可行性研究报告

Project Dawn



1st Edition on reversing

arvinsc@foxmail.com;511679327@qq.com

Dawn-Team



**前 言**

提出此项目时，项目组现有的成员（[Dawn-Team Group](#_项目提请信息)，20170910）刚刚完成web软件产品搭建的基本技术（Java EE、Spring framework、Web前端开发与数据库技术）和软件工程以及UML的基本理论，对于软件体系结构与设计和其他一些使用到的现代化的技术，仍然处于学习研究阶段。因此，在系统化工程化的构建上缺乏实践经验，体现在对于项目的分析缺乏思路、对于文档的编写缺乏规范、对于产品的构建缺乏良好的设计。本文档的编写时期，也是属于上述阶段之中，故在文档中出现的不足和缺陷，希望读者能够通过GitHub这一开源社区平台，[向我们提出Issue](#_项目提请信息)，我们会及时地验证出现的问题并尽可能及时地给出解决方案。

项目来源于笔者对于一个面向能源分析的ETL系统搭建过程的思考和总结。在传统的数据挖掘过程中，需要对于原始数据依次进行E(Extract)、T（Transform）、L（Loading）这三个步骤，才能把数据中的知识发掘出来进而为商业系统提供支持和决策，才能使得数据的价值能够得以发挥。上述三个步骤常见于各种BI（Business Intelligence）系统之中，而笔者曾经参与开发的ETL系统，面向的对象仅仅是能源行业，是更为通用化BI的一个小方面。相对于这样一个不完整的产品，本项目希望能够开发构建一个通用型的数据分析平台，不仅能够比较广泛地适应多种多样的数据源，进行分析和呈现，也要能够对外提供一套API，使得其他的专有平台（如上述面向能源行业的ETL系统）能够方便快捷地使用到本系统，使其开发使用和推向实际应用的过程变得简单便捷。同时，在全球各大厂商推出其各自的AI产品后（例如：来自Google的AlphaGo以及TensorFlow、来自Apple的siri、来自Microsoft的cortana以及其Cognitive Service等），各大科技企业和公司都着手构建自己的AI版图。本项目也希望能够利用相关的技术，提供相关的服务。项目特别对于非文档型数据处理有所涉及，意在能够利用Deep Leaning的技术提供在线便捷的数据处理工作；项目也考虑设计一套API，意在为一些需要数据处理的平台（不仅限于BI，也包括潜在的AI平台如自动驾驶平台等）提供方便的基础服务，也为本项目组日后对于这一领域的探索奠定基础。

对于项目组成员自身而言，本项目也是作为一个学习、训练和交流的机会。如上文所说，项目组成员均是学生，在全世界看来，还算不上初出茅驴的毛头小子。但也正是因此，我们正有时间和精力，在所热衷于的学科上，付出不断的努力，为这样一个令人兴奋的学科添砖加瓦。

最后，衷心感谢项目组中的成员，他们是：



Arvin Si.Chuan



Ash Morgan

**目 录**

[1 引言 1](#_Toc492766766)

[1.1 编写目的 1](#_Toc492766767)

[1.2 项目背景 1](#_Toc492766768)

[1.3 定义 2](#_Toc492766769)

[1.4 参考资料 2](#_Toc492766770)

[2 可行性研究的前提 3](#_Toc492766771)

[2.1 要求 3](#_Toc492766772)

[2.2 目标 6](#_Toc492766773)

[2.3 条件、假定和限制 6](#_Toc492766774)

[2.4 可行性研究方法 8](#_Toc492766775)

[2.5 决定可行性的主要因素 9](#_Toc492766776)

[3 对现有系统的分析 10](#_Toc492766777)

[3.1 处理流程和数据流程 10](#_Toc492766778)

[3.2 工作负荷 10](#_Toc492766779)

[3.3 费用支出 10](#_Toc492766780)

[3.4 人员 11](#_Toc492766781)

[3.5 设备 11](#_Toc492766782)

[3.6 局限性 11](#_Toc492766783)

[4 所建议技术可行性分析 12](#_Toc492766784)

[4.1 对系统的简要描述 12](#_Toc492766785)

[4.2 处理流程和数据流程 12](#_Toc492766786)

[4.3 与现有系统比较的优越性 12](#_Toc492766787)

[4.4 采用建议系统可能带来的影响 12](#_Toc492766788)

[4.5 技术可行性评价 13](#_Toc492766789)

[5 所建议系统经济可行性分析 14](#_Toc492766790)

[5.1 支出 14](#_Toc492766791)

[5.2 效益 14](#_Toc492766792)

[5.3 敏感性分析 14](#_Toc492766793)

[5.4 法律因素 15](#_Toc492766794)

[5.5 用户使用可行性 15](#_Toc492766795)

[6 其他可供选择的方案 16](#_Toc492766796)

[7 结论意见 17](#_Toc492766797)

# 引言

## 编写目的

本文档意在通过系统化的方法，对意图建设的软件产品（下称产品），从技术、资金和人员的角度，进行详尽可靠的分析和阐述。本文档应该作为产品立项的可靠性依据、作为潜在开发人员（包括产品经理、设计和实现者以及质量保证人员）进行产品开发和产品确认的重要参考资料。

## 项目背景

### 建议的软件名称

产品定位为通用数据处理平台，建议的软件项目名称为：Dawn。

其中，数据预处理部分的名称代号为：PancakeofMountHuang（PMH）。

### 项目提请信息

|  |  |
| --- | --- |
| 提出者 | 邱依强（[Arvinsc@foxmail.com，https://github.com/ArvinSiChuan](mailto:Arvinsc@foxmail.com，https://github.com/ArvinSiChuan)）  https://qr.api.cli.im/qr?data=https%253A%252F%252Fgithub.com%252FArvinSiChuan&level=H&transparent=false&bgcolor=%23ffffff&forecolor=%23000000&blockpixel=12&marginblock=1&logourl=&size=280&kid=cliim&key=b214c056596640d648ee2da9ea736e6a |
| 开发者 | [Dawn-Team](https://github.com/orgs/Dawn-Team/people)  https://qr.api.cli.im/qr?data=https%253A%252F%252Fgithub.com%252Forgs%252FDawn-Team%252Fpeople&level=H&transparent=false&bgcolor=%23ffffff&forecolor=%23000000&blockpixel=12&marginblock=1&logourl=&size=280&kid=cliim&key=254ebaa409c1835eeedc64c8e1572bc3 |
| 项目资料存储位置 | [Dawn-Team/dawn](https://github.com/Dawn-Team/dawn) |
| 所面向的用户 | 对于数据处理、分析和呈现，有较为经济的方式的需要的个人、教育机构和企业 |
| 软件实现单位 | [Dawn-Team](https://github.com/Dawn-Team/dawn)  https://qr.api.cli.im/qr?data=https%253A%252F%252Fgithub.com%252FDawn-Team&level=H&transparent=false&bgcolor=%23ffffff&forecolor=%23000000&blockpixel=12&marginblock=1&logourl=&size=280&kid=cliim&key=fae4ebfb20425644c13665dadeff0e8b |
| -- | -- |

### 与其他软件的关系

作为一个全新的数据处理平台，产品主要与数据库系统软件有较为密切的关系；在产品开发的后期，当提供对外API时，产品作为被依赖对象与其他软件有联系。

## 定义

|  |  |
| --- | --- |
| **名称** | **说明** |
| Dawn | 产品名称与项目代号 |
| Dawn-Team | Dawn-Team，Dawn开发组成员，详细清单位于：<https://github.com/orgs/Dawn-Team/people>  https://qr.api.cli.im/qr?data=https%253A%252F%252Fgithub.com%252Forgs%252FDawn-Team%252Fpeople&level=H&transparent=false&bgcolor=%23ffffff&forecolor=%23000000&blockpixel=12&marginblock=1&logourl=&size=280&kid=cliim&key=254ebaa409c1835eeedc64c8e1572bc3 |
| PancakeofMountHuang（PMH） | 数据预处理子系统部分的开发代号与名称 |
| BI | Business Intelligence |
| -- | -- |

## 参考资料

|  |  |  |  |
| --- | --- | --- | --- |
| **编号** | **资料名称** | **类型** | **详细信息** |
| 1 | Echarts | 官方网站 | <http://echarts.baidu.com/>  https://qr.api.cli.im/qr?data=http%253A%252F%252Fecharts.baidu.com%252F&level=H&transparent=false&bgcolor=%23ffffff&forecolor=%23000000&blockpixel=12&marginblock=1&logourl=&size=280&kid=cliim&key=8328ee27fc4fed1b44bfb4cdf2bb6eaa |
| 2 | 数据集 | 政府开放数据 | <http://www.datashanghai.gov.cn/home!toHomePage.action>    <https://www.data.gov/> |
| -- | -- | -- | -- |

# 可行性研究的前提

## 要求

### 产品功能

产品功能主要包括：数据（文档型）预处理（下称数据预处理）、数据分析、数据呈现、非文档型数据处理、开放性接口设计。以下是详细说明：

数据预处理。作为数据分析的前提，数据预处理需要完成对原始数据的数据清洗以及对于数据的在分析前所需要预先的运算。数据清洗即对文档型数据中的缺失值和异常值，通过适当的算法进行处理，以满足后续数据分析的需要的过程；数据所需要的预先的运算是指，为了满足特定的数据分析需要（如：原始数据已知功耗与对应的时间，可以预先运算得出对应的功耗），对已知数据进行一定的加工运算以满足后续数据分析需要的过程。

数据分析。数据分析是指，对已经通过了数据预处理的数据，使用诸如聚类、回归分析、关联分析等相关算法，对数据进行深入地发掘而获取其中更多的信息和知识的过程。

数据呈现。数据呈现即对数据分析的结果，根据其使用的算法，选用合适的方法（例如采用散点图、条形图、和饼图），进行有效的呈现（例如以可视化的方式呈现）。

非文档型数据处理。面向机器学习和BI的数据常常不仅限于文档型数据，而常常包括如图像、音频、视频甚至复杂的声光电传感器数据，这些数据不能通过简单的数据预处理进行有效地清洗和转换，媒体数据处理旨在解决这样的相对复杂的问题（例如处理一个人脸图片数据集中的图片尺寸、图片内容以更加符合对其进行数据分析的需要）。

开放性接口设计。在系统自身功能和稳定性相对完善的情况下，为丰富本系统的运作和扩展易用性，对已有的功能进行一定的包装，对外设计一套可以直接使用系统核心组件的过程，即为产品的开放性接口设计。

### 产品性能

除传统的关于产品在一定限制条件下对一定负载能够进行有效的响应外，产品需要对进入系统的大量数据处理、数据分析和接口使用有良好的调度处理能力，需要能够在将要出现高负载的情况下及时告知用户并保证数据的安全性和产品的稳定性。

### 输入与输出

从全局视角观察，系统的输入是用户给出的一份或者多份原始数据，输出是对于数据进行发掘而得出来的信息和知识。从各个子系统的角度观察，有如下的输入输出关系：

从数据预处理系统以及非文档型数据处理系统看，以用户提供的数据源（如文档型的Excel、CSV文件和DB连接，非文档型的图像和声音文件）作为输入，以处理后的规范化的结构化数据集（文档型数据）或特征值、非文档型数据集（非文档型数据）作为输出。

从数据分析的角度看，预处理后的结构化数据作为输入，通过一定算法得出的结果、结论作为输出。

从数据呈现看，数据分析的结果作为数据呈现的输入，多样的数据图表、文字表达作为输出。

从开放性接口看，使用接口的系统所传入的命令和参数作为输入，根据被调用的接口而返回的数据（如JSON、URL、状态码等）作为输出。

### 基本数据流程与处理流程



1. 基本数据流程和处理流程

由图1所示，整个系统的工作流程根据请求来源的不同，可以分为由系统内部请求所导向的基础性功能处理路径和由外部系统调用开放API进而相对独立使用基础性功能的处理路径。需要进一步说明的是，在系统视角（全局视角）看,基础性功能应当按照路径一步一步进行（由于数据的输入输出存在着一定的依赖关系，见[2.1.3输入与输出](#_输入与输出)），才能够作为一次完整的数据处理，而相对比之下，调用开放性API的过程可以独立使用基础性功能中的任何一个模块，在使用后可以立即结束整个数据处理流程。而从每一个独立的子系统视角看，子系统可以独立于其他系统完成其自身的功能，而不受制于其他子系统，在整个系统的实现中，可以选择性对系统内开放单个子系统的独立性功能。简而言之，要处理一整套数据处理任务，是需要各个子系统协同工作，缺一不可，但是，用户可以灵活地根据需要单独使用其中的一个部分；与此同时，相对与系统内部，为外部调用本产品的系统提供服务的开放性API则是直接工作在单个的独立的子系统中的，子系统之间的联系是需要外部调用者负责的。

### 安全与保密要求

安全性与保密要求在产品中主要表现为两个方面，一是对用户自身使用系统（下称：用户导向的安全）的用户名、密码、邮箱、使用记录和Cookie信息的安全保障和系统数据的故障转储处理，另一方面是对于用户投入到产品中的数据（下称：数据导向的安全）的安全性保证。

用户导向的安全。用户在使用产品期间（包括非在场情况，即处于对系统并非实时操作的状态）的数据，需要设计严格的安全与保密机制，包括用户信息和使用记录的保存与读取、使用系统的安全验证、中止使用系统（停止服务与完全注销<非退出登录>用户）的功能、在客户端使用cookie和本地存储的告知服务以及数据备份的功能。

数据导向的安全。在用户使用产品时，对于用户向系统输入的数据（用户用于数据预处理或非文档型数据处理的数据），在用户输入的过程中应当提供风险告知的服务，对于这样的数据，系统应该尽量保证其不被其他未经过用户授权的外部系统访问，关心其在系统中的生存情况（是否被意外清除等）。

### 与其他软件系统之间联系产生的需要

产品在使用过程中会产生大量的数据，根据其产生对象，主要分为用户导向的数据（因为使用系统而产生<而与所分析的数据本身无关>的数据）和数据导向的数据（因为需要对数据进行分析而由用户引入的数据或与用户引入数据相关的数据）。对于用户导向的数据，需要使用关系型数据库系统进行数据的保存与修改，对于数据导向的数据，根据需要可能使用到关系型数据库系统与非关系型数据库系统。对于这两类数据库系统的选择，需要根据实际需要进行适当的选择；在数据库系统之间的联系，应当尽可能地提高效率。

### 产品完成交付的期限

产品时间表计划如下（均为完整的开发生命周期）：

|  |  |  |
| --- | --- | --- |
| 系统功能划分 | 计划交付日期 | 备注 |
| 数据预处理 | 2017-11 | 开发代号：PancakeofMountHuang |
| 数据分析 | 2018-01 |  |
| 数据呈现 | 2018-02 |  |
| 非文档型数据处理 | 2018-04 |  |
| 开放型API | 2018-05 |  |
| -- | -- | -- |

## 目标

### 人力资源的节约

产品将数据从原始数据到数据呈现集成，将复杂的数据预处理过程做到可视化自动化的处理，相比于当前独立使用各种软件进行处理甚至手工处理，能够减少使用者在各种软件之间学习和相互联接的时间，能够重复、批量地处理数据以尽量地节约人力资源在这样的数据处理中的投入。

### 生产效率的提高

产品需要达到能够直观给予用户在操作上的便捷、在数据处理上的批量化、自动化，减少用户对数据处理结果不满意而返工的概率，提高生产效率。同时，当前的专业化而免费的数据分析和可视化工具都需要通过一定的学习，大多数时候需要使用者自行编程实现，产品希望能够尽可能减少使用者在这方面的学习时间，消除使用者在使用系统时对编程的依赖。

同时，目前在市面上提供服务的数据分析系统，大多采用收费制或部分收费制，对于这样的市场现状，产品希望能够本着学习交流的目的，减轻对数据分析这样类别的软件的使用成本，进而提高使用者的生产效率。

此外，产品在实现过程中，在保证稳定性的前提下，将大量采用优秀的先进的技术，旨在提供更好的产品结构和运行效率。这些技术包括但不仅限于：HTML5、CSS3、Bootstrap、Echarts、Ajax、WebSocket、NoSQL、Apache common。

### 管理信息服务的提高

在数据处理的批量化、自动化过程中，产品需要能够有效处理高耗时任务，能够给用户明确的任务执行进度，提高用户对于自己提交的数据处理任务的管理能力。

## 条件、假定和限制

### 产品运行的寿命

产品作为一个数据分析的工具，在需求的变化情况上，不同于很多信息管理系统。因此，在满足基本的数据分析需要的基础上，应当尽可能长地保证本产品的可用性，增加分析算法的种类和产品在数据分析上能够应用的范围。所以，在产品寿命的限制上，本系统应当作为一个长期支持的软件产品，在支持期限的规定上，对于目前构建的版本和功能，提供三年的维护和支持，对于后续的功能性更新和软件的迭代，在后续版本开发时再定义。

### 经费的来源和使用限制

产品作为一个自行设计开发的开源软件，所有的经费（人力和物力）均由项目组的成员自行提供。在此，特别说明物力经费的限制。在项目开发期间，所采用的计算机和软件产品，均由项目组成员自行提供；在项目开发完成后（发布第一个发布版本），由阿里云计算提供计算服务，其费用在本产品的生命周期中，项目组成员提供不超过￥100/月的经费。

### 法律和政策限制

在目前产品触及的范围内，从法律和政策上分析，主要分为对于用户的限制和产品自身发布运营的限制。

对于用户的法律法规的限制，主要有以下这些方面：

1. 用户对于自身身份信息应当保有隐私权，可以选择在任意时刻清除其使用信息；
2. 用户有权利选择其在本系统中作为分析用的数据的保存时间；
3. 本产品有义务保护使用者的信息不被泄露，保护其数据（使用系统产生的数据和用户用于数据分析的数据）的安全性，但出现不可抗力的情形下除外。

对于产品自身发布运营的法律法规限制，主要有一下方面：

1. 产品在上线运营时，应当本着技术交流的原则，不应当出现营利性的内容；
2. 产品上线运营所需要的域名、IP等资料，应当遵守相关规定。其中，本项目拟发布和运营的IP，由[项目提请者](#_项目提请信息)提供。

### 硬件、软件、运行环境和开发环境的限制

产品运行所必要的基础硬件环境应当至少满足：

|  |  |
| --- | --- |
| **项目** | **内容** |
| CPU | 单核 1Ghz |
| 内存 | 1GB |
| 富余硬盘空间 | 16GB |
| 互联网带宽 | 100Mbps |
| -- | -- |

软件环境应当至少满足（针对于用户端）：

|  |  |
| --- | --- |
| **项目** | **内容** |
| 浏览器 | Chrome 60+ 或其等效兼容浏览器 |
| -- | -- |

运行环境应当至少提供：

|  |  |
| --- | --- |
| **项目** | **内容（以下内容均为发布版<Release Version>）** |
| JDK | 1.8 |
| Spring | 4.3 |
| MySQL | 5.7 |
| NoSQL数据库系统 | 适用于K-V存储和文档型存储的优秀产品如：Redis、MongoDB |
| OS | 安全稳定的服务器版本OS，如：Ubuntu Server |
| -- | -- |

开发环境：

|  |  |
| --- | --- |
| **项目** | **内容** |
| IDE | Myeclipse 2015+ |
| Git | Github Platform |
| Word | 2013+ |
| MySQL Workbench | MySQL Workbench 6.3 CE+ |
| -- | -- |

### 可以利用的信息和资源

|  |  |  |
| --- | --- | --- |
| **类型** | **内容** | **备注** |
| 数据可视化 | Echarts，D3.js |  |
| -- | -- | -- |

### 建议开发软件投入生产的最迟时间

根据项目[计划时间表](#_产品完成交付的期限)，整个产品投入生产的最迟时间为2018年05月，其中各子部分的计划时间表，参照项目[计划时间表](#_产品完成交付的期限)。

## 可行性研究方法

### 经济可行性分析

成本效益分析从经济角度评价开发一个新的软件项目是否可行，它主要包括：＂成本－收益＂分析和＂短期－长远利益＂分析。在具体的分析中可以使用投资回收期、纯收入和货币的时间价值作为度量参数评价投资效益。

### 技术可行性分析

技术可行性分析研究根据当前拥有的资源和技术条件下能否给定的时间内实现用户需求中提出的功能。在技术可行性分析可以分解为物理分析、功能分析、信息分析和动态分析，在这一过程中通常利用数据流分析方法从功能、数据（信息）结构对物理系统进行分解和建立静态和动态模型，并将当前的设计和已有的方案进行分析和比较，选取最合适的开发路线，保证项目可以顺利进行。

### 人力估计

人是开发中的一个非常重要的资源，它除了具有一般资源的特性外还具有不确定性。软件是一种特殊的商品，它是智力的结晶，它需要多种可见资源和参与者的相互配合，但人力资源和其他资源的一个区别在于不会由于人力的增加就会使得生产率得到线性提高，过多的投入人力会增加通信费用和成本开支，因此，必须合理安排人员的结构和数量。

### 社会环境分析

社会环境的可行性至少包括两种因素：市场与政策。其中市场又可以分为未成熟的市场、成熟的市场和将要消亡的市场，政策主要涉及到法律和知识产权。环境分析就是明确系统的目的和限制条件；使用单位的状况、经营方针和组织机构、使用单位的计算机利用情况；相关的硬件、软件及其它接口部分；用户的操作环境及操作要求；习惯、法律、制度上对软件的制约；开发能具备的基数条件和设备条件。

### 复杂项目中的可行性分析方法

可行性分析的目的是用最小的代价在尽可能短的时间内确定问题是否可以解决。如果说需求分析决定软件的设计质量，那么可行性分析则决定了项目的开发路线。除了确认用户需求和对问题进行精确定义外，可行性分析还有粗略地估算成本；估算可能取得的效益；提出包括人员、硬件、软件等在内可能需要的资源，制定资源需求和使用清单最大限度的为管理和决策提供准确的信息支持。

## 决定可行性的主要因素

根据可行性分析的研究方法，决定可行性的主要因素包括：经济、技术、人员和社会环境。由于本项目作为开源项目，故其中最主要的因素应为技术因素和人员因素。

# 对现有系统的分析

## 处理流程和数据流程

本项目计划开发一套全新的数据分析产品，故没有完全匹配的完整成型的处理流程和数据流程，产品整体视角的处理模型，均是由科学研究中使用相关的软件而逐步形成的。

在各个子系统的设计和实现模型中，可以从一些优秀的处理平台借鉴思想、处理流程和数据流程。在数据预处理中，可以参考Airflow（Airbnb出品，参见“图2. Airflow Documentation”）;在数据运算处理中，可以学习NumPy（参见“图3. NumPy官方网站”），由Apache Common Math来实现；在数据可视化中，可以直接参考和使用百度[Echarts.js](#_参考资料)。



1. [Airflow Documentation](https://airflow.incubator.apache.org/index.html)



1. [NumPy官方网站](http://www.numpy.org/)

## 工作负荷

由于给出的[硬件环境的最低要求](#_硬件、软件、运行环境和开发环境的限制)并不高，所以在实际工作时，本系统实际能够承受的工作负荷可能比实际计划中的低，可能出现分析任务大量排队的情况。

## 费用支出

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目** | **单价（￥）** | **数量** | **小计（￥）** | **备注** |
| 生产系统服务器 | 15/月 | 1台 | 15 | 阿里云华北1可用区B，已经可用，到期日期为2018-08-27 |
| 生产系统公网流量 | 0.226 /Mbps/小时 | -- | -- | 按量付费 |
| -- | -- | -- | -- | -- |

## 人员

|  |  |  |
| --- | --- | --- |
| **人员职责** | **数量** | **备注** |
| 前端开发 | 1 | 可兼顾其他职责 |
| Java Web Server端开发 | 1≤X≤2 | 可兼顾其他职责 |
| 数据库系统设计 | 1 | 可兼顾其他职责 |
| 软件测试 | 1≤X≤2 | 可交叉兼顾其他职责 |
| -- | -- | -- |

## 设备

|  |  |  |  |
| --- | --- | --- | --- |
| **项目** | **类型** | **数量** | **备注** |
| 生产服务器 | 服务器 | 1 | 可用 |
| 开发用计算机 | 计算机 | 2 | 可用，由项目成员自行提供 |

## 局限性

当前可用的数据分析软件产品，绝大部分为C/S结构的产品，并且，在这样的产品中，有相当一部分的产品只能完成一个数据分析任务中的一部分，不能够将完整的数据分析处理任务连接起来。在少部分可以完整执行数据分析任务的软件中，几乎都需要对软件进行授权，即非免费提供软件产品。

综上所述，需要进行一次开发以填补这些可用系统的短板、缺陷，与他们的市场进行竞争。

# 所建议技术可行性分析

## 对系统的简要描述

系统主要功能为提供数据处理、分析和呈现的服务，主要包括四个核心的子系统和两个附加的子系统。其中核心的子系统包括：数据预处理、数据分析、数据呈现以及运行系统所需要的管理功能（如用户管理、作业调度等）；附加的子系统为：非文档型数据处理、开放性API。其中详细的描述参见[2.1“要求”一节](#_要求)，下面的内容主要说明其中简要的结构和组成。

数据预处理。此子系统需要的提供的功能有：对数据源（文档、关系型数据库）的支持、对多种非正常的数值数据提供数据预处理和数据转换的接口以及算法、对数据预处理后的数据进行存储（文档、数据库）的接口以及实现、对数据预处理任务进行调度的接口和实现。

数据分析。此子系统需要提供的功能有：对数据进行分析的算法接口和实现（包括但不仅限于聚类、回归分析、分类）、对分析结果进行存储的接口以及算法、对数据分析任务进行调度的接口和实现。

数据呈现。此子系统应当包括：对多种图表呈现的接口和实现、对呈现进行管理的接口和实现。

非文档型数据处理包括：对于多种非文档型数据源进行支持的接口和实现、对非文档型数据源进行处理的算法的接口支持。

开放性API。对外部系统调用此系统进行控制的调度接口，对于外部系统使用本系统的适配器角色支持的接口和实现。

## 处理流程和数据流程

参见[2.1.4. 基本数据流程与处理流程](#_基本数据流程与处理流程)

## 与现有系统比较的优越性

与现有系统相比，产品的优越性主要体现在：免费、基于服务器/浏览器（B/S）结构、系统化集成和系统化调度管理。

## 采用建议系统可能带来的影响

### 对设备的影响

由于产品拟采用B/S结构，故对于使用产品的用户，不需要要求有性能相对较好的计算机系统、不需要用户对设备的OS进行选择，只需要有能够访问到本项目的浏览器和网络支持即可。

### 对现有软件的影响

对于互联网中已有的数据分析处理软件，产品作为其B/S结构的集成软件，对其会有一定的竞争，会对其占有率产生影响。

### 对用户的影响

对于用户而言，产品可以减轻用户在处理数据分析任务时需要学习软件、学习编程以及对数据处理步骤进行互相连接沟通的负担，使用户将注意力转移到更加有价值的数据本身而非对于工具对的使用上，提高其工作的效率，缩短从数据中发现知识、挖掘价值的时间。

### 对经费支出的影响

本项目为自发的开源项目，基于已有的设备和人员即可进行开发和生产，故对于经费支出没有特别的影响。

## 技术可行性评价

在[限制条件](#_条件、假定和限制)下，利用已有的技术，[功能目标](#_产品功能)基本能够完成，但是存在几个需要注意的环节。一是，在最低限制条件下，系统将会承担比较大的负荷，可能导致系统运行缓慢，在实际设计开发环节中，应当加强资源调度的控制和性能优化处理。二是，对于实际的数据源，可能存在多种多样的数据形式，也存在各种各样的数据异常，使得数据的预处理工作变得比较复杂。

要较好地处理系统性能的瓶颈问题，一方面尽可能地提升生产系统的资源配备，一方面在开发时需要选择更有经验的开发人员。要处理好数据源多种结构的问题，一方面需要把数据源支持接口设计得尽可能地可扩展，另一方面，对于数据源要有一个基础的结构要求而不能使得输入的数据源过于零散。

综上所述，在规定的期限内，大体能够完成整个系统的功能，但是在系统测试和缺陷修补上，可能会存在一些不足。

# 所建议系统经济可行性分析

## 支出

### 基建投资

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目** | **单价（￥）** | **数量** | **小计（￥）** | **备注** |
| 生产服务器 | 15/月 | 12台 | 180 | 在已有系统中已存在 |
| **总计** | **--** | **--** | **180** | **--** |

### 其他一次性支出

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目** | **单价（￥）** | **数量** | **小计（￥）** | **备注** |
| 暂无 | -- | -- | -- | -- |
| **总计** | **--** | **--** | **--** | **--** |

### 经常性支出

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目** | **单价（￥）** | **数量** | **小计（￥）** | **备注** |
| 生产系统公网流量 | 0.226 /Mbps/小时 | -- | -- | 按量付费 |
| **总计** | **--** | **--** | **--** | **--** |

## 效益

本项目作为开源项目，产品作为学习交流性质的免费软件，故不会给项目组成员带来任何实际性的经济收益。

## 敏感性分析

在最低[硬件限制条件](#_硬件、软件、运行环境和开发环境的限制)下，可能存在系统高负荷运转而造成响应缓慢的问题。在这个方面，应当保证用户至少能够操作产品的界面并与服务器通信，使其知晓自己提交的任务处于处理队列中；应当能够保证系统至少执行作业队列中的一个任务，并且，任务队列中的任务不能够出现丢失的情况。社会因素可行性分析

## 法律因素

系统开发中会使用到大量的已有的开源项目库，在使用之前，应当核准其开源证书，保证其能够用于开源项目开发、教育和科研目的以及商业应用。

对于用户传入用于分析的数据，若有存在侵犯他人权利或者违反相关法律法规的情况，由具体使用者承担相关法律责任。这种情况下，系统有提前告知用户承担法律责任和风险的义务。

## 用户使用可行性

产品面向对于数据分析处理有相对经济的需要的用户。对于使用者，需要有对于计算机操作的简单知识和技能，需要对于数据分析、处理和呈现的相关概念有所知晓。

# 其他可供选择的方案

对于解决集成化的数据处理、分析和呈现问题，有如下一些解决方案：

1. [网易大数据](https://bigdata.163yun.com/?channel=B_baidu_sjfx_01514)



1. [数加 · 大数据开发套件](https://data.aliyun.com/product/ide?spm=5176.7946122.388261.313.AjuoHF)



1. [tableau](https://www.tableau.com/zh-cn)



1. [数析网](http://www.tjsql.com/)



在这些相对比较成熟的方案中，不乏有比较适合的平台。但没有直接采用的原因，有以下几个方面。一是，大公司平台（如1、2）有一个相对较高的门槛（如需要购买昂贵的服务器资源、需要单独对平台再进行开发学习等）；二是，对于一些较小的平台（如4），其功能支撑不够完善，例如可视化图表种类偏少等；三是，这些平台都需要进行购买，生产成本较高。

# 结论意见

【结论意见可能是：

1. 可着手组织开发；
2. 需待若干条件（如资金、人力、设备等）具备后才能开发；
3. 需对开发目标进行某些修改；
4. 不能进行或不必进行（如技术不成熟，经济上不合算等）；
5. 其他。】