数据治理(Data Governance)是组织中涉及数据使用的一整套管理行为，是对数据资产管理行使权力和控制的活动集合，通过一系列信息相关的**过程**来实现决策权和职责分工。过程即谁(Who)能根据什么信息，在什么时间(When)和情况(Where)下，用什么方法(How)，采取什么行动(What)。

我认为数据治理是我们一整个项目的大框架，涉及：

**元数据，**描述数据的数据，可以是表的名字，schema，总行数、存储位置、历史数据、资源查找、文件记录等，是一种电子式目录。在数据仓库系统中，包括**描述哪些数据在数据仓库中**的元数据、**定义要进入数据仓库中的数据和从数据仓库中产生的数据**的元数据、**记录数据抽取工作时间安排(根据业务事件进行的)**的元数据、**记录并检测系统数据一致性和执行情况**的元数据、**衡量数据质量**的元数据。

**Idea**:我认为，在本项目中，元数据可以涉及数据处理过程(抽取、转换、载入)每个阶段的时间、被处理的数据条数、表结构变化，从Amozon网站中筛选电影信息和非电影信息时通过某个指标(比如类型、时长、名称)筛除的数据条数，查询预处理时合并、分离了哪些数据，这些可以作为**溯源查询**的依据；在用户的查询行为也可以产生出一些元数据，包括查询某些相同、类似信息的频度(次数)，查询发生的普遍时间等，用于分析用户行为，即用户最感兴趣哪些信息，用户喜欢在什么时间点访问网站等等；三种存储模型上的执行时间也属于元数据，用于分析系统性能，作为存储优化的依据。

**数据血缘**：数据在产生、处理、流转到消亡过程中，数据之间形成的一种类似于人类社会血缘关系的关系。包括的特征是**归属性**(特定的数据归属特定的组织或者个人)、**多源性**(同一个数据可以有多个来源，一个数据可以是多个数据经过加工而生成)、**可追溯性**(数据的生命周期，数据从产生到消亡的整个过程)、**层次性**(对数据的分类、归纳、总结等数据描述信息构成层次)。

**Idea**:关于数据归属性，可以分析数据分布在哪台服务器、属于哪个数据库；关于数据的多源性，可以分析电影数据经过哪些加工生成的；可追溯性，记录数据处理、筛选、产生的过程，用于溯源查询；关于层次性，可以分析存储在数据库中的结构化数据血缘关系的层次结构(所有者-数据库-表-字段)、文件服务器存储的数据的层次机构(所有者-文件服务器-文件目录-文件)。

**数据血缘的可视化**，通过绘图(数据血缘关系图)的方式，描述数据处理过程的信息节点、数据流转线路、清洗规则的节点、转换规则的节点、归档销毁规则的节点。

详细参见https://www.sohu.com/a/161142366\_99934777

**数据血缘的使用场景**：数据溯源，评估数据价值(从数据受众、数据更新量级、数据更新频次几个方面进行)，数据质量评估(从数据血缘关系图中分析数据清洗的标准清单)，数据归档、销毁(处理无用数据)。

**数据质量**，准确性如何、实时性如何、一致性如何、相关性如何，对于数据质量的衡量可以涉及一个点或多个点，取决于数据使用者所要达成的目的。如：关于查询，要考虑查询结果的准确性和出结果的时间，即实时性；当优化存储，降低查询时间时，要考虑到数据之间的相关性，数据之间的相关性程度也意味着针对某种查询，比较适合何种存储模型；当存储数据时，要考虑数据的一致性。

**如何保证数据质量**：详细参见

https://www.esensoft.com/industry-news/data-governance-2155.html

**数据资产编目**：相当于一个数据目录，包括数据来源、数据内容、数据分布、管理部门、使用对象，有利于数据资产的快速浏览，来告诉数据消费者都有哪些数据、数据在哪和数据是什么样的。主要受众是ETL工程师、系统工程师、BI工程师。

具体数据资产目录如何编织参见

http://www.360doc.com/content/20/0410/14/39716884\_905055882.shtml

**数据安全**：涉及**数据梳理与识别**(明确保密和敏感数据的分布情况)、**数据安全认责**(采用“谁生产、谁管理、谁负责”的数据认责原则进行数据归属权确认)、**数据分类和分级**(依据数据的来源、内容和用途对数据资产进行分类，根据数据的价值、敏感程度、影响范围进行敏感分级)，**数据的访问授权**(角色只可查询某些数据，对其他数据不可见)、**数据安全的生命周期管理**(在数据的规划、设计、创建、存储、使用、销毁的各个阶段应设置相应的管控点和管理流程，这个可以结合数据血缘关系图)。

**Idea**:关于数据安全方面，我能想到的是哪些数据开放查询，而哪些数据不开放查询，比如一些分析用户行为的数据不开放查询，如查询某些相同、类似信息的频度(次数)，查询发生的普遍时间等。或者可以从不同的查询，返回不同的字段来体现，比如关于电影的查询返回哪些字段、屏蔽(不显示)哪些字段，关于导演、演员的查询又返回哪些字段、屏蔽(不显示)哪些字段。

**成本管理**：避免重复建设(减少冗余)、减少低效sql、表的生命周期管理。

**数据可用性**：数据备份、灾难可抗拒、易用可维护、兼容可扩展等。数据的一致性、准确性、完整性、时效性及实体同一性。

参考的博客：<https://www.sohu.com/a/161142366_99934777>

<http://www.360doc.com/content/20/0410/14/39716884_905055882.shtml>

<https://blog.csdn.net/kuangfeng88588/article/details/99689979>

<https://www.esensoft.com/industry-news/data-governance-2155.html>