Benefits of migrating to the modern data stack

Szukics Ádám
102956
ISCTE
szukiadam@gmail.com

Abstract

Data is getting more and more important for every single company, be it a consultancy, a mid-sized company or one of the biggest enterprises globally. Different teams have tried different methods to craft insights from their data, one of the most up-coming technology stack being the so-called “modern data stack” (MDS). This research aims to find out what are the benefits of adopting this stack, which companies should think about migrating to it and in which circumstances, thus providing useful information to data professionals.

1 Introduction

According to the results of a survey made in 2017, companies with the greatest overall growth in revenue earnings receive a significant proportion of that boost from data and analytics [1]. Data is an ever growing asset in our world, and it seems like the degree of its growth is only rising.

Making sense of this amount of data is a recurring problem for every company, regardless of its size. Even though we have come far from the early data infrastructure of the the 90s, companies are still not quite where they would like to be, and it is hindering their performance. The number of KPIs and metrics companies can invent is only going to increase, and making sense of this mess is something we are only striving for.

However, as we have been trying to solve this problem for a really long time, it seems that we are getting closer to an ideal solution. Previous failures and success stories show us the way, even if the advancements are not enormous.

These advancements and failures are what the MDS represents. It’s our experience in the form of tools that are supposed to help upcoming startups and small to mid-sized companies. In this paper we will look at how this modern data stack (MDS) is set up, which pillars it relies on and how it might help companies achieve their goals regarding data-related tasks.

This remainder of this paper is organised as follows. Section 2 outlines the reasons we needed to conclude a multivocal literature review instead of a systematic literature review. Section 3 describes the research questions and the study protocol. Section 4 presents results of the study. Section 5 summarises our study and recommends ideas for further investigation.

2 Research Methodology

In this section we will present our research method, the reasons we chose to conduct a multivocal literature review (MLR) and the criterias we used to filter the literature.

2.1 Initial Search

After conducting the initial search on the modern data stack, we realised that the amount of publications on this topic in

academia is insufficient. This was not surprising, as the term itself is really new and academia is usually really slow to react to industry trends. In order to dig deeper and find answers to our question, we decided to continue the research by conducting a multivocal literature review.

2.2 Multivocal Literature Review

Systematic Literature Review (SLR) studies are becoming popular in software engineering due to their trustworthiness. However, as SLRs only include formal literature, they are usually lacking in terms of practicality: it is not always clear how one should act with the newly acquired information. For a practical and fast-changing industry as software engineering is, it's key to act rapidly on insights.

The ability to include grey literature in MLRs is obviously a great asset, but in order to maximise its effect researchers needed rules. As the popularity of MLRs rose we were given guidelines on how to conduct an MLR [2].

2.3 Criteria to include grey literature in our review

Based on suggested guidelines for including grey literature in reviews, we asked ourselves the following questions.

#	Question	Possible answers	Our answer
1	Is the subject "complex" and not solvable by considering only the formal literature?	Yes / No	Yes
2	Is there a lack of volume or quality of evidence, or a lack of consensus of outcome measurement in the formal literature?	Yes / No	Yes
3	Is the contextual information important to the subject under study?	Yes / No	Yes
4	Is there a large volume of practitioner sources indicating high practitioner interest in the topic?	Yes / No	Yes

3 Research method

In this section we will describe what questions we wanted to answer with this study and in what way we found the literature we used. We will list the search engines and the keywords we used to narrow down the result set.

3.1 Research Questions

This MLR was conducted in order to understand how the MDS improves the current data infrastructure landscape, how it differs from their previous solutions and which companies can benefit the most by migrating to it. To specify the aim of this paper, we formulated the following questions:

1. What are the characteristics of the MDS?
2. Which companies benefit the most from the MDS?
3. What are the main expected advantages of adopting MDS?

4. What are the challenges of the MDS?

3.2 Study protocol

The study protocol defines the systematic way we managed to find the literature for the review. We will list the databases and the keywords we used to find meaningful sources. We will also present why other sources were ruled out in the process.

For this MLR we used Google's search engines to find applicable literature.

- Google Search (http://www.google.com/to locate grey literature (white papers, blogs, articles)
- Google Scholar (http://www.scholar.google.com/) to discover available academic literature
- Reddit (http://www.reddit.com/) to examine the public opinion on tools and frameworks

We chose Google's search engine as the idea of the MDS if fairly new and we didn't expect any academia database or journal to conclude a paper on this topic. We knew this beforehand and thus realised that this study will mainly focus on grey literature.

To understand how the developer and data community thinks of the new tools our main source was Reddit. We decided to avoid using Stackoverflow which has previously

been used to survey satisfaction of developers, because high-level concepts and tooling questions are generally not common on that platform as opposed to real life use cases and coding puzzles.

We used 4 search terms in Google in order to find the most relevant articles.

1. "Modern data stack"
2. " "Experience" modern data stack"
3. "Companies and their data stack"
4. "Modern data stack experience"
5. "Modern data stack for companies"
6. "Modern data stack company "case study" "
7. "Dbt :reddit.com"
8. "Snowflake" reddit.com

3.3 Selection criterias

After the initial search we had to skim through the result and exclude the irrelevant papers from our list. We agreed on the following configuration:

- Inclusion criterias
  - Literature discussing the MDS
  - Literature discussing the benefits of the MDS
  - Literature discussing the disadvantages of the MDS
  - Literature discussing the opinion of people from the industry
  - Literature discussing the future of the MDS
- Exclusion criterias
  - Literature that is unavailable
  - Literature that is not providing any new information for the study

The following criterias are lacking in one important aspect: it is not free of bias. Vendors that are providing tools for the MDS might try to make their product seem more appealing than it actually is. However, these vendors seemed to be self-critical and managed to start these companies providing these tools because they were unhappy with previous tools, so for this reason we will include their articles as well.

Another important aspect of our source collection was snowballing. Whenever we found a post that had a cross-reference to another page, blog or conference talk, we added those to the list as well.

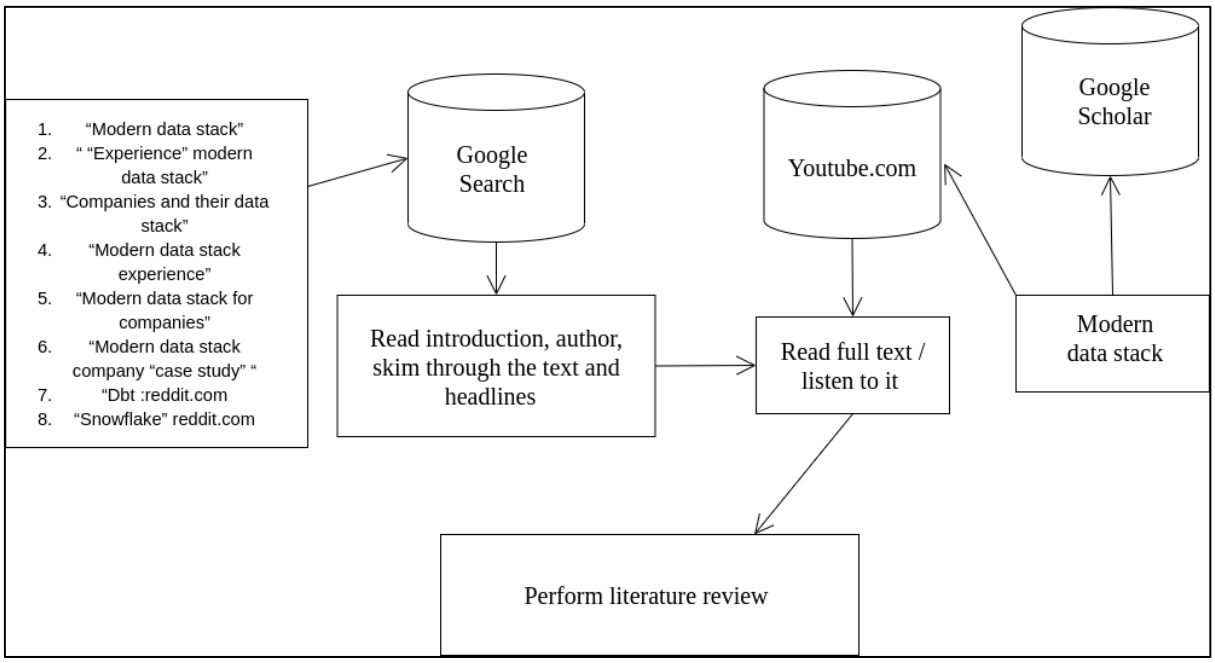

Fig 1. Research method 1

4 Results

In this section we will take a look at the results of our search process followed by our review of these materials and how they help to answer our research questions.

4.1 RQ1 - Characteristics of the modern data stack in comparison to its predecessors

We will define what the MDS is, what are its core building blocks and how it compares to previous solutions. We will also take a look at the most commonly used tools for each layer and how it has improved on their precursors.

4.1.1 Defining the modern data stack

Defining the MDS is something many people have tried, with varying success. Blog articles and interviews start by stating what it is according to persons beliefs, and it is usually different for employees at different departments and hierarchical levels [3]. One thing that authors seem to agree on is its use-case and structure: a new framework to move data around an organisation by using an ingestion tool to collect data and write it to a cloud data warehouse, where we can transform it from its raw form into usable pieces and finally using a BI tool to analyse and visualise the data. Even though new tools emerge every year, the core of the MDS doesn't seem to change [4].

4.1.2 Shift from on-premise to cloud data warehouses

The biggest catalyst for the rise of the MDS was the appearance of cloud data warehouses. In 2012 Redshift kicked things off by first appearing on the market and providing businesses a cheap alternative to store their data in the cloud. Redshifts' main charm was its speed: OLAP systems heavily outperformed OLTP systems that were popular that time, for example Postgres. Back at the time, Redshift was AWS's fastest-growing service ever. [4]

This improvement in speed managed to change the ecosystem. Previously everything revolved around the amount of time it took you to do something with your data: BI dashboards took minutes to load; running analytical queries

took way too long for not-so-small datasets. In order to solve this problem, teams tried to optimise for time: they first transformed their data and only later loaded it into a warehouse.

This change shook the data sphere by storm. Innovative data professionals jumped on the opportunity and founded their companies to complement the newly favoured product. Later others have also tried to join in on the cloud data warehouse race, and it doesn't seem to stop, especially if we consider the amount of funds these companies are getting [5].

4.1.3 Unified ingestion tools

Previously data teams had multiple sources where they wanted to collect data from, each of them requiring a different way of extraction. Different teams solved the same problems over and over and maintaining these pipelines took a lot of engineering time from other important tasks. To unify this collection process and provide an out of the box solution Fivetran stepped into the field and several companies are taking their services instead of building them in-house [6].

4.1.4 ETL to ELT

Cloud data warehouses not only brought speed, but also cheap storage. Earlier data engineers used to transform their data before they loaded it somewhere because storing the same data in different versions (raw and cleaned) was expensive. With the rise of the data warehouses this problem is gone: you can store your data for cheap and clean it there due to the fast computational possibilities. This change also brought some best practices with itself thanks to dbt, which appears to be an immovable object in the MDS [7].

4.1.5 Separate analyst teams and self-serve insights

One of the core tenets of the MDS is to make data analysis possible for every employee in the company. Even though some professionals argue that the tools of the MDS have hindered this approach, generally the idea was to make data accessible for everyone by providing tools that let people create visualisations and dashboards without the need to know SQL. This move to step away from Excel has been painful in the short term, but as every data tool revolves around SQL, it should benefit every person in the long run.

4.2 RQ2 - Companies benefiting the most from the modern data stack

We provide a brief overview of the target audience of the MDS and differentiate companies based on their type and their size in order to have a broad understanding of what they want to accomplish and their means to do that.

4.2.1 Company types

In order to be able to define best use-cases of the MDS we need to differentiate the companies that might adopt it. In general, businesses might succeed for three reasons: their novel business idea is a niche and it breaks the market; they excel at operations and manage to be more efficient and faster than their competitors; technology, specifically algorithms are their main drive. All three can benefit from using the tools of the MDS, however the way and scale by which they do it will greatly differ [8].

Companies coming up with a revolutionary idea are usually not reliant on data, so the main usage of the MDS for them is to monitor: how many sales did they have last month, what is the churn rate of their users or how many new clients can they expect after a happy customer. Based on these insights they can adopt their strategies, for example how much money they spend on a certain marketing channel. Their main goal is to find a tool that will fulfil their needs with as little resource as possible.

Operational corporations however optimise for time-to-value. They need to act fast and constantly up their game on various fronts. Continuous innovation is an often referenced phrase in these companies: they need to adopt and iterate quickly. In the 21st century data has become one of the main

drivers of innovation, thus setting up an infrastructure that enhances agility is very much beneficial.

Technical companies are more and more reliant on their algorithms and machine learning models crave for one thing: quality data. Until these companies manage to set up their infrastructure with quality data they cannot dream of success. The MDS might help them by accelerating the speed of building this infrastructure, however later on they might need to replace it with in-house tools in order to avoid stagnation.

4.2.2 Company size and their analytical needs

Businesses need analytics in order to power their decision making and adjust their strategies from top to bottom. However, the amount of it they need - and can manage - differs depending on the size of the company. In the early stages you can get away with simple analytics such as Google Analytics or using a built-in reporting service. But as you grow, more data sources will be needed [9], complexity and maintenance costs will grow. This is where you need to set up the fundamentals of your data infrastructure. As you grow even further, developing and maintaining the previous pipelines will be a pain. In order to mitigate this pain, it's advised to set up your infrastructure based on the MDS. The best practises that these tools provide will be beneficial for you in the long run [10].

Firms with more than a thousand employees can still benefit from the tools the MDS provides. Data teams face challenges regardless of company size and best of breed tools with a helping community is always beneficial along the bumpy road. Most of these challenges orbit around scalability, reducing costs, improving on time-to-value and enabling more people to do analysis [11] [12] [13] [14] [15].

4.2.3 In-house tools of IT giants

Huge tech companies tend to drive innovation and as a cause of that they are usually the first to hit roadblocks. To overcome these obstacles they built in-house tools that can specifically address these issues. As a result of this, they are usually not the target audience of the MDS ecosystem.

One can argue that these colossal companies are the main source of the MDS. In the tech world it's common to see professionals leaving a company and then building a product similar to what they have worked on for their previous employer. This is overall favourable for the industry: battle tested and widely accepted tools can emerge.

4.3 RQ 3 - Advantages of the MDS

In this section we will take a look at the best features of the MDS and why it's useful for companies in order to be able to decide whether it is useful for your case or not.

4.3.1 Improves agility

The speed at which companies make decisions depends on many factors: how agile the board is, how often they need to respond to changes or how long it takes for the dashboards to load. In case the market requires quick acting the difference between the dashboard loading for hours or seconds can be drastic. Thanks to the speed of the cloud data warehouses, the MDS offers quickness and allows companies to act rapidly upon insights.

4.3.2 Lowers total costs

Having a complete data team is often useful, however not always necessary. From the lowest layer of the MDS to the visualisations in the BI tools, every tool has built-in automations that make its usage effortless. Moving your data warehouse to the cloud not only frees up human resources, but also money. Low and declining costs of storage and computing will save you not only money, but also headaches [16].

Most of the layers contain at least one open-source tool, which makes using the MDS possible even if your local legislations are strict

and prohibit the use of foreign SaaS vendors [17].

4.3.3 Enables self-serve analytics

Prior to the modern BI tools we have now, if someone was interested in the data and wanted some analysis of it they needed a data person to help them do that. Now with the emergence of simplified tools such even non-SQL savvy employees can dig into the data and do their own research. With a proper data culture this can enable everyone working for the company to track their implemented features and try to improve on them if needed [18].

4.3.4 Avoids vendor lock-in

The MDS consists of technologies with relatively standard connection points, so swapping them around is really easy. The ability to change one tool to another made decision making for team leaders much easier. Being reliant on one tool has many disadvantages, but most importantly it might completely destroy your product if you rely on it too much [19].

4.3.5 Centralised data (single source of truth)

Analysing monthly annual users, looking at the income of different months is a task most companies face in their everyday life. One struggle they have is their data sources: how can they know which origin should they believe if they show different numbers? Keeping up the quality of the data for various sources and services is a demanding task, if not an impossible one. In order to solve this issue the MDS offers to bring all your data into a centralised place, where you can keep track of it. Engineers have an easier time keeping up and validating the quality of the data, so data trust grows, and as that grows, so does the impact of the analysis that we want to accomplish [20] [9] [21].

4.3.6 Employee satisfaction

One of the main reasons for potential hirees to turn down an offer is because they would have to maintain a legacy system in their job. However, one aspect every new candidate likes is the idea of getting some experience with tools that will probably dominate the market in the near future. As it is with the MDS, most tools are new and have the potential to be widely adopted, making candidates who have some experience with them potential targets for recruiters. This seems to be the case with dbt and Snowflake, two important tools in the MDS [22] [23] [24].

Another thing that used to annoy data professionals is the never ending cycle of data questions. Analysts and people from other departments were constantly lost on where certain data lives or what it represents. As the company grew so did the amount of questions, resulting in engineers spending most of their time on daunting tasks such as answering these questions instead of innovative and fulfilling work. As a result some of these employees had to find fitting challenges somewhere else [25].

4.4 RQ 4 - Drawbacks of the MDS

In this section we will take a look at what hardships can you expect when migrating to the MDS during the process and after you have set the foundations of it.

4.4.1 Too many tools to choose from

Having multiple options is one thing, choosing the correct solution for your case is another. Sadly, the MDS offers many options, but does not help in the decision making. Should you use Redshift, Snowflake or BigQuery as your cloud data warehouse? What are the differences, what are the benefits of using each tool? It is easy to get lost in the maze of the possibilities without any proper guidance [26].

4.4.2 Integrating many tools is still an overhead

The MDS provides many tools for many unique problems, and as things stand, it will continue to evolve even further. Every data problem (transformation, data discovery, data

governance etc.) might have its own tool, but without connecting them together they will not work. Integrating these tools into each other will cause problems, even if we try to guide the people doing it [27] [28] [29].

4.4.3 Batch based

The entire MDS ecosystem is built around batch-based operations, but data streams are just as important, especially if we want to be as fast as possible. Batches are good for analytics, but the next step is still missing from the tools of the MDS.

5 Conclusion

This MLR presents the research we did on MDS to find out what characterises it exactly, which companies can benefit the most from adopting it, what are the advantages of migrating to this stack and the challenges that a company can expect if they decide to go with it.

We used Google Scholar, Google Search, YouTube, and Reddit to find literature. After the collection process we applied our inclusion and exclusion criterias. As the term and the tools connected to it are new, we haven't found any academic research paper. Most of the literature we found were blog posts, articles, YouTube videos from conferences or interviews and forum discussion on Reddit.

We found that the MDS is made of four main components: a cloud data warehouse which is the foundation of the whole stack and the tool that started it all, an ingestion tool which you can use to collect data and load it into your preferred storage provider, a transformation tool that you can use to clean your data and transform it into your desired format and finally a BI tool where you can create visualisations and dashboards to make your data interpretable.

We realised that the companies that could benefit the most from using it are small to mid sized companies with tech teams, but not technically oriented. For others it might be useful as well, but it mainly depends on the use case.

As future work it would be interesting to conduct a survey on how many companies are using it and why certain companies decided against it. It would also be interesting to see after a couple of years of usage how employees and company leaders are describing their experiences with it.

References

[1] „https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/catch-them-if-you-can-how-leaders-in-data-and-analytics," [Online].

[2] M. F. M. V. Vahid Garousi, „Guidelines for including grey literature and conducting multivocal literature reviews in software engineering," 2019.

[3] „https://benn.substack.com/p/metrics-layer," [Online].

[4] „https://blog.getdbt.com/future-of-the-modern-data-stack/," [Online].

[5] „https://erikbern.com/2021/11/30/storm-in-the-stratosphere-how-the-cloud-will-be-reshuffled.html," [Online].

[6] „https://www.fivetran.com/case-studies," [Online].

[7] „https://www.confluent.io/blog/changing-face-etl/," [Online].

[8] „https://sifted.eu/articles/startup-ai-data-strategy/," [Online].

[9] „https://www.fivetran.com/case-studies/canva-builds-360-degree-customer-view-with-fivetran," [Online].

[10] „https://blog.getdbt.com/the-startup-founder-s-guide-to-analytics/," [Online].

[11] „https://www.fivetran.com/case-studies/case-study-ignition-group," [Online].

[12] „https://www.fivetran.com/case-studies/case-study-docusign]," [Online].

[13] „https://www.fivetran.com/case-studies/case-study-strava," [Online].

[14] „https://www.fivetran.com/case-studies/memrise-makes-online-learning-smarter-with-fivetran," [Online].

[15] „https://medium.com/eureka-engineering/evolution-of-eurekas-data-platform-918ee7f787dc]," [Online].

[16] „https://www.fivetran.com/blog/what-is-the-modern-data-stack," [Online].

[17] „https://www.datafold.com/blog/the-modern-data-stack-open-source-edition," [Online].

[18] „https://www.metabase.com/blog/The-Modern-Data-Stack/," [Online].

[19] „https://www.youtube.com/watch?v=AqrTojlYjac&list=PLrSbb3LJ2TFrfyyJzU7MzWi_De34id0yz&ab_channel=Grouparoo, " [Online].

[20] „https://monzo.com/blog/2021/10/14/an-introduction-to-monzos-data-stack," [Online].

[21] „https://www.fivetran.com/case-studies/canva-builds-360-degree-customer-view-with-fivetran," [Online].

[22] „https://www.reddit.com/r/dataengineering/comments/bqwv7p/an_interview_about_how_dbt_enables_your_data/," [Online].

[23] „https://www.reddit.com/r/dataengineering/comments/pomqcg/dbt_looker_whats_it_all_about/," [Online].

[24] „https://www.reddit.com/r/dataengineering/comments/ihpt0f/opinion_on_snowflake/," [Online].

[25] „https://mattturck.com/dbtprefect," [Online].

[26] „https://sisudata.com/blog/modern-analytics-stack," [Online].

[27] „https://benn.substack.com/p/the-modern-data-experience," [Online].

[28] „https://mode.com/get-dbt/," [Online].

[29] „https://www.holistics.io/dbt-integration," [Online].