INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Information Systems Architecture

Statistics Portugal
Department of Methodology and Information System
Information Infrastructure Service

Pedro Cunha

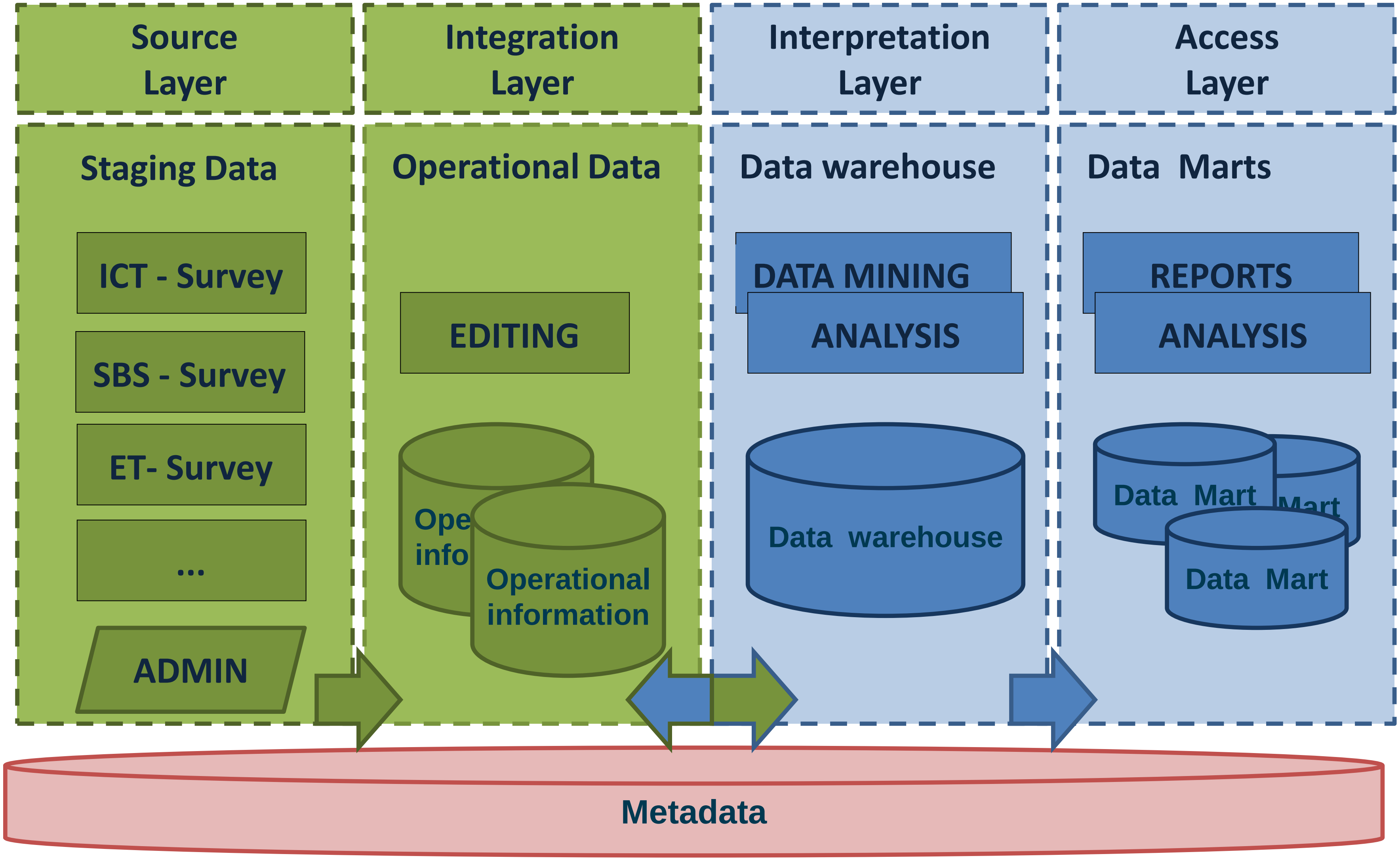
Sep 2013 • Final workshop of the ESSnet DWH• Amsterdam

Overview

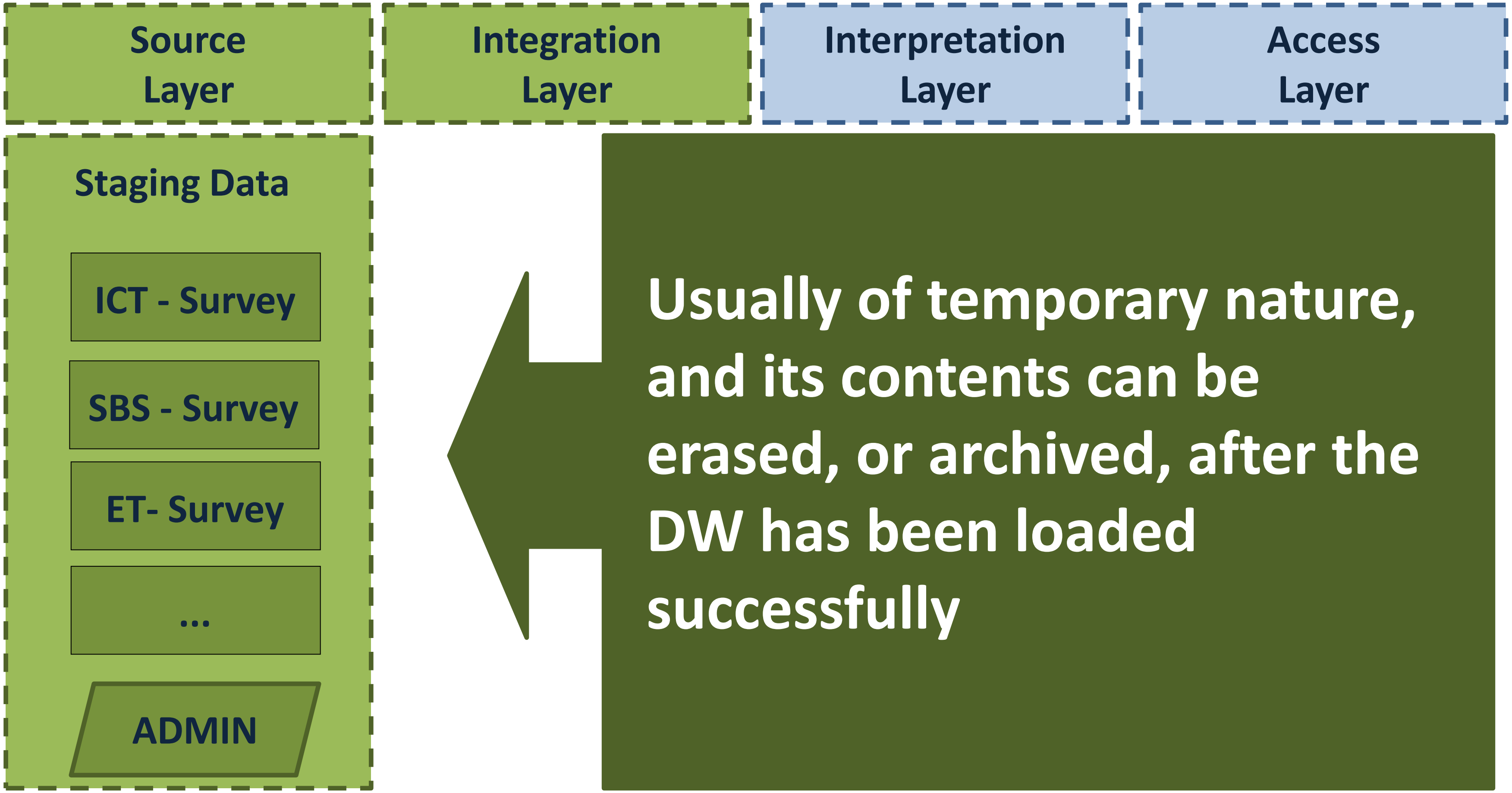
Information Systems Architecture

- **Layered approach**
- **Data model**
- **Relations with Metadata**
- **Case of use – External trade**

Information Systems Architecture

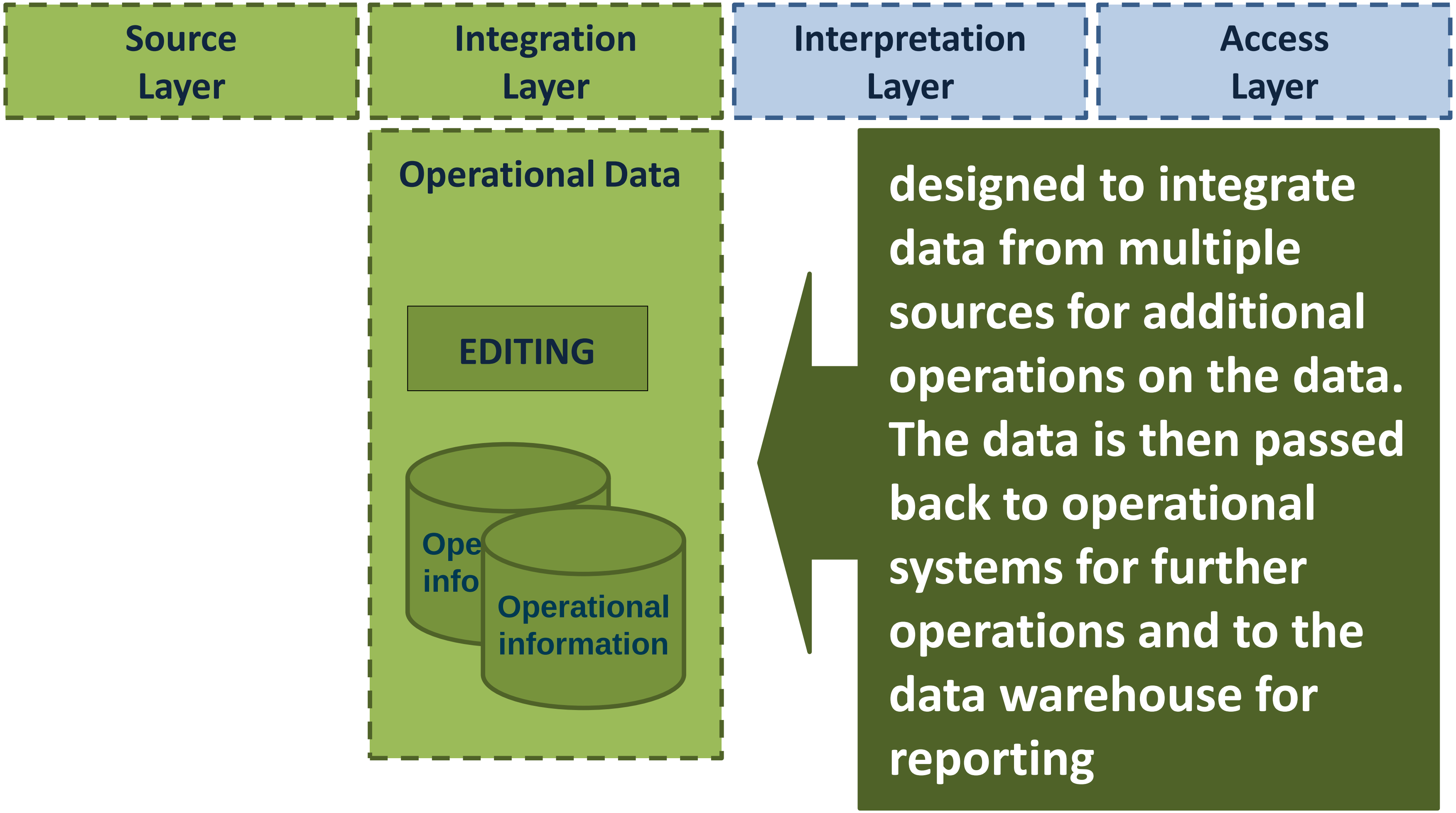


Information Systems Architecture



Usually of temporary nature, and its contents can be erased, or archived, after the DW has been loaded successfully

Information Systems Architecture



Source Layer

Integration Layer

Interpretation Layer

Access Layer

Operational Data

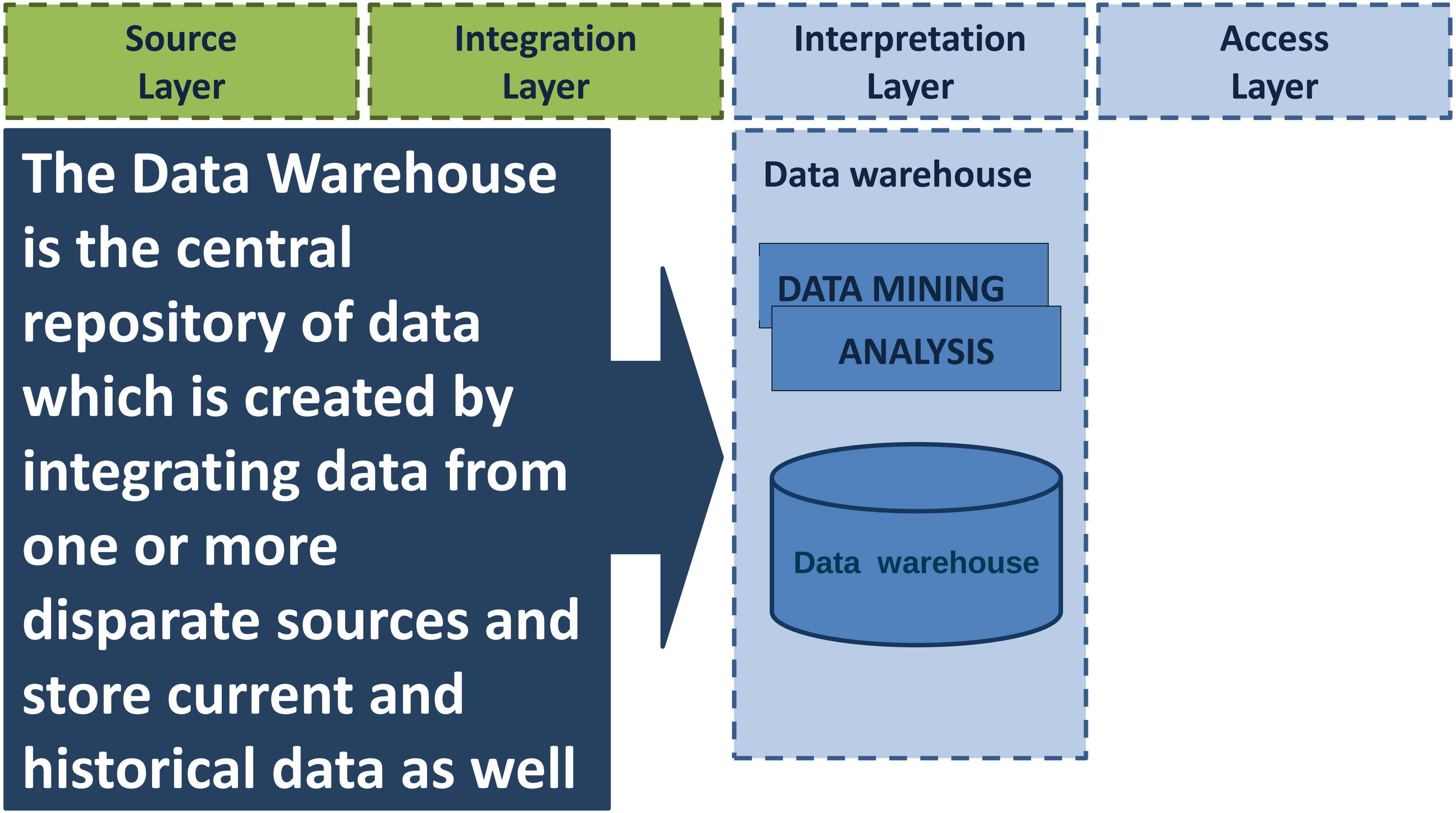
EDITING

Ope
info

Operational information

designed to integrate data from multiple sources for additional operations on the data. The data is then passed back to operational systems for further operations and to the data warehouse for reporting

Information Systems Architecture



Source Layer

Integration Layer

Interpretation Layer

Access Layer

The Data Warehouse is the central repository of data which is created by integrating data from one or more disparate sources and store current and historical data as well

Data warehouse

DATA MINING

ANALYSIS

Data warehouse

Information Systems Architecture

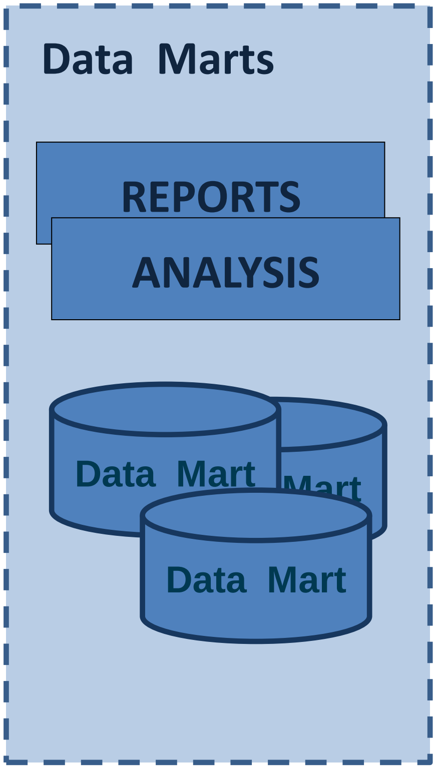
Source Layer

Integration Layer

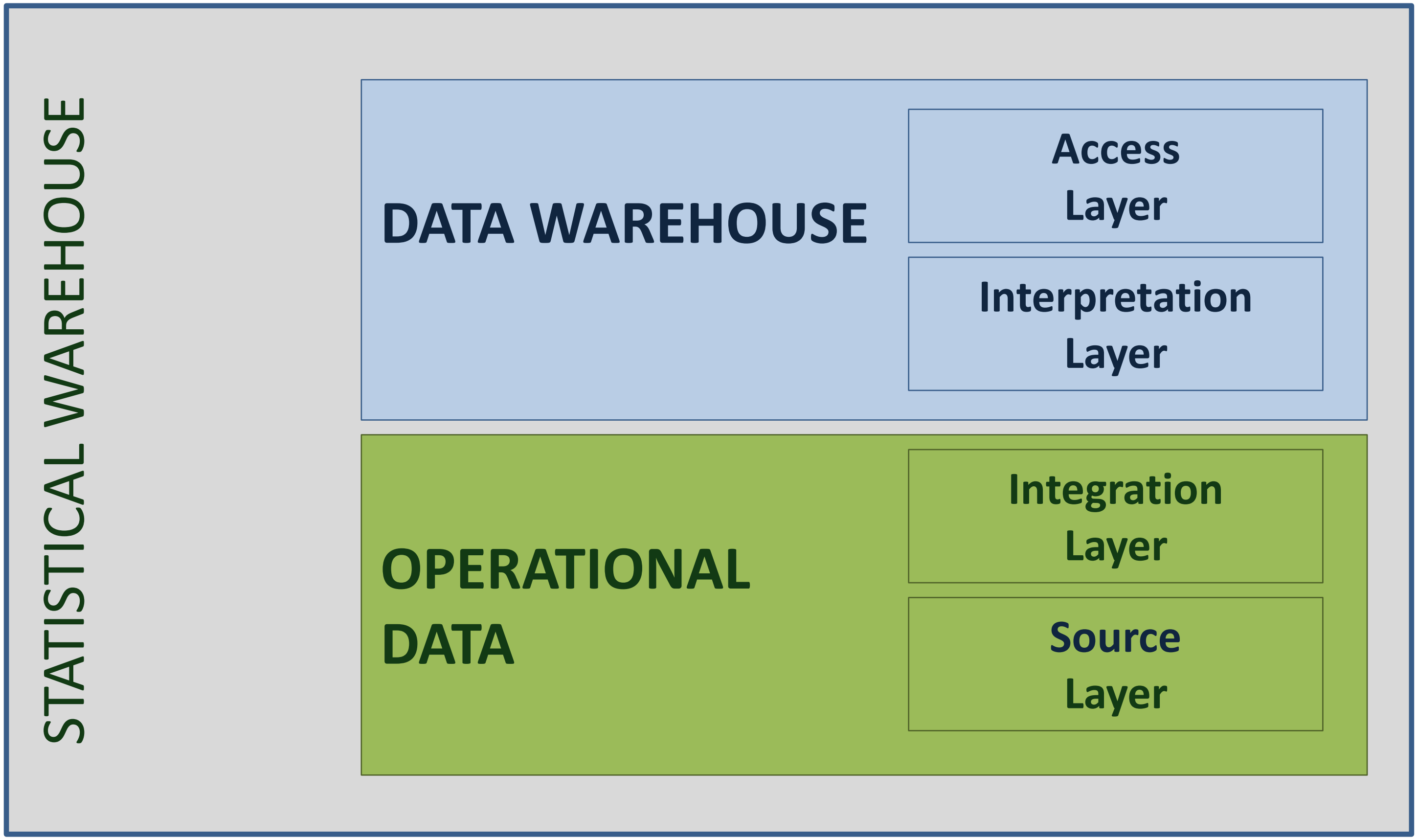
Interpretation Layer

Access Layer

Data marts are used to get data out to the users. Data marts are derived from the primary information of a data warehouse, and are usually oriented to specific business lines.



Layered architecture



STATISTICAL WAREHOUSE

DATA WAREHOUSE

Access Layer

Interpretation Layer

OPERATIONAL DATA

Integration Layer

Source Layer

data are accessible for data analysis

Used for acquiring, storing, editing and validating data

Layered architecture

These reflect two different IT environments:

- **An operational where we support semi-automatic computer interaction systems and**
- **An analytical, the warehouse, where we maximize human free interaction.**

Source Layer

The Source layer is the gathering point for all data that is going to be stored in the Data warehouse.

Source Layer

Input of source layer:

Internal sources - mainly data from surveys carried out by the NSI, but it can also be data from maintenance programs used for manipulating data in the Data warehouse

External sources - administrative data which is data collected by someone else, originally for some other purpose.

Source Layer – Data Model

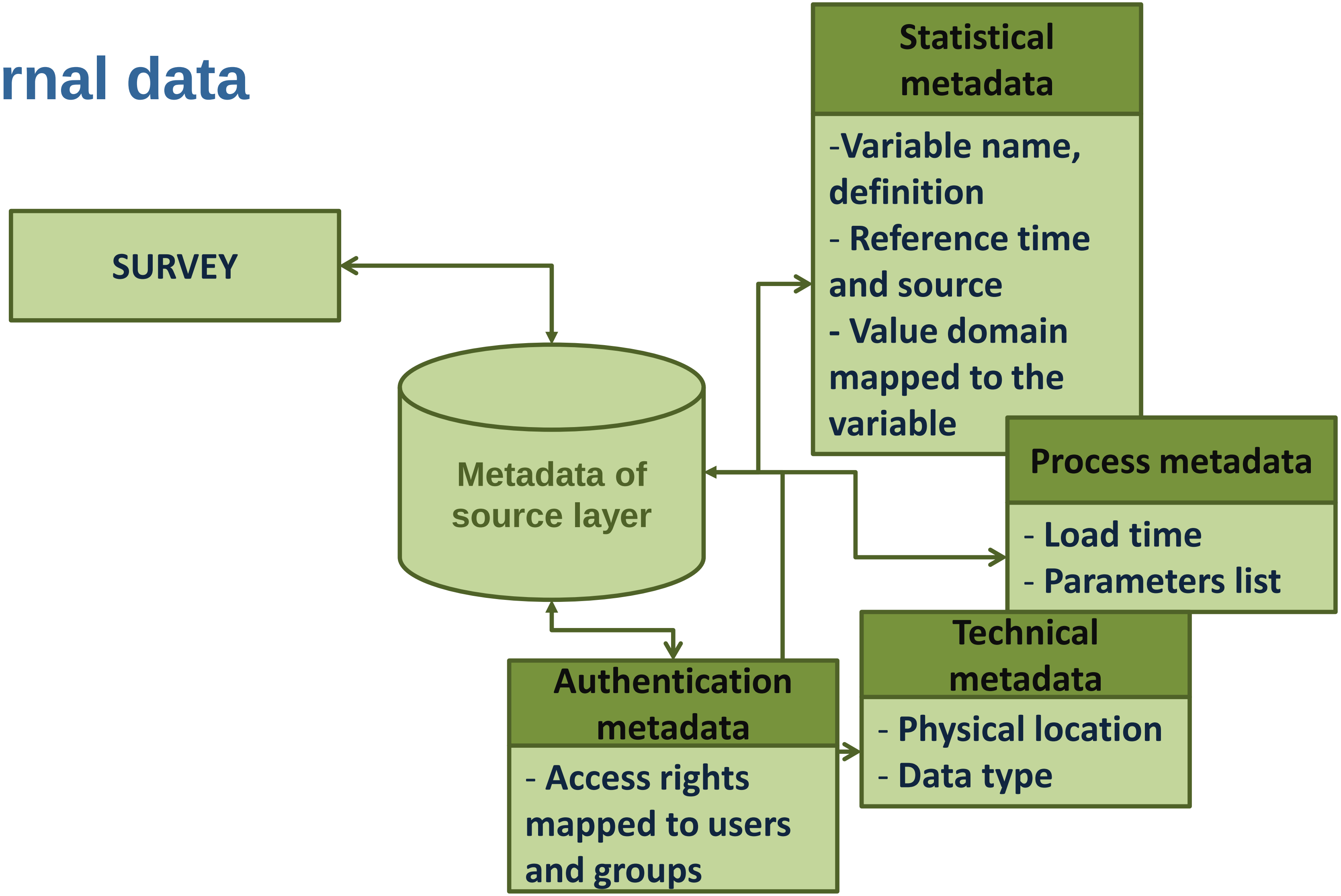
There is no pre-defined data model in source layer. Data model depends on how data is collected and on the design of each NSI data collection process. Could be a well structured data model or just simple flat files.

Source Layer and Metadata

The source layer, being the entry point, has the important role of gatekeeper, making sure that data entered into the SDWH and forwarded to the integration layer always have matching metadata of at least the agreed minimum extent and quality.

Source Layer and Metadata

Internal data



SURVEY

Metadata of source layer

Statistical metadata

-Variable name, definition
- Reference time and source
- Value domain mapped to the variable

Process metadata

- Load time
- Parameters list

Technical metadata

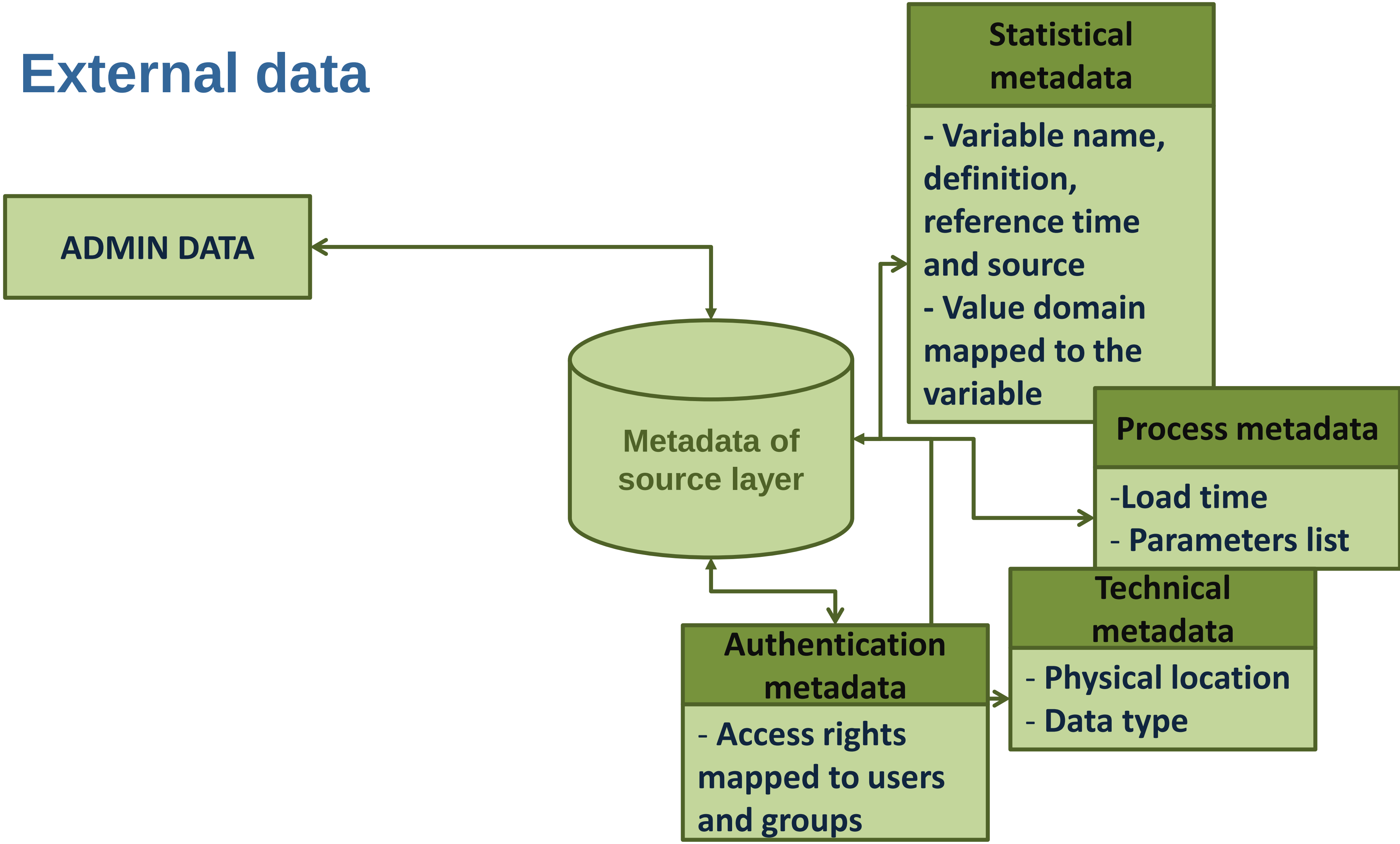
- Physical location
- Data type

Authentication metadata

- Access rights mapped to users and groups

Source Layer and Metadata

External data



ADMIN DATA

Metadata of source layer

Statistical metadata

- Variable name, definition, reference time and source
- Value domain mapped to the variable

Process metadata

-Load time
- Parameters list

Technical metadata

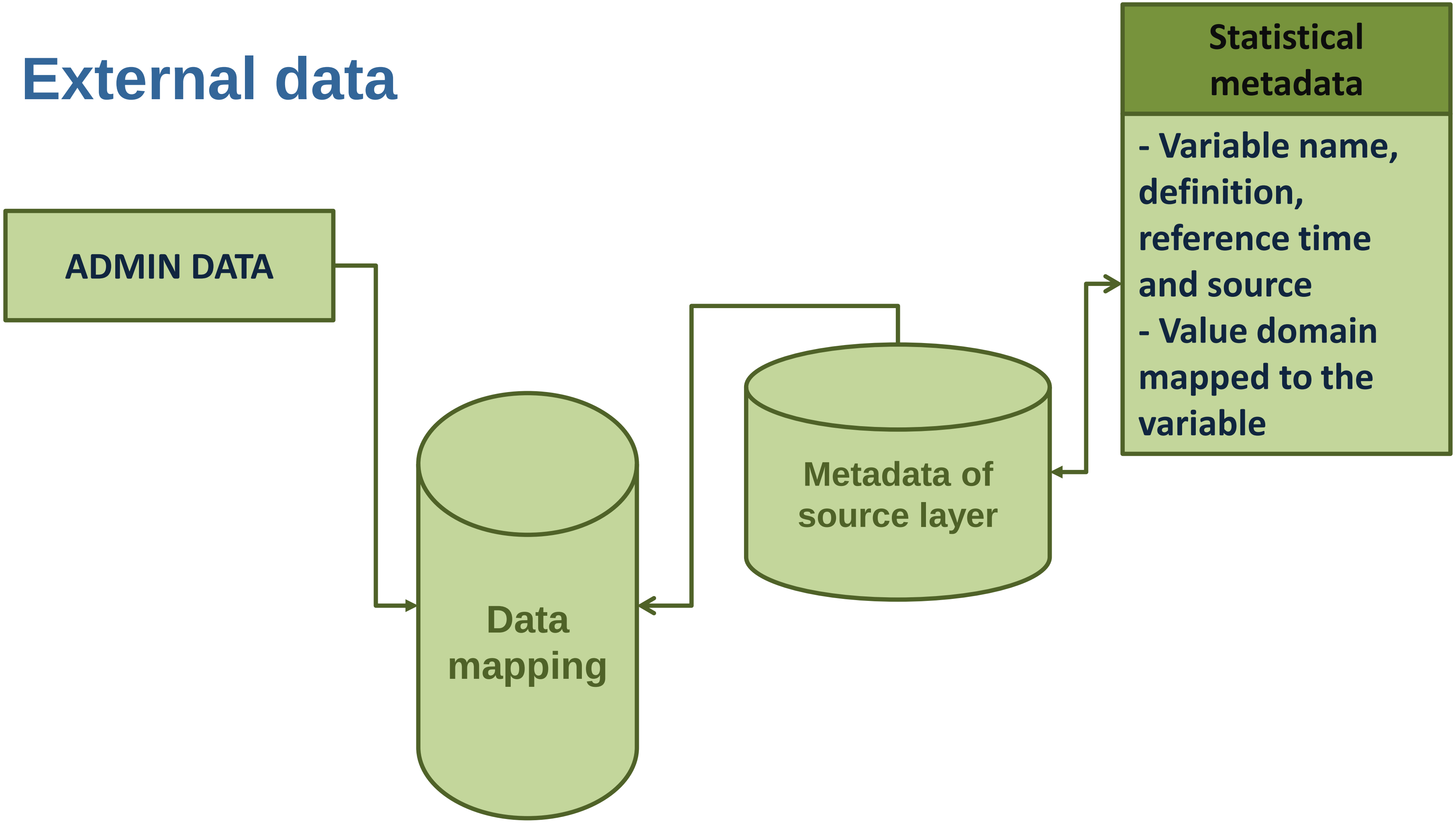
- Physical location
- Data type

Authentication metadata

- Access rights mapped to users and groups

Source Layer and Metadata

External data



ADMIN DATA

Data mapping

Metadata of source layer

Statistical metadata

- Variable name, definition, reference time and source
- Value domain mapped to the variable

Source Layer and Data Mapping

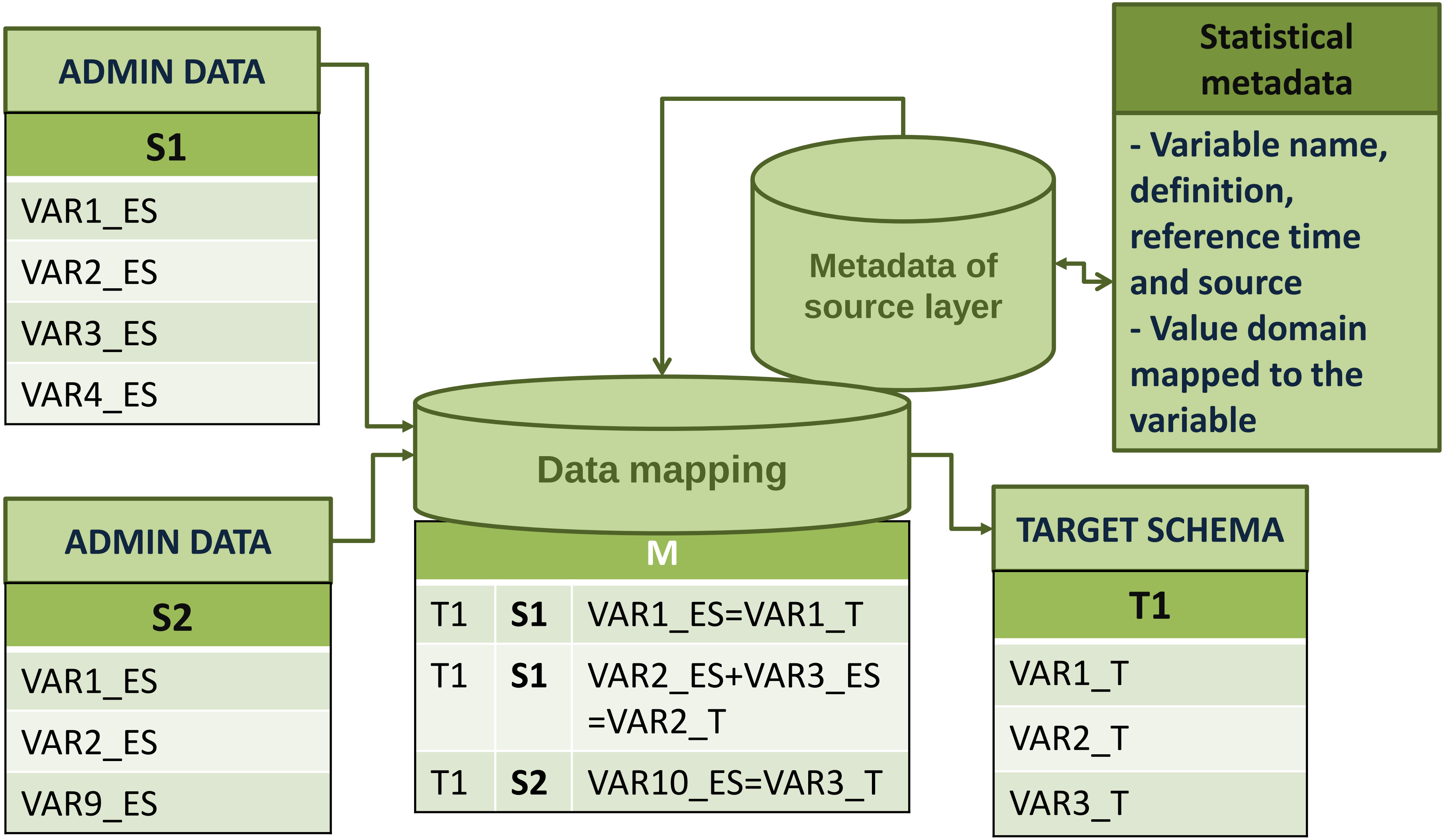
Involves combining data residing in different sources and providing users with a unified view of these data. These system are formally defined as triple <T,S,M> where:

T is the target schema,

S is source schema

M is the mapping that maps queries between source and the target schema.

Source Layer and Data Mapping



ADMIN DATA

S1
VAR1_ES
VAR2_ES
VAR3_ES
VAR4_ES

ADMIN DATA

S2
VAR1_ES
VAR2_ES
VAR9_ES

M		
T1	**S1**	VAR1_ES=VAR1_T
T1	**S1**	VAR2_ES+VAR3_ES =VAR2_T
T1	**S2**	VAR10_ES=VAR3_T

TARGET SCHEMA

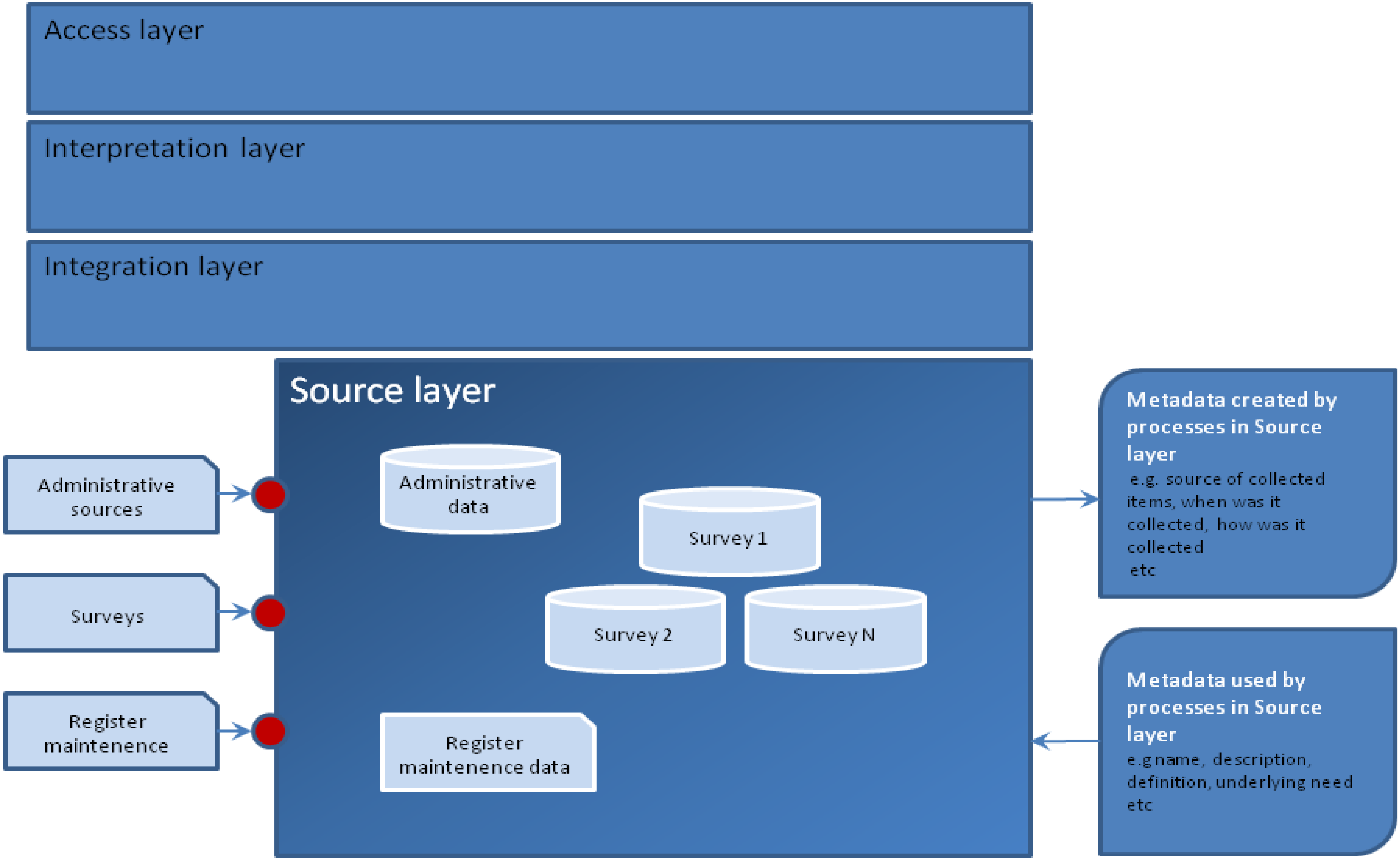
T1
VAR1_T
VAR2_T
VAR3_T

Statistical metadata

- Variable name, definition, reference time and source

- Value domain mapped to the variable

Source Layer



Metadata created by processes in Source layer
e.g. source of collected items, when was it collected, how was it collected
etc

Metadata used by processes in Source layer
e.g name, description, definition, underlying need
etc

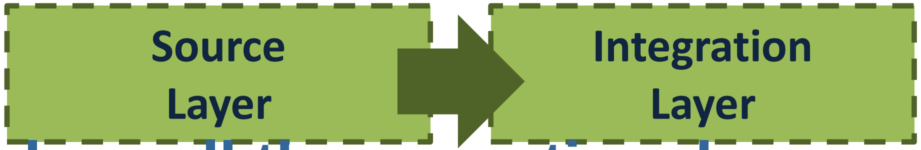
Integration Layer

Represents an operational system used to process the day-to-day transactions of an organization.

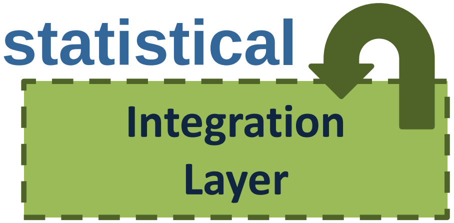
The process of translating data from source systems and transform it into useful content in the data warehouse is commonly called ETL (Extraction, Transformation, Load).

Integration Layer

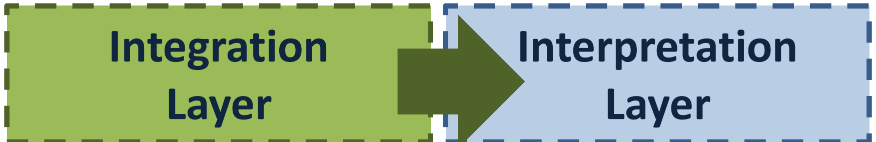
In the Extract step, data is moved from the Source layer and made accessible in the Integration layer for further processing.



The Transformation step involves all the operational activities usually associated with the typical statistical production process.



As soon as a variable is processed in the Integration layer in a way that makes it useful in the context of data warehouse it has to be Loaded into the Interpretation layer and the Access layer.

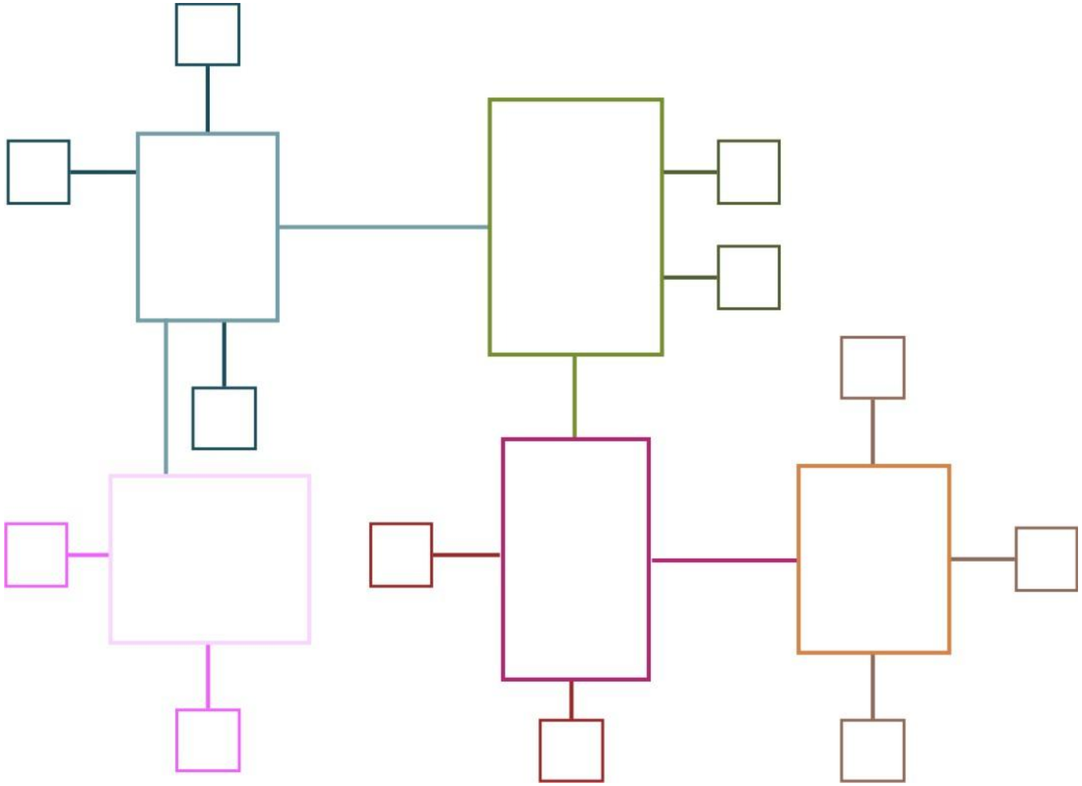


Integration Layer – Data Model

Since the focus for the Integration layer is on processing rather than search and analysis, data in the Integration layer should be stored in generalized normalized structure, optimized for OLTP (OnLine Transaction Processing).

Integration Layer – OLTP

OLTP refers to a class of applications that facilitate transaction for data editing, in which systems responds immediately to user requests.



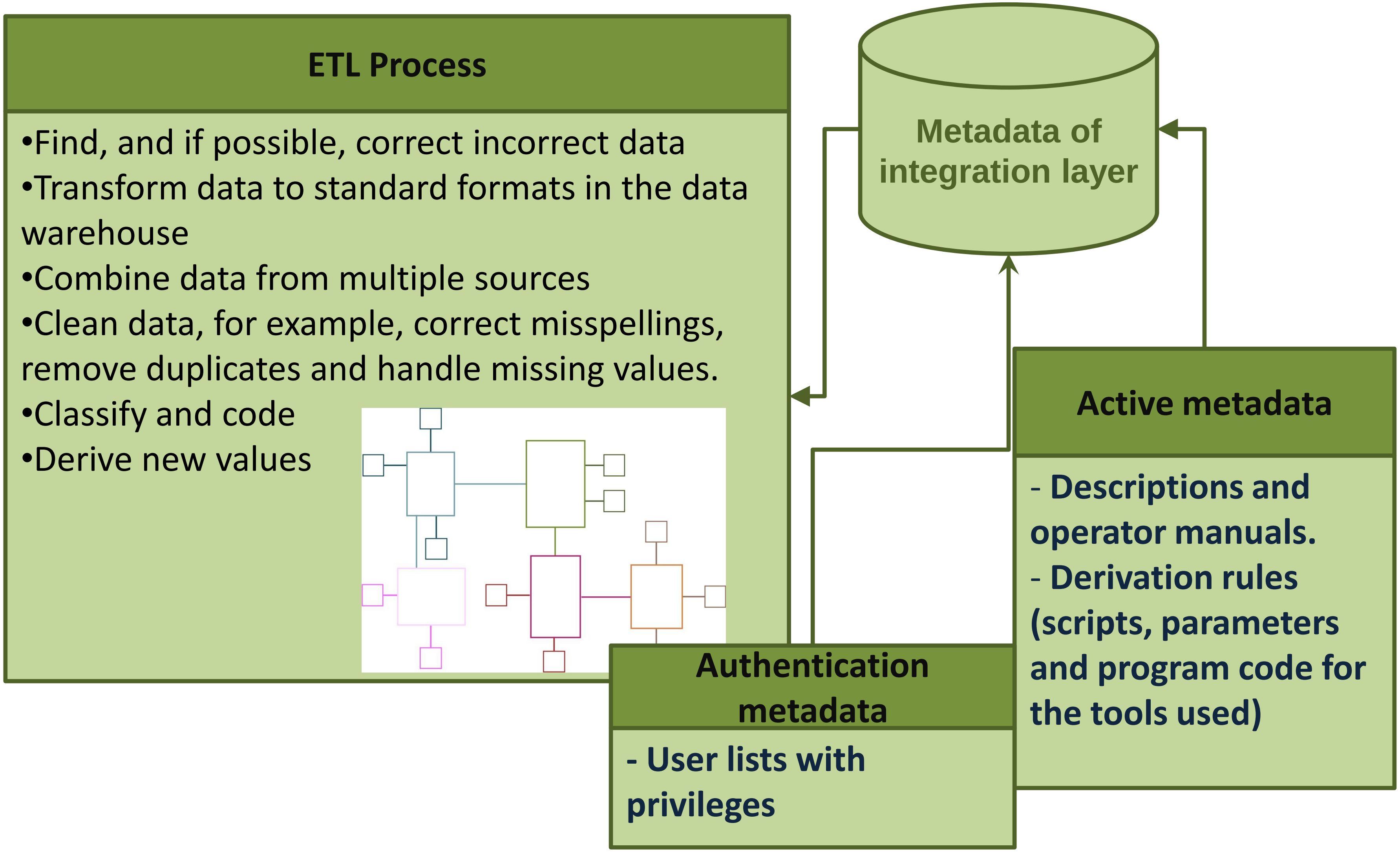
OLTP - Characteristics

Source of data	Operational data
Purpose of data	To control and run fundamental business tasks
Processing Speed	Typically Very Fast
Database Design	Highly normalized with many tables
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability
Age Of Data	Current
Queries	Relatively standardized and simple queries. Returning relatively few records
Data Base Operations	Insert, Delete and Update
What the data Reveals	A snapshot of ongoing business processes

Integration Layer and Metadata

ETL tasks need to use active metadata, such as descriptions and operator manuals as well as derivation rules being used, i.e. scripts, parameters and program code for the tools used.

Integration Layer and Metadata



ETL Process

- Find, and if possible, correct incorrect data
- Transform data to standard formats in the data warehouse
- Combine data from multiple sources
- Clean data, for example, correct misspellings, remove duplicates and handle missing values.
- Classify and code
- Derive new values

Metadata of integration layer

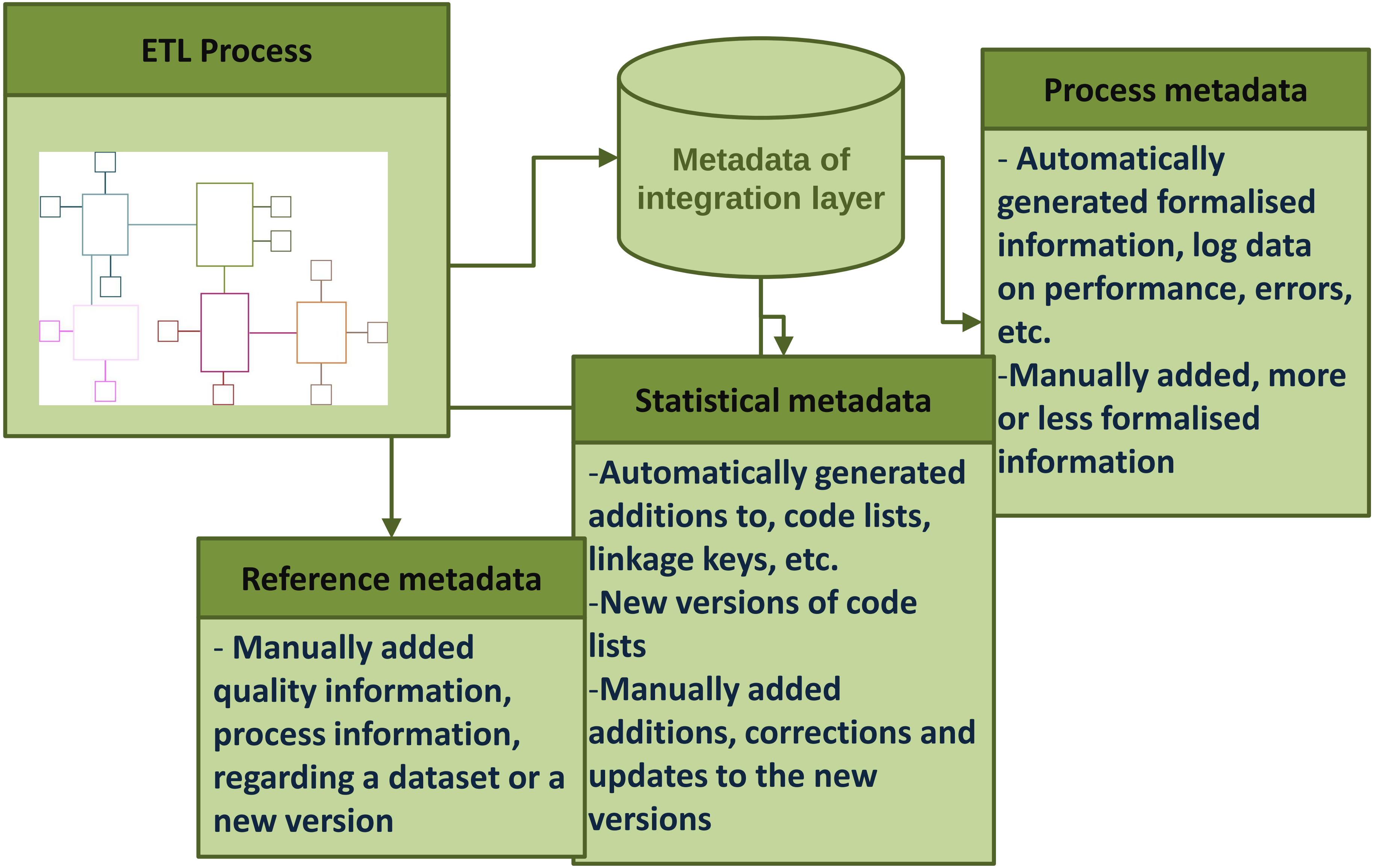
Active metadata

- Descriptions and operator manuals.
- Derivation rules (scripts, parameters and program code for the tools used)

Authentication metadata

- User lists with privileges

Integration Layer and Metadata



ETL Process

Metadata of integration layer

Process metadata

- Automatically generated formalised information, log data on performance, errors, etc.

-Manually added, more or less formalised information

Statistical metadata

-Automatically generated additions to, code lists, linkage keys, etc.

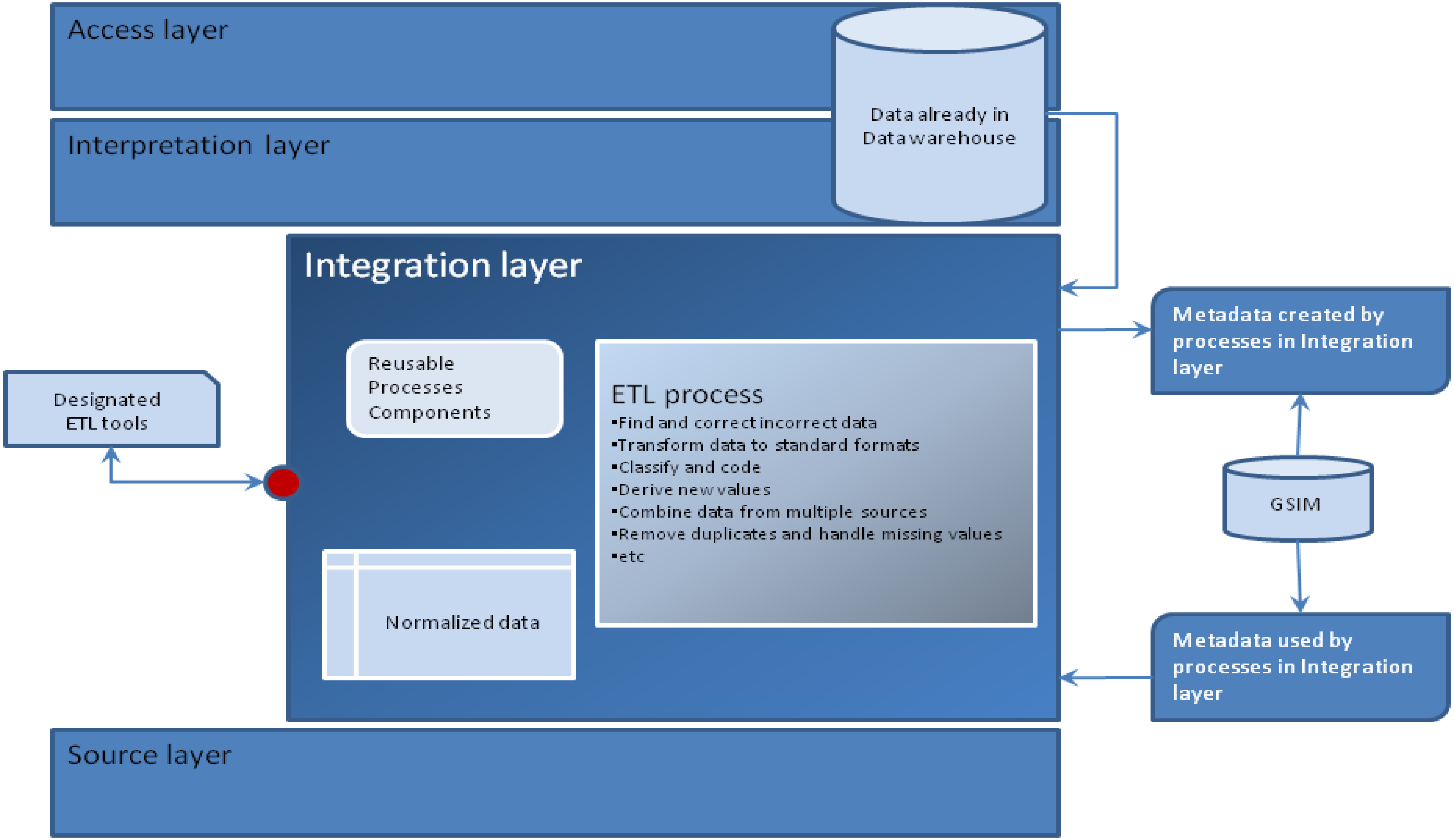
-New versions of code lists

-Manually added additions, corrections and updates to the new versions

Reference metadata

- Manually added quality information, process information, regarding a dataset or a new version

Integration Layer



Access layer
Interpretation layer
Data already in Data warehouse
Integration layer
Reusable Processes Components
ETL process
▪Find and correct incorrect data
▪Transform data to standard formats
▪Classify and code
▪Derive new values
▪Combine data from multiple sources
▪Remove duplicates and handle missing values
▪etc
Normalized data
Designated ETL tools
Metadata created by processes in Integration layer
GSIM
Metadata used by processes in Integration layer
Source layer

Interpretation Layer

Contains all collected data processed and structured to be optimized for analysis and as base for output planned by the NSI.

Its specially designed for statistical experts and is built to support data manipulation of big complex search operations.

Interpretation Layer

Typical activities in the Interpretation layer:

- **Basis analysis**
- **Correlation and Multivariate analysis**
- **Hypothesis testing, simulation and forecasting,**
- **Data mining,**
- **Design of new statistical strategies,**
- **Design data cubes to the Access layer.**

Interpretation Layer – Data Model

Its underlying model is not specific to a particular reporting or analytic requirement.

Instead of focusing on a process-oriented design, the design is modelled based on data inter-relationships

Interpretation Layer – Data Model

Although data warehouses are built on relational database technology, it’s database design differs substantially from the online OLTP database.

Interpretation Layer – OLAP

OnLine Analytical Processing (OLAP):

- **Subject orientated**
- **Designed to provide real-time analysis**
- **Data is historical**
- **Highly De-normalized**

↓

multi-dimensional and are optimised for processing

very complex real-time ad-hoc read queries

OLAP - Characteristics

Source of data	Consolidated data; OLAP data comes from the various OLTP Databases
Purpose of data	To help with planning, problem solving, and decision support
Processing Speed	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Design	Typically de-normalized with fewer tables; use of star schemas.
Backup	Regular backups
Age Of Data	Historical
Queries	Often complex queries involving aggregations
DB Operations	Read
What the data Reveals	Multi-dimensional views of various kinds of statistical activities

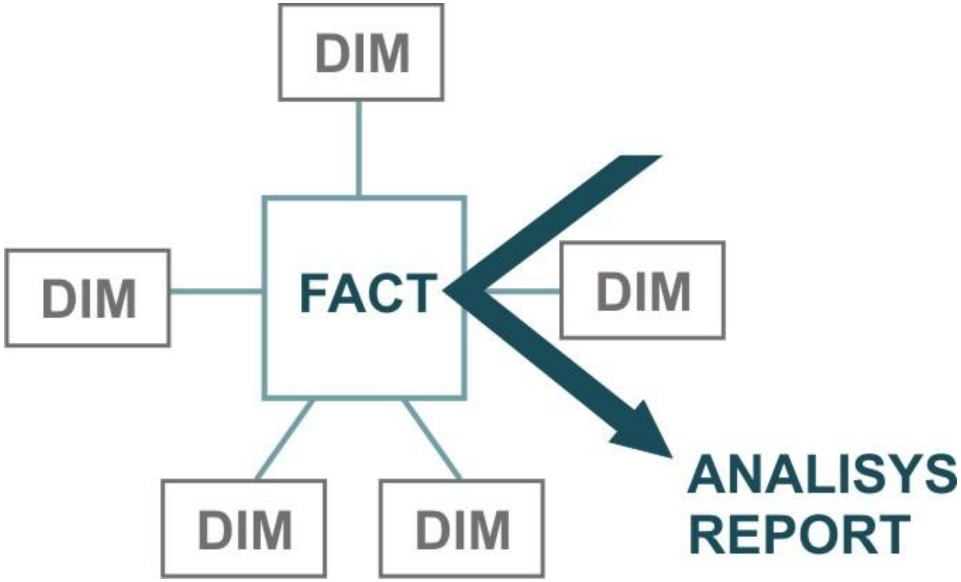
Interpretation Layer – Data Model

In this layer a specific type of OLAP should be used:

ROLAP - Relational Online Analytical Processing - uses specific analytical tools on a relational dimensional data model which is easy to understand and does not require pre-computation and storage of the information.

Interpretation Layer – Data Model

A star-schema design should be implemented with central Fact Tables (metrics or measures) related to Dimension Tables (De-normalised Labels – provide context to the facts/metrics/measures).



Interpretation Layer – Data Model

A dimension is a structural attribute of a cube that has a list of members, all of which are of a similar type in the user's perception of the data.

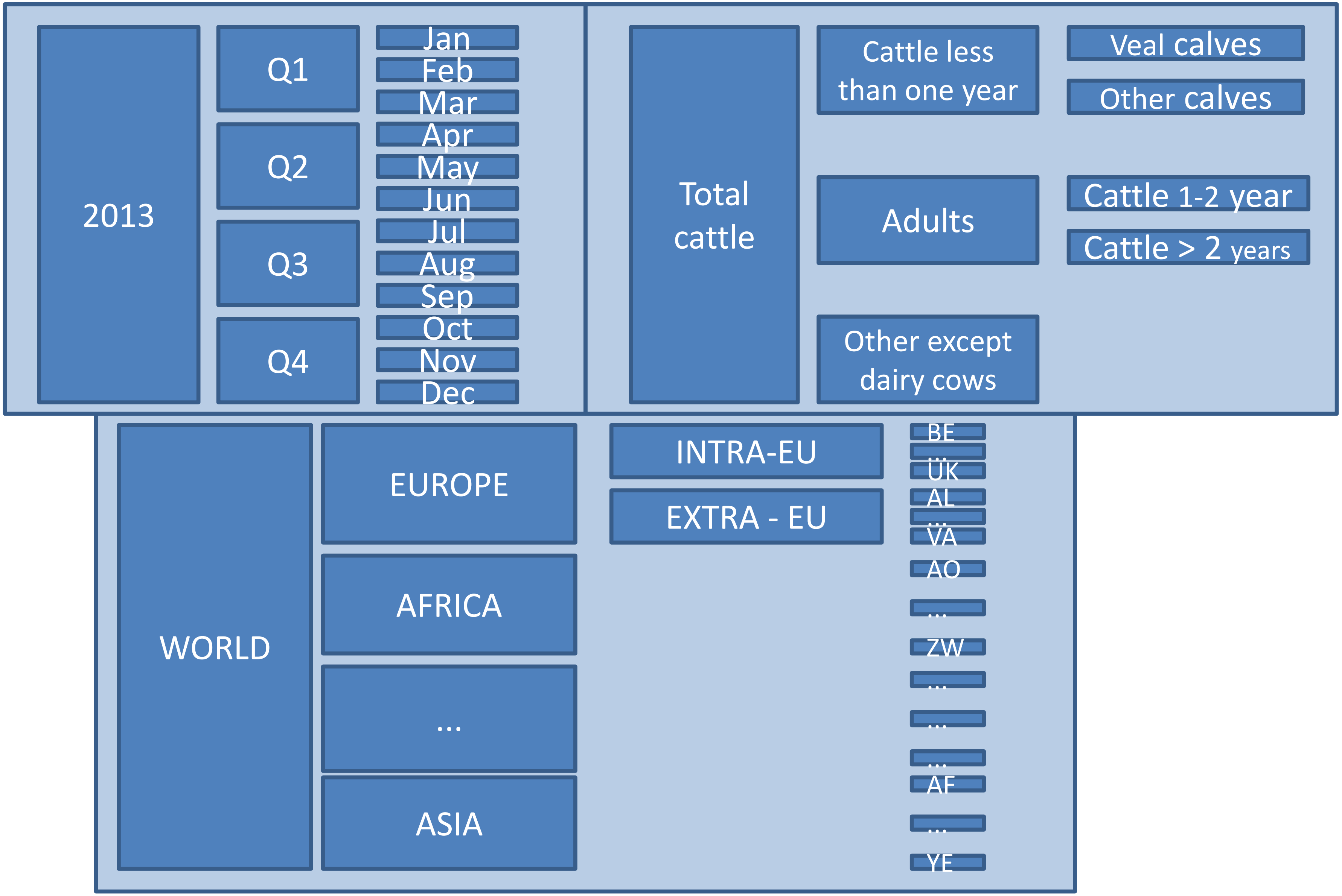
For example, all months, quarters, years, etc., make up a time dimension; likewise all cities, regions, countries, etc., make up a geography dimension.

Interpretation Layer – Data Model

Dimension could have hierarchy, which are classified into levels.

For example, in a "Time" dimension, level one stands for days, level two for months and level three for years.

Interpretation Layer – Data Model



2013

Q1 Q2 Q3 Q4

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Total cattle

Cattle less than one year

Veal calves

Other calves

Adults

Cattle 1-2 year

Cattle > 2 years

Other except dairy cows

WORLD

EUROPE

AFRICA

...

ASIA

INTRA-EU

EXTRA - EU

BE ... UK AL ... VA AO ... ZW ... ... AF ... YE

Interpretation Layer and Metadata

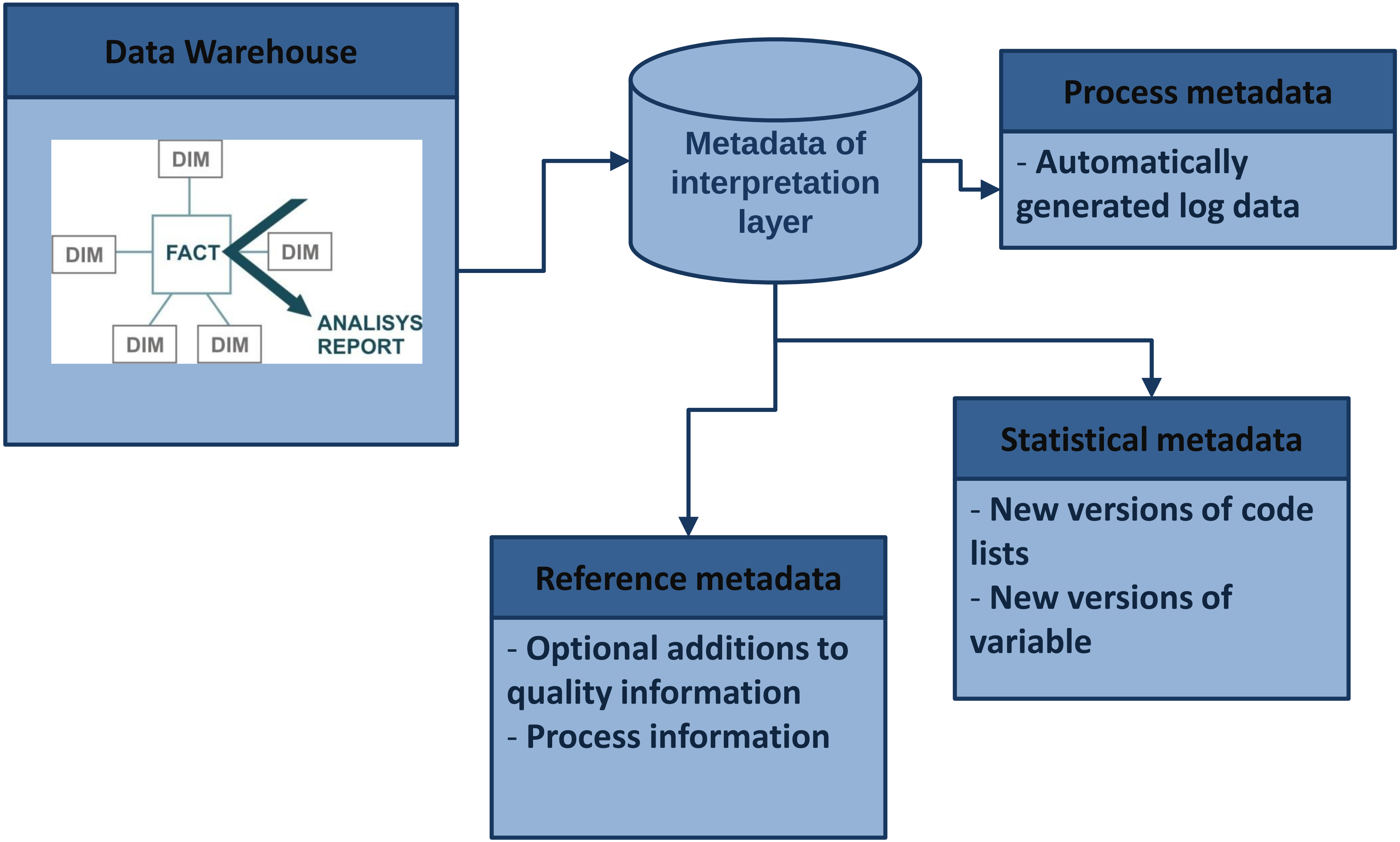
Stores cleaned, versioned and well-structured final micro data.

Once a new dataset or a new version has been loaded few updates are made to the data.

↓

Metadata are normally added, with few or no changes being made.

Interpretation Layer and Metadata



Data Warehouse

Metadata of interpretation layer

Process metadata

- Automatically generated log data

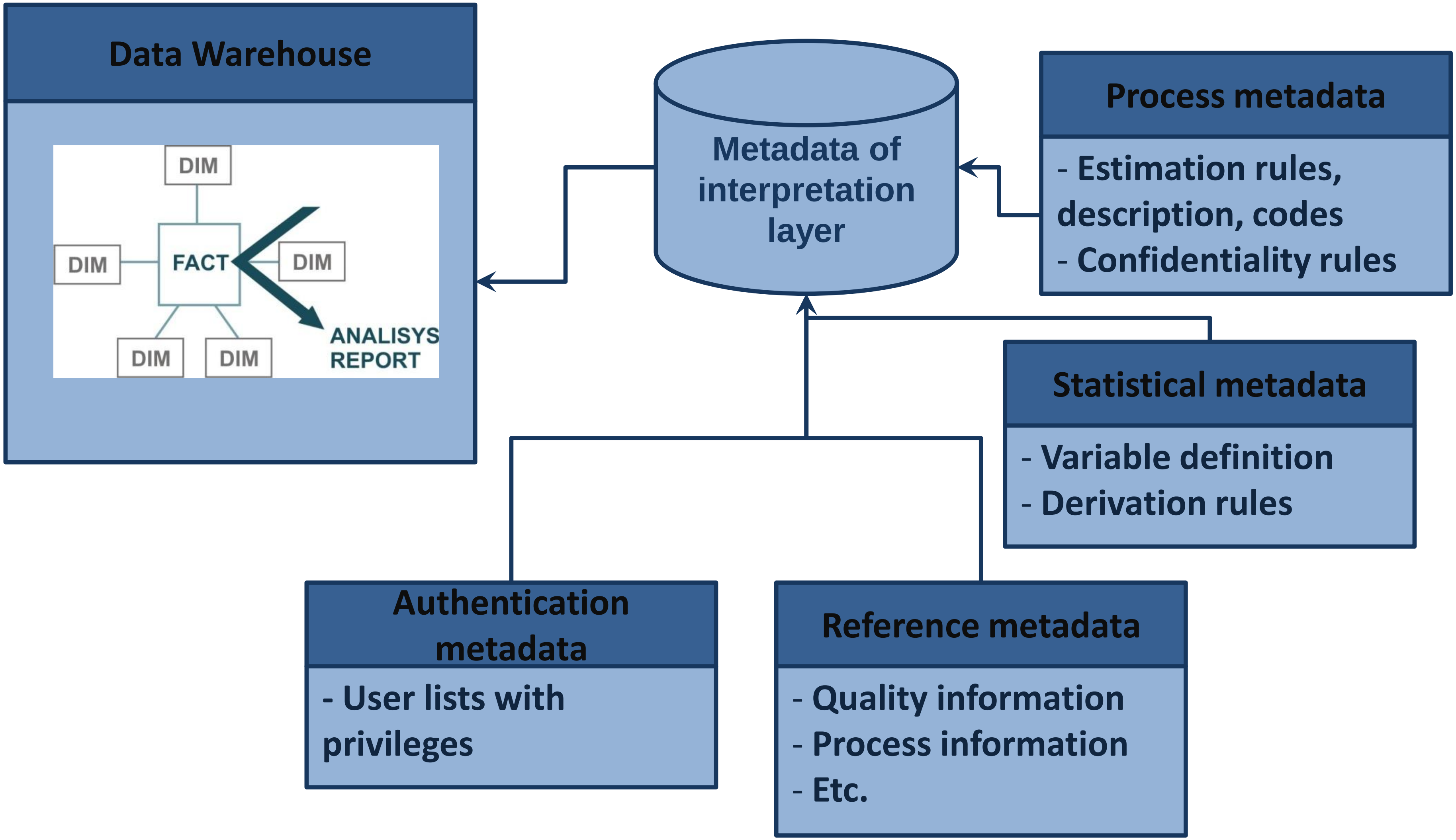
Statistical metadata

- New versions of code lists
- New versions of variable

Reference metadata

- Optional additions to quality information
- Process information

Interpretation Layer and Metadata



Data Warehouse

Metadata of interpretation layer

Process metadata

- Estimation rules, description, codes
- Confidentiality rules

Statistical metadata

- Variable definition
- Derivation rules

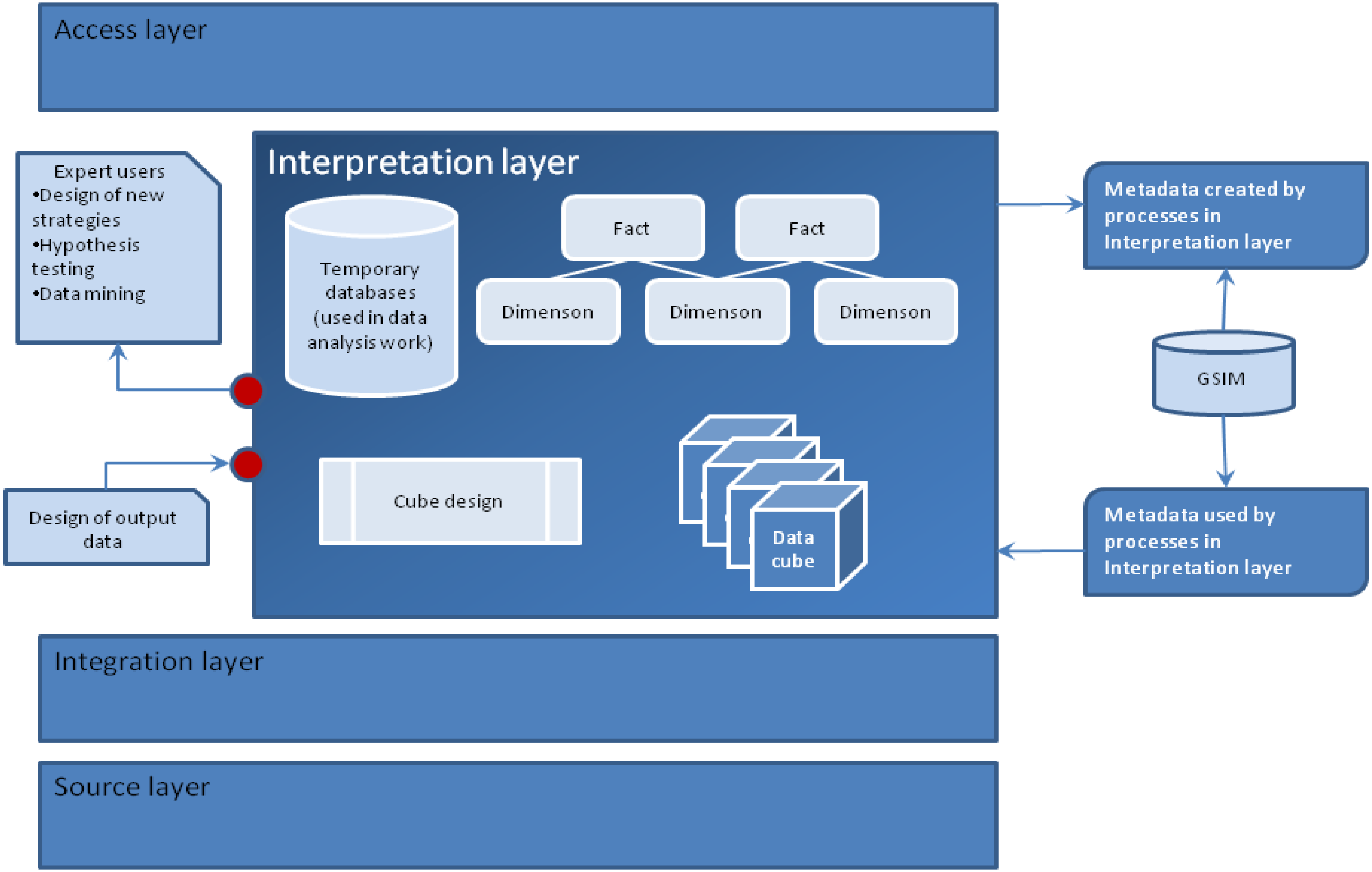
Authentication metadata

- User lists with privileges

Reference metadata

- Quality information
- Process information
- Etc.

Interpretation Layer



Access Layer

Is for the final presentation, dissemination and delivery of information.

Is used by a wide range of users and computer instruments.

The data is optimized to present and compile data effectively.

Access Layer – Data Mart

Is a simple form of a data warehouse that is focused on a single subject (or functional area).

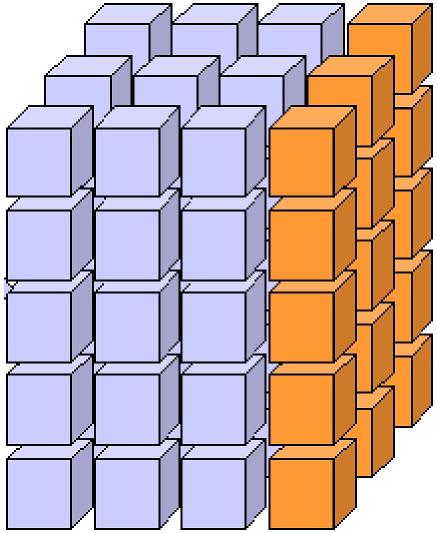
Data may be presented in data cubes with different formats, specialized to support different tools and software.

Access Layer - Data Model

Generally the data structure are optimized for MOLAP (Multidimensional Online Analytical Processing) that uses specific analytical tools on a multidimensional data model

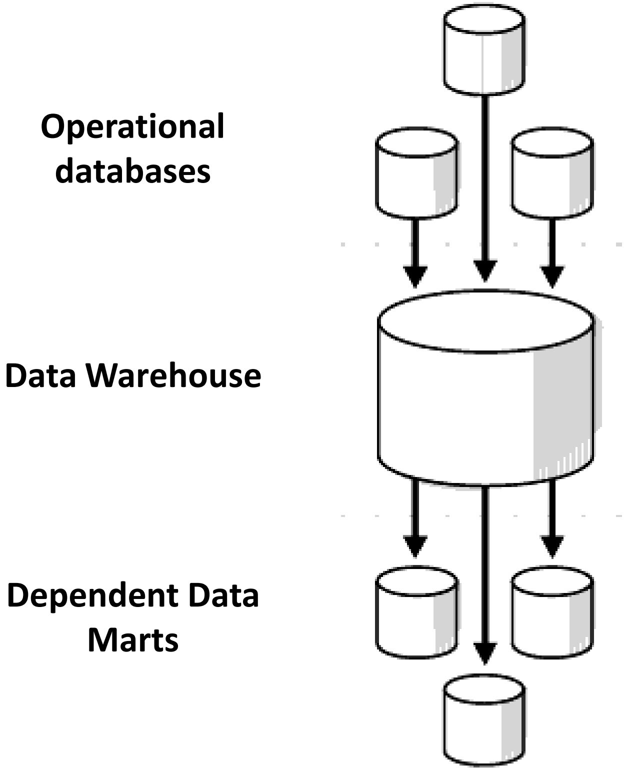
Access Layer - Data Model

Usually it requires pre-computation and storage of information in an optimized multidimensional array storage



Access Layer - Data Model

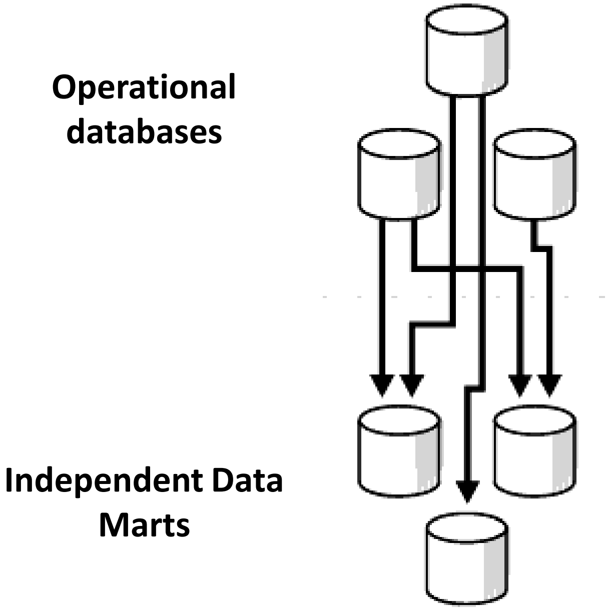
There are 2 basic types of data marts:



Operational databases

Data Warehouse

Dependent Data Marts



Operational databases

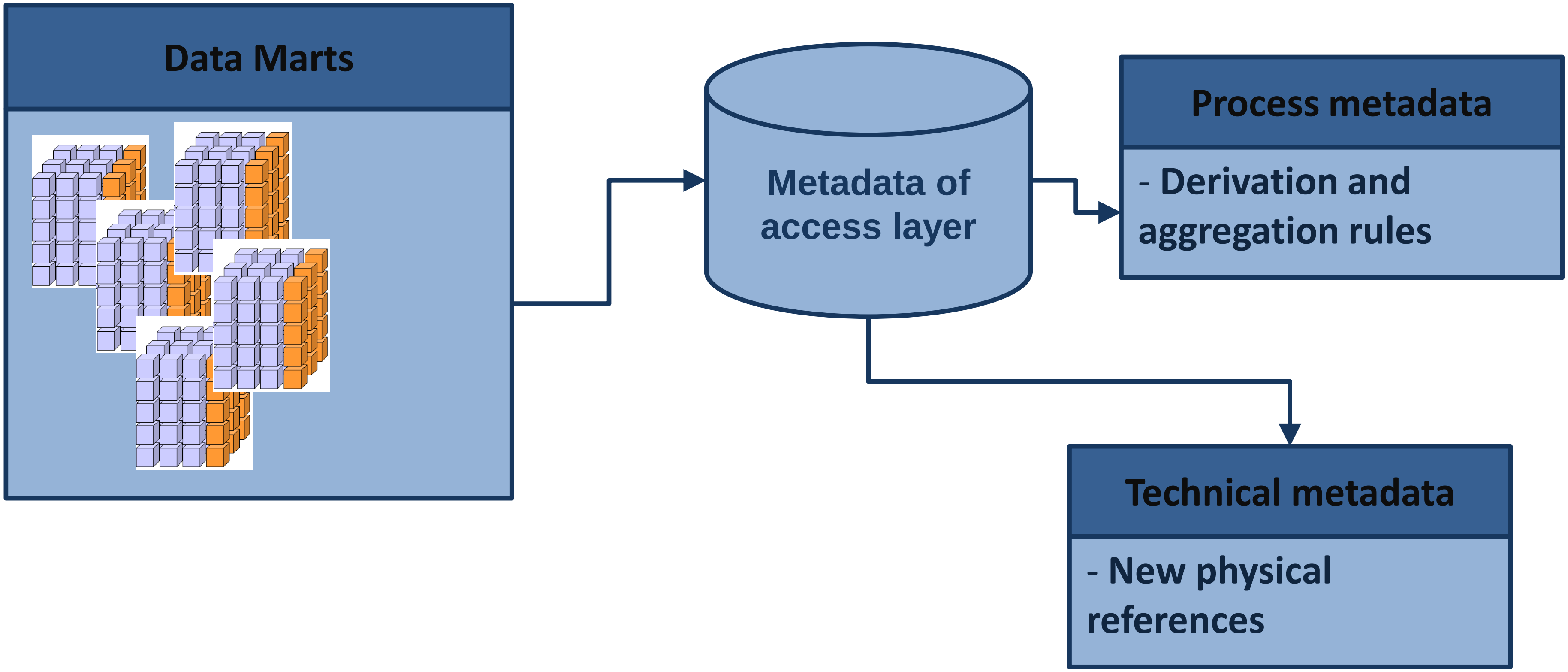
Independent Data Marts

Access Layer and Metadata

Loading data into the access layer means reorganising data from the analysis layer by derivation or aggregation into data marts.

Metadata that describe and support the process itself (derivation and aggregation rules), but also metadata that describe the reorganised data.

Access Layer and Metadata



Data Marts

Metadata of access layer

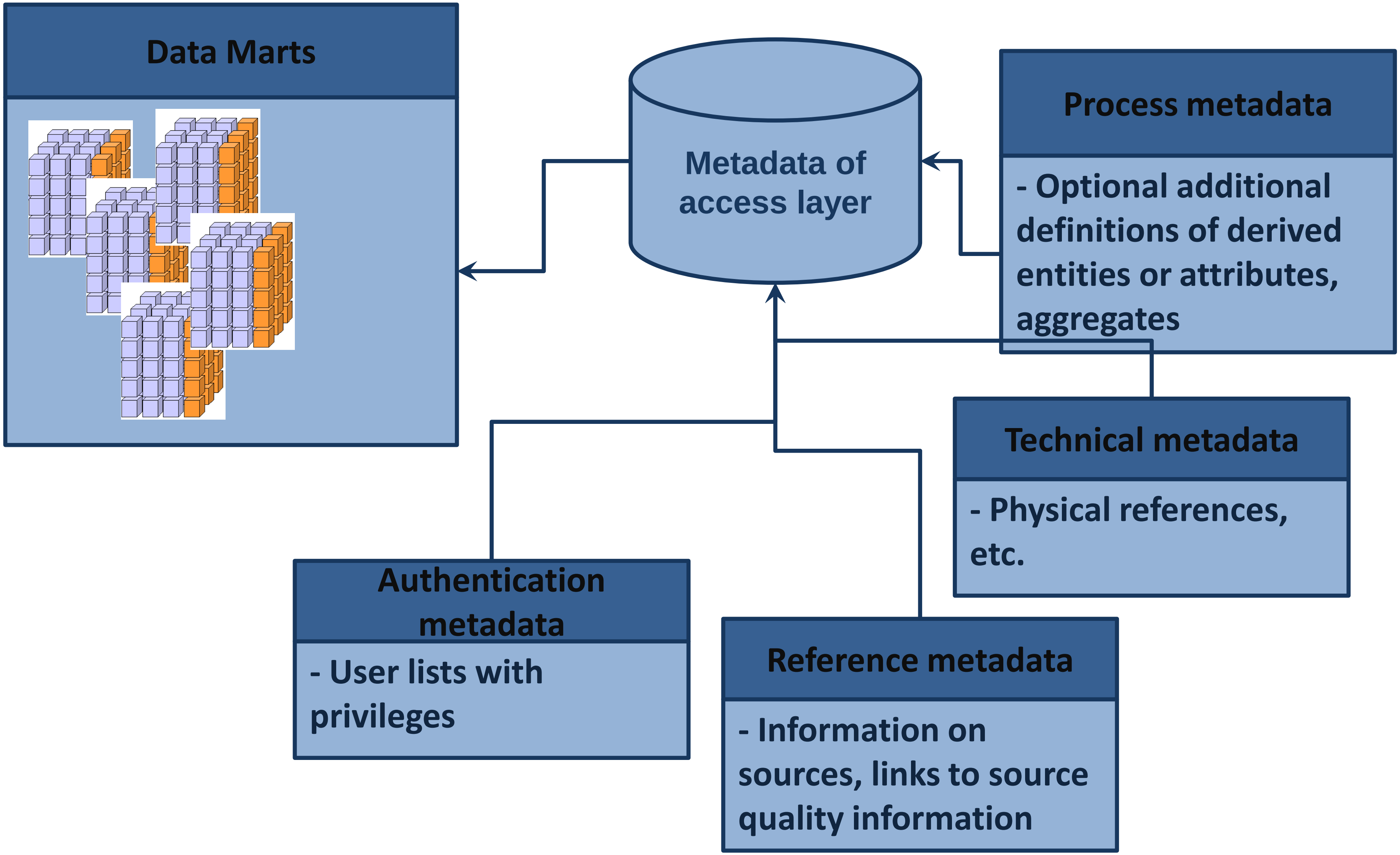
Process metadata

- Derivation and aggregation rules

Technical metadata

- New physical references

Access Layer and Metadata



Data Marts

Metadata of access layer

Process metadata

- Optional additional definitions of derived entities or attributes, aggregates

Technical metadata

- Physical references, etc.

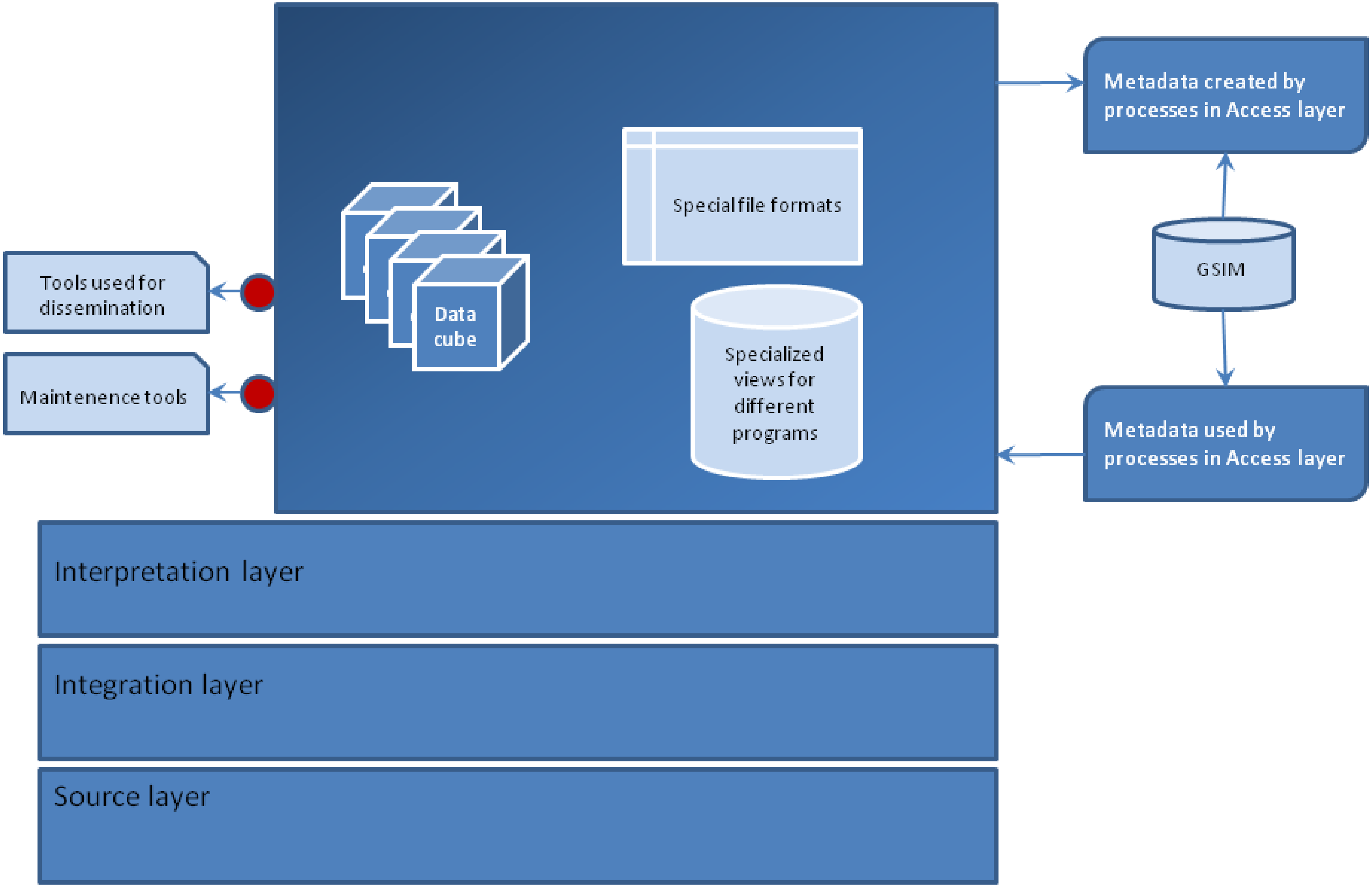
Authentication metadata

- User lists with privileges

Reference metadata

- Information on sources, links to source quality information

Access Layer

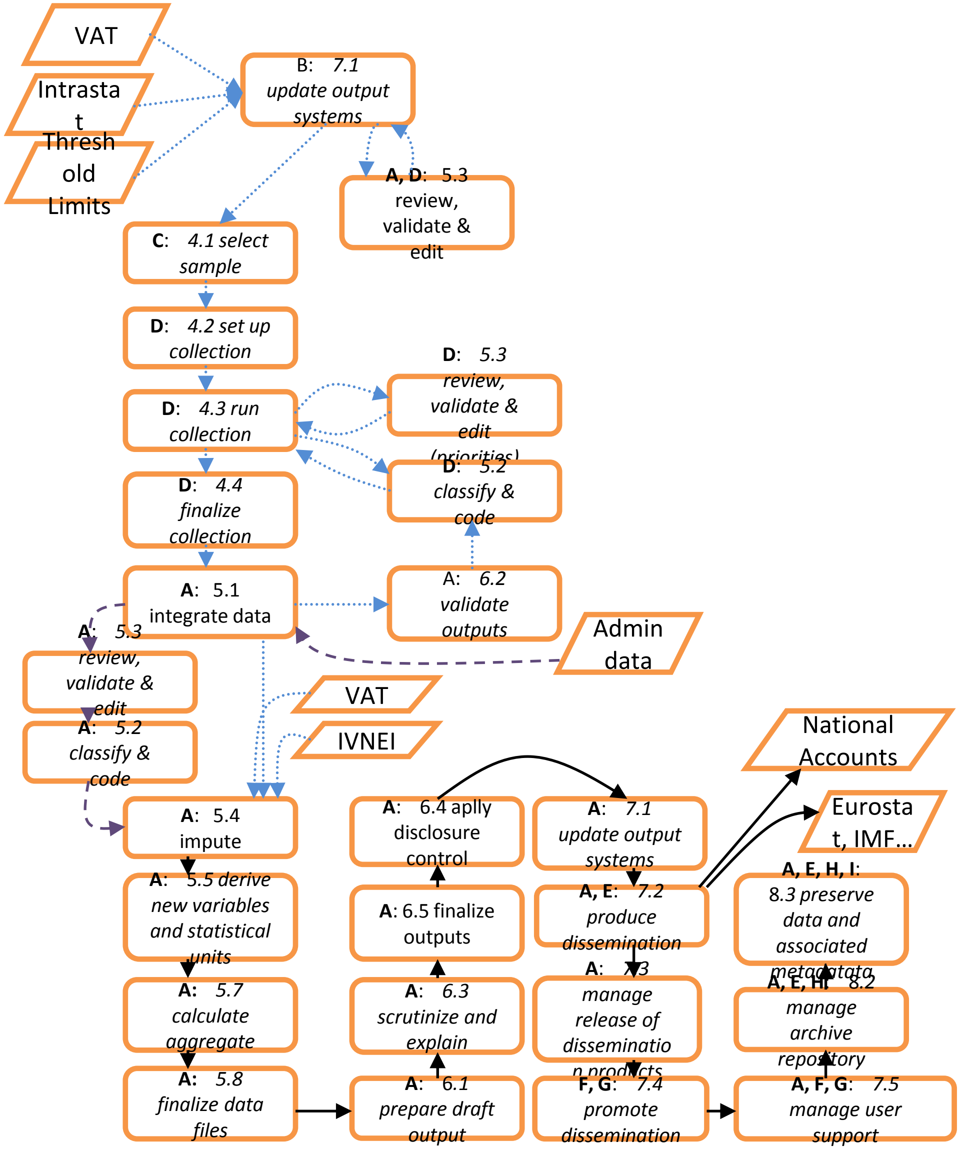

Data cube
Special file formats
Specialized views for different programs
Tools used for dissemination
Maintenence tools
Metadata created by processes in Access layer
GSIM
Metadata used by processes in Access layer
Interpretation layer
Integration layer
Source layer

External trade – case study

External Trade statistics track the value and quantity of goods traded between EU Member States (intra-EU trade) and between Member States and non-EU countries (extra-EU trade).

Intra-EU is survey based and extra-EU is admin base.

External trade – case study



VAT

Intrastat

Threshold Limits

B: *7.1 update output systems*

A, D: 5.3 review, validate & edit

C: *4.1 select sample*

D: *4.2 set up collection*

D: *4.3 run collection*

D: *5.3 review, validate & edit (priorities)*

D: *5.2 classify & code*

D: *4.4 finalize collection*

A: 5.1 integrate data

A: *6.2 validate outputs*

Admin data

A: *5.3 review, validate & edit*

A: *5.2 classify & code*

VAT

IVNEI

A: 5.4 impute

A: *5.5 derive new variables and statistical units*

A: *5.7 calculate aggregate*

A: *5.8 finalize data files*

A: *6.1 prepare draft output*

A: *6.3 scrutinize and explain*

A: 6.5 finalize outputs

A: 6.4 aplly disclosure control

A: *7.1 update output systems*

A, E: *7.2 produce dissemination*

A: *7.3 manage release of dissemination products*

F, G: *7.4 promote dissemination*

A, F, G: *7.5 manage user support*

National Accounts

Eurostat, IMF…

A, E, H, I: *8.3 preserve data and associated metadata*

A, E, H: *8.2 manage archive repository*

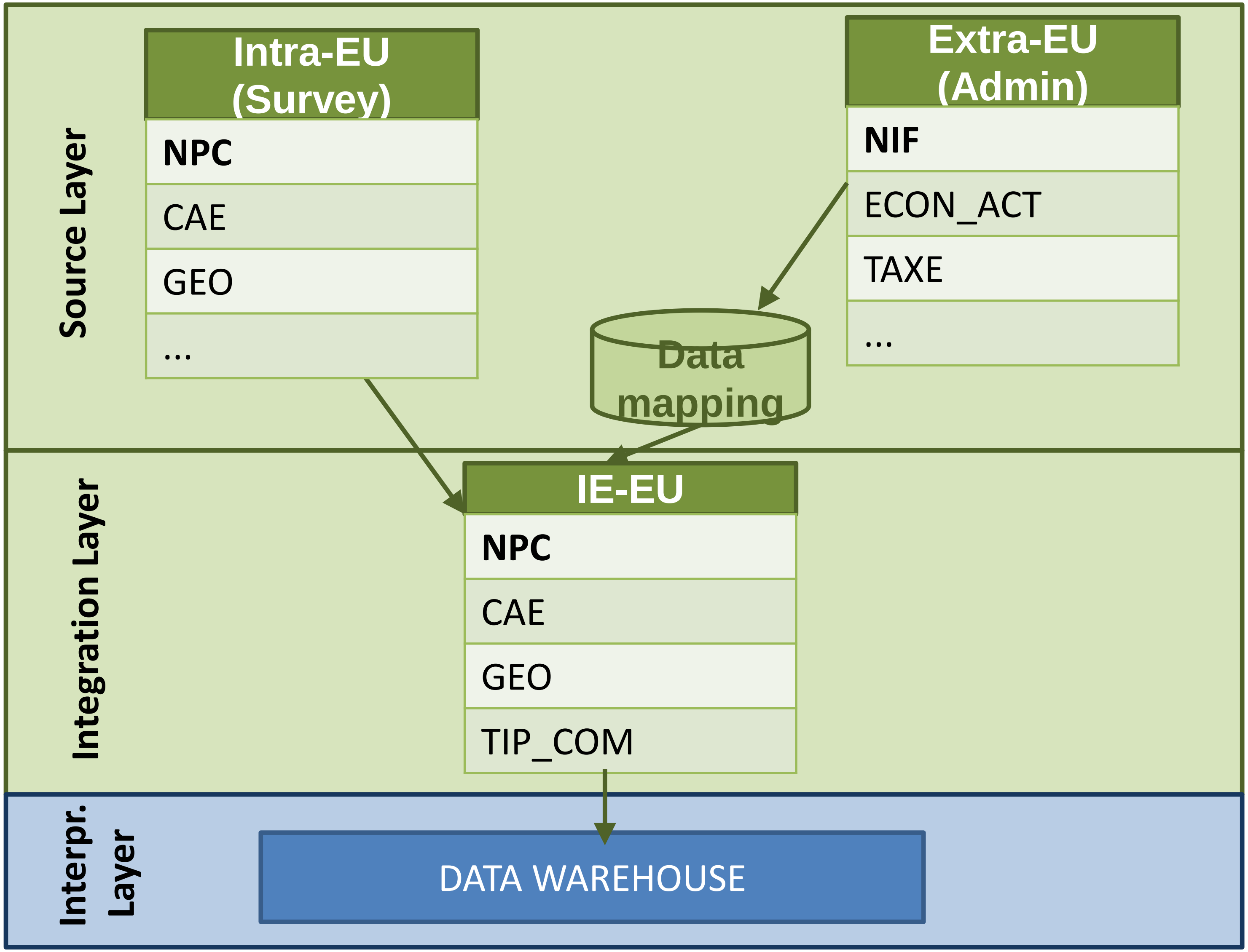
A: External trade statistical unit
B: Business Register
C: Statistical Methods unit
D: Information Collection department
E: Data Warehouse unit
F: Dissemination unit
G: PR unit
H: Software Development unit
I: Metadata unit
(Blue) Intra union (Purple)Extra union (Black) Both

External trade – case study

For demonstration purposes, let's make it simple:

- **Information about Enterprises (NPC, Economic Activity (CAEV3.0), GEO)**
- **Imports and exports (Quantity and Value) of products (Type of goods CN8) and country every month.**

External trade – case study



Source Layer

Intra-EU (Survey)
- NPC
- CAE
- GEO
- …

Extra-EU (Admin)
- NIF
- ECON_ACT
- TAXE
- …

Data mapping

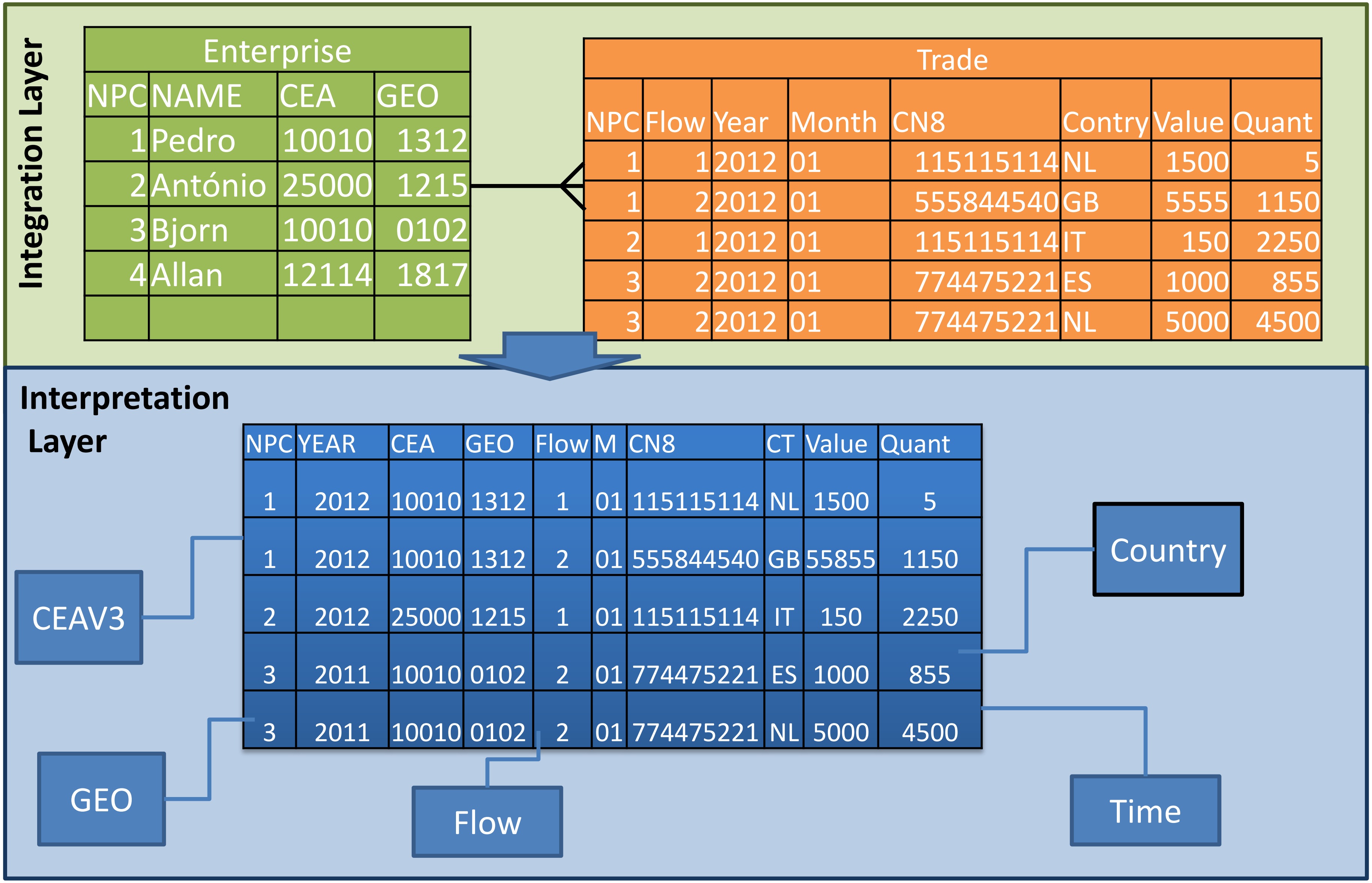
Integration Layer

IE-EU
- NPC
- CAE
- GEO
- TIP_COM

Interpr. Layer

DATA WAREHOUSE

External trade – case study



Integration Layer

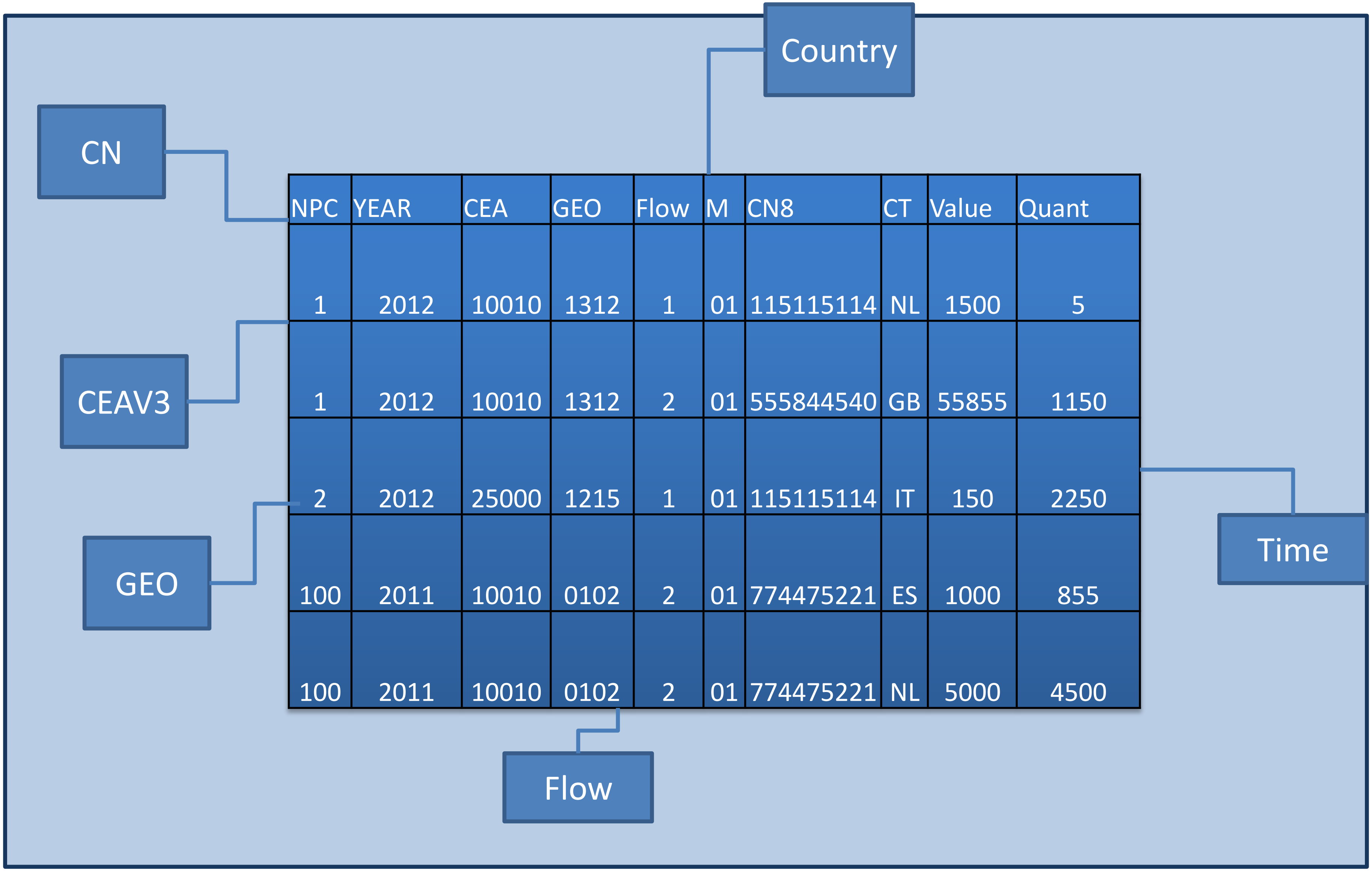
Enterprise			
NPC	NAME	CEA	GEO
1	Pedro	10010	1312
2	António	25000	1215
3	Bjorn	10010	0102
4	Allan	12114	1817

Trade							
NPC	Flow	Year	Month	CN8	Contry	Value	Quant
1	1	2012	01	115115114	NL	1500	5
1	2	2012	01	555844540	GB	5555	1150
2	1	2012	01	115115114	IT	150	2250
3	2	2012	01	774475221	ES	1000	855
3	2	2012	01	774475221	NL	5000	4500

Interpretation Layer

NPC	YEAR	CEA	GEO	Flow	M	CN8	CT	Value	Quant
1	2012	10010	1312	1	01	115115114	NL	1500	5
1	2012	10010	1312	2	01	555844540	GB	55855	1150
2	2012	25000	1215	1	01	115115114	IT	150	2250
3	2011	10010	0102	2	01	774475221	ES	1000	855
3	2011	10010	0102	2	01	774475221	NL	5000	4500

External trade – De-normalizing data



Country

CN

CEAV3

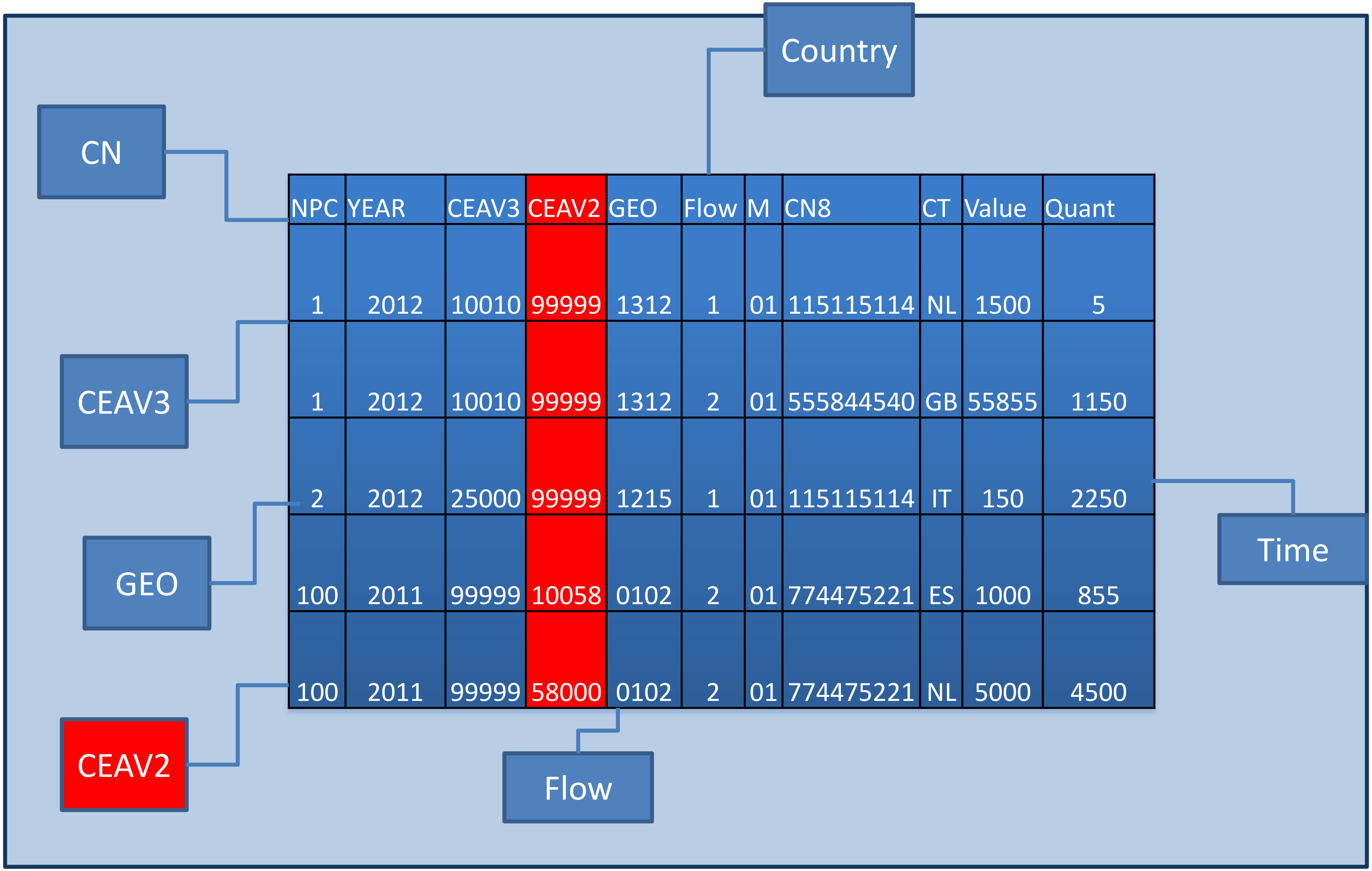
GEO

Time

Flow

NPC	YEAR	CEA	GEO	Flow	M	CN8	CT	Value	Quant
1	2012	10010	1312	1	01	115115114	NL	1500	5
1	2012	10010	1312	2	01	555844540	GB	55855	1150
2	2012	25000	1215	1	01	115115114	IT	150	2250
100	2011	10010	0102	2	01	774475221	ES	1000	855
100	2011	10010	0102	2	01	774475221	NL	5000	4500

External trade – De-normalizing data



NPC	YEAR	CEAV3	CEAV2	GEO	Flow	M	CN8	CT	Value	Quant
1	2012	10010	99999	1312	1	01	115115114	NL	1500	5
1	2012	10010	99999	1312	2	01	555844540	GB	55855	1150
2	2012	25000	99999	1215	1	01	115115114	IT	150	2250
100	2011	99999	10058	0102	2	01	774475221	ES	1000	855
100	2011	99999	58000	0102	2	01	774475221	NL	5000	4500

Country

CN

CEAV3

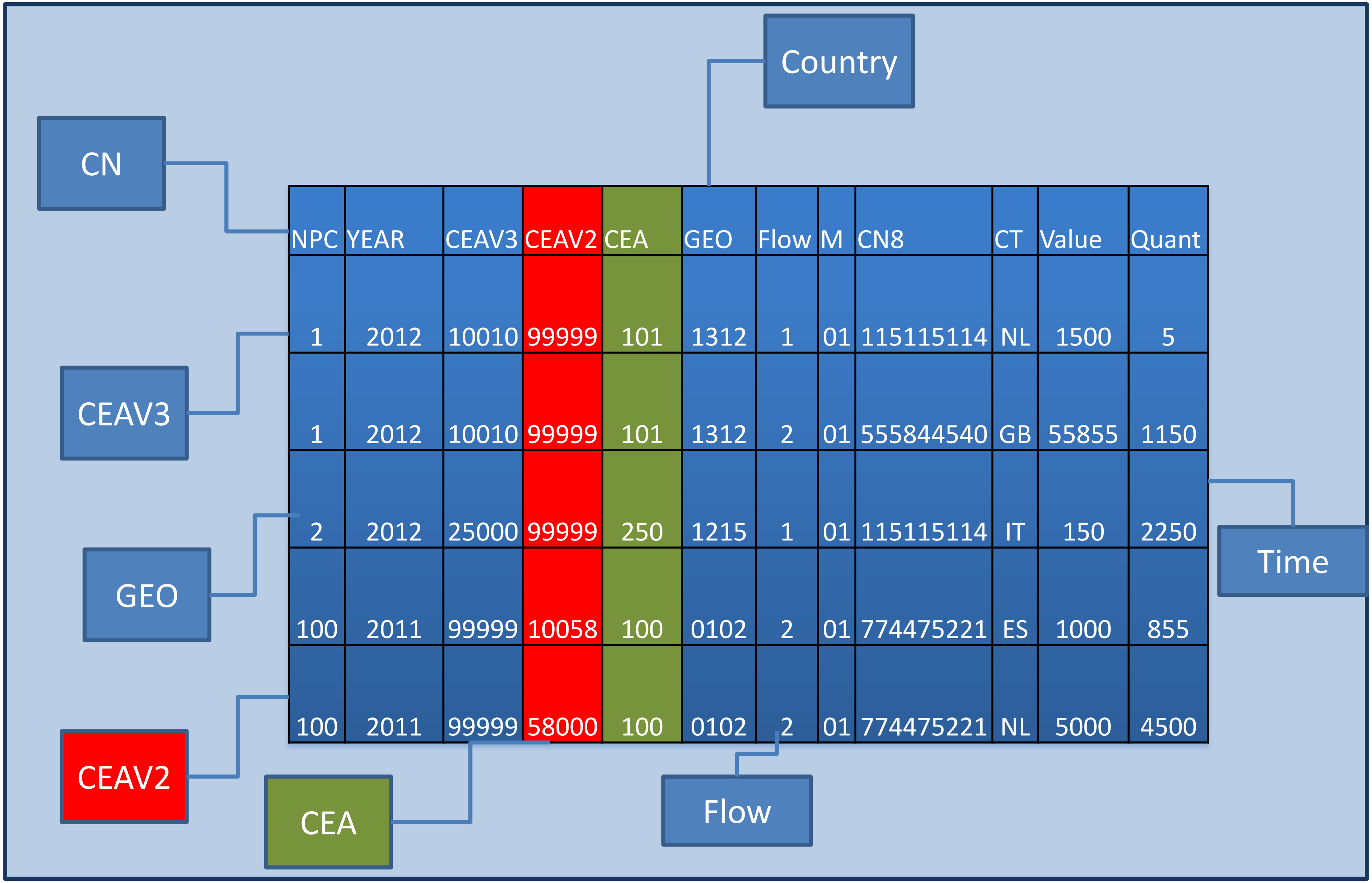
GEO

CEAV2

Flow

Time

External trade – De-normalizing data



NPC	YEAR	CEAV3	CEAV2	CEA	GEO	Flow	M	CN8	CT	Value	Quant
1	2012	10010	99999	101	1312	1	01	115115114	NL	1500	5
1	2012	10010	99999	101	1312	2	01	555844540	GB	55855	1150
2	2012	25000	99999	250	1215	1	01	115115114	IT	150	2250
100	2011	99999	10058	100	0102	2	01	774475221	ES	1000	855
100	2011	99999	58000	100	0102	2	01	774475221	NL	5000	4500

Country

CN

CEAV3

GEO

CEAV2

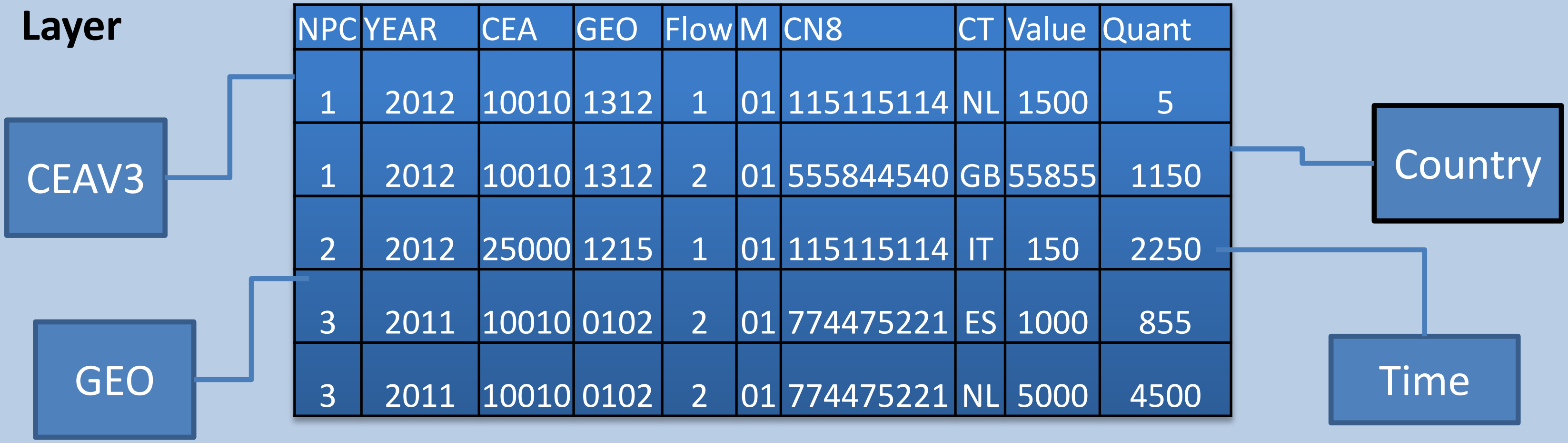
CEA

Flow

Time

External trade – case study – Creating Data Marts

Interpretation Layer



NPC	YEAR	CEA	GEO	Flow	M	CN8	CT	Value	Quant
1	2012	10010	1312	1	01	115115114	NL	1500	5
1	2012	10010	1312	2	01	555844540	GB	55855	1150
2	2012	25000	1215	1	01	115115114	IT	150	2250
3	2011	10010	0102	2	01	774475221	ES	1000	855
3	2011	10010	0102	2	01	774475221	NL	5000	4500

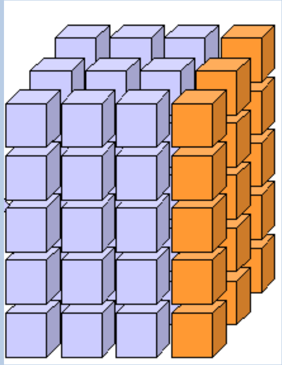
CEAV3

GEO

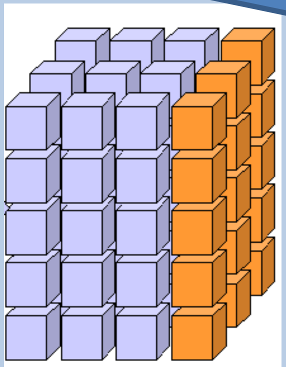
Country

Time

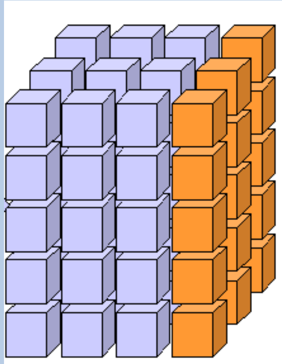
Access Layer



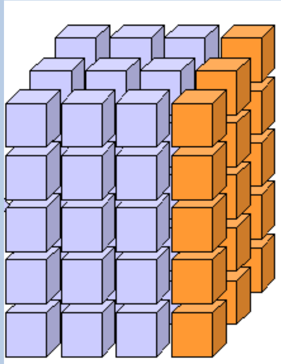
Imports (Month, CN2, Country)



Exports (Month, CN2, Country)

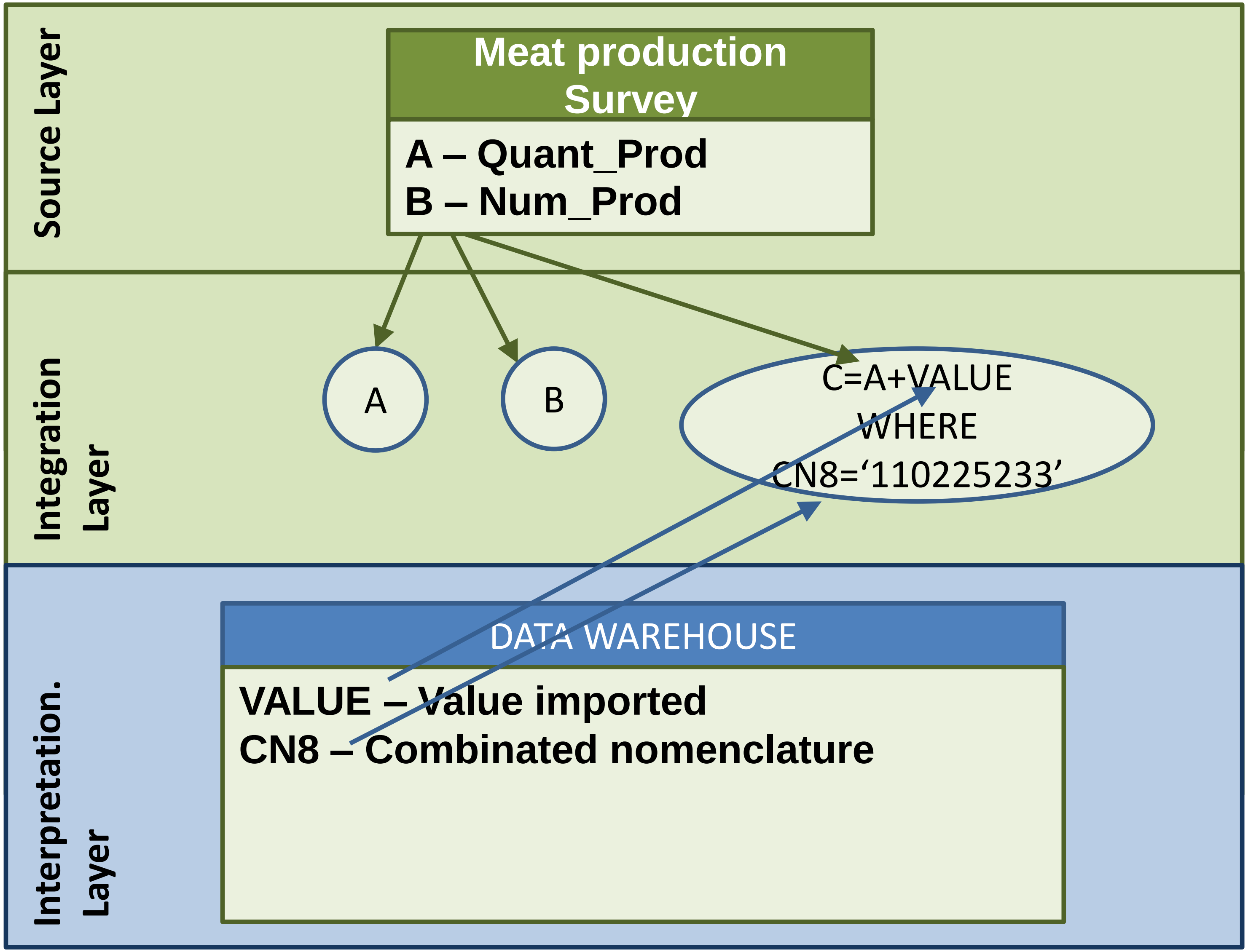


Intra-UE (Year, CN8, Flow)

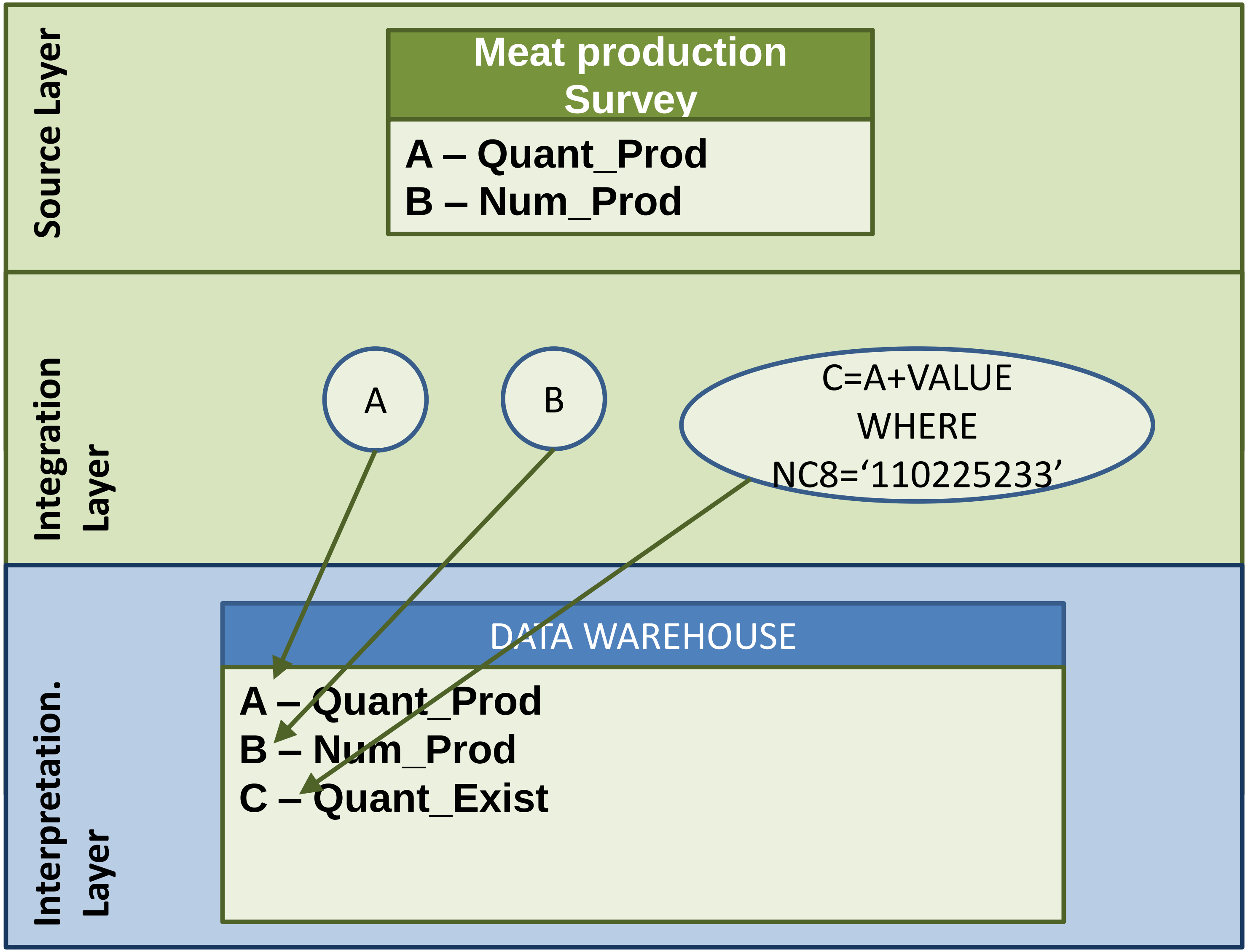


Extra-UE (Year, CN8, Flow)

External trade – Reusing Data to calculate new variables



External trade – Reusing Data to calculate new variables



Thank you for your attention

Questions ?