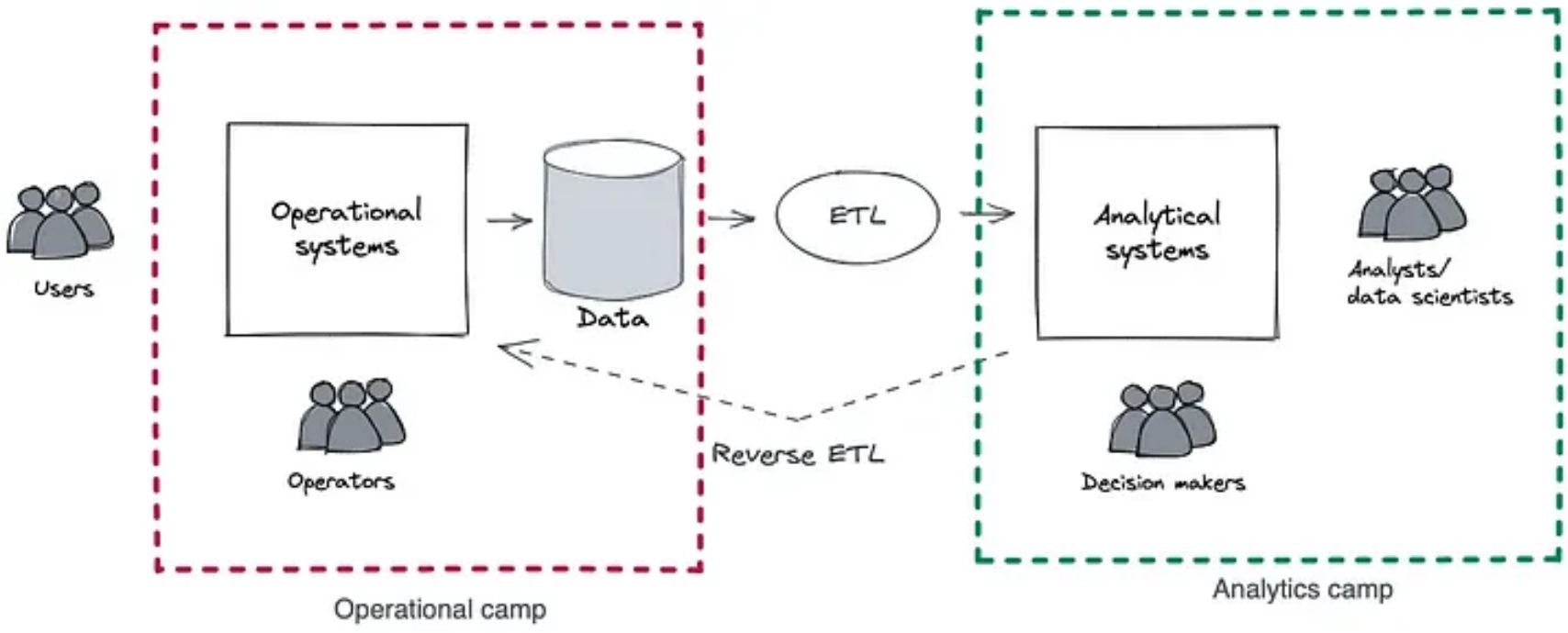
WHAT ARE DATA CONTRACTS?

Why do we need data contracts?

In any data-driven organization, we can identify two camps: the operational camp and the analytics (data) camp.



The **operational camp** consists of the machinery and operators interacting with customers and end-users.

These interactions generate data as a byproduct, which we refer to as operational data.

The **analytical camp** includes the people and infrastructure necessary to process operational data and generate insights.

These insights are then utilized by the operational team to enhance performance and grow the business.

How operational data is moved to the analytics camp?

Extract, transform, and load (ETL) jobs replicate operational data to the analytical data infrastructure, preventing analytical queries from consuming compute and storage resources on operational data stores, which are needed to generate revenue for the business.

WHAT ARE DATA CONTRACTS?

The great data divide

However, these two groups often struggle with issues of data ownership, visibility, and lack of trust.

The ETL pipeline was broken due to an unexpected upstream change done by the operational team.

Characters from https://gramener.com/comicgen/

Lack of data ownership - The software developers who build operational systems don't take ownership of the data they produce. They are also not very aware of the data dependencies on the analytics side and are not responsible for maintaining them. So, an upstream change on the operational side could potentially break several systems on the analytical side.

Downstream data quality issues - This happens when the data being brought into the analytics side, say into a data warehouse, isn't in a format that is usable by data consumers.

This divide can lead to inconsistencies, misunderstandings, and inefficiencies when it comes to how data is handled and used within an organization.

WHAT ARE DATA CONTRACTS?

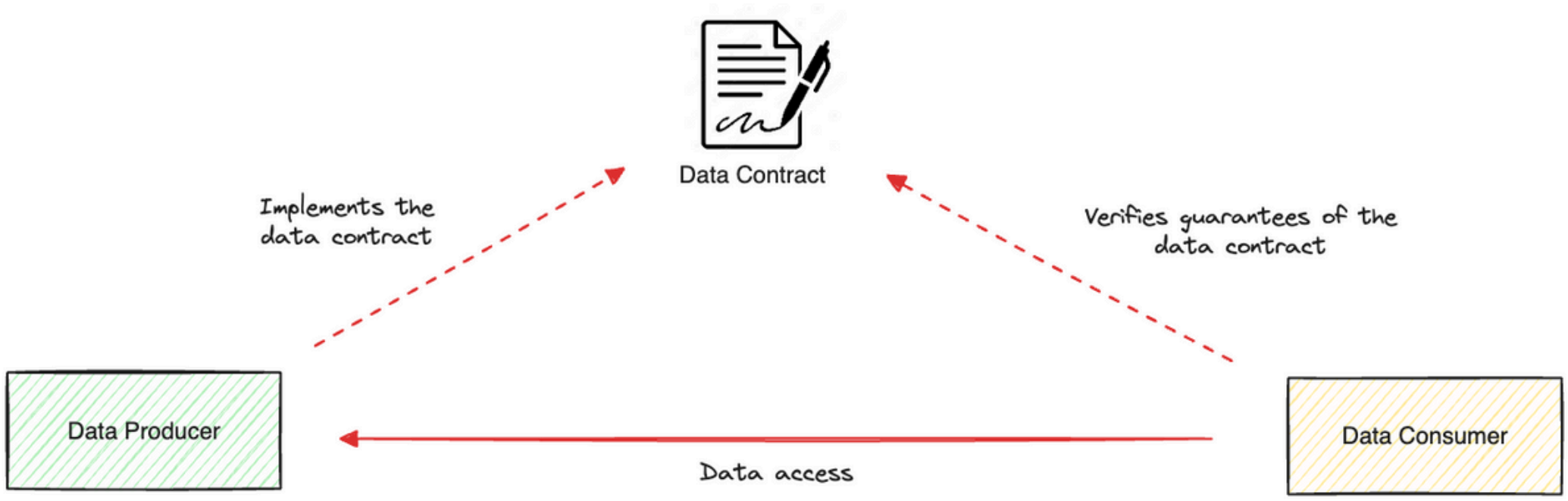
So, how can we bridge this divide? Can we have a clear and concise agreement to ensure that both parties have a shared understanding of the data's format, quality, and usage rules?

This is where the data contracts come in.

What is a data contract?

A data contract is a formal agreement between data producers and data consumers that defines the expectations and requirements for the data being shared. Think of it as an API but for data.

The aim? To ensure all parties involved have a clear understanding of the data's format, quality, and the rules governing its use.



In reality, a data contract is a machine-readable document formatted in YAML, JSON, or a similar kind. Despite the existence of an industry standard for data contracts, it's ultimately up to the data producers and consumers to agree on a format with which they are comfortable.

A schema registry is the closest example of data contracts applied in the real world.

WHAT ARE DATA CONTRACTS?

The following is an extract of a data contract for successful customer orders in the webshop.
We took it from:
https://datacontract.com/examples/orders-latest-nested/datacontract.html

```yaml
dataContractSpecification: 0.9.3
id: urn:datacontract:checkout:orders-latest-nested
info:
  title: Orders Latest (Nested)
  version: 1.0.0
  description: "Successful customer orders in the webshop. \nAll orders since 2020-01-01.\
    \ \nOrders with their line items are in their current state (no history included).\n"
  owner: Checkout Team
  contact:
    name: John Doe (Data Product Owner)
    url: https://teams.microsoft.com/l/channel/example/checkout
terms:
  usage: 'Data can be used for reports, analytics and machine learning use cases.

    Order may be linked and joined by other tables

    '
  limitations: 'Not suitable for real-time use cases.

    Data may not be used to identify individual customers.

    Max data processing per day: 10 TiB

    '
  billing: 5000 USD per month
  noticePeriod: P3M
models:
  orders:
    description: One record per order. Includes cancelled and deleted orders.
    type: table
    fields:
      order_id:
        ref: '#/definitions/order_id'
        title: Order ID
        type: text
        format: uuid
        required: true
        primary: true
        unique: true
        description: An internal ID that identifies an order in the online shop.
        pii: true
        classification: restricted
        example: 243c25e5-a081-43a9-aeab-6d5d5b6cb5e2
      order_timestamp:
        type: timestamp
        required: true
```

WHAT ARE DATA CONTRACTS?

What makes a data contract? What's in it?

In addition to general agreements about intended use, ownership, and provenance, data contracts include agreements about:

The explicit structure of the data being shared. This includes the organization, format, and type of data. The schema provides a blueprint that outlines how the data is organized and the type of data that can be expected in each field.

Semantics refers to the meaning of the data. They capture the rules of each business domain, including the definitions of each field, their possible values, and their relationships with other fields. Understanding the semantics is crucial in correctly interpreting the data.

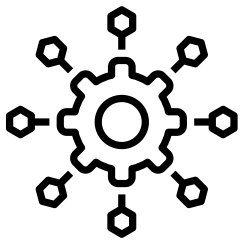
For example, in a dataset of e-commerce orders, the fulfillment date can never be earlier than the order date.

Commitments about the availability and freshness of data in a data product. An example SLA would be the maximum expected delay (in the case of the real-time data stream) for late-arriving events.

Additional information that helps describe, locate, or manage data. Metadata could include details about when and how the data was collected, who owns it, how it should be referenced, and any restrictions on its use. This information is crucial for ensuring proper data governance, as it helps data consumers understand the context and restrictions of the data they are using.

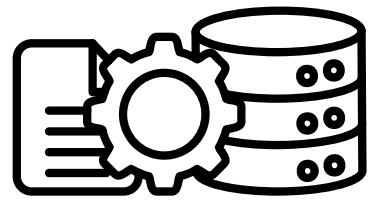
WHAT ARE DATA CONTRACTS?

Data contracts in practice



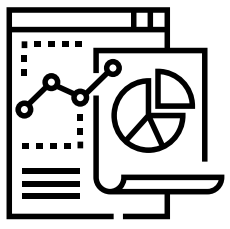
APIs and Microservices Communication

Data contracts ensure that all services agree on the structure, format, and meaning of the data being exchanged, preventing integration issues and reducing the likelihood of bugs caused by schema mismatches.



Data Integration and ETL Processes

Data contracts provide clear definitions and standards for the data, ensuring that reports and analyses are based on accurate and consistent data, leading to more reliable and actionable insights.

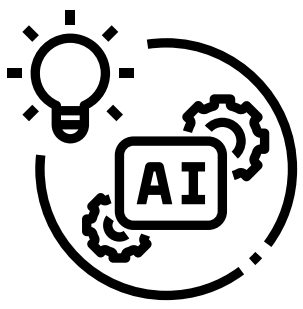


Analytics and Reporting

Data contracts help standardize the data ingested from different sources, ensuring that the data follows a consistent schema and quality standards, which simplifies the transformation and loading processes.

Regulatory Compliance

Data contracts define the rules and constraints around data usage, storage, and access, ensuring that data handling practices comply with legal and regulatory requirements. This enhances data governance and reduces the risk of non-compliance.



Machine Learning and Data Science

Data contracts ensure that the data used for training and inference is of high quality and consistent, which improves model accuracy and reliability. They also help in tracking and maintaining the lineage of data, ensuring the reproducibility of experiments and results.

Follow Me @dunithd /in/dunithd/