关键字，按照任务不同状态进行筛选，支持图形视图与列表视图展示 全部任务，对所有任务进行新增、删除、修改与刷新状态。

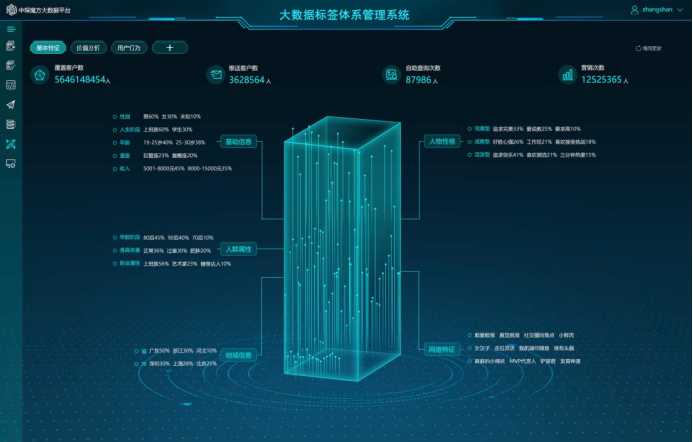
2.2 设计平台 设计平台包含新建或修改转换（Trans）任务与新建或修改调度（Job）任务。 转换任务支持大数据 HDFS、数据库表、字段、Csv 文件、Excel 文件、文本文件等多种 方式输入输出数据，支持 20 多种转换方式进行自定义设置转换，支持脚本、转换、ES 批量 加载、流查询、Switch/Case、分组、数据监测等多种功能。

2.3 资源中心 资源中心支持数据系统、服务器、服务器集群、Hadoop 集群、Spark 引擎、执行引擎、 文件模板管理等多种设计和管理功能

1. 全网数据采集子系统 全网数据采集子系统由多个组件构成，通过组件组合进行抓取流程创建，通过参数配置 进行抓取规则定义，以简单灵活的方式快速生成数据采集任务，方便各类网站信息的采集。 数据抓取支持“抓取方式、模拟消息头、cookie、代理 IP、token、抓取间隔”等配置。 数据抓取规则，支持增量获取和指定获取。数据解析支持元素提取和正则提取。数据保存， 可选用 csv、txt 以及各类型数据库保存等。



1. 标签管理子系统 标签管理系统是一款基于中琛源全品类基础维度 ATNT （All-Type-Natrue-Tag）模型、先 进特征挖掘和信息可视化技术，构建用户、产品、渠道等实体画像，并实现智能关联的产品。 它充分整合公司多触点、全渠道数据，能够支持智能推荐和精准营销等上层数据应用



1. 算法子系统 算法子系统主要由由“公共算法库、自定义算法、AB 测试工具”等模块构成。内置了 常用算法和模型，并支持算法编写和模型训练功能，同时提供测试数据方便验证。 其中自定义算法允许通过可视化方式进行算法自定义，并支持在线调试以及训练过程数 据可视化，对于深度学习训练过程的特征提取或梯度下降更是直观可见。AB 测试工具能够 极方便的进行不同算法模型的效果对比分析



**社区版：**指开源社区开发、编译出来的版本，通常遵守Apache 2.0或类似的协议，只要遵守协议即可免费用于各种商业目的。如Apache Hadoop。

**发行版：**一些厂商在开源代码的基础上，修复bug、加入自己的优化、整合管理工具，并提供相应的技术支持和培训认证等，如此产生的版本称为[发行版](https://www.zhihu.com/search?q=%E5%8F%91%E8%A1%8C%E7%89%88&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":2144027639}" \t "_blank)。通常，发行版只有用于研究目的时可以免费试用，用于生产时必须付费。如Hadoop领域的发行版有Cloudera的CDH、Hortonworks的HDP（两家已合并）、MapR的MapR、国内华为的FusionInsight等。

还有一些组件，其社区版和发行版是由同一家公司维护的，如Elasticsearch就是由名为Elasticsearch B.V.的公司开发，并同时提供免费版本和不同级别的收费版本。

社区版优点：完全免费 完全开元 社区活跃

缺点：版本管理比较麻烦 缺乏部署、运维等工具

发行版优点：版本管理清晰 兼容性、稳定性有提升 工具丰富

缺点：企业版要钱 部分或全部不开源

1. **产品优势**
2. **亿信ABI一站式数据分析平台**

1.高效的数据处理  
  
一键连接百种数据源，包括文件数据源及接口数据源。多样的建模方式和丰富的数据处理组件高效便捷地实现了数据仓库的构建及数据的抽取、清洗、转换等操作。

2.智能分析引擎  
  
支持类Excel表达式语法、内置语法解析器、SQL翻译引擎，遵循DW/BI的经典理论，形成亿信ABI独有的智能分析引擎技术。

3.高性能分析  
  
自主研发的适合大数据量计算、跨源跨库高速分析场景的高速缓存引擎，具备高效计算能力，数据查询与分析的速度远高于普通的数据库（如Oracle、Mysql等），千万数据的分析计算仅需毫秒级完成，完美解决大数据时代的性能难题。

4.集成开发API  
  
提供上千个API接口，方便用户扩展，便于与第三方系统集成，缩短项目实施周期，降低成本。

**二、中琛魔方元大数据平台。**

1. 强大的数据处理能力  
   具备流式处理、内存计算、批量处理、数据检索、数据统计、图计算、机器学习、图像识别、语音处理、自然语言处理能力。

2. 丰富的异构数据支持能力  
支持各种关系型数据、非关系型数据（Json、GIS、视频、语音、图片等）的存储和处理分析。

3. 简单易用的数据分析能力  
无需技术背景，数据分析人员即可通过可视化界面轻松完成各类业务分析，同时也可使用内置算法模型，简单操作即可实现对数据深层价值的挖掘。

4. 高效快捷的部署对接能力  
系统支持快速部署，迅速打通企业内部的各类数据源以及多种架构的业务系统。并可根  
据具体需求进行功能扩展，灵活方便，无须组建专门的技术团队。  
5. 全面可靠的安全保障能力  
平台提供网络安全、应用安全、数据安全、存储安全和账号安全等多重安全保障，通过先进的安全理念和系统架构确保用户源数据安全。实现用户数据标签化，保障企业数据安全性，真正实现数据流通的可能

**学术论文**

1. 张文明.基于云计算的Hadoop大数据平台挖掘算法及实现研究[J].无线互联科技,2021,18(19):110-111.
2. 冯国礼,李蓉,吴双,柴育峰. 一种改进FP-growth的电网大数据平台危险源分析算法
3. 刘佳耀,王佳斌.Slope One算法的改进及其在大数据平台的实现[J].计算机工程与应用,2020,56(01):83-91.
4. 陈阳雪. 基于大数据平台的MOOC混合推荐算法的研究及应用[D].杭州电子科技大学,2017.
5. 杜小芳,陈毅红,王登辉,卢思阳.大数据平台上的并行CART决策树算法[J].西华师范大学学报(自然科学版),2021,42(02):196-201.DOI:10.16246/j.issn.1673-5072.2021.02.015.
6. 胡婕. 基于大数据平台的贝叶斯网络增量学习方法研究与应用[D].北京邮电大学,2017.
7. 安建民,周一波,彭送庭.客户分析大数据平台方案研究[J].现代商业,2020(16):128-130.DOI:10.14097/j.cnki.5392/2020.16.059.
8. 张伟.医疗大数据平台数据高并发方案设计与关键技术分析[J].信息技术与网络安全,2018,37(04):18-22.DOI:10.19358/j.issn.2096-5133.2018.04.004.
9. 王俊艳,冯新华.面向通信运营商的大数据建设方案分析[J].信息通信,2015(10):239.
10. .“平台+行业应用” 大唐电信推出大数据整体解决方案[J].通信世界,2017(13):33.DOI:10.13571/j.cnki.cww.2017.13.010.