**Extracted Data**

1. Raw Data

* Definition: Raw Data refers to the unprocessed, original data as it is directly extracted from the customer's databases or internal corporate systems. This data has not undergone any form of cleansing, transformation, or manipulation.
* Characteristics:
  + May contain errors, inconsistencies, duplicates, missing values, and outliers.
  + Retains the original formats, data types, and structures as stored in the source systems.
  + Includes all attributes and records without any filtration or modification.
* Role in ETL Process:
  + Serves as the initial input for the ETL service.
  + Forms the basis for all subsequent data processing steps such as cleansing, merging, and transformation.

1. Meta Data

* Definition: Meta Data is the information that describes and provides context about the Raw Data. It includes details about the structure, schema, origin, and other attributes that help in understanding and processing the data
* Characteristics:
  + Contains schema definitions like table names, column names, data types, and relationships.
  + Includes data provenance information such as source identifiers, extraction timestamps, and version numbers.
* Role in ETL Process:
  + Facilitates schema mapping and alignment during data merging and transformation.
  + Aids in detecting inconsistencies and anomalies in the data.
  + Essential for compliance and auditing purposes.

1. Access Data

* Definition: Access Data consists of information related to user permissions, roles, and access controls associated with the datasets. This data ensures that only authorized users can interact with sensitive data.
* Characteristics:
* Includes user credentials, roles, and access permissions.
* Maintains logs of data access requests and actions taken.
* Supports compliance with security policies and regulatory requirements.
* Role in ETL Process:
* Ensures that data handling and processing are performed according to the organization's security protocols.
* Helps in auditing and monitoring who accessed the data and what actions were performed.
* Critical for maintaining data confidentiality and integrity.

**Data During Processing**

1. Cleansed Data

* Definition: Cleansed Data is the Raw Data that has undergone data cleansing processes to correct errors and remove inconsistencies. This includes handling missing values, eliminating duplicates, and correcting data formats.
* Characteristics:
  + Free from obvious errors and inconsistencies.
  + Standardized data formats and units.
  + Improved data quality, making it more reliable for analysis.
* Role in ETL Process:
  + Enhances the accuracy and reliability of the data.
  + Prepares data for effective merging and transformation.

1. Merged Data

* Definition: Merged Data is the result of combining Cleansed data from multiple sources into a single, cohesive dataset. This process resolves discrepancies, aligns schemas, and ensures data consistency across different datasets.
* Characteristics:
  + Consolidated view of data from various sources.
  + Resolved conflicts such as differing data formats or overlapping records.
  + Unified schema and structure.
* Role in ETL Process:
  + Provides a comprehensive dataset for analysis.
  + Eliminates data silos by integrating disparate datasets.

1. Transformed Data

* Definition: Transformed Data is the Cleansed and Merged Data that has been converted to meet specific business logic and application requirements. This includes data normalization, enrichment, and format conversions.
* Characteristics:
  + Data structured according to the needs of the target application.
  + Enhanced with additional calculated fields or aggregated metrics.
  + Converted into standardized units or classifications.
* Role in ETL Process:
  + Ensures compatibility with machine learning algorithms and analytical models.
  + Applies business rules and logic necessary for accurate predictions.
  + Facilitates efficient data storage and retrieval.

1. Anonymized Data

* Definition: Anonymized Data is data that has been processed to remove or obscure personal identifiers and sensitive information, ensuring individual privacy and compliance with data protection regulations.
* Characteristics:
  + Personal identifiable information (PII) is masked or removed.
  + Maintains the utility of the data for analysis while protecting privacy.
  + Complies with legal and regulatory requirements such as GDPR or Federal Law No. 152 of the Russian Federation.
* Role in ETL Process:
  + Enables the safe use of data in environments where sensitive information must be protected.
  + Reduces the risk of data breaches and misuse of personal data.
  + Essential for ethical data handling practices.

**Output Data**

1. Upload-ready Data

* Definition: Upload-ready Data is the final product of the ETL process, consisting of Cleansed, Merged, Transformed, and Anonymized Data that is ready to be uploaded to the data module for use in machine learning applications.
* Characteristics:
  + Fully processed and compliant with all data quality and privacy standards.
  + Formatted and structured to meet the requirements of downstream applications.
  + Accompanied by updated Meta Data for traceability and governance.
* Role in ETL Process:
  + Serves as the input for machine learning models and analytics tools.
  + Ensures that data consumers have access to high-quality, reliable data.
  + Supports versioning and change management for continuous improvement.