Abstract
The challenging task present in the data integration process is compromising the attribute value confliction. Data heterogeneity is the main reason for this attribute value conflicts. In this paper, a determination framework is developed to resolve the numerical value conflicts. Instead ad-hoc approach, this framework uses one new efficient systematic approach to handle this attribute value conflicts. This systematic approach considers the consequence of incorrect numerical values and selects a numerical value which minimizes the error cost for various applications.

I. Introduction
As long as the data integration process remains one of the most difficult task in the data management. In the challenges are present three levels. There are schema heterogeneity, entity heterogeneity, data heterogeneity.

A. Challenges

1. Schema Heterogeneity
Schema heterogeneity means that , the use of different structures or different names for the same information. For example, The personal details about Mr.JohnSundhar contain the details of age, sex, address, designation, salary in one data source. In that same data source, the information about Mr.JohnSundhar is stored using the name of Mr.John. This kind of problem arises because of schema heterogeneity. We can eliminate this problem by using the unique identifier value.

2. Entity Heterogeneity
Entity heterogeneity means that the information about the same real world entity is stored in different data sources with different identifier value. For example, the id value of Mr.JohnSundhar is 40. But the information about Mr.JohnSundhar is stored in different data source with the id value 145.

3. Data Heterogeneity
Data heterogeneity caused by the use of data inconsistencies in the absence of schema and entity heterogeneity. For example, the salary of Mr.JohnSundhar is stored as Rs.5,00,000 in one data source and Rs.5,50,000.

4. Comparison of Three Challenges
Hence, the schema heterogeneity and entity heterogeneity are not exist or already resolved. The question is , which salary value is chosen for Mr.JohnSundhar’s salary ? And the another question is , how to reduce the error cost ?. We have discussed about the three challenges present in the data application problems. Among these, schema heterogeneity has more attention. There are numerous schema matching techniques are developed to resolve the schema heterogeneity [11, 15]. Suppose if the schema heterogeneity is the reason for numerical value conflicts that resolved systematically using data conversion rule. Then we start to discuss about entity heterogeneity, the resolving techniques for entity heterogeneity are entity identification and matching [7,13] and duplication detection and removal [13,15].

B. Problem Statement
We divide the attributes into two categories. There are categorical attributes and numerical attributes. The above algorithm is suitable only for categorical attributes not for numerical attributes for various reasons. First, for categorical attributes we can decide whether the observed value is correct or wrong. But it is not possible for numerical attributes. For example the customer Mr.Sakthi’s true state is Tamilnadu. But the observed value of the customer Mr.Sakthi’s state is Kerala. Then we can decide the observed value is wrong. But in case of numerical attributes, The customer Mr.Sakthi’s true income is Rs.5,00,000. But the observed values of Mr.Sakthi’s salary are Rs.5,00,500 and Rs.2,00,000 from different data sources. Here both are considered as inaccurate values and we can’t decide whether the observed values are correct or wrong. Here we are considering the magnitude of error. If we choose the first value (Rs.5,00,500) that will not effect in the decision making. But if we choose the second value (Rs. 2,00,000) this leads to incorrect decisions. Second, categorical attribute values are mutually exclusive, but numerical attribute values observed in the different data sources based on different methods or different levels of rounding or different ranges. Third, queries based on categorical attributes use discrete values in selection conditions. But queries based on numerical attributes make their selections using the value intervals.

C. Determination Framework
Finally, a determination framework is developed to correct the unknown attribute values. It select the value that minimize
the total expected error cost for all data usage problems under considerations. To minimize the error cost, it uses probability distribution function of the true attribute values and cost of errors for all problems. The major steps of framework: Estimate the probability of all true values, deriving the cost associated with each candidate values and selecting the value that minimizes the expected error cost. To make proposed framework be feasible, reduce the no of candidate values and complexity of probability estimation.

II. Existing System
One of the major challenges of data integration is to resolve conflicting numerical attribute values caused by data heterogeneity. In addressing this problem, existing approaches proposed in prior literature often ignore such data inconsistencies or resolve them in an ad hoc manner. Ad hoc manner is not a effective one. It consumes more time and produce very high error cost. Challenges exist at three different levels: schema heterogeneity, entity heterogeneity, and data heterogeneity.

III. Proposed System
A determination framework is developed to correct the unknown attribute values. It select the value that minimize the total expected error cost for all data usage problems under considerations. To minimize the error cost, it uses probability distribution function of the true attribute values and cost of errors for all problems. The major steps of framework: Estimate the probability of all true values, deriving the cost associated with each candidate values and selecting the value that minimizes the expected error cost. To make proposed framework be feasible, reduce the no of candidate values and complexity of probability estimation.

B. Heterogeneous Database
Heterogeneous database does not contain all the data present in the master record. Heterogeneous database contain only less no of entries compare to master record. It contains only the data which are all used by the user. For example, The Sakthi pvt ltd maintaining the database for their employees and their activities. For every find of new products from the Sakthi pvt ltd, it asking the innovate ideas from their ideas. At that time, employee updates their profile. Updating means in the sense of using the data. So that data bring to heterogeneous database. Because of this we can call the heterogeneous database as used database. If the higher official of Sakthi pvt ltd wants to analyse employee that organization, he no need to check for the entire employee details i.e., master record. This is because very large master record. So it is not feasible and consumes more time. At that time it is enough to check the heterogeneous database. And the advantage it consumes less time compare to master record.

C. Unused Database
Here the inconsistent data present in the master record is separated in a separate database. That database is called unused database. Unused data means that in the sense of the data which are all not used by the user. Because consistent data only used by the user. This is very useful to handle the challenge of data heterogeneity. This is because here inconsistent data are separated. So the burden of using the inconsistent data getting reduced. For example, If the manager of Sakthi pvt ltd wants to check the performance of each and every employee who are all working skill fully and who are all not working not properly. Not properly working employee profile will not get updation. That kind of employee profile is automatically copied to unused database. Suppose the data present in the unused database getting used, it automatically moved to heterogeneous database and entry present in the unused database getting removed.

D. Probabilistic Database
To improve the accuracy, quality and speed, a new concept is introduced. That is maintaining one special database that is probabilistic database. It contain very limited no entries. It contains the entries which are frequently used by the user. Compare to heterogeneous database it contain less no entries only. This is very useful to speed up the process. Because this probabilistic database contain very limited no entries, it’s very to analyse this. For example, If the manager of the Shakhtry pvt ltd wants to give the incentive to the employees who are all working very sincerely and putting hard work to the organization, at that time the manager want to check the overall employee details i.e., master record. This is very large and so this process consumes more time. But this probabilistic database holds only details of employees who

Fig. 1: Architecture of Framework

IV. Functional Aspects

A. Database Integration
Database integration means that the process of adding the data into the data source. Database integration process is carried out from the users only. Only the users are integrating the data to the various data source or database. Database integration is not a simple task and all. At the time of integration, we have to aware of the data that we are going to integrate. As per the figure-1 the web environment contains more distinct data sources. Basically it contain n no of databases in web environment. Based on the category of the data, the user going add is routed to the distinct data source. This routing is based on the very clear analyse about the integrating data. All the present in the n databases copied into a new large data source that is called master record. This is for the purpose of processing the data. On every successful integration, the integrated data brought into master record.
working sincerely and smartly to the organization. So the manager needs not to analyze the master record and it’s enough to analyze the probabilistic database. This is very small compare to master record. So this process getting finished very soon. Its increase the performance too.

E. Numerical Value Conflicts

When the user giving queries to database, it will search and retrieve the suitable data for the query. But the main problem of the query process is conﬂict. If the query passed to master record means more amount of conﬂict will occur. This because we already defined that master record is very large. To avoid that we going another three databases, heterogeneous database, unused database, probabilistic database. Among these, the probabilistic database only is getting higher preference. So whatever the query the user giving its first routed to probabilistic database. The question is the result is not present in the probabilistic database means what will do? At that that is routed to heterogeneity database. The same question is the result is not present in the heterogeneous database means what will do? At that time its routed to unused database. This kind of query retrieving will reduce the numerical conﬂict in a signiﬁcant manner. But the second problem is, more than one true values retrieved for same query. The question is, at what manner choose the true one? If more than one true value comes for a single query means we are going for probability distribution function. This will choose most accurate for that query among the multiple results.

V. Probability Derivation

Consider a numerical attribute $C$, which is stored in $N$ databases $D_1, D_2, D_3, \ldots, D_N$. For a speciﬁed entity instance, the considered value of $C$ in databases $DN$ is denoted as $CDS$. For several reasons, the selected attribute for same entity may be different. We calculate the probability density function for the true values $C$, if $C$ is considered as continuous, i.e.

$$f(C_T=b|C_{D1},C_{D2},\ldots,C_{DN}) = \frac{f(C_{i}=b|C_{D1},C_{D2},\ldots,C_{DN})}{\sum_{i=1}^{N} f(C_{i}=b|C_{D1},C_{D2},\ldots,C_{DN})}$$

Where $C_{Dj}$ represents true probability density function of $C$. Similarly probability mass function, we

$$P(C_T=b|C_{D1},C_{D2},\ldots,C_{DN}) = \sum_{i=1}^{N} P(C_{Dj}|C_T=b)$$

VI. Conclusion

Finally, we come to the area of the conclusion of this paper. We already the three challenges data heterogeneity, schema heterogeneity and entity heterogeneity. Among these, Data heterogeneity is the major challenge and it’s very to handle too. In this paper, a decision-theoretic framework is proposed to handle this data heterogeneity. This framework also handle the another problem numerical value conﬂicts. The division of three databases from master record namely heterogeneous database, unused database and probabilistic database is new method to handle this numerical value conﬂicts. The new query processing introduced by this framework is very efﬁcient and reduce the time consumption. This is because the queries from user are not passed to the large master record. Its preferring probabilistic database only first which small. So it’s take only less time to retrieve. Suppose the selection condition not matched in the probabilistic database only its go for next level database.

This framework also simplifying the process of checking and verifying the database. This is suitable large scale organization in which we need to analyse the employee performance. This framework also support the wide range of application such as direct marketing, management of health care details and management of employees etc.,

References


Mr. Arumuga Arun received B.Tech degree in the department of Information Technology from Vel Tech Multitech Dr.RR and Dr. SR Engineering college, Chennai in 2011. And pursuing M.E degree in the department of Computer Science and Engineering in P.B College of Engineering, Chennai in the year of 2011-2013.

Mr. Anbazhagan received B.E degree in the department of Computer Science and Engineering from Rajarajeswari Engineering college Chennai and received M.Tech in the department of Information Technology from Sathyambama university, Chennai. Now he is associate professor in P.B College of Engineering, Chennai.