2021

元数据文献综述

报告人：郭志超

目录

大数据时代

随着计算机技术的发展，数据的产生量与日俱增，收集这些庞大而复杂的数据，现有的数据库变得难以处理。针对大数据时代，需要新的架构、技术、算法来管理和提取数据隐藏的价值。

数据特征

在现在的IT行业，各个企业对大数据这一概念都有着不同的解读。但大家都普遍认为，大数据有着4V特征：

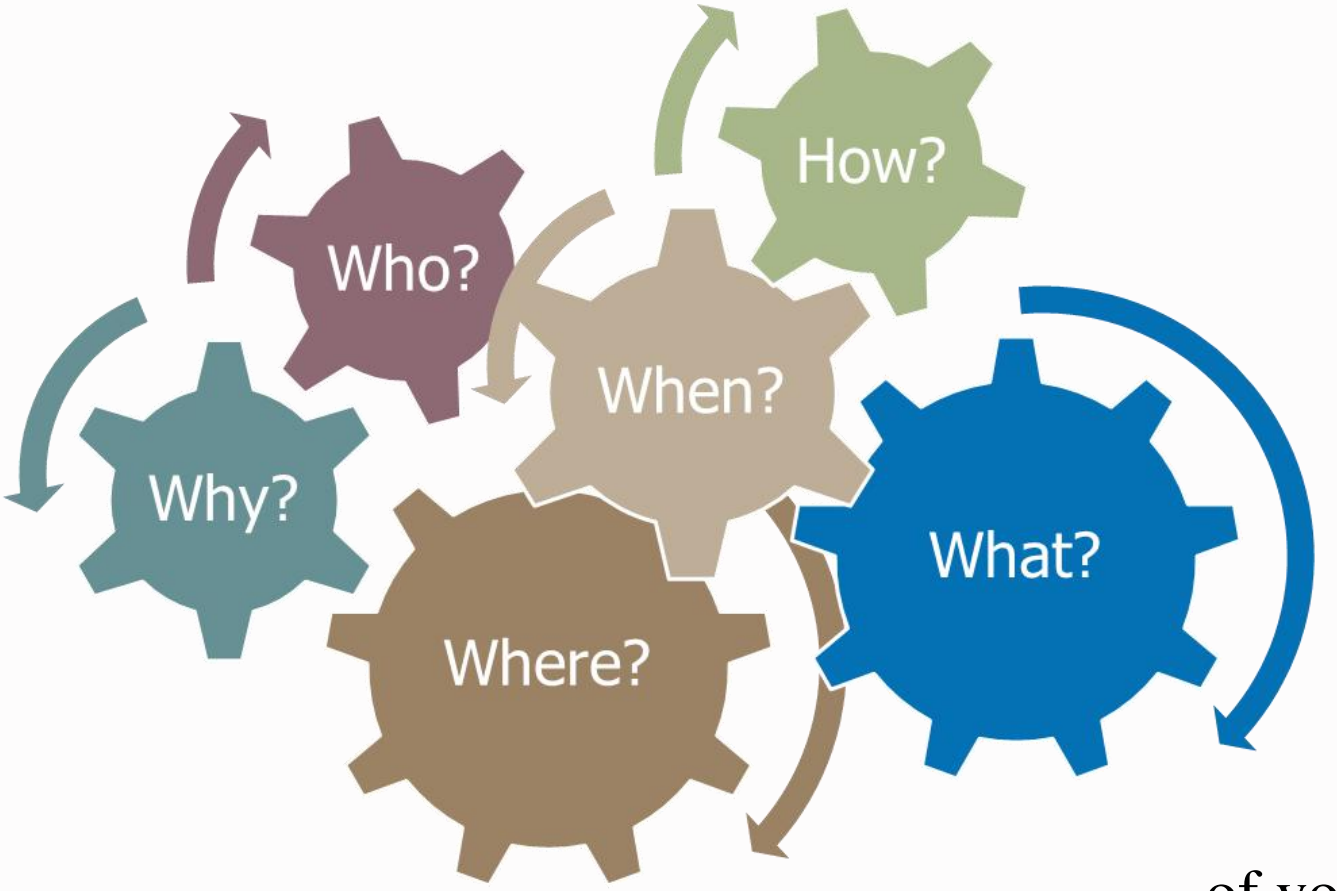
- Variety(种类多)
- Velocity(速度快)
- Value（价值密度低）
- Volume（大容量）

元数据的产生

对于这个海量而多样的数据信息。我们应该怎样高效而快速的进行处理和分析呢？这个时候元数据就应运而生了。元数据被用来定义数据的结构的数据。它汇总了当前数据的基本信息，使我们可以更轻松的查找、使用和重用数据实例。一言以蔽之就是描述数据的数据。在生活中的方方面面，只要有数据存在的地方，就有其数据对应的元数据。因为有元数据的存在，人们才能更好地理解数据，管理数据，更好的挖掘数据的价值。定义好元数据是进行数据分析和管理的前提。

什么是元数据

元数据是关于数据和信息的数据



of your data

数据的基本信息

Who	• 作者 • 管理者
Where	• 研究领域 • 获取数据
What	• 数据主要内容 • 使用了哪些源数据
How	• 数据的生成过程 • 数据的分布
When	• 数据产生时间 • 数据时效
Why	• 数据产生的原因 • 数据缺失原因

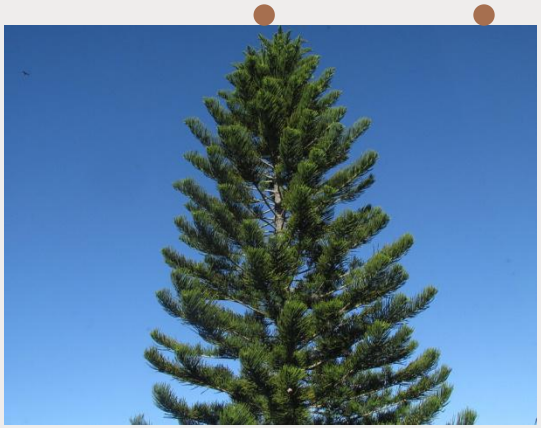
数据的标签


title
time period
abstract
supplemental information
(file) size
LOW-SODIUM
V8
100% VEGETABLE JUICE

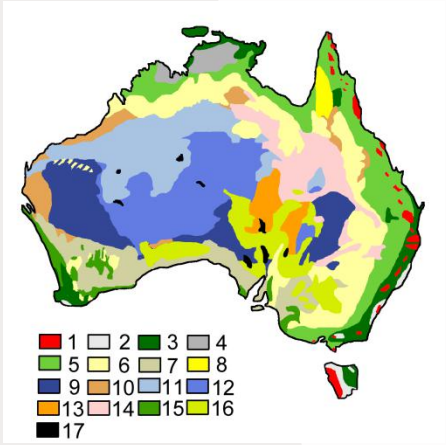
元数据为数据内容提供目录

元 数 据 是 数 据 的 一 部 分

DATA



content



context

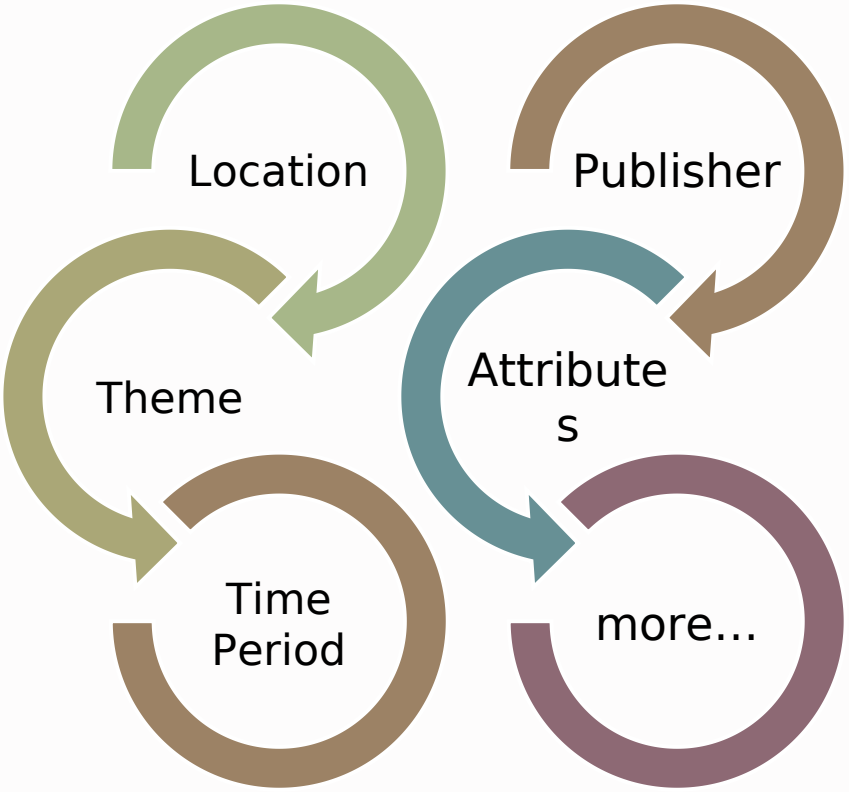
Metadata

Title
Responsible Party
Purpose
Abstract
Geographic Extent
Time Period of Content
Sources Used
Process Description
Positional Accuracy
Entities and Attributes…….

phenomena —*abstraction*→ **map / image / database** —*abstraction*→ **metadata**

元数据可用于数据发现

元数据使用户能够按顺序搜索数据:



几种常用的元数据标准

1. Dublin Core Metadata Initiative(都柏林核心集)

2. XML Schema（XML 文档的结构）

3. RDFS（RDF词表）

数据中台

数据中台是一套可持续“让企业的数据用起来”的机制，一种战略选择和组织形式，是依据企业特有的业务模式和组织架构，通过有形的产品和实施方法论支撑，构建一套持续不断把数据变成资产并服务于业务的机制，而元数据系统一般集成在数据中台中。

战略定位

企业需要从最高层的战略上明确数字化转型和建设中台的意图，这样才有可能真正将数据中台落地

组织保障

企业人力要提供配套的组织保障，包括以CEO、CIO、CTO、CDO为主的高层管理层配套中层管理层、基层执行层的全套组织体系，建立数据人才组织架构

一站式工具

选用适用、适配、成熟、完整的一站式大数据平台工具，利用工具对整个战略提供保障，并对全链路的数据采集、开发、质量和流程进行保证

元数据管理

数据增长速度的加快激发了人们对可从元数据中获得的潜在商业价值的新兴趣。存在各种数据结构，既带来机遇，也带来挑战。元数据管理框架应运而生，元数据管理框架提供了一个组织框架来协调存储在各种系统中的离散数据集。它还提供了描述信息的组织共识，通常分为业务、运营和技术数据。

元数据管理的核心是使人们能够使用基于Web等用户界面识别特定数据片段的属性。该属性可能是文件名、作者、客户 ID 号等。因此，请求文档的人能够查看和理解数据的不同属性、数据所在的企业系统以及创建这些属性的原因。

元数据管理

数据采集

- 数据的表结构信息
- 数据的空间存储，读写权限和其他统计数据
- 数据的血缘关系
- 业务数据

元数据管理

元数据管理平台-Apache Atlas

Atlas的架构方案应该说相当典型，基本上这类系统大致都会由元数据的收集，存储和查询展示三部分核心组件组成。此外，还会有一个管理后台对整体元数据的采集流程以及元数据格式定义和服务的部署等各项内容进行配置管理。

元数据管理

元数据管理平台-Apache Atlas

对应到Atlas的实现上，Atlas通过各种hook/bridge插件来采集几种数据源的元数据信息，通过一套自定义的Type 体系来定义元数据信息的格式，通过搜索引擎对元数据进行全文索引和条件检索，除了自带的UI控制台意外，Atlas还可以通过Rest API的形式对外提供服务。

元数据管理

元数据管理平台-Apache Atlas

总体而言，Atlas的实现，从结构原理的角度来说，还算是比较合理的，但从现阶段来看，Atlas的具体实现还比较粗糙，很多功能也是处于可用但并不完善的状态。此外Atlas在数据审计环节做的工作也不多，与整体数据业务流程的集成应用方面的能力也很有限。Atlas项目本身很长时间也都处于Incubator状态，截至2012-12-03，一直都在更新，但是目前功能还不够完善，因此还需要大家一起努力来帮助它的改进。对于一些自研项目或者是一个很好的开始，可以基于Apache Atlas针对自己的业务进行二次开发。

元数据管理

元数据管理平台-Cloudera Navigator Data Management

另外一个比较常见的解决方案是Cloudera CDH发行版中主推的Navigator，相比Atlas而言，Navigator的整体实现更加成熟一些，更像一个完整的解决方案，不过，Navigator并不是开源的。Navigator的产品定位是数据管理，本质上也是通过管理元数据来管理数据，但周边工具和配套设施相对完善，和Cloudera Manager管理后台的产品集成工作也做得比较彻底。相比Atlas来说，Navigator的整体组件架构也更加复杂一些。

元数据管理

元数据管理平台-Cloudera Navigator Data Management

总体而言，Navigator和Cloudera Manger的产品集成工作做得相对完善，如果你使用CDH发行版全家福套件来管理你的集群的话，使用Navigator应该是一个不错的选择。不过，如果是自主管理的集群或者自建的大数据开发平台，深度集成定制的Navigator就很难为你所用了，但无论如何，对于自主开发的元数据管理系统来说，Navigator的整体设计思想也还是值得借鉴的。

元数据展望

随着各个公司的业务不断扩展，数据源不断增多。每个业务都是相互独立的，很多工作都要从头做起，在计算机技术领域现象就是重复的造轮子。而元数据管理平台可以把这些重复的开发工作复用起来，上层业务可以直接基于元数据管理平台已有的功能去实现，这样达成目标的路径更短更高效，对于业务可以做到敏捷开发，对于需求可以做到快速响应。所以元数据管理平台在未来一定可以得到广大企业和组织的应用，如果可以实现一个可以普遍接入的元数据管理平台，在未来一定可以拥有广大的市场。

1. “Metadata.” Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/metadata. Accessed 7 Dec. 2021.
2. Zeng, Marcia. Metadata Types and Functions. NISO. 2004 [5 October 2016].（原始内容存档于2016-10-07）.
3. Directorate, OECD Statistics. OECD Glossary of Statistical Terms - Reference metadata Definition. stats.oecd.org. [2018-05-24].
4. National Information Standards Organization (NISO). Understanding Metadata (PDF). NISO Press. 2001 [2016-11-14]. ISBN 1-880124-62-9.（2014-11-07）.
5. Dippo, Cathryn. The Role of Metadata in Statistics (PDF). Bureau of Labor Statistics.
6. A Guardian Guide to your Metadata. theguardian.com. Guardian News and Media Limited. 12 June 2013 [2016-11-14].
7. ADEO Imaging: TIFF Metadata. [2013-05-20].
8. https://blog.csdn.net/colorant/article/details/79549785
9. https://blog.csdn.net/cafebar123/article/details/79944247
10. https://blog.csdn.net/xiaohai798/article/details/80868955
11. https://whatis.techtarget.com/definition/metadata