Assessment 3: Data processing trends literature review

The impact on analytics and data science of the trend towards cloud-based data warehousing

MA5831 – Advanced Data Processing and Analysis using SAS

13848336 Nikki Fitzherbert

There is a broad consensus in both the academic and non-academic literature that data has value, and insights from data inform decisions and action that can have social, economic and environmental impacts. It is also well-established in the literature that data has a significant value-add impact on the operations of organisations and public institutions (Diepeveen & Wdowin, 2020). As a result, organisations in the private sector are increasingly focused on exploiting data for competitive advantage and public sector institutions such as governments are using it to evaluate and improve decision making and policy development (Productivity Commission, 2017).

In addition, the growing amount, variety and volume of data held and generated by organisations has been driving an evolution of data management and processing technologies. One of these has been the trend towards cloud computing and specifically, the trend toward cloud-based data warehousing, which has had significant implications for data science and the role of the data scientist.

The following review begins by discussing in some depth how big data led to the appearance of the cloud-based data warehouse in the early 2010s and the take-up of the technology by both the private and public sectors. It then briefly discusses what this has meant for the subjects of data security, data privacy and data sovereignty, which is particularly pertinent in an era of rising concern from the general public about their personal privacy and under what circumstances personal information can be collected and used. The review concludes with an assessment of what some of the impacts have been on the role of the data scientist with the evolution of the data warehouse from a primarily on-premise platform to one primarily based in the cloud.

Data warehouses are not a new technology. They first emerged in about the 1990s as a tool that enabled organisations to integrate and analyse data from one or more transactional databases and produce strategic insights to support decision-making. In other words, data warehouses was developed as a type of decision support system (Ponniah, 2010). Bill Inmon, who is generally considered the father of data warehousing, defined a data warehouse as a "a subject-oriented, integrated, non-volatile and time-variant collection of data in support of management's decisions" (as cited in Ponniah, 2010, p. 23).

Data warehouses were designed from the outset to handle large amounts of data and information, and have evolved over time to support a much wider range of applications and sophisticated analytics such as artificial intelligence and machine learning (Oracle, n.d.-b). However, it was also becoming increasingly clear that data warehousing technology was going to have to evolve again in order to overcome the challenges presented by the amount, variety and speed at which data was being collected and generated by even the early 2010s (Kraynak & Baum, 2020; McKinsey Global Intitute, 2011).

There are some that had argued, even as far back as 2004, that the data warehouse had no place in the modern data environment and would shortly become a little-used data management solution (Learn Data Vault, n.d.; McKendrick, 2019). For example, Carter (2004) argued that the cost,

complexity and utility of the data warehouse would become marginalised as “distributed intelligence” became more prevalent and enabled the introduction of portable and real-time business intelligence platforms.

This was not a problem being faced solely by data warehouses. There was an overabundance of digital information being captured from sources as diverse as mobile phones, social media sites and computers as well as internal operational systems, which could be used to reveal valuable strategic insights that otherwise might have remained hidden ("Data, data", 2010; Lee, 2017). The issue then was the development and re-design of tools and technologies that could manage, process and analyse such data efficiently and cost-effectively (Kaisler, Armour, Espinosa, & Money, 2013; Thabet & Soomro, 2015). One such solution was the introduction of data warehouses located in the cloud, which made its first appearance with Amazon Redshift in 2012 (Henschen, 2012) – others have included data lakes and data hubs, but an exploration of these other architectures are outside the scope of this review.

Cloud-based data warehouses have been attractive to organisations largely because of their scalability, reduced cost overhead, and shortened implementation time (Transforming Data with Intelligence, 2016). Conventional data warehouses tended to be the exact opposite – requiring significant amounts of time and resources to alter (Rehman, Ahmad, & Mahmood, 2018). They have also been attractive due to a number of benefits shared with cloud computing in general such as theoretically-infinite low-cost storage, the ability to outsource data management and security to the product supplier, and the possibility of only having to pay for the storage and computing resources actually used (Kraynak & Baum, 2020).

In addition, the trend toward the cloud has also had a significant impact on some of the core features of the data warehouse. For example, data warehouses have traditionally been based on a relational database structure that uses an extract-transform-load (ETL) process to ingest data, but cloud-based data warehouses using a non-relational or NoSQL-based design have also appeared (Bicevska & Oditis, 2017). These are better suited to storing unstructured data and large volumes of any data type as they are document- rather than schema-oriented (Hecht & Jablonski, 2011) and therefore also tend to be associated with an extract-load-transform (ELT) process, which prioritises the rapid ingestion of data over a clean and standardised dataset (Kumar, 2020). Nonetheless, Hecht and Jablonski (2011) also observed that there was a wide variation in the data models, query languages and Application Programming Interfaces (APIs) used.

The ability of organisations to take advantage of all the potential benefits of adopting a cloud-based data warehouse has also very much depended on the cloud computing approach chosen (Kraynak & Baum, 2020; Litchfield & Althouse, 2014). As with a conventional data warehouse that has generally been designed according to Inmon’s top-down approach or Kimball’s bottom-up approach (Sansu, 2012), most modern data warehouse solutions fall into one of three main cloud computing approaches: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) or Software-as-a-Service (SaaS) where IaaS is very similar to having an on-premise data warehouse and the SaaS approach transfers all responsibility for provision and maintenance to the product supplier (Kraynak & Baum, 2020). As a result, in some instances there may be very little difference in the capability and the way data is being processed and managed compared to a conventional data warehouse (Snowflake, n.d.-a).

Despite the significant advantages offered by modern cloud-based warehouses, one of the longest-standing concerns about migrating data and systems into the cloud has been that of data security and data sovereignty (Maayan, 2018; Panoply, 2017; Robinson, 2020; Snowflake, n.d.-b).

This concern about data security is certainly justified, with the 2020 Australian Community Attitudes to Privacy Survey indicating that data privacy remains a significant concern for Australian citizens. The survey identified the biggest risks as identity theft and fraud, and data breaches and security. Furthermore, data breaches have often ended up in the media (for example Hodge, 2019), which can lead to a loss of reputation for the organisation(s) involved if nothing else (Productivity Commission, 2017).

Data sovereignty is a concept derived from the historic power of a state of absolute and exclusive control within its borders, and refers to the fact that data is subject to the laws and regulations of the country where it is physically located (Taylor, 2020). In an environment where cloud computing has become increasingly prevalent, this is an issue because the laws and regulations relating to data privacy and security have tended to vary widely between different jurisdictions (Vaile, Kalinish, Fair, & Lawrence, 2013). As a result, there are often strict rules regarding where data can reside and under what circumstances it can move across jurisdictions, which have not been without criticism. For example, Cory (2017) argued that they hinder global digital trade and the economies of the countries that implement them.

One solution that has been developed in response has been that of the hybrid cloud. A hybrid cloud allows organisations to integrate existing infrastructure and data with cloud-based applications and software. This has meant organisations could keep sensitive data on local servers and therefore maintain compliance with data sovereignty laws, but still take advantage of the processing power and analytics capabilities of the cloud-environment (Cloud Standards Consumer Council [CSCC], 2017). Furthermore, it allowed organisations to access new technologies and analytics capabilities with which to generate new insights without first having to deploy the software on their on-premise systems.

However, data security is potentially a bigger problem with a hybrid cloud, as responsibility is no longer totally with the organisation nor with the cloud service provider (CSCC, 2017). In contrast, in many cases data could be more secure in a solely cloud-based environment as cloud-service providers generally have far more resources and expertise directed towards data and system security than many organisations (Low, 2020; Steier, n.d.; Vahie, 2016). For example, cloud systems offer security protocols that many organisations may not have been able to deploy themselves or were too cost-prohibitive to even consider - such as web-application firewalls or hardware security modules (Tripwire, 2018).

As has already been highlighted, cloud-based data warehousing was developed as one way for organisations to manage and process the volume and variety of data being collected (Low, 2020). The development, and increasing prevalence and integration of cloud-based data warehouses has invariably also had a significant impact on the role of the data scientist.

Interestingly, the term “data scientist” first appeared around late 2009 to refer to someone who could combine the skills and techniques from the fields of computer science, mathematics, statistics and data mining to derive insights from large data sets (Press, 2013). This was about the same time that the term “big data” started to be used ("Data, Data", 2010).

The general definition of what a data scientist is has not changed much since then. They use a combination of domain expertise, programming, and data analytics to uncover, analyse, process and model structured and unstructured data, and then interpret and communicate those results to stakeholders. Although the literature defines the role of a data scientist as being quite different to that of other related disciplines (Doyle, 2020), in reality the lines have often become quite blurred

and data scientists find themselves performing tasks closer to the purview of those other roles (Imarticus, 2017).

Cloud-based data warehouses have allowed organisations to deviate from the traditional data warehouse architecture and offer organisations a complete and integrated solution from the data ingestion layer right through the data analysis and machine learning model deployment layer (Oracle, n.d.-a). This has meant that data scientists are able to leverage off the cleansed organisational data contained within the data warehouse to address the problem they are trying to solve, or if necessary, query raw data contained within the data lake (CITO Research, 2014).

The rapidly evolving big data and cloud computing environment has also meant that the role of the data scientist now carries some additional expectations in addition to the traditional ones of being able code, analyse and model data using programming languages such as Python, R and SQL, as well as communicate the results of the research to a non-technical audience. For example, data scientists have needed to become competent with tools associated with the analysis of big data, know how to access and query data located in modern data management systems, and be comfortable using APIs (Castrounis, 2020).

Cloud-based data warehouses have meant that organisations have become better equipped to use data to solve problems, uncover insights and achieve their strategic objectives. In particular, they have made it easier for data scientists to collaborate on projects irrespective of their physical location (Field, 2019). Nonetheless, data scientists have also needed to become increasingly aware of the broader data environment they operate within, such as with respect to their obligations regarding data security and privacy because it is nearly always the end user that is the weakest link (Ranjan et al., 2018).

This literature review explored some of the research and observations on how the trend toward cloud-based data warehousing was impacting analytics and data science. It identified that the cloud-based data warehouse had been a platform developed in response to the growing volume and variety of digital data along with other modern technologies such as data hubs and data lakes. However, the review also concluded that whilst the modern platform was attractive to organisations primarily because it could scale up and down in response to changing organisational demand, concerns about data privacy, data sovereignty and the desire to maintain sensitive data on private systems meant that cloud service providers also needed to offer a hybrid model.

The literature review also looked at some of the possible impacts this trend was having on the role of the data scientist. It found that although the general definition of a data scientist had not changed, there was a greater expectation that a data scientist was proficient in tools and languages associated with big data analytics. Also, the increasing emphasis on protecting personal information along with the lack of standardisation across countries in this space has meant that data scientists have needed to become more aware of their data security obligations in their work.

Word count: 2,200 words

References

Bicevska, Z., & Oditis, I. (2017). Towards NoSQL-based Data Warehouse Solutions. *Procedia Computer Science, 104*, 104-111. doi:https://doi.org/10.1016/j.procs.2017.01.080

Carter, M. (2004). The death of data warehousing. Retrieved from http://www.looselycoupled.com/opinion/2004/carter-dw-bp0311.html

Castrounis, A. (2020). What is data science, and what does a data scientist do? Retrieved from https://www.innoarchitech.com/blog/what-is-data-science-does-data-scientist-do

CITO Research. (2014). *Putting the data lake to work: A guide to best practices.* Retrieved from https://hortonworks.com/wp-content/uploads/2014/05/TeradataHortonworks_Datalake_White-Paper_20140410.pdf

Cloud Standards Consumer Council. (2017). *Hybrid cloud considerations for big data and analytics.* Retrieved from https://www.omg.org/cloud/deliverables/hybrid-cloud-considerations-for-big-data-and-analytics.htm

Cory, N. (2017). *Cross-border data flows: Where are the barriers, and what do they cost?* Retrieved from http://www2.itif.org/2017-cross-border-data-flows.pdf

Data data everywhere: A special report on managing information. (2010, February 27). *The Economist*. Retrieved from http://faculty.smu.edu/tfomby/eco5385_eco6380/The%20Economist-data-data-everywhere.pdf

Diepeveen, S., & Wdowin, J. (2020). *The value of data: Accompanying literature review*. Retrieved from https://www.bennettinstitute.cam.ac.uk/publications/?research_projects=10

Doyle, L. (2020). What does a data scientist do? Retrieved from https://www.northeastern.edu/graduate/blog/what-does-a-data-scientist-do/

Field, S. (2019). Unlocking enterprise collaboration with cloud-based data warehouses. Retrieved from https://www.techerati.com/features-hub/opinions/power-of-data-with-cloud-built-data-warehouses/

Hecht, R., & Jablonski, S. (2011). NoSQL evaluation: A use case oriented survey. In R. Buyya & C.-Z. Xu (Eds.), *2011 International Conference on Cloud and Service Computing* (pp. 336-341). doi:10.1109/CSC.2011.6138544

Henschen, D. (2012). Amazon debuts low-cost, big data warehousing. Retrieved from https://www.informationweek.com/software/information-management/amazon-debuts-low-cost-big-data-warehousing/d/d-id/1107568?

Hodge, R. (2019). 2019 data breach hall of shame: These were the biggest data breaches of the year. Retrieved from https://www.cnet.com/news/2019-data-breach-hall-of-shame-these-were-the-biggest-data-breaches-of-the-year/

Imarticus. (2017). What is the role of the data scientist? Retrieved from https://blog.imarticus.org/what-is-the-role-of-a-data-scientist/

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In R. H. Sprague (Ed.), *46th Hawaii International Conference on System Sciences* (pp. 995-1004). doi:10.1109/HICSS.2013.645

Kraynak, J., & Baum, D. (2020). *Cloud data warehousing for dummies*. (2nd ed.). Retrieved from https://resources.snowflake.com/ebooks/cloud-data-warehousing-for-dummies

Kumar, R. (2020). ETL vs. ELT: How to choose the best approach for your data warehouse. Retrieved from https://www.softwareadvice.com/resources/etl-vs-elt-for-your-data-warehouse/

Learn Data Vault. (n.d.). Data warehousing is dead! Retrieved from https://learndatavault.com/data-warehousing-is-dead/

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons, 60*(3), 293-303. doi:https://doi.org/10.1016/j.bushor.2017.01.004

Litchfield, A. T., & Althouse, J. (2014). *A systematic review of cloud computing, big data and databases on the cloud.* Paper presented at the Americas Conference on Information

Systems, Savannah. Retrieved from https://aisel.aisnet.org/amcis2014/ServiceSystems/GeneralPresentations/1/
Low, L. (2020). Data warehousing in the cloud. Retrieved from https://www.cio.com/article/3540416/data-warehousing-in-the-cloud.html
Maayan, G. D. (2018). The difference between a traditional data warehouse and a cloud data warehouse. Retrieved from https://www.dataversity.net/difference-traditional-data-warehouse-cloud-data-warehouse/#
McKendrick, J. (2019, April/May). Rethinking the future of data warehousing. *Database Trends and Applications*. Retrieved from https://www.actian.com/wp-content/uploads/2019/04/Rethinking-the-Future-of-Data-Warehousing.pdf
McKinsey Global Intitute. (2011). *Big data: The next frontier for innovation, competition and productivity.* Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.pdf
Oracle. (n.d.-a). Oracle modern data warehouse. Retrieved from https://www.oracle.com/autonomous-database/modern-data-warehouse/
Oracle. (n.d.-b). What is big data? Retrieved from https://www.oracle.com/big-data/what-is-big-data.html
Panoply. (2017). *The evolution of the data warehouse*. Retrieved from https://learn.panoply.io/the-evolution-of-the-data-warehouse
Ponniah, P. (2010). *Data warehousing fundamentals for IT professionals* (2nd ed.). Hoboken, N.J.: Wiley.
Press, G. (2013, May 28). A very short history of data science. *Forbes*. Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#36a5cce455cf
Productivity Commission. (2017). *Data availability and use: Inquiry report*. (Report No. 82). Retrieved from https://www.pc.gov.au/inquiries/completed/data-access/report/data-access.pdf
Ranjan, R., Rana, O., Nepal, S., Yousif, M., James, P., Wen, Z., . . . Dustdar, S. (2018). The next grand challenges: Integrating the Internet of Things and data science. *IEEE Cloud Computing, 5*(3), 12-26. doi:10.1109/MCC.2018.032591612
Rehman, K. U. U., Ahmad, U., & Mahmood, S. (2018). A comparative analysis of traditional and cloud data warehouse. *VAWKUM Transactions on Computer Sciences, 15*(1), 34-40. doi:10.21015/vtcs.v15i1.487
Robinson, M. (2020, January 3). The responsible stewardship of personal data is a key challenge for the 2020s. *Forbes*. Retrieved from https://www.forbes.com/sites/forbestechcouncil/2020/01/03/the-responsible-stewardship-of-personal-data-is-a-key-challenge-for-the-2020s/#2bfa56d47556
Sansu, G. (2012). Inmon or Kimball: Which approach is suitable for your data warehouse? Retrieved from https://www.computerweekly.com/tip/Inmon-or-Kimball-Which-approach-is-suitable-for-your-data-warehouse
Snowflake. (n.d.-a). *Beyond Hadoop: Modern cloud data warehousing*. Retrieved from https://resources.snowflake.com/data-lake/beyond-hadoop-modern-cloud-data-warehousing
Snowflake. (n.d.-b). *The data-driven enterprise done right*. Retrieved from https://www.snowflake.com/wp-content/uploads/2017/04/Snowflake-Data-Driven-Enterprise.pdf
Steier, S. (n.d.). To cloud or not to cloud: Where does your data warehouse belong? Retrieved from https://www.wired.com/insights/2013/05/to-cloud-or-not-to-cloud-where-does-your-data-warehouse-belong/

Taylor, R. D. (2020). “Data localization”: The internet in the balance. *Telecommunications Policy, 44*(8), 102003. doi:https://doi.org/10.1016/j.telpol.2020.102003

Thabet, N., & Soomro, T. R. (2015). Big data challenges. *Journal of Computer Engineering and Information Technology, 4*(3), 10. doi:http://dx.doi.org/10.4172/2324-9307.1000135

Transforming Data with Intelligence. (2016). *Why your next data warehouse should be in the cloud*. Retrieved from https://www.talend.com/resources/why-your-next-data-warehouse-should-be-in-the-cloud/?ty=content

Tripwire. (2018). *18 expert tips for effective and secure cloud migration.* Retrieved from https://www.tripwire.com/-/media/tripwiredotcom/files/book/18_expert_tips_for_effective_and_secure_cloud_migration.pdf

Vahie, V. (2016). 99 problems but the cloud ain't one: Cloud security. Retrieved from https://www.itproportal.com/2016/08/18/99-problems-but-the-cloud-aint-one-cloud-security/

Vaile, D., Kalinish, K., Fair, P., & Lawrence, A. (2013). *Data sovereignty and the cloud: A Board and Executive Officer's guide*. Retrieved from http://cyberlawcentre.austlii.edu.au/data_sovereignty/CLOUD_DataSovReport_Short.pdf