Business intelligence (BI)

Automating processes and information channels in order to transform relevant data into actionable insights that are easily available to decision-makers

Application programming interface (API)

A set of functions and procedures that integrate computer programs, forming a connection that enables them to communicate

forming a connection that

Data warehousing specialists

People who develop processes and procedures to effectively store and organize data

Data governance professionals

People who are responsible for the formal management of an organization's data assets

Data life cycle

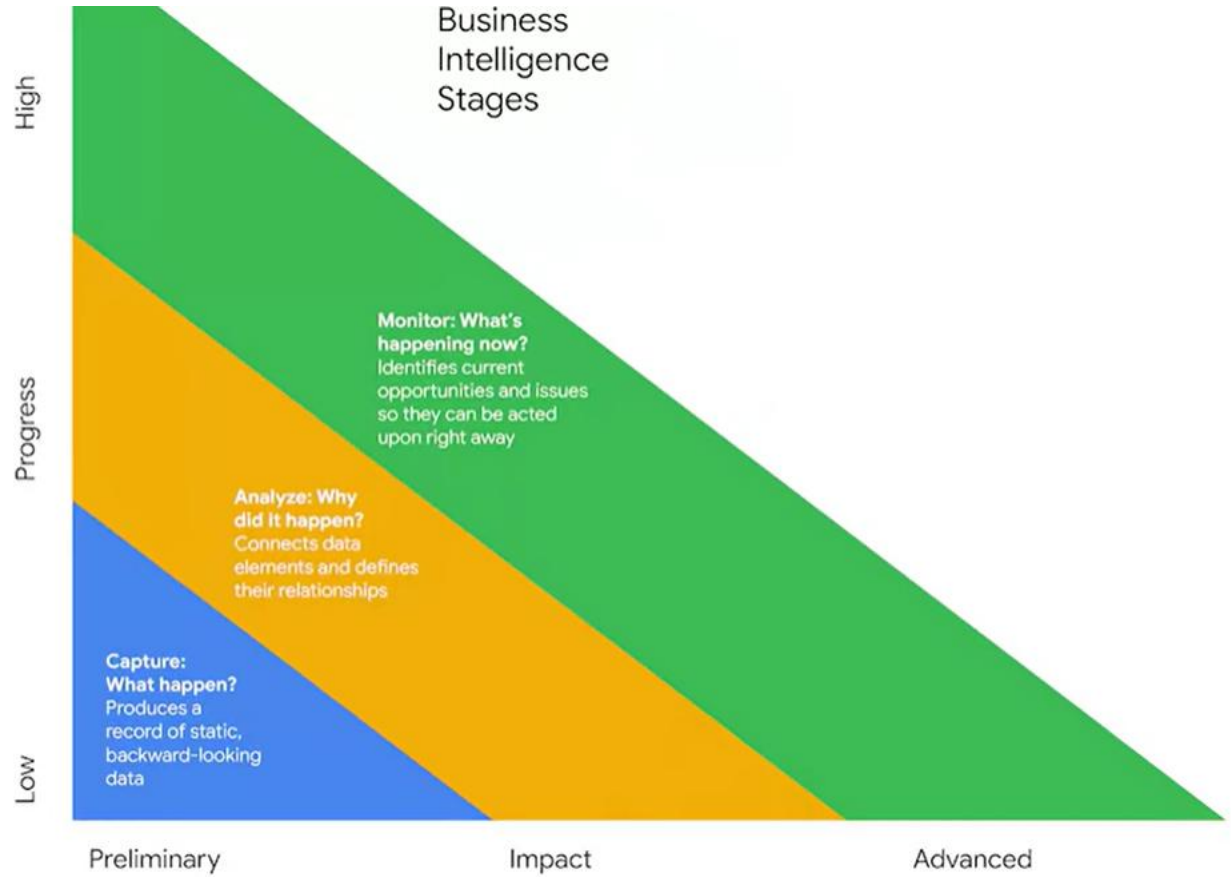
The sequence of stages that data experiences, which include plan, capture, manage, analyze, archive, and destroy

Data analysis process

The six phases of ask, prepare, process, analyze, share, and act

Data maturity

The extent to which an organization is able to use its data in order to extract actionable insights



Business intelligence governance

A process for defining and implementing business intelligence systems and frameworks within an organization

What solutions are we using, and how?

Which of them bring value?

What types of solutions do we plan to implement?

How will we deliver them?

How will we support them?

Which dashboards, reports, and other solutions will be most effective?

Do different users, teams, and departments require different technologies?

Which technologies do we have access to?

Can we access others, if needed?

How will we measure success?

Key performance indicator (KPI)

A quantifiable value, closely linked to business strategy, which is used to track progress toward a goal

Dashboard

An interactive visualization tool that monitors live, incoming data

Effective questioning

- Understanding the difference between effective and ineffective questions
- Knowing what types of questions lead to the best insights
- Using questioning to confirm that you fully understand stakeholder expectations

Deliverable

Any product, service, or outcome that must be achieved in order to complete a project

Bias

A conscious or subconscious preference in favor of or against a person, group of people, or thing

Fairness

A quality of data analysis that does not create or reinforce bias

There are a few questions you can keep in mind to help guide your communications with stakeholders and partners:

- **Who is your audience?** When communicating with stakeholders and project partners, it's important to consider who you're working with. Consider all of the people who need to understand the BI tools and processes you build when communicating. The sales or marketing team has different goals and expertise than the data science team, for example.

- **What do they already know?** Because different users have different levels of knowledge and expertise, it can be useful to consider what they already know before communicating with them. This provides a baseline for your communications and prevents you from overexplaining yourself or skipping over any information they need to know.

- **What do they need to know?** Different stakeholders need different kinds of information. For instance, a user might want to understand how to access and use the data or any dashboards you create, but they probably aren't as interested in the nitty-gritty details about how the data was cleaned.

- **How can you best communicate what they need to know?** After you have considered your audience, what they already know, and what they need to know, you need to choose the best way to communicate that information to them. This might be an email report, a small meeting, or a cross-team presentation with a Q&A section.

Create realistic deadlines. Before you start a project, make a list of dependencies and potential roadblocks so you can assess how much extra time to give yourself when you discuss project expectations and timelines with your stakeholders.

Know your project. When you have a good understanding about why you are building a new BI tool, it can help you connect your work with larger initiatives and add meaning to the project. Keep track of your discussions about the project over email or meeting notes, and be ready to answer questions about how certain aspects are important for your organization. In short, it should be easy to understand and explain the value the project is bringing to the company.

Communicate often. Your stakeholders will want regular updates. Keep track of key project milestones, setbacks, and changes. Another great resource to use is a changelog, which can provide a chronologically ordered list of modifications. Then, use your notes to create a report in a document that you share with your stakeholders.

As a BI professional, it's your responsibility to remain as objective as possible and try to recognize the many sides of an argument before drawing conclusions. The best thing you can do for the fairness and accuracy of your data is to make sure you start with data that has been collected in the most appropriate, and objective way. Then you'll have facts that you can pass on to your team.

A big part of your job will be putting data into context. Context is the condition in which something exists or happens; basically, this is who, what, where, when, how, and why of the data. When presenting data, you'll want to make sure that you're providing information that answers these questions:

- WHO collected the data?
- WHAT is it about? What does the data represent in the world and how does it relate to other data?
- WHEN was the data collected?
- WHERE did the data come from?
- HOW was it collected? And how was it transformed for the destination?
- WHY was this data collected? Why is it useful or relevant to the business task?

Metric

A single, quantifiable data point that is used to evaluate performance

Key performance indicator (KPI)

A quantifiable value, closely linked to business strategy, which is used to track progress toward a goal

Metrics support KPIs

KPIs support overall business objectives

Strategy

A plan for achieving a goal or arriving at a desired future state

Tactic

A method used to enable an accomplishment

BI monitoring

Building and using hardware and software tools to easily and rapidly analyze data and enable stakeholders to make impactful business decisions

~~"Why did our recent video go viral?"~~

"How many times was our video shared on social channels the first week it was posted?"

Action-oriented questions encourage change.

~~"How can we get customers to recycle our product packaging?"~~

"What design features will make our packaging easier to recycle?"

Relevant questions matter, are important, and have significance to the problem you're trying to solve.

~~"Why does it matter that Pine Barrens tree frogs started disappearing?"~~

"What environmental factors changed in Durham, North Carolina, between 1983 and 2004 that could cause Pine Barrens tree frogs to disappear from the Sandhills Regions?"

"What environmental factors changed in Durham, North Carolina, between 1983 and 2004 that could cause Pine Barrens tree frogs to disappear from the Sandhills Regions?"

S pecific

M easurable

A ction-oriented

R elevant

T ime-bound

Data bias

A type of error that systematically skews results in a certain direction

Entry-level data analytics professional; recently completed the Google Data Analytics Professional Certificate

Accomplished [X]
As measured by [Y]
By doing [Z]

Data availability

The degree or extent to which timely and relevant information is readily accessible and able to be put to use

Data integrity

The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

Data visibility

The degree or extent to which information can be identified, monitored, and integrated from disparate internal and external sources

Data availability factors

- Data integrity
- Data visibility
- Update frequency
- Change

Types of bias in business intelligence

- Confirmation bias
- Selection bias
- Historical bias
- Outlier bias

Tips for addressing bias

- Record prior beliefs and assumptions
- Use a randomized set of data
- Gather more data and research about the opposing hypothesis
- Be cognizant of outliers

Vanity metrics

Data points that are intended to impress others, but are not indicative of actual performance and, therefore, cannot reveal any meaningful business insights

Data manipulation

The process of changing data to make it more organized and easier to read

Experiential learning

Understanding through doing

Unstructured data

Data that is not organized in any easily identifiable manner

Structured data

Data that has been organized in a certain format such as rows and columns

Data model

A tool for organizing data elements and how they relate to one another

Design pattern

A solution that uses relevant measures and facts to create a model in support of business needs

Schema

A way of describing how something, such as data, is organized

Common schemas

- Relational models
- Star schemas
- Snowflake schemas
- NoSQL schemas

Relational database

A database that contains a series of tables that can be connected to form relationships

Primary key

An identifier in a database that references a column or group of columns in which each row uniquely identifies each record in the table

Fact table

A table that contains measurements or metrics related to a particular event

Dimension table

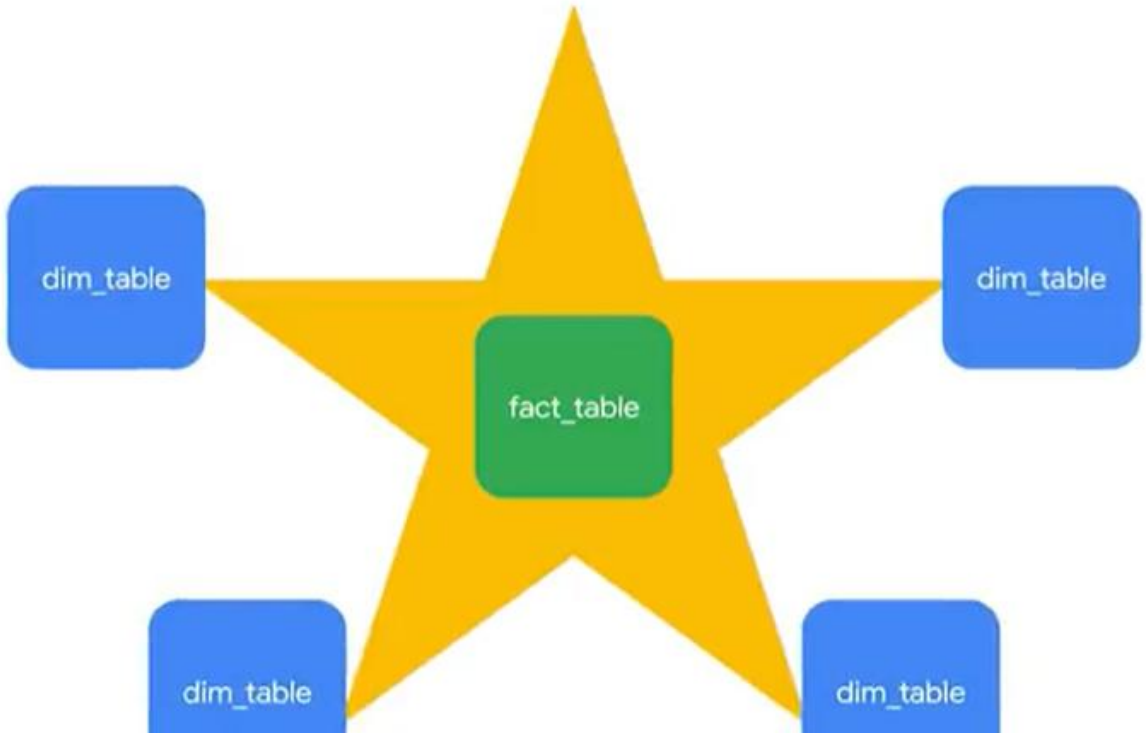
The table where the attributes of the dimensions of a fact are stored

Schema

A way of describing how something, such as data, is organized

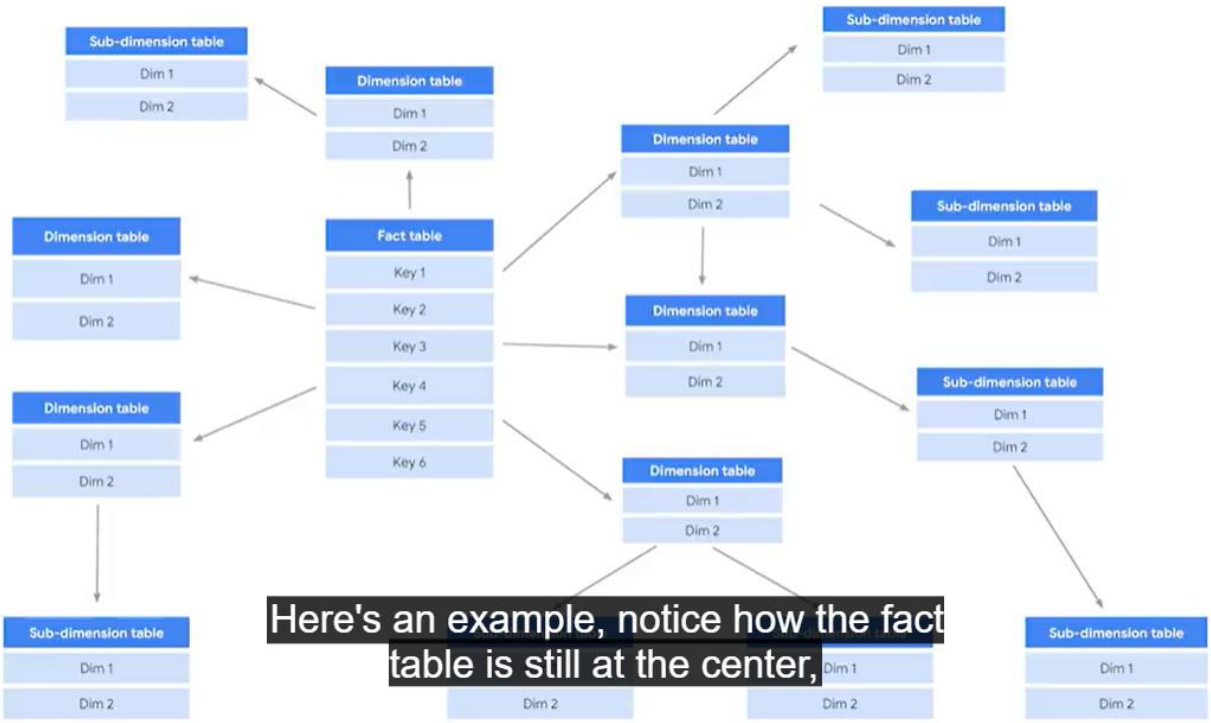
Star schema

A schema consisting of one fact table that references any number of dimension tables



Snowflake schema

An extension of a star schema with additional dimensions and, often, subdimensions



Here's an example, notice how the fact table is still at the center,

Database migration

Moving data from one source platform to another target database

Type of databases

- OLTP
- OLAP
- Row-based
- Columnar
- Distributed
- Single-homed
- Separated storage and compute
- Combined

separated storage and compute and combined databases.

Online Transaction Processing (OLTP) database

A database that has been optimized for data processing instead of analysis

OLAP (Online Analytical Processing)

A database system that has been optimized for analysis in addition to processing and can analyze data from multiple databases

Row-based database

A database that is organized by rows

Columnar database

A database organized by columns instead of rows

Single-homed database

Databases where all of the data is stored in the same physical location

Distributed database

A collection of data systems distributed across multiple physical locations

Combined systems

Database systems that store and analyze data in the same place

Separated storage and computing systems

Databases where less relevant data is stored remotely, and relevant data is stored locally for analysis

[redacted]

They are using a data lake. A data lake is a database system that stores large amounts of raw data in its original format until it's needed.

Data warehouse

A specific type of database that consolidates data from multiple source systems for data consistency, accuracy, and efficient access

Considerations for designing a data warehouse

- Business needs
- The shape and volume of the data
- What model the data warehouse will follow

Logical data modeling

Representing different tables in the physical data model

A database schema should include

1. The relevant data
2. Names and data types for each column in each table
3. Consistent formatting
4. Unique keys

Data pipeline

A series of processes that transports data from different sources to their final destination for storage and analysis

Extract, transform, and load (ETL)

A type of data pipeline that enables data to be gathered from source systems, converted into a useful format, and brought into a data warehouse or other unified destination system

Target table

The predetermined location where pipeline data is sent in order to be acted on

Extraction

- Access source systems
- Read and collect the necessary data
- Make the data useful for analysis

Transformation

- Consider the structure and format of the destination
- Consider business case requirements
- Validate, clean, and prepare the data for analysis
- Map the data types from the sources to the target systems

Loading

- Data is delivered to its target destination
- Data can exist within multiple locations and in multiple formats

Key performance indicator (KPI)

A quantifiable value, closely linked to business strategy, which is used to track progress toward a goal

Google DataFlow

A serverless data-processing service that reads data from the source, transforms it, and writes it in the destination location

Object-oriented programming language

A programming language that is modeled around data objects

Interpreted programming language

A programming language that uses an interpreter -usually another program- to read and execute coded instructions

Compiled programming language

A programming language that compiles coded instructions that are executed directly by the target machine

Discrete data

Data that is counted and has a limited number of values

Continuous data

Data that is measured and can have almost any numeric value

Nominal data

A type of qualitative data that is categorized without a set order

Internal data

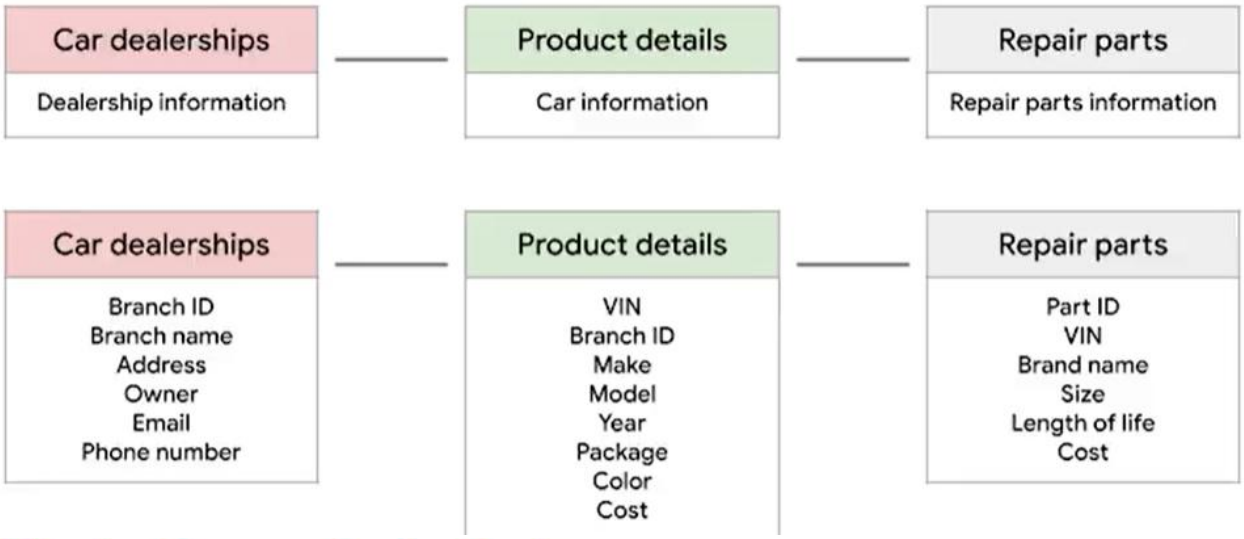
Data that lives within a company's own systems

External data

Data that lives and is generated outside of an organization

Unstructured data

Data that is not organized in any easily identifiable manner



Relational database

A database that contains a series of related tables that can be connected via their relationships

Primary key

An identifier that references a column in which each value is unique

Foreign key

A field within a table that is a primary key in another table

Primary key:

- Used to ensure data in a specific column is unique
- Uniquely identifies a record in a relational database table
- Only one primary key is allowed in a table
- Cannot contain null or blank values

Foreign key:

- A column or group of columns in a relational database table that provides a link between the data in two tables
- Refers to the field in a table that's the primary key of another table
- More than one foreign key is allowed to exist in a table

Data mart

A subject-oriented database that can be a subset of a larger data warehouse

Subject-oriented

Associated with specific areas or departments of a business

Data lake

A database system that stores large amounts of raw data in its original format until it's needed

ELT (extract, load, and transform)

A type of data pipeline that enables data to be gathered from data lakes, loaded into a unified destination system, and transformed into a useful format

Database performance

A measure of the workload that can be processed by a database, as well as associated costs

The five factors of database performance

1. Workload
2. Throughput
3. Resources
4. Optimization
5. Contention

First, we'll start with workload.

Workload

The combination of transactions, queries, data warehousing analysis, and system commands being processed by the database system at any given time

the combination of

Throughput

The overall capability of the database's hardware and software to process requests

Resources

The hardware and software tools available for use in a database system

Optimization

Maximizing the speed and efficiency with which data is retrieved in order to ensure high levels of database performance

Contention

When two or more components attempt to use a single resource in a conflicting way

Factor	Definition	Example
Workload	The combination of transactions, queries, data warehousing analysis, and system commands being processed by the database system at any given time.	On a daily basis, your database needs to process sales reports, perform revenue calculations, and respond to real-time requests from stakeholders.All of these needs represent the workload the database needs to be able to handle.
Throughput	The overall capability of the database's hardware and software to process requests.	The system's throughput is the combination of input and output speed, the CPU speed, the machine's ability to run parallel processes, the database management system, and the operating system and system software.
Resources	The hardware and software tools available for use in a database system.	The database system is primarily cloud-based, which means it depends on online resources and software to maintain functionality.
Optimization	Maximizing the speed and efficiency with which data is retrieved in order to ensure high levels of database performance.	Continually checking that the database is running optimally is part of your job as the team's BI professional.
Contention	When two or more components attempt to use a single resource in a conflicting way.	Because this system automatically generates reports and responds to user-requests, there are times when it may be trying to run the queries on the same datasets at the same time, causing slowdown for users.

Response time

The time it takes for a database to respond to a user request

Query plan

A description of the steps a database system takes in order to execute a query

Index

An organizational tag used to quickly locate data within a database system

Data partitioning

The process of dividing a database into distinct, logical parts in order to improve query processing and increase manageability

Fragmented data

Data that is broken up into many pieces that are not stored together, often as a result of using the data frequently or creating, deleting, or modifying files

Workload

The combination of transactions, queries, data warehouse analysis, and system commands being processed by the database system at any given time

Quality testing

The process of checking data for defects in order to prevent system failures

1. Completeness
2. Consistency
3. Conformity
4. Accuracy
5. Redundancy
6. Integrity
7. Timeliness

Completeness

Confirming that the data contains all desired components or measures

Consistency

Confirming that data is compatible and in agreement across all systems

Conformity

Confirming that the data fits the required destination format

Accuracy

Confirming that the data conforms to the actual entity being measured or described

Redundancy

Moving, transforming, or storing more than the necessary data

Integrity

Confirming the data is accurate, complete, consistent, and trustworthy throughout its life cycle

Data mapping

The process of matching fields from one data source to another

Timeliness

Confirming that data is current

Common issues

There are also some common issues you can protect against within your system to ensure the incoming data doesn't cause errors or other large-scale problems in your database system:

- **Check data mapping:** Does the data from the source match the data in the target database?
- **Check for inconsistencies:** Are there inconsistencies between the source system and the target system?
- **Check for inaccurate data:** Is the data correct and does it reflect the actual entity being measured?
- **Check for duplicate data:** Does this data already exist within the target system?

- Schema validation
- Data dictionaries
- Data lineages

Schema validation

A process to ensure that the source system data schema matches the target database data schema

Schema validation properties

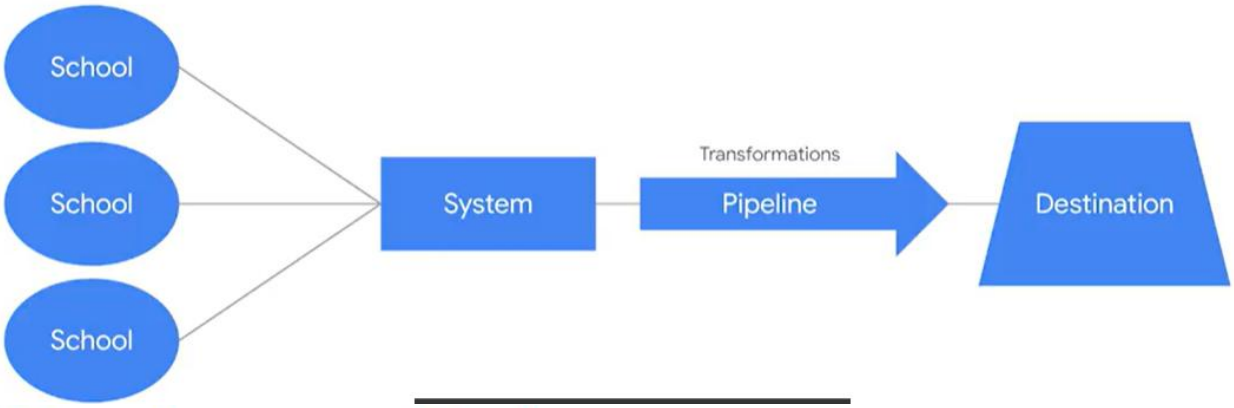
1. The keys are still valid
2. The table relationships have been preserved
3. The conventions are consistent

Data dictionary

A collection of information that describes the content, format, and structure of data objects within a database, as well as their relationships

Data lineage

The process of identifying the origin of data, where it has moved throughout the system, and how it has transformed over time



Business Rule

A statement that creates a restriction on specific parts of a database

Business rules affect

- What data is collected and stored
- How relationships are defined
- What kind of information the database provides
- The security of the data

Data integrity

The accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle

A strong analysis depends on the integrity of the data

Data replication

The process of storing data in multiple locations

Data transfer

The process of copying data from a storage device to memory, or from one computer to another

Data manipulation

The process of changing data to make it more organized and easier to read

Other threats to data integrity

- Human error
- Viruses
- Malware
- Hacking
- System failures

Metadata

Data about data

Metadata is used in database management to help data analysts interpret the contents of the data within the database

3 common types of metadata

- Descriptive
- Structural
- Administrative

Descriptive metadata

Metadata that describes a piece of data and can be used to identify it at a later point in time

Structural metadata

Metadata that indicates how a piece of data is organized and whether it is part of one, or more than one, data collection

Administrative metadata

Metadata that indicates the technical source of a digital asset

Experiential learning

Understanding through doing

Interview preparation

- What is the role?
- What are the typical questions and types of interviews this company does?
- What is the team or the organization for this role?
- What are the projects that help this team to grow and excel?

Types of Dashboards

Often, BI professionals will tailor a dashboard for a specific purpose. The three most common categories are:

- **Strategic**: focuses on long-term goals and strategies at the highest level of metrics
- **Operational:** tracks short-term performance and intermediate goals
- **Analytic:** consists of the datasets and the mathematics used in these sets

Low-fidelity mockup

A simple draft of a visualization that is used for planning a dashboard and evaluating its progress

Common data problems

- More data than you need
- Not enough data
- The wrong datasets
- Data from the wrong time period
- Lacking understanding of data availability

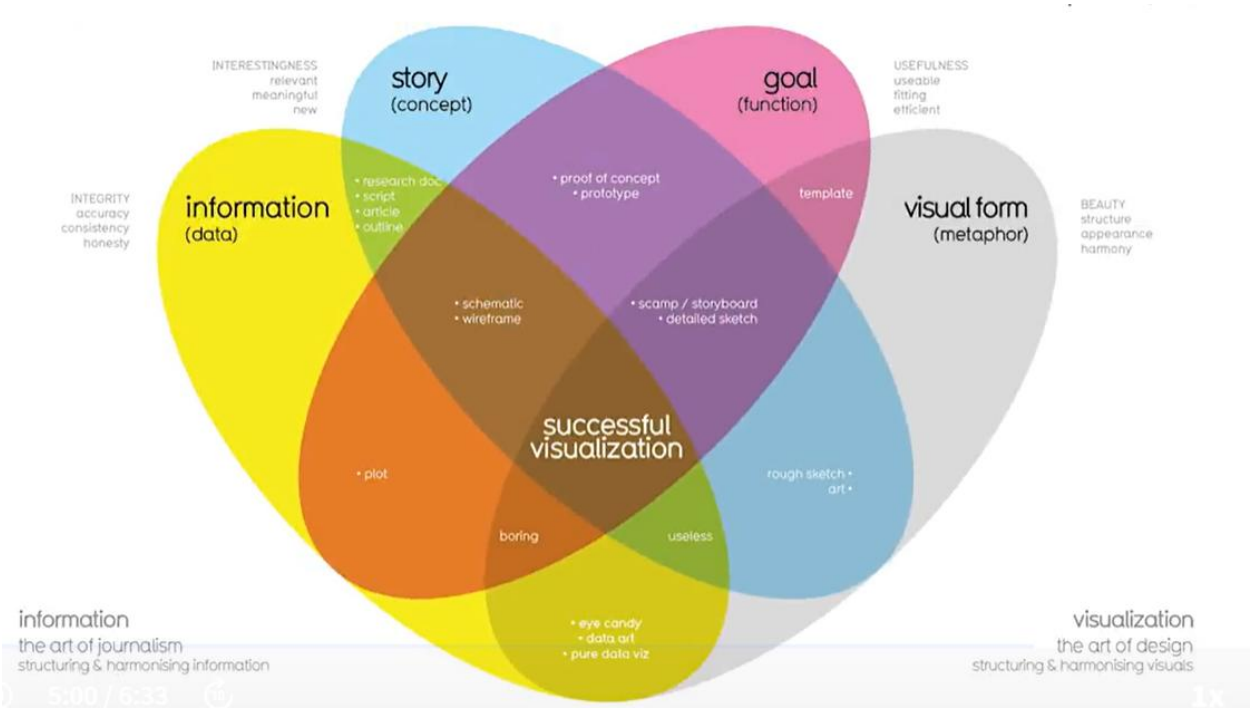
- What data should be displayed?
- Where should the data come from?
- How will you access the data?
- How will you import the data into the dashboard?
- Do you have permission and access to the data?

Tool problem

A dashboard issue involving the hardware or software being used

Audience problem

A dashboard issue caused by failing to adequately consider the needs of the user



Pre-aggregation

The process of performing calculations on data while it is still in the database

Trade-off:

More steps in the pipeline; faster performance for users

Trade-off:

Fewer steps in the pipeline; slower performance for users

Trade-off

Pre-aggregated data is faster to work with; less flexible in certain situations

Dimension

A qualitative data type that can be used to categorize data

Measure

A quantitative data type that can be either discrete or continuous
Discrete data has a limited number of values

Continuous data can have almost any numeric value

Encoding

The process of translating dimensions and measures into visual representations of the data

Accessibility alert:

Refer to color accessibility guidelines or use double-encoding

- Aesthetics
- Design
- Privacy
- Processing speed

Processing speed

How quickly a program can update and load a specified amount of data

Contributors to processing speed

- Volume of data
- Number of measures
- Number of dimensions

To reduce load and increase speed

- Start broadly, then narrow scope
- Remove irrelevant metrics
- Change the calculations
- Configure the amount of preloaded data
- Filter data early

Privacy permissions

- Public availability
- Object-level permission
- Row-level permission

Public availability

A privacy setting that allows anyone to access a dashboard

Object-level permission

A privacy setting that controls the availability of a single item in a dashboard

Row-level permission

A privacy setting that controls the availability of specific rows of a table or dataset in a dashboard

Aggregation

Collecting or gathering many separate pieces into a whole

Data aggregation

The process of gathering data from multiple sources in order to combine it into a single summarized collection

- Puzzle pieces = data
- Organization = aggregation
- Pile of pieces = summary
- Putting the pieces together = gaining insights

Data can also be aggregated over a given time period to provide statistics such as:

- Averages
- Minimums
- Maximums
- Sums

Subquery

A query within another query

Nominal data

A type of qualitative data that is categorized without a set order

Ordinal data

A type of qualitative data with a set order or scale

Internal data

Data that lives within a company's own systems

External data

Data that lives and is generated outside of an organization

Structured data

Data organized in a certain format such as rows and columns

Unstructured data

Data that is not organized in any easily identifiable manner

3 data storytelling steps

1. Engage your audience
2. Create compelling visuals
3. Tell the story in an interesting narrative

Dashboard

A tool that organizes information from multiple datasets into one central location for tracking, analysis, and simple visualization through tables, charts and graphs

Materials

- Dataset
- BI planning documents
- Business requirement

Business intelligence presentation

A communication with stakeholders about their needs or project status

BI presentations can be

- Emails
- Phone calls
- Meetings
- A mixture