International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

DATA TRANSFORMATION NORMALIZATION TO DENORMALIZATION IN CLOUD

Ankit Srivastava
Ankit1985sri@gmail.com

Abstract

Data integration is the process of combining data from multiple sources into a single view for users. One example of data integration is ensuring that a customer support system has the same customer records as the accounting system.Data integration is the process for combining data from several disparate sources to provide users with a single, unified view. Integration is the act of bringing together smaller components into a single system so that it's able to function as one. Normalized data refers to a database design technique that organizes data in a way that reduces redundancy and improves data integrity. The primary goal of normalization is to eliminate data anomalies and inconsistencies by organizing data into well-structured tables that adhere to certain rules. Denormalized data refers to a database design approach where data from multiple tables is combined into a single table. The purpose of denormalization is to optimize data retrieval and improve performance, especially in scenarios where read operations significantly outnumber write operations.

Index Terms— Data, Normalized, Denormalized, Data Integration, Types of data

I. INTRODUCTION

Data is a collection of discrete or continuous values that convey information, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted formally. A datum is an individual value in a collection of data. Data are usually organized into structures such as tables that provide additional context and meaning, and may themselves be used as data in larger structures. Data may be used as variables in a computational process (1,2).

The two main types of data are:

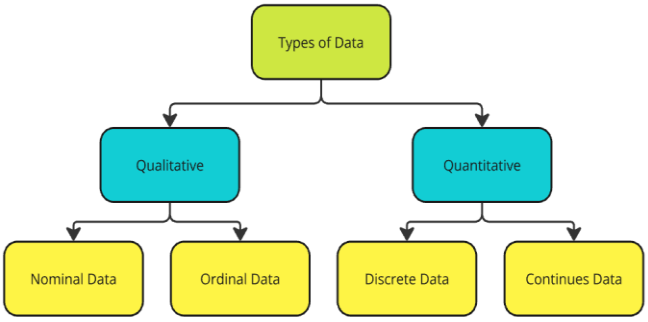

Figure 1

Qualitative Data - A type of data that can be measured and expressed in numbers. Quantitative data is used to answer questions like "how many, "'how much," and "how often"

Quantitative Data- A type of data that is exploratory and seeks to find out what potential consumers think and feel about a given subject.
The data is classified into four categories:

- Nominal data - A type of qualitative data that describes a characteristic or other descriptive factors. Nominal data is neither numerical nor can it be ranked in a hierarchical order.
- Ordinal data - A statistical type of data that has a set order or scale. The distance between categories is not established with ordinal data
- Discrete data- A type of quantitative data that is made up of whole integers and cannot be divided into parts. For example, the number of people in a population is discrete data.
- Continuous data - Continuous data includes values that can assume any number within a given range or interval, often depicted by fractional numbers

Data Integration - In modern days we get data from multiple source system and integrate the data using ETL system. We have now API's available to retrieve the data from source system. Traditionally we get denormalization data and convert it into normalization (3NF) to store it into Database. The Kimball methodology is a structured approach to designing, developing, and deploying data warehouses and business intelligence systems. It was developed in the 1980s by Ralph Kimball and colleagues at Metaphor Computer Systems.

First normal form (1NF), which eliminates duplicate columns within a table and ensures that each column contains atomic (indivisible) values.

Second normal form (2NF), which meets the requirements of 1NF and removes partial dependencies by ensuring that all non-key attributes are fully functionally dependent on the primary key.

Third normal form (3NF), which meets the requirements of 2NF and eliminates transitive dependencies by ensuring that non-key attributes are not dependent on other non-key attributes.[3].

Data integration is the process of achieving consistent access and delivery for all types of data in the enterprise. All departments in an organization collect large data volumes with varying structures, formats, and functions. Data integration includes architectural techniques, tools, and practices that unify this disparate data for analytics. As a result, organizations can fully view their data for high-value business intelligence and insights.

Dimensional modeling is a process in which the business requirements are used to design dimensional models for the system.
Physical design is the phase where the database is designed. It involves the database environment as well as security.

Extract, transform, load (ETL) design and development is the design of some of the heavy

procedures in the data warehouse and business intelligence system. Kimball et al. suggests four parts to this process, which are further divided into 34 subsystems. [4]
Data consolidation uses tools to extract, cleanse, and store physical data in a final storage location. It eliminates data silos and reduces data infrastructure costs. There are two main types of tools used in data consolidation.

ETL- ETL stands for extract, transform, and load. First, the ETL tool extracts the data from different sources. Next, it changes the data according to specific business rules, formats, and conventions. For example, the ETL tool could convert all transaction values to US dollars, even if the sales were in other currencies. Finally, it loads the transformed data to the target system, such as a data warehouse.

ELT- ELT stands for extract, load, and transform. It is similar to ETL, except that ELT switches the final two data processes on the sequence. All the data is loaded in an unstructured data system, like a data lake, and transformed only when required. ELT takes advantage of cloud computing's processing power and scalability to provide real-time data integration capabilities.

Data replication, or data propagation, creates duplicate copies of data instead of moving data physically from one system to another. This technique works well for small and medium businesses with few data sources. For example, a retail hardware business could use enterprise data replication to copy specific tables from its inventory to its sales database.

Data virtualization does not move data between systems but creates a virtual unified view that integrates all the data sources. The storage systems do not transfer data between databases during data virtualization. Instead, it populates the dashboard with data from multiple sources after receiving a query.

Database administrators use several data denormalization techniques depending on the scenario.
Introducing a redundant column/Pre-joining tables
This technique can be used when there are expensive join operations and data from multiple tables are frequently used. Here, that frequently used data will be added to one table.

Data Denormalization brings several advantages for organizations.
Improve user experience through enhanced query performance
Querying data from a normalized data store may require multiple joins from different types of tables, depending on the requirement. When the data grows larger, it will slow down the performance of Join operations. It can negatively impact the user experience, especially when such operations are related to frequently-used functionalities.

Data denormalization allows us to reduce the number of joins between tables by keeping frequently accessed data in redundant tables.
Reduce complexity, keep the data model simple

Data denormalization reduces the complexity of queries by reducing the number of join queries. It enables developers and other application users to write simple and maintainable codes. Even novice developers can understand the queries and perform query operations easily.

Plus, this simplicity will help reduce bugs associated with database operations significantly.[5]

II. SUMMARY

With new cloud database system storage technique has transformed. With Pay per use databases like Amazon aurora, Azure and Snowflake storage techniques and data transformation has changed. In on premise database we get data from different system store it in different stage table create fact and dimension table. Created primary and foreign key relationships and loaded in final table. With new cloud database technology techniques, data is stored in single table instead of fact dimension table. These are fast processing we can do all the transformation in one table and store all data in single table. Example one table for each system eg: claims, provider, member etc. it will take less time to retrieve and process and cost less. It will also make life of reporting and ETL team easy as they will not need to do all the joins and only need to have transformation logic and filter the condition. It doesn't require team to be technical and will take less time for them to develop the code.

III. CONCLUSION

With new technology involving and costing, storage and retrieving of data has changed over time, now companies can concentrate on storing of data only without thinking about server maintenance, server security, insurance and other costs and can buy the storage as needed. It drastically reduces the cost as its pay as per use. This has lead to change in storage strategy also, since storage cost has decreased and company need to pay for computation also so its better to store the data in denormalize form as there is no joins on table and data retrieval would be fast, it also will take less time in developing ETL pipelines and reporting dashboard.

REFERENCES

1. OECD Glossary of Statistical Terms. OECD. 2008. p. 119. ISBN 978-92-64-025561.
2. Australian Bureau of Statistics. 2013-07-13. Archived from the original on 2019-04-19. Retrieved 2020-03-09.
3. Pure Storage, https://blog.purestorage.com/author/pure/
4. Kimball, Ralph; Ross, Margy; Thornthwaite, Warren; Mundy, Joy; Becker, Bob (2008), The Data Warehouse Lifecycle Toolkit, Wiley Publishing Inc
5. Shanika Wickramasinghe https://www.splunk.com/en_us/blog/learn/data-denormalization.html