第二章 数据化运营的数据来源

“巧妇难为无米之炊”，对于数据工作者来说数据便是所有数据工作的基础。企业的数据化运营数据来源复杂，从数据结构类型上来说，包括**结构化**和**非结构化**数据；从数据来源方式来说，即有导出的**数据文件**、**数据库**，又有**流失**、**API**等复杂系统接口和**外部资源数据**。

2.1 数据化运营的数据来源类型

2.1.1 数据文件

数据文件是存储数据的文件，广义上，任何文件中存储的信息都可以称为数据；狭义上，数据文件中以数字或者文本形式存储的结构化的数据记录才是数据。

结构化数据记录大多数来源自数据库，也有来源自系统或者工具的工作过程或者返回结果。常见的数据格式类型包括：

- txt
- csv
- tsv
- xls
- xlsx
- xml
- html
- doc
- sql

2.1.2 数据库

数据库（DataBase）是按照数据结构来组织、存储和管理数据的仓库。数据库广泛应用于CMS、CRM、OA、ERP、DSS、财务系统、数据仓库和数据集市、进销存管理、生产管理、仓储管理等各类企业事务中。

数据库的主要应用包括：

- 数据的定义、存储、增加、删除、更新、查询等事务性工作
- 数据传输、同步、抽取、转换、加载等数据清洗工作
- 数据计算、关联查询、OLAP等分析型工作
- 数据权限控制、数据质量维护、异构数据库、多系统通信交互等工作

数据库按类型分为：

- 关系型数据库：DB2、Sybase、Oracle、PostgreSQL、SQL Server、MySQL等
- 非关系型数据库
  - 面向高性能并发读写的键值数据库：Redis、Tokyo Cabinet等
  - 面向海量文档的文档数据库：MongoDB、CouchDB等
  - 面向可扩展性的列式数据库：HBase、Riak等
  - 面向图结构的图形数据库：Neo4J、InfoGrid、Infinite Graph等

2.1.3 API

API（Application Programming Interface）是应用程序编程接口，数据化运营中的API通常分为服务型API和数据型API。

- 服务型API：可基于预定义的规则，通过调用API实现特定的功能。
- 数据型API：通过特定的语法，通过向服务器发送数据请求，返回特定格式的数据或者数据文件。

API通常返回的数据格式为XML或者JSON。

- JSON：一种轻量级的数据交换格式，由流行的JavaScript编程语言创建，多应用于Web数据交互，其格式简介、结构清晰，使用键值对的格式存储数据对象。
- XML：一种可扩展标记语言，听统一的方法来描述和交换独立于应用程序或者供应商的结构化数据。

2.1.4 流式数据

流式数据是指实时或者接近实时处理的大数据流。常见的流式数据处理使用Spark、Storm和Samza等框架。流式数据常见的系统例如：

- 在线个性化推荐系统
- 网站用户实时行为采集和分析
- 物联网机器日志实时分析
- 金融实时消费反欺诈
- 实时异常人员识别等

按照数据对象来区分，流式数据可分为两大类:

- **第一类是用户行为数据流。**主要围绕“人”产生的数据流，例如网站或者APP内部因为浏览、搜索、评论、分享、交易等等产生的数据流。
- **第二类是机器数据流。**主要围绕“物”产生的数据流，例如机器在运行过程中产生的日志数据、物联网传感器的监控数据等等。

2.1.5 外部公开数据

外部公开数据指公开的任意第三方都能获取的数据。

2.1.6 其他数据

- 调查问卷
- 购买
- 合作

结束

本章无代码编写内容，从下章开始将会有大量的代码部分。