International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Data Warehouse Success

Atul Bhati, Prof. Ramarcha Kumar

Galgotias University,India

1 Introduction

Database theory has become widespread, but in the beginning, there was neither an enterprise prevalent vision nor a more global vision to solve a variety of requests of the organization. In 1980s, subsequently a more sophisticated notion of database emerged, which was defined as Data warehouse by IBM researchers Barry Devlin and Paul Murphy. Data warehouse definition provides a broader picture of operational system in supporting dealing with a wide range of requests and queries. (Rouse, 2018). For business executives, data warehouse brings significant competitive benefits for their enterprises. On the other hand, data warehouse has helped managers and many other end users to overcome traditional roadblocks with more specific business information. Nowadays, technology has grown with data warehouse correspondingly. With the rapid increase in data use together with the transition to big data era, data warehouse architecture has significantly taken a huge shift from traditional onsite warehouses towards cloud-based data warehouses. A crucial factor affecting the evolution of modern data warehousing is the cloud service. A large portion of companies use the cloud-service data warehouse as an effective way to a have a low-cost storage, easy scale up/scale down capability, flexibility, loss prevention, sustainability and more.

The objective of this thesis was to present a fresh perspective view of modern data warehouse and an idea of utilizing the advantages of cloud-base data warehouse so as to solve case company internet advertisement problem. To be more specific, the case company was looking for a solution to track the delivery process and goals of services that the company provides. The solution would improve business decision making and subsequently return more successful business outcomes. (Because of business secret, the author cannot disclose the company's real name.

Data warehouse, enterprise data warehouse (EDW) and data mart

A data warehouse is constructed by combining data from numerous different sources with the purpose of analyzing data for business ideas. This step is known as the most crucial part for any data warehouse operation. Data warehouse is seen as a repository to store extracted information which is necessary for business solutions.

An enterprise data warehouse often refers to a data warehouse for a company, typically it is a place for all company data or at least majority of that. An EDW

makes data more solid from multiple sources and also accessible for different responsible departments in a company. In other words, data is normalized and standardized in order to meet various reporting purposes in business operation. (Adams, 2018). A traditional platform for an EDW or plain data warehouse has been on premise servers. Besides, operational database such as general ledger (GL) and other intra company services are traditionally hosted on-premise.

A data mart is a miniature, subdivision of the data warehouse system that concentrates on a particular business unit such as marketing, sales, logistics or a decision support requisition is intended to meet an immediate requirement. To build a working data mart model, firstly, it is crucially important to build a set of clean consistent dimension tables. Secondly, data mart designers need to ensure that fact tables created from joining up dimension tables do not have dimensional data, adequately, just the facts (Standen, 2008). Dimension tables mentioned here mean tables that store the objects involved in a business intelligent effort whereas fact tables store the data corresponding to a specific business process. Each row in fact tables contains the measurement data associated with a single event occurred within the business process. (Chapple, 2018). A data mart may or may not be dependent on these other data marts in an organization and is also targeted to meet an immediate requirement. If data marts are using the same dimensions and facts, they will be linked together (SIlvers, 2008).

Data presentation layer

The data presentation layer generates required data to end users in many formats depending on their demands. For instance, this layer may provide product or service insight data querying possibility and even support developing automated or ad-hoc reports. (Wainstein, 2018). Usually Business Intelligence and analytics tools are used in this layer.

Business intelligence (BI) is the ecosystem of skills, technologies, analytics and human expertise to enables an organization to get insight into its critical operations through reporting and analysis tools. BI applications may consist of awide range of components dashboard, scorecard, drill down, slicing and dicing data, spreadsheets, tabular reports, and charts.

Analytics is the practice of supporting decision-making through number- crunching which takes BI to the higher level. It assists process for instance customer segmentation, department spending. Furthermore, in order to support effective analytics, some tools and techniques are often used such as data mining, regression modelling or statistical analysis (Haertzen, 2012).

Back to 2005, tools such as Google Analytics (GA), were some of the first to offer mainstream access to analytics which plays an important role in obtaining rich insights about website traffic for enterprise. The reports and dashboards within GA provide comprehensive information to various questions that some website owners did not even consider possibility to make a beneficial contribution to their enterprise. Figure 3 below illustrates multiple benefits of BI and analytics to enterprise.

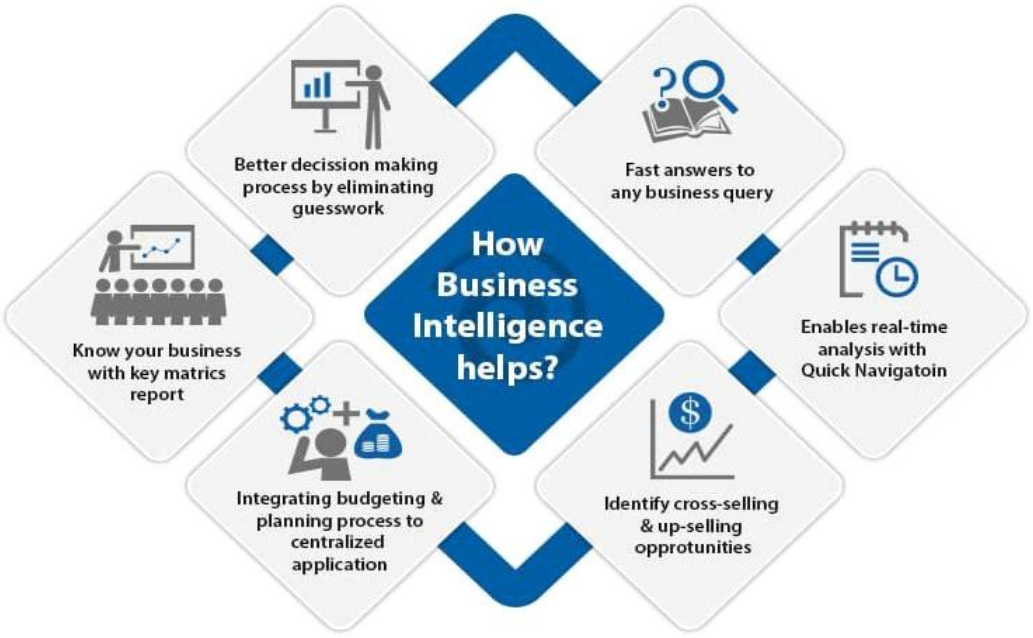

Figure 3. Benefits of business intelligence and analytics (Hoppe, 2016)

As illustrated in Figure 3, business intelligence and analytics turn data into insights and actions. Indeed, instead of relying on a great deal of guesswork, business managers can utilize insights to accelerate the decision-making process. In addition, analytics data plays an important role in creating faster reports, analysis or plans. Reports generated from BI tools contain key metrics and can be used to answer any business query. Thanks to real-time analysis withquick navigation provided in some BI tools, strategic decision making is no longerblocked, and business managers react to market changes promptly and more accurately.

- **Operational model and data warehouse model**

An operational database model is generally known to stock and administer data in real time for an enterprise. They have a critical impact to data warehouse system as well as business operations due to the fact that they contribute as the essential source for a data warehouse. Those databases are usually in form of SQL or NoSQL-based. The crucial characteristic of operational model is their orientation toward real-time transaction. In this database, records can be added,deleted and adjusted in real-time. The comparison between operational databaseand data warehouse is shown in Table 1.

Table 1. Operational database *and* data warehouse (Ricardo and Urban, 2017)

Operational databases	**Data warehouses**
For data retrieval, updating and management	For data analysis and decision making.
Focus on data in	Focus on data out
Stored with a functional or process orientation	Stored with a subject orientation
Represent current transaction	Read historical data
Generally, update regularly	Non-volatile
Complex data structure (relational database)	Multi-dimensional data structure or relationalformat
Use OLTP (online transaction processingsystem)	Analytical software such as data mining tools, reporting tools and OLAP (online analytical processing)
Support a limited area within the organization	Provide view of entire organization
Sources from operational domain	Combine from multiple sources (include operational system)

1.1 **Cloud data warehouse architecture**

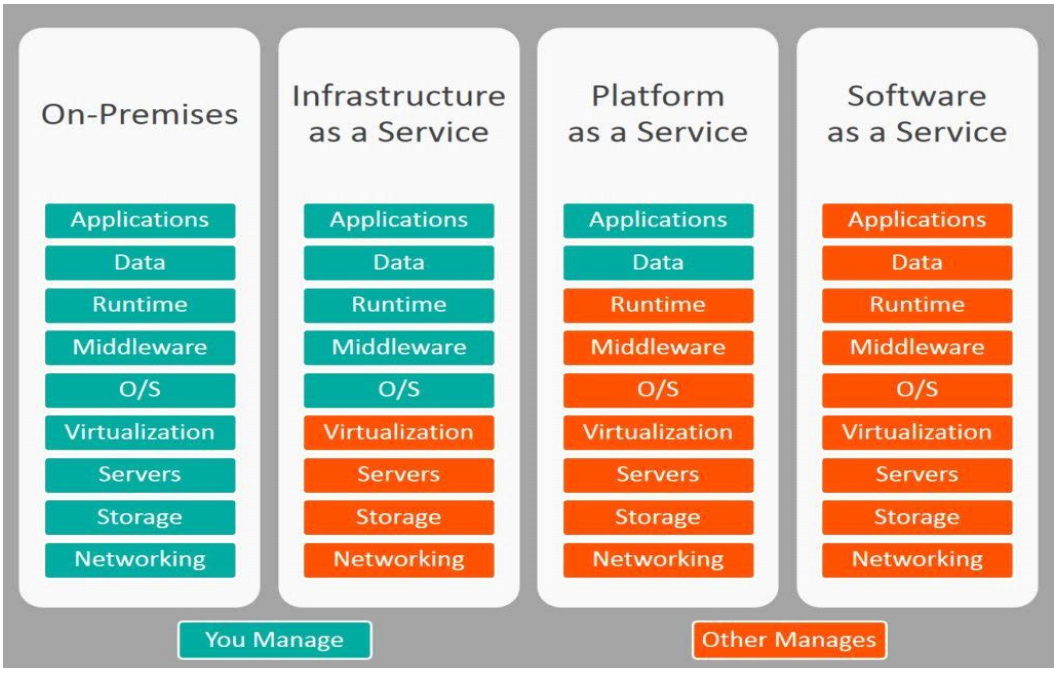

Figure 5. Summary of key differences of On-premises, IaaS, PaaS, SaaS infrastructure (Watts,2017)

Technology operations provide customers' access to their data warehouse deployment through their subscription or usage-based model. However, the following cloud approaches provide considerably various product abilities including PaaS (platform as a service), IaaS (infrastructure as a service) and SaaS (software as a service). The following Figure illustrates some remarkable differences among various cloud data warehouse approaches.

As shown in Figure 5, IaaS provides virtual space, storage and resources through a server helping to create a place called "infinite scalability''. Operating system and applications are installed and updated by customers. They manage all aspects of data warehouse hardware and software. This gives them more flexibility in using resources for what purpose. With this type of cloud approach, businesses can easily leverage level up or down infrastructure to meet demand at an effective price without building internal servers for storage and support. IaaS is on the increase due to the explosion of artificial intelligence (AI), Business

Intelligence (BI), Internet of Things IoT and cloud-based products - all of which require large amounts of storage space and computing power.

PaaS provides hardware as well as software as a cloud service. The provider manages all hardware deployment, software installation such as operating systems, databases, web servers, and programming environments. However, customer is responsible for software management and utilization. In addition, it allows customers to focus on specific applications, terminal services rather than wasting time on the operating system.

SaaS gives opportunity for customers accessing to cloud-based software without managing the infrastructure and the platform it is running on. The data warehousecompany supplies all software and hardware as well as all aspects of managing them. (Antonopoulos and Gillam, 2017)

Generally, an enterprise has a proper selection based on the benefits and drawbacks of the different offerings based on usage, security and availability. The architecture of a modern data warehouse nowadays varies greatly depending on their target to the market. Currently there are four popular technology vendors providing these services consisting of Amazon redshift, Microsoft azure, BigQuery and Snowflake.

Amazon Redshift was officially published in 2012 contributing to the larger cloud- computing platform Amazon Web Services. It is truly a virtual version of a traditional data warehouse and is used for large scale database with business intelligence tools. Redshift was built based on the massive parallel processing (MPP) ParAccel by Actian so as to manage tremendous scale database. It means that columnar storage technology is used for parallelizing and distributing queriesacross various nodes to take full advantage of all accessible sources. (Thelwel, 2015)

Big Query is a RESTful web service that is capable of interactive analysis of extensively large datasets in connection with Google Storage. The heart of Big Query is a novel query engine built on Google's Dremel project. Google call Big Query as an external version of Dremel query service software which is conducted to query billions of rows of data in just a few seconds. The other consideration that sets Big Query apart is that the whole process of provisioning, assigning, and maintaining of resources, is fully taken care of automatically by Google. This makes Big Query ideal for small organizations or teams that prioritize ease of use over maximum performance. (Inc, Google, 2018)

4 Case company

4.1 Business concept

Case company is a job board and recruiting media company that gathers and displays all the job postings in one place. Job seekers come to company not only for job searches but also for many useful pieces of career advice, news about working life and recruitment. Not only that, the case company helps multiple enterprises attract active and passive job candidates through bolstering their brands via video, news on social media as well abundant external website advertisements.

Case company is using internet display advertising network to drive traffic to job advertisements hosted in the case company main website. As a result, job advertisements get attention from readers/viewers, and a portion of those will potentially apply for the job and they become applicants, naturally some of these get employed in the end of the recruiting process. Therefore, one of the strategic goals of the case company is to produce relevant and suitable applicants to the customer companies. This raises the question of managing the job advertisements' delivery process efficiently, results in the recruitment of tracking nearly real-time delivery process and target guarantee for each job advertisement.

4.2 **Current situation**

This section will introduce the case company's current advertisement management situation, scattered data and data analysis challenges and solution for that.

4.2.1 Current internet advertisement management

At case company, campaign team has to manage several thousand internet advertisements daily and currently the team is using only excel sheets to manage the advertisement data. These data have become gradually larger as case company enjoys its higher demand of services caused by the increase of B2Bclients and end users. It means data flow is painfully slow and there is a lack of automated analytics reports. To be more specific, manual tasks

include checking advertisement spends and performing status, documenting them into excel sheets, necessarily, implementing some corresponding charts/ graph based on the documented data. To avoid these time-consuming tasks and improve the KPI, creating automated reports that present advertisement spends and performing status plays an important role in campaign management.

4.2.2 Internal data analytics burden

The internal data in case company is already huge, due to the fact that several thousand job advertisements come to the database weekly, as averagely about a thousand job advertisements daily. Additionally, internal database contains not only job entry data but also other data with both structure or semi-structure types. If analytics procedure implements in the current internal database, it will unexpectedly create an unavoidable burden in internal database. While the tech team always gives higher priority to optimize the back-end performance, cloud data warehousing candidate potentially offers great service for analysis, as it was designed to do analytics work.

A few studies have proved that when using cloud data warehouse, analytics processing is separated from the main internal transactional database, leaving the transactional database free to focus only on transaction. According to Inmon (2002), databases have divided into two categories, classified by the needs of their users. While serving operational needs such as transaction processing is the focus of the first category, serving informational or analytics needs is achievable in the other category. Inmon (2002) also pointed out that the split occurred for several reasons, and one of those reasons is the data serving operational needs is physically different from the data serving in-formational or analytics needs.

Thanks to the scalability and BI tools integration of cloud data warehouse, case company does not have to worry about the size of data storage as well as how to analyze and visualize data more accurately and efficiently. Figure 8 below depicts current data situation in case company.

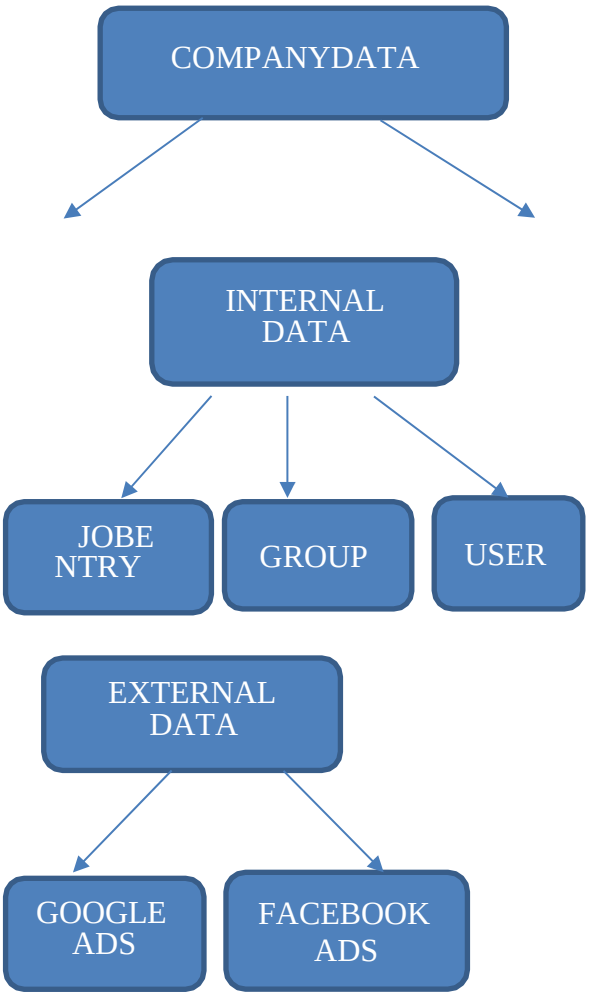

Figure 8. Case company X database scenario

As can be seen in Figure 8, case company has 2 main sources of data flow, namely, internal and external data. Internal data contains data for example job entry, user and group tables while external data are obtained from Facebook and Google ads services. The central issue addressed here is the integration of two data sources together as they have different schemas and data models. Practically, moving external database into internal database does not seem to be a good solution because it has been noted earlier as creating more burden to already heavy transactional internal database.

4.2.3 Solution for analytics data

In this section, a solution for case company's data analytics issue is introduced. To answer to the burden analytics data question described previously, project leader and business managers in company have discovered a straight forward answer, which is moving all necessary data into cloud data warehouses. Figure

the fields in the CSV data files that are staged.

Create virtual data warehouse

To create a virtual warehouse, snowflake provides 2 options: using web-based worksheet or snowSQL CLI. The author selected the second option, codes that create 2 virtual warehouses described in Figure 11.

```
CREATE WAREHOUSE X_COMPUTE_WH
WITH WAREHOUSE_SIZE = 'XSMALL'
WAREHOUSE_TYPE = 'STANDARD'
AUTO_SUSPEND = 600
AUTO_RESUME = TRUE
MIN_CLUSTER_COUNT = 1
MAX_CUSTER_COUNT = 2
SCALING_POLICY = 'STANDARD'
COMMENT = ''
CREATE WAREHOUSE X_MANUAL_WH
WITH WAREHOUSE_SIZE = 'XSMALL'
AUT0_SUSPEND = 300
AUTO_RESUME = TRUE
MIN_CLUSTER_COUNT = 1
MAX_CUSTER_COUNT = 2
SCALING_POLICY = 'STANDARD'
COMMENT = ''
```

Figure 11. Creating virtual warehouse with snowSQL.

Figure 11 illustrates how to create two new data warehouses; noticeably, manual warehouse uses smaller "XSMALL" size and set up with smaller auto - suspend time than compute warehouse does. Auto-suspend time is used to define a period of time when there is no activity and the warehouse can be suspended to avoid consuming unexpectedly wanted credits. Auto-resume sets to True to enable warehouse to automatically resume when new queries are submitted.

Conclusion

The incredible pace of change in the data center today is making companies more challenging. They are struggling to get a business model which is able to take advantage of all advanced technologies with the expectation of rapid accessibility to service. IT has experienced the same phenomenon but at a much faster pace. As a consequence, cloud computing was born as a solution for this thriving market. It has created new paradigms that align with other trends such as Big Data, Virtualization or Security.

This thesis has provided a comprehensive review about traditional data warehouse as well as cloud-based data warehouse and various key considerations between them in theory. Besides, the author also introduces the implementation and management of snowflake computing for the case company in order to get insights of job advertisement campaigns. This implementation will hopefully be a general tutorial for everyone who wants to create a cloud computing data warehouse in Snowflake from scratch since it covers outright detail from creating a new cloud data warehouse to implementing relevant ETL processes along with the given data which are scattered among the internet. Furthermore, although the project seldom mentions about logging report, setting up bash scripts that automates log report broadcasting plays an important role in monitoring the whole ETL process.

Overall, Snowflake is an attractive proposition as a Cloud Data Warehousing solution for case company. It has provided some distinct advantages over legacy technologies as outlined above. More specifically, Snowflake has proven to be a prospective platform for starting a production-ready data warehouse from day one with almost zero extra overhead for configuring the platform itself. Moreover, thanks to Snowflake native support for processing JSON (a de factor for the most internet-based APIs out there), the service itself is easy to integrate to different SaaS services such as Google Analytics and Facebook Business Manager. However, for creating and maintaining a larger cloud data warehousing solution, a visual ETL tool would bring back a great deal of advantages such as an overview of the whole process and bug reporting. Since Snowflake is a cloud database, it does not offer an ETL tool.