INFS 4020 – Big Data Concepts

Assignment 2 – Big Data Strategy Proposal

[Wangjun Shen]

[2022/11/02]

Executive Summary

Since the production of pizza can be standardized and there are no technical barriers, the quality of service will become the core factor of market competitiveness. Whether employees perform well and are willing to work long-term will determine customer satisfaction with the service. Based on Domino's Pizza's industry situation and the fact that it has a large amount of data, this report gives recommendations for using big data to assess, analyse and predict employee performance and long-term work intentions. A preliminary big data architecture is designed and provided, and the functions and technologies that can be implemented by the different components of the architecture are also analysed and demonstrated. The results of two fictitious visualizations are also provided to illustrate how management's decision-making is aided. Although the use of big data can bring commercial benefits, technical difficulties and policy constraints will also be challenges.

Introduction

Different catering brands will provide pizzas with different tastes and textures for differentiated competition, but if one pizza hits the market, other restaurant brands will be quick to imitate it to compete because there are no technical barriers to making pizza. Therefore, providing good service will become a key factor in whether to occupy more markets.

Domino's Pizza, a multinational pizza delivery chain, became the world's largest pizza seller by sales in February 2018 (“Domino’s Unseats Pizza Hut as Biggest Pizza Chain,” 2018). Huge commercial scale, numerous storefronts and numerous users all generate massive amounts of data. This report will provide technical suggestions related to big data, using existing data and suggested data collection, to improve Domino's Pizza's competitiveness in the pizza market. In addition to discussing the source of data and big data technology, it will also dialectically discuss the pros and cons of adopting relevant recommendations.

To understand this report, big data related expertise is not required, but a certain understanding of the business is required.

Key priority considerations

Pizza is not luxury or luxury food, and pizzerias are not like luxury hotels, so the core of providing high-quality service is the performance of the staff, not other factors. At the heart of using big data to make Domino's Pizza competitive lies in hiring high-performing employees who can work long-term.

Therefore, data about existing employees needs to be collected and mined, as well as analysed to obtain a final evaluation of the performance of existing employees, and through techniques such as statistical learning, a set of criteria is obtained to predict the performance of job candidates after entry.

Data sources

The main sources of data are as follows: employee database, official website data, third-party websites, business transaction systems.

When employees are hired, their personal information will be collected. Whether the employee is a full-time employee or a part-time employee, the employee's education level, and the employee's punch-in record (number of late arrivals, year of employment, total working hours and average monthly hours, frequency of taking leave, renting, or owning a house, etc.). These data can be considered as labelled data for making predictions about employee performance.

It is not difficult to collect the supervisor's evaluation of the employee and the mutual evaluation of the employees, and these evaluations will also be stored in the employee database.

After the input data (employee characteristics) is obtained, the output data (employee performance evaluation results) needs to be obtained.

Another important and more realistic data comes from consumers. If you google "domino's pizza", it's not hard to find the official website:

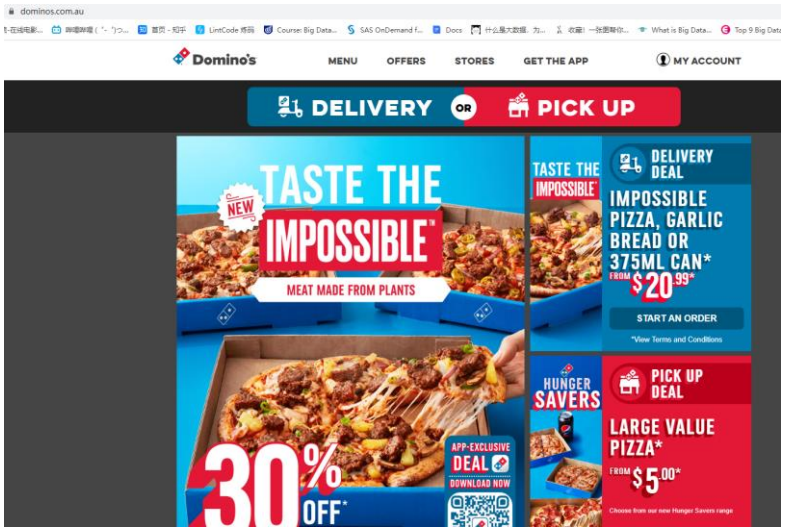

Figure 1 Screenshot of Domino's Pizza Australia's official website
by author

The site offers ordering and advertising, but if data on employee performance needs to be collected, a "rate employee" feature needs to be added.

Here is a simple design diagram provided by the author of this report:

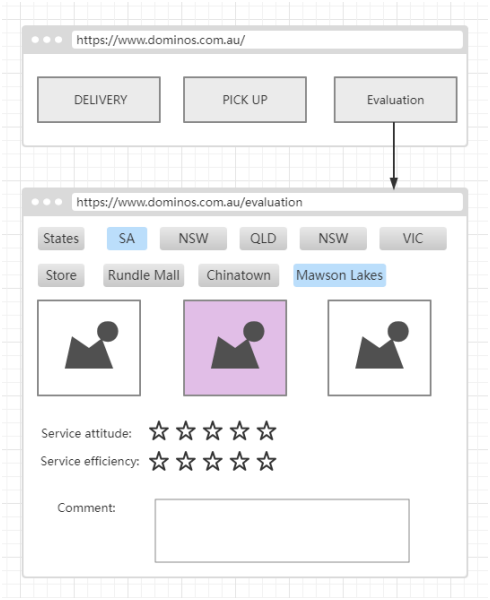

Figure 2 UI Design for New Function
by author

By providing the above-mentioned new functions or similar functions, user evaluations about employees can be well collected. In addition to the official website, the official mobile application also needs to provide similar functions so that users can evaluate employees.

Business transaction systems can also provide data for evaluating employee performance. Customers who bought pizza can rate the order with stars. If a customer uses takeout, his experience will be mainly affected by two factors: production and delivery. On the other hand, if it is picked up, then his experience is only affected by the production, and both are largely related to the

performance of employees and are rarely affected by other factors. From the perspective of 3v of data, the accuracy of these data is relatively high.

In addition to internal data and data generated by official platforms, data from third-party platforms can also be collected and considered.

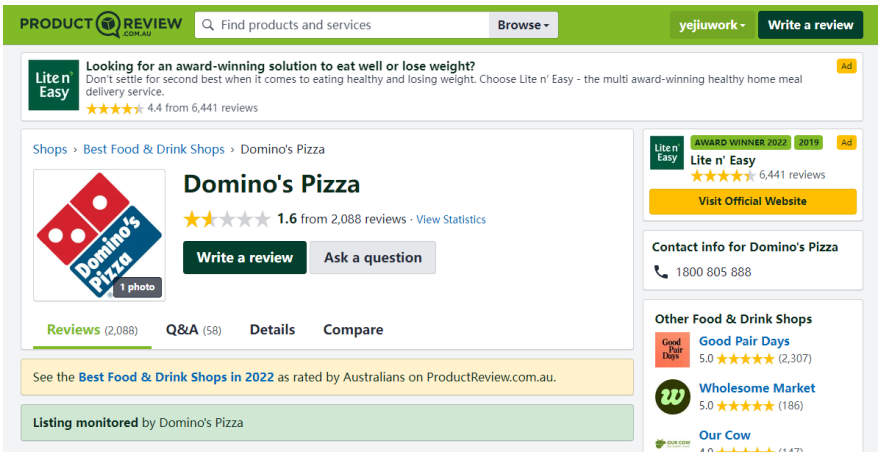

Figure 3 Screenshots from third-party platforms
by author

The above picture is a specific example, similar websites can be included. By using tools such as crawlers, data can be crawled from these third-party websites in batches for analysis.

Big Data architecture

The following figure shows the proposal for the design of the big data architecture.

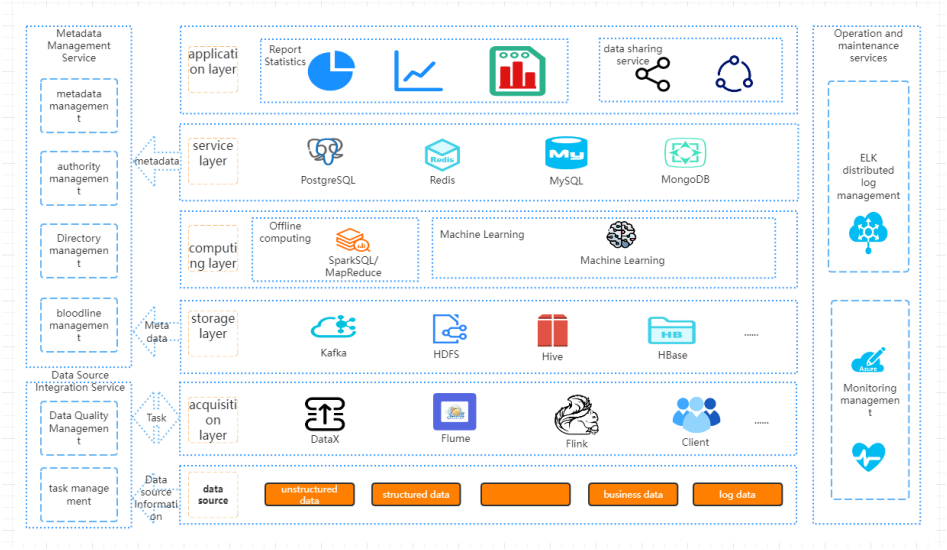

Figure 4 Big Data Architecture
by author

The task of data collection is to collect and store data from various data sources to the data storage, during which some simple cleaning may be done.

When an enterprise provides network services, website logs account for the largest share. Website logs are stored on multiple website log servers. Domino's Pizza recorded more than 85,000

structured and unstructured data per day through the group's point-of-sale systems as early as 2016, and today the amount of data collected daily is much higher than that (Marr, n.d.). Generally, flume agent is deployed on each website log server to collect website logs in real time and store them on HDFS. There are also various types of business databases, including MySQL, Oracle, SQL Server, etc. At this time, a tool that can synchronise various data to HDFS is required. Taobao's open source DataX is a good solution. DataX can automatically set up the appropriate data communication mechanism, which can convert the data stream more easily, thereby improving the efficiency of developers (Coviello, Rao, Sankaradas, & Chakradhar, 2022).

In terms of data storage, in addition to local storage, cloud storage also needs to be considered. HDFS is the most perfect data storage solution for data warehouse/data platform in big data environment.

For offline data analysis and calculation, that is, the part that does not require high real-time performance, Hive is the first choice. Hive provides rich data types and built-in functions and supports operations using SQL language. This makes Hive higher than MapReduce in statistical analysis of structured data.

Spark has been very popular in the past two years. After practice, its performance is indeed several times faster than MapReduce, and it is getting better and better in combination with Hive and Yarn (Shi et al., 2015). Therefore, it is necessary to support the use of Spark and Spark SQL for analysis and calculation. Because Hadoop Yarn already exists, it is actually very easy to use Spark without deploying a Spark cluster separately.

A core part of the big data framework is the use of machine learning to fit employee evaluation results and employee characteristic variables and use the fitted model to predict candidate performance and long-term job intentions. The functions of this part can also be implemented using Spark, which provides the open-source distributed machine learning library MLlib for calling. From the test results of some data scientists:

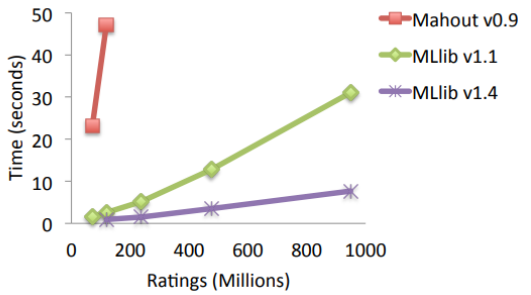

Figure 5 Benchmarking results for ALS
(Meng et al., 2016)

The machine learning capabilities provided by Spark have excellent performance when dealing with moderate data volumes.

Then there is the function of data sharing, which is where the results are stored after data analysis and calculation. The results of previous analysis and calculations through Hive, MR, Spark, and Spark SQL are still on HDFS, but managers cannot directly obtain data from HDFS to make decisions. Therefore, a place for data sharing is needed, and data is obtained from this place through the terminal application for further processing.

In the data warehouse/data platform, there are many kinds of programs and tasks, such as: data collection tasks, data synchronisation tasks, data analysis tasks, etc. In addition to timing scheduling, these tasks also have very complex task dependencies. For example, the data analysis task can only be started after the corresponding data collection task is completed, and the data synchronisation task can only be started after the data analysis task is completed. This requires a very complete task scheduling and monitoring system. As the centre of the data warehouse/data platform, it is responsible for scheduling and monitoring the allocation and operation of all tasks.

End-User Application and Visualisation

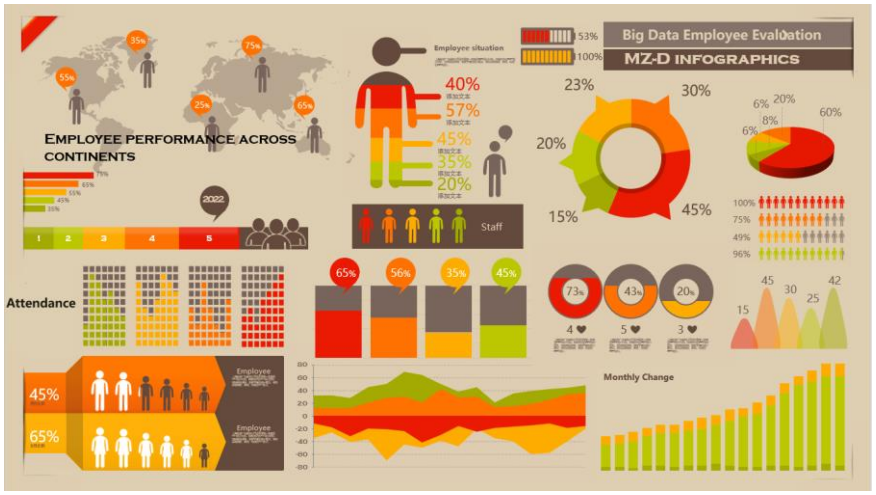

Figure 6 Employee performance in 2022
fictional by the author

The image above is an example of big data visualisation. This example gives a visualisation of the final assessment of employees across continents in 2022. Based on the results of this visualisation, management can know the distribution of employees with the same rating in different continents, what is the overall proportion of employees with a certain rating, and how the performance of employees changes in different months, etc. The results can assist management to make decisions related to layoffs or hiring in a certain time and a certain region.

Figure 7 Data visualization for employees
fictional by the author

The image above shows the visualisation results for a specific employee. This visualisation shows the employee's ability radar chart, the proportion of different levels of evaluations obtained, and the changes in different evaluations over time. Managers can use this data to determine how the employee's overall performance is and predict how the employee will perform next. Combining these data, the final judgement of the employee's performance and whether to stay or not.

Benefits and challenges

The benefits of using big data to assess and predict employee performance and long-term willingness to work are clear. Managers can use visualisation tools to intuitively understand and compare employee evaluation results, thereby eliminating bad employees and recruiting good ones. This reduces the trial costs of hiring employees, thereby increasing business profits.

There are some technical and non-technical challenges in using big data. The following figure shows the results of a survey on the challenges of big data:

Figure 8 Big Data Challenge
("Four Big Data Challenges," n.d.)

Among the challenges listed above, regarding the technical challenges, lack of employees with skills related to big data is the primary challenge for Domino's Pizza as an established pizza chain. The structure of big data is the second important challenge. Big data systems must be tailored to the specific needs of the organisation. This is a DIY business. IT teams and application developers are required to organise a set of tools from all available technologies. The third technical challenge is how to ensure the security of users' data. Domino's Pizza has many branches and users. The existence of a large amount of data will inevitably lead to hackers trying to steal this data for profit.

Fortunately, several of these challenges can be solved by using managed cloud services, but IT managers need to keep a close eye on cloud usage to ensure costs don't get out of hand. Additionally, migrating on-premises datasets and processing workloads to the cloud is often a complex process for organisations.

There are two non-technical questions. The first is about the accuracy of the data. From a raw 3V perspective, data accuracy refers to the degree of certainty of a data set. Uncertain raw data collected from sources such as third-party platforms and web pages can lead to serious data quality issues that can be difficult to pin down. Bad data leads to inaccurate analysis and can damage the value of business analytics as it can lead to executives' distrust of the data. The amount of indeterminate data in an organisation must be accounted for before it can be used in big data analytics applications. IT and analytics teams also need to ensure they have enough accurate data to produce valid results. The second non-technical challenge is policy. The European Union passed the General Data Protection Regulation (GDRP), which came into effect in May 2018, and the law puts the risk of illegal collection of some employee data [7]. Likewise, the use of certain characteristics to judge and select candidates may lead to discrimination and prejudice, which is also against the law in some countries.

Conclusion

For industries without technical barriers, such as pizza making, the performance and long-term retention of employees will greatly affect the quality of service customers enjoy. Domino's Pizza, an established pizza chain, has captured the largest share of the pizza market. Many storefronts, employees and customers will generate a large amount of data. Using big data tools to collect, analyse and visualise this data will help management to evaluate old employees and recruit new employees, thereby improving service quality and seizing more markets share.

This report provides a preliminary big data architecture, and analyses and discusses the functions of the different parts and the tools used, but whether it is feasible requires further consultation with experts in the field of big data. Two examples are provided to demonstrate the results of the data analysis and how the results can assist management staff.

Although there are benefits of using big data, which can reduce costs and improve competitiveness, there are also technical and non-technical challenges that need to be solved.

As a pizza chain with a long history, Domino's Pizza needs to embrace big data technology in the era of big data and find and solve difficulties in practice.

References

Coviello, G., Rao, K., Sankaradas, M., & Chakradhar, S. (2022). DataX: A System for Data eXchange and Transformation of Streams. *Intelligent Distributed Computing XIV*, 319–329. https://doi.org/10.1007/978-3-030-96627-0_29

Domino's Unseats Pizza Hut as Biggest Pizza Chain. (2018, February 20). Retrieved from adage.com website: https://adage.com/article/cmo-strategy/dominos-unseats-pizza-hut-biggest-pizza-chain/312463

Four Big Data Challenges. (n.d.). Retrieved November 4, 2022, from Transforming Data with Intelligence website: https://tdwi.org/blogs/tdwi-blog/2013/10/four-big-data-challenges.aspx

Marr, B. (n.d.). Big Data-Driven Decision-Making At Domino's Pizza. Retrieved November 4, 2022, from Forbes website: https://www.forbes.com/sites/bernardmarr/2016/04/06/big-data-driven-decision-making-at-dominos-pizza/?sh=4e73dca52b8e

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., … Talwalkar, A. (2016). MLlib: Machine Learning in Apache Spark. *Journal of Machine Learning Research*, *17*, 1–7. Retrieved from https://www.jmlr.org/papers/volume17/15-237/15-237.pdf

Shi, J., Qiu, Y., Minhas, U. F., Jiao, L., Wang, C., Reinwald, B., & Özcan, F. (2015). Clash of the titans. *Proceedings of the VLDB Endowment*, *8*(13), 2110–2121. https://doi.org/10.14778/2831360.2831365