Design and Implementation of a Data Warehouse for Facility Management of a Public Television Broadcaster in Sri Lanka.

Rumesh Mohan Vibath Algama Sachini Jayawardena Sajith Sanjeewa

*Master of Data Analytics* *Master of Data Analytics* *Master of Data Analytics* *Master of Data Analytics*

*Faculty of Graduate Studies Faculty of Graduate Studies Faculty of Graduate Studies Faculty of Graduate Studies*

*University of Kelaniya University of Kelaniya University of Kelaniya University of Kelaniya*

*Sri Lanka Sri Lanka Sri Lanka Sri Lanka*

[rumeshmohan@gmail.com](mailto:rumeshmohan@gmail.com) [algama.vibhath@gmail.com](mailto:algama.vibhath@gmail.com) [sachinidjayawardena@gmail.com](mailto:sachinidjayawardena@gmail.com) [sanjeewasajith@gmail.com](mailto:sanjeewasajith@gmail.com)

***Abstract -* Any organization requires a proper operational system to maintain its day-to-day activities. But an organization without strategy fails in its vision and goal, finally ending its run in the competitive world. Analytical systems, basically data warehouses, are the pillars of analysis that collect all information around the organization required by the top management to analyze, operate and persist their organization in this competitive world using their strategic actions. The aim of this study is to create a data warehouse that integrates data from various sources, extract them appropriately and load them into it for future analysis with ease and accuracy that helps the organization’s stakeholders to make correct decisions.**

***Keywords – BI, Broadcast, Data Marts, ETL, OLTP, OLAP, IOT, SCD***

1. **INTRODUCTION**

There are three main broadcasters in radio and television communication. They are Public broadcasters, Private broadcasters and Community broadcasters. Public broadcasters are fully or partially funded by the respective governments, private broadcasters make money by selling advertisements and community broadcasters often operate as cable networks. Since 1982, Terrestrial Television has been popular among the Sri Lankan community.

Later, with the advancement in technology, satellite television became popular among certain groups of people. But, now television broadcasting is more popular on the internet with the enhancement of IOT (also referred as OTT - Over the top). Broadcasters use two main categories of materials or contents for broadcasting. They either produce their own contents or purchase content from production houses or third parties. Own contents produced within the broadcasters’ premises or outside is known as the Electronic Field Production (EFP) of television programs.

One of the leading public sector television broadcasting organizations in Sri Lanka has been selected for this study which is a pioneer in this field. The organization capitalized several divisions for the success of their major business. Administration, Engineering, News, Finance, Commercial, General Programmes, Education Programmes, Channel Eye are the main divisions which consist of several other key units. With the collaboration of few or more divisions, the organization continues its business processes. The organization produces their own broadcasting contents or purchase value based on air time, viewership and sells airtime for advertisements to make money.

The purpose of this study is to provide accurate and meaningful information for the top management for better decision making regarding the content production process which is a core component of their entire business process. The existing standalone Transaction Processing System (TPS) only provides the Management Information System (MIS) with necessary required reports. It also could not accommodate ad-hoc queries required by the management when necessary. Production based processes are mainly involved in engineering division and this information is not available for the finance department for costing.

Further, some of the advanced facilities are not equally distributed among the producers. In most instances, actual cost of the program production is significantly higher than the estimated budget in the program proposal. This has led towards financial concerns especially in defining feasible rates for sales advertisements. Also the existing systems generate reports based on individuals and their relevant divisions. In this project the proposed data warehouse is expected to reduce and rectify miscommunications, mistakes and errors as mentioned above enhancing better and feasible decision making.

1. **LITERATURE REVIEW**

The data warehouse design principles are applied and used in various forms with respect to the organization and its structure. In a data-driven management decision making process, a data warehouse is essential as it facilitates the Business Intelligence (BI) systems. A data warehouse system is designed from the existing transaction processing data systems where a common data model is created from various different source systems.

In a data warehouse model the heterogeneous data is selected, extracted and then transformed by data integrations, data cleansing, identification of data quality inconsistencies and conversions that are needed to make the data fit for the model and finally loading them to the data warehouse. Using this data the user can perform queries and analysis required for decision making. Online Analytical Processing (OLAP) systems are used in the analysis process which enable visuals, report generation and support ad-hoc reporting [6].

The selected study refers to the implementation of a data warehouse for a facility management system in a broadcasting company. When considering previous studies made on Facility management utilization, in the research paper of “Business Intelligence in Facility Management: Determinants and Benchmarking Scenarios for Improving Energy Efficiency” by Gavin and Marcinkowski discusses how the electrical power consumption in retail facilities have been analysed and optimized for better utilization using a data warehouse and a business intelligence model [6].

1. **NEED OF A DATA WAREHOUSE**

Today's television industry is becoming more uncertain due to the competitive environment (business pressure), customer intimacy and diversified mechanisms and mediums of video streaming. Videos are available on demand with different high quality formats on different platforms which are more attractive to the viewers. In order to persist these challenging factors, this organization needs to utilize its production process in an optimal cost effective manner. Further it must maintain the quality of production content which is suitable for television transmission as well as other high demand platforms (Example: Youtube). In this study we expect to conclude how the following questions can be answered using the concepts of data warehousing.

Q1. What facilities are highly utilized in program production?

Q2. Which productions have over utilized facilities more than the regular estimates?

Q3. Which productions do not utilize facilities that are reserved for a particular period?

Q4. How much estimate and actual cost differentiation are in program production?

Facility Management Systems receives a large number of reservation requests daily for 250+ facilities. Physically and Logically separated two Online Transaction Processing (OLTP) databases receive data from three separate TPSs for the entire process. For answering the above four main questions, a large volume of data in above two databases need to be processed and stored for quick reference. It is observed that there is no such mechanism for reporting and analysing rather than creating ad-hoc reports as instances. Data warehouse facilitates the requirement by storing large volumes of meaningful data for generating decision making information. This will help the top management to guide the operational staff for optimal utilization of facilities with minimal cost.

1. **DATA SOURCES**

Existing data sources of Financial MIS, Program Production MIS and Facility Management MIS are identified as the main data sources. Program production proposal forms and facility request forms could be used as other data sources too. Figure 1 demonstrates how the current TPS works.

**Figure 1 : Current TPS Flow.**



1. **DATA WAREHOUSE REQUIREMENTS.**

Organization has different users including top management, middle management, divisional heads, unit heads, producers, assistant producers, booking officers, marketing personnel etc. Each user has different roles and requirements identified by categorizing them into functional and non-functional requirements.

1. **FUNCTIONAL REQUIREMENTS.**

Functional requirements define what the organization expects from implementing a data warehouse. It explains the way how a particular data warehouse functions based on the user specified requirements.

**Data Visualization:** Different reports based on different queries with graphical visualization to make a clear picture of the data and the requirement of the user must be met that is complying with the industry standards.

**Data Integration:** Data from multiple different databases and data sources have to be integrated and made available to the users via the data warehouse that can be processed for all kinds of required and detailed analysis.

**Data Validation:** Data Validation is a process that is incorporated before using the collected data. Methods like scripting and usage of various other essential integrated tools can be used for doing data validation. By ensuring the integrity of datasets, the quality of results and efficiency of the overall process is significantly enhanced.

**Security:** User Authorization is defined based on the designation and the level of authority defined based on the job position. Row-level security for user based or division based access can be used throughout the data warehouse. Authentications are provided by the organization and they can be utilized in the system which are secured and private for employees based on their IDs, biometrics or other feasible measures.

1. **NON-FUNCTIONAL REQUIREMENTS**

These are user requirements to use the data warehouse which outlines its operational capabilities and specifications that mainly interprets how a data warehouse works and operates.

**Performance of ETL Process:** ETL process should be performed without or with minimum impact to the source system as impacts on TPS are crucial in routine functionality. Indexing, sharding, partitioning and query optimization etc can be used to increase the performance of the data warehouse.

**Interoperability:** The system should be compatible with the hardware and software we use for the implementation of the above infrastructure. Possibility of viewing analytical tools and reports via mobile platforms and other IOT devices. Sources that we obtain from various heterogeneous platforms should be processed and handled by the hardware as well as the software all the time with minimum failures. When using different systems that are highly cohesive and less coupled we can achieve robustness and autonomous integrity of the system which paves way for high interoperability among components of the data warehouse.

**Disaster Recovery:** Analytical systems do not generate data. Even if the system crashes, the staging environment will be able to regenerate facts and dimensions. Therefore the existing infrastructure could be enough and be able to retrieve lost data.

**Scalability:** Could be able to accommodate a large volume of data with current and future requirements. Easy update and continuous improvement with less cost with changing timeframes and requirements.

1. **DATA WAREHOUSE DESIGN**
2. **Kimball’s Approach**

In this data warehouse design we have utilized the Kimball approach which is also known as the bottom-up procedure. In this approach the data marts are created and placed before the data warehouse where data from various sources goes via staging area and into the data marts which are connected to the data warehouse. These data marts provide reporting facilities of individual business divisions. The main benefit of this approach is its cost and time consumption is comparatively low to other approaches. It allows easy adding of new data marts that correspond to different or new divisions that are being added to the organization and later extending the data warehouse as required.

**Kimball & Ross 9 Step Methodology:**

This is a 9 step method which consists of separate tasks in order to design an effective data warehouse [4][5].

1. **Choosing the process:** Choose the main subjects that we are going to transform into the data warehouse.
2. **Choosing the grains:** Grains refers to the data represented by the fact table. Fact table is a collection of data that comes from the main processes. Dimension tables related to the fact table can be determined after determining the grain from the fact table. Grains can be obtained from facilities allocated by the users and utilization of those facilities.

1. **Identifying and confirming dimensions:** Dimension tables are determined based on the design requirement of the data warehouse to be built. Table 1 shows the list of identified dimensions.

**Table 1 - Dimensions Table**

| **Dimension Name** | **Description** |
| --- | --- |
| **DimProducer** | Main actor in the facility utilization of program production process. This dimension can be used as SCD Type2 as designation and the division is changed. Eg: Assistant Producer → Producer → Senior Producer |
| **DimProduct** | This is the base table for Programme Production data. Name of the Programme, PP\_Number(Programme ProductionNumber), etc.  ProgrammeType, ProgrammeLanguage, ProgrammeTransaction, ProductionDivision, ProductionChannel tables in the OLTP can be amalgamated into this single dimension.This also includes estimated cost for each facility that is to be utilized for the production (Each production utilizes one or more facilities). |
| **DimFacility** | This consists of all the available facilities and related attributes. FacilityName, Description and UnitPrice are key attributes. This dimension can also be used as SCD Type2 as the unit price changes periodically or yearly. |
| **DimTotals** | This includes the total time and cost for a particular program production. (Sum of all the facilities per production) |
| **DimDate** | Information about timeline attributes are included. |

ProgrammeCostCategory and ProductionNature tables could be isolated as they do not have any identified analytical data.

1. **Choosing the facts:** Based on the identified grains,FactBooking has been chosen for the data warehouse design.

**Table 2 - Fact Table**

| **Fact Name** | **Attributes** |
| --- | --- |
| **FactBooking** | PP\_Number, Producer\_SK, Product\_SK, Facility\_SK, Totals\_SK, Date\_SK, EstimatedCost, BookingCost, UtilizeCost |

1. **Store precalculations in the fact table:** Calculate and measure the parameters and attributes that are going to be used in the fact table. We expect that every time the ETL happens, attributes are calculated.

1. **Round out the dimension tables:** Attributes are chosen for the identified dimensions in the fact. Table 3 defines the identified attributes for each dimension.

**Table 3 - Attributes of Dimensions**

| **Dimension Name** | **Attributes** |
| --- | --- |
| **DimProducer** | Service\_Number, Producer\_Name, Producer\_SK, IsActive |
| **DimProduct** | PP\_Number, PP\_Index\_No,  Product\_SK, Producer\_Name, Title\_Of\_Programme, Title\_Of\_Series, Actual\_No\_Episode, Duration,  Group\_Description, Cost\_Description, Language\_Desc, Programme\_Status, Programme\_Type\_Desc,  Division\_Description, Estimated\_Amt\_Dr  Channel\_Description, Programme\_Nature\_Telecast, Prog\_Year, Date\_Of\_Telecast, Name\_Of\_Sponsor, Name\_Of\_Participation, Target\_Audience, Objectives, Contents |
| **DimFacility** | Item\_No, Facility\_SK, Item\_Name, Cost\_Description, IsActive |
| **DimTotals** | PP\_Number, Totals\_SK,Total\_Estimate\_Cost, Total\_Actual\_Cost |
| **DimDate** | Date\_SK, Date, Day, WeekDayName, Month,MonthName,Year, IsPoyaday, IsHoliday, IsWeekend |

1. **Choosing the database data timeframe:** In this step we decide how long the data from the sources is going to process. At present, sources retain data since 2015, we are planning to process last five years data from 2018 to the end of 2022.
2. **Determine the need of SCDs:** As the chosen dimensions could be changed in future, we define a few dimensions as Slowly Changing Dimensions Type 2 to track the changes. When the dimensional attributes are changed it causes the creation of new dimension records with different surrogate keys. Table 4 defines the SCD Type2.

**Table 4 - Dimensions with SCD Type2**

| **Dimension Name** | **Attributes** |
| --- | --- |
| **DimProducer** | Designation, Previous\_Designation, DesEffective\_Date, DesExpiration\_Date,  Division, Previous\_Division, DivEffective\_Date, DivExpiration\_Date |
| **DimFacility** | Unit\_Price, Previous\_Unit\_Price, Unit\_Price\_Effective\_Date, Unit\_Price\_Expiration\_Date |

1. **Design the physical database:** All gathered facts, dimensions and attributes are drawn into a physical database model which is the final product of the data warehouse.
2. **DATA WAREHOUSE ARCHITECTURE**

**Figure 2 : Kimball’s Model Based Data Warehouse Architecture**



Figure 2 exhibits the proposed physical architecture of the data warehouse. Some divisions have a fairly larger volume of data than others. Therefore data marts are determined based on the divisions of the organization.

1. **PROPOSED DESIGN**

**Figure 3 : Proposed Logical Design**

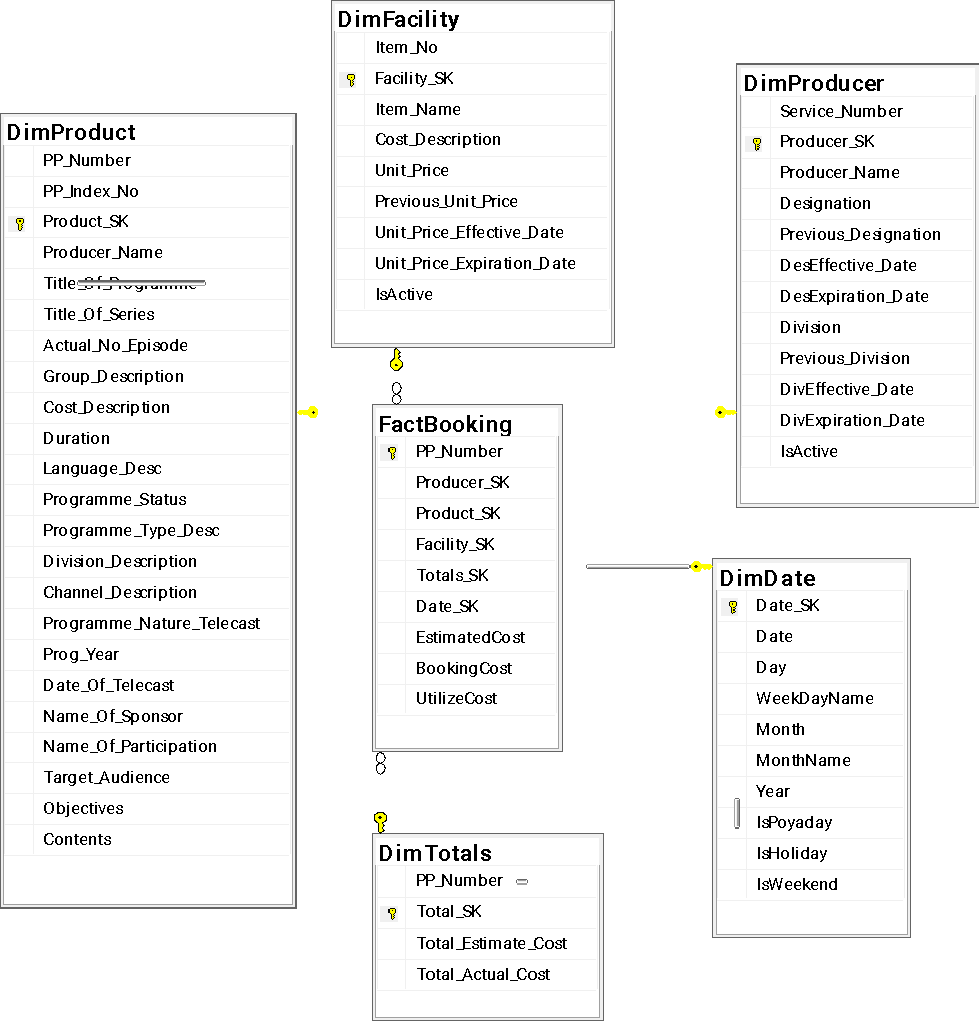


Figure 3 demonstrates the logical design of the data warehouse. We have identified star schema modeling, which holds quantitative data. Fact table is surrounded by multiple dimensions and linked by the surrogate key. Further it is simple and straightforward fetching data from transactional schema to analytical schema and also facilitates designing OLAP cubes.

1. **TECHNOLOGY STACK FOR THE PROPOSED SOLUTION**

While proposing the technologies to realize the proposed data warehouse solution, we examined the following dimensions based on organizational requirements and constraints.

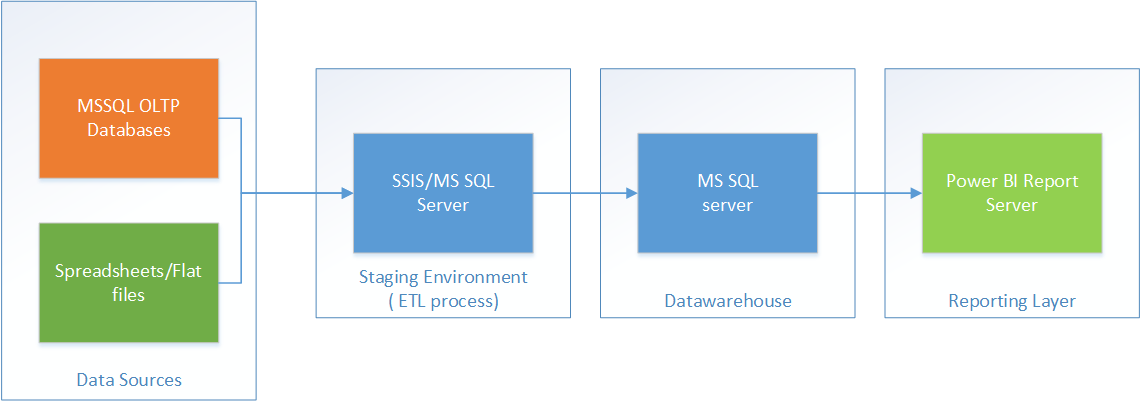
* **Expected storage data types**: The types of data to be stored in the data warehouse. All data sources considered in this solution are of structured nature. Therefore, a solution that would favour structured data with a relational database approach would be beneficial.
* **Scale of the solution**: In this dimension we consider the scale of the data to be ingested and analyzed. We estimate that the solution outlined in this paper will only ingest and analyze around 1TB to 2TB of data during its expected operational life.
* **Expected performance**: In this dimension, we consider the query performance that is expected from the system. This dimension also considers the ability to scale up as the ingested data volume increases over time. The use case for the data warehouse outlined in this paper is non critical and therefore the solution is trying to achieve a maximum query time of 5 minutes per query.
* **Maintenance Effort**: The effort to be put in by the organization for the operation and maintenance of the proposed solution. The organization that is being discussed in this paper does not have a dedicated data engineering team. For this reason, it would be favorable to use a set of technologies that have a shallow learning curve.
* **Cost of the solution**: In this dimension we are concerned with the operational, maintenance cost as well as the implementation cost for the proposed solution. Due to organization policies, the use of subscription based cloud technologies is discouraged.
* **Integrations**: In this dimension, we consider the effort required for the integration of the proposed solution with the already existing systems and infrastructure. When organizational policy and existing hardware infrastructure are taken into account, it is observed that an on-premise solution on top of existing hardware is more feasible.

Based on above dimensions, the following technologies were chosen to realize the proposed data warehouse solution.

* **Microsoft SQL Server Integration Services (MS SSIS)** tool is proposed as the ETL tool. The existing expertise in administration of MSSQL server software, its minimal code approach to the ETL process and widely available documentation are considered as contributing factors to this decision.
* **MSSQL Server** was chosen as the data store for the data warehouse. There were few other competitors such as Oracle and SAP (SAP BW/4HANA) that provide on-premise data warehouse solutions. Existing in-house expertise for MSSQL server deployment, administration and high cost of implementation as well as the need for specialized hardware were big driving factors behind this choice.
* **Microsoft Power BI report server** on-premise was chosen for the reporting layer. The MSSQL server used for data warehouses can be seamlessly integrated as the backend to provide KPIs and reporting functionalities that are defined in section V.

The below diagram shows the integration between the proposed technology stack.

**Figure 4 : Proposed Technology Stack**

****

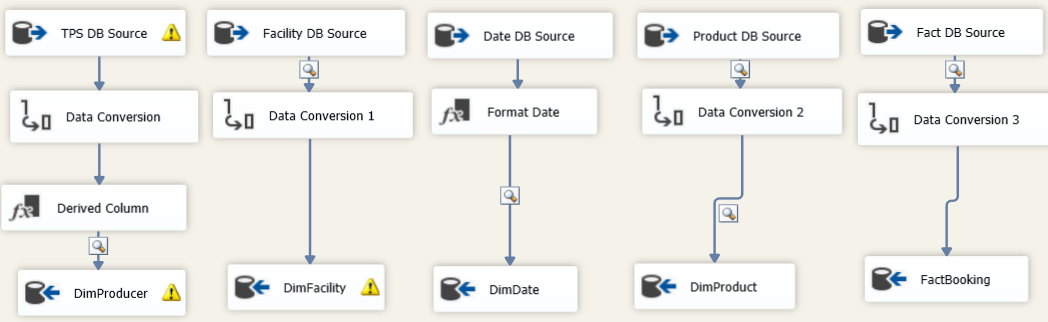
1. **EXTRACT-TRANSFORM-LOAD (ETL) PROCESS**

Extract-Transform-Load (ETL) extracts data from various sources, transforms them into different formats aligning to business rules and then loads it to a data warehouse or similar component [1].

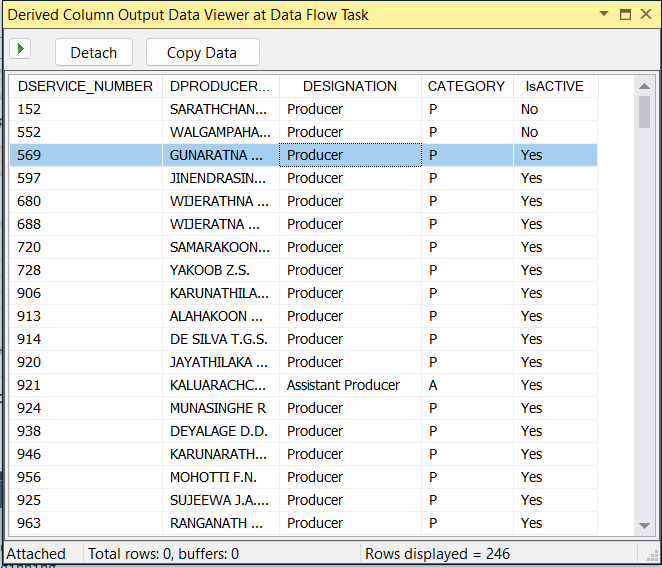
The solution outlined in this paper uses SSIS to extract data from two MS SQL relational databases and transform them into star schema. The ETL process starts with the creation of a SSIS package, which is considered as a unit of execution in SSIS. The source databases are configured as OLE database sources using the connection manager of the SSIS tool. In the data flow section of the SSIS package definition, the solution makes use of features such as merge join transform to join tables from independent data sources. Once this transformation process is completed, the transformed data is loaded to the destination data warehouse which is configured as an OLE DB destination. In addition to this, basic audit logs will be enabled to monitor the insertions to the destination data warehouse.

**SSIS Process:**

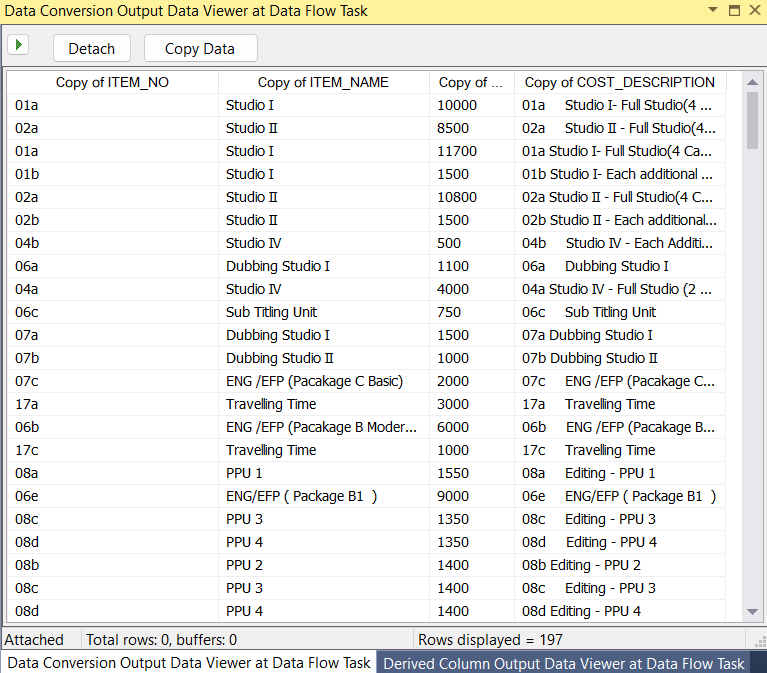
**Figure 5: ETL Processes for Dimensions and Fact Before Surrogate Key Lookups**



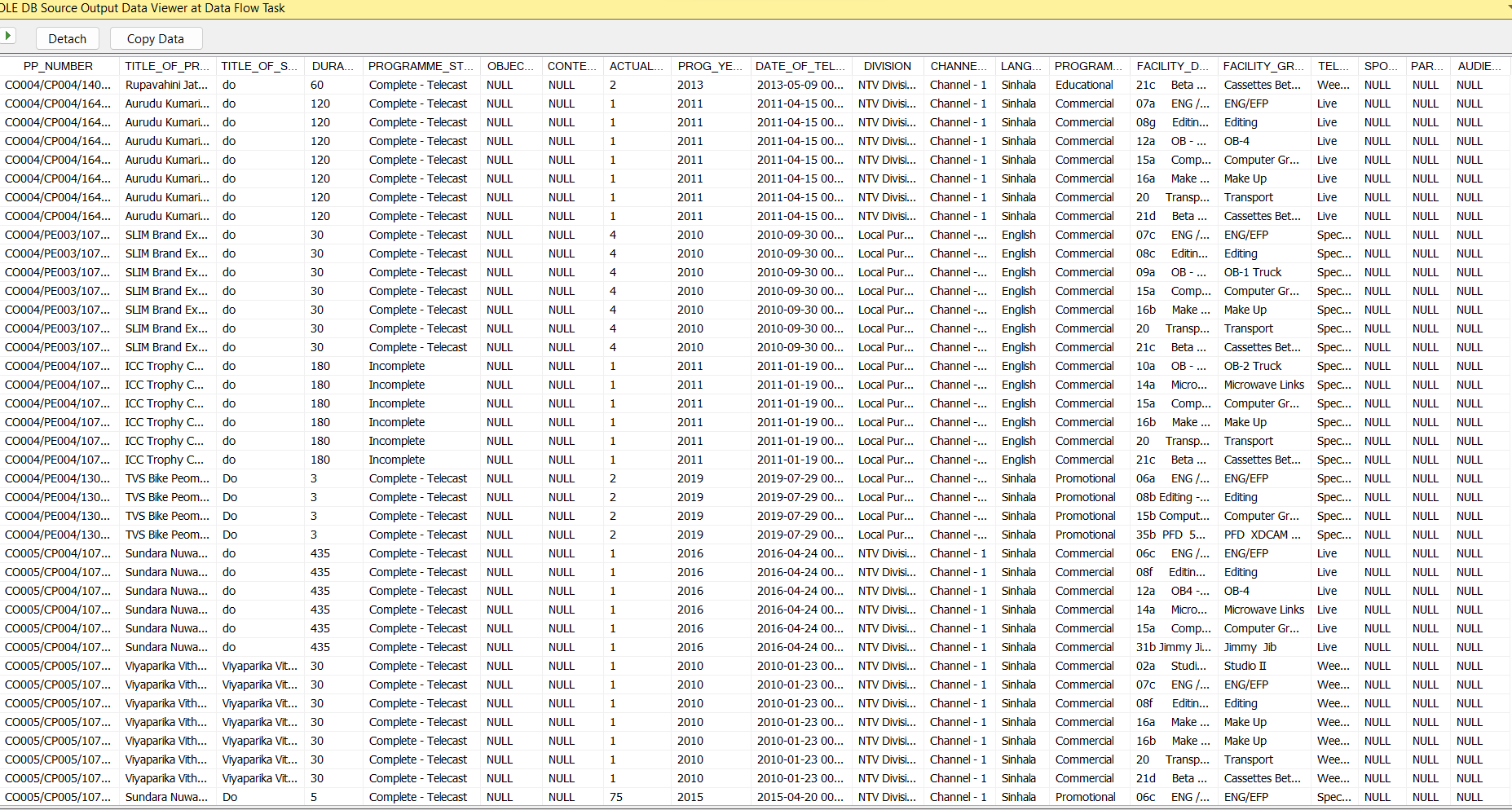
**Figure 6 : ETL Processes Output DimProducer**



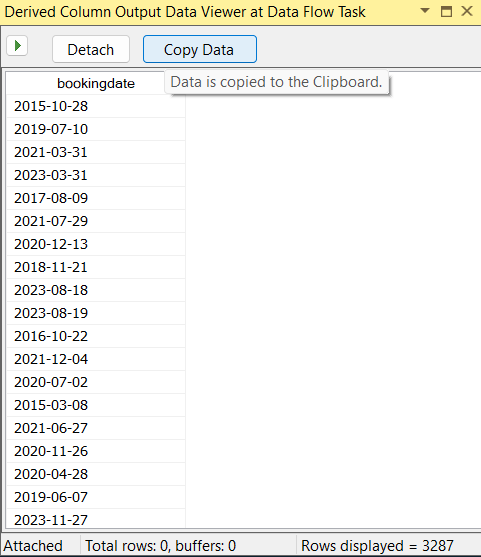
**Figure 7 : ETL Processes Output DimFacility**



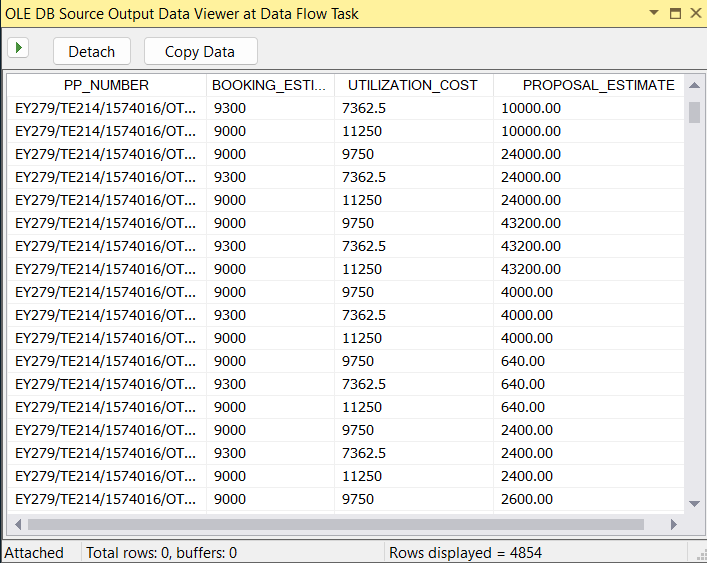
**Figure 8 : ETL Processes Output DimProduct**



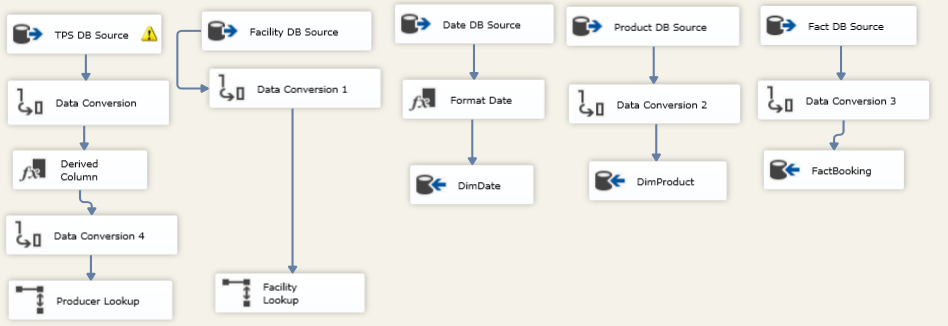
**Figure 9 : ETL Processes Output DimDate**



**Figure 10 : ETL Processes Output FactBooking**



**Figure 11: ETL Processes Few Surrogate Key Lookups**



Finally needed to write all dimension data into the fact table.

1. **ANALYSIS**

| **Business Problem** | **Reports & KPIs** |
| --- | --- |
| Q1. What facilities are highly utilized in program production? | The utilization cost per each facility has been measured and analysed. |
| Q2. Which productions have over utilized facilities more than the regular estimates? | The productions with the highest amount of over utilization have been measured by comparing utilized cost vs estimated cost per each program production. |
| Q3. Which productions do not utilize facilities that are reserved for a particular period? | The productions with no utilization have been analysed by filtering the utilized cost of program productions for a given month. |
| Q4. How much estimate and actual cost differentiation are in program production? | The cost difference of actual utilization cost and estimated cost have been analysed by comparing the cost difference of actual and estimated with the actual cost. |

The following figures show the analysis of the program production utilizations and production programs for decision making purposes of the management.

**Figure 12: Analysis of Utilization of Facilities in Production**

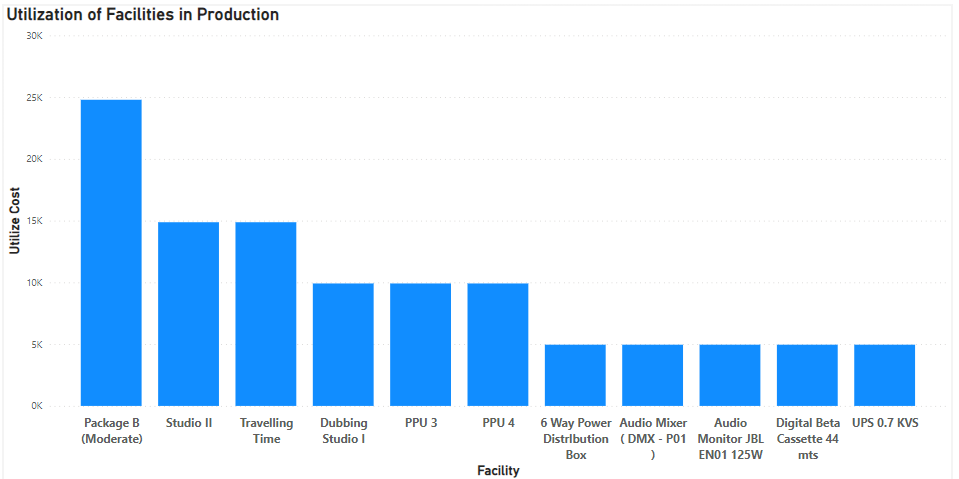
****

Figure 12: The analysis of utilization of facilities depicts the mostly utilised facilities of the broadcasting company. These insights can be used to understand how the facilities are used for program production utilizations and what facilities are mostly used in the program productions.

**Figure 13: Analysis of Program Productions Utilization**

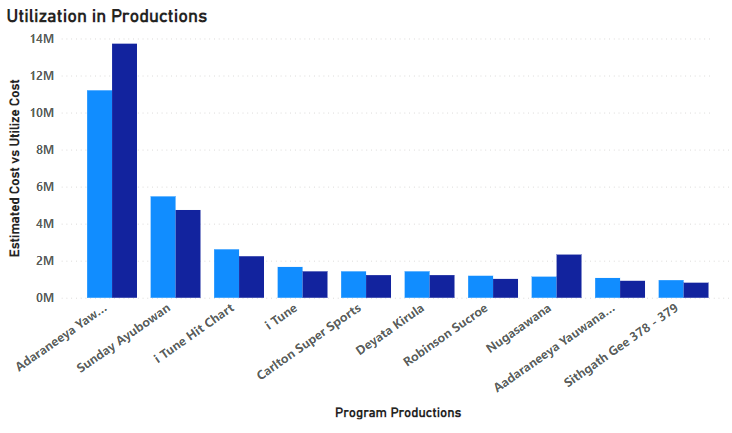
****

Figure 13: Utilization in productions shows the program productions with high utilization cost compared to its estimated cost. This analysis will help the management to understand how the actual cost incurred surpasses the estimates for different program productions and come up with more accurate estimates with regards to program productions created by the company.

**Figure 14: Analysis Program Productions with no Utilization**

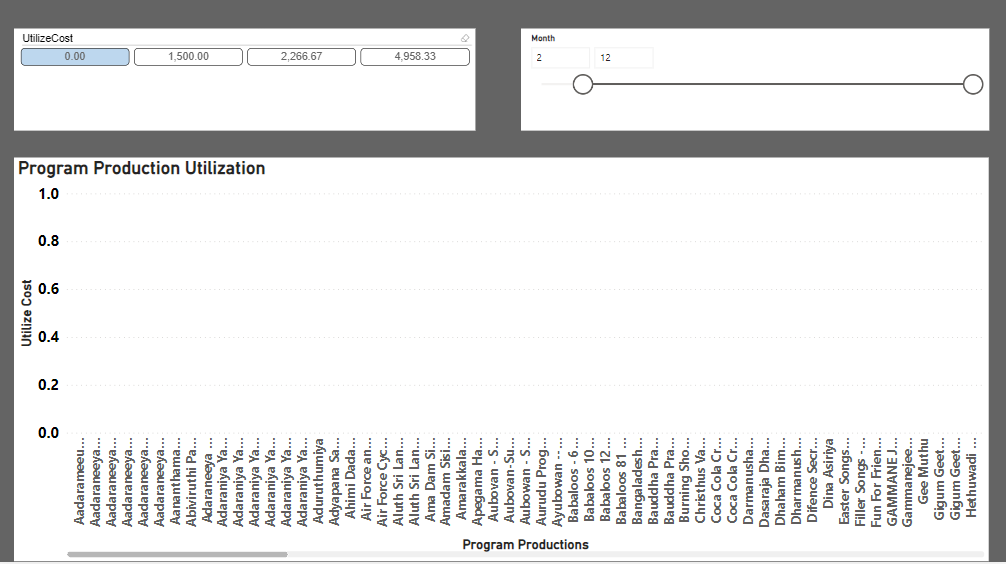
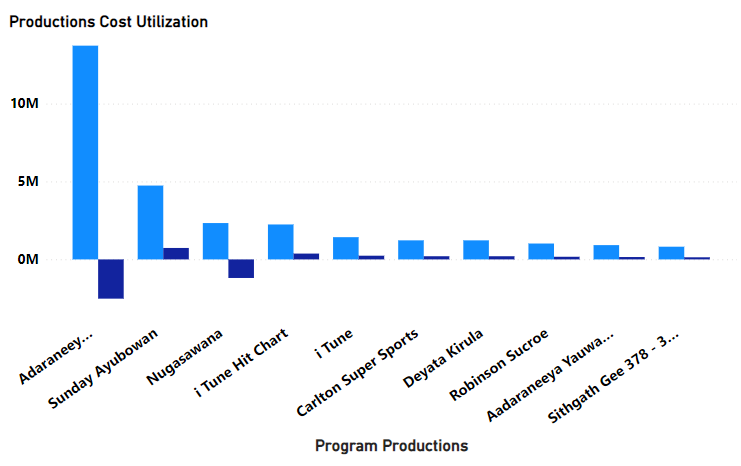


Figure 14: The analysis of program productions with no utilization helps in understanding what are the programs that the facilities are booked but not have been used efficiently for a given month or kept idle. By identifying these program productions will allow the management to prioritise the productions that need the facilities more often than these productions and schedule the facility booking times accordingly.

****

**Figure 15: Analysis of Cost Difference in Utilization Costs in Program Production**

Figure 15: The analysis of cost difference in utilization costs in program production, would help in understanding which programs had more cost difference with compared to the estimated cost and as shown in figure 6 these two analysis will help in making more facility booking time and other resources available for the program productions with high cost difference such as Adaraneeya Yawwanya and Nugasewana as these have high cost differences compared to other program productions.

1. **CHALLENGES**

Building a productive data warehouse for this type of scenario is highly challenging as it is difficult to find existing cases. We mapped real world production environment instances which match this study.

As designing and implementing any analytical system this also has technical and non technical challenges [7]. Most of the time users are reluctant to explain what they exactly need. Source data is kept in multiple database technologies. Data tables are designed in an ad-hoc manner, therefore there are many inconsistencies, incomplete, duplicate and reasonable amounts of null data.

Further in OLTP there are no properly defined key columns as well as relationships, but used by the TPS. There is no unified data capturing process as some instances utilization data is arrived into OLTP before the facility allocation for pending approvals.

1. **CONCLUSION & FUTURE WORKS**

The data warehouse designed for this program production process can be used as the basis to gain insights on how the facilities are reserved, how they have been utilized, and the difference between the total estimated and utilization cost. This can be categorized based on the persons, facilities, programmes for a particular timeline. Output can be implemented in the form of reports and graphical visualizations for the decision makers.

It is really important to understand that building a data warehouse requires a considerable amount of time, effort and cost. The benefits obtained from the data warehouse can be evaluated from the KPIs we have defined already. But still the amount of cost we spent on building the data warehouse and the revenue we have obtained afterwards from it should be acceptable as well as profitable to the organization ultimately. Implemented data warehouses should help in further innovative ideas as far as it is relevant to technological advancements.

In this paper, we have described the model for implementing a data warehouse system to a government owned television broadcasting corporation. Considering the selected business process of the television broadcaster we can further implement other functional areas and business processes as required in future or when necessary.

For future development we need to fine tune this project to facilitate addition and attrition of data sources to avoid major redesign. Data marts can be proposed for different departments.

**REFERENCES**

[1] A Survey of Extract–Transform–Load Technology by Panos Vassiliadis. [International Journal of Data Warehousing and Mining (IJDWM)](https://www.igi-global.com/journal/international-journal-data-warehousing-mining/1085) Vol. 19, 2009, 01-27

[2] Ideal Strategy to Improve Data Warehouse Performance by Fahad Sultan & Dr. Abdul Aziz. (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 409-415

[3] Fundamentals of database systems. 4th Edition. Persons international and Addison Wesley. Ramez Elmasri and Shamkant B. Navathe

[4] A. A. G. O. K. A. K. O. S. Nyoma Darma Kotama, "Design of Data Warehouse for University Library using Kimball and Ross 9 Steps Methodology," International Journal of Engineering and Emerging Technology, vol. 4, no. 1, pp. 37-40, 2019.

[5] C. J. C. S. K. D. PPG Dinesh Asanka, "Design and Implementation of Data Warehouse for a Higher Educational Institute in Sri Lanka," 2021.

[6] B. Gawin and B. Marcinkowski, “Business intelligence in facility management: Determinants and benchmarking scenarios for improving energy efficiency,” *Information Systems Management*, vol. 34, no. 4, pp. 347–358, 2017.

[7] S. Orlando, R. Orsini, A. Raffaeta, A. Roncato, and C. Silvestri, “Trajectory Data Warehouses: Design and Implementation Issues,” Journal of Computing Science and Engineering, vol. 1, no. 2. Korean Institute of Information Scientists and Engineers, pp. 211–232, 2007.