International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-1, January 2022
https://doi.org/10.55524/ijircst.2022.10.1.18
Article ID IRP221014, Pages 102-106
www.ijircst.org

Sturdy Data Warehouse for Complex Data of Travel Survey

Madhav Singh Solanki [1], and Mrinal Paliwal [2]

[1, 2] Assistant Professor, Department of Computer Science Engineering, Sanskriti University, Mathura, Uttar Pradesh

Correspondence should be addressed to Chandra Sekhar Sanaboina; madhavsolanki.cse@sanskriti.edu.in

ABSTRACT: The data warehouse enables information to be organized in order to ease data handling from one sphere to the other and to promote knowledge acquisition. The requirement for a consistent organization and compatibility across various data sources grows as the quantity of data grows, making it more difficult to conduct thorough analyses within short periods. This study offers a trip data warehouse that uses dimensional modeling to promote a more comprehensible structure, comparable findings, quicker data access, and faster publishing of summaries. The use of multivariate representation to transport information helps to improve construction while also integrating, augmenting, and improving data. It performs data processing and validation in an automated manner. The development of transportation planning tools is anticipated to lead to the creation of a multivariate representation for trip data. In future, there is a pragmatic scope of extensive research in this field.

KEYWORDS: Data Warehouse, Data warehouse, Knowledge management, Metadata, Travel survey.

I. INTRODUCTION

It is very much critical for the different players engaged in holiday preparation to create a precise image of the wandering habits of the population that utilizes the conveyance methods, whether it is to improve competence and sustainability or to prepare for change. The growing quantity of data, the growing amount and involvedness of computer programs or platforms utilized, and the growing responsibility of patrons for their choices all create concerns regarding the effective use of intellectual capital in this field [1].

A. Paradigms and Definitions

This section begins with terminology related to travel data collecting, and then moves on to the major paradigms related to travel data management. It then moves on to a discussion of how to build high-performance data warehouses using dimensional modeling techniques taken from the area of business intelligence. Finally, the relationship between data warehousing and visualization elements such as animated trip maps is discussed.

a. Origin-Destination Travel Surveys

Travel behavior research relies heavily on data from origin-to-destination surveys. Defendants are requested to give their political and social background and to describe the locations visited by them through a certain time period, including the activities they engaged in as well as the means of transportation they used to get there, during interviews. This data allows for an impartial assessment of the population's travel habits.

b. Household vs Personal Review

It is common in a domiciliary review to gather the statistical data of every household member as well as travels they took within the given time period. Typically, only one household member is needed to participate in these surveys, and that individual is also obliged to provide statistics for the extra affiliates of the home. When conducting a personal review, just one person is interviewed, and the objective is typically to gather travel data for that person alone. Nonetheless, collecting basic demographic information on additional household members is common practice for survey comparability, modeling (household structure being a key predictor of behaviors), or sample weighting reasons.

c. Harmonization of Travel Surveys

The European Co-operation in Scientific & Technological (ECST) initiative has suggested a documentation of referencing to unify transportable survey techniques so as to deal with the creation of survey forms and methodology especially in European countries. This attempt to standardize is promising. Despite the fact that one of the document's goals mainly was to provide coherent facts, the methods for doling out, verifying, and storage of data were not specified.

d. Travel Objects Modeling

Trépanier suggested an object connotation drawing expressing the items gathered in a archetypal Origin-Destination domiciliary study, as illustrated in Fig 1, using a completely disaggregated method to verify, analyze, and model data from travel surveys [2]. For data generated by transportation reviews, the department in authority of local reviews in the province of Quebec suggested a prototypical relation called SAQE a few years ago [3]. It has the same components as the previous one, but it also includes tables related to the interview procedure. The data model also includes a list of the characteristics gathered for each item. Valiquette researched and developed the idea of trip chains to better comprehend the sequence of excursions and the impact they have on one another throughout the day [4]. The SAQE model does not have trip chains. They are,

however, produced after the data gathering procedure is completed. As the construction of travel databases enables conveyance organizers to conduct disaggregated studies, there is a demand of increasingly comprehensive and relevant solutions across surveys, as well as data formats that are easier to comprehend and offer quicker access to enhanced and verified data.

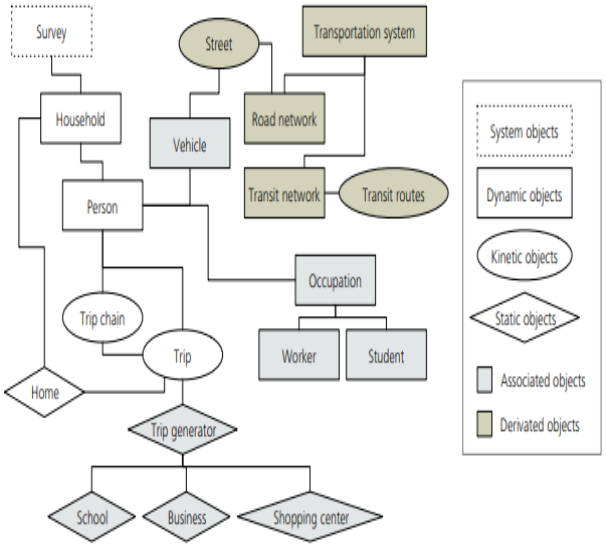

Figure 1: ER illustration of items relating to travel comportment analysis [3]

II. DISCUSSION

A. Data Warehouses Modeling

A DW is a kind of a data fountain solution intended to make the analysis and improvement of verified data from various sources easier and faster [5,6]. The regularized prototype and the dimensional archetype suggested by Kim et al., both based on the relational database management system (RDBMS) model, are the two primary methods to data management inside a data warehouse [7,8]. To prevent information redundancy, the normalized model adheres to the normal forms. It does, however, increase the number of tables, result in numerous joining procedures during queries, and does not make it easier to understand the outcomes just by observing the schema. The discussed prototype was created to make data exploration and elucidation easier, but it came along with expense of few of the data redundancy, more storing capacity, and a longer preliminary handling time. Because the tables have already been partly deformalized and in certain instances aggregated for specific analyses, enquiries from a data warehouse depending on the discussed prototype are considerably quicker. As a result, the database doesn't need to connect all of the tables in a regularized relational structure. With a few exceptions, the dimensional model suggests just permitting single level of distinction across the entries (single join). The main difference between the two models is shown in Figure 2. The regularized prototype on the left, for example, has triple level of distinction among both the sales transaction table and the representative table, whereas the discussed prototype only has single level of distinction.

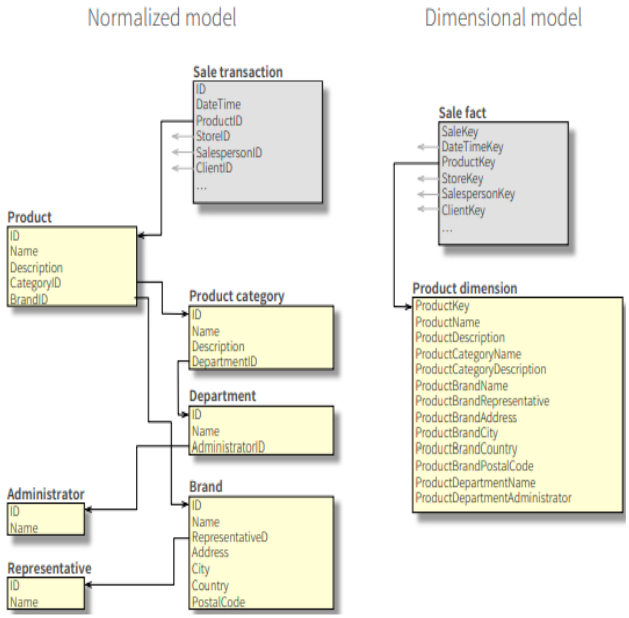

Figure 2: Association among the regularized proposed prototype [3]

a. Dimensional Modeling

Dimensional modeling is the best way to achieve two primary goals: presenting comprehensible data and rapidly performing the required queries. Star schemas are another name for dimensional models. A detail table reflecting the fundamental items of the area of learning is located in the middle of the star. For example, a piece of information in the detail table is created for each sale of a particular product at a specified time in a retail company. The dimensions may be found at the star's points. In the instance of a retail transaction, the model will have one table for dates, another for goods and descriptions, a third for salespeople, a fourth for consumers, a fifth for shops, and so on. figure 3 depicts a star schema for a retail company based on a dimensional model. For example, the characteristics of the connected dimensions are incorporated into the table for the product dimension, while same attributes are split into many different tables on the normalized side.

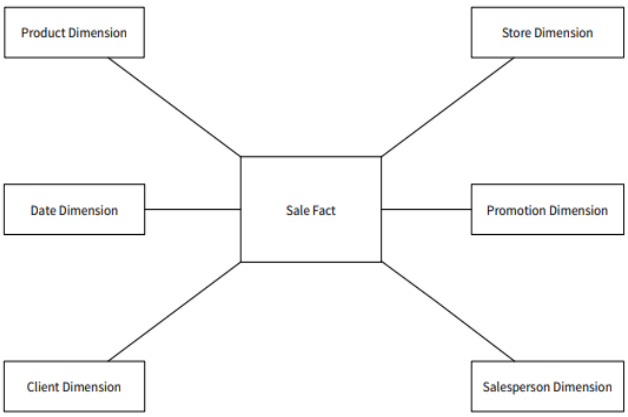

Figure 3: Illustrates an instance of a multivariate star schema of a business relating to retail sale

b. Facts and Dimensions

Merely assessable values (typically the numerical one) and foreign keys connect the fact to its related dimensions in the fact database. Because it is the table with the most entries, the guideline of not adding text or superfluous information greatly lowers its size. Simply detail tables ought to be linked to supplementary tables most of the time (dimensions). Written reports, numerical characteristics, and typically pre-calculated variables that enhance the data for better comprehension and analysis are all included in the dimension tables. The amount of records in the multivariate tables is very modest compared to the number of entries in the fact table since they are linked using keys in the detail table and the similar multivariate is usually associated with more than one fact. It's worth noting that the dimension tables have been de-normalized. When it comes to a product explicit feature, for example, details on product categories are straightly integrated in the dimension. As a consequence, the data is recurrent for every product in a given classification. Such model upsurges the scope of the multivariate tables, but it also provides for a better knowledge of the dimensional characteristics and improved query performance.

c. Data warehouse and Visualization

Indeed, the construction of a data warehouse depending on multivariate modeling enables research groups and even the common people to choose unique datasets for their own investigations. Visualization is a fact-finding activity that may promote formulation of novel research ideas while also preventing non-specialist misinterpretations when properly structured. The most essential requirements for cherishing a visualization item and aiding the comprehension of complicated physical and human events are clarity, accuracy, and efficiency. Geovisualization, in particular, combines the geographical aspect with visuals and is a critical means for spatial data investigation and analysis. To explain transportation-related problems, whole research groups and organizations of visualization experts investigate and propose visualization items. Simply making data more available via efficient visualization objects may persuade transportation management organizations to fund specific initiatives for the development of visualization technologies [9,10].

d. Travel Data warehouse

The suggested modeling method is first described to contextualize the development of a travel data warehouse. The characteristics that guarantee data comparability from one survey to the next are specified, as well as the data enrichment procedure. Also mentioned are the dispensation and authentication techniques utilized to transform the collected data to the data warehouse. The features of the various tables in the multivariate prototype are ultimately specified after that.

e. Multivariate Modeling of Travelling Dataset

The initial stage in creating a multivariate model is to identify the system's fundamental objects, or granularity levels. These are the components that make up the fact tables. From the most basic to the most complex, the following items are found in the study of travel behavior: the home, the individual, the trip chain, the trip, the visited location, and the trip segment. The journey segment is not usually the primary target in travel behavior analysis. In reality, certain analyses may be satisfied with greater granularity if they do not need such exact information. The context determines whether the various items in a system are classified as facts or dimensions. It is possible to discover both factual and dimensional things. The categorization is determined by the user's requirements. Directional modelling, in most circumstances, simply provides for a yet another distinction among data and proportions tables. There is an exemption to the principle because the timestamp aspects may indeed be accessible by keywords in the attribute values, leading to a second level of abstraction.

f. Analogous Characteristics

One among primary goals of establishing a transportable warehouse of data is to make data comparisons across surveys easier. A collection of similar characteristics is suggested to do this. The first feature of these characteristics is their longevity (they're also known as durable attributes), which means they're not intended to be changed often, if at all. If one of the similar characteristics is to be changed, all surveys recorded in the data warehouse must be updated in order to preserve data comparability across time. The second aspect is that all options are included to allow for thorough comparisons. A similar mode property, for example, should contain all conceivable modes, not only locally but also globally. To prevent ambiguity, special attention must be made in the definition and suggested options for each similar characteristic. In this study, several similar characteristics and dimensions were suggested. The final selection of similar characteristics and dimensions to be included in the data warehouse, on the other hand, must be the product of a potential agreement among stakeholders in the area of trip surveys.

g. Information Amelioration

The radical development of data warehouse usually being preceded by a information amelioration procedure. Numerous variables are pre-computed upon import in this dimensional model proposal for trip data, and several additional characteristics are introduced to simplify and enhance the analyses. The minimum lifespan of the residential area, the percentage of females and males, the proportion of children, employees, and retired people, the proportion of visitors in each age group, the mean and total amount of journeys, the average and cumulative distances travelled by all family members, as well as the least number of cars needed in the residence.

h. Data Validation

The process of cleaning and loading data into a data warehouse is known as ETL (Extract, Transform, and Load). When new data sources are added or the format of the imported files changes, the ETL module must be changed. The initial version of the ETL module was built using data from different online surveys as well as data from numerous telephone polls conducted in the Quebec area as part of this research project. The data is enhanced and loaded into the data warehouse once it has been verified and processed. The necessary fields and the

criteria of validity of the various information stated by the responders are listed in a configuration file. When one of the rules is broken, a record is created in the audit table of the collection database. All validations are adaptable, and they may differ from survey to survey based on the administrators' and analysts' requirements.

i. Data Consistency

Data consistency is a major issue when de-normalizing data from the assemblage catalogue into the data warehouse magnitudes. Whenever a subscriber updates the title of an event, for instance, the activity aspect and the columns that hold event identifiers must be changed. In a dimensional model, however, it is the duty of the ETL module to ensure that the data is consistent by modifying the values in all impacted tables. In this sense, all users must utilize the data warehouse in read-only mode. Indeed, the ETL module must be the only one having write access to all tables.

j. The Dimensions of Different Fact Tables

Each fact and dimension are specified in this section. The fact tables are then provided in a schema with their dimensions and key characteristics (domiciliary, individual, visited location, trip chain etc.).

k. The Survey and Sample Dimensions

All of the fact tables have two dimensions connected with them. The first is the survey dimension, which contains information about the survey for which the various trip items were gathered. Second, each home and respondent that takes part in the survey has a dimension that describes the sample. Attributes on recruitment and collecting techniques, sample sizes, and recruitment durations are all included in this dimension.

l. The Date Dimensions

Several characteristics have been introduced to the date dimensions, allowing for various categorization and purifying techniques related with dates. The same contain the day's name, the day's count in the week i.e., may be 1, 2, 3…, and 7, in the the months - possibly 1, 2, 3…, and 31, and in complete year, as well as the week number in the month (1-4 or 5) and the year (possible values from 1 to 366). (1-52 or 53). Another option specifies whether the date falls inside the week (Saturday or Sunday). Finally, in their respective fields, the year, month, and day are given individually. It is feasible, for example, to run a query that groups trips by kind of day (weekday or weekend) without needing to first group by day of the week. This simplifies searches and groups, particularly when several additional kinds or filters are used in the query. The date dimension also ensures data quality and comparability by standardizing all dates used in the data warehouse. Because each date has just one entry in the date table, all dimensions and facts that relate to the same date have the same date key.

m. Some Additional Time Measurements

The further dimensions of time enable grouping and sorting by epochs or time intervals of the day simpler. A measurement counter for 24 hours available in a day (0 through 23) is produced, while another for the full 1440 minutes of a day is created. We may discover characteristics that indicate the time, with or without the minutes, in a variety of forms in these tables. For instance, there are separate entries for period in 24-hour worldwide set-up & 12-hour acceptable set-up. A different support shows the minutes from the time after mid-night, which makes it easier to do sequential analyses that don't need conversion or special methods to calculate time intervals. The extended time dimensions are similar, except they contain hours ranging from 0 to 47 hours (0 to 2879 minutes). Because the format specified by the GTFS for transit schedules may exceed 24 hours, the inclusion of the additional period measurements provides greater consistency in the analysis of night journeys.

n. Geographic Object Dimensions

Survey administrators may submit geographic data using the import interface. The items in the file, as well as their categories, will be put into the data warehouse's geographic object dimension table after it has been imported. Each record having a geographic type property (home, typical location of work or study, visited place, etc.) will be connected to all geographic objects of the polygonal type associated with it during data processing (by means of a spatial intersection). Each imported geographic object category is given its own column. All queries on the related fact tables will therefore have access to the properties of the geographical objects.

o. Web Channeling Aspect

The trip web channeling aspects are suggested for gathering data for routing computations for a specific trip. Each trip's routing is computed and saved automatically for each of the major means of transport. This allows for comparisons of various modes for the same trip that aren't restricted to the respondent's stated mode, and travel durations for each mode may be input into modal choice models. The boarding and alighting stations, as well as the route and form of public transportation are all note down. When average rate are given for every group, modified trip times for walking and cycling are also supplied based on the respondent's demographic data.

p. Household Fact

The fact table is linked with a dimension reflecting the home location for each household. The features of the household are included in the household dimension.

q. Person Fact

All eligible household members and responders are included in the person fact table. Each person has a home dimension, a household dimension, two dimensions for the usual places (one for work and the other for study, both optional), and a person aspect that involves the person's characteristic features (primary profession, ownership of a driver's license, ownership of a transit pass, interview language etc.). This person dimension also includes pre-calculated descriptive statistics on the travels taken, as well as similar characteristics that are the same for all surveys kept in the data warehouse (comparable age group, comparable profession, etc.). The journey from home to the typical location of work/study, as well as the trip from the usual place of work/study back home, are both included in the network routing dimensions. Finally, the survey respondent's person

dimension (sometimes known as a proxy when this person gives responses for other members) is added.

r. Visited Place Fact

The visited place table follows the same logic as the person-made table, and contains a home, household, person dimensions, and the dimensions of the individual's regular place of work/study. A dimension defining the activity carried out at the visited location, as well as the date and time dimensions of arrival and exit, are also included.

s. Trip Chain Fact

The trip chain table is connected with beginning and finish time and date of the trip chain, as well as the primary activity carried out at the anchor point of the trip chain, in addition to the dimensions mentioned above (home, household, person, regular locations). The trip chain dimension contains characteristics such as the amount of rounds, the quantity of tours, different modes, segments, the length of the primary activity, and the overall distance traveled, as described by Valiquette.

III. CONCLUSION

The growing complication of handling, treating and maintaining data gathered via travel studies and reviews has prompted a need to streamline data validation and storage. Despite the fact that completely disaggregated modeling correctly depicted the objects connected with travel behavior, the growing quantity of data gathered from various sources has only added to the management problem. The primary goal of the platform described in this article was to make the processing of trip data easier, more flexible, and quicker, while also allowing data enrichment and automatic creation of visualization objects as a solution to this issue. The construction of a trip data warehouse has been suggested, with the goal of simplifying the integration, validation, and visualization of travel data, thanks to the adaption of dimensional modeling widely used in business intelligence. The issue of survey comparability has to be investigated further. Despite the fact that the created platform makes it simpler to handle similar characteristics, the problems connected with survey comparison must be evaluated and monitored. There will be many future improvements to the platform and the trip data warehouse. A conversation with travel survey managers from across the globe would aid in standardizing similar characteristics and selecting the finest data enrichments to be included by default in all surveys imported into the platform. The goal of sharing travel data that has been enriched and validated using the proposed structure is to make it easier to analyze travel behaviors, feed models and simulations, and especially to improve and extend the analysis power of travel data that has been collected in the past and will be collected in the future.

REFERENCE

[1] AR. Real-time big data warehousing and analysis framework. In: 2018 IEEE 3rd International Conference on Big Data Analysis, ICBDA 2018. 2018.

[2] Sioui L, Morency C, Trépanier M. How Carsharing Affects the Travel Behavior of Households: A Case Study of Montréal, Canada. Int J Sustain Transp. 2012;

[3] Bourbonnais PL, Morency C. A robust datawarehouse as a requirement to the increasing quantity and complexity of travel survey data. In: Transportation Research Procedia. 2018.

[4] Sicotte G, Morency C, Farooq B. Comparison Between Trip and Trip Chain Models: Evidence from Montreal Commuter Train Corridor. 2017;(June). Available from: https://www.cirrelt.ca/DocumentsTravail/CIRRELT-2017-35.pdf

[5] Ren S, Wang T, Lu X. Dimensional modeling of medical data warehouse based on ontology - 2018 {IEEE} 3rd {International} {Conference} on {Big} {Data} {Analysis} ({ICBDA}). 2018 IEEE 3rd Int Conf Big Data Anal. 2018;

[6] M Kirmani M. Dimensional Modeling Using Star Schema for Data Warehouse Creation. Orient J Comput Sci Technol. 2017;

[7] Kimball R, Ross M. The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling. Wiley. 2013.

[8] Kimball R, Reeves L, Ross M, Thornthwaite W. The Data Warehouse Lifecycle Toolkit Table of Contents. Architecture. 2008;

[9] Silva SF. A web visualization tool for historical analysis of geo-referenced multidimensional data. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2008.

[10] Nogués A, Valladares J. Business Intelligence Tools for Small Companies. Business Intelligence Tools for Small Companies. 2017.