Theoretical Analysis and Warehouse Design for Retail Return Fraud Analytics

02-19-2018

Anurag Jain

Faisal Ahmad

Mohsin Asif

Naren Sham

Smarak Das

University of Cincinnati, Lindner College of Business

– Data Warehousing and BI

Table of Contents

Executive Summary:

Theft, either by consumers or employees, is an ongoing concern for many retailers. This theft costs retailers nearly 1.5% of their total sales per year. There are multiple ways in which customers and employees commit fraud. SAM's Club is facing similar problems and in this report we are providing a possible to solution to proactively detect fraud. We have suggested an Enterprise Data Warehouse using Hub and Spoke architecture with two InfoCubes, Store Visits and Item Scan enabling swift and effective decision making. There are, however, some limitations to the structure we have suggested. It is not creating a unit of measure dimension for the currencies in all transaction tables because the geographies in consideration right now are only from the US and hence the currency is assumed to be in USD. Secondly, we do not have inventory tracking in this model for processing return. We discuss the possible solution to these limitations towards the end of this report. As a future improvement, we can automate the update of dimension tables that would enable near real-time analysis of daily transactions. The decision support system is currently being provided by SAP BusinessObjects which provides tabular data to the decision makers to view their metrics. This can be improvised by using third party visualization tools and connecting the data-marts through ODBC connectors, thus helping business users understand their numbers better and take effective decisions.

Introduction:

Retail fraud, also called as shoplifting, is a serious offence according to the US law. A customer or an employee, if caught and proven guilty, can be sentenced up to five years in prison or fined up to 10,000 USD. Despite this law, retail fraud is on the rise and accounts for 1.5% to 7% of all the retail sales putting a major dent in company income. A rise in theft suggest that retailers must proactively prevent theft even before it happens by taking appropriate measures. One of the steps companies could take is employing the power of business intelligence using cutting edge BI and warehousing tools such as SAP BW/BI for speedy analysis of daily transactions and returns data.

In this report we have analyzed the database populated with real sales data of a major named as SAM's Club. After analyzing the database, we suggest building a hub & spoke data warehouse design using SAP BW/BI. This data warehouse will feed our data mart, having a snow flake schema, that will contain important facts regarding daily sales and returns. We have defined the grain of the fact table after rigorous team analysis to address many tough business questions regarding fraud. We aim to deliver real time analytics to our customer Sam's Club, so its management can prevent fraud even before it causes major dent in their sales.

Literature Review:

According to our literature review, we found that retail fraud in the retail industry ranges from 1.5 to 7% per annual sales of an average company. This is a major concern for many retailers, but they still haven't taken serious steps to curb this menace. Today, retailers routinely find

themselves fighting against the manipulation of their financial statements, POS transactions, collusion among vendors, shoplifting and refund fraud, and many different other schemes. All reports, suggest that implementing analytics solutions especially related to data warehousing to analyze daily sales could drastically reduce theft in stores.

According to a recent Time Money article, the retail industry, in 2016, faced losses of nearly 50 billion USD due to retail fraud mainly because of customer and employees. In a survey on 83 major retailers who own multiple brands in USA, The National Retail Federation measured that inventory losses increased to 1.44% of the entire retail sales in USA. These losses measured in the survey included shoplifting, theft by employees, admin error, and vendor fraud. It was revealed that 36.5% of missing inventory was due to shoplifting by outside customers. Around 30% of inventory shrinkage was due to employee theft followed by Administrative errors accounting for 21.3% of inventory shrinkage; Vendor fraud accounted for 5.4%. The average cost of each shoplifting incident is 799 USD whereas the cost of each employee theft is estimated to be 1923 USD. Still many retailors did not take any action against such thefts.

In a similar article, Forbes reported that retailers in United States are losing $60 Billion a year to shrinkage. $57 Billion in 2014, this is an increase of nearly 5%. Additionally, employee theft was identified as the biggest reason for retail losses. American retailers generally put losses owing to staff ahead of losses owing to external shrink. The study was conducted on 91 retailers with annual sales totaling $844.6 billion.

All the major articles on preventing retail fraud emphasize on employing latest technology and tools to curb losses. Apart from installing cameras across the store, RFID scanners, and assigning

unique bar codes to each product, these articles emphasize on employing business intelligence and data warehousing tools to detect fraud. Some of the major vendor and their tools

Vendor	Tool
Teradata Corporation	Teradata EDW
SAP	BW/BI
Oracle	Exadata
Amazon Web Services	Redshift
IBM	Cognos
Informatica	Informatica Power Center, MDM

Theoretical Analysis:

Although there are many ways fraud is committed in a retail store, but the major ones are committed by customers and store employees. In this section we will discuss various customers and employees commit fraud.

Customer Fraud:

Risk	Mitigation Strategy
Wardrobing: This is the phenomenon where customer buys a product with an intention to return the product for full refund after using it. For instance, a student may buy a book for an open book exam and return it after the exam.	▪ An item must be returned to same store where it was purchased within 7 days of the purchase with original receipt and product tags ▪ A customer may not return more than a fixed number of times within a specified time

Return of stolen Items: Customer may legitimately buy a product from a store and return to the store with an original receipt to grab another product of the same specifications and return it for cash.	Assign unique product SKU to each product and track returned item against that code.
Receipt fraud: Customer may use a legitimate receipt of some other customer and try to get a refund.	Returned without member card and the customer who is trying to return on another receipt will have to bring the same card it which sale was made
Returning product to the wrong store: Customer may buy a product from a cheaper store and try to return it to a different store for higher price refund.	Item must be returned to the store where it was purchased from
Credit Card Fraud: Customer may buy a product from a stolen credit card and try to return it for cash refund.	Return will be issued to the same credit card it was purchased from and there will not be any cash refund without verification from the owner of the credit card.

Employee Fraud:

Risk	**Mitigation Strategy**
Discount Fraud: When an employee misuses his power to use discount codes/coupons to sell items to his friends/family or sells items at higher prices online.	Number of discounts given by an internal employee should be limited to a certain number.

Not recording a return: When an employee does not intentionally record a return and steals the item while money goes out of company in refund.	Money cannot be paid out unless the return is recorded into the system
Price modification: May misuse his rights to lower the price of the product to buy it and increase price later to the usual price.	Price changing must be validated by two separate employees and proper rights should be allocated to prevent unauthorized price change
False return: Employee enters a false return and pays out money to someone when no actual item is returned.	Return must be processed at a separate counter and must be reconciled with the physical item at the end of the day

Database Analysis:

Whenever a new customer visits a store and checks out at the cashier, a transaction is generated and is stored in a normalized database. The aim of this database is to implement data integrity and efficient storage of data. These databases are designed for fast storage of data and avoid redundancy. To attain this, redundant data is split in to a different table and are related to each other using foreign keys. Although this database architecture is good for capturing data efficiently, it is not good for performing analysis on historical data. The select performance is poor thus we need to convert this database into a flat structure which contains less relations and avoids multiple joins. For this reason, dimensional model is preferred for running analysis.

To convert our normalized model to a dimensional model, we analyzed which columns are getting populated on every transaction and which are not. The ones which are growing with every transaction are used in the fact table and the ones which are categorical are used in dimension

tables. We aggregate the fact values to get summarized results. This reduces our effort of aggregating during querying for analysis and gives more meaningful insights. Data from different source systems are accumulated and cleansed. Finally, analysis is performed on this aggregated model.

Data Warehouse Architecture:

We suggest employing the Dependent Data Mart and Operational Data Store (Hub-and-Spoke) architecture. Here, we source our data from the various OLTP (Transactional) systems and stage the data. We then perform our transformation and cleansing of the data. The transformed data is moved to the Enterprise Data Warehouse (EDW), where it is stored in a uniformly consistent structure. The data in the EDW is stored with historical data and the it can only be updated here. The EDW is the source for our data marts which is connected to end user presentation tools for decision making support.

We use SAP BW for our cleansing, ETL, data warehousing and data mart creation. In our design, we have two target Infocubes – *Visit* and *Item Scan*, which act as our data marts. The Infocubes have conformed dimensions of member, operator, items, brand, category and store between them. The *Visit* Infocube contains the item details of each member visit and the *Item Scan* contains the details with granularity of each item for every member visit. The development method starts with loading the master data and dimension tables (characteristics) initially and then loading the transactional data on a regular basis. The master data is initially defined and their text and attribute values are set. This data is rarely updated, in case of new data definitions.

The transactional data is loaded daily at the close of business from the source systems. A job is run post the cleansing and transformations performed in the PSA. A process chain can be scheduled to load the transformed data into the EDW (Data Store Object). The data is stored indefinitely in the DSO with the characteristics linked to the corresponding transaction table fields.

The data marts get their inputs from the EDW and are used for performing analysis and specific business questions about the OLTP data at hand. Here, we use a visit data mart which analyses the daily visit data, the operator performing the transaction, the member visiting, the store and the sale details of the transaction. The item Infocube provides analysis about the item level details, such as each item bought, the member buying and item cost.

A drill through analysis of the data marts can be performed by creating an additional data mart from the data marts, to answer questions related to each item within a visit by a member (shown below). The facts are both recording data at the granularity of a second. The data in the dimension tables are to be rarely updated. They are updated on an incremental basis every month, or by a need to update basis, if an additional product is added into the catalog, or a new employee joins the ranks. The transactional data, as mentioned, is updated daily from all the available source systems during non-business hours such as middle of the night. This is because the daily transactional data needs to be available for analysis at the data mart as soon as possible to enable decision making.

Larger Infocubes are accommodated by the creation of *multiproviders* available in the SAP BW environment. As they store, they correspond to the logical data mart which only store the queries or views required for analysis. This helps in saving storage space and allows ad-hoc analysis.

The summary of our architecture is provided below:

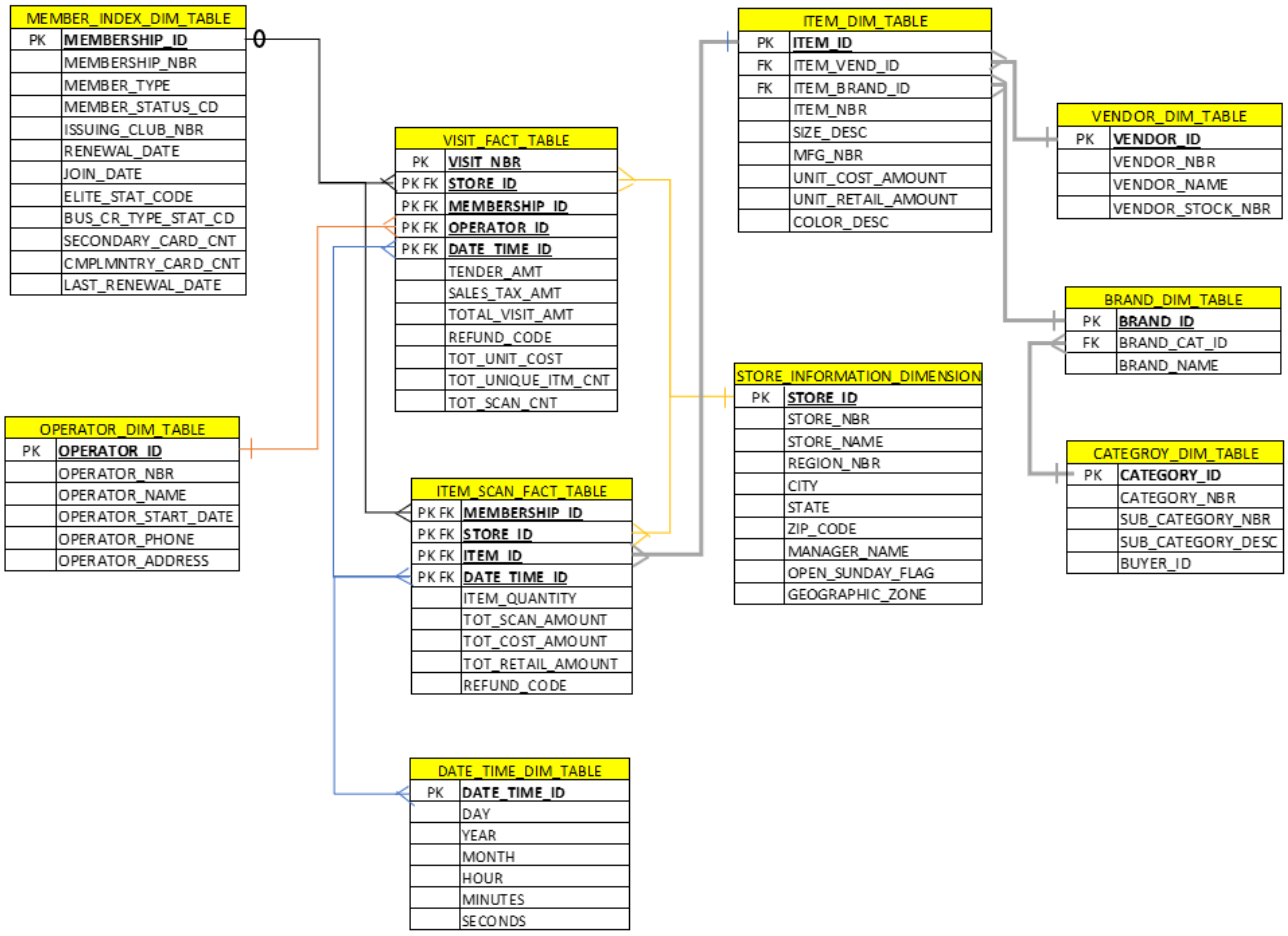
- Hub and Spoke Architecture with a staging layer and enterprise data warehouse
- Subject oriented Infocubes created for analysis with a snowflake structure, conformed dimensions and uniform granularity.
- Dimensions updated monthly and daily incremental refresh of transactional data.
- Provision for additional infocubes on top of the visit and item infocubes for drill through analysis
- Process chains to enable serial updating of data to the infocubes
- Provision of logical data marts in the form of multi providers

DW Model

We have proposed a two fact tables dimensional model with total seven dimensions. The two fact tables are "VISIT_FACT_TABLE" which captures the details of a store visit and "ITEM_SCAN_FACT_TABLE" which stores the details based on Items purchased. "ITEM_SCAN_FACT_TABLE" and "VISIT_FACT_TABLE" is connected to four dimension tables namely "MEMBER_INDEX_DIM_TABLE" which stores the details of every store member, "OPERATOR_DIM_TABLE" stores Operators details present at the store, "DATE_TIME_DIM_TABLE" which stores the conformed date and time dimension of transaction and "STORE_INFORMATION_DIMENSION" that captures the detailed information of that store. The other fact table "ITEM_SCAN_FACT_TABLE" is connected to "MEMBER_INDEX_DIM_TABLE", "DATE_TIME_DIM_TABLE", "STORE_INFORMATION_DIMENSION" and "ITEM_DIM_TABLE" which has all the details about the purchased item.

The “ITEM_DIM_TABLE” is further connected to the “VENDOR_DIM_TABLE” which has the vendor details for every item “BRAND_DIM_TABLE” and “CATEGORY_DIM_TABLE” which store the brand and category details

We have implemented the “Hub and spoke model” with “type 2” changes architecture with snowflaking and hierarchical schema structure.



Using this model, we can answer below business questions:

Question	Response
Are there certain stores that have more returns than others?	This can be found by referring to STORE_ID in the ITEM_SCAN_FACT_TABLE, we can analyse for stores with a high count of refund code
Are there certain times that are more likely to have returns?	To find certain times that are more likely to have returns we can analyze the ITEM_SCAN_FACT_TABLE for DATE_TIME_ID which can help us find times with highest number of returns
Are there certain employees who process more returns?	If there are certain employees who accept more returns, then they can be found by analysis of VISIT_FACT_TABLE with high OPERATOR_ID count
Are there certain customers who return more items?	This can be found by looking for high MEMBERSHIP_ID count against returns in the VISIT_FACT_TABLE
Are certain products returned more often than others?	These questions can be answered by querying ITEM_SCAN_FACT_TABLE and its dimensions.

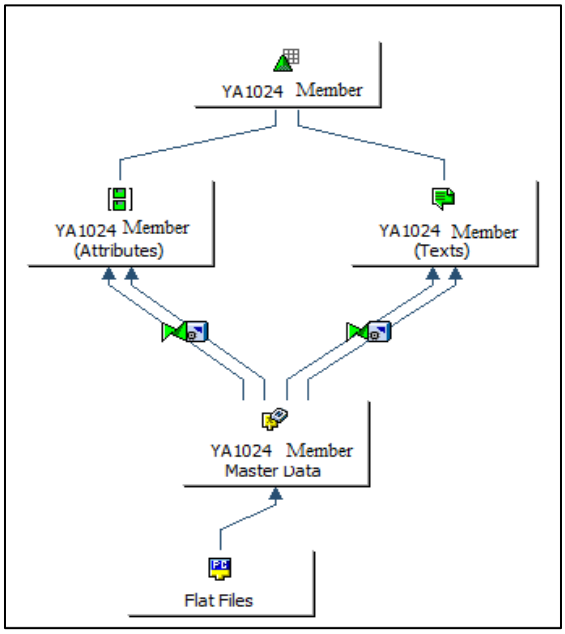
Answering these questions will help us in taking actions and making strategies to avoid fraud.

Key Figures and Characteristics

The Infocubes, Info-objects, Key figures, and characteristics that Sam's club data warehouse will be using are attached in the *Appendix A* of this report.

ETL Process for Sam's Club EDW

- **Master Data Flow:** The master data of each dimension must be defined through the SAP BW Master data flow. The Member Index Dimension is created by loading its characteristics Texts and Attributes as referred from the UA_SAMSCLUB_SMALL tables.



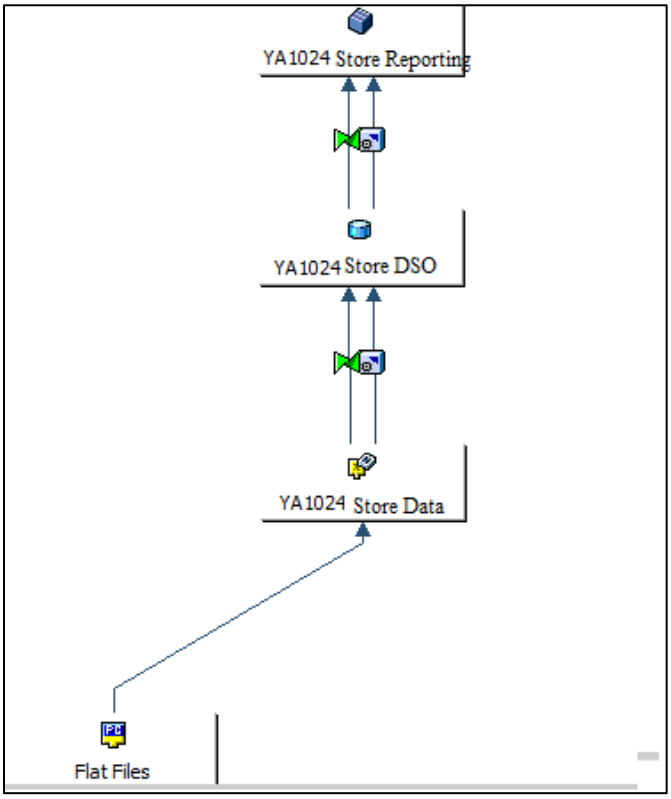
The primary steps for creating the Master Data Flow are as follows:

- **Create Data Flow (Extract)**: The data flow is created from the extracting of transactional database tables from the Sam's club database.
- **Add Member InfoObject:** The member InfoObject is mapped to the flow and used for Metadata management.
- **Create Data Source:** The data source is created and scrubbing of the data by removing NULL values is done in this process.
- **Create Transformation for Attributes:** The attributes of Member Index and different dimensions are transformed and mapped in this flow.
- **Load and data transfer:** The attributes data from the transactional database tables are loaded into the dimensional tables.

- **Create Transformation for Texts:** The textual fields are transformed and mapped in this flow.
- **Load and data transfer:** The textual data from the transactional tables are loaded into the dimensional tables.

Transactional Data Flow

The proposed EDW architecture contains two Fact Tables, Visit Fact and Item Scan Fact. These two fact tables should be created as two InfoCubes for analyzing and reporting the data using the above dimensional and transactional table data.

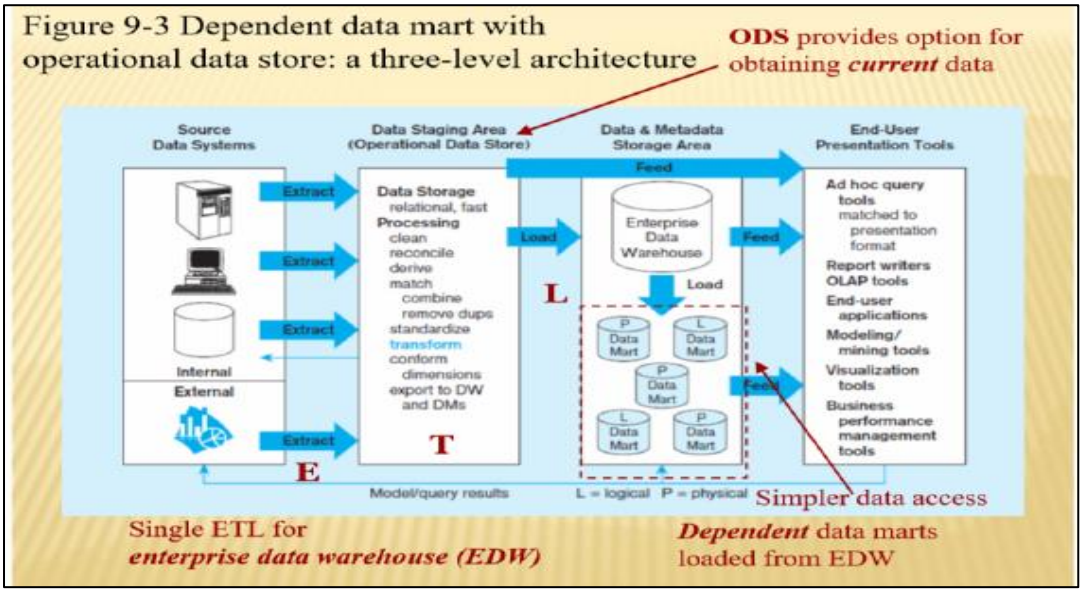


The primary steps for created the Master Data Flow are as following: -

- **Create Data Flow (Extract)**: The data flow is created from the extracting of transactional database tables from the Sam's club database.

- **Create Data Source:** The data source is created and scrubbing of the data by removing NULL values is done in this process.
- **Create InfoPackage and load PSA:** The InfoPackage is created and data are loaded from the PSA.
- **Create DataStoreObject:** The DSO of each fact tables are created for storing the data of each Store visits and Item scan fact tables.
- **Create Transformation:** The transactional table fields are mapped and transformed using this step to each fact key figures and characteristics.
- **Load and data transfer to DSO:** The data from the transactional tables are loaded into the DSO.
- **Activate Data in DataStoreObject:** The DSO are activated using this step.
- **Assign InfoCube:** The InfoCube which is created for each Fact Table, Store Visit and Item scan are assigned to the DSO flow.
- **Create Transformation:** The transformation from DSO to Infocube of both Store visit and Item scan is performed in this step. Total Cost and Retail is calculated by Unit price and quantity.
- **Load and data transfer:** The data is then loaded from DSO to Infocubes in this step.

Hence, from each of the above steps we can conclude the ETL processes used for the EDW:



- **Extract:** The Data flow creation in Master data and transaction data flow from the transactional data source of Sam's club database table.
- **Metadata management:** The creation of InfoObject and creation of new Data source is the step where the Mapping of the metadata is performed.
- **Scrub and Cleanse:** The step to create the data source and load the data from the transactional tables also involves the data cleansing step where null and empty values are removed.
- **Transform:** This step is performed during the attribute mapping in Master data and DSO to InfoCube transformation flow where Quantity and UnitCost or UnitRetail is used to calculate the total cost of each item scan.
- **Load:** This step is performed in both Master data creation and Transactional data flow to start the data transfer process where the data from DSO is loaded into the InfoCubes.

Discussion and Conclusion

Our suggestions will enable SAM's Club to create an Enterprise Data Warehouse using Hub and Spoke architecture with two InfoCubes, Store Visits and Item Scan enabling swift and effective decision making enabling SAM's club to answer tough business questions and detect fraud as quickly as possible. There, however, are some limitations to the structure we have suggested. It is not creating a unit of measure dimension for the currencies in all transaction tables because the geographies in consideration right now are only from the US and hence the currency is assumed to be in USD. This limitation can be overcome by adding a unit dimension in case the business expands geographically. Also, the addition of Inventory tracking would enable a more

robust analysis of Retail Return Fraud, both from the customer and employee standpoint, which has been provisioned for in our Data Warehouse model.

For future improvement on the current design, the system can be integrated with automation processes. An automated update of dimension tables would enable near real-time analysis of daily transactions. For instance, the addition of an operator as soon as they are onboard into the dimension table. Similarly, the addition of a new product on its release to the market will enable an automated update to the item dimension table. The decision support system is currently being provided by SAP BusinessObjects which provides tabular data to the decision makers to view their metrics. This can be improvised by using third party visualization tools and connecting the data-marts through ODBC connectors, thus helping business users understand their numbers better and take effective decisions.

APPENDIX – A

Store visit info cube

Key figures

SOURCE FIELD	**InfoObject Description**
VISIT_NBR	The Visit Number generated for each visit
STORE_ID	The Store ID of the store
MEMBERSHIP_ID	The membership ID of the customer
OPERATOR_ID	The employee ID of the store
DATE_TIME_ID	The date and time of transaction
TENDER_AMT	The total tender amount
SALES_TAX_AMT	The sales tax amount
TOTAL_VISIT_AMT	The total visit amount
REFUND_CODE	The Refund code
TOT_UNIT_COST	The total unit cost
TOT_UNIQUE_ITM_CNT	The total unique item count
TOT_SCAN_CNT	The total scan item count

Dimensions

DIMENSIONS	**Description**
MEMBER_INDEX	This dimension stores customer info
OPERATOR	This dimension stores employee info
DATE_TIME	This dimension store transaction data time
STORE_INFORMATION	This dimension has the store info

Item_Scan Info Cube

Key Figures

SOURCE FIELD	InfoObject Description
STORE_ID	The Store ID of the store
ITEM_ID	The Item Id scanned in each visit
MEMBER_ID	The Customer Id of the member
DATE_TIME_ID	The date and time of transaction
ITEM_QUANTITY	The total number of items
TOTAL_SCAN_AMOUNT	The total scan amount
TOTAL_COST_AMOUNT	The total cost amount of item
TOTAL_RETAIL_AMOUNT	The total retail amount of item
RETURN_CODE	Return Code of the item

Dimensions

DIMENSIONS	Description
STORE_INFORMATION	This dimension has stores info
DATE_TIME	This dimension stores date info
MEMBER_INDEX	This dimension store member info
ITEM	This dimension stores of Item info

Store_Information Dimension Characteristics

SOURCE FIELD	InfoObject Description
STORE_ID	The Store ID key of the store
STORE_NBR	The store id of the store
STORE_NAME	The Store name
REGION_NBR	The Region number
CITY	The city
STATE	The state
ZIP_CODE	The zip code
MANAGER_NAME	The manager name
OPEN_SUNDAY_FLAG	Open Sunday flag
GEOGRAPHIC_ZONE	Geographic zone

Date_Time Dimension Characteristics

SOURCE FIELD	InfoObject Description
DATE_TIME_ID	The date time ID
DAY	The calendar day
YEAR	The calendar year
MONTH	The month
HOUR	The hour of transaction
MINUTES	The minute of transaction
SECONDS	Seconds of transaction

Item Dimension Characteristics

SOURCE FIELD	InfoObject Description
ITEM_ID	The item id
ITEM_VEND_ID	The vendor id
ITEM_BRAND_ID	The brand id
ITEM_NBR	The item number
SIZE_DESC	The item size
MFG_NBR	The manufacturing number
UNIT_COST_AMOUNT	The unit cost
UNIT_RETAIL_AMOUNT	The retail cost
COLOR_DESC	The item color

Member_Index Dimension Characteristics

SOURCE FIELD	InfoObject Description
MEMBERSHIP_ID	The membership id
MEMBERSHIP_NBR	The customer number
MEMBER_TYPE	The membership type
MEMBER_STATUS_CD	The member status
ISSUING_CLUB_NBR	The issuing club id
RENEWAL_DATE	The renewal date
JOIN_DATE	The join date

ELITE_STAT_CODE	The elite status
BUS_CR_TYPE_STAT_CD	The business credit card status
SECONDARY_CARD_CNT	The secondary card count
CMPLMNTRY_CARD_CNT	The complementary card count
LAST_RENEWAL_DATE	Last renewal date

Operator Dimension Characteristics

SOURCE FIELD	**InfoObject Description**
OPERATOR_ID	The operator id
OPERATOR_NBR	The employee number
OPERATOR_NAME	The employee name
OPERATOR_START_DATE	The employee start date
OPERATOR_PHONE	The employee phone
OPERATOR_ADDRESS	The employee address

Sources:

http://time.com/money/4829684/shoplifting-fraud-retail-survey/

http://www.monitis.com/blog/top-5-data-warehouses-on-the-market-today/

http://deloitte.wsj.com/cio/2014/03/17/using-analytics-to-detect-retail-fraud/

https://www.forbes.com/sites/nicoleleinbachreyhle/2015/10/07/new-report-identifies-us-retailers-lose-60-billion-a-year-employee-theft-top-concern/#44d706c080eb

Decision Support and Business Intelligence Systems 9th Edition by Turban, Sharda, Delen