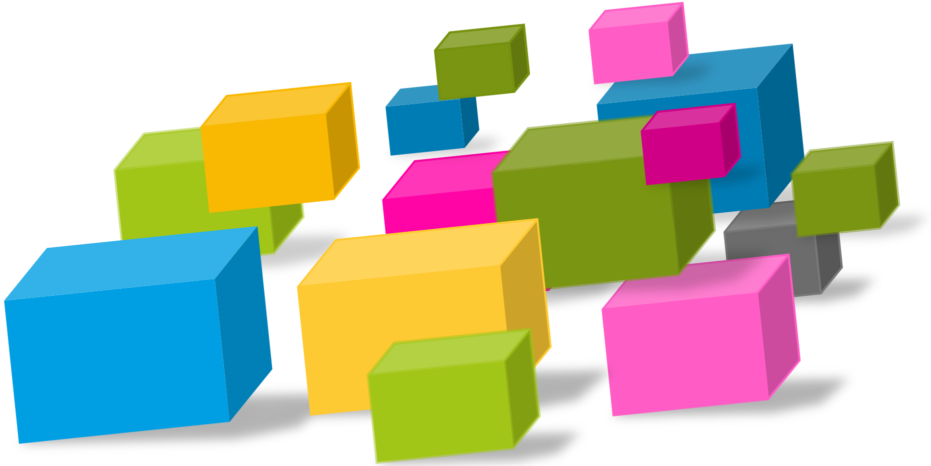
Data Types and Principles

Kim Horn

Version 0.31 13 December 2021



Types of Data

Types of Data

- Transactional Data
- Historical Data
- Reference Data
- Master Data
- Metadata
- Multi-Hop Data

Transactional Data

- The data a companies everyday Business Transactions and Processes produce.
- It is fine-grained information that represents the details of an enterprise enabling the main business capabilities and use cases.
- Examples: accounts, sales, orders, inventory, invoices, logistics, fulfillment.
- This data is normally stored as a current state, e.g. current bank account balance. The values do change over time, and past values are usually lost.
- Transactional systems may emit events, that are of business importance to users or other systems, e.g. email for bank account in debit.
- These system usually apply ACID properties to their data rather than BASE.
- In most cases the system are optimised for transaction rate rather than add hoc search and queries.
- Organisations may receive source data from suppliers, that gets processed to provide higher quality data. That data may be transactional, including updates, deletions and inserts.

Master Data

- Master Data represents the common business objects that need to be agreed on and shared throughout an enterprise.
- It is required to support key capabilities and is used across business processes, organisations, between operational systems, and decision support systems.
- It is some of the most valuable data that a business owns, and so may be sold as a product.
- It can be viewed as key core data that is used across the enterprise to facilitate the business capabilities, for example it may contain core information for, suppliers, customers, products and accounts.
- In many cases this data is kept in many different repositories, of unknown quality. Knowing what data is up to data or of known quality is complex. Users may not know where to source data. These issues make it difficult for organisations to change, systems (architectures) become brittle, cost becomes prohibitive, and agility is lost. It can provide considerable benefit if this core data is managed.
- Master data captures the key things that all parts of an organization must agree on, both in meaning and usage. Managing its metadata is critical.
- A single source of metadata offers:
    - An authoritative source of information
    - Ability to use the information in a consistent way
    - Ability to evolve the master data, and management of the master data, to meet new or changed business needs.

Historical Data

- Represents the accumulation of transaction and master data over time. Commonly the record of state changes.
- Most often used for analytics, insight discovery, reporting, audit and regulatory compliance
- Usually transformed to reorganize the data by subject.
- Events may be collected recording important business happenings or the state changes of systems over time.
- Databases normally just store current state data but record history separately from the database as change logs.
- Usually loaded into a data warehouse, from the source systems, via ETL from CDC.
- Transactional data that is streamed produces a historical feed of data.

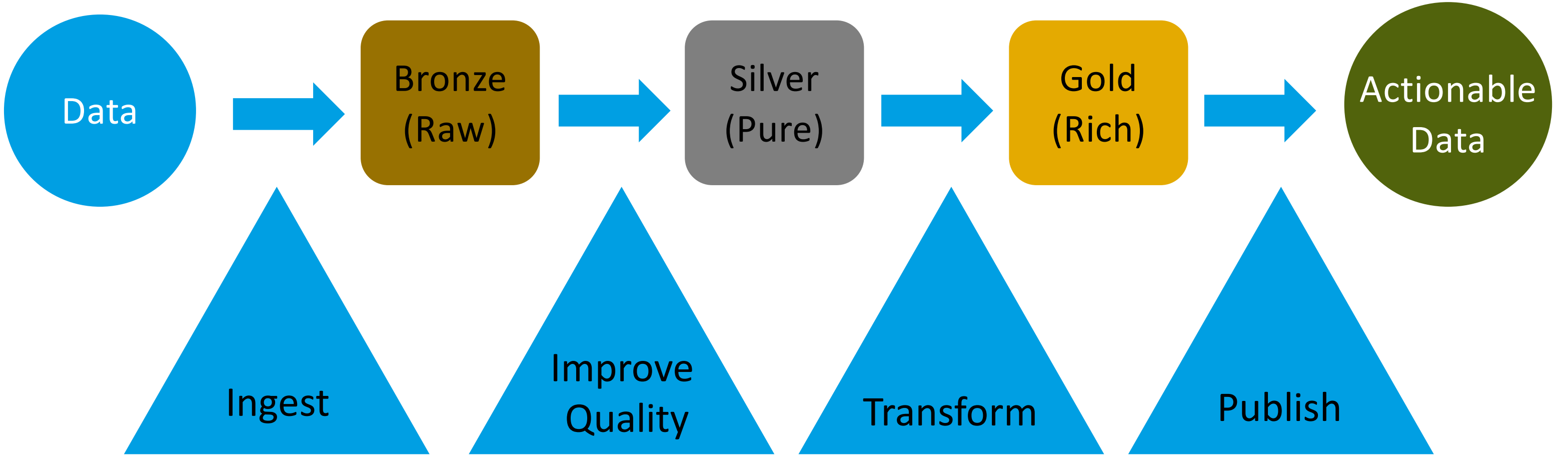
Reference Data

- Reference data is focused on defining and distributing collection of agreed valid common values for things such as codes and abbreviations.
- It is used to ensure common and consistent values for attributes of master data.
- It enables accurate and efficient processing of operational and analytical activities by enabling processes to use the same defined values of information
- Examples: Country codes, Color codes, Product codes. Models, Identity types, Account types, subscriptions, stock codes.
- There are many industry standards that provide agreed sources of reference data.

Meta Data

- Descriptive information, data, that is useful for people or systems to understand and describe data.
- It can be created, managed, stored , and preserved like any other data.
- Metadata generally has little value on its own. It only has meaning in context of the data it describes.
- There are three kinds of metadata:
  1. **Descriptive metadata** consist of information about the content and context of your data. Examples: title, creator, subject keywords, and description (abstract)
  2. **Structural metadata** describe the physical structure of compound data. Examples: camera used, aperture, exposure, file format, and relation to other data
  3. **Administrative metadata** are information used to manage your data. Examples: when and how it was created, who can access it, software required to use it, and copyright permissions
- Benefits of Metadata:
  - **Finding Data:** it much easier to find relevant data. Most searches are done using text (like a Google search), so formats like audio, images, and video are limited unless text metadata is available. Metadata also makes text documents easier to find because it explains exactly what the document is about.
  - **Using Data:** To use a dataset, you need to understand how the data is structured, definitions of terms used, how it was collected, and how it should be read.
  - **Re-using Data:** To re-use data it needs to be found and used, but often at a higher level of trust and understanding. Re-using data often requires careful preservation and documentation of the metadata.

Multi-Hop - Data Types



Multi-Hop Architecture - Data Types

A common Data Lake architecture that uses ‘Tables’ that correspond to different quality levels in the data engineering pipeline, progressively adding structure to the data. It is an improvement on the Lambda Architecture.

It relies of 3 stages of processing that can combine batch and streaming workflows:

- **“Bronze” tables** - data ingestion, the raw data, the **“single source of truth”,** the data lake. Often called ‘Landed’ data, or ‘Raw’ data.
- **“Silver” tables** – cleaned, filtered, transformation and feature engineering performed. The data may be modelled into canonical formats. Has gone through operational processes to ensure data quality, security, metadata and other business defined SLAs are met. Often called ‘Pure’, ‘Sandbox’ or ‘Pond’ Data
- **“Gold” tables** - provide business level aggregates often used for reporting and dashboarding. You can join it with other datasets, apply custom functions, perform aggregations, implement machine learning, and more. The goal of this step is to get *rich data*, the fruit of your analytics work. It is finalised data ready for consumption, e.g. for machine learning training or prediction. Often called ‘Production’ or ‘Refined’ Data

Note the ‘Table’ in this case is a special structure based on a Data Lake approach to storing data. It is not a relational DBMS table.

Multi-Hop Architecture - Data Types

A common Data Lake architecture that uses ‘Tables’ that correspond to different quality levels in the data engineering pipeline, progressively adding structure to the data. It is an improvement on the Lambda Architecture.

It relies of 3 stages of processing that can combine batch and streaming workflows:

- **“Bronze” tables** - data ingestion, the raw data, the **“single source of truth”,** the data lake. Often called ‘Landed’ data, or ‘Raw’ data.
- **“Silver” tables** – cleaned, filtered, transformation and feature engineering performed. The data may be modelled into canonical formats. Has gone through operational processes to ensure data quality, security, metadata and other business defined SLAs are met. Often called ‘Pure’, ‘Sandbox’ or ‘Pond’ Data
- **“Gold” tables** - provide business level aggregates often used for reporting and dashboarding. You can join it with other datasets, apply custom functions, perform aggregations, implement machine learning, and more. It is finalised data ready for consumption, e.g. for machine learning training or prediction. Often called ‘Rich’, ‘Production’ or ‘Refined’ Data

Note the ‘Table’ in this case is a special structure based on a Data Lake approach to storing data. It is not a relational DBMS table.

Data Principles

Principle 1: Data is a Valued Asset

Statement: Data is an asset that has value to the enterprise and is managed accordingly.

Rationale: Data is a valuable corporate resource; it has real, measurable value. In simple terms, the purpose of data is to aid decision-making for the enterprise; it is a source of knowledge. Accurate, timely data is critical to accurate, timely decisions. Most corporate assets are carefully managed, and data is no exception. Data is the foundation of our decision-making, so we must also carefully manage data to ensure that we know where it is, can rely upon its accuracy, and can obtain it when and where we need it.

Implications:

- Since data is an asset of value to the entire enterprise, data stewards accountable for properly managing the data must be assigned at the enterprise level
- Stewards must have the authority and means to manage the data for which they are accountable
- We must make the cultural transition from "data ownership" thinking to "data stewardship" thinking
- The role of data steward is critical because obsolete, incorrect, or inconsistent data could be passed to enterprise personnel and adversely affect decisions across the enterprise
- Part of the role of data steward, who manages the data, is to ensure data quality.
- Procedures must be developed and used to prevent and correct errors in the information and to improve those processes that produce flawed information. Data quality will need to be measured and steps taken to improve data quality - it is probable that policy and procedures will need to be developed for this as well.
- A forum with comprehensive enterprise-wide representation should decide on process changes suggested by the steward

Principle 2: Data is a Product

Statement: As Data is an asset it can be managed as a product. It has value to our customers.

Rationale: Data is a valuable resource not only for the enterprise but also for our customers. Customers can use our data to improve their decision making, operations, and to create new products. The processes of an enterprise refine and transform data and produce more value from that data. Data has generative effect producing more data. All this data has immense value. Data is not only the foundation of our decision-making, but of our customers.

Implications:

- As an example, a history of order data can be used to generate purchase recommendations. The rules for these recommendations are new data that can be sold as a product. Customers can now create their own recommendations, as a product for their customers.
- Data quality and validity becomes more important, and needs to be measurable and reportable.
- Data products are created as side effects of a companies operations or intentionally via data science workflows, specifically through the application of models, usually predictive or inferential, to a domain-specific dataset.
- Data will need a mechanism for publication and transmission to clients. This may involve real-time, streaming or batch distribution.
- Data as a product will need to be versioned. The product will be improved, upgraded, and changed over time.
- Generated data has to be carefully selected for sharing, as it can easily become unwieldy, loose its integrity and become too costly to manage.

Principle 3: Data is Shared and Re-used

Statement: Users and customers have access to the data necessary to perform their duties; therefore, data is shared across enterprise functions and organisations. Data is re-used across the enterprise for other purposes.

Rationale: Timely access to accurate data is essential to improving the quality and efficiency of enterprise decision-making. It is less costly to maintain timely, accurate data in a single application, and then share it, than it is to maintain duplicative data in multiple applications. The enterprise holds a wealth of data, but it may be stored in isolated and incompatible stovepipes. The speed of data collection, creation, transfer, and assimilation is driven by the ability of the organization to efficiently share these islands of data across the organization. Shared data will result in improved decisions since we will rely on fewer (ultimately one virtual master) sources of more accurate and timely managed data for all of our decision-making.

Implications:

- To enable data sharing we must develop and abide by a common set of policies, procedures, and standards governing data management and access for both the short and the long term.
- We must adopt common methods and tools for creating, maintaining, and accessing the data shared across the enterprise.
- We will also need to develop standard data models, data elements, and other metadata that defines this shared environment and develop a repository system for storing this metadata to make it accessible.

Data sharing will require a significant cultural change.

- This principle of data sharing will continually "bump up against”:
  - The principle of data security - under no circumstances will the data sharing principle cause confidential data to be compromised.
  - The principles of autonomy and of decoupling – sharing data can result in coupling systems, that then looses their autonomy.
- Data made available for sharing will have to be relied upon by all users to execute their respective tasks. Shared data will become the enterprise-wide "virtual single source" of data.

Principle 4: Data is Fit for Purpose

Statement: Data must be of known quality and fit for both its primary purpose and potential secondary/tertiary uses.

Rationale: The quality of data will affect the outcomes of its use. Quality includes factors such as accuracy, validity, reliability, timeliness, relevance and completeness. These factors need to be observable and managed. Data does not necessarily need to be good or high quality, but that quality needs to be known, and suitable for the use. Systems may be able to deal with poor quality data and refine it. Indeed, data may progress along stages from low to medium to high quality, suitable for different purposes along the way.

Implications:

- The quality of information should be regularly monitored to ensure that it at least meets the level that has been assessed as necessary for its purpose(s).
- The data quality is managed and captured as metadata.
- The purpose of the data drives its sharing and defines its value as an asset.
- QA and test data needs to be identified and differentiated from production or customer data. Metadata allows it to be easily identified and removed, if it is not relevant to the purpose.

Principle 3: Data is Accessible

Statement: Data is accessible for users to perform their functions.

Rationale: Wide access to data leads to efficiency and effectiveness in decision-making and affords a timely response to information requests and service delivery. Using information must be considered from an enterprise perspective to allow access by a wide variety of users. Staff time is saved and consistency of data is improved.

Implications:

- Accessibility involves the ease with which users obtain information
- The way information is accessed and displayed must be sufficiently adaptable to meet a wide range of enterprise users and their corresponding methods of access
- Access to data does not constitute understanding of the data - personnel should take caution not to misinterpret information
- Access to data does not necessarily grant the user access rights to modify or disclose the data, or rights to change the systems providing that data.
- Access to data by separate teams (organisations) or systems is via:
  - Agreed and managed contracts. These contracts specify Qualities of Service of the data, and the working relationship of the provider and consumer.
  - Agreed Integration patterns, that do not break the provider or consumer systems or breach their principles.

Principle 4: Data has a Trustee

Statement: Each data element has a trustee accountable for data quality.

Rationale: One of the benefits of an architected environment is the ability to share data across the enterprise. As the degree of data sharing grows and business units rely upon common information, it becomes essential that only the data trustee makes decisions about the content of data. Since data can lose its integrity when it is entered multiple times, the data trustee will have sole responsibility for data entry which eliminates redundant human effort and data storage resources.

Note: A trustee is different than a steward - a trustee is responsible for accuracy and currency of the data, while responsibilities of a steward may be broader and include data standardization and definition tasks.

Implications:

- Real trusteeship dissolves the data "ownership" issues and allows the data to be available to meet all users' needs This implies that a cultural change from data "ownership" to data "trusteeship" may be required.
- The data trustee will be responsible for meeting quality requirements levied upon the data for which the trustee is accountable
- It is essential that the trustee has the ability to provide user confidence in the data based upon attributes such as "data source"
- It is essential to identify the true source of the data in order that the data authority can be assigned this trustee responsibility This does not mean that classified sources will be revealed, nor does it mean the source will be the trustee.
- Information once captured should be immediately validated as close to the source as possible.
- Quality control measures must be implemented to ensure the integrity of the data.
- As a result of sharing data across the enterprise, the trustee is accountable and responsible for the accuracy and currency of their designated data element(s) and, subsequently, must then recognize the importance of this trusteeship responsibility

Principle 5: Data never changes. It is Immutable.

Statement: Data as an enterprise asset is stored as it is, it should never change

Rationale:. Transactional systems change (update and delete) the current state of entities. The history and change of those entities is lost. Instead, data should only be stored and never mutated. Updates becomes inserts of a new version of the data. These new versions are collected as a sequence of immutable events. This log of events is the primary data, and the current state is reconstructed by replaying the logs. With immutable data returning to the past is an inherent aspect of the data. By making it immutable, it inherently takes care of human fault tolerance to at least some extent and takes away errors with regards to data loss and corruption. It allows data to be selected, inserted, and not updated or deleted.

Transactional systems as a side effect, store a log of the change events, for example Change Data Capture (CDC), rather than storing the current state primarily. With transactional systems returning to a point in the past requires external operational functions.

Implications:

- For example, a bank account can be modelled as a sequence of events, such as a debit or credit, with a dollar amount and a timestamp. A credit of $10.00, on 1st July never changes. Summing these events over time produces the current balance. The balance never changes but is a function of immutable data.
- Data should be stored in its raw format from source systems.
- Immutable data allows you to track its lineage. To do this the processes using the data need to be observable, allowing metadata to be captured, e.g. a timestamp, sequence number
- Immutable data means that data is not lost, and future new use cases can be catered for.
- Immutable data means that processes (operations) on that data should be idempotent; the same input will consistently produce the same output (no side effects). Functional programming facilitates the ability to make data processes reproducible.
- Immutable data provides many advantages: all data is forever accessible, system behaviour is predictable, and systems easier to manage

Principle 6: Data is Understood

Statement: Data has known and observable quality and lineage, and its context is understood.

Rationale: Data should have known quality, its authenticity and validity known. To do this its lineage must be understood, and the context it was created in understood. To do this metadata is critical to know the accuracy of data, its completeness, its consistency, timeliness, and relevancy. Data of know quality can be trusted. Master data provides a know level of trust across the organisation.

Implications:

- Metadata needs to be managed providing quality categorisations. For example, data may be given Bronze, Silver or Gold meta-labels, classifying its quality. These may combine completeness, consistency and accuracy.
- At a minimum metadata needs to be managed about the history and lineage of the data, when (timestamp) it was produced by whom (process or supplier), and its context (what is the domain concept).
- Data lineage on its own is the ability to trace and understand why data was mutated a certain way at a specific step in a process. Data may be obtained as raw data from suppliers, refined and transformed. The ability to track the lineage of data is important.
- Data will change over time and the concepts may change. Legacy systems producing data will be replaced and so will their data. Recording the legacy status of the data allows it to be transformed and used correctly.
- One approach to lineage is to collect metadata that tags each data record with its input source(s) and which version of code processed it. This creates a graph of which row record points to which previous data sources and so on.
- Over time data models change, their concepts mutate and are modified. Tracking the lineage of data over time using metadata models will allow this to be understood and mapped.
- Understanding the data will ensure that only the most accurate and timely data is relied upon for decision-making.

Principle 7: Data is Bound to a Context

Statement: Data is not isolated or universal conceptually, it has meaning only within a context; a domain.

Rationale: A piece of data has numerous assumptions behind it, it exists in a domain, that defines the context and its meaning. To understand data the domain has to be also understood, and to this this it must be bounded. When data meaning starts to change then you have approached the boundary of the domain. Data in one context may not be relevant in another; it may not be applicable to a different matter in another domain. Understanding the context allows data to be re-used appropriately across the organisation.

Implications:

- When data moves from system to system its context changes, and so does its meaning.
- The language used to define the data may be bound to the context.
- The desire for Canonical Data can be misleading; In many cases there is no universal form of data. For example, consider 'Account' data; a savings account is different to an investment account that is different to an IT account. Creating a universal 'Account' may create ineffective and overly complex data, that is no longer suited to a domain.

Across Principle Implications

The three principles :

- data is an asset;
- data is shared;
- data is easily accessible

Are highly related.

The implication is that there is an education task to ensure that all organizations within the enterprise understand the relationship between value of data, sharing of data, and accessibility to data.

Legacy Systems

As legacy systems are replaced, we must adopt common data access policies and guidelines to ensure that data in new applications remains available to the shared environment and that data in the shared environment can continue to be used by the new applications.