to dive deeper into this area. As market basket analysis can be highly useful for retailers, it is important to research the best way of developing a system for this purpose.

1.2 AIM AND RESEARCH QUESTIONS

The aim of this thesis is to study different data warehousing architectures as well as different data mining methods that are applicable to transactional data to gain insight about customer purchase patterns, namely in the form of market basket analysis. There are various methods available, and it can be quite challenging to know which methods should be used in different scenarios. The aim is to then design a generic data warehouse that can be used as a starting point when designing a data warehouse solution for market basket analysis in the retail industry. The resulting model is not meant to be a complete data warehouse model for a retail company, but rather a part of a solution that can quickly deliver results. The architecture should be easily scalable and support various future BI/Analysis needs. The designed system will also include a market basket analysis solution. The goal is to finally develop the proposed system to test its functionality and analyze the results. I try to answer the following research questions with this thesis:

1. *How should a data warehouse be designed so that it can support market basket analysis?*
2. *How should automated market basket analysis be applied?*

With this thesis, I hope to shed light upon the benefits and drawbacks of the different methods available in literature.

1.3 LIMITATION

My thesis will be limited to data warehousing models and market basket analysis in the retail industry. The thesis will focus on existing theories and models introduced in literature. Challenges in big data will not be part of this thesis. Additionally, I will not try to improve or do performance testing on different algorithms. I will further limit the algorithms discussed in this thesis to Affinity Analysis and Market Basket Analysis. Although data presentation is a crucial part of the models that are studied, the thesis will

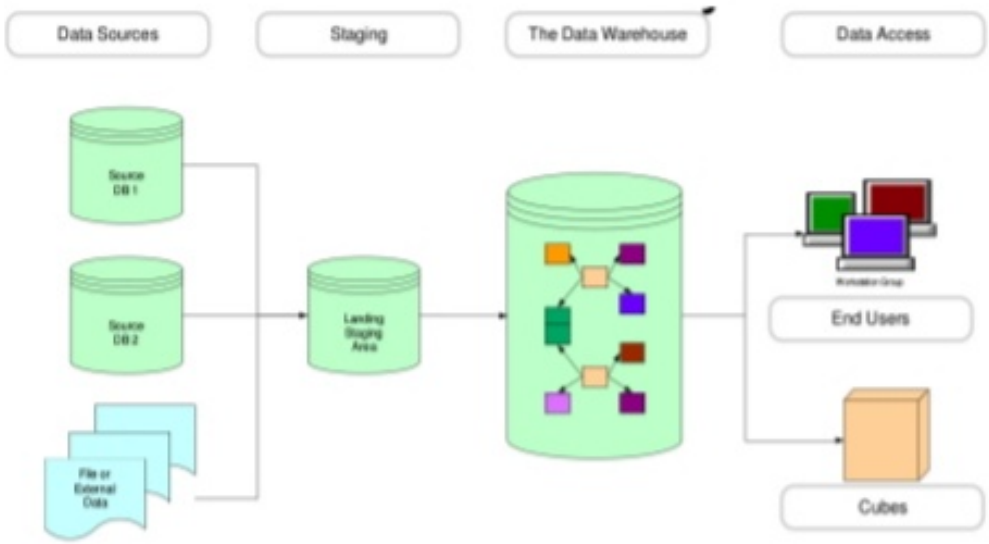

Figure 2. Two-layer architecture. (Abramson)

Another option is a three-layer architecture illustrated in Figure 3. This architecture has been introduced by Inmon. In this architecture, the middle layer holds the atomic raw data that is modeled in 3NF. The goal of this layer is to capture all data in the organization, and it is based on the sources. This layer reminds more of a large operational database. On top of this normalized layer, there is a data mart layer. This data mart layer is most often based on dimensional modeling (Linstedt & Olschimke, 2016).

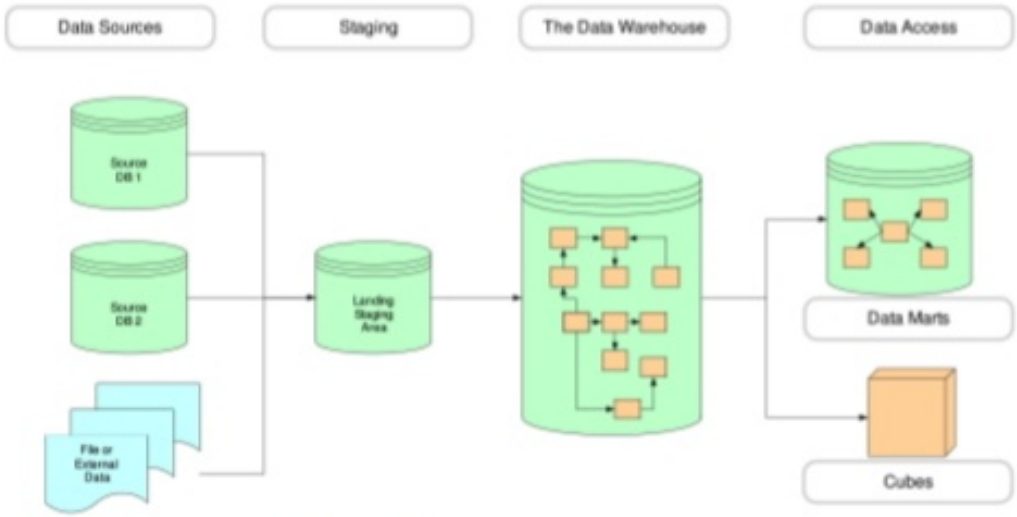

Figure 3. Three-layer architecture. (Abramson)

These architectures and approaches are discussed further in the following chapters.

6.1 FUTURE RESEARCH

How well does the star schema support expansion to more data and different types of analytics?

Further research should focus on investigating how well a data warehouse based on dimensional design can handle expansion, and how non-relational data can be incorporated to the model. It is also of interest whether a dimensional design will provide the best foundation for an ever increasing variety of analytics, as some other type of solution might prove to work better, especially when the volume and variety of the data increases.

How effective is an automated system compared to ad-hoc analytics?

It is questionable how useful an automated market basket analysis system will be in practise. This is most likely highly dependent on the business and the data. The results might prove to not be useful, and incorporating these results might even lead to poor decisions and campaigns. This means that in many cases, simple ad-hoc analysis might prove to be more cost-effective with greater results.

How effective is automated use of the MBA-results?

The results of an automated MBA-system can be used for many purposes. It is also possible to build sophisticated applications and further analytics that use the results. Some examples are recommender systems, marketing emails and group pricing. Further usage options and their effectiveness should thereby be researched.

based on the Kimball approach, meaning that the layer is based on dimensional modeling.

4. Analytics layer: In this layer MBA is performed. Any future BI/Analytics will be built in this layer.
5. Presentation/End users: The data is delivered in a usable and easily understandable format for end users in a way that can help in decision making.

Figure 29. The proposed architecture.

An example of a retail POS-system is illustrated in Figure 30. It is fairly simple to extract the needed data from the source tables into the staging area and load the data into a star schema. In another situation, the data might have to be delivered as flat files, and in this case the staging tables are based on the columns in the flat file(s).