Mystery of Missing Million Dollars

Overview:

A leading US Based Health Care Company gets Producer/Broker payment, commission, demographics data in the form of flat files from 5 different source systems and the data is loaded into their Enterprise data warehouse. This data is used by downstream systems, regional data marts, and business users for reporting, payment and balancing purposes.

Every month data is extracted in the form of flat files using ETL tool-Informatica and loaded into Data warehouse flowing through different stages. In each stage, cleansing of data is done and different business rules are applied as applicable.

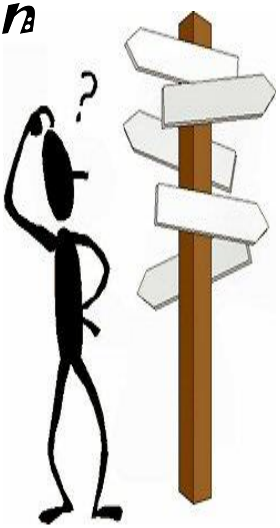
Tetrasoft Helps solve - Mystery of missing million dollars for US based Health Care Company

Challenge: "Missing $16Million"

After one of the monthly loads, an issue was identified in data loaded from few of the Source Systems. Producer/Broker payment amount was showing up as $27 million in the Enterprise Data warehouse but it was showing up as $11 million in Source systems. There was huge confusion all around related to the missing $16 million dollars.

Data has to be released for reporting & downstream extraction else it could cause a domino effect of impacting the SLA's throughout the Enterprise. But if data is released as is it would lead to inaccurate reporting of data from various regional data marts which pull data from Enterprise data warehouse and also cause balancing issues. Root cause for this discrepancy has to be identified immediately and data has to be fixed to avoid any SLA breach and to keep Company's repute intact.

Through diligent and thorough analysis Tetrasoft team identified that existing code converts single record into multiple records while allocating sequence number based on key columns when loading data from LZ to Staging Area. Due to unique sequence numbers allocated to these records they are not considered as duplicates while loading from Staging Area to Enterprise Data Warehouse and hence duplicates existed in the main Production tables. Upon further analysis it was identified that this issue existed with other Producers/brokers in historical data for few source systems. Solution was to clean up the existing data and to put in a fix to avoid such scenarios in future.



Analysis of Alternatives

To clean up existing data and eliminate duplicates below 2 options were considered.

Option 1: Fix the data by eliminating duplicates using Qualify Function

Option 2: Fix the data by eliminating duplicates using GroupBy /Having Condition with VT table through Sub Queries

Recommended Solutions

Opted for Option 1 as compared to Option 2 it is more performance effective and will have minimal impact on CPU utilization

Implementation

Historical data was cleansed by eliminating duplicates for all the Producers/brokers within few source systems where issue existed. Sanity checks were performed after data cleansing and data was released for reporting & downstream extraction in a timely manner so that any SLA's were not breached and accurate data was available for reporting & balancing purposes. A post load validation process was introduced to identify and delete any duplicate records to ensure this scenario is not repeated in future.

> "30% improved performance & reduced CPU utilization"

Business Value

- **D**iligent and critical Root cause analysis and fix ensured that accurate data is released for reporting & downstream extraction in a timely manner so that any SLA's were not breached.
- **B**y providing a permanent solution ensured that the issue doesn't resurface in future