Lightweight DWH Data Analysis for SMEs

Lukas Dötlinger, Manuel Penz, Markus Reiter & Stephanie Widauer

University of Innsbruck, Austria

Abstract

With an increasing level of digitisation, the amount of data, that companies are dealing with is increasing rapidly across all business sectors. Nowadays, especially small and medium enterprises are struggling to properly process all of their data as they do not have a proper infrastructure to manage it. Therefore, the usage of data warehouse (DWH) systems is becoming more popular in the SME sector, since it provides an easy way to collect the entire business data in one place, while gaining better possibilities for analytical insights. This lead to a high development of dedicated DWH services, which specifically target SMEs. Those are typically designed as software as a service (SaaS) with a flexible pay-as-you-go model. This paper aims to highlight the specific needs for such specialised systems and compares different cloud-based DWH services for their suitability within small and medium enterprises while using a predefined review methodology. The conclusion reflects on the current state of DWH services and their suitability for smaller and medium companies.

1 Introduction

A data warehouse (DWH) is a special type of database system that focuses on reporting and analysis of its data. Implementing such a system reduces the complexity to access business data in an analytical way and is an important step to achieve business intelligence (BI). Enterprises typically bundle the data of all their operational databases together in one data warehouse. All departments within a company still use their own database for day to day production, as they don't use it for analysing and reporting of their business process. Hence, the DWH system is mostly used in the management layer of an enterprise which deals with internal reporting and business analytics.

With increasing digitisation of business processes and communication, the amount of data that companies are collecting is increasing rapidly across all business sectors. Therefore, a data warehouse system is becoming more interesting for many small- and medium-sized enterprises (SMEs) as they require a systematic approach to analyse their business data in a productive way. In the context of this research, companies with at most 250 employees are considered an SME, also known as small and medium business (SMB).

The increasing demand for such systems has lead to an increase in development for specific data warehouse solutions targeting SMEs. This paper aims to compare and analyse such systems for their suitability in the context of a small or medium enterprise. Furthermore, a comparison of different systems is used to give a general baseline for a lightweight implementation of a DWH in a SME. The work also tries to highlight the essential features a data warehouse is required to have if being applied in an SME. Additionally, the review process of existing services is thoroughly described, to achieve a

high reproducibility of the results.

Section 2 discusses some related work and gives an overview of previous approaches. Afterwards, section 3 presents some factors of success for a data warehouse implementation at an SME and highlights the specific needs for such a use case. The methodology for the review of DWH services is described in section 5. Finally, a conclusion and outlook for some future work is presented in section 7.

2 Related Work

As the amount of digital data is ever increasing for all companies, the effort to manage it grows exponentially. To structure that data, most business have the option of implementing a data warehouse. Although there are many existing DWH implementations, only some are actually applicable to the setting of a smaller business. This is due to the fact that SMEs tend to lack certain expertise and have a limited budget as well as a small amount of spare employees for IT. This further aggravates the use of many traditional DWH solutions as they are tailored towards big enterprises and include many features which are not relevant for an SME. [1]

Furthermore, on-premise data warehouse systems require a certain level of storage and computing power to fully utilise the advantages of the software. Those solutions often have a high up-front cost, making them a less ideal solution for an SME. Therefore, many vendors of DWH solutions offer their product as a cloud-based subscription service, which is well received by customers. [2]

Cloud-based data warehouses enable smaller business to fully utilise all needed features of the DWH technology, as they require considerably less time and expertise to set up and configure. Additionally, there is no up-front cost as there is no local infrastructure required. Many popular vendors also offer a pay-as-you-go subscription, which gives customers flexibility to try out their system for a very low fee. The reduction in cost and a static monthly payment model is the main reason why smaller companies can even consider implementing a DWH in their business process, as this was previously the main barrier. [3]

Business intelligence (BI) is a systematic approach that compromises a set of tools and guidelines which helps a business analyse its operations and report on different statistics. Systems that offer BI features use a data warehouse for analytical queries. While this is a common approach within global enterprises for years, SMEs rarely adopt any form of business intelligence due to lack of expertise and capabilities to deploy such systems for it. [4]

SMEs therefore require a simple and streamlined approach as they are not able to analytically deal with big data in their existing systems. With the addition of cloud computing, SMEs no longer need complex hardware and software systems, which have high maintenance costs, to deal with big data. Vajjhala and Ramollari argue that cloud services also encapsulate some of the complexity, giving SMEs the chance the acquire flexible computing power for data warehouses to use business intelligence tools. [5]

3 DWH Needs for SMEs

The data generated and captured by an SME is its most important asset. Since the amount of available data is constantly growing, the only solution to avoid data management problems is to use a dedicated data warehouse (DWH) system. Not every SME may need such a storage solution right from the start, but there are some common signs which show a business when it would be more efficient to switch over or start with a data warehouse system.

Heavy reliance on spreadsheets is for example one critical sign why SMEs should use a DWH System. Spreadsheets are very commonly used in pretty much every business and its different departments to track data. While in most cases they seem to be pretty universal, a lot of these spreadsheets can grow to immense sizes and can become unmanageable. Combining these large files manually to create a report takes a lot of time, not to mention the fact that every department may also rely on different spreadsheets.

Spreadsheets are designed to take a specific amount of data divided into rows and columns. Continuous adding of data can lead to "spreadsheet overwhelming". The file itself can become either sluggish or prevent the user from adding rows and/or columns altogether. Therefore, a data warehouse system can definitely increase the productivity, especially when combining multiple spreadsheets.

If employees in different departments work on these spreadsheets and one person needs to wait on specific information to create a report or analyse data, this takes too much time just to wait on other employees. With a DWH on the other hand, data gets added directly into one centralised data location and analysis can be performed at any minute.

Discrepancies in data and reports can be the result of different departments creating their own data and reports. The difference in the results can be time consuming to sort out and for SMEs this can lead to costly mistakes. In most cases, the reason is caused by adding different, sometimes untrustworthy data sources. If the point of data discrepancy is reached, it may be time to remedy this problem by looking into a data warehouse system which ensures mistakes like duplicate data are eliminated.

If the time invested in creating reports is too much, then SMEs should decide on using a DWH System. Ideally, such reports can be created with a few clicks and prevent employees from going to different sources to check if the data is already updated. Since data warehouses consolidate data, all departments can rely on a single source for data. Maintenance can be further simplified by using the ability of such systems to set up automatic updates if the source data gets changed or updated in order to guarantee the data which departments rely on is always correct.

4 Methodology

The conclusion about the current state of data warehouse solutions for SMEs is drawn from the several service evaluations. To achieve a high degree of reproducibility, this evaluation and testing has been done by a fixed methodology that thoroughly describes each step. As the review of those services has been conducted by multiple researchers, a predefined evaluation approach was necessary to collect the same type of data and test for a fixed set of use-cases.

Upon selection of a service to review, an account is created to check if a free trial version is even available. If this was not the case, the service was excluded from the review, since just evaluating the specifications and documentation was not seen as sufficient to reason about the suitability for an SME.

The initial setup process of an account and the data warehouse is also discussed within the review, since the whole process should not be too complex for IT staff that are not familiar with the underlying technology of a data warehouse. This is due to the fact that such highly specialised staff might not exist in an SME that needs a simple data warehouse service solution.

Afterwards, the review focused on the features of the service. Therefore, the following questions were asked for each:

- How many different data sources are supported?
- Does the service include analytical features?
  - If not, how many data destinations are supported?
  - If it does, how many analytical visualisation tools are supported?
- What are the different pricing tiers and which one is recommended for SMEs?

In addition to those questions, any other information like underlying services or legal certifications were noted within a services review. Furthermore, the overall design of the services interface is rated for usability.

While those aspects are mostly very objective, some subjective bias by the researcher cannot be excluded, especially for the initial setup of a ser-

vice. The complexity of the service and its interface itself is in generally a completely subjective perception and therefore might have been different for other researchers. However, this is not seen as a major problem, since the core suitability of a service for an SME is determined by its feature set and price.

5 DWH Services for SMEs

For small and medium businesses, probably the most important aspect when choosing a data warehouse system is cost, both for the initial development and for the ongoing maintenance of such a system.

Nowadays, software as a service (SaaS) can provide many advantages over traditional services. The pay-as-you-go model is very friendly towards small businesses which could not otherwise easily justify the upfront cost for servers and related costs for hosting a data warehouse.

This means that, in many cases, SaaS is the most cost-effective and also the simplest solution for small businesses to opt for. When comparing them to traditional services, SaaS products virtually don't need any setup time and can be deployed instantly.

In the case of data warehouse systems, SaaS is also commonly referred to as data warehouse as a service (DWaaS).

Given the advantages above, we focus in this section on some concrete DWaaS products and review what they have in common, how they differ and whether they are in fact suitable for small businesses.

5.1 Segment (by Twilio)

Segment is a customer data platform for collecting, cleaning and controlling data across multiple services in one central location. It offers a *Free* plan which supports two data sources, 1,000 API calls and more than 300 integrations, a *Team* plan starting at $120/month for 10 users with unlimited data sources and 10,000 API calls up to $1,125/month for 100,000 API calls, and a *Business* plan for custom usage requirements. The scalable *Team* plan should be sufficient for most SMEs.

After first logging into Segment, the user can choose the team they are working on (Engineering, Marketing, Founder/Executive, Product, Analytics) and select the first data source, e.g. a website, programming language or HTTP API. Next, data destinations have to be selected, e.g. Google Analytics, Intercom, etc. Finally, the user gets to the dashboard, which provides an overview of all data sources and destinations and a way to add new ones.

In total, Segment supports 98 different data sources and 650 data destinations at the time of writing. Additionally, creating custom data sources and destinations by building JavaScript function that access the corresponding API. Also, by supporting programming languages as data source and webhooks as data destinations, virtually any software can be integrated.

Segment does not offer any analytics capability on its own but is meant to simplify data collection and distribution by managing all data sources and destinations in a single place, therefore reducing complexity and increasing flexibility. For example, website analytics can be switched from Google Analytics [6] to GoSquared [7] without changing the website itself.

5.2 Panoply

Panoply is a data warehouse solution built on AWS Redshift and offers four plans: *LITE*, *STARTER*, *PRO* and *BUSINESS*. The last of them is specifically aimed at SMBs, according to their own website. There is a free version available for testing which has the functionality of the *LITE* plan for a period of 14 days. After logging in for the first time, the website prompts the user to create a data warehouse with a unique name. Next, the user is prompted with the possibility of adding a data source, which can also be skipped. Afterwards, the user has full access to the instance.

Panoply offers 122 data sources that have been integrated by the Panoply team. Additionally, there are 131 data sources which are developed by partners that can be added. In total, 253 different data sources are supported. To analyse the collected data, Panoply offers the integration of 43 visualisation tools, of which 42 are BI tools.

Panoply is a full solution to synchronise, store and access a company's data while also providing analytical features. In addition to the supported visualisation tools, data can be structured and viewed in a traditional tabular form.

The different pricing tiers are differentiated by three main parts: amount of data sources, storage space and support. The suggested SMB solution, called the *BUSINESS* plan, includes 10 distinct data sources, 100 GB of storage and support with a reaction time of less than an hour. It also includes data governance features, yet the storage itself is based in the USA, without any other option. All plans offer an unlimited amount of users. More storage and data sources is possible for the *Enterprise* plan, which is adaptive to a

company's needs. This adds the possibility of storing the data in one of 19 different countries.

5.3 Tableau

Tableau is a visualisation software which mainly focuses on data visualisation and data reporting. This tool is founded by Tableau Software and is now owned by the cloud computing solutions company Salesforce.com. Tableau uses a machine-learning based analysis engine which helps model automated structured data tables and displays statistical findings.

Analytics or Business Intelligence is a cycle with different steps: Transactions need to be stored securely before data analysts periodically analyse the data. Afterwards, these insights are shared across the company where then senior colleagues make decisions before the outcomes get monitored from managers who change product offers. Tableau has six key products which aim at improving the workflow for every of the six previously mentioned steps by offering a simple to use, connected platform.

For storing data, Tableau does not use a dedicated database but a special file type called `.tde` or more recent `.hyper`. Since data is not stored correctly all time due to a missing category or a wrong scanned item, these mistakes need to be fixed manually every time or in bulk periodically. The Tableau Prep Builder tool allows users to clean, shape and prepare data by making it ease to delete, move or even merge fields from different data sources. After cleaning data, Tableau Desktop and Tableau Public connect to the clean data and are able to analyse it. Tableau Desktop allows users to connect to basically any data source they want, like Excel spreadsheets, billion row

databases and even web APIs. In total, both Tableau Desktop and Tableau Public support more than 80 different data sources including hundreds of web data connectors used to grab data from e.g. HTML, JSON & XML. Analysis can be done by drag-and-drop and analysts can also ask questions inside the tool and get the corresponding answer displayed as data. This really simplifies business intelligence without the need of data warehouse expertise. Tableau Public on the other hand can do the same as the desktop version but users can only share dashboards and insights with others also using the public version - basically a community, non profit, free edition of Tableau Desktop. After preparing and connecting to data in order to build reports, analysis cannot be done by a single person, so in order to share these reports across the business, a safe and secure tool is needed. Also, collaboration between analysts and being able to withstand numerous requests should be key factors of such a tool. Tableau Server and Tableau Online offer all these aspects. The former is deployed on-premise while the latter is hosted on Tableau Cloud. One downside of Tableau Cloud which has to be considered is the storage limitation of 100GB per site. Employees on the go can access data within the Tableau Mobile apps. Tableau offers data visualisation out-of-the-box by dragging and dropping columns and rows into desired fields and then choosing the desired visualisation type, e.g. histogram, box-and-whisker and any further combination between 30 different types are possible.

Except for Tableau Public, which is the only free version built to share data in the Tableau ecosystem, every other tool needs to be purchased. In this case, Tableau offers a package for individuals containing Desktop, Prep Builder and Server or Online, which costs $70/month per user. In case of

teams and organisations, packages are available from *Viewer* to *Explorer* and *Creator* ranging from $12-$70/month per user, where the first one contains only basic functionality and the last one includes the full package.

Taking all these different packages into account, the best solution for SMEs would be the Tableau Creator package with Tableau Online in order to avoid having to manually set up and maintain a Tableau Server instance.

5.4 Snowflake

Snowflake is a data warehouse solution built on top of Amazon Web Services, Microsoft Azure or Google Cloud Platform. It aims to fulfil the majority of data analytics needs, such as data storage, data processing, data integration and it provides analytics solutions. Snowflake offers multiple editions of cloud data platform service: *Standard Edition*, *Enterprise Edition*, *Business Critical Edition* and *Virtual Private Snowflake (VPS)*. The *Standard Edition* offers unlimited access to all standard features in the platform. Additionally the *Enterprise Edition* contains features designed especially for the needs of large-scale organisations and enterprises. Furthermore, the *Business Critical Edition* and the *Virtual Private Snowflake* are for organisations who are dealing with extremely sensitive data. The editions contain strict requirements and provide higher levels of data protection. Snowflake also offers a 30-day trial with $400 worth of free usage. Before the login the user has to choose which cloud platform he wants to use.

The Snowflake platform uses a unique architecture consisting of three layers: database storage, query processing and cloud services. In the database storage layer, data is loaded into the platform. After the loading process, the

data will be optimised, compressed and stored in cloud storage. Snowflake manages all aspects of how the data is stored: the organisation, structure, metadata, statistics and many more. The data objects are then accessible through SQL query operations in the platform. In the query processing or virtual warehouse layer, the query execution is performed. This is done by using "virtual warehouses". Each virtual warehouse is an independent compute cluster which means that each warehouse has no impact on the performance of other virtual warehouses. Warehouses come in different sizes ranging from XS (extra small) to XXXL, depending on the organisations needs. Each size comes with different credits which are important for pricing. Additionally the warehouses deliver efficient BI solutions with an array of BI products, in total there are 23 different BI tools. The cloud service layer is responsible for the coordination of activities across the platform, such as authentication, infrastructure management, query parsing and optimisation and metadata management.

The pricing of theses layer depends on their actual usage. Snowflake offers two different pricing options. The first one is "On Demand": Customers are charged a fixed rate for the services that are consumed and are billed every month. The second one is "Pre-Purchased Capacity". A company can pre-purchase capacity which is then consumed on a monthly basis. Furthermore, Snowflake offers a pricing overview for each cloud platform and region on their website. As an example, a standard-level data warehouse running on Amazon Web Services in the European region will cost $2.70 per credit. In addition, 1TB of on-demand storage costs $45 per month.

In summary, the Snowflake data warehouse platform offers all the tools

necessary to store, retrieve, analyse and process data. The platform provides good solutions for small businesses as well as for big organisations.

6 Comparison

In this section we provide a brief comparison about the pricing models and features provided by the four different DWaaS solutions discussed in section 5.

One big advantage of DWaaS products is that service providers configure and manage hardware and software resources, and therefore the customer only provides the data and pays for the used services. The four different DWaaS products reviewed in this paper offer four different pricing models. However all of them offer a free trial in order to test their platform. Table 1 provides an overview of the different pricing options for each DWaaS product.

Panoply offers monthly and annual plans starting at $200 per month. Their pricing model is simple: users pay for the amount of data sources and storage space they use. There are no extra costs for adding users or the number of queries to run, therefore costs are more predictable than with "on-demand" models like Snowflake. The pricing is based on data consumption per second, which means users only pay for the compute power and storage they use. This structure works well for experienced users who know their average data consumption but this could be disadvantageous for new users with little experience. Furthermore, Segment's pricing tiers are based on the number of data sources and customers that are being tracked. Tableau offers different packages, which are only billed annually.

In comparison to the other services, Segment does not offer any data

	Pricing
Panoply	- monthly and annual plans starting at $200 - users pay for the number of data sources and amount of data
Segment	- pricing tiers based on number of data sources and customer or visitors - users pay for the number of data sources and amount of data
Tableau	- different packages starting from $12 per user per month - billed annually
Snowflake	- pricing based on data consumption per second - works well for experienced users

Table 1: Overview pricing models for Panoply, Segment, Tableau and Snowflake

analysis tools. It is designed as a centralised collection point that distributes the company's data to different destinations. Therefore, the service is limited by API calls.

The other services also include data analysis tools and therefore function as a central collection hub for a company's data. Different plans are limited by data storage and the amount of simultaneous data sources, that can be used.

Overall, not each service is ideally suited for all types of SMEs, yet the different pricing tiers and their features cover a wide variety of them. From the review we can conclude that all data warehouse services are considered suitable for an SME.

7 Conclusion

Concluding from this work, we can say that small and medium enterprises are generally in the need for a data warehouse system, yet they need to rely on dedicated services offering a specialised solution since they cannot afford an on-premise DWH. Cloud-based SaaS offerings provide a viable and affordable opportunity to use such a data warehouse in a pay-as-you-go fashion.

The reviewed cloud services show that there are multiple different systems which are ideally suited for SMEs. All services have a variety of plans designed for different sizes of smaller businesses. Furthermore, the integration with possible data sources is well supported across all of the reviewed systems. The majority of the reviewed services also support data analysis with the integration of different business intelligence tools. Additionally, custom queries written in SQL are also possible for all services, giving SMEs maximum flexibility when accessing their data.

Overall, we can conclude that cloud based data warehouse services are a very good and efficient solution for small and medium enterprises to improve data management and analysis. The services fully tackle the complicated needs of SMEs, while offering an attractive price structure. Therefore we can assume that there exists an almost ideal service for most SMEs.

References

[1] Raghavendra Raj, Shun Ha Sylvia Wong, and A. Beaumont. "Business Intelligence Solution for an SME: A Case Study". In: *KMIS*. 2016.

[2] Alessandro Agostino, Klaus Søilen, and Bart Gerritsen. "Cloud solution in Business Intelligence for SMEs –vendor and customer perspectives". In: *Journal of Intelligence Studies in Business* 3 (Dec. 2013), pp. 5–28. DOI: 10.37380/jisib.v3i3.72.

[3] Sérgio Fernandes and Jorge Bernardino. "Cloud Data Warehousing for SMEs". In: Jan. 2016, pp. 276–282. DOI: 10.5220/0005996502760282.

[4] Matteo Golfarelli, Stefano Rizzi, and Iuris Cella. "Beyond Data Warehousing: What's next in Business Intelligence?" In: *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP*. DOLAP '04. Washington, DC, USA: Association for Computing Machinery, 2004, pp. 1–6. ISBN: 1581139772. DOI: 10.1145/1031763.1031765. URL: https://doi.org/10.1145/1031763.1031765.

[5] Narasimha Vajjhala and Ervin Ramollari. "Big Data using Cloud Computing - Opportunities for Small and Medium-sized Enterprises". In: *European Journal of Economics and Business Studies* 4 (June 2016), pp. 129–137. DOI: 10.26417/ejes.v4i1.p129-137.

[6] *Google Analytics*. URL: https://analytics.google.com/ (visited on 01/07/2021).

[7] *GoSquared*. URL: https://www.gosquared.com (visited on 01/07/2021).