A Comparative Analysis of Data Integration and Business Intelligence Tools with an Emphasis on Healthcare Data

Joseph George[1], Dr. M.K Jeyakumar[2]

[1]*Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Kumaracoil, Tamilnadu, India*

[2]*Professor, Department of Computer Applications, Noorul Islam Centre for Higher Education, Kumaracoil, Tamilnadu, India*

[1]*jg.joseph@hotmail.com*

Abstract *- The heart and soul of any Business Intelligence suite is its ETL (Extraction Transformation and Loading) capability. Business Intelligence (BI) helps the organizations to take informed decisions with the help of properly formulated, known and unknown facts. Integrating data from multiple sources and transform it into a holistic data view for analysis is the core feature of any BI platform. There exists many tools and technologies to facilitate ETL and Data Warehouse activities. These tools vary a lot in terms of maturity and usability. ETL is expensive in terms of computing resources and time. A poorly designed ETL steps can take the entire BI solution into a toss.*

In current scenario, the term ETL is getting diminished and the term Data Integration Tool (DIT) is getting more popular and widespread. In this paper we will treat both terms synonymously. In an ETL or rather Data integration effort, business processes and domain knowledge are vital. This research paper is doing an analysis of the popular Data Integration tools with an emphasis on healthcare Data Warehouse domain. Healthcare industry is unique thus the data associated with it is much more unique than any other sector. This study is the primary step towards the end to end development of a healthcare business intelligence or Clinical Business Intelligence model.

Keywords - *Data Integration Tools, ETL, Business Intelligence, Data Warehouse, Clinical Business Intelligence, Data Visualization, Data Mart*

Abbreviations: ETL, Extraction Transformation and Loading; DW, Data Warehouse; PHI, Personal Health Identifiers; HIPAA, Health Insurance Portability and Accountability Act; BI, Business Intelligence

I. INTRODUCTION

As per the World Economic Forum report, it is estimated that approximately 463 exabytes of data will be created per day by 2025[1]. The speed at which business environment is changing is increasing day by day and here comes the real scenario of survival of the fittest. Organizations who can predict and adjust to the changing business environments remain in the race and others are becoming part of history. The once who anticipates trends and adjust themselves to the changing environment by modifying strategies and finding new opportunities survive. The vital partner in strategic development and formation is the new oil, which is "The Data"[2].

Data is spread everywhere, it can be either the data captured within the organization or the data which resides outside the business organization. The real value of data comes once we integrate all these together and transform it into a meaningful insight. This is the point where organizations achieve competitive advantage and the entire activities altogether termed as BI[3].

Business Intelligence (BI) is not an isolated technology but an umbrella of technologies and processes which comes together to deliver a desired outcome. Data warehouse is one of the major players in any BI spectrum[4]. BI is not at a luxury anymore and it becomes a necessity for the organizations in current competitive era[5].

ETL is the heart of any DW project and it is estimated that 60-80 % of BI activity is related to ETL. All ETL tools aim to process more data in less time[6]. This research paper in organized in the below format. First, we will touch base ETL in general and then with a focus on healthcare data. Then we will do a comparative study of the market leading ETL tools. HIPAA compliance and healthcare specific requirements will be touch based in this section. Finally, we shall come to a conclusion based on the business requirements and data scenario.

II. ETL

Any transaction system, like an airline booking system or e-commerce system, is always optimized for the best possible response time and performance. To achieve this, the underlying database will be normalized at its best. These systems are designed for day to day activities and at a time, these systems concentrate on a specific event or record[7]. When it comes to reporting or analysis scenario, the requirement is the analysis of huge amount of data predominantly historic data. Of course, performance is a factor here, but the expected response time is higher than that we get from an operational system.

For data analysis, we need to collect data from multiple sources, which are mostly optimized for day to day business operations, to have a holistic view of the business. ETL is the group of processes which extract data from multiple internal and external data sources specific to a business (Source Systems), do the transactional operations which includes cleansing, unifications, format change etc.(many a times a staging is utilized), then finally load into the destination , mostly a Data Warehouse.

In a very high level, ETL has 3 core processes which are the accumulation of source data , process it for consistency and integrity then load for further utilization[8]. But ETL can do much more than this. That is the reason in current scenario, we call it as Data Integration tools. Even Gartner in its reports use the term Data Integration Tools in the magic quadrant reports[9]. Modern data integration tools expand its reachability to the new requirements of hybrid and cloud integration and even data augmentation. When it comes to cloud data integration compared to on-premise, the cloud processes even eliminate the staging area as they have the huge computing power for the transformations in place itself. Figure 1 shows the graphical representation of a data integration tool.

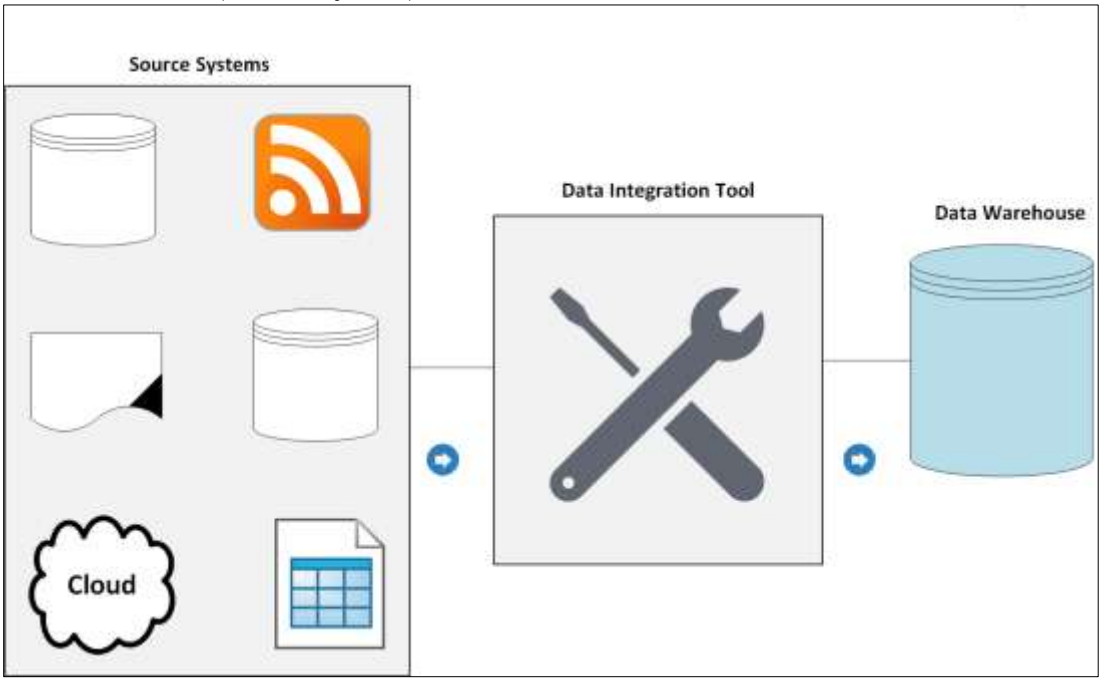

Fig. 1 Data Integration Activities

The data integration covers all the tasks and business rules associated right from data flow design, data transportation till the administration and operation of the tasks with metadata. The Data Warehouse institute rephrased the ETL acronym to EMTDLEA (Extract-Move-Transform-Document-Load-Exchange-Administer) [6].

The first part, which is the extract, will collect data from various operational systems of the business. This includes structured, unstructured, semi structured, big data, spreadsheets, CSV file, Data feeds etc. The source types and structure depend on the nature of the business and the scope of the data integration and the resulting DW and BI[10].

Even though transformation is termed as a single process, it is in fact a process container which includes a number of sub processes[11]. The sub process list includes, but not limited to the below list:

- Data Cleaning

- Standardization of data
- Data accuracy verification and integrity check
- Removal of duplications and anomalies
- Sorting and reordering
- Data quality improvements task
- and the list goes on.

Loading, which is the final process, loads the data into data warehouse. The loading methods depends on how and when you do the activity. It could be full loading, incremental loading or even a refresh loading[12]. With the vast computing power available in cloud now a days, even ELT is getting popular. First, we load the data and the same instance, without the need of a staging space, will do the transformation as the computing power is huge. A kind of ELTL is employed here.

III. MATERIALS AND METHODS

There exists hundreds of ETL tools and many of them satisfy only a portion of the ETL activities . In this research we are considering only those tools which are full-fledged data integration tools. For this, Gartner Magic Quadrant report of Q4 2019 is taken as a base.

The second benchmark used in this research is the KLAS report of Healthcare Business Intelligence and Analytics [13]. Even though KLAS report is describing the BI and Analytics part, there is a greater dependency between BI and DI Tools[10][14].

The comparison parameters which we will be utilizing for the comparison of Data Integration tools will be based on the above two reports, Gartner Magic Quadrant report and the 2020's Best in KLAS Report. Only the prominent tools, which are featured in the "Leaders" and "Challengers" magic quadrants are considered.

A. DIT Comparison Criteria

Data Integration Tools by definition is much more than an ETL Tool. Apart from the traditional extraction transformation and load, DIT includes the architectural methods, best practices and tools which consume, convert, consolidate and facilitate data from and across the various sources and formats[15]. DIT does not define a boundary for its operations, it could be within the organization and beyond. Current data spectrum for an organization could literally reside anywhere on the globe. Any activity which is associated with "the data" comes under the DIT platform[16].

A wide variety of criteria starting from the basic data capture ability up to SOA (Service-oriented Architecture) enablement capabilities are considered in the comparison process of the solutions. Listed below the core parameters used for the evaluation in the study. As specified, the last two criteria are specific to healthcare domain.

- Connectivity range and mode (Source and Destination)
- Advanced metadata management
- Data governance structure
- Master Data Management capabilities
- Data conversion and transformation capabilities
- Development environment
- Data delivery methods (batch data movement, utilization of data virtualization, synchronization etc)
- Change data capture (CDC) support
- Hybrid integration capability (on premise, cloud and intercloud)
- HIPAA compliance [17]
- Health Level Seven International (HL7) interaction capabilities

There exists a huge number of solutions, opensource and proprietary which satisfies partially or fully the selected criteria. All the solutions were compared in a neutral way concentrating only on the features without any bias or prejudice.

Based on the DIT capabilities and the Data Integration scenarios, the blow platforms came on top of the outcome of the comparison process.

- Informatica PowerCenter
- Oracle Data Integrator
- SAS Data Management
- IBM InfoSphere
- Talend Data Integration
- Microsoft SQL Server Integration Services (SSIS)/ Azure Data Factory
- SAP Enterprise Data Management
- Qlik/Attunity

B. Healthcare Business Intelligence & Analytics capabilities

Based on the Business intelligence and Analytics capability for healthcare domain, KLAS report defines the best in KLAS options and the top solutions in this category. As we know, Data integration without Business Intelligence and Visualization capabilities is incomplete. Since this research is concentrating specifically on healthcare data spectrum, lets evaluate the solutions in that directions too [18].

Healthcare deals with sensitive data and always demands highest level of privacy and confidentiality. Here comes the importance of PHI (Personal Health Identifiers) and HIPAA compliance (Health Insurance Portability and Accountability Act).

The HIMSS Analytics Adoption Model for Analytics Maturity (AMAM) is an evidence of the capability of a solution to utilize the data at its best[19]. Among the solutions listed above, SAS is the only HIMSS Adoption Model for Analytic Maturity (AMAM) certified solution [20].

Combining the DIT features and Healthcare specific Analytics and Business intelligence capabilities, now we have shrunk the shortlisted to the below narrow list.

- Microsoft
- Qlik
- SAS
- Oracle
- Informatica

Table 1: Gartner /KLAS comparison of DIT and BI solutions

Solution	Gartner Magic Quadrant (August 2019)	KLAS Ranking (2020)
Microsoft	Challenger	4
Qlik	Challenger	6
SAS	Leader	-
Oracle	Leader	-
Informatica	Leader	-

IV. RESULTS AND CONCLUSION

Microsoft's Data Integration solution is known as SQL Server Integration Services (SSIS). Microsoft Cloud platform "Azure" also offers cloud data integration feature, which is the Azure Data Factory (ADS). ADS is a hybrid Data Integration Platform [21]. One of the strong criteria which makes Microsoft to stands out is its hybrid deployment capability. SSIS and ADS combinations makes the scalability and performance unmatched. The best performance of SSIS is visible when the environment is Microsoft Centric. In contrary, complex data integration activities could lead to performance issues and challenges too.

The visualization platform called Power BI, is one among the best visualization tool. A combination of SSIS and Power BI makes a perfect match for Business Intelligence. A huge number of healthcare organizations are running their Clinical Business Intelligence on this platform.

Qlik

Qlik is well known for its Data Visualization and BI capabilities. Qlik acquired "Attunity" Data Integration Platform in Q2 2019. A combination of Attunity and Qlik make the end to end solution to an unmatched performance. Needless to say, Qlik as a standalone product for visualization is second to none. Attunity's strength lies in its simplicity to handle complex tasks. Attunity has a strong Change Data Capture functionality.

SAS

In Data Integration and Visualization category, SAS delivers one of the best performance. It is one of the best solutions for data manipulation, whether is integration or visualization. Self BI is the core strength of SAS.

Informatica

Informatics is a true leader in data integration spectrum. Application integration and data migration are the key strengths of Informatica. Informatica Intelligent Cloud Service (IICS) is the platform which Informatica utilize for Hybrid and multi cloud integration. Many organizations build their Data Warehouse and Data Marts on Informatica. The data transformation speed is amazing in Informatica. The development environment is a multi-user scenario and the learning curve is much less in informatica compared to other solutions. Informatica being a dedicated DIT platforms, is widely used in large enterprises and in mid sized organizations. It has a good catalog of integrated products, which together deliver the best data integration experience.

When it comes to integration with other technologies like Python, R etc. Informatica lacks a bit. Also the built in reporting capabilities are also not up to the mark

Oracle

When it comes to Oracle Data Integrator and Visualization, it has an edge on the technical capabilities and the vast integration availability. There exists a lot of aggregation functions which makes the Data Warehouse development much easier. The connection adaptors to ODBC and other data base engines are much stable.

As with any other Oracle platform, here also the learning curve is long and the usage is bit difficult for a new techies. A very deep knowledge is required to work with these platforms.

If we look into only visualization capabilities, Tableau is one of the best preferences. But in terms of data integration capability spectrum, it lacks the competency. In those scenarios, we use the DIT to develop the Data Warehouse and utilize the specialized visualizations tools to execute the visualizations. This could be the best approach we should adopt, as the business requirements and the size of the organization vary one to one.

ACKNOWLEDGEMENT

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

CONFLICT OF INTEREST

The authors confirm that there are no known conflicts of interest associated with this paper.

REFERENCES

[1] J. Desjardins, "*World Economic Forum. How much data is generated each day?*," 2019. Accessed: May 09, 2020. [Online]. Available: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/.

[2] Bernard Marr, "*Here's Why Data Is Not The New Oil*," *www.forbes.com*, 2018. https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/#3d4745013aa9 (accessed May 11, 2019).

[3] C. Ballard and D. M. Farrell, "*Dimensional Modeling : In a Business Dimensional modeling for easier data performance"*, 1st ed., vol. 1, no. 1. Menlo Park, CA.: IBM Redbooks, 2006.

[4] M. R, Kimball;R, "*The Data Warehouse Tool Kit - Dimensional Modelling"*, 3rd ed. 2013.

[5] B. Wieder and M. L. Ossimitz, "*The Impact of Business Intelligence on the Quality of Decision Making - A Mediation Model*," in Procedia Computer Science, 2015, vol. 64, doi: 10.1016/j.procs.2015.08.599.

[6] W. Eckerson and C. White, "*Evaluating ETL and Data Integration Platforms*," The data warehouse institure Journal(TDWI Research), vol. 1, pp. 1–38, 2013, [Online]. Available: http://download.101com.com/tdwi/research_report/2003ETLReport.pdf.

[7] P. Ponniah, "*Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals"*, vol. 6. 605 Third Avenue, New York: John Wiley & Sons, 2001.

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

[9] Gartner, "*Magic Quadrant for Data Integration Tools,Gartner Reprint*," Stamford, 2020. Accessed: May 01, 2020. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-1OA35PNQ.

[10] N. Biswas, S. Chattapadhyay, G. Mahapatra, S. Chatterjee, and K. C. Mondal, "*A new approach for conceptual extraction-transformation-loading process modeling*," International Journal of Ambient Computing and Intelligence, vol. 10, no. 1, pp. 30–45, 2019, doi: 10.4018/IJACI.2019010102.

[11] 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.

[12] R. P. Deb Nath, K. Hose, T. B. Pedersen, and O. Romero, "*SETL: A programmable semantic extract-transform-load framework for semantic data warehouses*," Information Systems, vol. 68, pp. 17–43, 2017, doi: 10.1016/j.is.2017.01.005.

[13] KLAS Research, "*2020 Best In KLAS Healthcare Business Intelligence & Analytics*," 2020. Accessed: May 01, 2020. [Online]. Available: https://klasresearch.com/best-in-klas-ranking/healthcare-business-intelligence-and-analytics/2020/97.

[14] MicroStrategy, "*Architecture for Enterprise Business Intelligence*," p. 443, 2012, [Online]. Available: https://www.microstrategy.com/Strategy/media/downloads/products/Analytics/MicroStrategy-Architecture-for-Enterprise-BI.pdf.

[15] P. Russom, "*Next Generation Data Integration*," *TDWI Research*, p. 35, 2011, doi: 10.1145/1216993.1216994.

[16] T. C. Ong *et al.*, "*Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading*," BMC Medical Informatics and Decision Making, vol. 17, no. 1, p. 134, Dec. 2017, doi: 10.1186/s12911-017-0532-3.

[17] HHS.gov, "*HIPAA for Professionals | HHS.gov*," www.hhs.org, 2017. https://www.hhs.gov/hipaa/for-professionals/index.html (accessed May 16, 2020).

[18] B. Bergeron, "*Developing a Data Warehouse for the Healthcare Enterprise: Lessons from the Trenches"*, 3rd ed. Healthcare Information and Management Systems Society (HIMSS)., 2018.

[19] HIMSS, "*Adoption Model for Analytics Maturity | HIMSS Analytics*," HIMSS Analytics, 2020. https://www.himssanalytics.org/amam (accessed May 15, 2020).

[20] HIMSS, "*HIMSS Analytics*," AMAM, 2020. https://www.himssanalytics.org/work-with-certified-organizations/amam (accessed May 15, 2020).

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