ISDS 556 - DATA WAREHOUSING AND FOUNDATIONS OF BUSINESS INTELLIGENCE

TEAM 1 - DW/BI SYSTEM DESIGN PROJECT

Team Members:

- **Abhinay Sariswal**
- **Aditi Adhik Patil**
- **Aishwarya Chavan**
- **Amit Mane**
- **Ankita Jaiswal**

Overview

The project is about building a data warehouse/business intelligence solution for a software firm. The data warehouse or business intelligence solution can change the whole business process in an organization. Data warehousing coupled with business intelligence provides an excellent opportunity to transform operational data into a more useful resource. The solution will be used to integrate the data from the operating system and external data sources. It will also help managers to make informed business decisions. It will give a broader view and in-depth knowledge of the status of the organization. The solution will also help to integrate several department data based on revenue, profit, and market value. The project's success will bring a measure of effectiveness to the company which will be useful for analysis and decision making.

Company Description

The organization for which the solution is to be developed is a document capture and content management company. The company delivers its technology on a private or public cloud platform that turns the unstructured content into actionable information. The company has over 1500 customers globally. The company of this firm vary from large organizations, government agencies, small and medium business, wholesale and residential customers. Company has high enterprise clientele as compared to residential consumers and small-medium businesses.

Need for BI

The company has a broad channel of networks from partners which forms a basis of their external data. They also have a CRM system which stores customer and lead information, transactional and sales data. The existing CRM system is an elementary legacy system which is limited to running basic preconfigured reports. The reporting limits are strict and can be customized minimally. The operational data within the organization is unstructured and hard to interpret. The management is in need of a BI solution as it will give a strategic perspective for making business decisions. Adopting a BI solution will help to rationalize the large volume of data residing in the company systems into advanced reports that will deliver insights. These insights can form the basis of predictions which will allow the company to prepare for the future.

KPI

The BI system will be developed primarily for the sales process. In future, the solution can be deployed for the rest of the processes too. The reason for choosing this process is because the CRM system connects the leads or prospects from marketing, invoices/transactions from accounting to the customers. The customer information consists of their details, opportunities, quotes tied with the sales department. The primary KPI that will be focused on is the revenue. Revenue is one of the critical factors that drive the performance of the organization. Revenue is a key metric for any organization to monitor since it is an essential part of growth projections and is instrumental in strategic decision-making. Monitoring this metric over multiple time periods will give a clear indication of growth trends.

BI Users

The management or the senior executives will be the primary users of the reports since their goal is to maximize their earnings by expanding and improving their business. The critical reports can be scheduled periodically on a weekly basis to check the live status of the customers and on a monthly basis to check the performance over time.

The reports will be generated using the proposed BI system and will include various data visualization. The reports will showcase historical sales trends, revenue over a period of time, profit earned, the market value for the investment, customers which form a major portion of revenue and so on. Since the senior management is well aware of the benefits that analytics can provide to improve strategic decision making, revenue generation, and customer satisfaction, they are willing to invest time and money.

High Level Enterprise Bus Matrix

The high level enterprise bus matrix consists of vertical list of business process and a horizontal list of dimensions as shown in the Figure 1 below. The column header represent the dimensions and the row headers are the events on which the fact tables will be changed. The cross mark represents a relation between fact and dimension.

	Dimensions									
		Customer		Product			Employee			
Business Processes	Date	Customer	Customer Type	Product	Category	Vendor	Employee	Region	Product domain	Purchase details
Sales	X	X	X	X	X	X	X	X	X	X
Orders	X	X	X	X	X		X	X	X	
Training and setup	X	X	X	X	X		X	X	X	X
Client requirement analysis	X	X	X	X	X				X	
Demand Forecast	X		X	X	X	X		X	X	
Sales Forecast	X		X	X	X			X	X	
Revenue	X	X	X	X	X	X	X	X	X	X
Profit	X		X	X	X	X	X	X	X	

Figure 1 - Bus Matrix

Data Modelling

To meet the requirements of the project, we are using the star schema model. The model consists of a Sales data fact table and various dimensions. The fact table includes various attributes that are required to analyze sales data with the minimum number of joins. That, in turn, would improve the performance when querying the data warehouse.

The goal of the project is to provide a system which will help the user to analyze the overall company sales, product, employees and customer data as needed. With this model, the user can query that requests the total sales and quantity sold for a range of products in a specific geographical region for a specific time period can typically be answered in a few seconds or less regardless of how many hundreds of millions of rows of data are stored in the data warehouse database. Moreover, sales fact table contains information about service provided with the sale, order details, lead, Salesperson, date, time, account, etc. In accordance with the data model, future value for the company sales can be forecasted to increase the revenue of the company e.g., expected gross revenue, costs, expected purchasing frequency, probability of gaining additional revenue, etc.

The type of schema considered for this model is transaction type. This schema covers the most atomic level of sales transaction of the system. The lowest level of grain considered is the fact table recording every item in every sale transaction. Here, Grain = each sale that happened and Fact = Units sold, Number of sales (=1)

The reason for selecting the star data model is because it provides simplicity and high query performance to the data warehouse. It also reduces the time required to load large batches of data into the database; and it enforces referential integrity. The data model is denormalized to avoid complex joins in queries, hence, snowflaking is not used. In this data model, the only degenerate dimension is Sale_ID which is present in the fact table. There are several slowly changing dimensions in this model. Product price changes over time, Type I SCD, Customer details such as the age, gender, salary, marital status, number of children change over time. The slowly changing dimension workhorse technique, type II, is inappropriate for tracking attribute changes in large dimension tables with millions of rows. Therefore, the more frequently changing customer attributes are placed into their own separate dimension table i.e. a mini-dimension. In the model, we can see that the new mini-dimension contains customer demographics. The mini-dimensions grain is one row per demographics profile, whereas the grain of the primary customer dimension is one row per customer.

The time dimension family consists of dimensions such as hours, minutes, seconds etc. The date dimension family consists of day, month, year, quarter, etc. Our KPI being revenue, it is generated monthly, weekly or daily which is associated with time and date. Hence, from business point of view, if management wants to analyze current sales and previous sales, this can done easily by trivial selection process using the date and time dimension.

Below diagram represents a logical dimensional model designed for handling of all dimension families:

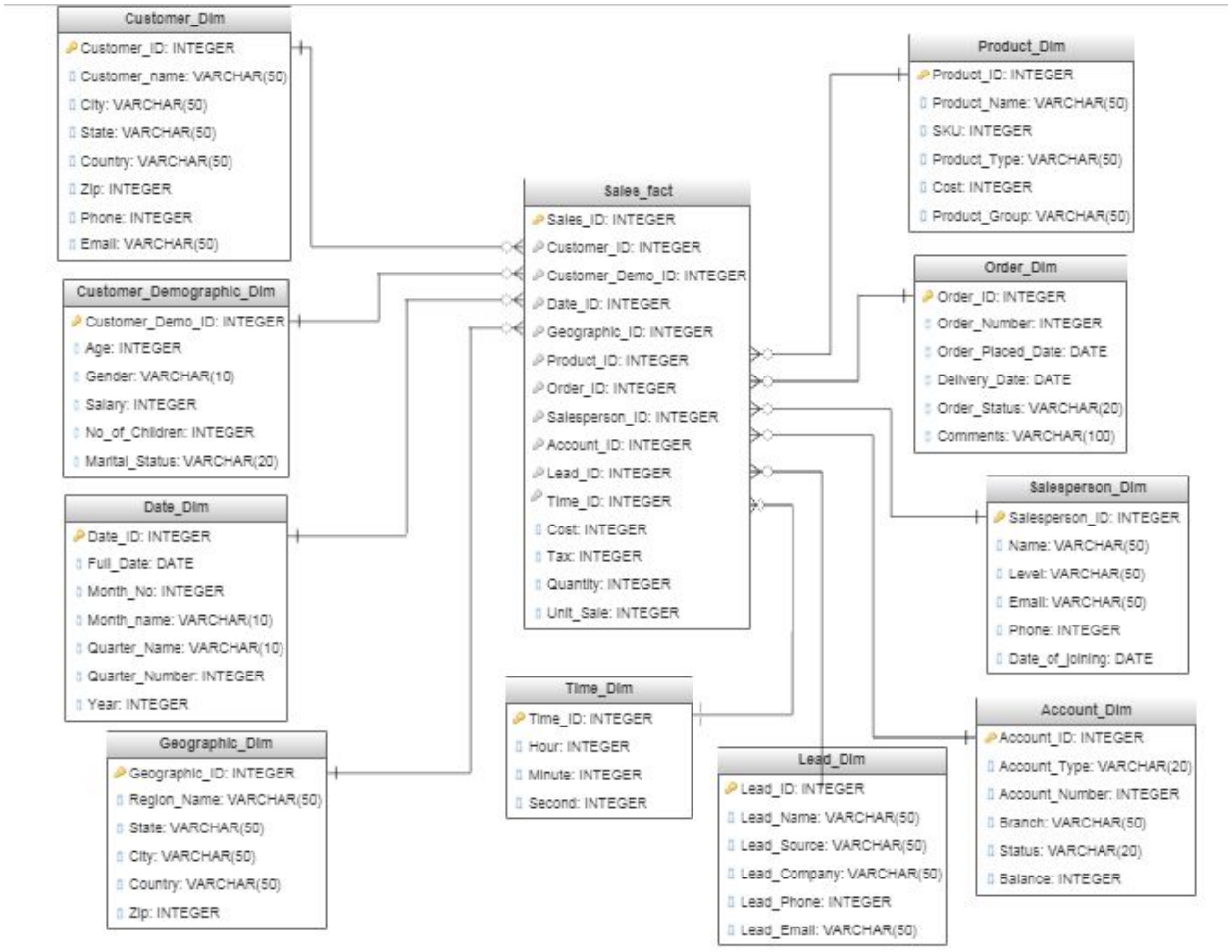


Figure 2 - Star Schema

Data Analysis

Data analysis is one of the most important steps while building a data warehouse/business intelligence system for an organization. First, it is important to decide and finalize the purpose of the data warehouse and how it will be beneficial to the organization.

Our primary aim is to develop and build a data warehouse system for a software firm which will help the firm to manage, organize and transform its data into more resourceful information which can be used to unfold many facts.

The firm has different business processes of which the data warehouse will manage the below-mentioned data

- Sales
- Orders
- Training and setup
- Client requirement analysis
- Demand forecasting
- Sales forecasting
- Revenue
- Profit

We also have data of different dimensions like customers, products, and salesperson which are as follows:

- Customer will consist of Customer, Customer Type
- Products will consist of Product Category and Vendor
- The Salesperson will contain details about Employee and Region
- Along with these, we will also store the Date, Product Domain, and order details.

The data will be stored at the most atomic level with the sales and orders information recorded individually for all the transactions. The grain of the data is One fact table per day per individual product per each customer and every item in every sale transaction at the lowest level of grain while recording individual transactions and every item in every sale.

The firm has three different sources of data

- Internal Data (CRM)
- Legacy Data
- External Data (New Prospects)

All the data sources are credible as the data collected from CRM, and Legacy data are in-house collected data, and the external data is collected from leads generated by the sales people and other firm associated.

Source System	Source Attributes	Table Name	Measure of Interest
CRM	Customer	Customer_dim	Measure total sales for every customer
CRM	Product	Product_dim	Generate total sales reports for individual product from different categories
Legacy Data	SalesPerson	SalesPerson_Dim	Measure total sales for individual employee and performance
Legacy Data	Region	Region_Dim	Measure total sales for individual region and performance and tax calculations.
External Data	Client Requirement analysis	Lead_Dim	Use New prospects to analyze and find client requirements to acquire new clients in the future
CRM	Sales	Sales_fact	Calculate total sales to generate quarterly reports and forecast reports

Figure 3 - Source System Mapping to Measure of Interest

This data will be used by the BI applications to generate different reports which will be used to find different facts and information by analyzing the data. This will help in generating future forecasts and demands for the firm to allocate its resources to gain maximum market share in the industry and target all the customers in the market. The data will be acquired from different sources to generate reports and forecasts, but the data will have unwanted and unnecessary information which has to be removed and the data cleaned before producing any of the reports and forecasts.

As a part of building a data warehouse we have to maintain metadata that contains structure and source of raw data, data about the data, the data model and rules for replication, distribution, exception handling and any other details required for mapping the data warehouse its inputs and outputs. As the system will become more complex, the amount of data will grow and more complex data analysis will be done. Powerful tools will be required to analyze the data, maintain the data warehouse.

ETL

As the area for real-time data inclines more on the extreme side; the endless flow of new data in a large amount and real-time reporting, traditional ETL system requires a lot of work and remodeling to support a transaction of relational data into a dimensional model. The goal of our ETL process is to produce high-quality data and identify different types of dependencies exist in the data so that the data can be easily integrated into a data warehouse.

The ETL team has the following objectives to meet this goal:

- **High availability:** data streaming as the name suggests should always be available in a constant flow. Data distribution and replication are essential to ensure malleability of the data collection process, data loss should be recovered and always available whenever required.
- **Low latency and reliability:** ETL must ensure reliability to produce correct and consistent data and speed required to provide new data to meet the business needs.
- The ETL must **ensure scalability** for improving the performance and mitigating the risks.

The ETL team can achieve this goal step by step and built the facts and conformed dimensions by using the bus matrix designed by the team.

Below are some of the issues and its solutions which may be encountered during the ETL process:

- **Data cleaning/inconsistency issues** may arise if the mappings are not proper and complete. The aggregations should be carefully tested and evaluated to ensure the data integrity. In order to mitigate the major issues encountered in the live database, we can plan to have a backup server to ensure that the business does not suffer highly. This will help the company when there is a major loss of functionality or critical data issues. It can act as a high availability server as well to deviate the traffic if users are more in future.
- **Multiple data stores:** We can use data integration tools from Microsoft BI Stack instead of manual ETL process execution to ensure that the data from multiple sources such as CRM, legacy system, lead generation system can be efficiently handled.
- **Performance loss during the ETL execution**: Customer management should be notified and the user base should be communicated accordingly to inform expected delays in performance during the ETL job execution. To mitigate these issue we can design staging tables for performance consuming tasks to reduce the time of job execution.

Some of the transformation issues that we can encounter with data are discussed below

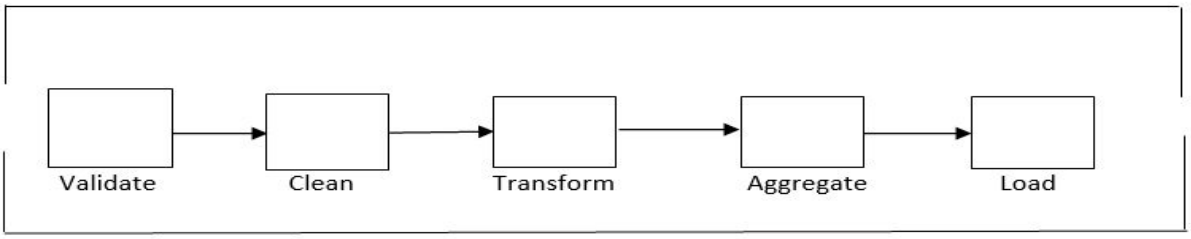
.

- **Record level problems:** We see this problem when there are multiple instances of the same record with variety of information coming from different sources. The duplicate records can be handled within the ETL process before the load or we can also handle it after it has been loaded into the warehouse. For example, we can run merge process to avoid data redundancy after the ETL process is completed. The data aggregation should also be carefully handled.
- **Value problems:** Value problems may arise when interpreting data from different sources. We should ensure that values from different sources have been standardized before the aggregation. If data is loaded or aggregated without proper referencing or formating the values will not be accurate and the reporting may fail.

The Extract, Transform, and Load (ETL) process for company's data warehouse will be performed to extract data from multiple sources such as a firm's data stores, several kinds of documents or web pages. Then, the extracted data are transformed, cleansed and loaded to the warehouse. Such transformation includes checks and filters to ensure the data adequately propagated to the warehouse concerning business rules and regulations as well as the normalized data will be denormalized into a schema to fit the data into the target data warehouse.

The following are the steps for the ETL:

1. **Extract Source Data:** ETL system extracts the data from transactional sources such as CRM or legacy database and load them to the staging schema. Staging is useful to illustrate the framework and structure of the source data.



ETL

2. **Transform Dimensions**: During this stage, dimensions are created and their surrogate keys are used to populate the fact table. Before loading the dimension tables, data cleansing and validation checks are performed to ensure a sanitized extracted input data. Validation check are of utmost priority when it comes down to maintaining data integrity. The journey of the data from the source system to the target system must be thoroughly tested. The transformed values should be tested for the expected data values. Data type, length and constraints must be checked

in the metadata. The focus during the validity testing must be on the completeness and the accuracy of the data. The data transformation should be tested using multiple SQL queries covering all scenarios. ETL must be tested for incremental data changes. Make sure no spam data is loaded and check if all the keys are in place. Data integrity results should pass the expectations after analyzing the inserts and updates added on top of the old data. The final check is to test the BI web application to ensure that the navigations and system functionality are not hampered, and everything is working as expected after the sanity test.

For transforming data into the dimension tables, the ETL team perform set of customized operations on the data. For instance, from the sales fact table, we can perform product level aggregation to calculate total revenue or total product sales.

3. **Load into Dimension:** The last step of the ETL process will be to load the data into star schema as facts and dimensions of the target data warehouse. Due to the large volume of data, loading process will be carried out in the night in order to optimize the performance. To handle load failure, ETL team will configure load recovery mechanisms to restart the process from the point of failure without any data loss.

The time for scheduling an ETL is a very critical decision to make. It is preferred to be scheduled during that time of the day when the concurrent active user count in the database is low. It is a time consuming process when the database has big data and needs to be monitored due if there are more number of tasks in an ETL. It impacts the performance of the system and SQL transactions due to the update process of the ETL server. When the data is in the loading phase of the process the warehouse tables with the update process gets locked and may take time if the end user is trying to update the same. The memory and CPU utilization of the ETL server needs to be monitored along with that of the warehouse database. Speed of transactional processing in the database might be slower during the ETL process. Depending to the business hours of the affiliates the ETL should be scheduled at such a time that the business end users are least impacted. Based on the decision of how much time lag the business users can bear to serve the business reporting needs, the frequency of scheduling needs to be agreed upon. We can do it twice a day, but we prefer to run the entire ETL process once a day. The time duration consumed by the ETL process needs to be analysed after the go live to further decide if the frequency needs to be changed. It totally depends on the business reporting requirements and urgency. However, during emergencies specific ETL tasks can be run separately to achieve a specific reporting need to view the data as close as possible to the live OLTP (online transaction processing) database.

Architecture

Various transactions are performed on various systems. The data warehouse architecture shown below is divided in four layers viz, Data Source Layer, ETL or Data Integration Layer, Presentation Layer and Application Layer. The components in these layers are performed by Microsoft Business Intelligence (MS BI) Stack . The MS BI Stack has tools like SQL Server Integration Services (SSIS), SQL Server Analysis Services (SSAS) and SQL Server Reporting Services (SSRS).

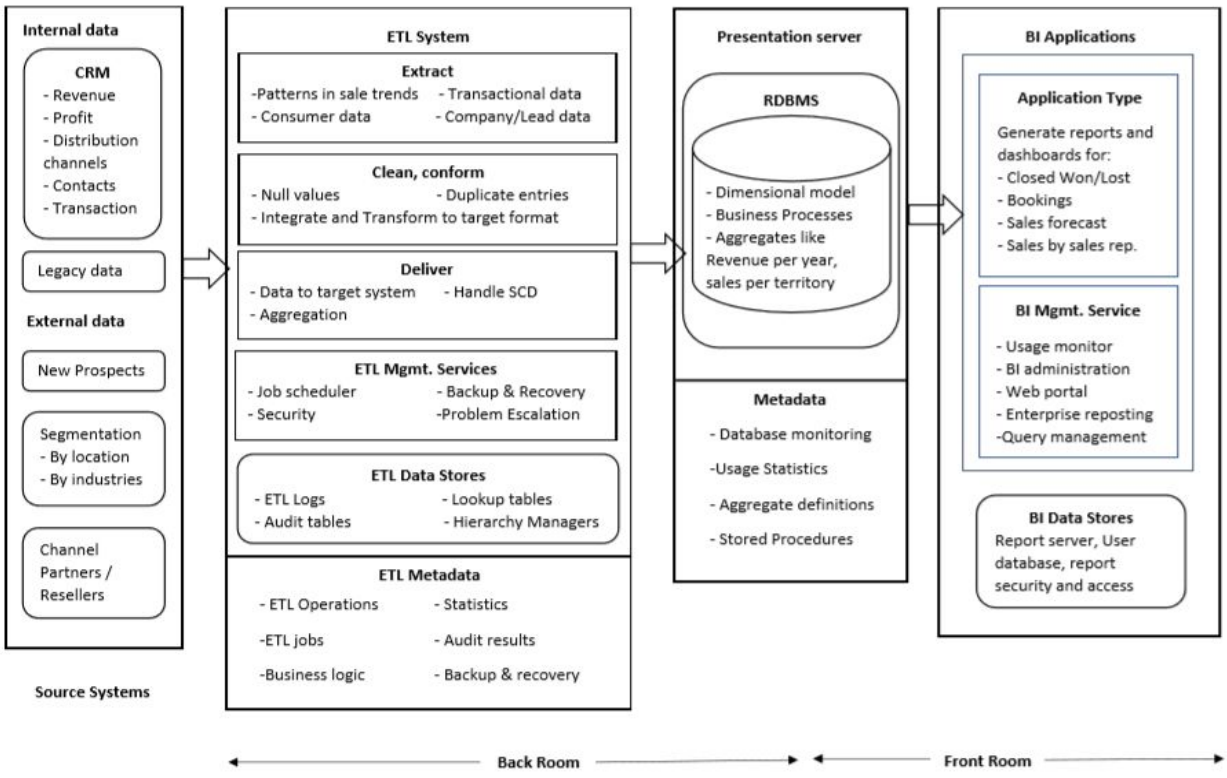


Figure 4 - Data warehouse Architecture

Data Source Layer : The main source of data comes from Salesforce CRM. It consists of customer accounts linked with the information such as opportunities, leads, revenue, transactions for the customer. It can also be used to understand the pipeline of prospective sales making forecasting more accurate. CRM and legacy data consisting historical information form our internal sources of data. The external data comes from new prospects from marketing, market segmentation and partners.

ETL / Data Integration Layer : This layer extracts, transforms and loads raw data from the data source to create an Operational Data Source (ODS). The data obtained from data source layer is

in different formats with different granularity therefore there are various transformations that are to be applied to the data to get it in the desired format. SQL Server Integration Services (SSIS) performs all the backroom technology.

The data from source systems is extracted, cleaned for duplicates and null values. Then, the transformations will generate surrogate keys for all the dimensions. SSIS makes transforming/handling the Slowly Changing Dimensions simpler. The other transformations like generating look up, derived columns for aggregation in fact table and data conversion tasks will also be in this data warehouse. Logical columns to capture the load date will also be generated in the transformation. The flat files consisting of transformed and cleaned data will be then loaded in the target system.

Presentation Layer : The presentation server will have all the data that will be stored for end user query and reports. The operations will be carried out by SQL Server Analysis Services (SSAS). The primary aim of SSAS is to derive sensible inferences from the widely dispersed data which can help the organization to work intelligently and in a cross functional way. It will create and manage multi-dimensional relational structures known as cubes. The users can then generate reports from these cubes. The Analysis services provide functionality like Aggregation, KPI, Calculated fields that will help in finding business answers.

Application Layer : The SQL Server Reporting Services (SSRS) performs all the front room operations. This layer will generate reports, dashboards and scorecards for conducting analysis on the data obtained from the previous layers. Drill through, drill down, roll up, parameterized, cached, ad hoc and linked are some of the types of reports that can be created using SSRS. The following reports specific to the company can be generated : total bookings for current year, revenue generated per quarter, number of closed won/lost sales, sales per region, sales forecast and much more.

Budget plan

The budget described below is for the period of 12 months which is the time we are anticipating for the completion of the project. The Chief Finance Officer have issued a budget of $500,000 for the next 12 months.

A good technical team is required for a successful BI project. Therefore, the BI team consists of 7 employees who are highly specialized. The team consists of:

- 4 ETL and BI developers - will develop packages and load data from source system into tables which is then transformed into data mart.
- 2 Technical Architects - will organize the data and maintain the ETL infrastructure at the desired project level.
- 1 Quality Assurance Tester - to keep timely track on project and keep a check on requirements.

The budget is also used towards different high end equipment such as servers, computers etc. and required software and licenses.

Following table shows a detailed description of the budget spent on different items:

Item	Price	Number of License	Total Price
Microsoft BI Stack	$ 8092 + $199 / license	7	$ 9,485.00
Virtualization Server 12-Core 128GB RAM 12TB RAID Dell PowerEdge R710 (used)	$1,241	3	$ 3,723.00
Windows Server 2012 R2 Standard edition	$882	7	$ 6,174.00
Dell XPS 8900 6th Generation Intel Core i7-6700 Processor, 16 GB DDR4 Ram 1 TB	$799	7	$ 5,593.00
Dell S2240M 21.5 inch monitor	$129	7	$ 903.00
Microsoft Office 365	$99	7	$ 693.00
Employee Salary (Full time)			
ETL and BI developers	$70,000	4	$ 280,000.00
Technical Architects	$65,000	2	$ 130,000.00
Quality Assurance Tester	$60,000	1	$ 60,000.00
Installation Charges	$2,000	-	$3,000
Total			**$ 499,571.00**

Figure 5 - Budget Plan

Conclusion

This report presents the information and requirements needed for the management to consider for implementing the data warehouse and business intelligence solution. The proposed solution meets the tactical, operational and strategic needs across the organization.

The focus is currently on a single business process for making the project feasible. The DW/BI solution will help the management to record sales information more systematically. The main purpose of any business is to maximize its profit. We believe that an investment in a data warehouse with profitability as the KPI will help them achieve their goal. The data can be used for utilizing the data mining capacity and to provide business users a statistical knowledge about the business for better decision making.

Recommendation

Strategic decision making at the right time and occasion plays a key role in determining success of an organization. Hence, it is important for the executives to consider the relevant data available. The best source of integrated data is available with a data warehouse system coupled with business analytics.

Currently, this business proposal is geared towards one business process - Sales. The management can also consider deploying this solution for the rest of the processes.

The management should also make decision on investing in data warehouse based on the cost and benefits of the investment. They should perform various financial calculations to see how their profitability will increase over the years and how the expenses of the data warehouse will be recovered over the years. However, from a long term perspective, we believe that the data warehouse will definitely be beneficial. Once that is done, the management should arrange for funds and install the data warehouse as soon as possible.