Which architecture is the future of Big Data, Data Warehouses or Data Lakes

Ryan Donovan
X16104269@student.ncirl.ie

ABSTRACT

Data warehouses have been around since the 1980's and have aided many enterprises in their Business Intelligence efforts. In last few years a new data collection concept has appeared, called a "data lake". Both are used to hold and analyse data to be used by BI and more. There is not a concise agreement with community on which is better or which might be the "future" technology. I will be discussing the pro and cons of both in order to answer the question: "Which architecture is the future of Big Data, Data Warehouses or Data Lakes?" I will be compare both in terms of cost (time and money), querying and schema, maintenance and metadata, data governance, and architecture matching for specific business types. I have gathered data from journals, interview, conferences, and case studies. There has been little experimental research into data lakes field.

KEY TERMS

A **data lake** is a central data repository used by enterprises to hold vast amounts of data in its raw format using low-cost storage. It can handle any type of data: structured, unstructured or semi-structured. It can be queried to answer important business questions in place. Watson (2015) stated that Hadoop is usually used for the data storage platform for lakes. A **data warehouse** is a collection of data that is used by businesses to aid with important business decisions. It needs to be integrated, subject-focused, time variant, and non-volatile (Inmon, 2005). A warehouse only houses structured data.

COST

Data warehouses generally take time to set up and also come with a big price tag as compared to a data lake, of which can be implemented quickly and with a lower price. To construct a data warehouse a great deal f planning has to be done and it could take years to implement. A data lake can be implemented much faster than a warehouse. One reason for the speed is that data is generally thrown into a lake with no format or care of future business use, as opposed to careful placement of formatted data into a warehouse. Halter stated that a lake can be waiting area for high-volume data of which can be put into a warehouse at a later time (Soares, Kutemperor, Kromer and Halter, 2016). Its cheaper to store data then to process it so it could be good idea to hold it in that waiting zone, although Breur (2015) believes this waiting area is nothing new and looks similar to staging area in warehouses that have been around for past couple of decades. Most data lakes use Hadoop for its platform and this is significantly cheap to set up and inject data. It starts to become expensive when data is to be taken out and queried because there is a lack of talent to currently perform these operations easily. In a warehouse, the querying is less expensive because the data is structured and it takes less skill to perform the queries. Also, if new classification of data was to be added to a warehouse, then you might have to unload data and then re-perform ETL, of which would be costly and time-consuming. This is easily done with a data lake architecture (Roski, Bo-Linn and Andrews, 2014). Data lakes have the advantage of being able to take a huge amount of data in quickly in any type of structure. A warehouse can’t deal with this new influx of unstructured/semi-structured data unless it is cleaned up in ETL process.

QUERYING AND SCHEMA

The lake can be queried with Hadoop on any structure data, but the warehouse can strictly only use structured data. This is a drawback because some companies are looking to query unstructured data from social media platforms and other places in order to gain new insights. A data warehouse must have a schema upfront for BI needed by an organization. Data lakes don’t need schemas up front, but do need a schema when acrtually querying the data. Most data lakes use Hadoop along with MapReduce, Hive, Spark, and Pig to enact a query schema (Soares et al., 2016).

MAINTENANCE AND METADATA

Data can be injected into a lake but it needs to be maintained with metadata for it to have some worth. Its sometimes difficult to decipher incoming raw data and maintaining its usefulness becomes very tedious, especially when more and more data sources are gathered (Terrizzano, Schwartz, Roth and Colino, 2015). Being meticulous with meta data is warranted by many in big data field in order that the lake doesn't become congested with meaningless data (bad metadata). The data must be quality, be trusted, and be legal to use (Terrizzano et al., 2015) and metadata should document these facts. There is a benefit as metadata is collected because there is more insight into new queries that weren't though of before.

DATA GOVERNANCE

Data governance entails what is done with data once its gathered. A warehouse would have an implemented governance plan when it was created and would only need to be periodically updated. A lake needs to be very careful in governance so it can maintain a plan for its security, metadata, privacy, and legal obligations on data (Varanasi, Ring, D'Antoni, Barnes and Armstrong, 2016). It is usuallt difficult to maintain governance in a lake but it is argued that it is worth it because the low cost to store data. Also, you get returns of faster analysis and answers to unthought of queries (Terrizzano et al., 2015). Fitzgerald (2015) also adds that scaling of a data lake is in direct proportion to how data governance and sovereignty is solved.

ARCHITECTURE TYPE FOR ENTERPRISES

It seems that data lakes and data warehouses might be useful in certain organizational structures. If a business already has a warehouse for analysing structured data, then there is no reason to change to a lake because of cost and time implications. Most big organizations use a data warehouse architecturs so a lake should be used when it’s worthwhile. The goals of the company and its financial state should dictate their wanted architecture. If a business has mostly unstructured data, then a lake would be a better fit. Also, if a company wants to have a lot information at hand and be able to “play around with it” in order to prototype quickly, then data lake is the way to go (Soares et al., 2016). A warehouse should be continually used unless there is a string reason to not use it.

CONCLUSION

There seems to be a use for data warehouse and for lakes. Warehouses are more expensive but most enterprise already have them implemented, so it would be costly and perhaps useless to switch to a lake (especially with the lake being such a new concept). On the other hand, lakes seem to be of great use for the huge influx of data (Big Data) that is unstructured or semi-structured. Accordingly, Hadoop aids in being able to store and analyse data (help from MapReduce and the like) for important BI insights. O’Leary (2014) argued that lake architecture is unclear but it that was in 2014 and there seems to be a clearer vision for lakes. More research could be done on how to organize the data lake to transform unstructured data into structured form because its easer to analyse and I think most people can query better with structured data.

Bibliography

Breur, T. (2015), ‘Big data and the internet of things’, *Journal of Marketing Analytics,* 3(1), pp. 1-4.

Bucur, C. (2015), ‘Using big data for intelligent businesses’ 2015 *Proceedings of the Scientific Conference (AFASES).* Brasov, Romania, 28-20 May 2015, 2, pp. 605-612.

Dyché, J. and Davenport, T.H. (2013) ‘Big data in big companies’, *International Institute for Analytics,* pp. 1-31.

Fitzgerald, M. (2015) ‘Gone Fishing – For Data’, *MIT Sloan Management Review,* 56(3), pp. 1-9.

Inmon, W. H. (2005) *Building the data warehouse.* 4th ed. USA: Wiley Publishing Inc.

McKendrick, J. (2015) ‘THE FUTURE OF DATA WAREHOUSING’, *Information Today,* 29(2), pp. 11-13.

O'Leary, D.E. (2014) ‘Embedding AI and Crowdsourcing in the Big Data Lake’, *IEEE Intelligent Systems,* 29(5), pp. 70-73.

Roski, J., Bo-Linn, G.W. and Andrews, T.A. (2014) ‘Creating value in health care through big data: opportunities and policy implications’, *Health affairs (Project Hope),* 33(7), pp. 1115-1122.

Singh, K., Paneri, K., Pandey, A., Gupta, G., Sharma, G., Agarwal, P. & Shroff, G. (2016) ‘Visual Bayesian Fusion to Navigate a Data Lake’ 2016 *19th International Conference on Information Fusion.* Heidelberg, Germany, 5-8 July 2016, pp. 987-1004.

Soares, D., Kutemperor, N., Kromer, M. and Halter, O. (2016) ‘Dipping a Toe into Data Lakes’, *Business Intelligence Journal,* 21(2): pp. 40-46.

Stein, B. and Morrison, A. (2014) ‘The enterprise data lake: Better integration and deeper analytics’, *PwC Technology Forecast: Rethinking integration,* (1), pp. 1-9.

Terrizzano, I., Schwartz, P., Roth, M. and Colino, J. E. (2015) 'Data Wrangling: The Challenging Journey from the Wild to the Lake' 2015 *Conference on Innovative Data Systems Research (CIDR).* Asilomar, California. 4-7 January 2015, pp. 1-9.

Varanasi, S., Ring, T., D'Antoni, J., Barnes, S. and. Armstrong, R. (2016) 'When It's Time to Hadoop', *Business Intelligence Journal,* 21(1): pp. 32-28.

Watson, H. J. (2015) 'Data Lakes, Data Labs, and Sandboxes'. *Business Intelligence Journal,* 20(1): pp. 4-7.

Yadav, S., Shroff, G., Hassan, E. & Agarwal, P. (2015) 'Business data fusion' 2015 *18th International Conference on Information Fusion (ISIF).* Washington, DC, 6-9 July 2015, pp. 1876-1885.