CUSTOMLYTICS

eBooks

Introduction to Data Warehousing

This information was written by the Customlytics team for a blog post series on the Customlytics App Marketing Blog.

Table of contents

1 Why you should invest in a data warehouse

Every big company such as Netflix, Spotify or Airbnb has it and is proud of having it and speaks about it: A data warehouse.

The big question is why does your company need data warehousing? Here is our answer: It's a time saver providing efficiency and allowing you to scale analytics. E-commerce companies like Amazon, Netflix, Lyft are "trendsetters" when it comes to using data as a competitive advantage. While doing so they put pressure on their competitors to keep up with them.

Something similar is happening for the mobile app market, independent from the vertical.The market has matured over the last years and data has become a competitive edge for many developers and marketers.

1.1 The evolution of analytics

All of them have gone a long way from using (many) Excel files until they had their own data warehouse and BI-Tool. (see Ben Webers Medium post series or his book. The demand for data grows as the business do, in order to allow stakeholders across all functions to make informed decisions to further grow the business.

An often described analytic evolution is a scenario where a developer or a marketing person takes care of ad-hoc data requests outputting a CSV-file for C-/Middle - management. Using data creates an even higher need for data. If a top-level question show problems or surprising results. People want to understand why they see the unexpected result and request more data.

There will be a point when the time a developer has to spent or a marketing person has to wait with data requests has become his full-time job. This would be the time to consider building a data warehouse.

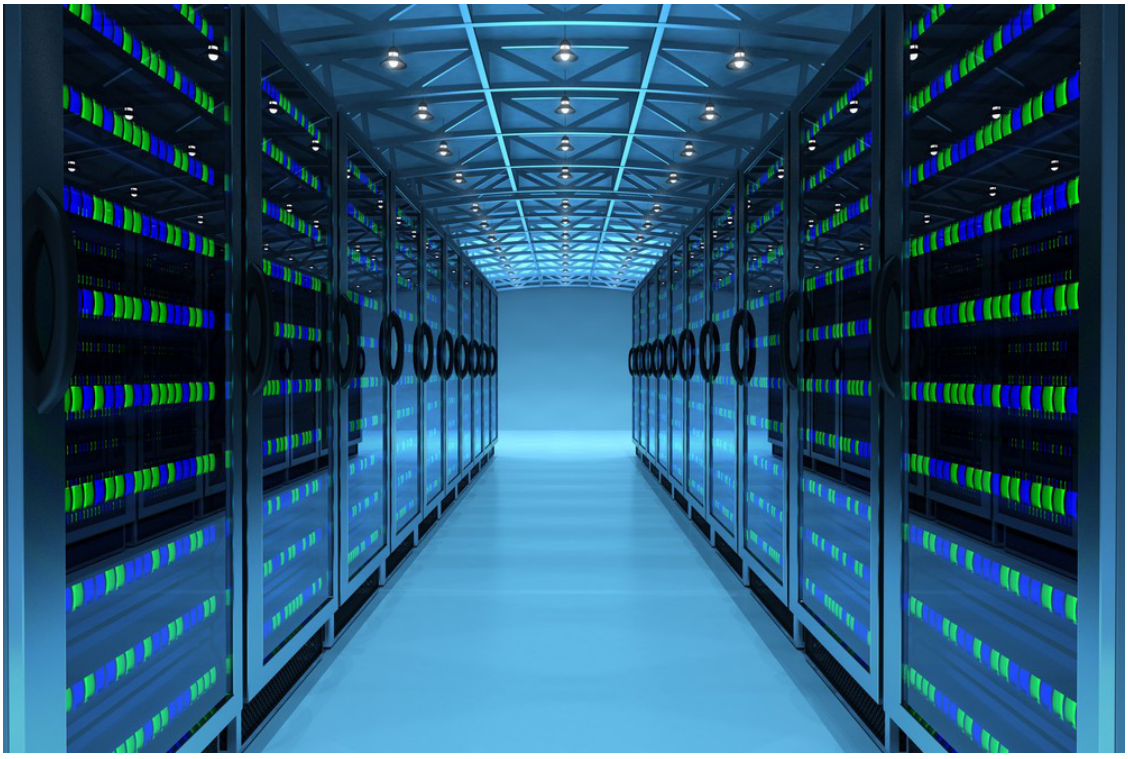
1.1 The evolution of analytics (continued)

A typical argument against it is: “Why should I invest in a data warehouse and a BI/A-team, which doesn’t bring any value?”

Well, the question has it’s right of existence. A BI-/A-Team is truly not generating any direct profits or return on investment. But if you consider it as an investment into work efficiency and time, it pays off.

Having dedicated analytic resources has the following benefits:

- Free time for developers pulling data and build reports
- Easier and timelier access to data for all stakeholders
- Building a scalable and adaptable data structure in order to prevent expensive code refactoring to get the data needed
- Minimizing negative impacts on the production server, since data is replicated in the data warehouse



1.2 Head to the cloud or not?

After you decided to follow down the path of building your data warehouse you will be faced with the decision of having a cloud or on-premise solution.

The Customlytics team believe the answer is simple: A cloud solution is the answer. Why? Because it scales with your demand and doesn't require huge stuff to operate. Another good argument for a cloud solution is the development and further competition in the cloud market. It's becoming easier and cheaper to build a scalable and maintainable data warehouse. Which makes it a viable solution even for small companies.

Cloud providers also offer credits, which allow you to build and test your setup during the initial phase:

- AWS Activate
- Google Cloud Platform
- Azure Microsoft
- IBM Cloud

Now let's touch on the efficiency argument from earlier. As your company grows you have to incorporate more and more data sources to perform the requested analytics or enrich your data.

This quickly becomes a time-consuming and tedious task to do it for every requested report. The natural next step is to automate those tasks. Also making the data available on-demand as opposed to on-request brings you naturally to a data warehouse.

A data warehouse serves as a source of truth for your company. It's the place where all first and third-party data is combined and waiting to be analyzed and consumed to generate more insights.

1.2 Head to the cloud or not? (continued)

An example of third party data is listed below:

- Production server logs
- Mobile Measurement Partner (MMP) data & marketing data
- iTunes/ Google Play stats
- Finance data
- Mapping IP to Geo

1.3 The basis matters

All data in the world doesn't help if the data doesn't provide the answer to your business questions.

So defining the right KPIs for your business is crucial as deciding for the right technologies.

Across a company, different stakeholders require different information at different frequencies or levels of detail. All these thoughts should be flow into the architecture of your data warehouse because it will enhance the latency a stakeholder has to wait until he gets the data needed to make a decision.

2 The Why, When, How and Whom of data warehousing

We think a data warehouse has required the asset to stay competitive and cope with the increasing amount of information. Further, the latest developments in the cloud market allow even startups to build their analytics platform. Which was previously an asset only big companies with financial and personal resources could afford. But not anymore.

We want to challenge you with questions to help you figure out some crucial answers, which have an impact on the architecture of your data warehouse.

2.1 When to start?

We argue that data is needed from the get-go! Might it be to convince investors, check your company KPI's or simply check server/app performance.

2.2 How to start?

There are two answers to this question, which depends on your company. If you are right at the start of your journey you should think about which KPI's and data are needed and how to collect this data. Along with a scalable and modifiable way to collect and store the data, because the demands and constraints will change over time.

The other answer is when you are in an established company. Take a look around and check which data is already available and what are KPI's requested by different stakeholders to see if all questions can be answered with the available data.

It is a good idea to care about a good data structure early on. You want to avoid the need to refactor bigger parts of your code-base and infrastructure when new requirement appears.

With such a structure you can start with an Excel/ script reporting and later transition to a cloud or on-premise solution.

2.3 When to start?

Investing in a data warehouse is an investment in efficiency and the future. First, your developer might be able to satisfy the need for data with excel/ script reporting. But data creates the need for even more data! This point can't be more stressed as 3rd party data has to enrich your own 1st party data in order to make decisions.

Second, excel and scripts are not endlessly scalable. With the growth of the company, the demand for data is growing. Switching to a cloud solution allows even smaller companies with little effort and budgets to build a scalable data warehouse.

2.4 Who to start with?

In our opinion, you have two options (ordered by our preference):

- Data Engineer - models and defines data sets, writes scalable ETL processes and data collection code, ensures data is clean and well structured
- Data Scientist - applies advanced statistics to data to discover new insights, builds Machine Learning models, figures out ways to optimize business processes using data

Without data, an Analyst or Data Scientist can't work. Which makes the Data Engineer the first choice. Since Data Scientists are (considered) to be to a degree a "Jack of all Trades" they are a potential option for a first hire.

Don't forget to think about which technology you like to use. If you already use products of one of the cloud providers it makes sense to utilize their data warehouse products.

Be open about this part, people tend to have a favorite tech stack and like to stick to them.

3 How to avoid pitfalls working with your data warehouse

In the third and final chapter of the Customlytics data warehouse ebook, we will finally talk about the daily routines of managing a data warehouse. Prior to this article, we encouraged you why you should invest in a data warehouse, and we investigated the architecture of a data warehouse.

3.1 General aspects to consider when managing a data warehouse

Once you have set up the data warehouse and the data is coming in steadily, what are the aspects you should take care of? Here is a list of essential details that I think are crucial, and which will be put into perspective in the following:

- Monitoring - Enables you to inspect system performance and quickly spot if things are breaking down.
- Business Intelligence (BI) tool - make the collected data accessible and visible for everyone.
- Documentation - As part of good data governance, we suggest that you document your data warehousing.
- Integration of new data sources - think about how to integrate new data sources with minimal effort.
- Data discovery - with the increasing number of data sources and tables it becomes harder for everyone to find the right dataset for the right purpose so having the right tools in place is important.

Now, let‘s put these aspects in order and perspective by digging deeper into the means of maintaining your data warehouse.

3.2 Daily routines of managing a BI tool

Often, the decision which BI tool to use in order to visualize data, dashboards and reporting has already been taken alongside the decision to actually build a data warehouse. Most definitely someone on the team has gained experience with some of the BI software that is out there in the market and can recommend a tool according to features and price model. So what aspects do you have to take into consideration on a daily basis for a BI tool?

First of all, **user and user rights management**. This means granting users access to the BI tool and making sure they have the correct rights according to their role in the company. Stakeholders in your company want to get the data and insights according to their needs. For instance, finance needs to see different dashboards as opposed to the sales department.

Secondly, depending on if you decided to go with an on-premise solution, regular updates have to be performed. Choosing a cloud solution for your Business Intelligence, however, saves you a lot of time and hassle because you don’t have to take care of updates.

3.3 Documentation is the foundation

The **documentation** is something you should start along with the creation of your data warehouse and should encompass design decisions, a technical description, database schemas and a description of all data processing pipelines. The document should include description and imagery at varying levels of detail for different types of teams and stakeholders.

The main goal of good documentation is to ensure that everyone is headed in the same direction. Detailed documentation is necessary for colleagues working on or with the data warehouse. Whereas a more abstract form of documentation can be provided for other stakeholders in the company and presentations. But documentation is worthless if it isn’t kept up to date. So yes do it, it will save you a lot of trouble and time.

3.3 Documentation is the foundation (continued)

The next point we want to mention is monitoring. It needs to be done or otherwise you realize at some point no data has been ingested. This can be very crucial for your business and money will be lost. That's why keeping a close eye is important. This brings us to the point of what should be monitored:

- All data imports (batch or real-time ingest) (i.e. API or server errors)
- Data processing job(s) (i.e. failed scheduled queries)
- Velocity and volume of the data.

If you use Google Cloud Platform you can utilize Stackdriver or check the email notification checkbox while setting up Pub/Sub or a scheduled query in BigQuery.

But sometimes **not an error indicates that something is going wrong**. Sometimes a **drastic change in the amount of data you receive is already indicating a problem** and requires you to take action. Consider the following example: a change was applied to an API you use and afterward the amount of data drops, which might be a sign that you don't receive the full data. So make sure the amount of ingested data is also monitored.

A situation you will encounter and should plan ahead is to integrate an additional data source. So what could be done to make the process easier?

Make a list of all the tables you plan to create or change. This list will serve you as a checklist for the Q&A part and updating the documentation. So don't miss anything.

Also, spend some time thinking about how to orchestrate the new data import to the existing one. There could be some constraints (timewise) from the tables and their data sources you plan to change.

3.3 Documentation is the foundation (continued)

I would recommend making a staged deployment. First, deploy the new data import and make sure it works. Afterward, start to change the data pipeline. At last, don't forget to keep the documentation up to date.

Last but not least I want to talk about **data discovery**, which will become an issue the longer the data warehouse exists and the more first and third-party data sources are added. Examples for first-party data are Facebook or Google Adwords and a third party example is MaxMinds IP2Geo database, which maps IPs to geolocation.

If you have a well-documented data warehouse you are halfway through it. The information provided in your database just needs to be accessible to everyone in a way that they can find the right dataset for the question, they are looking to answer.

4 Conclusion

From a technical perspective, a data warehouse is an investment in your company's marketing efficiency. For Marketers, a data warehouse is the source of knowledge where they find all information easily they need to make data-driven decisions.

Thinking about how to incorporate data into daily workflows is something to be done sooner than later. It is easier to do while the company grows than doing it later on. The needed structures and habits can be incorporated while scaling the company.

4 Conclusion (continued)

The initial phase can be summarized in three steps:

- Think about what your KPIs are and if additional data is needed on how to acquire those data points.
- Make the initial hire and start a data team.
- Research the tool landscape and get some ideas about what you like to use and how already used technologies can be utilized in your endeavor. (i.e. Firebase and Google Cloud Platform with Big Query and Looker)

One last thing, don't forget to build your monitoring system, for data imports, ETL pipelines or cost monitoring.

Yet, the work isn't done once you have established a data warehouse. The biggest part is about to come: Maintaining the data warehouse. Documentation is important and shouldn't be neglected because it serves as a starting point for other tasks (like data discovery or integrating additional data sources).

Making the move to a data warehouse strategy and don't know where to start?
We here at Customlytics cover all app marketing topics, from implementing and guiding you through the technical set up all the way to crafting a powerful marketing strategy. We're here to help if you need actionable tips for your data warehousing strategy. Drop us a line via email info@customlytics.com.