Data Warehousing – Maximum Timeline Availability of CDR Data

Deependra Swaroop Sharma*
Dept. of Computer Science and Engineering,
Suresh Gyan Vihar University,
Jagatpura, Jaipur-Rajasthan, India
*Email: *er.deependra {at} gmail.com*

Dinesh Goyal
Dept. of Computer Science and Engineering,
Suresh Gyan Vihar University,
Jagatpura, Jaipur-Rajasthan, India

Sagar Sharma
IT PROFESSIONAL

Abstract: **In this paper we have proposed a research in the field of Telecom data warehousing using which we can provide the maximum time line of CDR data (CDR-Call Data Records which are records generated when two or more parties communicate over cell phone or landline either verbally, by sending sms , etc) to the Law Enforcement Agencies(LEA's) to help them resolve the criminal cases. To prove the logic used, we have considered two approaches to data warehousing. First approach is the commonly used approach in Telecom industry which we can only provide one day old CDR data to the LEA's and no recent data of current date can be provided, which is a constraint with the data warehousing architecture used. To provide the maximum possible time line of CDR data to LEA's we have proposed a new Data Warehousing Architecture which has been discussed in this paper.**

Keywords: Database, Data warehousing, maximum time line of CDR data availability in telecommunications, Optimized Data warehousing architecture

I. Introduction

In this paper we will present a comparison between the two techniques of data warehousing where a superior technique will supersede the other by providing maximum possible time line of data availability in Telecom operations. While investigating a case Law Enforcement Agencies(LEA's) like police force, honorable courts etc , need the most recent data for investigation . e.g., after any bomb blast, of after a murder , or after a abduction or any other criminal offence , the CDR data that is usually made available by the telecom companies is of Sysdate -1 , and this is made available a day after any other criminal offence. During this time period the criminals get a change to save themselves. With the optimized model of telecom data warehousing presented in this paper CDR data of few hours old can be made available to the LEA's to help them work more efficiently[1].

A logical data record consisting of many columns produced by a telephone exchange agency network is called CDR, also known as call data record. In case of manual telephone exchange CDR is equal to the paper document that were written and timed by operators for long-distance calls.

II. Description of Existing/Generally followed Architecture of Data warehousing

Initially we consider a database in which there is a table for each day of month to capture the CDR data. Each table contains many columns some of them are CDR_DISTINCTION, MOBILENO, CALLEDCALLINGNO, TIME_OF_CALL & DATE_OF_CALL,DURATION_OF_CALL etc , also each database tables contains CDR data of type: HOME , ROAMERS DATA and OUTROAMER DATA , which means all the Home data, Rom data and Outroamer data will be stored in a common table with general name MOBILE_DATA_MONTH_YEAR_DAY , and in particular like MOBILE_DATA_JUL_2012_01.

In this approach loading of CDR data can only be carried after business hours i.e, not during day , because loading is directly carried out in main tables by making indexes unusable due to which no data can be extracted and if such loading process is carried out during business hours this will lead to business impact and with this loading process with the start of each day only Sysdate -1 old data can be made available, and therefore at start of each business day and if LEA's require the current day data to investigate any case , the telecom services providers are not able to make it available. Also using this approach large amount of space is required. In general in this scenario there will be 30(or 31) tables for a month , and about 360 tables for a year, 150 indexes for a month (if there are 5 indexes on a table) and 150*12=1800 indexes for a year which required a large table space also.

To overcome this disadvantage of loading architecture we propose a new database architecture using which 01-02 hour old CDR data can be made available and same has been confirmed by IT professional Sagar Sharma after implementing the proposed architecture in his organization. This is the real time solution which after my proposal is being used most of the telecom companies [1].

III. Proposed Architecture for Data warehousing in Telecommunications

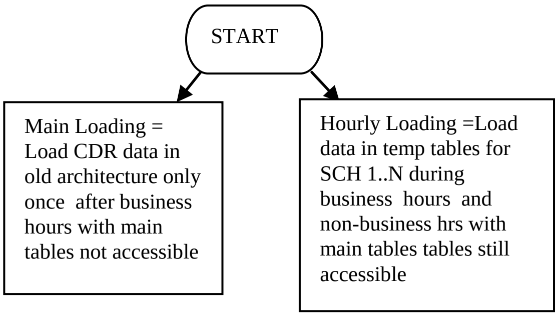
To overcome the drawbacks of the above structure we propose a database structure where we will have only 3 MAIN PARTITIONED TABLES, PARTITIONED ON THE BASIS OF DATE_OF_CALL, created for entire year, viz, Home table, Rom table and the Outroamer table with general names like

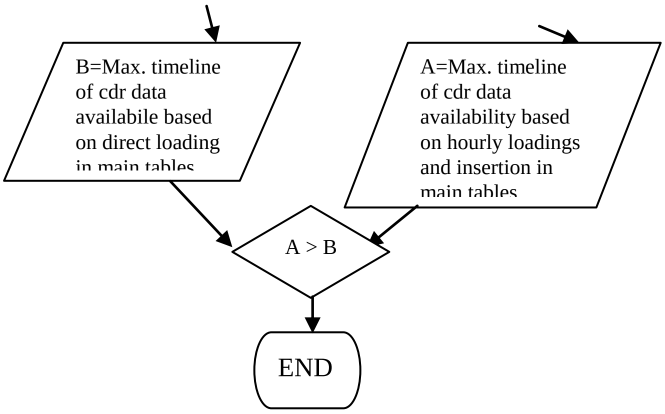
a)BASE_DATA_YEAR,
b) ONMOVE_IN_DATA_YEAR,
c) ONMOVE_OUT_DATA_YEAR) and three simple non-partitioned temporary tables for entire year each for home, rom and outroamer CDR data with general names like referred to as temporary tables:
d)TEMP_BASE_DATA_YEAR,
e)TEMP_ ONMOVE_IN _DATA_YEAR
f)TEMP_OUTR_DATA_YEAR.

Here we will have only 3 indexes each over main tables and temp tables. Following the said scheme will save a lot of table space as compared to the above said database structure:
a) As there will only be 18 indexes for entire year as compared to 1800 indexes,
b) As there will only be six tables as compared to 360 tables for a year.

Loading will only be carried out in the temporary tables as per the schedule during day time also , i.e , loading in temporary tables will be carried out more than once , (depending on the amount of data to be loaded and thus varies) depending on schedule fixed for a day .No loading will be carried out in the Main tables. In main tables only insertion of data loaded in temporary tables to main tables will be carried out during non-business hours , i.e at night .
3. After Insertion,the count of data inserted in the main tables will be calculated and compared with the count of data present the temporary tables and is counts are equal temporary tables will be truncated will be available for the next day's loading[3].
4. While loading in day time will be INPRG, in the temporary tables, the temporary tables will not be used in the extraction query because while loading data in temp tables indexes over temp tables are dropped and when the loading is NOT INPRG, the temporary tables will be included in the CDR extraction Query, because after loading indexes are recreated over the temp tables [2].

II. Flow chart of Proposed Solution :


START
Main Loading = Load CDR data in old architecture only once after business hours with main tables not accessible
Hourly Loading =Load data in temp tables for SCH 1..N during business hours and non-business hrs with main tables tables still accessible


B=Max. timeline of cdr data availabile based on direct loading in main tables
A=Max. timeline of cdr data availability based on hourly loadings and insertion in main tables
A > B
END

and this is what we wanted to achieve**.**

IV.GRAPHICAL REPRESENTATION:

CDR data to be loaded is always made available by the agencies in increasing order of time. So, if more data is loaded during a day then it will mean that timeline of data availability has been increased.
In the real time environment where loading was carried out with new architecture after replacing the old architecture it was found that the timeline of data availability has been increased with greater amount of data being loaded**.** We experimented with 30GB of data to be loaded and found that with the old architecture since no loading was carried out during business hours no data was loaded and therefore the entire day only one day old data was used. But while using the new structure since hourly loading was carried out in temporary tables multiple times in a day (we carried out hourly loading 2 times in a day) current day data was also available for the legal agencies. Using the old architecture 0 GB of data was loaded for current day , but with the new architecture 20 GB of data was loaded and made available for extraction within the business hours during 2 consecutive hourly loading schedules.

Source Table :

Table 1

Data Volume loaded	Size in GB
V1	0
V2	20

Where V1 = volume of data loaded with old architecture during business hours

And V2= volume of data loaded with new architecture during business hours

Graph:

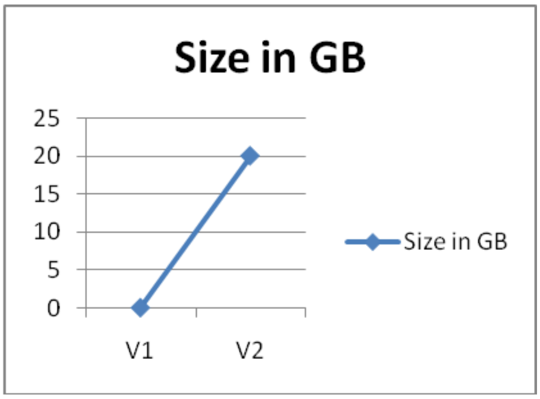


Fig 1.

V. Conclusion :

It was found that while using the new database architecture and without degrading the system performance during any time of business hours on any day below equation of performance existed, because it is clear that in the new structure loading was carried out multiple times during a day in temporary tables as compared to the original database structure where loading was carried out only after business hours .

(Max.time_line_of_CDR_data_availability_in_proposed_database_architecture_during_all_business_hours)> (Max.time_line_of_CDR_data_availability_in_old_database_ architecture _during_all_business_hours)

VI. SCOPE OF FUTURE WORK

In the above stated work, we have provided a better technique of CDR loading, using which CDR data of few hours old can be made available, but by using the proposed data warehousing real time CDR data availability was not made possible.

- **IMPLEMENTATION OF REAL TIME AVAILABILITY OF CDR DATA** as compared to few hours old data that has been made available to optimize the performance of database is one of the candidates of my upcoming PhD work.
- **AUTO ADDITION OF PARTITIONS FOR A MONTH:** Also, while using partitioned tables, we have to manually add the partitions for every month before the start of First date of month. Implementing the auto addition of partitions will also be considered in my PhD work.
- **POSTING OF DATA ON BASIS OF A SMS**: In the above logic while extracting the data from a database, someone have to be present to work on the database. Implementation of the logic where the CDR data can extracted from the server and posted to the email id of concerned authority of the Telecom Company just on sending a SMS by the authority to a short code, and not being physically present, will also be done in my PhD.

VII. INTENDED AUDIENCE

This paper involves practically applied and very advanced concepts of data ware housing. Also while writing this paper below section of IT industry was kept in mind as per below list :

- DBA
- Technical Leads
- Senior Developers
- Project Managers
- Research Scientists

VIII. REFERENCES

[1] Data Warehouse Performance Management Techniques, Andrew Hold sworth, Oracle Services, Advanced Technologies, Data Warehousing Practice. 2/9/96

[2] Jennifer Widom, Research Problems in Data Warehousing, Int'l Conf. on Information and Knowledge Management, 1995.

[3] P. Raj, Database Design Articles, Overviews, and Resources.

[4] Oracle® Database Performance Tuning Guide10g Release(10.2)Part Number B14211-03