**Data warehouse development for Banking Sector**

Lisha G and Dr. Siddhaling Urolagin

https://www.linkedin.com/in/siddhaling-urolagin www.researchreader.com https://medium.com/@dr.siddhaling https://github.com/siddhaling

**Introduction**

Data warehousing is the electronic storage of a large amount of information by a business or organization. A data warehouse is designed to run query and analysis on historical data derived from transactional sources for business intelligence and data mining purposes. Making better decisions faster can make the difference between surviving and thriving in an increasingly competitive market place. The financial services industry needs to respond to challenge such as globalization, de regularization, and customer expectations. Over the years, researches are proven that when it comes to data warehousing and OLAP technology it means we are storing data, reuse the data, and retrieve the data for a long period.

In the banking sector, the OLAP technique is used for accessing information at any time and anywhere. Nowadays, banks are storing more information than ever before. Therefore, the decision-makers must have the right information at right time to help them more informed and intelligent decisions to generate the reports. A data warehouse is a type of data management system that is designed to enable and support business intelligence activities, especially analytics. Data warehouses are solely intended to perform queries and analyses and often contain a large amount of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications. A data warehouse centralizes and consolidates a large amount of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision making. Over time it builds a historical record that can be invaluable to data scientists and business analysts. A typical data warehouse often includes the following elements i)a relational database to store and manage data ii) an extraction, loading, and transformation (ELT) solution for preparing the data for analysis iii) statistical analysis, reporting, and data mining capabilities iv) client analysis tools for visualizing and presenting data to business users v) other, more sophisticated analytical applications that generate actionable iv) information by applying machine learning and artificial intelligence (AI) algorithms

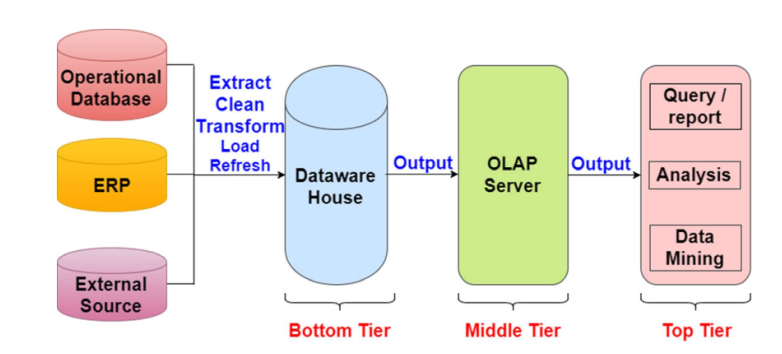
**2. Literature Review**

The data warehouse is being used in many industries like manufacturing, production area, telecommunication, logistics, health sectors, and financial sectors. To make process decisions, performing a sophisticated data analysis by organizations is very much important in a current competitive world. The usual databases that come with the operating system do not perform the requirement analysis as these are targeted only for basic operations. There is a data warehouse identified as it is more useful. A data ware is seen in following environments were integrated and a total of enterprise is required, for strategic information where enterprises current and historical data is needed, and where can be pulled decision-making transaction without hindering operation system, where information can be rendered consistently to organizations, where could provide a flexible and interactive source of information. The data warehouse is a computing environment rather than thinking like single software or hardware and a user can be in touch with data when they need to make a better decision.

when seeing into the trends of globalization and competition for the internet world, many departments in enterprises or organizations or the companies located in various regions, bring up the competition and reaction ability. there are numerous data warehouse systems available in the enterprise. In distributed decision-making requirements. data in the integrated data warehouse is addressed to enable data exchange [4]. Rendering OLAP data into useful information as in a data warehouse is a very good approach. this will help departments to take decisions making. this can be achieved based on data analysis techniques like multidimensional analysis and data mining. Designing a data warehouse process considers a complex task and requires a systematic and structured way to promise its success. hence different methodologies are proposed to perform such kind of designing process which then be classified into at requirement level, data drive, and hybrid way.[5]. Data sources are being seen from different areas like hardware and software and network and operational databases for storing the data. The data will be extracted from these sources using customized application programs or commercial extraction and transforms and then loads this is known as the ETL process. the extracted data then fed into the staging area where it should be transformed [6]. the data warehouse can be categorized into main three phases: First is the ETL process, then data modeling, and finally efficient and flexible access to data [7]. data quality in the data warehouse and the cube depends on the quality of the ETL process [8]. There are many types of data warehouse schemas. Here will be discussing star schema. instar schema, each group of dimensions are placed in the dimension table and transaction-related data are placed in the fact table. In a star schema, the fact table is placed in the center rounded by the dimension tables. a dimension table contains master details and a Fact table can grow to billions of rows with attributes like, foreign key and it will be referencing to dimension table [9]. one of the most known techniques for the database which is designed in the view of end-user querying is dimensional modeling. in this paper [10] discuss the options of deployment and acceptance of conversion Model (CM) to provide the details of the fact table and dimension tables. in the phase of designing, if expects structured and systematic analysis on each stage before moving to the next one waterfall method would be ideal. in the spiral method, short intervals of successive releases will be generated. This is considered a good option for data warehouse development, specifically for data marts, turnaround time calculated in this method is short. if any modification required, it can be achieved quickly, new designs of technologies can be also be accepted inappropriate manner. [12]. The relation between the data warehouse and OLAP is essential in the discovery process [13] the main features observed in OLAP and that are regularly by referenced as Slice- and-Dice, Pivot, Drill down/up [14] OLAP functions gives access to the data facts and with the data captures and the ability to view the data in multiple ways helps to the actual status of the business growth

**Data warehouse architecture**

Banking information-based data warehouse management system needs to implement features like, first solve customer relationship management, the business system will organize the data into a customer-centric data warehouse, the customer structure, customer behavior, customer groups and market activities analytics and management. While the results of the analysis and email system closely integrated and timely market information will be passed to the customer. Secondly, the companies analyzed in terms of revenue, cost, budget, and so on. From a variety of different angles, using statistical analysis techniques, to provide relevant key performance indicators and income statements. While revenue development, project status, budget, and other aspects of the use of detailed analysis of a situation, management leaders can gain business development reports. Taking into account the need for all aspects of the above systems, based on the data warehouse management information systems from basic banking information network, warehouse management, data warehousing and data presentations, the interaction between them, together form a structured data warehouse-based bank information system environments.



**Figure 1: Data warehouse architecture**

The figure I provided above depicts the three-tier architecture of the data warehouse

Data warehouse usually has a three-tier architecture that includes a) Bottom tier (data warehouse server) b) Middle tier (OLAP server) c) Top tier (front end tools)

**Bottom tier** of the architecture is a database server. It is the relational database system. Backend tools and utilities use to feed the data into the bottom tier. These backend utilities perform the task like extract, clean, load, and refresh functions.

**Middle Tier**: OLAP server is implemented in the middle tier in either of the following ways: By Relational OLAP or by multi-dimensional OLAP. Relational OLAP (ROLAP) is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations. Multi-dimensional OLAP directly implements the multi-dimensional data and operations.

**Top tier** is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools, and data mining tools.

There are 5 main components of a data warehouse 1) database 2 ) ETL tools 3) Meta Data 4) Query tools 5) Data marts

**Database** is the central component of the data warehouse architecture. And it stores all enterprise data and make it arrangeable for reporting.

**ETL** tools help to extract data from different sources, transforming it into a suitable arrangement, and loading it into a data warehouse. They include the functionalities like eliminates unwanted data in operational data from loading into the data warehouse. Search and replace the common name for data arriving from different sources.in case of missing data, populates with default data. De duplicated repeated data arriving from multiple sources. These ETL tools have to deal with the challenges of database and data heterogeneity.

**Metadata** is data about data that defines the data warehouse. It is used for building, maintaining, and managing the data warehouse.

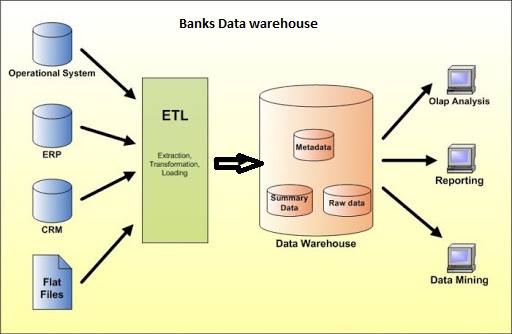
**Query tools** allow users to interact with the data warehouse system. business users cannot work with databases directly. they extract the report with the assistance of several tools. Some of these tools include: Query and reporting tools help users produce corporate reports in the form of excel sheets or interactive visuals, application development tools help create tailored reports, data mining tools, Data mining tool is a process of discovering meaningful new correlation, patterns by mining large amount data. OLAP tools allow users to analyze the data using elaborate and complex multidimensional views.

OLAP: in the data ware design and implementation OLAP has an important role. OLAP could give quick result to the analytical queries which are multidimensional. Data mining techniques are used with OLAP as an assisting tool. OLAP has made successful role in data warehouse due to the following: Fast speed, Analysis ability, Multidimensional

**Data warehouse architecture on Banking Sector**

The data warehouse architecture for the banking sector described in the paper has 4 layers and different processes to complete the task. This architecture will help to control and monitor how many customers are onboarding yearly wise in different regions on one particular bank. External Data sources, OLAP, and front end interface are defined in a good manner.

Data sources can be from different types like excels file, text file, or operational database. Then it will be transported into targeted systems.



**Figure 2: Data warehouse architecture in Banks**

4. **System framework and implementations**

1. Data collection process

In the bank operation process, data received from multiple sources and collects the data in the following few ways

Flat files like excel files, word files, could contain branch details and staff details

And relational database where resides dynamic and huge amount of transaction data. [16]

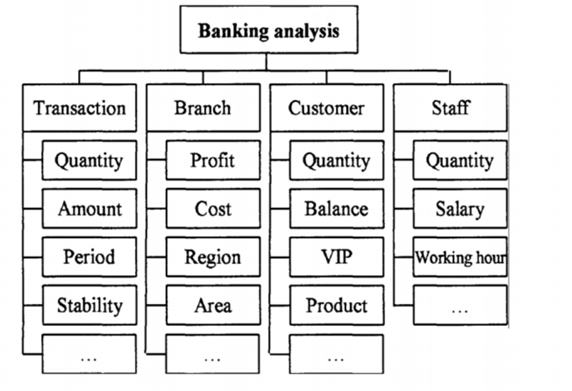
Data will be first entered daily using online transaction processing (OLTP) environment and stored in the operational database. This could be in a common relational database like Oracle, SQL Server, MySQL, DB2etc. before data gets transferred into the data warehouse from the operational database, it should go through three main functions: extraction, transformation, and loading.

In the first stage, data will be extracted from the available external and internal sources. The data relative to pas periods loads into an empty data warehouse. The subsequent incremental data extractions will be updated into the data warehouse and will available over time. The selection of data to be imported into the data warehouse is based on DW design, this further depends on the information needed for business intelligence analysis and decision support systems that operate in the application domain.

The next stage is the transformation phase. This is intended to improve the quality of data extracted from different sources. Where correction is required in the inconsistent and missing data. During data cleaning stage the values recorded in different attribute with the same meanings, missing data, and existence of junk values these will removed during the cleaning process. Other major operations inside the transformation phase are converting character set, normalizing, or DE normalizing data into the desired dW schema.

Next is the loading phase, the cleansed data will be loaded into DW tables. In the traditional data warehouse architecture model, new data loading occurs only at a certain interval. When the data warehouse is taken offline, data is integrated during a relatively long-time-interval.

To standardize the data analysis, a data warehouse is organized in small units called data mart. Datamart is dedicated to studying a specific problem. To understand locate the data for DW users, information related to the DW system and content is required. This information is called metadata. The metadata consists of attributes of the Data warehouse original source of data. The metadata could consist of business definitions, data alerts, organizational change.

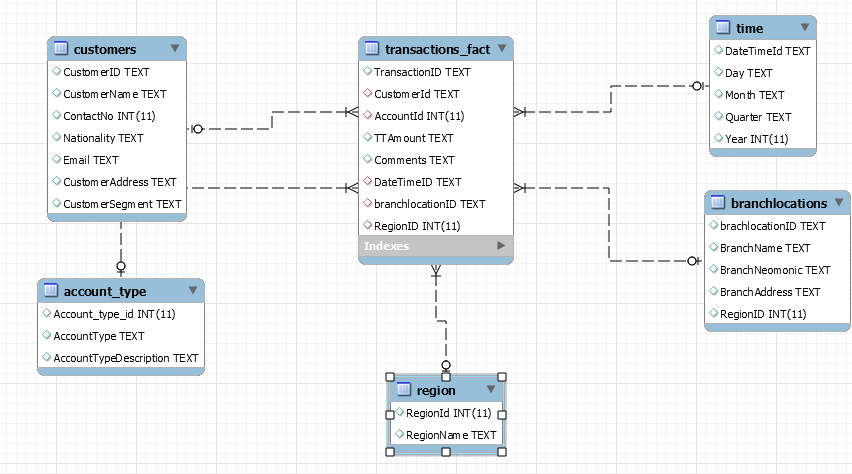


**Figure 3: Banking analysis**

Figure 3 depicts the area of analysis that can be performed on the individual segment.

**5. Schema Design**

Here is the Star Schema designed for banks data warehouse



**Figure 4: example of star schema**

The star schema has a centralized data repository stored in the fact table. The schema splits the fact table into a series of the denormalized dimension table. The fact table contains aggregated data to be used for reporting purposes while the dimension table describes the stored data

The denormalized designs are less complex because data is grouped. The fact table uses only one link to join to each dimension table. The star schema’s simpler design makes it much easier to write complex queries.

The DW schema consists of majorly with two elements: facts and dimensions. Facts consist of the matrices of events. Dimensions are used to analyze the results of the metrics.

**6. Experimental Results**

Figure 4 shows the star schema designed for the banking institute. It consists of a fact table as transactions and dimensions as a customer, branch locations, Region, time, and account type. The customer dimension table consists of a customer name, customer address, customer segment, nationality, email, and contact number. In branch, the location dimension contains the details of the location where the branches exist. Region dimension has the name of Region and the time dimension contains the details of year, month, and quarter details.

|  |
| --- |
| **Figure 5: dimensions populated in Pentaho BI interface** |

Figure 5 helps to identify the dimensions included in the reporting tool and the measures

|  |
| --- |
|  |
| **Figure 6: graph generated on the total number of branch locations for All customer in all branch an in All-time range** |

Figure 6 displays a horizontal bar representation of the total number of branch locations included. This includes all customer details of all branch and all-time dimensions

|  |
| --- |
| **Figure 7: Total volume of All branch location all customers all accounts and all times** |

Figure 7: displays the total number of records included for all customers in all branch location and all-time dimension

|  |
| --- |
| **Figure 8: Total Volume in all account type** |

Figure 8 displays a vertical area view of the total number of records found in all account type

|  |
| --- |
| **Figure 9 Total number of Regions** |

Figure 9 displays the total number of regions and following is the MDX query generated on total count of Regions

**MDX Query**

select NON EMPTY {[Measures].[CountofRegion]} ON COLUMNS, NON EMPTY {([Customer].[All Customers], [BranchLocation].[All BranchLocations], [Region].[All Regions], [New Dimension 4].[All New Dimension 4s])} ON ROWS from [BankData\_Cube1]

|  |
| --- |
|  |
| **Figure 10: Graph generated on All customer on selected/distinct branch location plus All customer in Selected Region** |

Figure 10 displays the pie chart representation of the total number of branch locations for all customers and the total number of Regions for all the customers. And here is the MDX query generated for the same.

**MDX Query**

Select NON EMPTY Crossjoin({[Measures].[CountofBranchlocation], [Measures].[CountofRegion]}, {[Customer].[All Customers]}) ON COLUMNS,NON EMPTY {([BranchLocation].[All BranchLocations], [Region].[All Regions], [Time].[All Times], [New Dimension 4].[All New Dimension 4s])} ON ROWS from [BankData\_Cube1]

|  |
| --- |
| **Figure 11: Graph generated on All customer on selected/distinct branch location plus All customer in Selected Region** |

Figure 11 displays the horizontal bar representation of the total number of branch locations and the total number of regions for all the customers. And here is the MDX generated for the same.

select NON EMPTY Hierarchize({([Measures].[CountofBranchlocation], [Customer].[All Customers]), ([Measures].[CountofRegion], [Customer].[All Customers])}) ON COLUMNS, NON EMPTY Hierarchize(Union({([BranchLocation].[All BranchLocations], [Region].[All Regions], [Time].[All Times], [New Dimension 4].[All New Dimension 4s])}, Crossjoin([BranchLocation].[All BranchLocations].Children, {([Region].[All Regions], [Time].[All Times], [New Dimension 4].[All New Dimension 4s])}))) ON ROWS from [BankData\_Cube1]

|  |
| --- |
| **Figure 11: All customer in all branch locations** |

Figure 11 displays the pie chart representation of all the customer of all branch locations. And here is the mdx query generated

select NON EMPTY {([Measures].[TotalVolume], [Time].[All Times], [AccountType].[All AccountTypes])} ON COLUMNS, NON EMPTY {([Customer].[All Customers], [BranchLocation].[All BranchLocations])} ON ROWS from [bankdataCube]

|  |
| --- |
| **Figure 12: Drill through table screen** |

Figure 12 displays the total number of branch location and the total number of the region in vertical bar representation

**Conclusion**

The data warehouse is useful for decision-makers to gain insights by accessing different data sets. Building a data warehouse is required an effort. The efficient tools available in the market helps to build the report inefficient way and on the user requirement. Multi-dimensional modeling is one of the most popular and effective techniques in data warehousing. the design presented here proves that star schema can be used in a large amount of data. The reports generated using the Pentaho BI give a detailed view of the available statistics. using the report generated in the BI tool helps to see the current volume of each region and use a reference to the management to decide if there is a problem with the business results. The process of making a report using the Pentaho BI is quite easy and available in formats such as Excel, PDF. The design of the data warehouse included in this paper can be used to analyze the growth of the business.

**References**

1. <https://www.researchgate.net/publication/224353129_Banking_Intelligence_Application_of_data_warehouse_in_bank_operations>
2. [Bank of America Case Study: The Information Currency Advantage (vldb.org)](http://www.vldb.org/conf/1998/p641.pdf) <http://www.vldb.org/conf/1998/p641.pdf>
3. <https://www.inetsoft.com/company/banking_business_intelligence_case_study/>
4. <https://www.researchgate.net/publication/228922190_LITERATURE_REVIEW_ON_DATA_WAREHOUSE_DEVELOPMENT>
5. <https://www.researchgate.net/publication/320148512_Data_warehouse_conceptual_design-a_literature_survey>
6. <https://silo.tips/download/31-data-warehouse-methodology-literature-review-on-data-warehouse>
7. <http://docsdrive.com/pdfs/medwelljournals/jeasci/2017/4208-4214.pdf>
8. <https://www.sciencedirect.com/science/article/pii/S1877050919314097>
9. <https://www.researchgate.net/publication/290787413_Star_Schema_Advantages_on_Data_Warehouse_Using_Bitmap_Index_and_Partitioned_Fact_Tables>
10. <https://www.researchgate.net/publication/322410950_Dimensional_Modeling_Using_Star_Schema_for_Data_Warehouse_Creation>
11. <https://link.springer.com/chapter/10.1007/978-3-642-36926-1_14>
12. <http://www.ijsrp.org/research-paper-0313/ijsrp-p1597.pdf>
13. <https://ieeexplore.ieee.org/abstract/document/8749017>
14. <https://www.irjet.net/archives/V3/i11/IRJET-V3I1158.pdf>
15. <https://www.researchgate.net/publication/226951406_Online_Analytical_Processing_OLAP_for_Decision_Support>
16. <https://link.springer.com/chapter/10.1007/978-3-642-57681-2_8>