**实际问题：**

政府/企业机构/个人都有很多数据，但：

1-在数据存储上：异构结构，信息孤岛式存储，各类数据无法相互连通。

2-在数据利用上：

* 个人和小企业数据量小，虽然他们也很想利用这些数据，但没有利用数据的技术，也不知道怎样利用。
* 政府和大企业有大量数据，但他们对数据的利用很局限，达不到所期望的要求。

3-在数据服务的提供上：基本采用项目的模式，主要服务于大客户，各类公司产品水平参差不齐，领域性很强——虽带来专业的好处，但没有横向的交流。中小企业及个人基本无法得到这样的服务。

以上可知，目前的数据存储形式、利用模式、技术(虽然有大数据技术，但不注重分[http://www.csdn.net/article/2014-06-05/2820102]析)已不能满足需要。

**在大数据时代互联网环境下，需要构建新的以服务为中心的数据应用模式。才能激活数据的效能，为客户提供更加优质的服务。**

**事业：**

**目标：**

激活数据、交易数据、利用数据。

**手段：一个开放的数据服务平台**

提供数据存储/分析/交易的平台。

为中小用户——个人和中小商户——提供数据分析的平台。

为大数据提供者提供数据交易的平台。

**宗旨：**

开放和安全并行。

中立于任何机构和公司。

专注于数据分析（不做少做数据采集的工作）。

**希望最终形成以提供数据服务（分析和交易）为目的开放的商业生态环境。**

**平台构成：**

从架构：数据存储归类平台、数据分析平台、数据交易计数平台。

从用户角度看：数据提交系统，数据分析展示报告系统，数据交易系统。

如下图：



**1.1、数据存储归类平台：**

包括：

1. 数据与用户绑定(为计数做准备)；
2. 数据导入:第一次分析（对数据质量进行评估）：

对于结构化数据（rdb/格式文档[excel/xml/日志]）：

**自动的元数据分析——减轻元数据定义的用户工作，并且避免人为不专业的定义对数据分析的影响。包括：数据类型识别/语义识别{字典识别/人名识别/证件号识别/地名识别/地理信息识别},同时进行数据质量的第一次过滤**

对于非结构化数据（视音频，文档）：

利用标签/全文检索等对文档进行标记。

1. 数据存储：采用分布式方式进行数据存储。

**数据包括：结构化数据、非结构化数据(文本、视音频)、空间数据、特殊数据(如CAD等)**

更详细的说明，可参见“数据的架构”。

**1.2、数据分析平台：**



此处数据分析是平台的子部分，允许分析的开发者把分析的代码插到平台中，为平台用户提供服务，同时，用户使用了这些分析方法，也会计数。

**1.3、数据交易计数平台：**

计数包括三部分：使用计数，提供计数，分析工具计数。同时计数还要能够在进行传递。



目前考虑计数应该从如下方面考虑：

***使用者：***使用了才计数，此计数要能通过计数传递，计算出算法和数据的提供者分别的贡献。

***算法：***正向贡献——算法的使用频度，同类算法的比较优势；反向贡献——资源消耗；

***数据提供者：***数据的质量，数据的量，数据的使用频率；

**数据架构：**

在平台中，数据架构如下：



**1.1、参与者：**

此架构也说明了平台各参与者的作用。

**数据提供者：**他们提供数据，从使用数据者处获得计数。

**数据使用者：**他们使用数据或利用平台提供的数据分析结果，也可能直接购买数据。

大多数情况下，数据提供者也是数据使用者。

**分析算法提供者：**提供数据的分析算法(包括算法的组合、复杂算法流程的符合某一特定业务领域的参数方案等)，从数据使用者处获得计数。

*数据的整理和存储目前设计由平台提供商开发，是否也可以开放吗？*

**1.2、数据的流程：**

1-首先数据提供者提供数据；

2-数据经整理进行分布式存储(记录数据与提供者的关系)：同时进行数据的元数据分析，单次数据与积累数据的合并；

3-对数据进行分析、归纳，形成D2数据(中心层数据)；

4-提供数据：从中心层提取数据，与自有数据结合，分析出对用户有用的结果。

以上过程，隐藏了数据的间接交易的过程。

**1.3、数据交易：**

1-数据的展示：（数据的提供者）

·对自有的数据进行展示或进行加工(分析)后展示；

·对自有数据与组合区数据联立加工(分析)后展示；

2-数据获得：（数据的使用者）

·检索展示的数据(可能需要开发数据展示的检索系统)，找到自己的需要的数据

·交易数据（拍卖/根据制定好的规则交易）

**淘汰和遗忘机制：**

在平台中，随着平台的运营：数据的量、分析算法的量都会不断地增加，带来——

1-资源消耗增长；

2-无用现象：

·无用或过时的数据会增加，数据质量的下降

·无用的计算资源

因此，必须有淘汰和遗忘机制：

遗忘——根据使用反馈和时间衰变情况，把用处不多的数据和算法从总框架中去掉；

淘汰——根据使用反馈和质量(数据的质量/算法的质量)，把劣质的算法和数据从总框架中去掉。

从“数据架构”图看，在中心处需要淘汰这些数据，但在各用户自己的作用域内，不进行淘汰，这个权利留给各用户自己。

**难点：**

***拿到数据***——为什么要把数据拿出来？对数据提供者有什么好处？安全吗？

***分析数据***——怎样能够分析出有价值的结果（对于不同的人、不同的动机有不同的需求）。

***数据的安全***——让用户相信数据是安全的：独立存储与权限，授权的用户控制。

|  |
| --- |
| Kb:2^10=1024  Mb:2^20=1024\*1024=1048576~10^6  Gb:2^30=1024\*Mb=1073741824~10^9  Tb:2^40=1024\*Gb=1099511627776~1.1\*10^12  Pb:2^50=1024\*Tb=1125899906842624~1.1\*10^15  ......  Zb(Zettabyte 泽字节) = 2^10 EB = 1024 EB = 2^70 B；  Yb(YottaByte 尧字节):2^80=1208925819614629174706176~1.2\*10^24 = 2^10 ZB = 1024 ZB = 2^80 B；  这应该是大数据的上限。超过它，就要用其他方法处理了？？  **到2020年，全球数据使用量预计将暴增44倍，达到35ZB**  阿伏伽德罗常量:  6.02\*10^23=2^78.99 |