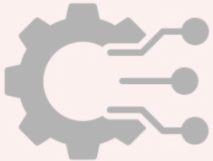
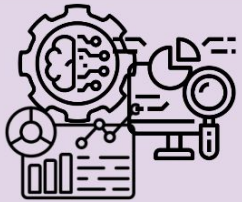
Big Data Management
Project P2

DTIM Research Group
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

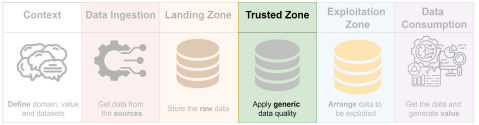
Contents

- Explaining the stages
  - Trusted Zone
  - Exploitation Zone
  - Downstream tasks
- Implementing the stages
  - Design
  - Databases
  - Transformations
- Data governance (additional criteria)
- Deliverables

Context	Data Ingestion	Landing Zone	Trusted Zone	Exploitation Zone	Data Consumption
					
Define domain, value and datasets	Get data from the **sources**	Store the **raw** data	Apply **generic** data quality	**Arrange** data to be exploited	Get the data and generate **value**

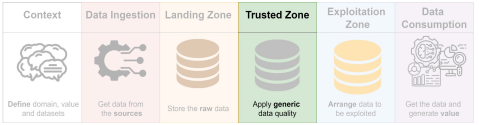
Stages Overview

Trusted Zone



- Raw data (stored in the landing zone) comes with all sorts of mistakes or inconsistencies that need to be address **before** it is used.

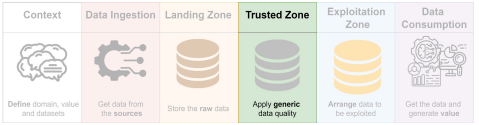
Trusted Zone

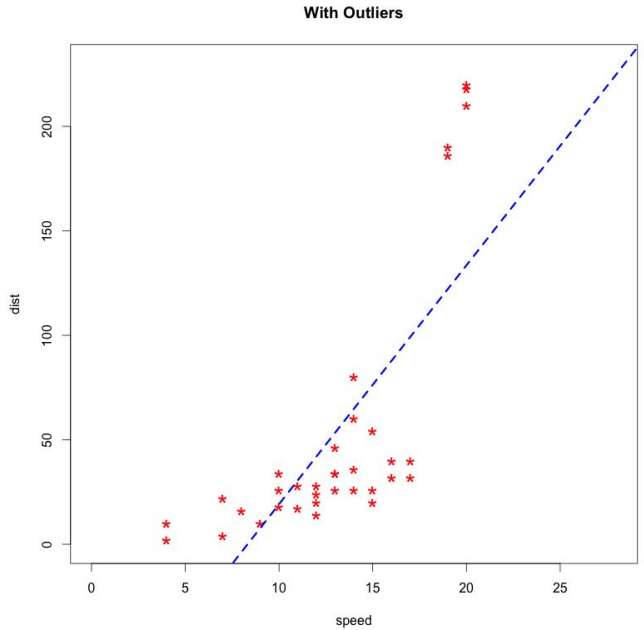


- Raw data (stored in the landing zone) comes with all sorts of mistakes or inconsistencies that need to be address **before** it is used.

- Our data might be used for a lot of tasks. Hence, we **can not** apply **specific** preprocessing techniques.

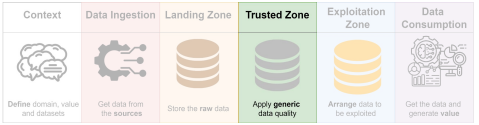
Trusted Zone

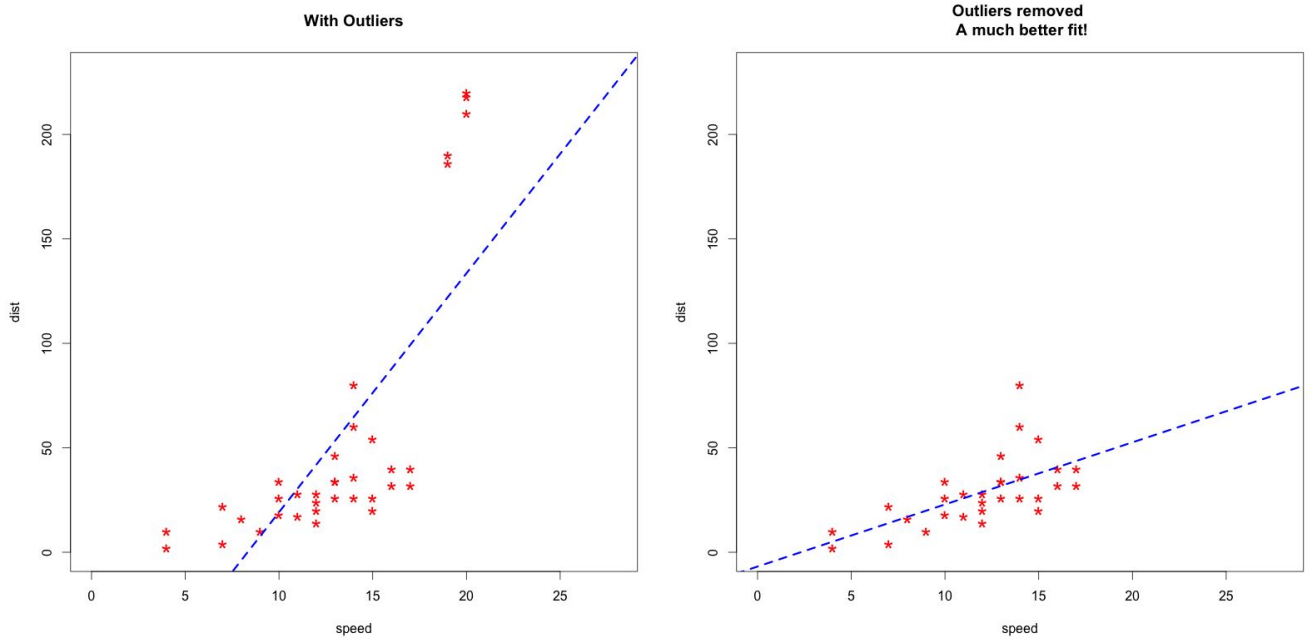




r-statistics: *Outlier Treatment With R*

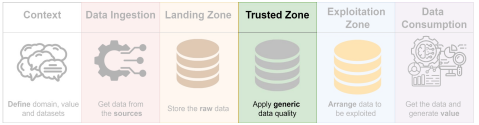
Trusted Zone





r-statistics: *Outlier Treatment With R*

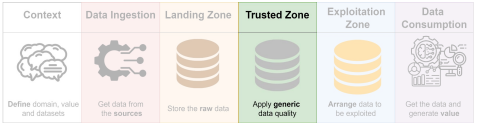
Trusted Zone



- Raw data (stored in the landing zone) comes with all sorts of mistakes or inconsistencies that need to be address **before** it is used.

- Our data might be used for a lot of tasks. Hence, we **can not** apply **specific** preprocessing techniques.
  - Be careful with: outlier removal, missing values imputation, normalization and some encodings

Trusted Zone

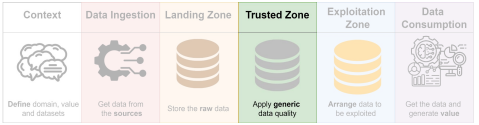


- Raw data (stored in the landing zone) comes with all sorts of mistakes or inconsistencies that need to be address **before** it is used.

- Our data might be used for a lot of tasks. Hence, we **can not** apply **specific** preprocessing techniques.
  - Be careful with: outlier removal, missing values imputation, normalization and some encodings

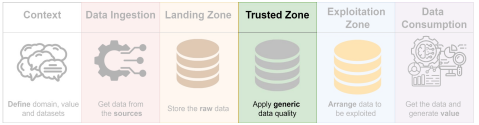
- What we have to apply are **generic** data quality tasks

Trusted Zone



- Examples of tasks (more are listed in the document):
  - **Deduplication**: removing repeated elements (e.g. based on primary key)
  - **Constraint validation, based on business rules** (e.g. non-negative ages or
  - **Text normalization** (e.g. lowercasing, removing punctuation).
  - **Check file integrity and remove/fix corrupted files** (mainly for unstructured data).

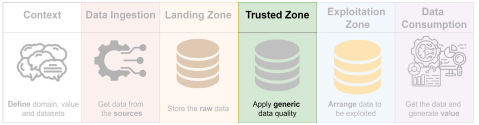
Trusted Zone



- Examples of tasks (more are listed in the document):
  - **Deduplication**: removing repeated elements (e.g. based on primary key)
  - **Constraint validation, based on business rules** (e.g. non-negative ages or
  - **Text normalization** (e.g. lowercasing, removing punctuation).
  - **Check file integrity and remove/fix corrupted files** (mainly for unstructured data).

- Our goal is to **"fix mistakes".**

Trusted Zone

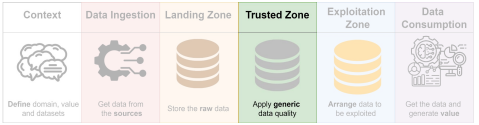


- Examples of tasks (more are listed in the document):
  - **Deduplication**: removing repeated elements (e.g. based on primary key)
  - **Constraint validation, based on business rules** (e.g. non-negative ages or
  - **Text normalization** (e.g. lowercasing, removing punctuation).
  - **Check file integrity and remove/fix corrupted files** (mainly for unstructured data).

- Our goal is to **"fix mistakes".**

- Note that you can implement any task you want if it is adequately justified.

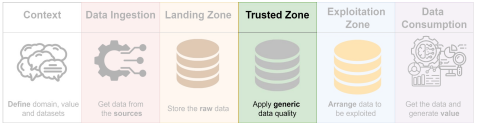
Trusted Zone - Tasks



You need to:

- Select the required data storage tools (i.e. databases, see section 2).
- Define where each of your data assets will be stored, and which transformations will be applied.
- Implement the pipelines that move the data from the landing to the exploitation zone. Do so with appropriate technology (see section 2).

Trusted Zone - Tasks



You need to:

- Select the required data storage tools (i.e. databases, see section 2).
- Define where each of your data assets will be stored, and which transformations will be applied.
- Implement the pipelines that move the data from the landing to the exploitation zone. Do so with appropriate technology (see section 2).

Factors that will **positively contribute** towards the grade:

- Thoroughness and correctness of the cleaning tasks

Exploitation Zone

- Data in the trusted zone is clean, but also **poorly organized**. The goal of the exploitation zone is to **provide data in the best possible way to facilitate analyses**.

Exploitation Zone

- Data in the trusted zone is clean, but also **poorly organized**. The goal of the exploitation zone is to **provide data in the best possible way to facilitate analyses**.

- Once again, we have to accommodate for a plethora of potential downstream tasks.
  - We aim to create **subject/domain-oriented structures** that can serve multiple purposes

Exploitation Zone

- Data in the trusted zone is clean, but also **poorly organized**. The goal of the exploitation zone is to **provide data in the best possible way to facilitate analyses**.

- Once again, we have to accommodate for a plethora of potential downstream tasks.
  - We aim to create **subject/domain-oriented structures** that can serve multiple purposes

- We are going to implement **two** separate processes to do so.

Exploitation Zone - Organizing

- For unstructured data this mostly consists on storing it in separated buckets defined based on the domain or some characteristic.

Exploitation Zone - Organizing

- For unstructured data this mostly consists on storing it in separated buckets defined based on the domain or some characteristic.

- For structured data this can mean:

Exploitation Zone - Organizing

- For unstructured data this mostly consists on storing it in separated buckets defined based on the domain or some characteristic.

- For structured data this can mean:
  - Table join

Country	GDP	HDI

Country	Military	Land

Country	GDP	HDI	Military	Land

Exploitation Zone - Organizing

- For unstructured data this mostly consists on storing it in separated buckets defined based on the domain or some characteristic.

- For structured data this can mean:
  - Table join
  - Vertical partition

Movie	Actor	A.Age	A.Birth	A...
Mx	a1	45	Madrid	...
My	a1	45	Madrid	...

Movie
Mx
My

Movie	Actor
Mx	a1
My	a1

Actor	A.Age	A.Birth	A...
a1	45	Madrid	...

Exploitation Zone - Organizing

- For unstructured data this mostly consists on storing it in separated buckets defined based on the domain or some characteristic.

- For structured data this can mean:
  - Table join
  - Vertical partition
  - Table union

A	B	C	D
1	...	...	...
2	...	...	...

A	B	C	D
3	...	...	...
4	...	...	...

A	B	C	D
1	...	...	...
2	...	...	...
3	...	...	...
4	...	...	...

Exploitation Zone - Organizing

- For unstructured data this mostly consists on storing it in separated buckets defined based on the domain or some characteristic.

- For structured data this can mean:
  - Table join
  - Vertical partition
  - Table union
  - Horizontal partition

A	B	C	D
1	B1	...	...
2	B1	...	...
3	B2	...	...
4	B2	...	...

A	B	C	D
1	B1	...	...
2	B1	...	...

A	B	C	D
3	B2	...	...
4	B2	...	...

Exploitation Zone - Organizing

- For unstructured data this mostly consists on storing it in separated buckets defined based on the domain or some characteristic.

- For structured data this can mean:
  - Table join
  - Vertical partition
  - Table union
  - Horizontal partition

- For semi-structured data this processing can be very complex

Exploitation Zone - New Data

- For unstructured data, this consists on generating **new information or artifacts** that enrich the data.
  - Summaries of textual data.
  - Classify into categories
  - Embeddings

Exploitation Zone - New Data

- For unstructured data, this consists on generating **new information or artifacts** that enrich the data.
  - Summaries of textual data.
  - Classify into categories.
  - Embeddings.

- For structured and semi-structured data, **KPI**s or other business-related **aggregates** can be obtained.
  - Total benefit per month.
  - Percentage of returning customers.
  - Average time from job posting to hire.

Exploitation Zone - New Data

You need to:

- Define the structure of the exploitation zone.
- Decide which additional data assets you are going to generate and how.
- Select the required data storage tools (i.e. databases, see section 2) and set them up
- Implement the pipelines that move the data from the trusted to the exploitation zone. Do so with appropriate technology (see section 2).

Factors that will **positively contribute** towards the grade:

- Effort in providing a coherent and nuanced organization to the data.
- Appropriate development of new data/artifacts.

Exploitation Zone - New Data

Important note: We acknowledge that the exploitation zone can be a bit confusing, and the data that you have might severely limit what you can do.

The goal of implementing this zone is, simply, that you play around with the problem of organizing the data in interesting ways to maximize its utility.

Data Exploitation

- At this point most of the data management aspects have been fulfilled, and **data is ready to be employed** in subsequent tasks or fed to external systems.

Data Exploitation

- At this point most of the data management aspects have been fulfilled, and **data is ready to be employed** in subsequent tasks or fed to external systems.

- **Priorities** are:
  - Think about how to exploit your data.
  - Think about which tools to do so (and set them up).
  - Establish the necessary connections to ship the data.

Data Exploitation

- At this point most of the data management aspects have been fulfilled, and **data is ready to be employed** in subsequent tasks or fed to external systems.

- **Priorities** are:
  - Think about how to exploit your data.
  - Think about which tools to do so (and set them up).
  - Establish the necessary connections to ship the data.

- Recall that BDM is not a modeling/analysis course.

Data Exploitation

- At this point most of the data management aspects have been fulfilled, and **data is ready to be employed** in subsequent tasks or fed to external systems.

- **Priorities** are:
  - Think about how to exploit your data.
  - Think about which tools to do so (and set them up).
  - Establish the necessary connections to ship the data.

- Recall that BDM is not a modeling/analysis course.

- You can use **placeholders** and **synthetic data**.

Data Exploitation- Tasks

You need to:

- Define which **mechanisms** you are going to use to exploit the data.
- Select the **tools** you will use to implement such mechanisms.
- **Ship** the data from the exploitation zone onto selected systems.
- **Implement** the designed tasks.

Data Exploitation- Tasks

You **need to**:

- Define which **mechanisms** you are going to use to exploit the data.
- Select the **tools** you will use to implement such mechanisms.
- **Ship** the data from the exploitation zone onto selected systems.
- **Implement** the designed tasks.

Factors that will **positively contribute** towards the grade:

- Quality and complexity of the implemented processes.

Implementing the stages

Design

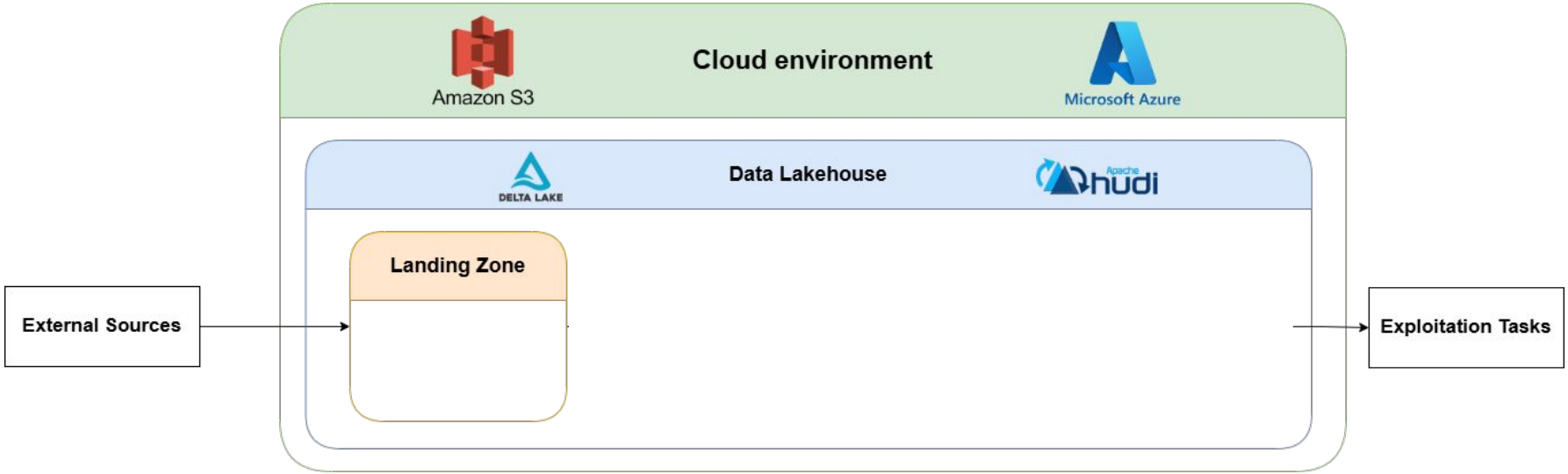
- The most important part is that **each zone has to store the data** after the respective transformations have been applied.

Design

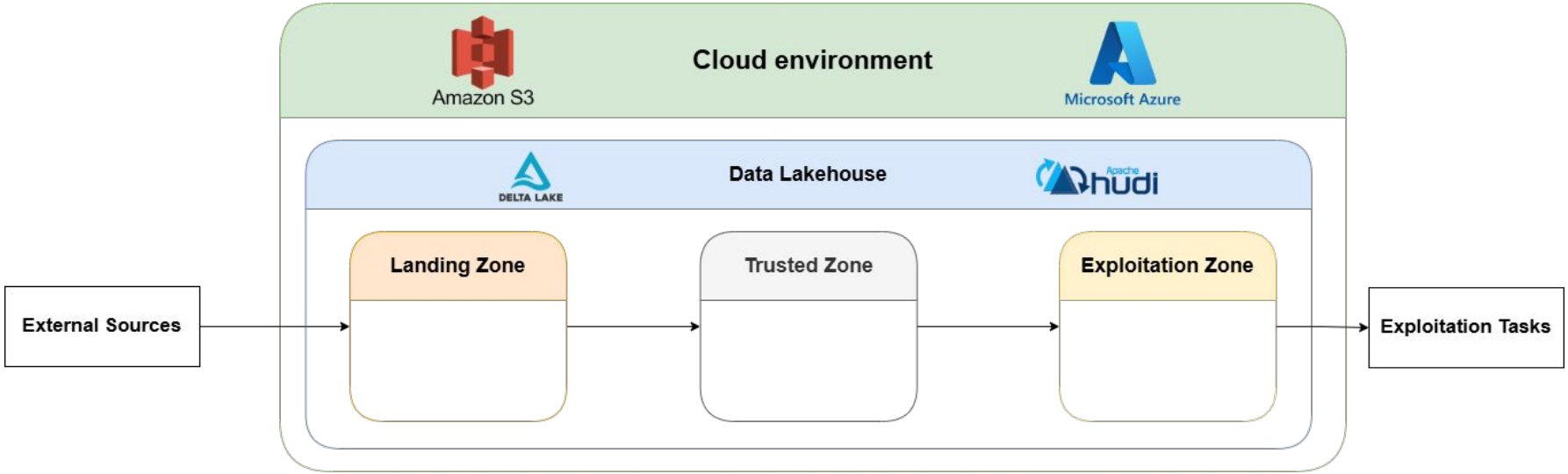
- The most important part is that **each zone has to store the data** after the respective transformations have been applied.

- We propose two main design alternatives

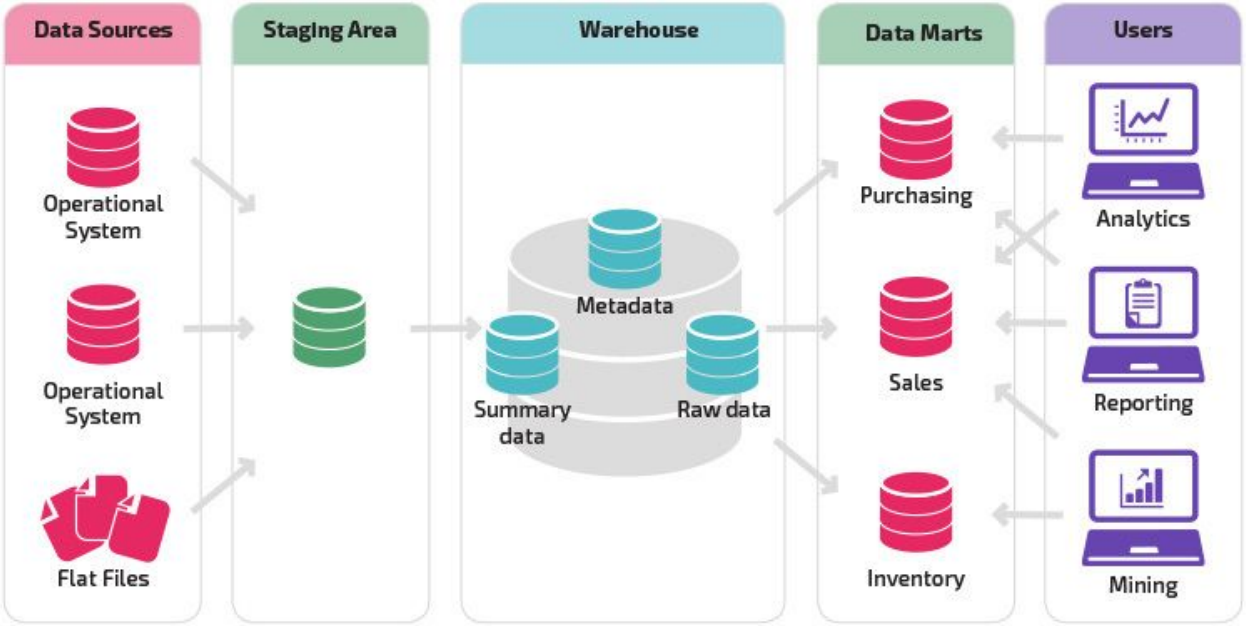
Design - Option 1: Extending your Implementation



Design - Option 1: Extending your Implementation

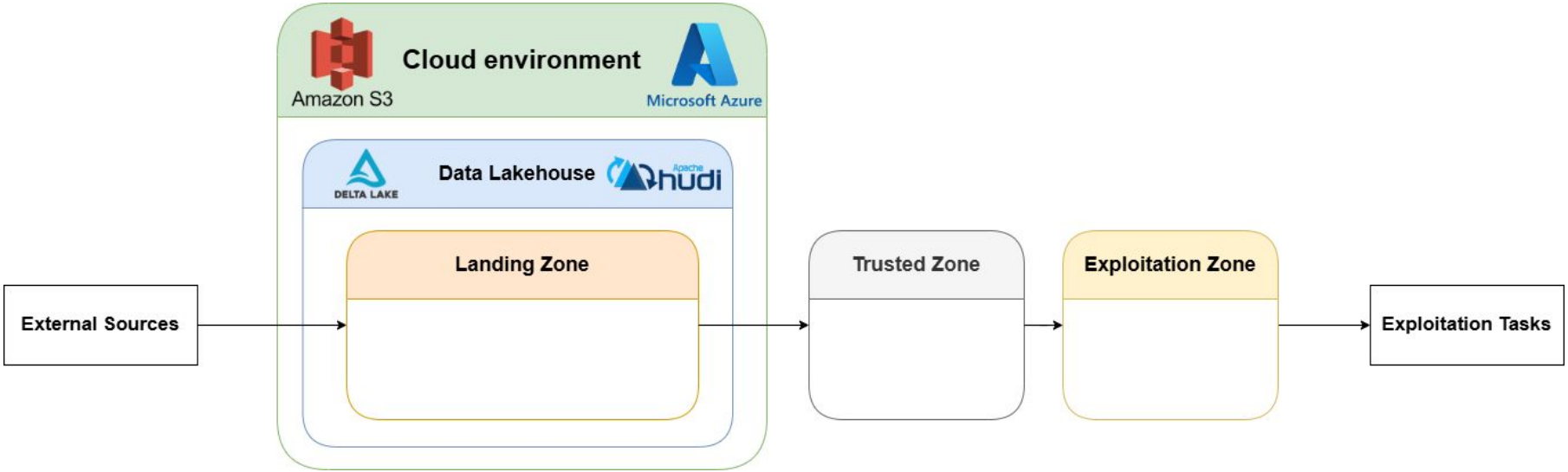


Design - Option 2: Moving the Data Out

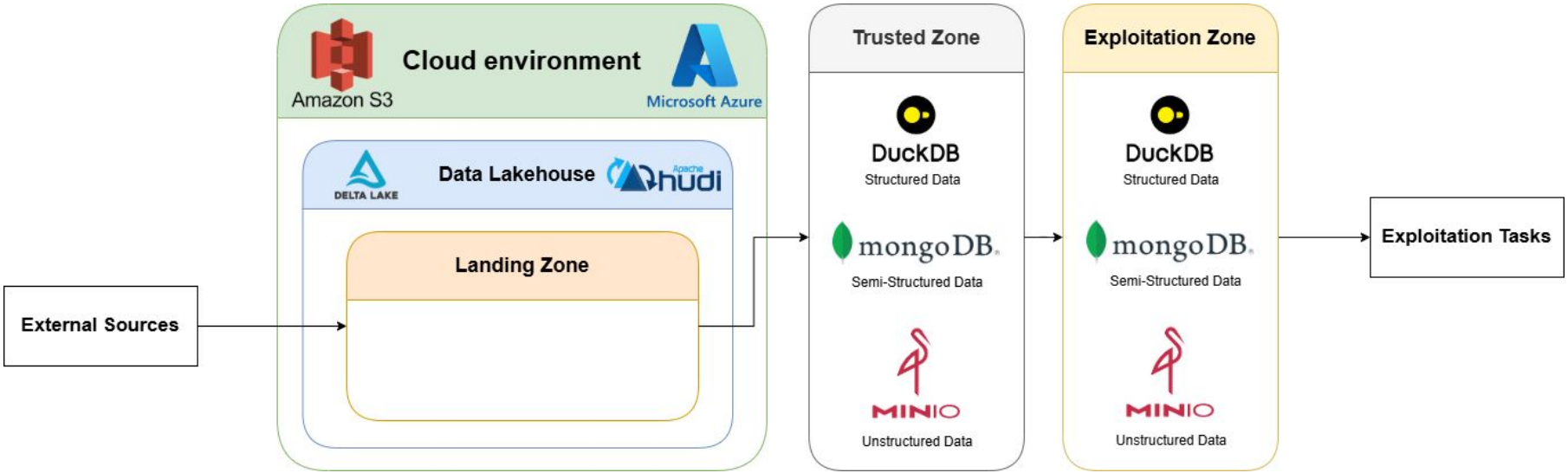


Panoply: *Data Mart vs. Data Warehouse*

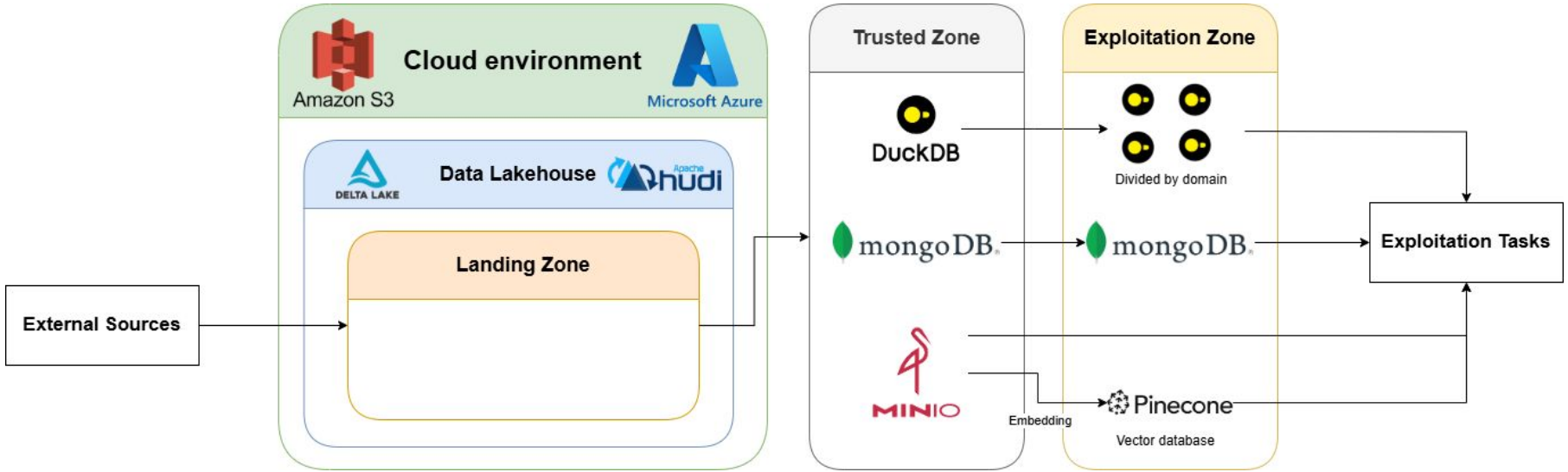
Design - Option 2: Moving the Data Out



Design - Option 2: Moving the Data Out



Design - Option 2: Moving the Data Out



Design - Tasks

You need to:

- Complete the architecture diagram (new zones, tools, operations, etc.)
- Be sure to adequately reason and justify the inclusion of each element.

Design - Tasks

You **need to**:

- Complete the architecture diagram (new zones, tools, operations, etc.)
- Be sure to adequately reason and justify the inclusion of each element.

Important: From an evaluation standpoint, both are perfectly acceptable. Nonetheless, a grade boost will be given to implementations that combine several technologies or define more interesting data management mechanisms.

Databases

Databases

- Relational Databases
  - Transactional (“traditional”): 
  - Analytical (“modern”): 

Databases

- Relational Databases
  - Transactional (“traditional”):   
  - Analytical (“modern”): 
- Key-Value Stores 

Databases

- Relational Databases
  - Transactional (“traditional”): 
  - Analytical (“modern”): 
- Key-Value Stores 
- Document Stores 

Databases

- Relational Databases
  - Transactional (“traditional”):   
  - Analytical (“modern”): 
- Key-Value Stores 
- Document Stores  
- Wide-Column  

Databases

- Graph Databases 

Databases

- Graph Databases 

- Time-Series Databases 

Databases

- Graph Databases 

- Time-Series Databases 

- Vector Databases 

Databases

- Graph Databases 

- Time-Series Databases 

- Vector Databases 

- Object Stores 

Transformations - Batch

Transformations - Batch

- Large-Scale Processing

Transformations - Batch

- Large-Scale Processing

- Small-Scale Processing

Transformations - Batch

- Large-Scale Processing

- Small-Scale Processing

- You need to **justify** what type of technology is appropriate to use depending on the **volume of data** you are working with.

Transformations - Streaming

- Should your process the data you stream?
  - *Conceptually*, not in a hot path, as data has to be sent immediately.
  - *Conceptually*, yes in a warm path, as we can tolerate a higher latency.

Transformations - Streaming

- Should your process the data you stream?
  - *Conceptually*, not in a hot path, as data has to be sent immediately.
  - *Conceptually*, yes in a warm path, as we can tolerate a higher latency.

- Tools to process a stream:

(In Java)

Data Governance (Optional)

Data Governance

- Data governance is a **high-level structured framework that formalizes the management** of data through policies, standards, processes, and ongoing monitoring.

Data Governance

- Data governance is a **high-level structured framework that formalizes the management** of data through policies, standards, processes, and ongoing monitoring.

- When designing data governance, it is important to understand how different scopes (intra/inter-organizational, quality, security, etc.) tie together to **maximize data value while minimizing risks and costs**.

Data Governance

- Data governance is a **high-level structured framework that formalizes the management** of data through policies, standards, processes, and ongoing monitoring.

- When designing data governance, it is important to understand how different scopes (intra/inter-organizational, quality, security, etc.) tie together to **maximize data value while minimizing risks and costs**.

- A good starting point is to reflect on your business domain and **define what you aim to achieve**.
  - For example: is securing sensitive customer data a concern? Or is it more important to ensure trust in ML outputs through robust data quality controls?

Data Governance - Data Products

- A common governance strategy in modern big data architectures is to view data not simply as individual datasets, but as **data products** grouped within data domains.

Data Governance - Data Products

- A common governance strategy in modern big data architectures is to view data not simply as individual datasets, but as **data products** grouped within data domains.

- A data product is a well-defined, reusable dataset or API that is treated with the **same care and discipline** as a software product.

Data Governance – Data Products

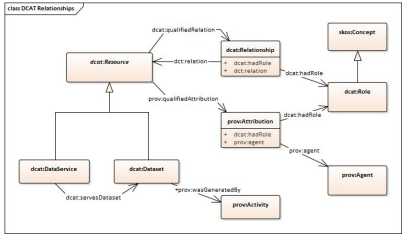
- A common governance strategy in modern big data architectures is to view data not simply as individual datasets, but as **data products** grouped within data domains.

- A data product is a well-defined, reusable dataset or API that is treated with the **same care and discipline** as a software product.

- Rather than governing every table or pipeline independently, organizations can attach policies, ownership, quality checks, and access control rules at the **level of the data product**.

Data Governance - Tasks

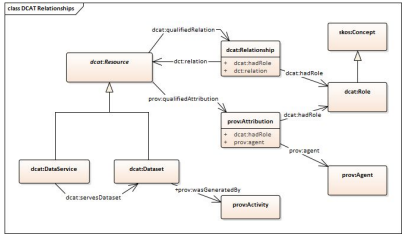
- Metadata Management



Apache Atlas

Data Governance - Tasks

- Metadata Management



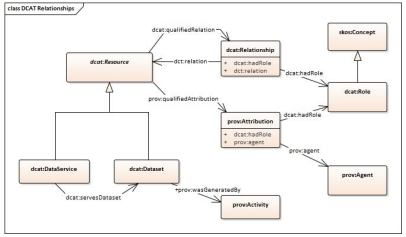
Apache Atlas

- Data Quality

great expectations

Data Governance - Tasks

- Metadata Management



Apache Atlas

- Data Quality

great expectations

- Data Security

Apache Ranger

Data Governance - Tasks

Over your resulting data architecture, begin to **design your data governance strategy**. As previously discussed, data governance includes organizational, strategic, and business-aligned scopes.

1. Begin by reflecting on the data domains that are relevant to your business. Based on these domains, **propose a set of data products** that should exist in your exploitation zone.

Data Governance - Tasks

Over your resulting data architecture, begin to **design your data governance strategy**. As previously discussed, data governance includes organizational, strategic, and business-aligned scopes.

1. Begin by reflecting on the data domains that are relevant to your business. Based on these domains, **propose a set of data products** that should exist in your exploitation zone.
2. Provide at least **one detailed example** of a data product and the metadata you would associate with it to support governance.

Data Governance - Tasks

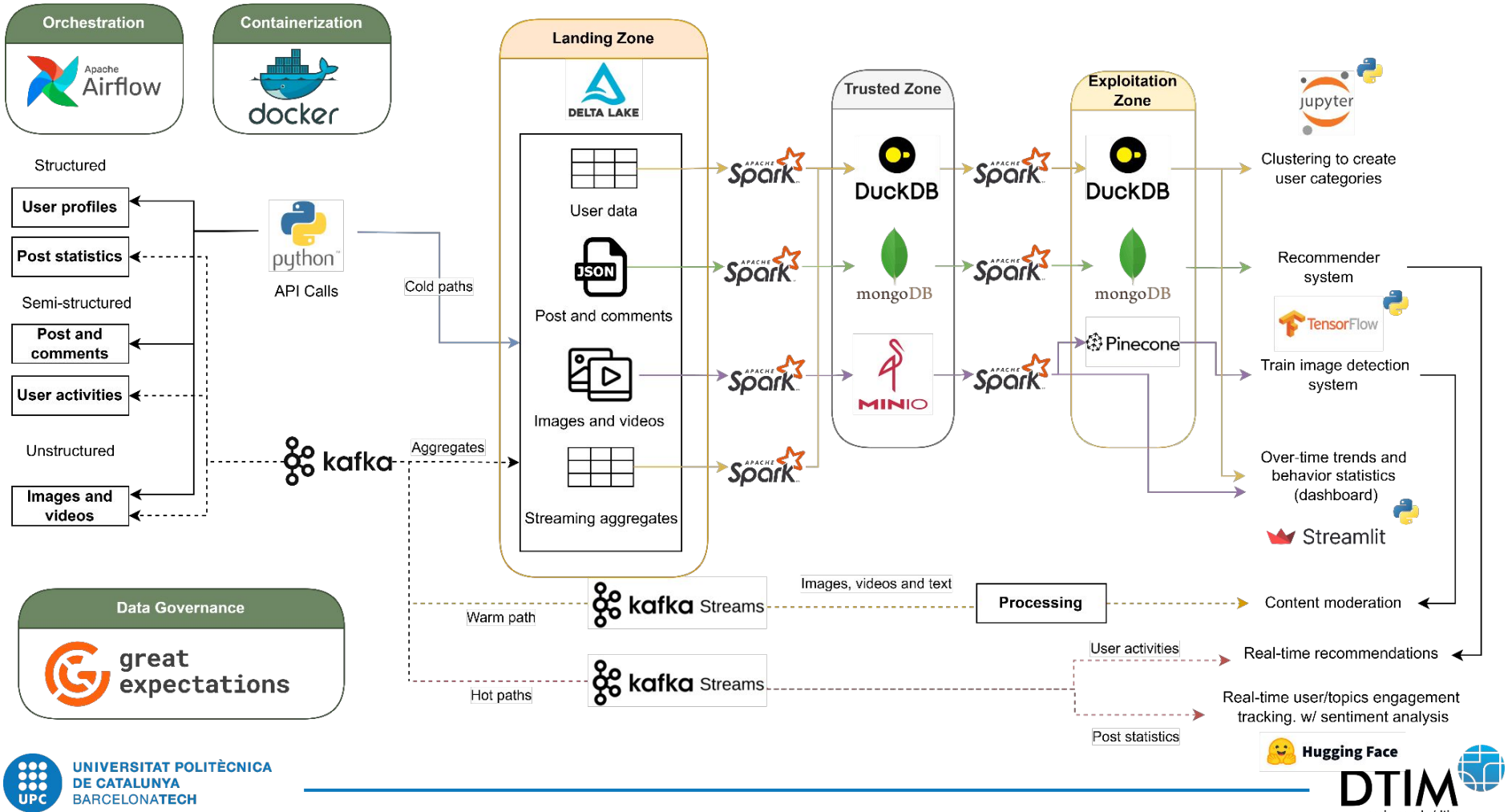
Over your resulting data architecture, begin to **design your data governance strategy**. As previously discussed, data governance includes organizational, strategic, and business-aligned scopes.

1. Begin by reflecting on the data domains that are relevant to your business. Based on these domains, **propose a set of data products** that should exist in your exploitation zone.
2. Provide at least **one detailed example** of a data product and the metadata you would associate with it to support governance.
3. Choose at least one **governance mechanism** (e.g., data quality validation, access control, lineage tracking, cataloging) and implement it over your architecture.

Deliverables

Main deliverable

- It has to implement:
  - Architecture design (updated version with the new zones, tools, processes, etc.)
  - Stage 3: trusted zone
  - Stage 4: exploitation zone
  - Stage 5: consumption tasks
  - Governance tasks (if applicable)


Orchestration
Apache Airflow
Containerization
docker
Landing Zone
DELTA LAKE
User data
Post and comments
Images and videos
Streaming aggregates
Trusted Zone
DuckDB
mongoDB
MINIO
Exploitation Zone
DuckDB
mongoDB
Pinecone
Structured
User profiles
Post statistics
Semi-structured
Post and comments
User activities
Unstructured
Images and videos
python
API Calls
Cold paths
kafka
Aggregates
Spark
jupyter
Clustering to create user categories
Recommender system
TensorFlow
Train image detection system
Over-time trends and behavior statistics (dashboard)
Streamlit
Data Governance
great expectations
Warm path
kafka Streams
Images, videos and text
Processing
Content moderation
Hot paths
kafka Streams
User activities
Real-time recommendations
Real-time user/topics engagement tracking. w/ sentiment analysis
Post statistics
Hugging Face

Follow-up deliverables

- First deliverable:
  - Architecture design (first version).

- Second deliverable:
  - Architecture design (updated).
  - Basic implementation of trusted zone, exploitation zone and consumption tasks.

- Remember the containerization and orchestration aspects