Lightweight DWH Data Analysis for SMEs

Lukas Dötlinger, Manuel Penz, Markus Reiter & Stephanie Widauer

University of Innsbruck, 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 a 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 data warehouse services for their suitability within small and medium enterprises, while using a predefined review methodology. The comparison is followed up by some performance examples for the tested systems. The conclusion reflects on the current state of DWH services and their general applicability for smaller companies and summarises the best services within the set of reviewed ones.

Contents

1 Introduction

A *data warehouse* (*DWH*) is a special type of database system, that focuses on reporting and analysing of it's 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, since they deal 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 a 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 at 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 about 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 6.

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 a 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 a 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 is a kind of privilege which has been used from larger companies longer than a decade ago but during the last 10 years more and more SMEs choose to use this technology since it has been developed rapidly. These data analysis tools are now more lightweight and accessible for smaller businesses and therefore used to turn data into informed decisions in order to face main competitors. Small Business Analytics is a technique and practice which measures a specific performance of a small company on an operational or strategic level. This technique is used on small datasets to gain insights on company processes. These insights can serve as key factors determining crucial decision-making processes. In most cases when organisations approach small data, they often overlook these insights. There are different reasons why small data should be treated seriously:

- Focus on target - While big data sees the performance, small data is more focused on improving results. Key performance indicators (KPI) need to be identified and people should get one indicator assigned to track the development.

- Actionable - Big data serves information on every metric of each department. The problem is that all this data can get too general and overwhelming which forces data analysts to make strategic and organisational changes.

- All about what is happening now - Getting information from small data is quite easy and the data-source acts immediately compared to big data. If one needs a historical insight or wants to combine old data with current data, it is not possible without big data.

- Delivered ready to be served - Small data serves information in a strategic way which makes it easier to manage it and coworkers are more likely to utilise reports that will deliver clearer insights on the data.

Even though small data is a part of big data, both can be used separately, dependent on the quantity of departments in the business. Every business, even SMEs need a clear overview of where they stand on the market to achieve business goals. [4]

PRESTA is a system used in a methodology that helps SMEs to measure the performance management. This system can generate specific data warehouses and it is responsible for data input and calculating norm values. Different organisations are seeking new systems to track the progress against plans and for flexible allocations of resources. Traditional systems include ERP and business intelligence systems but it is not sufficient. Business Performance Management (BPM) is a term to integrate key indicator of performance into management processes. The best known model for BPM is the Balanced Scorecard of Kaplan and Norton. Data around the company gets integrated into a DWH system in a consistent and reliable way. There are different tools needed in such systems like the front end, or the standard C/S tools. C/S tools are responsible for the development of customised data applications. Further EIS/DSS tools are needed to generate high-end multidimensional analysis which are more complex. Finally data mining tools are needed. Since different data mining products are exploratory to identify information without an initial question or hypothesis, specific data mining tools are needed to avoid this. In order to get a multidimensional view on the data it has to be organised in fact and dimension tables. To become a

generic performance management tool for a range of business situations a generic data model needs to be defined by having a list of critical success factors and performance indicators. These indicators (PI) variables are organised in facts and dimension tables. In the management analysis the PI measurement model has to be implemented in a way such that it delivers report of PI measures and also selected norm values. These reports are very helpful in determining business problems. The PRESTA system is planned to be a generic DWH based performance management system. Based on the input of the PIs it generates a DWH dimensional data structure and data gets input automatically. Considering that the operational database structure is far from complete, tables can also be input manually. Since manual input can lead to double data, data gets input on PI measurement variables level which means that measurements are kept at company level and loses detailed data. PRESTA is based on a web application and uses a MySQL database. In order for it to be become a full DWH based system, in the near future it will get a PRESTA based data warehouse where the system itself gets data feed directly from there. [5]

3 DWH Needs for SMEs

The data generated and captured by an *SME* is the most important asset available for the company itself. Since the amount of available data is constantly growing the only solution to not get into data management problems is to use a specific *Data Warehouse System (DWH).* It might not seem that every *SME* needs such a storage solution from the very first minute but there are different signs which show a business why 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. The spreadsheet itself is a very common used file type in pretty much every business and its different departments to track data. While in most cases it seems to be pretty universal, a lot of these spreadsheets can grow to immense size and can become unmanageable. Combining the fact of growing sheets across all departments, combining these files to create a manual report takes a lot of time, not to mention the fact that every department can also rely on different sheets.

Spreadsheets are designed to take a specific amount of data divided into rows and columns. Repetitive data adding can lead to *spreadsheet overwhelming*. The file itself can handle either sluggish or just prevent the user from adding rows and/or columns. Therefore a data warehouse system can definitely increase the productivity, especially if multiple different sheets get combined.

If employed in different departments work on these sheets and one person needs to wait on specific information to create a report or analyse data, it *takes too much time just to wait* on other employees. If the data person

needs to create a report gets added directly into one business centralised data source, analysing can be done at any minute. Furthermore other members in the same department also don't have to wait for data, due to an employee being too busy at the moment.

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 not trustworthy data sources. If the point of data discrepancy is reached it may be time for businesses to sort out this problem by looking into a data warehouse system which ensures eliminating mistakes like duplicate data.

If the *time invested in creating reports* is too much, then *SMEs* should decide using a DWH System. Ideally such reports can be created with 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 have to just turn to one source for data. Maintenance can be further simplified by using the ability of such systems to set up to automatically update if the source data gets changed or updated and it is guaranteed that 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 the 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 enough 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 existing in an SME needing 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?
- Dose 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 service. 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 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 *Busi-*

ness 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 build 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 if the *LITE* plan for a period of 14 days. After logging in for the first time, the website prompts a user to

create a data warehouse, that is required to have a unique name. Next, a 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 the integration of 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 source 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 sync, store and access a companies data, while also providing analytic features. Additionally 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 adapted to a companies 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 found by *Tableau Software Inc.* and is now owned by the Cloud Computing Solutions company *Salesforce.com*. *Tableau*

uses a machine learning based analysis engine which helps modelling 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 work-flow for every of the six previously mentioned cycle step by offering a simple to use, connected platform.

For storing data tableau does not use a dedicated database but uses 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 tool *Tableau Prep Builder* allows users to clean, shape and prepare data by making it ease deleting, moving or even merging fields from different data sources. After cleaning data, *Tableau Desktop* & *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-Spreadsheet, billion row databases and even to a web-API. In total both *Desktop* & *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 simply drag and drop and analysts can also ask questions inside the tool and get the corresponding answer displayed as data. *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 can't be done by a single person. To share these across the business, it needs a tool which is safe and secure, to prevent access from others. Also collaboration between analysts and being able to withstand numerous requests should be key factors of such a tool. *Tableau Server & Tableau Online* offer all these aspects and on top of it, in case of the first one, runs locally inside the business and the second one runs online, powered by the *Tableau Cloud*. One major downside of running the server in *Tableau's* cloud is the storage limitation to 100GB per site. Employees on the go can access data within the *Tableau Mobile* apps. Compared to other tools *Tableau* offers data visualisation out-of-the-box by dragging and dropping columns and rows into desired fields before choosing the desired visualisation type like histogram, box-and-whisker and any further combination between these 30 types are possible.

Except for the *Tableau Public* which is the only free version build to share data inside the Tableau-community, every other tool needs to be purchased. In this case *Tableau* offers a package for individuals containing *Desktop, Prep Builder & Server* or *Online* which costs $70 per user per month. In case of teams and organisations packages are available from *Viewer* to *Explorer* and *Creator* ranging from $12 up to $70 per user per month, where the first one is 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 and let the *Server* run in the SME's

datacentre in order to have more storage than 100GB.

5.4 Snowflake

Snowflake is a data warehouse solution build 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 analytic solutions. Snowflake offers multiple editions of cloud data platform service (1) Standard Edition (2) Enterprise Edition (3) Business Critical Edition (4) 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 edition 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: (1) Database Storage (2) Query Processing and (3) 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 an cloud storage. *Snowflake* manages all aspects on 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. The warehouse comes with different sizes ranging form XS (extra small) to XXXL, depending on the organisations needs. Each size comes with different credits which is important for the 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 the consumed on a monthly basis. Furthermore *Snowflake* offers an 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. And in addition one TB of on demand storage will cost $45 per month.

To sum up, snowflakes data warehouse platform offers all the tools necessary to store, retrieve, analyse and process data. The platform provides good solutions for small business as well as for big organisations.

6 Conclusion

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] Sandra Durcevic. "Guide for Big Data, Business Intelligence & Analytics for Small Business". In: *datapine.com* (Nov. 2018). URL: https://www.datapine.com/blog/business-intelligence-for-small-business/.

[5] Jeanne Schreurs, Dirk ROOX, and Rachel Moreau. "A data warehouse system in business performance management in SME's". In: (Dec. 2007).

[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).