

2013

编者按：这是一个大胆的想法，更是一个数字油田创新驱动的典型。我们从 2012 年就启动了数字油田数据正常化建设研究工作，核心是解决数据问题。就是说数字油田建设 10 多年，大家反映“不好使”，这是一个什么问题？暑假我参加数字油田调研活动，有幸拜见了王权主任，我们一起深入地讨论了“不好使”的问题，特别是他的数据建设中的“脱库”思想和“自标准”思想，对我们震动很大。这里将王权主任今天在智能数字油田论坛上发表的“自标准数据设想”在我们网站上转载，以飨读者，共同思考，创新驱动。

高倩

2013 11 4

自标准数据设想



2013-11-2

概述

自标准数据，Self-Standard Data，即自带标准或格式的数据体。数据提供者按照自己的标准或格式提供数据，并将该数据所使用的标准或格式与所提供的数据一起打包，数据使用者按照该标准或格式自行解读并使用数据。它是传统数据标准体系的重要补充和突破。

这一设想的初步想法本人于 2013 年 10 月形成。目的是为了解决“数据标准过严不易推行，过松不易整合”的问题。顺应“数据使用者急，积极性高；而提供者不急，积极性不高”的实际情况，按照客观规律充分调动数据使用者的主动性，减轻数据提供者的负担。理念一经提出即受到多位专家、学者支持，在智能数字油田开放论坛热烈讨论，在大庆油田也引起一部分人的关注，并计划在即将实施的系统中进行试验。下图是本人设计的自标准数据的标识图。



自标准数据 Self-Standard Data

定义

自标准数据，Self-Standard Data，即自带标准或格式的数据体。数据提供者按照自己的标准或格式提供数据，并将该数据所使用的标准或格式与所提供的数据一起打包，数据使用者按照该标准或格式自行解读并使用数据。它是传统数据标准体系的重要补充和突破。

性质

自标准数据打破了大家共同遵守统一标准的局限，给数据共享提供了更加切实可行的路径。其具有如下性质：

- （1）自标准数据是一种数据体，它既包含数据本身还包括数据格式；
- （2）自标准数据是元数据的一种特例，元数据与数据捆绑；
- （3）自标准数据是一种全新的数据共享模式，打破了传统的数据与标准脱离的局面；
- （4）自标准数据是一种客观、现实的数据管理策略，适应性强；
- （5）自标准数据是大数据的基本单元，采用自标准数据技术有利于大数据技术发展；
- （6）自标准数据是系统自治思想的应用。

与传统数据库等的区别

自标准数据离不开传统数据库及相关技术的支持，但仍有很多明显区别：

- （1）传统数据库里面的数据注重存储、查询、更新，而自标准数据更注重流动性，主要目的是数据共享；
- （2）传统数据库的元数据与数据分离，查询数据时在数据库系统上分析元数据，而自标准数据中的标准就相当于元数据，它不固定在数据库上，而是随着数据体一起流动；
- （3）传统数据库存储的数据量是积累性的，会越来越大，而自标准数据是增量性的，每次的体量可能变化不大，体量一般不大；

(4) 传统数据库的数据结构与数据本身加起来也可以看做是一种自标准数据体，只是体积大，不便于流动；反过来，自标准数据体可以看作是流动的数据库，只是体量较小；

(5) 传统数据库重视冗余，自标准数据不重视冗余，而重视时效性；

(6) 传统数据库结构是严格统一的，而自标准数据的格式和标准是允许自定义的；

(7) 传统数据库主要支持某个（些）专门软件，而自标准数据主要支持系统间数据共享；

(8) 与其他具体的大数据理论或技术相比，自标准数据主要是一种思想，可使用多种方式和技术实现。

起源

1998 年，大庆油田开展了一个项目——《勘探、开发、钻井数据一体化共享》。该项目目标是建立一个油田内部数据共享的平台。当时认识到，“数据使用者急，积极性高；而提供者不急，积极性不高。”鉴于此，为了实现项目目标，项目组决定顺势而为。本人是项目负责人，当时我提出一个想法，叫做“数据码头”，就是数据提供者把数据放在指定位置就不管了，使用者自己去取，去处理。使用者再产生的数据也放到码头上。这样就调动了使用者的主动性，也减轻了提供者的工作量。该想法得到了项目组的认可。但后来大庆油田重组，勘探和钻井的大部分业务与油田开发分离，此项目下马。

那时还是要求提供者按统一标准把数据放到“码头”上，提供者还要处理数据，所以没有把提供者的工作量减到最小。当时，还没有XML，也没元数据，也没想到用它来描述数据。

2013 年，大庆油田制定信息规划过程中，关于信息共享（十多年过去了，问题依然很多）进行了讨论。期间本人对“数据码头”思路进行了进一步的扩展，应用 XML，让提供者按照自己的格式提供数据，还要包含这些数据的格式。这样，这些数据就成为了“自标准数据”，使用者能读明白，想怎么用就怎么用。大家都方便。

2013 年 10 月，本人为长安大学数字油田论坛第三届大会准备题为《数据多了就智能！》的演讲材料时，与数字油田研究所所长高志亮教授、高倩博士进行了较深入的探讨，形成了较完整的思路。本届论坛上，各位专家学者对“自标准数据”给予了充分肯定。参加会议的多位专家学者建议，简化传统数据标准，简政放权，大力推广“自标准数据”，并认为“自标准数据”将成为大数据时代的有力的信息共享的支撑性技术。

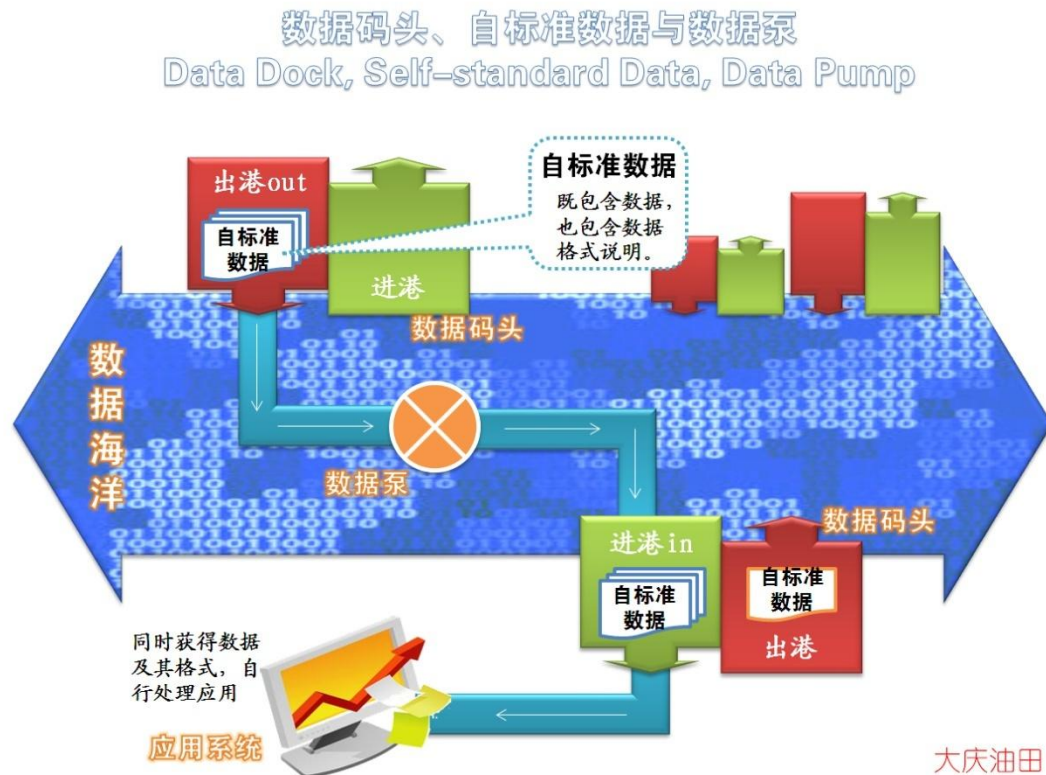
相关技术

配合自标准数据，本人进一步丰富了原来数据码头等概念，并进一步提出了数据泵的概念。

数据码头，Data Dock，即数据提供者放数据的场所。数据提供者把数据放在指定位置就不管了，使用者自己去取，去处理。使用者再产生的数据也放到码头上。

数据泵，Data Pump，是专门的抽取数据的部件，可以是集中的，

或分散的。它可以被看作是传统数据适配器的改进。其功能是：存取数据，全局统一管理资源目录、使用权限等。



应用前景

数据标准一直让人头疼，主要是难以统一。

客观上，太严格不好执行，太松不好整合。

更重要的，是主观的，产生数据的人不积极遵守标准，自己方便就行了。

特别是在油田上，产生数据的人都是主角，信息化主要处于弱势的服务地位，要求勘探开发主营业务人员主动遵守标准，是十分困难的。

自标准数据有望使这一问题得到较好解决。

另外，大数据的迅速发展，传统的数据库、数据仓库、数据银行的技术都面临一个信息共享的问题，但都具有提供者不主动、使用者主动的特点，可以应用自标准数据提高海量信息共享的主动性和有效性。

致谢

自标准数据还是一个很初步的想法，能否可行尚有很多疑问，需要深入探索。感谢自本人抛出这个想法以来各位专家学者的支持、建议和批评。特别感谢高志亮教授、高倩博士、程国建教授、张艳国教授、譙英教授、黄放明教授、王哲博士等！希望有关专家、学者、技术人员、管理人员继续提出宝贵意见。