J. Adv. Sci. Edu. Res. 2023: 4: 164-173, *ISSN*: 2583-0155 (ONLINE)
https://doi.org/10.56253/JASER.3.1.2023.164-173
Published: 22.11.2024*(http://jaser.rkmvccrahara.org/)*

FROM INFORMATION TO INSIGHTS: EXPLORING THE COST BENEFIT ANALYSIS OF SELECTED DATA WAREHOUSING TOOLS

[1*]Subham Das& [2]Swapan Khan

[1]MLISc Student, Department of Library and Information Science, Jadavpur University, Kolkata, West Bengal,

[2]Librarian, N.D College, Howrah, West Bengal,

*For Correspondence: realmesdas@gmail.com

ABSTRACT

As big data continues to grow at an unprecedented rate, it has become increasingly important for organizations to effectively manage and analyse large volumes of data. This includes libraries, which are facing the challenge of managing and analysing a vast amount of information. Data warehouse tools have emerged as an important solution for libraries to store and analyse their data.

This article begins by discussing the importance of data warehousing in the library sector and how it has led to the need for effective data management and analysis tools. The article then focuses on the six most popular data warehouse tools used in libraries: Amazon Redshift, Google Big Query, Microsoft Azure Synapse Analytics, Snowflake, Alma and OCLC WorldshareAnalytics.

The article provides a detailed comparison of these data warehouse tools based on their architecture, usability, cost-benefit, use cases in libraries, recognized clients, security features, and special features. The article also explains why only these five data warehouse tools are selected for the study, based on their popularity, market share, and recognition in the industry.

The article concludes by summarizing the strengths of each data warehouse tool, along with recommendations for libraries based on their specific needs and requirements. The article emphasizes the importance of selecting the right data warehouse tool for libraries, as it can significantly impact their data management and analysis capabilities.

Keywords: Data Warehousing Tools, Cost-Benefit Analysis, Amazon Redshift, Google BigQuery, Alma.

1. Introduction

Data warehousing is a system designed for storing, managing, and analysing large volumes of data from multiple sources(Chandra & Gupta, 2018). It involves collecting data from disparate sources and transforming it into a centralized repository for analysis and reporting. The primary goal of data warehousing is to provide organizations with a comprehensive view of their data, which can

help them make informed decisions, improve business operations, and gain a competitive advantage (March & Hevner, 2007).

In today's data-driven world, data warehousing has become increasingly important for businesses of all sizes and industries. By providing a single source of truth for data, data warehousing enables businesses to streamline their decision-making processes, improve operational efficiency, and gain insights that can help them stay ahead of the competition (Redman, 2013). With the increasing amount of data generated by businesses, data warehousing has become a critical component in managing and analysing data effectively.

2.Literature Review:

According to Laitinen & Saarti, (2012), a data warehouse is used to store historical and integrated data for decision support systems. Libraries can use this tool to combine different types of statistical data to measure their efficiency and impact on parent organizations. Even the applications of the data warehouse tools is varied in libraries (Yang & Shieh, 2016). But the concept of data warehousing is not that easy it has its own challenges to comprehend as discussed in L C & R K, (2023). According to Sureddy & Yallamula (2020),there are certain techniques to choose correct tool for an enterprise which needs to be explored first.

3. Objectives of the Study:

- To explore what are the different Data Warehousing tools available in the market for organizations to choose from.
- To evaluate the selected sample size on various parameters so that it easy for the first time users(specifically libraries) to select which tool they might like to use.

4.Research Methodology:

For this study, a survey based methodologies has been used.In terms of selection of sample the sample size is fixed to be 6 tools for the population of the survey out of which two has been selected on the basis of their library specialization usage and rest 60% of the population for the survey are the most popular companies (in terms of brand value) in this domain providing general data warehousing services.

The list of data warehouse tools considered for this study are as follows:

- Amazon Redshift
- Google Big Query
- Microsoft Azure Synapse Analytics
- Snowflake
- Alma
- OCLC World share Analytics

Out of the above chosen data warehouse tools, Alma& OCLC Worldshare Analytics are specialized tool for library specific needs and rest are general data warehousing tools used by varied sectors.

In terms of comparison between these tools a total of 7 points of comparisons are selectedaftr through research so that the cost benefit analysis can smoothly take place covering every aspects of the tools considered. The points of comparison are as follows:

- Year of Establishment
- Architecture
- Special Features
- Type of Clients
- Recommended for
- Pricing or Cost of Acquisition
- Key Benefits

5. Overview of Different Data Warehousing Tools:

i. Amazon Redshift:

Amazon Web Services (AWS) provides a cloud-based data warehousing service called Amazon Redshift, designed to handle petabyte-scale data warehouses efficiently. It uses a columnar storage architecture that supports structured, semi-structured, and unstructured data sources, integrating with various AWS services such as S3, EMR, and Data Pipeline for easy data loading and analysis.

To ensure data security and availability, Redshift offers automatic backups, snapshots, and encryption, giving organizations peace of mind when storing sensitive information.(*Cloud Data Warehouse – Amazon Redshift – Amazon Web Services*, n.d.) Moreover, the platform provides flexible pricing options such as on-demand and reserved instances to help organizations optimize their costs, making it an affordable option for organizations of all sizes.

Amazon Redshift also supports various business intelligence tools and SQL interfaces, enabling data analysts to perform complex data analysis tasks and derive valuable insights from large volumes of data. The platform's fast query processing and scalability make it a powerful solution that can handle complex data analysis tasks with ease.

Overall, Amazon Redshift is a powerful and scalable solution that enables organizations to efficiently and cost-effectively analyze large volumes of data using SQL and business intelligence tools. Its advanced features, including automatic backups, snapshots, and encryption, ensure data security and availability. Moreover, its flexible pricing options make it an affordable option for organizations of all sizes seeking to store and analyze large datasets.

ii. Google BigQuery:

Google Cloud provides a cloud-based data warehousing and analytics platform, known as Google BigQuery, which is a fully-managed service capable of processing and analyzing massive datasets efficiently and cost-effectively. The platform is designed to support structured, semi-structured, and unstructured data sources and uses a columnar storage format that enables efficient compression and high-performance query processing, resulting in fast query response times. (*BigQuery Enterprise Data Warehouse*, n.d.)

BigQuery offers automatic scaling, real-time streaming ingestion, and integration with other Google Cloud services such as Cloud Storage and Cloud Dataflow, which enable organizations to process and analyze large-scale data with ease. The platform also provides advanced security features, such as multi-layered security controls, data encryption, and secure access controls, ensuring data protection and regulatory compliance. Additionally, BigQuery supports a wide range of programming languages and analytics tools, making it easy for data analysts and developers to utilize the platform.

Flexibility is a significant benefit of BigQuery's pricing model, which offers various pricing options based on usage patterns, enabling organizations to optimize their costs. In summary, Google BigQuery is a robust and scalable data warehousing and analytics solution, capable of processing and analyzing massive datasets efficiently and cost-effectively. Its advanced features, high-performance query processing, and flexible pricing make it an attractive option for organizations seeking to store, process, and analyze their data.

iii. Microsoft Azure Synapse Analytics:

Microsoft offers Azure Synapse Analytics, a cloud-based analytics service that can process and analyze large volumes of data in a scalable and cost-effective manner. The platform integrates data integration, big data analytics, and business intelligence services to support various data sources, including structured, semi-structured, and unstructured data.

Azure Synapse Analytics uses a columnar storage architecture that facilitates fast query processing and seamless integration with Azure services such as Azure Data Lake Storage and Azure Data Factory. The platform offers robust security features such as encryption and access controls, ensuring that data is protected and adheres to regulatory compliance standards(*Azure Synapse Analytics | Microsoft Azure*, n.d.).

Flexible pricing options for the platform, including on-demand and provisioned capacity, allow organizations to optimize their costs and adjust their analytics workloads accordingly. Moreover, Azure Synapse Analytics provides a collaborative workspace where data analysts and engineers can work together to share insights easily, making it an ideal option for complex analytics projects.

In conclusion, Microsoft Azure Synapse Analytics is a powerful and scalable cloud-based analytics service that can efficiently process and analyze large volumes of data. Its advanced features, including fast query processing, seamless integration with Azure services, robust security, and collaborative workspace, make it an attractive option for organizations looking to derive valuable insights and drive business growth.

iv. Snowflake:

It is a cloud-based data warehousing tool.Snowflake has a unique architecture that separates storage, compute, and services. This architecture allows organizations to scale up or down their computing resources on-demand to meet their changing business needs, providing unparalleled flexibility and cost-effectiveness. In comparison, Amazon Redshift and Google BigQuery use a shared infrastructure, which can limit scalability and flexibility.(*Snowflake Documentation*, n.d.)

Snowflake's seamless integration with various third-party tools, including data integration and business intelligence tools, makes it easy for organizations to incorporate the platform into their existing workflows. This integration is more straightforward and customizable than that of Microsoft Azure Synapse Analytics, which requires the use of specific Microsoft tools for data ingestion and processing.

Snowflake's advanced security features, such as encryption, multi-factor authentication, and data masking, ensure data privacy and regulatory compliance, which is essential for organizations handling sensitive data. While Azure Synapse Analytics and Amazon Redshift also offer advanced security features, Snowflake provides more granular control over data access and auditing capabilities.

Overall, Snowflake's unique architecture, seamless integration with third-party tools, and advanced security features set it apart from other cloud-based data warehousing platforms.

v. Alma:

It is a cloud-based integrated library system (ILS) that offers a comprehensive suite of library management tools. One of its key features is its data warehousing capabilities, which allow libraries to store, manage, and analyze large volumes of data efficiently.

Alma's data warehousing capabilities enable libraries to consolidate their data from multiple sources into a single, unified platform. The platform supports a range of data formats, including structured, semi-structured, and unstructured data, making it easier for libraries to manage and analyze their data.

Alma's data warehousing capabilities also offer advanced data analysis and reporting tools that enable libraries to extract valuable insights from their data. The platform includes a range of pre-built reports and dashboards, as well as customizable options that allow libraries to generate reports specific to their needs.(*Cloud Data Warehouse – Amazon Redshift – Amazon Web Services*, n.d.)

In addition to its data warehousing capabilities, Alma offers a range of other tools to manage various library functions, including acquisitions, cataloging, circulation, and resource sharing. The platform also integrates with a range of third-party tools, including discovery systems, electronic resource management systems, and learning management systems, making it easier for libraries to manage their resources effectively.

Overall, Alma's data warehousing capabilities make it an ideal choice for libraries that need to manage and analyze large volumes of data efficiently. It's advanced reporting and analysis tools provide valuable insights that can help libraries make data-driven decisions and improve their services. Additionally, its integration with a range of third-party tools makes it a comprehensive and versatile solution for library management.

vi. OCLC WorldShare Analytics:

It is a cloud-based analytics platform designed to assist libraries in effectively managing and analyzing their data. One of the platform's key features is its robust data warehousing capabilities, which allow libraries to store, manage, and analyze large volumes of data in a centralized platform. By consolidating data from multiple sources, including ILS, circulation data, and electronic resources, libraries can make informed decisions based on a comprehensive and unified view of their data.

WorldShare Analytics supports a range of data formats, including structured, semi-structured, and unstructured data, making it easier for libraries to manage and analyze their data. The platform includes advanced reporting and analysis tools that enable libraries to extract valuable insights from their data. With a wide range of pre-built reports and dashboards available, as well as customizable options, libraries can generate reports specific to their needs and obtain the necessary information quickly and easily.(*WorldShare*, 2021)

In addition to its data warehousing and analysis capabilities, WorldShare Analytics includes tools to help libraries identify trends and patterns in their data, enabling them to make data-driven decisions to improve their services. The platform's data governance tools also allow libraries to ensure that their data is accurate, secure, and compliant with industry standards and regulations. By providing a secure and compliant data management solution, WorldShare Analytics empowers libraries to use their data to make informed decisions and improve the quality of their services.

In conclusion, WorldShare Analytics is an effective cloud-based analytics platform that provides libraries with data warehousing capabilities to store, manage, and analyze large volumes of data in a centralized platform. It's advanced reporting and analysis tools allow libraries to extract valuable insights from their data, and its customizable options ensure that libraries can generate reports specific to their needs. The platform's tools for identifying trends and patterns in data, as well as its

focus on data governance, enable libraries to make data-driven decisions and ensure that their data is accurate, secure, and compliant with industry standards and regulations.

Comparison of SelectedData warehousing Tools:

Point of Comparisons	Amazon Redshift	Google BigQuery	Microsoft Azure Synapse Analytics	Snowflake	Alma	OCLC WorldShare Analytics
Year of Establishment	2012	2010	2019	2012	2011	2013
Architecture	Columnar	Columnar	Massively Parallel Processing	Cloud-based, Columnar Storage	Cloud-based	Cloud-based, Data Warehousing
Special Features	Automatic backups, snapshots, encryption, flexible pricing	Serverless, scalable, integrates with other Google Cloud services	Integration with other Azure services, Built-in security, server less	Zero-copy cloning, data sharing, cross-cloud capabilities	Automation, analytics, open platform	Consolidation of library data, advanced analysis and reporting tools
Type of Clients	Libraries, Retail, Healthcare, Finance	Libraries, E-commerce, Healthcare, Finance	Libraries, Retail, Healthcare, Finance	Libraries, E-commerce, Healthcare, Finance	Libraries, Higher Education, Research	Libraries, Higher Education, Research
Recommended for	Large libraries with high data volumes and complex analytical needs	Libraries with significant Google Cloud Platform adoption or integration needs	Libraries with significant Azure adoption or integration needs	Libraries with complex data needs, including semi-structured data	Libraries in search for Integrated library system, resource management	Libraries using OCLC products or seeking a library-specific data warehousing solution

Pricing or Cost of Acquisition	Starts at $0.25 per hour	Pay-as-you-go pricing based on usage, starts at $0.01 per query	Pay-as-you-go pricing based on usage, starts at $1.10/hour	Pay-as-you-go pricing based on usage, starts at $40/TB/month	Pricing available on request.	Custom pricing based on library size and usage
Key Benefits	High scalability, cost-effective, easy integration with AWS services	Scalable, fast performance, cost-effective, serverless	Integration with Microsoft Azure services, security and compliance features	High scalability, cost-effective, fast performance	Integration with Ex Libris services, customizable workflows	Integration with library data sources, advanced reporting and analysis tools, data governance features

6. Cost-Benefit Analysis:

After going through the chart it is clear that each and every tool considered in the population have their own range of features which will benefit the customers of that specific needs at different price range and structure. But on an overall level the following are the one line analysis of each tool:

1. Amazon Redshift can be used in organizations where you want to try data warehousing tools for the first time but your staff works on a fixed shift schedule, such as in India, where it is typically 10:00-17:00 hrs. In this scenario, the cost structure will be advantageous to the organization because it is an hourly plan.
2. For organizations that already use Google services and have workers available 24/7 in various shifts, Google BigQuery will be beneficial. Since the pricing is based on the number of queries, many people can use it at any given time without worrying about the cost.
3. Although Microsoft Azure Synapse Analytics is more expensive than other options, it has the significant advantage of working faster and handling more queries than other options thanks to its MPP architecture. Therefore, it is not advised for beginners or new users who have only recently started using these tools. It is an advanced tool for institutions that must process large numbers of queries quickly.

4. Snowflake offers a monthly pricing plan, which is best for establishments with staff members who have fewer holidays and more workdays each week (such as private institutions).It is appropriate for institutions that have a long-term perspective on data warehousing analysis. Its ability to process semi-structured data will be a huge edge over competing products on the market.
5. Alma & OCLC are specialized tools for library data analytics, or data warehousing tools created to analyze data produced in libraries, but there is no clear indication of the cost of these tools; organizations must ask these companies for a quote after they assess the scope of the organization's requirements.

7. Conclusion:

After analysing various data warehousing tools available for libraries, it is evident that each tool has its own set of advantages and disadvantages. The decision to choose one tool over the other will ultimately depend on the specific needs and budget of the library. Data warehousing in libraries can provide valuable insights into user behaviour, resource usage, and collection development, which can be used to improve library services and outreach. These tools help to draw trends that we will never know through conventional processes. With various examples of other client libraries, it can be concluded that investing in data warehousing tools is a much more favorable option for a better and faster tomorrow for libraries. Although these technologies may take some time to reach every library, it is important to know about these tools to serve better.

8. References:

1. *Azure Synapse Analytics | Microsoft Azure*. (n.d.). Retrieved April 30, 2023, from https://azure.microsoft.com/en-in/products/synapse-analytics

2. *BigQuery Enterprise Data Warehouse*. (n.d.). Google Cloud. Retrieved April 30, 2023, from https://cloud.google.com/bigquery

3.Chandra, P., & Gupta, M. K. (2018). Comprehensive survey on data warehousing research. *International Journal of Information Technology*, *10*(2), 217–224. https://doi.org/10.1007/s41870-017-0067-y

4.*Cloud Data Warehouse – Amazon Redshift – Amazon Web Services*. (n.d.). Amazon Web Services, Inc. Retrieved April 30, 2023, from https://aws.amazon.com/redshift/

5.L C, M., & R K, S. (2023). A Review on Data Warehousing Concepts, Challenges and Applications. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 25–31. https://doi.org/10.32628/CSEIT239015

6. Laitinen, M., & Saarti, J. (2012). A model for a library-management toolbox: Data warehousing as a tool for filtering and analyzing statistical information from multiple sources. *Library Management*, *33*, 253–260. https://doi.org/10.1108/01435121211242290
7. March, S. T., & Hevner, A. R. (2007). Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, *43*(3), 1031–1043. https://doi.org/10.1016/j.dss.2005.05.029
8. *Snowflake Documentation*. (n.d.). Retrieved April 30, 2023, from https://docs.snowflake.com/en/
9. Sureddy, M. R., & Yallamula, P. (2020). *Approach to help choose right data warehousing tool for an enterprise*. *6*, 579–583.
10. *WorldShare: Enable shared efficiencies and innovation*. (2021, November 3). OCLC. https://www.oclc.org/en/worldshare.html
11. Yang, Y.-T., & Shieh, J.-C. (2016). *Data Warehouse Applications in Libraries—The Development of Library Management Reports*. 88–91. https://doi.org/10.1109/IIAI-AAI.2016.129.

Declaration:

"This is to declare that this article is developed by us, based on our personal study and research or, on primary / secondary published data and that We have duly acknowledged the use of all such previously published data in preparation of this report in a conventional manner and it is free from plagiarism. Further, We declare that this report has not previously been submitted for assessment / publication in any other journal." – Subham Das

Authors Contribution: Dr. Swapan Khan and Subham Das both worked simultaneously in this article where the former one give the idea about the topic the later one executed the whole study.

Consent of authors:

Consent of each author towards publication of the article is to be submitted separately along with the manuscript in the following format –

Declaration of Consent

"This is to declare that I have full consent in publishing the article entitled 'FROM INFORMATION TO INSIGHTS: EXPLORING THE COST BENEFIT ANALYSIS OF SELECTED DATA WAREHOUSING TOOLS' along with Dr. Swapan Khan as co-author. I do not have any conflict of interest in publication of the said article." – Subham Das.

"This is to declare that I have full consent in publishing the article entitled 'FROM INFORMATION TO INSIGHTS: EXPLORING THE COST BENEFIT ANALYSIS OF SELECTED DATA WAREHOUSING TOOLS' along with Subham Das as co-author. I do not have any conflict of interest in publication of the said article." –Dr. Swapan Khan.