# Data Engineering

204426

# Data Integration

# Data Integration

**Data Integration** is a process where data from many sources goes to a single centralized location, which is often a data warehouse.

# Data Integration

**Example:**

See Food, Inc. (SFI). SFI's product is a mobile app where users can take pictures of different items and identify whether the item in the picture is, or is not, a hot dog. SFI uses a lot of tools to run its business:

- Facebook Ads and Google Ads in order to acquire new users
- Google Analytics to track events on its website and in its mobile app
- MySQL database to store user information and image metadata (e.g. hot dog or not hot dog)
- Marketo to send marketing email and nurture leads
- Zendesk to perform customer support
- Netsuite for accounting and financial tracking

Each of those applications has a silo of information about SFI's operations. For SFI to get a 360-degree view of the business, all of that data needs to be combined in one place. That process is data integration.

# Data Integration

**Example:**

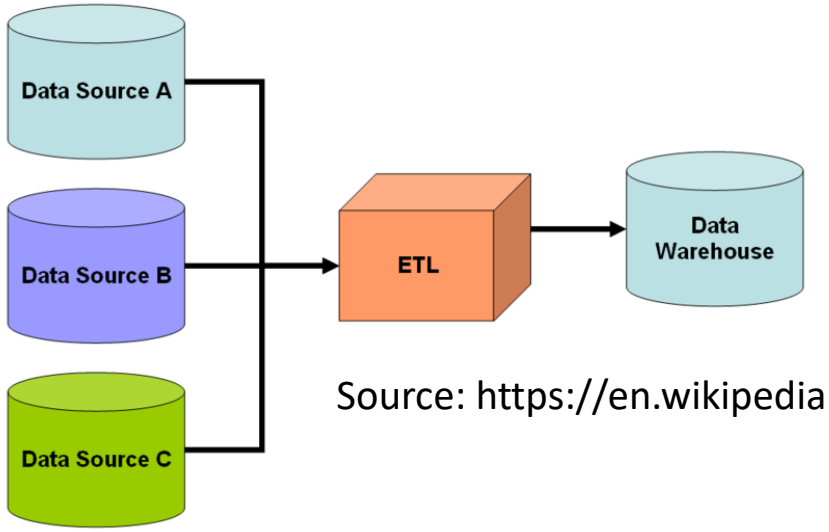
Suppose SFI is considering increasing its advertising budget, but it's not sure if it should spend more on Facebook or Google. It could ask whether the cost of acquisition is lower on Facebook or Google, but that misses out on whether there are differences between the kinds of users they acquire on the two different channels. Some additional questions the company might want to ask are:

- Do users from Facebook post more photos of hot dogs?
- Do users from Google file more customer support tickets?
- Which users are more likely to refer friends?

Each of these questions can be combined and further segmented to individual campaigns and variations on ad creative. These questions can only be answered when the data is integrated.

# Data Integration

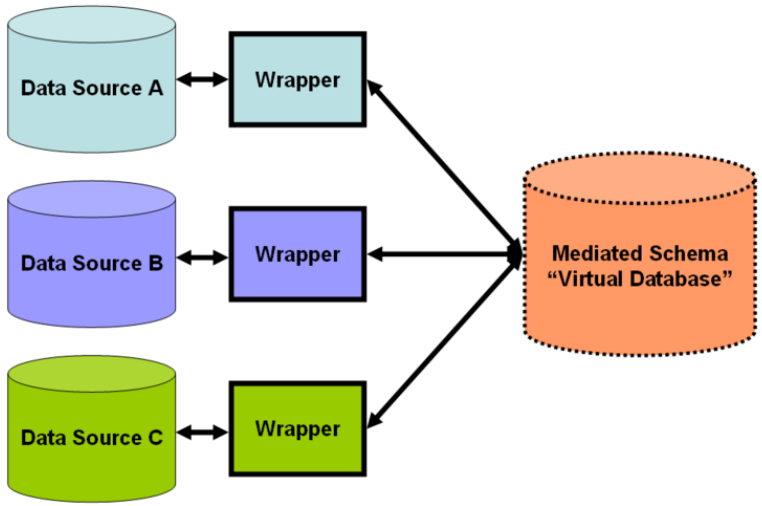
- The first data integration system: a data warehousing approach.
  - The data warehouse approach is less feasible for data sets that are frequently updated, requiring the extract, transform, load (ETL) process to be continuously re-executed for synchronization.
  - Difficulties arise in Constructing data warehouses when one has only a query interface to summary data sources and no access to the full data.




Source: https://en.wikipedia.org/wiki/Data_integration

# Data Integration

- **Mediated Schema**: Service-oriented architecture (SOA) approach.
  - providing a unified query-interface to access real time data.
  - allows information to be retrieved directly from original databases.




- Service-oriented architecture (SOA) approach.
  - Mappings between the mediated schema and the schema of original sources.
    - Mapping from entities in the mediated schema to entities in the original sources (the "Global-as-View" (GAV) approach)
    - Mapping from entities in the original sources to the mediated schema (the "Local-as-View"(LAV) approach)
  - Translating a query into decomposed queries to match the schema of the original databases.

Source: https://en.wikipedia.org/wiki/Data_integration

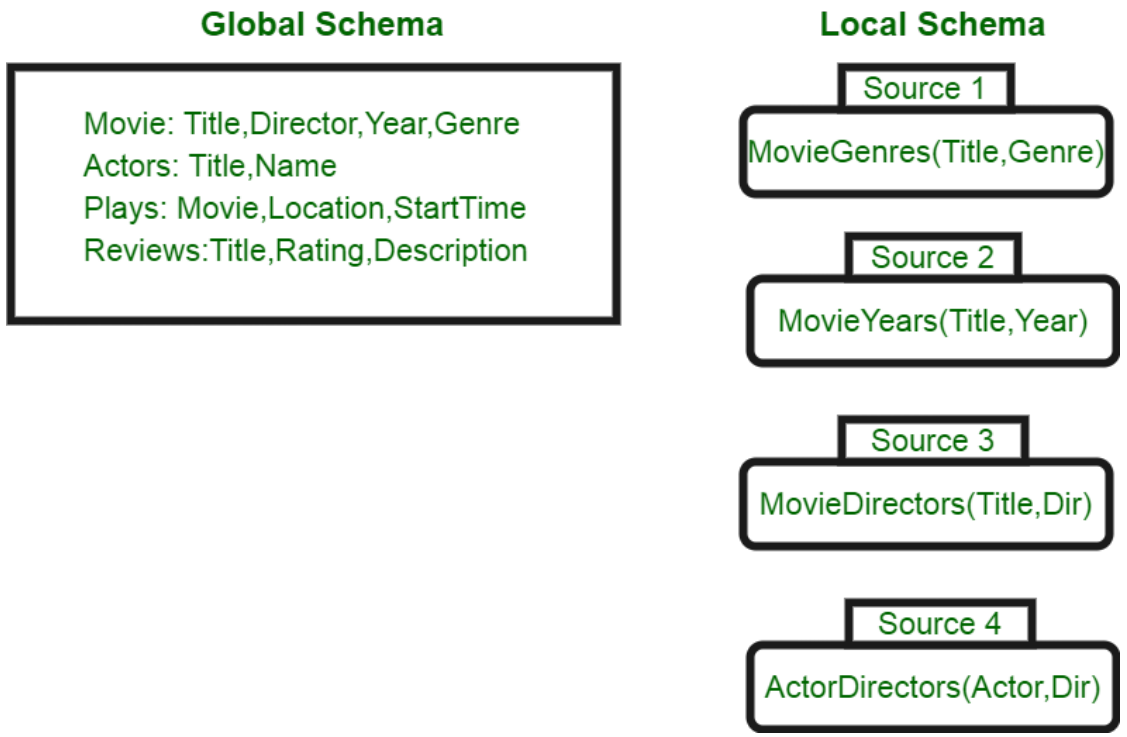
# Data Integration

- Local as View (LAV)
  - It describes each local schema as function over global schema.
  - Individual relations of local schemas of the data sources are expressed as views of the common global schema.
  - LAV can be thought of as source owners' view of system by describing which data of global database are present in source.
  - Local schema i.e. sources are described in terms of global schema using various expressions.
  - Source provides expressions to generate information from pieces of global schema.
  - With the help of Mediator, these expressions are collated to find all probable ways to answer query being fired.

Source: https://www.geeksforgeeks.org/what-is-gav-global-as-view/

# Data Integration

- Local as View (LAV) (Cont.)




Source: https://www.geeksforgeeks.org/local-as-view-lav/

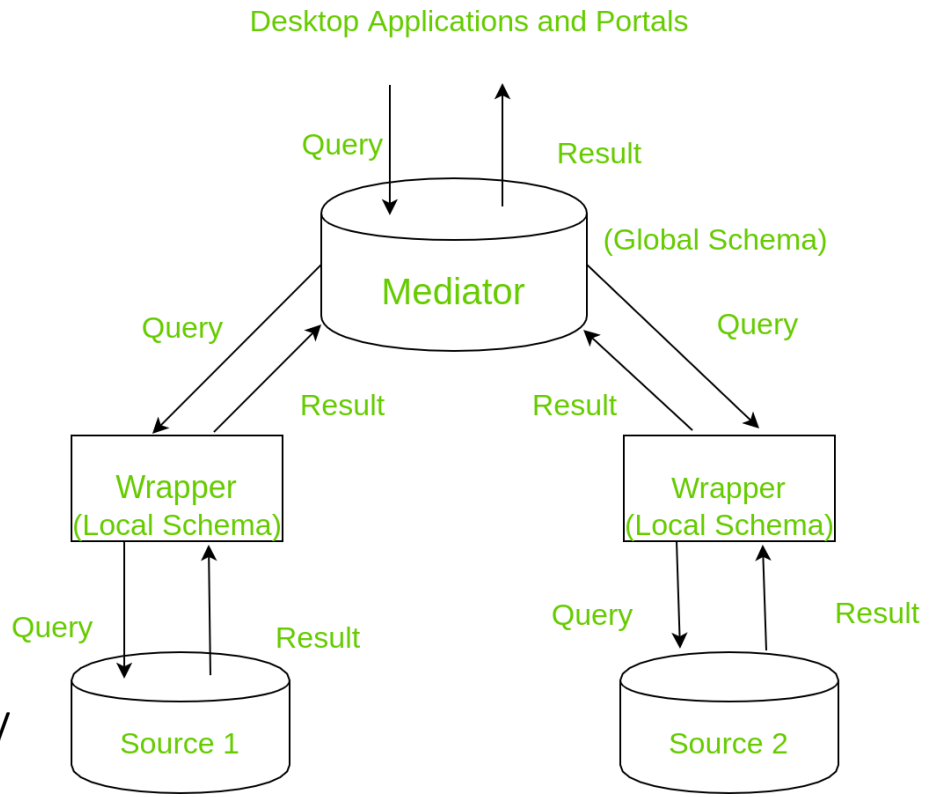
# Data Integration

- Global-as-View (GAV)
  - Global schema act as a view over source schema i.e the mediator schema is described in terms of local schema.
  - Given a query over the global schema, the mediator will follow the existing rules and templates to convert the query into source specific queries.
  - It sends the new queries to wrappers for execution.
  - Wrapper searches for all the possible expressions and how they can be combined to answer the given query.
  - Mediation involves a mediator which is a virtual view of the data and it doesn’t store any data as the data is stored at sources.

# Data Integration

- Global-as-View (GAV)  (Cont.)
  - Schema from various sources is combined forming a virtual schema of mediator.
  - When a user queries, it is mapped to multiple other queries and each query is sent to the sources.
  - Sources evaluate them and send back the results.
  - Results are merged together and sent to the end user.
  - This process is called mediation.
  - It uses wrappers which are responsible for performing the mapping of the queries.
  - They use templates (which are already created) who represent many queries and thus are made flexible.
  - If the mediator query matches a template then the results are returned, else not.

Source: https://www.geeksforgeeks.org/what-is-gav-global-as-view/



# Data Integration

- **GAV vs. LAV**

  **Global schema**:

  Movie(Title, Year, Director)
  European(Director)
  Review(Title, Critique)

  **Source 1**: $r_1(Title, Year, Director)$ since 1960, european directors
  **Source 2**: $r_2(Title, Critique)$ since 1990

  **Query**: Title and critique of movies in 1998

# Data Integration

- **GAV vs. LAV**

  **Global schema**:

  Movie(Title, Year, Director)
  European(Director)
  Review(Title, Critique)

  **LAV**: associated to source relations we have views over the global schema

  $$r_1(T, Y, D) \subseteq \{(T, Y, D) | \text{Movie}(T, Y, D) \wedge \text{European}(D) \wedge Y \geq 1960\}$$
  $$r_2(T, R) \subseteq \{(T, R) | \text{Movie}(T, Y, D) \wedge \text{Review}(T, R) \wedge Y \geq 1990\}$$

  **Query**: Title and critique of movies in 1998

  The query $\{(T, R) | \text{Movie}(T, 1998, D) \wedge \text{Review}(T, R)\}$ is processed by means of an inference mechanism that aims at re-expressing the atoms of the global schema in terms of atoms at the sources. In this case:

  $$\{(T, R) | r_2(T, R) \wedge r_1(T, 1998, D)\}$$

# Data Integration

- **GAV vs. LAV**

  **Global schema**:

  Movie(Title, Year, Director)
  European(Director)
  Review(Title, Critique)

  **GAV:** associated to relations in the global schema we have views over the sources

  $\text{Movie}(T, Y, D) \supseteq \{(T, Y, D) | r_1(T, Y, D)\}$
  $\text{European}(D) \supseteq \{(D) | r_1(D)\}$
  $\text{Review}(T, R) \supseteq \{(T, R) | r_2(T, R)\}$

# Data Integration

- **GAV vs. LAV**

  The query $\{(T, R) | \text{Movie}(T, 1998, D) \wedge \text{Review}(T, R)\}$ is processed by means of unfolding, i.e., by expanding the atoms according to their definitions, so as to come up with source relations. In this case:

$$\text{Movie}(T, 1998, D) \wedge \text{Review}(T, R)$$

$$\downarrow \qquad\qquad \downarrow$$

$$r_1(T, 1998, D) \qquad r_2(T, R)$$

# Data Integration

- **Data hub:** a collection of data from multiple sources organized for distribution, sharing, and often subsetting and sharing.
  - A data hub is designed on first principles not only to store data but also to unify and deliver data.
  - Unifying data means that the same data can be accessed by multiple applications at the same time with full data integrity.
  - Delivering data means each application has the full performance of data access that it requires at the speed of today's business.

# Data Integration

- **Data lake:** a system or repository of data stored in its natural/raw format, usually object blobs or files.
- A data lake is usually a single store of data including raw copies of source system data, sensor data, social data etc.
- A data lake can include
  - structured data from relational databases
  - semi-structured data
  - unstructured data
- A data lake can be established
  - "on premises" (within an organization's data centers)
  - "in the cloud" (using cloud services from vendors such as Amazon, Microsoft, or Google).

# Data Integration

| | DWH | Data Lake | Data Hub |
|---|---|---|---|
| Data storage | Yes | Yes | Yes |
| Indexing | Yes | No | Yes |
| Data latency | Bigger | Smaller | Smaller |
| All types of data | No | Yes | Yes |
| Inherent analytics | No | Yes | Yes |
| Optimized for machine learning | No | Yes | Yes |

Source: https://towardsdatascience.com/what-is-a-data-hub-41d2ac34c270

# References


- Maurizio Lenzerini (2002). Data integration: A theoretical perspective. http://www.diag.uniroma1.it//~lenzerin/homepagine/talks/TutorialPODS02.pdf
- What is a Data Hub? (2021). https://towardsdatascience.com/what-is-a-data-hub-41d2ac34c270
- Wikipedia (2019). Data integration. https://en.wikipedia.org/wiki/Data_integration