ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse131942020.pdf
https://doi.org/10.30534/ijatcse/2020/131942020

An Agile Dimensional Data Mart Architecture for Clinical Laboratory Towards the Development of an Evolving Enterprise Clinical Data Warehouse

Joseph George [1], Dr. M.K Jeyakumar [2]
[1]Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education Kumaracoil, Tamilnadu, India
jg.joseph@hotmail.com
[2]Professor, Department of Computer Applications, Noorul Islam Centre for Higher Education Kumaracoil, Tamilnadu, India
jeyakumarmk@yahoo.com

ABSTRACT

Data science and business intelligence streams are getting increasingly popular in current era and the reason behind this trend is the hidden potential of processed data. Data warehousing solutions, whether it is on premise or cloud based, are getting more and more popular, as it is creating the foundation for a crucial business intelligence solution. Organizations who are exploring the power of data, are well equipped with the business challenges moving forward. Even though the terms Data Warehousing and Business Intelligence are pretty common and widely used, unlike other industries healthcare is bit slow in reaping and adopting the power of the data, even though the data volume in healthcare is higher than any other industries. In this paper, authors are trying to develop a dimensional data model for one of the widely used healthcare data segment, which is the laboratory services. This is an effort towards the development of agile Enterprise Data Warehouse for healthcare. The base of this research is Kimball's agile development methodologies and a phased approach is utilized in the study.

Key words: Clinical Business Intelligence, Data Warehouse, Data Analytics, Agile Data Mart, Lab Informatics

1. INTRODUCTION

One of the most famous management quotes by Peter Ducker says "you can't manage what you can't measure.". Informed decision making requires precise data, which is refined to produce the intelligence and then wisdom. Healthcare industry is very unique in many aspects and thus the data associated with it too[1]. There exists two prominent data warehousing(DW) model, Inmon's Corporate Information Factory or the top down approach and Kimball's Dimensional Modelling or bottom up approach[2], [3]. Many industries are gone a long way with implementation of data warehouse and surprisingly healthcare is still way behind in the effective utilization of the collected data. Healthcare is the forerunner in terms of data collection and the amount and variety of data collected in healthcare is far ahead of any other industry[4].

Due to the changing environment and the data complexity, often a top down approach of having a comprehensive DW blue print to start with is often difficult in healthcare due to high initial investment and less return of investment[5]. Healthcare data shows the possibility of doing agile and independent data marts and accumulating them to the enterprise DW[6]. Healthcare has two major kinds of data sets, clinical and administrative. In administrative data set, we often talk about commercial aspects and the back-office activities. Clinical data is spread across various systems and one of the key elements of clinical data is the laboratory reports[7]. That is the reason why, that specific subject is taken for this study.

In this study, authors are developing an agile dimensional data model for one of the most important clinical data section, which is the laboratory data. The paper is organized in the following manner. First, we will be analyzing the lab workflow with a concentration on the data capture and processing activities. From there, the Kimball's agile model steps will be utilized to make the foundation of the research. The resulted data mart, which is in the form of a star schema, will be used for the visualization aspects, which will be again based on Power BI.

2. LABORATORY WORKFLOW

Clinical laboratories ate inevitable part of any healthcare environment, as it provides a wide range of support to physicians and care givers during the diagnosis and the treatment process. Starting from the blood drawing till the signing off the test result, there exist a well-coordinated process flow, which facilitate the timely reporting of the requested tests. One of the major section of "Objectives" category of SOAP Notes (Subjective, Objective, Assessment and Plan) is the Laboratory Data[8]. In most of the cases, majority of the tests are conducted within the inhouse laboratory and the complicated teste are sent out to reference labs or specialized diagnostic centers.

As a first step to analyze the data flow, let us have a deep dive into the lab workflow[9]. Figure 1 shows the different processes associated within the lifecycle of a lab test.

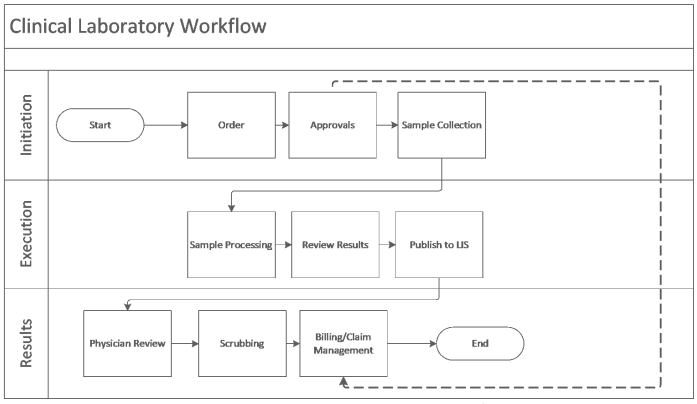

Figure 1: Clinical Laboratory Workflow

Sample Collection

Patient has a valid encounter and as a part of the treatment process, a set of lab tests are prescribed by the physician. This is normally recorded in the electronic medical records (EMR). This could be an order set or individual tests. The lab work flow starts with the sample collection, also the TAT (Turnaround Time) starts ticking from here[10]. Prior to sample collection, it is assumed that the approvals and the related insurance activities are taken care.

Sample Processing

Once the sample is ready and the accession number is generated, the processing can happen either inhouse or the complex tests to be performed outside. In case of external tests TAT to be mutually agreed.

Review of the Result

Lab devices are mostly integrated with the LIS (Laboratory Information System). Once the samples run through the devices, the output(results) are transferred to the LIS for review by the lab technologist. This is the phase where the Panic Value Notification or rerun decisions are taken.

Physician Review

Once the technologist verifies and release the result, it is available in the EMR for the physician to view.

Claims and financial activities

Laboratory informatics is not an independent area in care delivery, it has tight dependency with all other care giving processes. For data analytics purpose, we are creating Lab as a separate Data Mart which will in turn will be connected with other related data marts to constitute the enterprise data warehouse.

3. METHODS

As we know, needs of healthcare analytics are fast changing. To keep pace with healthcare's analytics needs in today's fast forward and speedy world, we need to use a data warehouse model that is agile enough. Ralph Kimball's dimensional model (which will be used in this research activity) proposes 4 different steps:

- Identification of Business Processes
- Grain Definitions
- Identification of Dimensions
- Identification of Facts

Thus, this research work is based on the Kimball theory, of the Bottom Up method, with a few modifications to achieve better quality, agility and results.

3.1 Step # 1 : Identification of Business Processes

When a set of activities are well coordinated and combined together to produce the desired output, we call it a process. Figure 1 shows the activities associated with a lab workflow. In an agile Data Warehouse development, which consists of related data marts, facts are derived from the key activities of the list of interrelated processes and the dimensions from the dependency of the activities. Let us find out the processes involved in the specified workflow.

Process 1: Encounter Centric

Any visit of a patient to the care facility can be termed as an encounter or episode. Every encounter will have a unique identifier. Encounter ID together with the patient identification number formulate the combined primary key identifier for any visit. Encounter activities start from the time of appointment booking and completes the process once the checkout is done. In a lab workflow specific scenario, the process starts when the physician order the test[11]. It could be a part of an order set or could be a stand-alone test. The output of this process is the input for the next process, which is the "lab test transaction".

Process 2: Test Examination (Clinical test examination)

Once the order is placed the lab test transaction (test examination) starts. It starts with the blood drawing and completes when the test result is pushed to LIS (Laboratory Information System). This is the key process of the laboratory workflow and the output of this process is the specific value of the prescribed test, which is often shows with a reference range[12].

Process 3: Revenue specific

One of the inevitable process is the financial related, say Revenue Cycle Management (RCM). It starts with the pre-authorizations and ends with the bill settlement. This is mostly covered in financial data marts and we are not considering it here[13],[14].

3.2 Step # 2 : Grain definitions

The grain of the lab workflow shall be defined as:
Patient – Encounter – Order set – Order – Test examination.
So the granular data level is the individual test and the data and the processes associated with it. The granularity of data in lab flow will be encounter specific and then lab examination specific[15]. Every test will have unique accession number and the atomic level of activity in the "subject of interest" is the individual test.

3.3 Step #3: Identification of Dimensions

In healthcare, dimensions are more or less same for all data marts be it RCM or OP or IP and Ancillary services. In this paper we are concentration only on those dimensions specific to the lab workflow.

- Date
- Invoice
- Patient
- Doctor
- Lab Staff
- External Lab
- Payer
- Procedure
- Diagnosis
- Department

3.4 Step #4 : Identification of Facts

In a DW implementation, Facts are identified from the process flow. The key activity of any process turned out to be a fact. Identification of Facts are often done by utilizing the activity theory. In order to identify the facts, the key question to be answered is " What are the KPIs we are trying to analyze with this data mart?". Table 1 lists the key performance indicators of a lab process[16]. The list of KPIs comes under three major categories of data section and the first two categories are directly being into the lab workflow. Based on that, we can identify two Facts in this scenario.

Episode Fact: This is the patient foot fall measure, treating every visit of a patient as an event. This fact table will be long enough to record every transaction details of the specific visit. The width of this table depends on the granularity and the dimension needs of the analytics. This is a "fact less" fact table or an event.

Lab examination Fact: Every tab test irrespective of whether it is being a part of an order set or not, will constitute the lab examination fact table[11].

3.5 Technology Used

In this research implementation , Microsoft Business Intelligence Stack is utilized. For ETL activity Microsoft SSIS (SQL Server Integration Services) is utilized[17]–[19]. SQL Server Database is the Data warehouse platform and for visualization purpose, Power BI is used.

Facts	Dimensions									
	Date	Invoice	Patient	Doctor	Lab Staff	External Lab	Payer	Procedure	Diagnosis	Department
Encounter	1	0	1	1	1	0	1	0	1	1
Lab Test Transaction	1	1	1	1	1	1	1	1	1	1
Billing Activities	1	1	1	1	0	0	1	1	1	0

Figure 2: Facts and Dimensions connectivity

4. RESULTS AND DISCUSSIONS

In healthcare environment a top down data warehouse implementation is practically much difficult, and the return of investment also is pretty hard to achieve. Reason being, the healthcare environment is a fast-changing industry in terms of analytics requirements and regulatory bindings. Ralph Kimball's bottom up approach is in alignment with the agile development and DevOps[20],[21]–[23]. A set of well-defined and interrelated data marts leading to the enterprise DW is the practical scenario in healthcare data arena. Kimball's 4 step approach is used for the identification of Facts, Dimension and grains.

Figure 2 shows the connection between the identified facts and dimensions of the Clinical Lab Workflow. The identified dimensions will be shared between multiple data marts in the enterprise data warehouse. Resulted Star Scheme with the identified Facts and Dimensions ate represented in Figure 3.

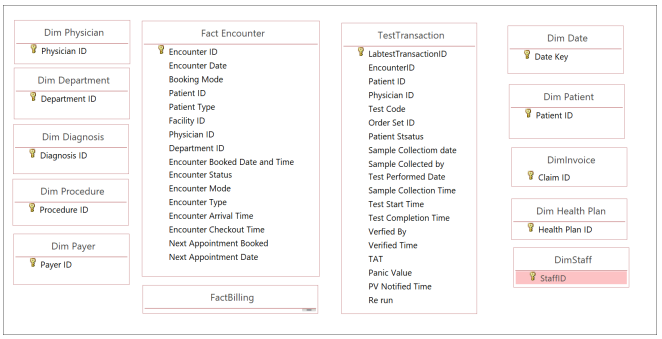

Figure 3: Star Schema for Clinical Laboratory Dimensional Model

4.1 Laboratory Dimensional Data Model design explained

The first consideration is how to implement the dimensional model. This can be implemented either in an RDBMS environment or in an OLAP environment. In this research we are considering the RDBMS environment and the resulted model is a star schema[24]. The same facts and dimensions can be implemented in a multi-dimensional database leading to the creation of OLAP Cubes.

Figure 3 shows detailed list of columns of the fact tables, but dimensions are represented just as a box to avoid congestion in the diagram. Facts and dimensions are linked with the primary key and foreign key. The connection matrix is explained in Figure 2. Wherever an intersection of Facts and

Dimensions is filled with "1", that means there exists a connection between the two tables.

Date dimension is treated in s different way altogether. Date is used as a "Role Play" mode. That means, the same date data will be used for different purposed. For that reason, multiple views will be created based on the need and the created view will be linked to the fact. For example, encounter date and the lab test completed date, both are "date" in nature. But for easy reporting and drill down, appropriate views will be created from the master Date dimension table.

4.2 Data Flow Diagram

Figure 4 shows the direction of data flow from the source to the data consumer. In this scenario, the source systems are the EMR (Electronic Medical records), LIS (Laboratory Information System) and the ERP (Enterprise Resource Planning) systems. ETL(Extraction – Transformation – Loading) activities on the source systems lead the data to the staging area and from there to the Lab dimensional model, which is a part of the enterprise DW[25]. All the dimensional data marts are connected to each other, in this case the lab mart will be joined to other marts like RCM, OPD, IP etc. The blue sockets are representing the agile nature and the modular socket feature.

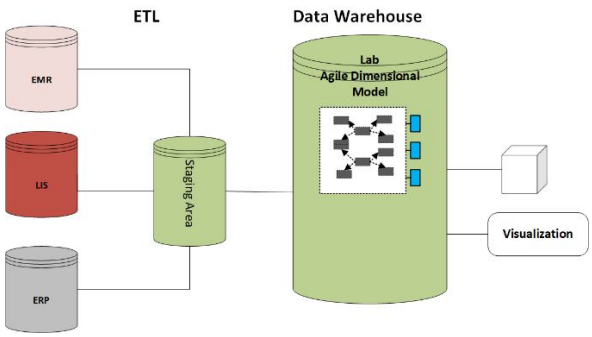

Figure 5: Data Flow Diagram – Lab Data Mart

4.3 Dashboards and KPIs

Business intelligence projects concludes its lifecycle with the brilliant visualization of the processed data. This presentation layer is often referred to as self BI and it demands the features of drilling down and rolling up along with different modes of slicing and dicing across the dimensions[26]. The major KPI reports often come in lab scenario are summarized in table 1. These are the most frequently uses KPIs and of course the proposed model is equipped with any sort of analysis of the fact datta across the defined dimensions.

Figures 6-7 represents the sample dashboards created in Power BI utilizing the underlying data marts. Please note sample data is used for the visualizations to comply to the regulations.

Table 1: Few Key Performance Indicators of Clinical Laboratory

Areas of interest		
Encounter Specific	Process Execution	Financial
Test Volume	Volume – Blood Transfusion	Revenue/Department/ Physician
Inpatient Vs Outpatient Test Volume	Test volume - Microbiology	Revenue/Facility
Tests within order set	Test Volume - Chemistry	Insurance Rejection Rate
Test volume per department	Test Volume – Hematology	Rejection Reason
Test volume per physician	TAT – Inhouse Tests	
Lab resource utilization	TAT – External Lab	
Test volume within normal range	Top Lab procedures	
Test volume done externally	Panic Value Notifications	
PoC (Point of Care) Test Volume		

Figure 6: Sample Dashboard 1

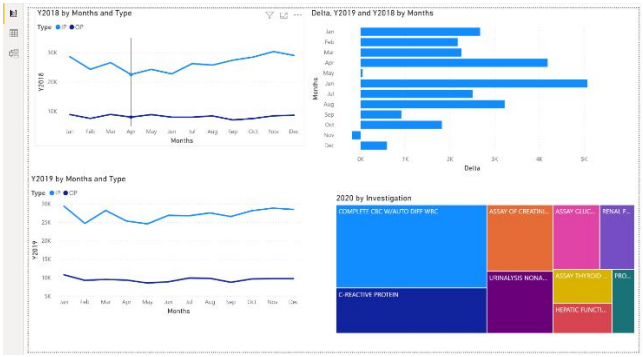

Figure 7: Sample Dashboard 2

5. CONCLUSION AND FUTURE WORK

Data is the new oil and the timely availability and interpretation of the accumulated data leads to informed decision making. This is true in any industry and any

department. Here comes the role of self BI. Static report generation and depending on IT team to create more and more reports, which leads to an ocean of unused and insignificant reports, is now moving away and the self BI is now getting into the driving seat.

It is estimated that a proper utilization of consolidated data mart can reduce the time for the root cause analysis by more than 90%. This is also leading to the identification of optimized process flow and better adoption[27].

Healthcare domain is the best example where agile DW and Self BI can be best utilized. Authors are in the process of implementing a layered DW approach, with includes the architecture to combine the individual data marts and to move the common entities into a shared space, thus leading to single source of truth with less redundancy.

ACKNOWLEDGEMENT

We would like to thank Mr. Ramachandran and Mr. Gopakumar for their support and motivation throughout the research work.

REFERENCES

[1] G. S. Nelson, T. Technologies, and C. Hill, **A Practical Guide to Healthcare Data : Tips , traps and techniques**, *Thinking data*, vol. 1, no. 1, pp. 1–20, 2017.

[2] I. Inmon;Claudia, *Corporate Information Factory*, 2nd ed. 605 Third Avenue, New York: John Wiley & Sons, 2001.

[3] M. R, Kimball;R, *The Data Warehouse Toolkit - Second Edition*, 2nd ed. 605 Third Avenue, New York: Wiley Computer Publishing, 2002.

[4] Dell Technologies, **Solutions for Healthcare**, 2019.

[5] M. Karami, A. Rahimi, and A. H. Shahmirzadi, **Clinical Data Warehouse**, *The Health Care Manager*, vol. 36, no. 4, pp. 380–384, 2017 doi: 10.1097/HCM.0000000000000113.

[6] S. Dhir, D. Kumar, and V. B. Singh, **Success and Failure Factors that Impact on Project Implementation Using Agile Software Development Methodology**, in *Advances in Intelligent Systems and Computing*, vol. 731, Springer Singapore, 2019, pp. 647–654.

[7] R. Jackups, **Answering Unanswerable Questions in the Clinical Laboratory with Data Warehouses**, *Clinical Chemistry*, vol. 65, no. 12, pp. 1471–1473, Dec. 2019 doi: 10.1373/clinchem.2019.311654.

[8] M. L. Bayot and P. Naidoo, *Clinical Laboratory*. StatPearls Publishing, 2020.

[9] C. and L. S. Institute, *CLSI. Process Management*, 1st ed. QMS18 Process Management, 2015.

[10] M. Miler, N. Nikolac Gabaj, L. Dukic, and A.-M. Simundic, **Key Performance Indicators to Measure Improvement After Implementation of Total Laboratory Automation Abbott Accelerator a3600**, *Journal of Medical Systems*, vol. 42, no. 2, p. 28, Feb. 2018

[11] A. N. Crowson, M. Harvey, and S. Stout, **Data warehouse strategies and the modern anatomic pathology laboratory: Quality management, patient safety, and pathology productivity issues and opportunities**, *Seminars in Diagnostic Pathology*, vol. 36, no. 5, pp. 294–302, Sep. 2019, doi: 10.1053/j.semdp.2019.05.001.

[12] A. Sterrett, M. Bailey, G. Saylor, and M. Raebel, **PS2-42: Developing High-Quality Laboratory Results for the Virtual Data Warehouse: The Importance of Single-site Quality Assessment**, *Clinical Medicine & Research*, vol. 10, no. 3, pp. 193–193, Aug. 2012 doi: 10.3121/cmr.2012.1100.ps2-42.

[13] M. A. C. R. Green, *A Guide to Billing and Reimbursement*, 14th ed. Clifton Park, NY: Delmar Cengage Learning, 2018.

[14] S. L. Visscher, J. M. Naessens, B. P. Yawn, M. S. Reinalda, S. S. Anderson, and B. J. Borah, **Developing a standardized healthcare cost data warehouse**, *BMC Health Services Research*, vol. 17, no. 1, p. 396, Dec. 2017, doi: 10.1186/s12913-017-2327-8.

[15] R. D. Aller, **The Clinical Laboratory Data Warehouse**, *American Journal of Clinical Pathology*, vol. 120, no. 6, pp. 817–819, Dec. 2003, doi: 10.1309/TXXABU8MW75L04KF.

[16] J. G. Franco *et al.*, **Key performance indicators score (KPIS score) based on clinical and laboratorial parameters can establish benchmarks for internal quality control in an IVF/ICSI program**, *Fertility and Sterility*, vol. 108, no. 3, p. e91, Sep. 2017 doi: 10.1016/j.fertnstert.2017.07.282.

[17] J. P. A. Runtuwene, I. R. H. T. Tangkawarow, C. T. M. Manoppo, and R. J. Salaki, **A Comparative Analysis of Extract, Transformation and Loading (ETL) Process**, *IOP Conference Series: Materials Science and Engineering*, vol. 306, no. 1, 2018 doi: 10.1088/1757-899X/306/1/012066.

[18] S. Young Lee, **Architecture for Business Intelligence in the Healthcare Sector**, *IOP Conference Series: Materials Science and Engineering*, vol. 317, no. 1, p. 012033, Mar. 2018, doi: 10.1088/1757-899X/317/1/012033.

[19] Microsoft, **Data Factory - Data Integration Service**, *Azure*, 2020. https://azure.microsoft.com/en-us/services/data-factory/ (accessed May 16, 2020).

[20] I. Bibik, *How to Kill the Scrum Monster*. Berkeley, CA: Apress, 2018.

[21] M. R, Kimball;R, *The Kimball Group Reader*, vol. 1, no. 1. Crosspoint Boulevard Indianapolis, IN: John Wiley & Sons, 2016.

[22] D. Linstedt and M. Olschimke, *Data Vault 2.0*. Elsevier, 2016.

[23] Microsoft, **Building a Modern Data Warehouse with Microsoft Data Warehouse Fast Track and SQL Server**, 2017.

[24] Y. Hu *et al.*, **A Nonrelational Data Warehouse for the Analysis of Field and Laboratory Data From Multiple Heterogeneous Photovoltaic Test Sites**, *IEEE Journal of Photovoltaics*, vol. 7, no. 1, pp. 230–236, Jan. 2017 doi: 10.1109/JPHOTOV.2016.2626919.

[25] Z. Azadmanjir, M. Torabi, R. Safdari, M. Bayat, and F. Golmahi, **A Map for Clinical Laboratories Management Indicators in the Intelligent Dashboard**, *Acta Informatica Medica*, vol. 23, no. 4, p. 210, 2015, doi: 10.5455/aim.2015.23.210-214.

[26] B. Harry, **54 Application of a Data Analytics Dashboard for a Single-Institution Laboratory Information System**, *American Journal of Clinical Pathology*, vol. 149, no. suppl_1, pp. S192–S192, Jan. 2018, doi: 10.1093/ajcp/aqx149.423.

[27] A. A. K. Hamoud, **CLINICAL DATA WAREHOUSE: A REVIEW**, *Iraqi Journal for Computers and Informatics*, vol. 44, no. 2, pp. 1–11, 2018.