Data Integration in Data Mining

**What is Data Integration?**

It has been an integral part of data operations because data can be obtained from several sources. It is a strategy that integrates data from several sources to make it available to users in a single uniform view that shows their status. There are communication sources between systems that can include multiple databases, data cubes, or flat files. Data fusion merges data from various diverse sources to produce meaningful results. The consolidated findings must exclude inconsistencies, contradictions, redundancies, and inequities.

Data integration is important because it gives a uniform view of scattered data while also maintaining data accuracy. It assists the data-mining program in meaningful mining information, which in turn assists the executive and managers make strategic decisions for the enterprise's benefit.

The data integration methods are formally characterized as a triple (**G, S, M**), where; **G** represents the **global schema**,

**S** represents the **heterogeneous** source of schema,

**M** represents the **mapping** between **source** and **global** schema queries.

**Why is the Data Integration Important?**

• Companies that want to stay competitive and relevant welcome big data and all of its benefits and drawbacks.

• One of the most common applications for data integration services and technologies is market and consumer data collection.

• Data integration supports queries in these vast datasets, benefiting from corporate intelligence and consumer data analytics to stimulate real-time information delivery.

• Enterprise data integration feeds integrated data into data centers to enable enterprise reporting, predictive analytics, and business intelligence.

**Data integration is particularly important in the healthcare industry.**

❖ Integrated data from various patient records and clinics assist clinicians in identifying medical disorders and diseases by integrating data from many systems into a single perspective of beneficial information from which useful insights can be derived.

❖ Effective data collection and integration also improve medical insurance claims processing accuracy and ensure that patient names and contact information are recorded consistently and accurately. Interoperability refers to the sharing of information across different systems.

**Data Integration Approaches**

There are mainly two types of approaches for data integration. These are as follows: **Tight Coupling**

It is the process of using **ETL** (**Extraction, Transformation, and Loading**) to combine data from various sources into a single physical location.

**Loose Coupling**

Facts with loose coupling are most effectively kept in the actual source databases. This approach provides an interface that gets a query from the user, changes it into a format that the supply database may understand, and then sends the query to the source databases without delay to obtain the result.

**Issues in Data Integration**

When you integrate the data in Data Mining, you may face many issues. There are some of those issues:

**Entity Identification Problem**

❖ As you understand, the records are obtained from heterogeneous sources, and how can you 'match the real-world entities from the data'.

❖ For example, you were given client data from specialized statistics sites. ❖ Customer identity is assigned to an entity from one statistics supply, while a customer range is assigned to an entity from another statistics supply. ❖ Analyzing such metadata statistics will prevent you from making errors during schema integration.

✓ Structural integration is completed by guaranteeing that the functional dependency and referential constraints of a character in the source machine match the functional dependency and referential constraints of the identical character in the target machine.

✓ For example, assume that the discount is applied to the entire order in one machine, but in every other machine, the discount is applied to each item in the order.

✓ This distinction should be noted before the information from those assets is included in the goal system.

**Redundancy and Correlation Analysis**

One of the major issues in the course of data integration is redundancy. Unimportant data that are no longer required are referred to as redundant data.

It may also appear due to attributes created from the use of another property inside the information set.

For example, if one truth set contains the patronage and distinct data set as the purchaser's date of the beginning, then age may be a redundant attribute because it can be deduced from the use of the beginning date.

Inconsistencies further increase the level of redundancy within the characteristic. The use of correlation analysis can be used to determine redundancy. The traits are examined to determine their interdependence on each difference, consequently discovering the link between them.

**Tuple Duplication**

Information integration has also handled duplicate tuples in addition to redundancy. Duplicate tuples may also appear in the generated information if the denormalized table was utilized as a deliverable for data integration.

**Data warfare Detection and backbone**

The data warfare technique of combining records from several sources is unhealthy. In the same way, that characteristic values can vary, so can statistics units. The disparity may be related to the fact that they are represented differently within the special data units. For example, in one-of-a-kind towns, the price of an inn room might be expressed in a particular currency. This type of issue is recognized and fixed during the data integration process.

**Data Integration Techniques**

There are various data integration techniques in data mining. Some of them are as follows:

**Manual Integration**

✓ This method avoids using automation during data integration. The data analyst collects, cleans, and integrates the data to produce meaningful information. ✓ This strategy is suitable for a mini organization with a limited data set. ✓ Although, it will be time-consuming for the huge, sophisticated, and recurring integration. Because the entire process must be done manually, it is a time consuming operation.

**Middleware Integration**

✓ The middleware software is used to take data from many sources, normalize it, and store it in the resulting data set.

✓ When an enterprise needs to integrate data from legacy systems to modern systems, this technique is used.

✓ Middleware software acts as a translator between legacy and advanced systems. ✓ You may take an adapter that allows two systems with different interfaces to be connected. It is only applicable to certain systems.

**Application-based integration**

✓ It is using software applications to extract, transform, and load data from disparate sources.

✓ This strategy saves time and effort, but it is a little more complicated because building such an application necessitates technical understanding. ✓ This strategy saves time and effort, but it is a little more complicated because building such an application necessitates technical understanding.

**Uniform Access Integration**

✓ This method combines data from a more disparate source.

✓ However, the data's position is not altered in this scenario; the data stays in its original location.

✓ This technique merely generates a unified view of the integrated data. ✓ The integrated data does not need to be stored separately because the end user only sees the integrated view.

**Data Warehousing**

✓ This technique is related to the uniform access integration technique in a roundabout way.

✓ The unified view, on the other hand, is stored in a different location. It enables the data analyst to deal with more sophisticated inquiries.

✓ Although it is a promising solution and increased storage costs, the unified data's view or copy requires separate storage and maintenance costs.

**Integration tools**

There are various integration tools in data mining. Some of them are as follows: **On-promise data integration tool**

An on-premise data integration tool integrates data from local sources and connects legacy databases using middleware software.

**Open-source data integration tool**

If you want to avoid pricey enterprise solutions, an open-source data integration tool is the ideal alternative. Although, you will be responsible for the security and privacy of the data if you're using the tool.

**Cloud-based data integration tool**

A cloud-based data integration tool may provide an **'integration platform as a service'**.

**Conclusion**

Data integration is the process of combining data from many sources. Data integration must contend with issues such as duplicated data, inconsistent data, duplicate data, old systems, etc. Manual data integration can be accomplished through the use of middleware and applications. You can even use uniform access or data warehousing. There are several tools available on the market that may be used to do data integration.

**DATA TRANSFORMATION:**

In data mining pre-processes and especially in metadata and data warehouse, we use data transformation in order to convert data from a source data format into destination data.

We can divide data transformation into 2 steps:

• **Data Mapping:**

It maps the data elements from the source to the destination and captures any transformation that must occur.

• **Code Generation:**

It creates the actual transformation program.

Data transformation:

• Here the data are transformed or consolidated into forms appropriate for mining. • Data transformation can involve the following:

♣ Smoothing:

• It works to remove noise from the data.

• It is a form of data cleaning where users specify transformations to correct data inconsistencies.

• Such techniques include binning, regression, and clustering.

♣ Aggregation:

• Here summary or aggregation operations are applied to the data.

• This step is typically used in constructing a data cube for analysis of the data at multiple granularities.

• Aggregation is a form of data reduction.

♣ Generalization :

• Here low-level or “primitive” (raw) data are replaced by higher-level concepts through the use of concept hierarchies.

• For example, attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.

• Generalization is a form of data reduction.

♣ Normalization:

• Here the attribute data are scaled so as to fall within a small specified range, such as 1:0 to 1:0, or 0:0 to 1:0.

• Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest-neighbor classification and clustering

• For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., income).

• There are three methods for data normalization:

1. min-max normalization :

o performs a linear transformation on the original data

o Suppose that minAand maxAare the minimum and maximum values of an attribute, A.

o Min-max normalization maps a value, v, of A to v0 in the range [new minA;newmaxA] by computing

Min-max normalization preserves the relationships among the original data values.

1. z-score normalization

o Here the values for an attribute, A, are normalized based on the mean and standard deviation of A.

o Value, v of A is normalized to v0 by computing



where A and σA are the mean and standard deviation, respectively.

o This method of normalization is useful when the actual minimum and maximum of attribute Aare unknown, or when there are outliers that dominate the min-max normalization.

1. normalization by decimal scaling:

o Here the normalization is done by moving the decimal point of values of attribute A. o The number of decimal points moved depends on the maximum absolute value of A. o Value, v of A is normalized to v0 by computing



where j is the smallest integer such that Max



Attribute construction:

o Here new attributes are constructed and added from the given set of attributes to help the mining process.

o Attribute construction helps to improve the accuracy and understanding of structure in high-dimensional data.

o By combining attributes, attribute construction can discover missing information about the relationships between data attributes that can be useful for knowledge discovery.

EG:The structure of stored data may vary between applications, requiring semantic mapping prior to the transformation process. For instance, two applications might store the same customer credit card information using slightly different structures:



To ensure that critical data isn’t lost when the two applications are integrated, information from Application A needs to be reorganized to fit the data structure of Application B.