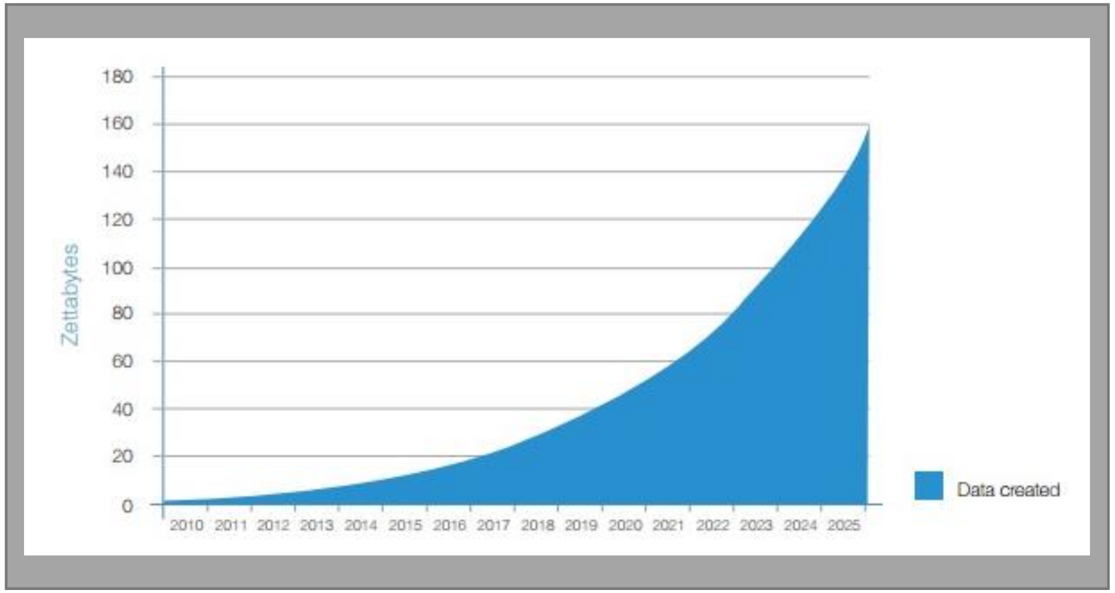
12th International Young Scientists Conference on Computational Science

Framework Architecture of a Secure Big Data Lake

Simar Muratov, Sergey Muravyov
University ITMO

Data Volume Growth

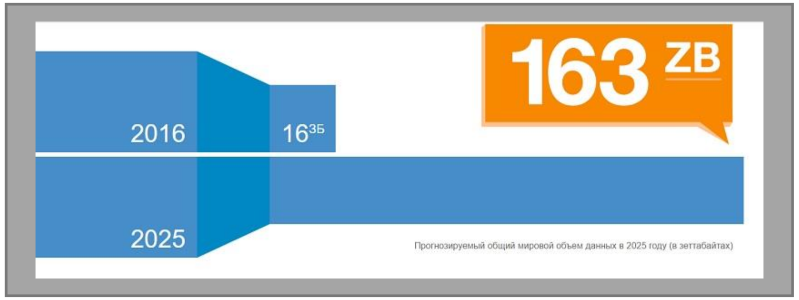


Annual growth in global data volume Seagate forecasts

- Unstoppable growth of data volume;

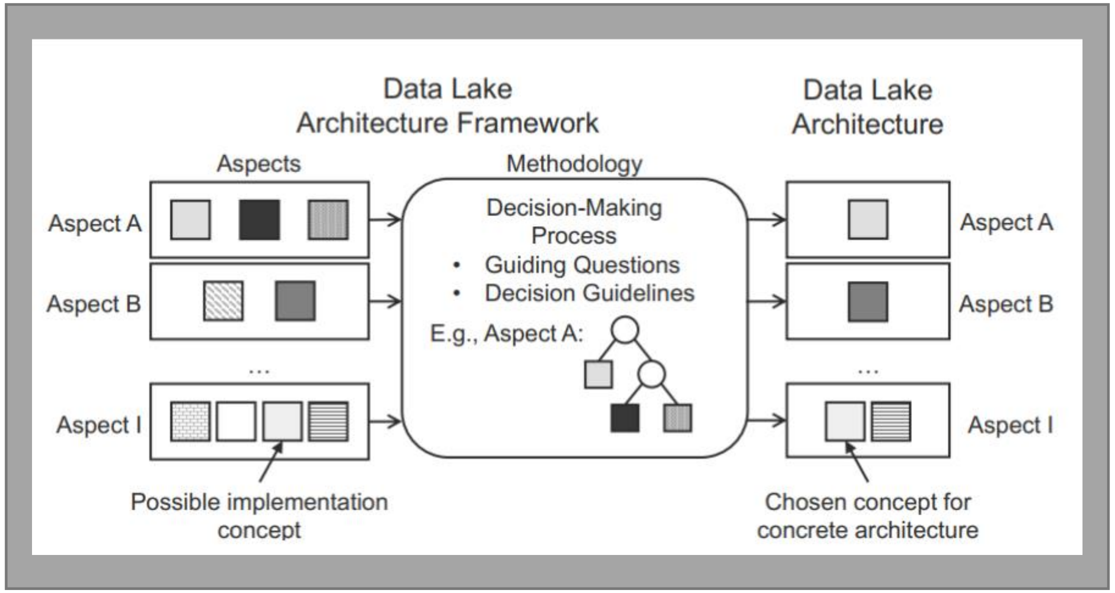
- High speed of progress of technologies for storing and processing ultra-large data;

- The lag of methods and means of protection of the above-mentioned technologies from the development of the technologies themselves.

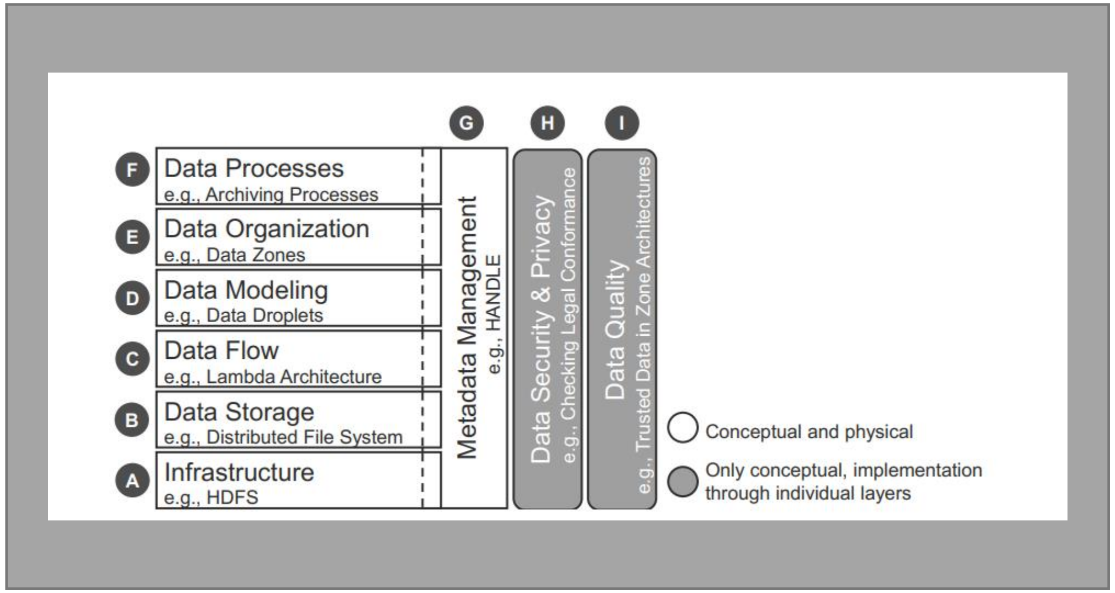


Projected data volume by 2025

Data Lake Architecture Framework

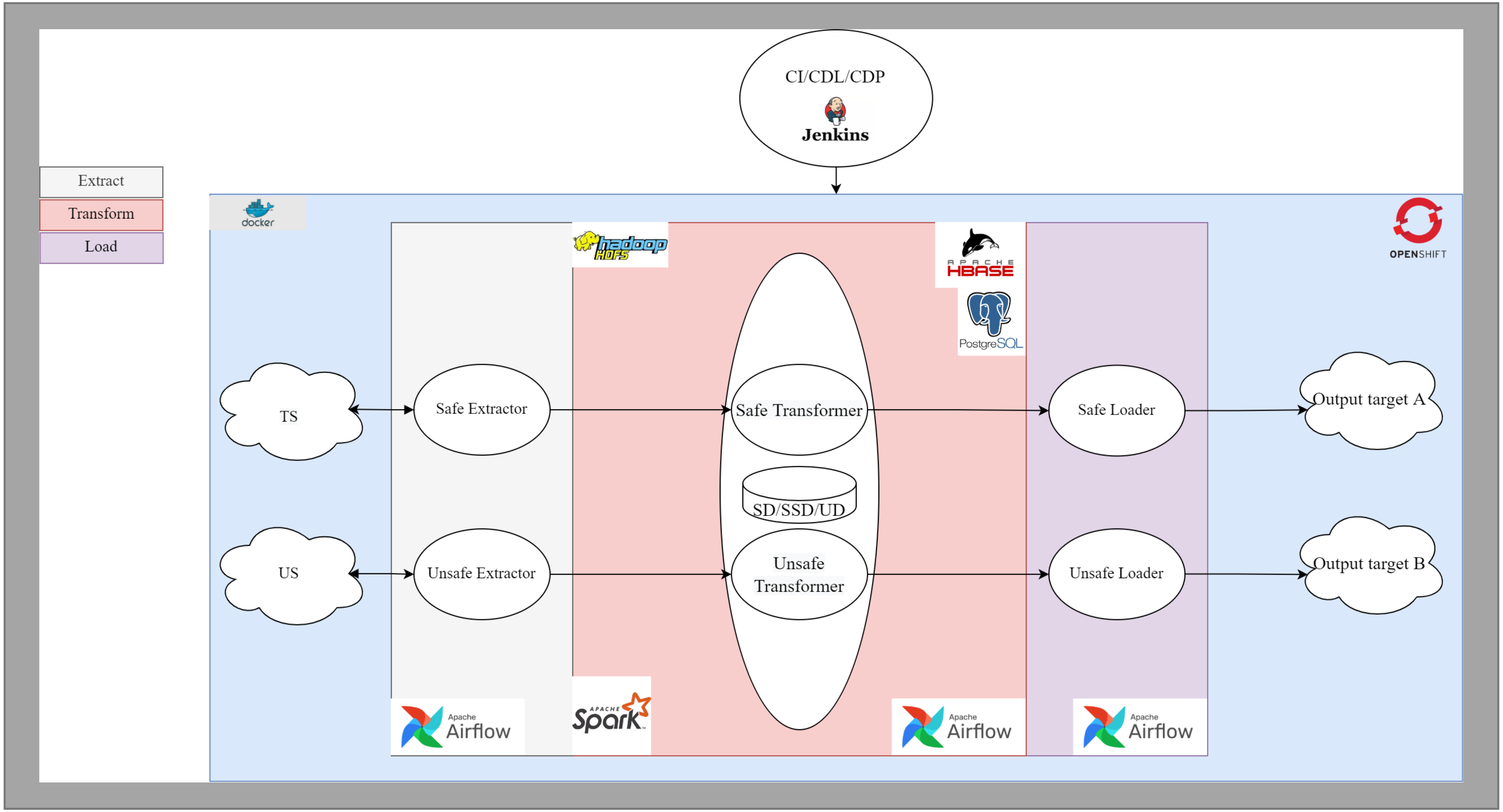


DLAF schema containing possible implementation concepts and configurations


G
H
I
F
Data Processes
e.g., Archiving Processes
E
Data Organization
e.g., Data Zones
D
Data Modeling
e.g., Data Droplets
C
Data Flow
e.g., Lambda Architecture
B
Data Storage
e.g., Distributed File System
A
Infrastructure
e.g., HDFS
Metadata Management
e.g., HANDLE
Data Security & Privacy
e.g., Checking Legal Conformance
Data Quality
e.g., Trusted Data in Zone Architectures
○ Conceptual and physical
● Only conceptual, implementation through individual layers

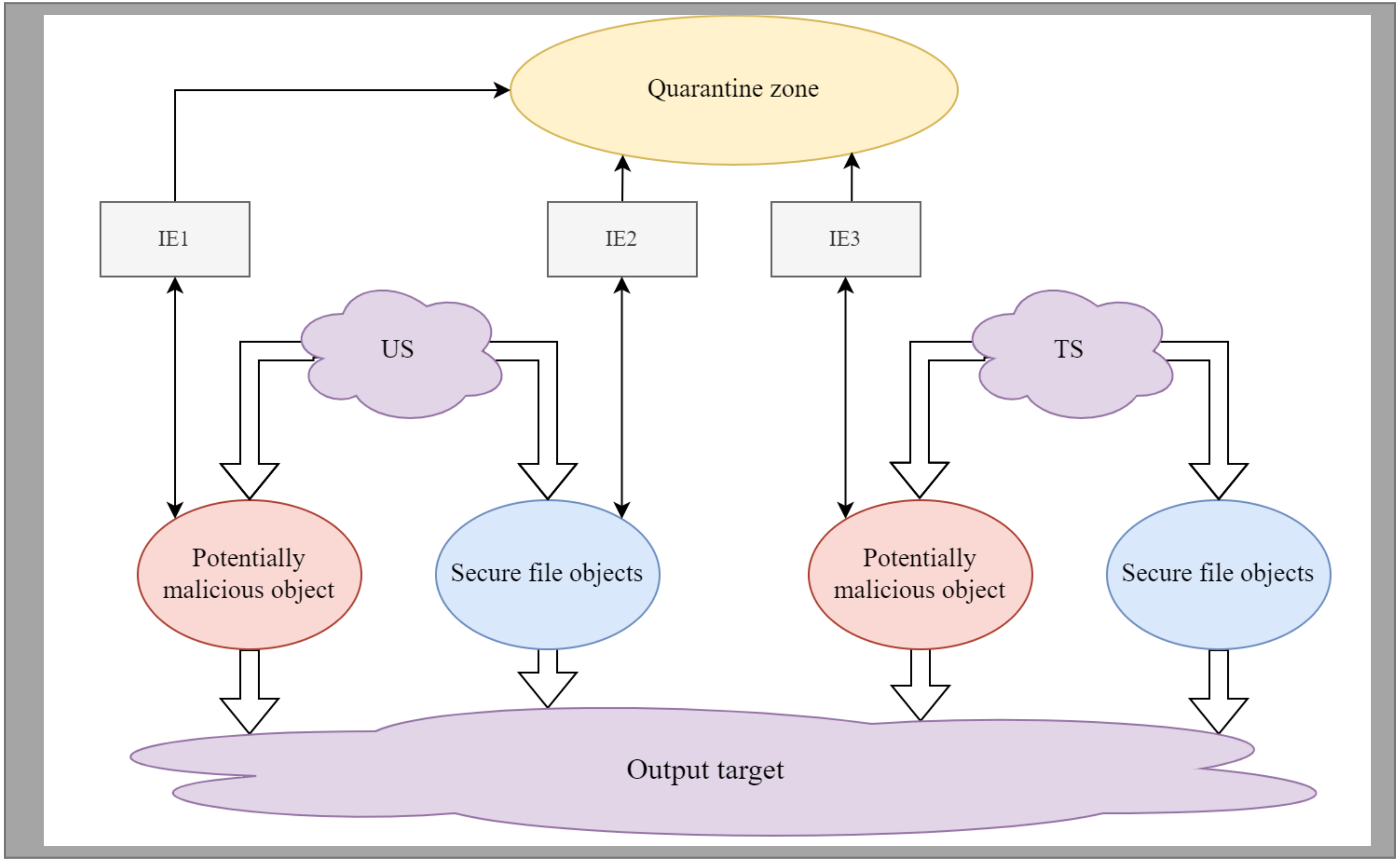
A DLAF framework consisting of nine aspects of data lakes to consider when creating a comprehensive data lake architecture

SDLAF Data Infrastructure Aspect



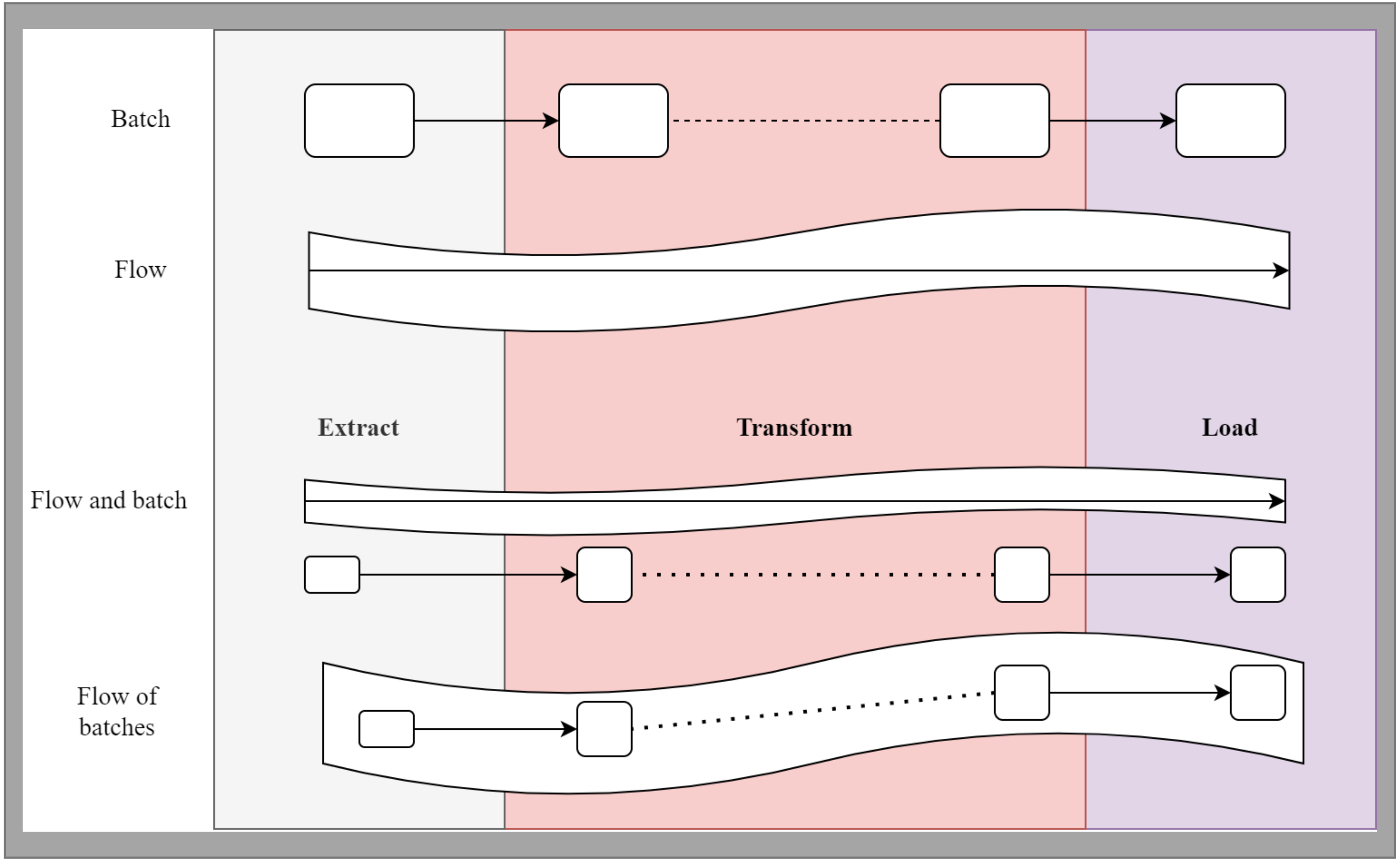
Secure data lake infrastructure

SDLAF Data Storage Aspect



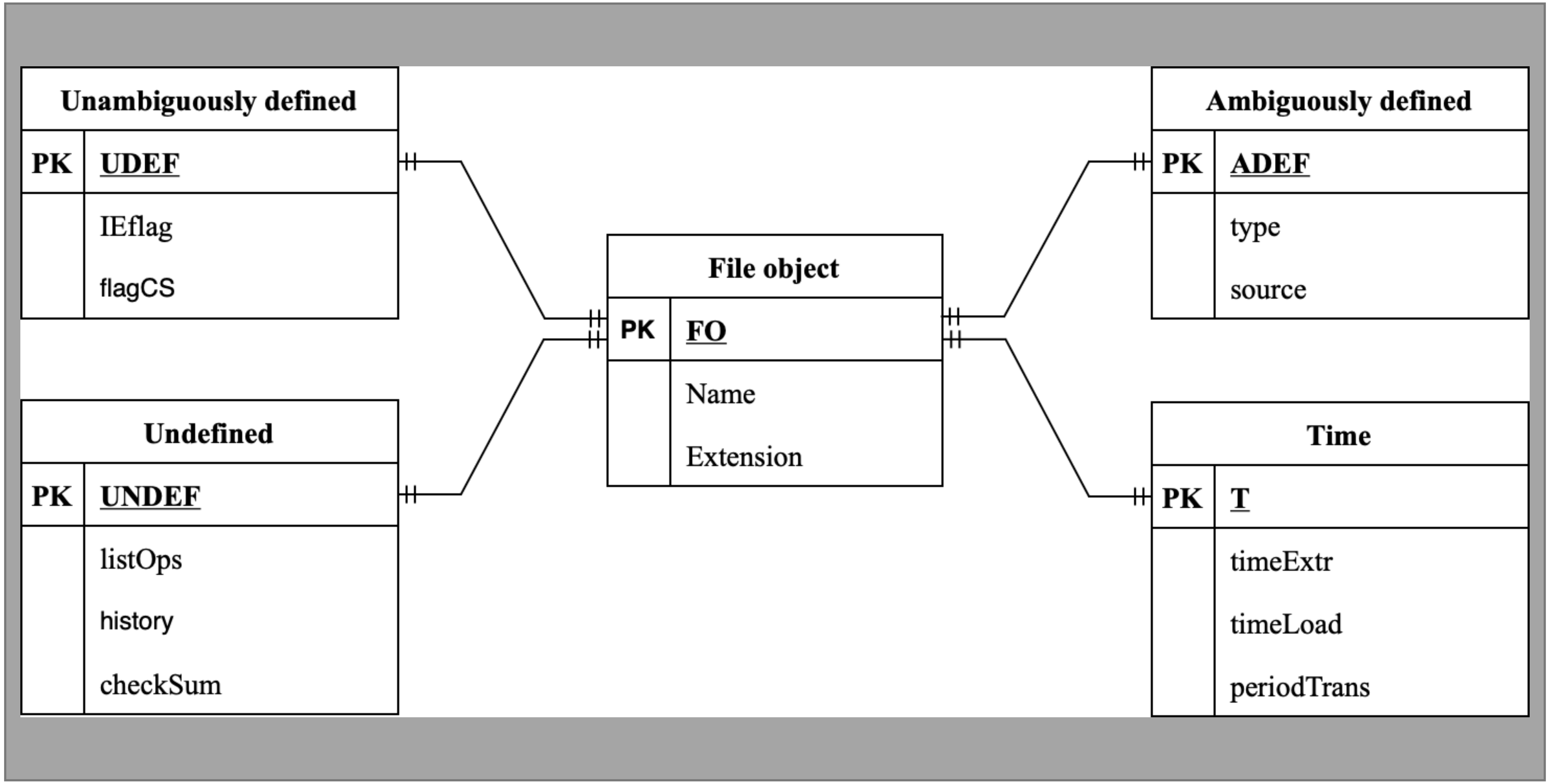
Secure data lake repository

SDLAF Data Flow Aspect



Secure data lake flow

SDLAF Data Modelling Aspect



Unambiguously defined

PK	**<u>UDEF</u>**
	IEflag flagCS

Undefined

PK	**<u>UNDEF</u>**
	listOps history checkSum

File object

PK	**<u>FO</u>**
	Name Extension

Ambiguously defined

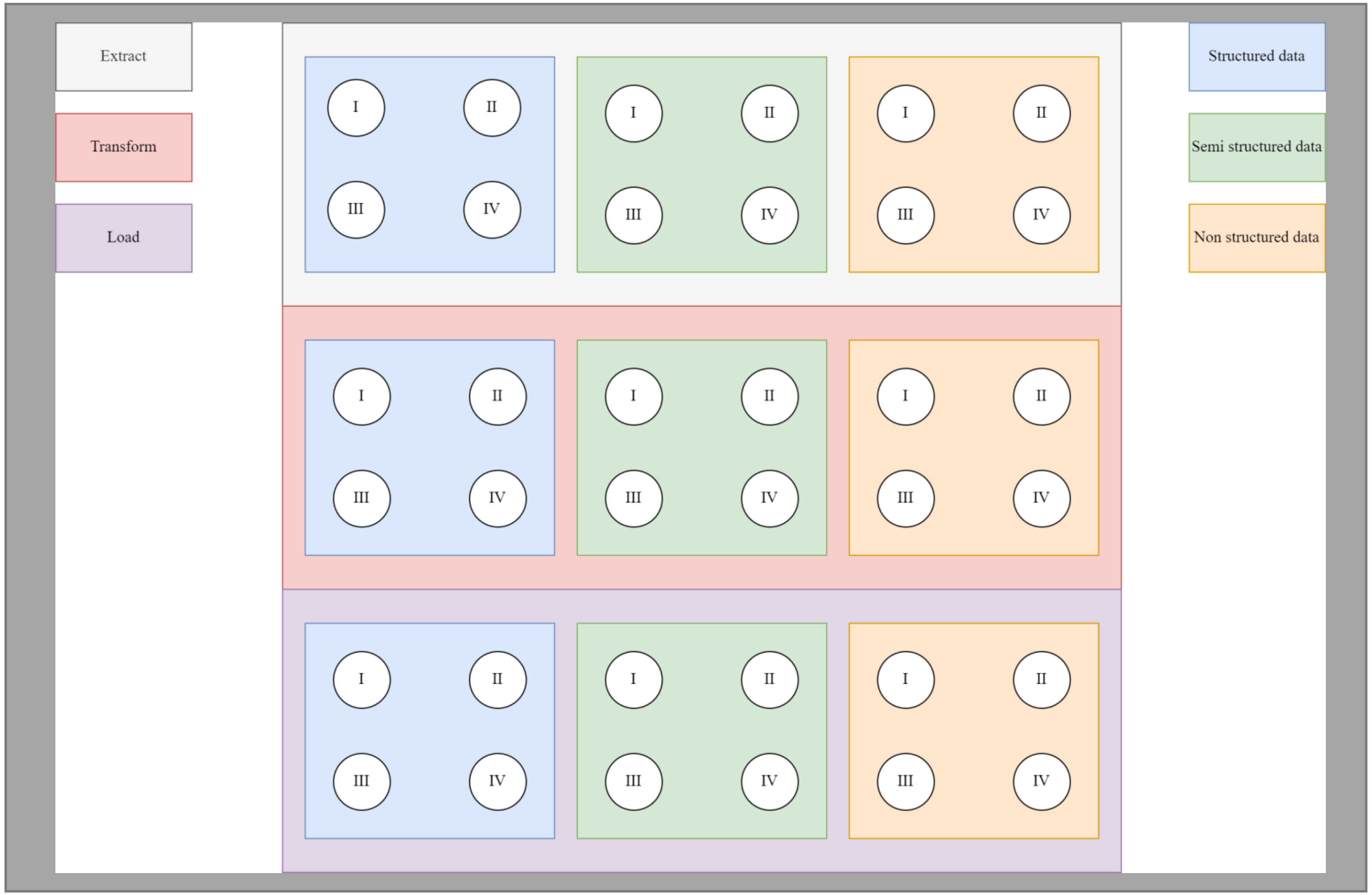
PK	**<u>ADEF</u>**
	type source

Time

PK	**<u>T</u>**
	timeExtr timeLoad periodTrans

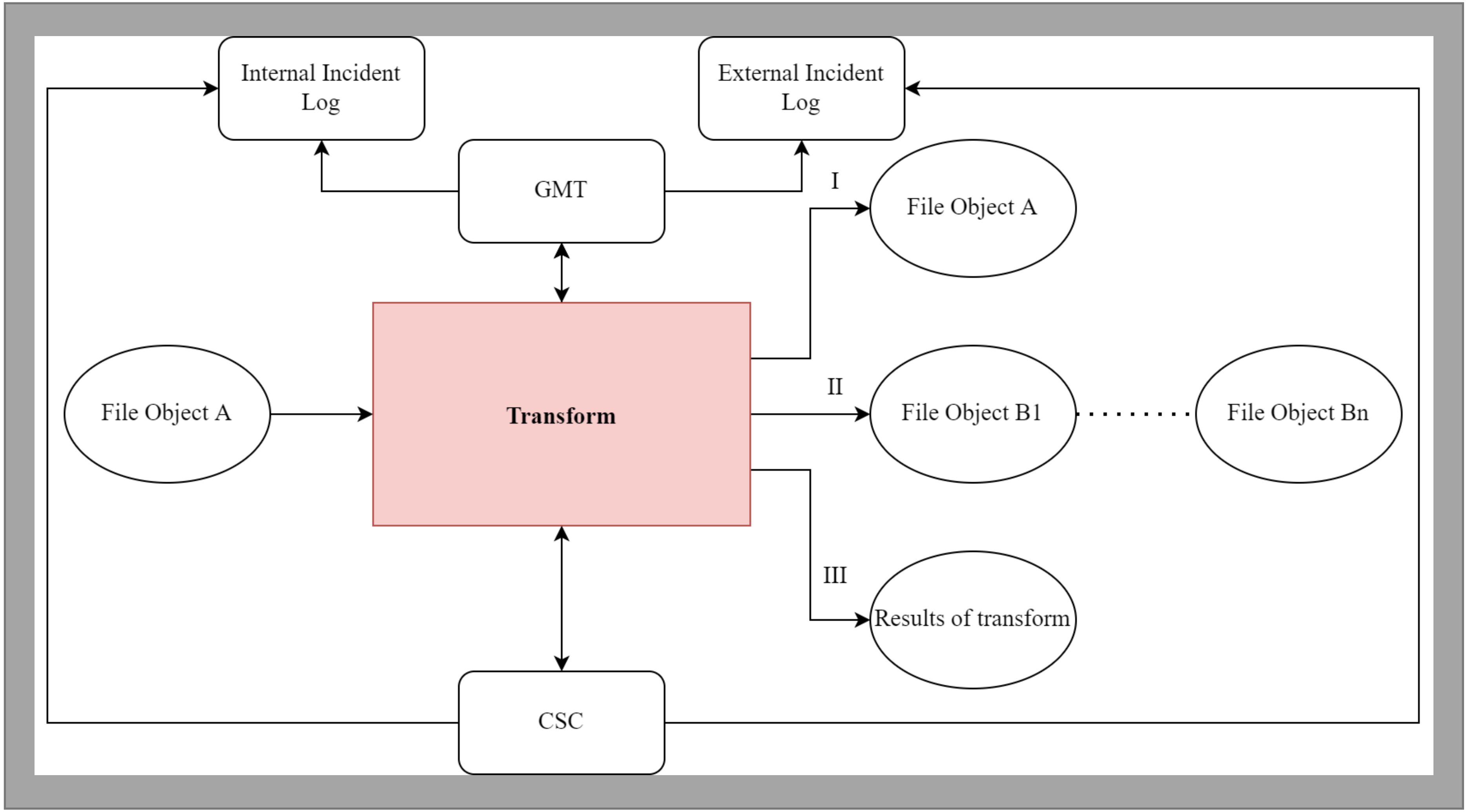
Secure data lake model

SDLAF Data Organization Aspect



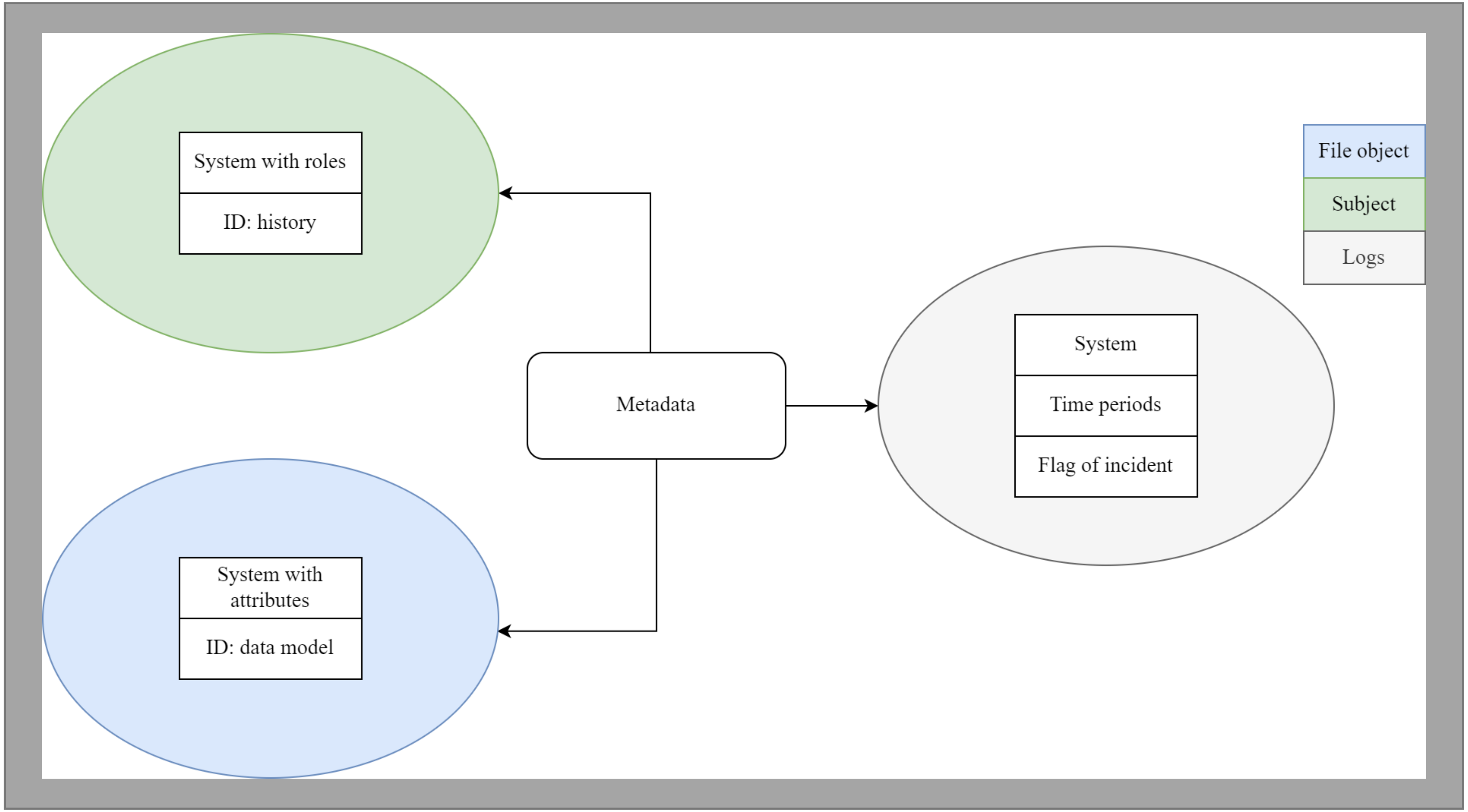
Secure data lake organization

SDLAF Data Processing Aspect



Secure data lake processing

SDLAF Metadata Management Aspect



Secure data lake metadata management

SDLAF Security Models

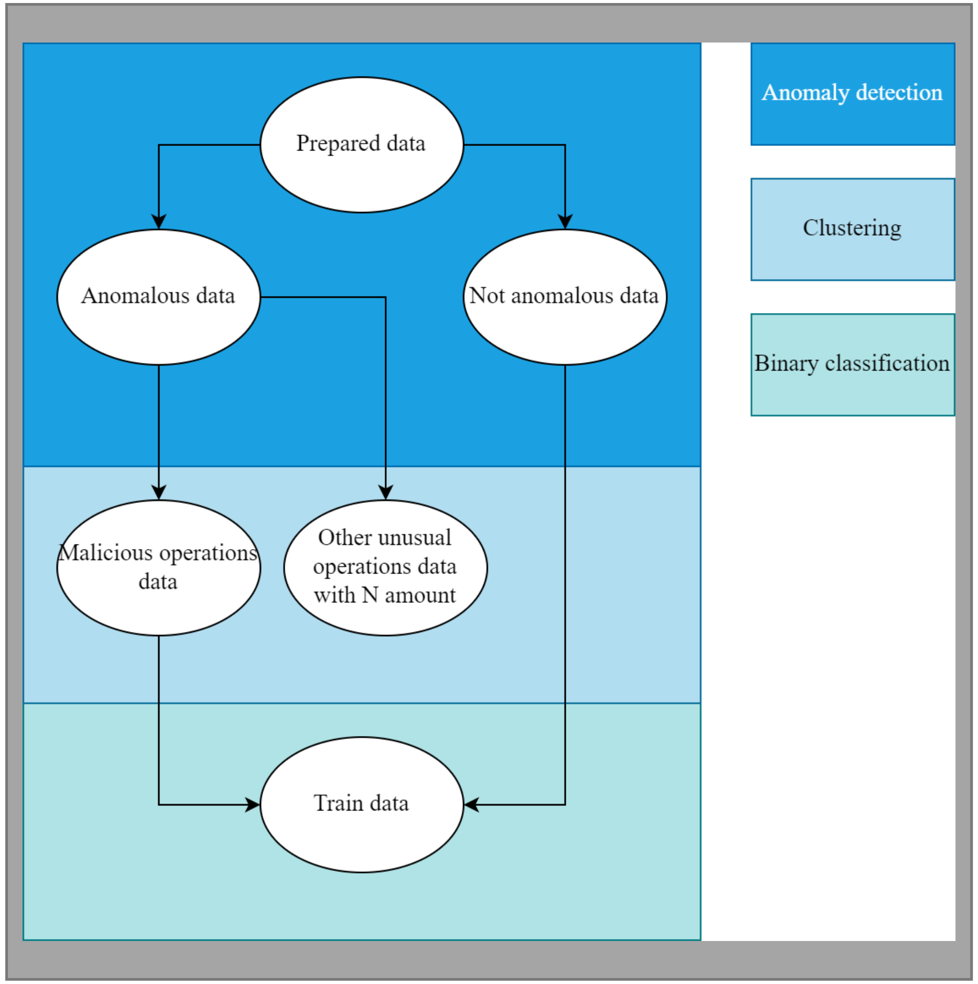
There are **four** security models to choose from:

- **DRBAC** - an implemented modified security model RBAC with an embedded Clark-Wilson discretionary integrity control model;
- **MRBAC** - an implemented modified RBAC security model with Ken Beeb's inbuilt mandate model of integrity control;
- **MABAC** - modified security model with ABAC;
- **XACML** - generally accepted standard for ABAC implemented in OPC without any modifications.

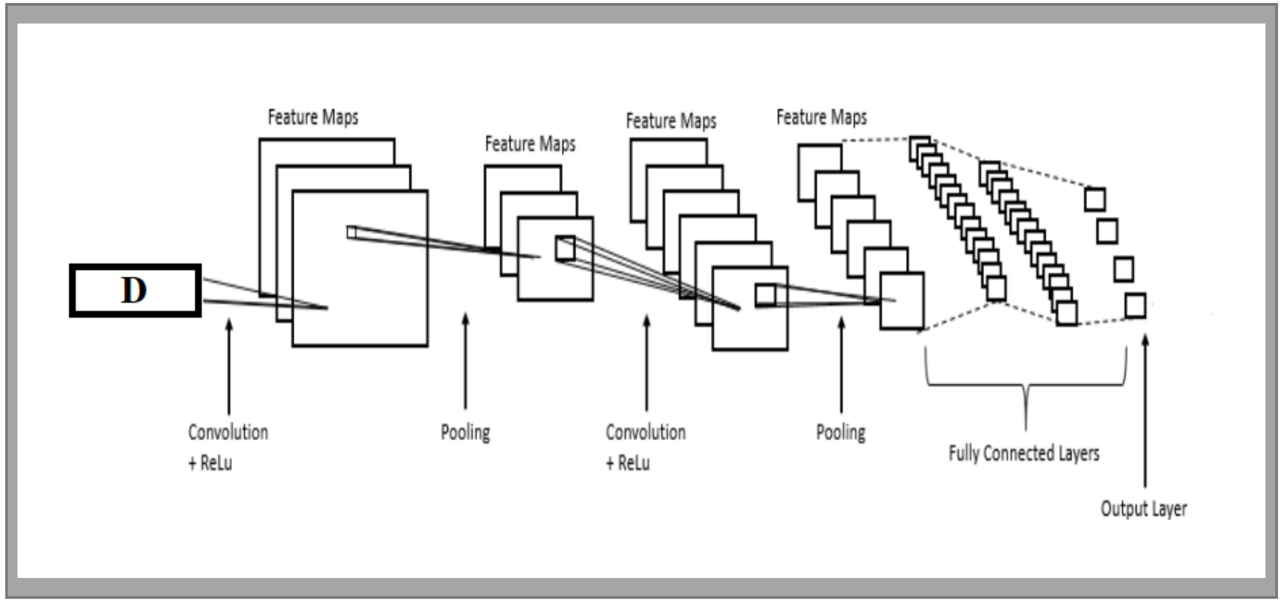
All the above MBs interact with the **two** main components of a secure OPC architecture:

- **GMT** - Global Monitoring Tool;
- **CSC** - Check Sum Controller - check sum handler.

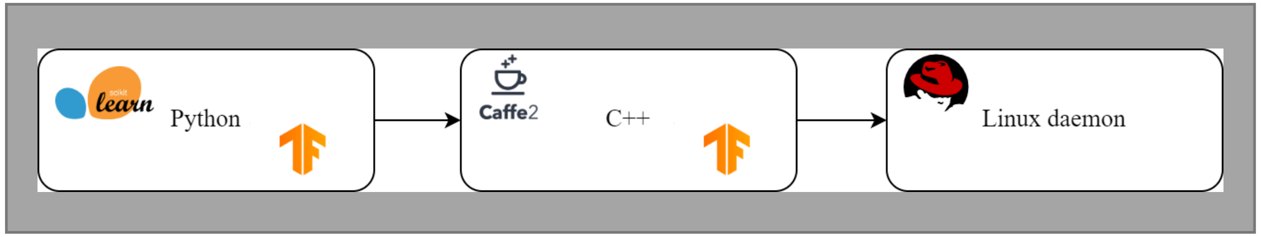
SDLAF Data Protection Aspect



The process of generating a dataset

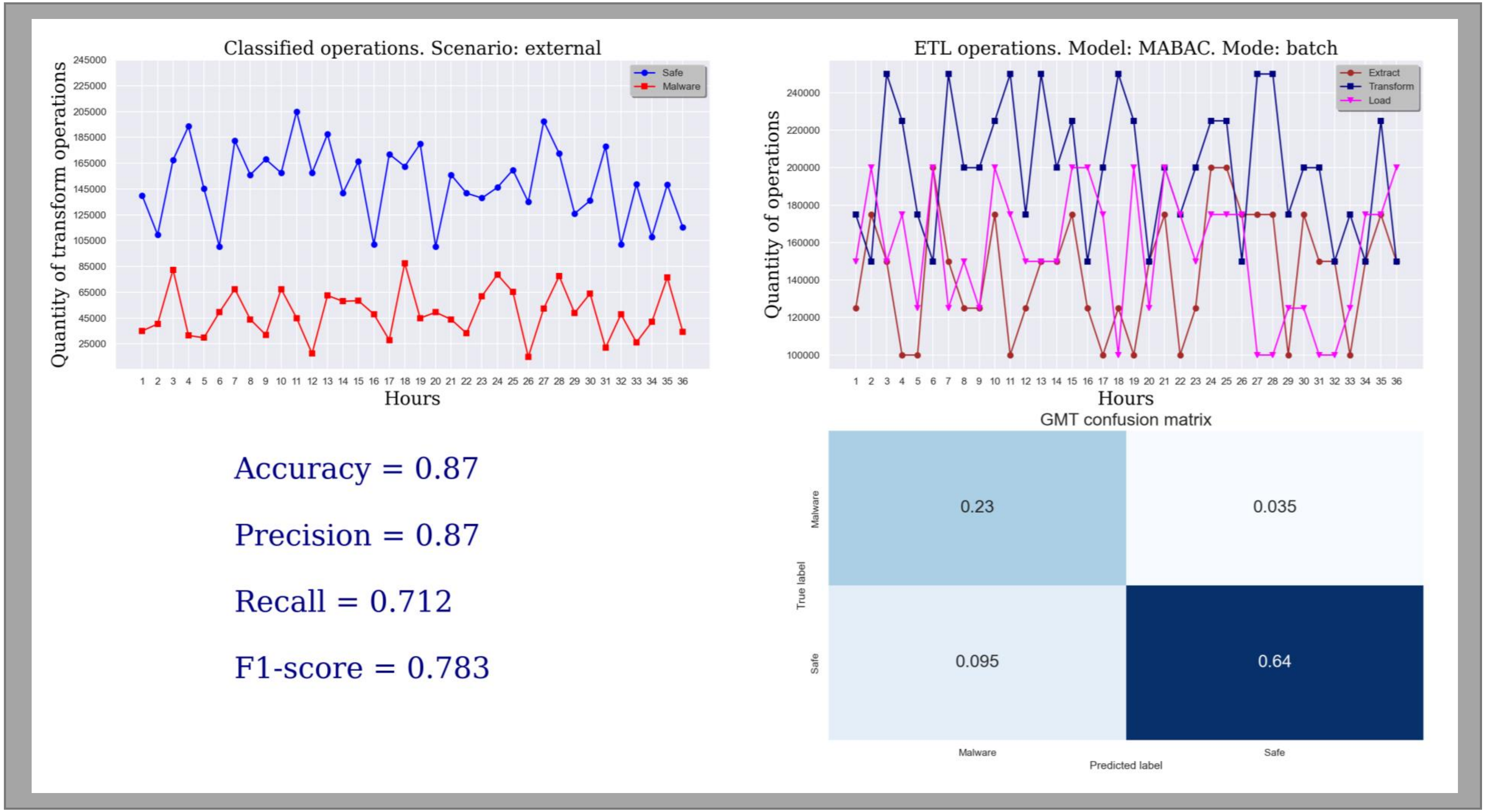


Architecture of CNN
classification of log code sequences



GMT model development process

Results



Accuracy = 0.87

Precision = 0.87

Recall = 0.712

F1-score = 0.783

GMT confusion matrix

	Malware	Safe
Malware	0.23	0.035
Safe	0.095	0.64

Dashboard of the MABAC model in batch mode and emulation of external attack scenarios

Conclusion

The developed Secured Data Lake Architecture Framework (SDLAF) provides the ability to design secure data lakes **without losing key benefits** and while **maintaining flexibility** for a wide range of business requirements.

The next stage of work – **GMT 2.0**:

- Data trait engineering of monitoring and journaling systems;
- Automatic anomaly detection;
- Adaptation to **any log sequence** regardless of system.

Thanks for your attention