Data and the Community: How a Data Science Ecosystem Impacts Society

Nicole Aguilera
Bellevue University
naguilera@my365.bellevue.edu

Osmond Oke
Bellevue University
ocoke@my365.bellevue.edu

Author Keywords
Data Science; Data Science Ecosystem; Society, Community; Data; Structure, Infrastructure

Abstract
A data science ecosystem is the structural backbone of all data science processes in an industry or organization. It is a collection of infrastructure, analytics, and applications used to capture and analyze data [3]. The data collected is used to provide companies and organizations with the data they require to better understand the behaviors of their customers or their own company tactics [1], in order to make necessary adjustments to marketing techniques, pricing modules, and even day to day operations. Using the term "ecosystem'' rather than the traditional "environment" to describe this domain, is primarily due to its adaptive characteristic; it is constantly evolving as newer technologies and technological processes arise. The goal of this study is to elaborate on the necessary steps and procedures that are taken into consideration when creating a suitable data science ecosystem for an organization, as well as how the selected combinations affect the society, or community, they are present in. This understanding creates room for a more informed decision when assembling data science ecosystems moving forward. Each corporation takes this ecosystem and adapts it to their own business so that it works well for them. Our research paper aims to highlight the necessary factors in constructing an ecosystem in this fashion, aid the understanding of the processes involved, and how

these processes and decisions are able to impact society.

Introduction

As stated before, a data science ecosystem is an infrastructure that connects the analytics and applications that are used to capture and analyze data within a company. It is made up of data sources, which generate data, data storage that stores and processes the data, and applications which are eventually shared with both internal and external consumers of the company. Data science ecosystems are a great way for corporations to organize the collecting and scrutinizing of data. Both traditional sources, for example, databases within the company on consumer purchases, and big-data sources, like blogs, websites, and social media sites, are involved in discovering new information. These points would then get transferred to be stored in any data warehouse or program, like Python or Hadoop, to be processed and cleaned. After the data is fully cleaned with all the appropriate variables and topics intact, it then moves on to the final phase: applications. Applications vary by company, but they can be any type of business analysis, packaged applications, or custom application that have been chosen as the means to carry out information to their appropriate customers. Applications can send data back to data sources if needed, as well.

When creating a data science ecosystem, a company needs to look at what software systems that they are already using in order to easily integrate it with normal operations. Data sources can come from pretty much anywhere, from popular data set sites that already have the collected data all on one file, to the establishment’s own collection system in which they have from past sales, reviews, or anything else the corporation aims to analyze. They can be structured or unstructured [15], but it is necessary to denote which your set is coming from in order to perform the proper cleaning techniques. Once establishing this, it is crucial to make sure the data storage program is able to support that format that the collected information is going to; for example, if a certain software does not support any Microsoft Excel files, it is going to be a problem if the data is organized in a Microsoft Excel file. In the data storage stage, different companies, yet again, their own unique ways of sorting data depending on what they want to achieve through the outcome of their analyses. It is also essential to consider how data will flow through the data ecosystem when contemplating the implementation of one in a business. Some algorithms take up too much time being transferred to the data when executing a particular analysis on data when going from the phase of processing in data storage to the final stage of analyzing it. Sometimes, depending on the situation, it is better to move the algorithm to the data by using machine learning algorithms [19]. This allows for faster performance and easier production deployment from the program itself.

Administration of Data Science Ecosystems

The steps that go into creating a Data Science Ecosystem are essentially selected to encompass the three main parts of a data scientist's work functions: gathering the data, wrangling the data, analyzing the data. While general misconceptions have led people to believe that a Data Scientist and a Statistician are quite similar, their responsibilities can easily be set apart; while a statistician takes data and runs a regression, a

Data Scientist gathers the data, runs the regression, communicate the results, reveal patterns and relations, and implement real change. This general overseeing of the Data Science process essentially makes a Data Scientist, and the Data Science Ecosystem showcases the tools used by the data Scientist, how they work, and more importantly, how they work together. The main factors of a Data Science Ecosystem include:

- The Data Source - Without the data, there is essentially no Ecosystem. The sources of this data can be grouped into three categories:
  - Databases - These are structured data houses held within a computer system. Some examples usually seen are Oracle RDBMS, MySQL, and even Microsoft Access.
  - Applications - Over the last decade, industries have gravitated towards cloud storage. This now common storage method allows businesses to store their data on remote storage systems, where it can be accessed, maintained and backed up by users over a network.
  - Third-party Data - This is data that is obtained by entities without direct relationships with the individuals or groups the data is collected on. An example of third-party data sources can be seen in surveys and questionnaires.
- Open Source Tools - These are software programs designed to carry out specific tasks, and in which their source code is published openly for use, or even modification from their original designs, for no extra charge.

These factors work hand in hand, as one factor is somewhat useless without the existence of the other.

Data Science Ecosystems in Society

Data is vital to any company for improvement; they have to analyze their own data in order to find out what their personal weak spots are and how it can be refined to boost productivity, sales or whatever outcomes they create in the future. Each data science ecosystem can vary depending on the company. One of the papers makes an example of having all of the different software systems through a visual network analysis [1]. This sorts all of the programs into modules, so they are able to see, visually, how the various systems are connected to one another. There is a belief that these ecosystems can potentially create a Digital Transformation and Sustainability model between data actors, the capability of big data analytics that cause the change through value, business and society [2], however, it had yet to be seen as prominent in research when this idea was constructed in 2018. On the industrial side, the ecosystemic approach that they have intertwines data science, IT, and business aspects [14], as seen below in Figure 1. This incorporates the technical side of data science and IT, with the more organizational side of business in order to create the flow of a successful corporation, from Kegl’s point of view.

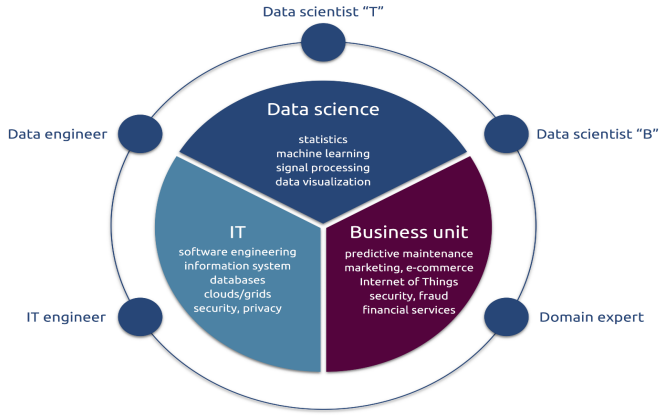

Figure 1: The business industry has many ways of organizing its ecosystems. This is just one example of how data science, IT, and business affairs coexist within a company together.

Conclusions

In conclusion, the importance of the Data Science Ecosystem is unrivaled, as it is essentially the backbone of the Data Science process. From the factors outlined in our Administration section, to the graphical illustration, Figure 1, it is clear that the elements of a Data Science Ecosystem rely on each other for smooth and proper function and operations. Now can the Data Science process be successful without an Ecosystem in place? The answer is no; once the data sources and tools come together, even in more informal conditions, they essentially constitute the Data Science Ecosystem that ensures the free flow of the Data Science process through the various levels. It is also very important to note that an organization can have multiple approaches towards building their Data Science Ecosystem, be it a more structured approach with various technologies already pieced together, or even "a la carte" where they pick and choose what works for them.

Acknowledgements

We would like to thank our professor, Shankar Parajulee, who has led us to this point of furthering our education in data science, and to our other professors in the department for guiding us while pursuing a career in the field. We would also like to thank Bellevue University for helping us to get there. Nicole would like to personally thank her father, Pete, for introducing her to the technical field which led her to this career path, and Osmond would like to thank Dr Ganapathiraman Raman for opening his eyes to the wonders of Information Technology and Data Science in his time as an undergraduate at The University of Texas at Arlington.

References

1. Majumdar, Dr. Archisman. "Data Science Ecosystem." Data Science Ecosystem, 24 Nov. 2017, pp. 1-5.
2. Pappas, Ilias O., et al. "Big Data and Business Analytics Ecosystems: Paving the Way towards Digital Transformation and Sustainable Societies." Big Data and Business Analytics Ecosystems: Paving the Way towards Digital Transformation and Sustainable Societies, 26 Oct. 2018, pp. 480-491.
3. Liu, Dr. Alex. "A New Ecosystem Approach to Improve Data Science Success." A New Ecosystem Approach to Improve Data Science Success, 28 Jan. 2019, pp. 1-11.
4. Evans, Barbara J, and Harlan M Krumholz. "People-Powered Data Collaboratives: Fueling Data Science with the Health-Related Experiences of Individuals." People-Powered Data Collaboratives: Fueling Data Science with the Health-Related Experiences of Individuals, 20 Dec. 2018, pp. 159-161.
5. Hoyt, Robert, and Victoria Wangia-Anderson. "An Overview of Two Open Interactive Computing Environments Useful for Data Science Education." An Overview of Two Open Interactive Computing Environments Useful for Data Science Education, 1 Oct. 2018, pp. 159-165.
6. "Innovation in Aging." Innovation in Aging, 13 Apr. 2020, pp. 1-1., academic.c.oup/innovateage/article-abstract/3/Supplement_1/S480/5617582/.
7. Federer, Lisa, et al. "The Medical Library Association Data Services Competency: a Framework for Data Science and Open Science Skills Development." The Medical Library Association Data Services Competency: a Framework for Data Science and Open Science Skills Development, pp. 304-309., doi:dx.doi.org/10.5195/jmla.2020.909.
8. Berman, Francine, et al. "Realizing the Potential of Data Science." Communications of the ACM, vol. 61, no. 4, 2018, pp. 67-72., doi:10.1145/3188721.
9. Kalidindi, Surya R., et al. "Vision for Data and Informatics in the Future Materials Innovation Ecosystem." Jom, vol. 68, no. 8, 8 Nov. 2016, pp. 2126-2137., doi:10.1007/s11837-016-2036-5.
10. C√°rdenas-Navia, Isabel, and Brian K. Fitzgerald. "THE BROAD APPLICATION OF DATA SCIENCE AND ANALYTICS." THE BROAD APPLICATION OF DATA SCIENCE AND ANALYTICS, 2015, pp. 25-32.
11. Maksimenkova, Olga, et al. "Using Data Expedition as a Formative Assessment Tool in Data Science Education: Reasoning, Justification, and Evaluation." International Journal of Emerging Technologies in Learning (IJET), vol. 14, no. 11, 2019, pp. 107-122., doi:10.3991/ijet.v14i11.10202.
12. Chen, Cathy Yi-Hsuan, and Wolfgang Karl H√§rdle. "Data Science and Digital Society." Proceedings of the International Conference on Business Excellence, vol. 11, no. 1, 2017, pp. 669-675., doi:10.1515/picbe-2017-0071.
13. Fattah, Ahmed. "Going Beyond Data Science Toward an Analytics Ecosystem: Part 2." IBM Big Data & Analytics Hub, IBM, 14 Mar. 2104, www.ibmbigdatahub.com/blog/going-beyond-data-science-toward-analytics-ecosystem-part-2.
14. Kegl, Balazs. "The Data Science Ecosystem: Industrial Edition." Towards Data Science, Medium, 22 May 2017, towardsdatascience.com/the-data-science-ecosystem-industrial-edition-938582427466.
15. Biewald, Lukas. "The Data Science Ecosystem." Computerworld, Computerworld, 24 Mar. 2015, www.computerworld.com/article/2899647/the-data-science-ecosystem.html.
16. Scott, Jim. "Accelerating the Data Science Ecosystem." Database Trends and Applications,

BDQ, 16 Dec. 2019, www.dbta.com/BigDataQuarterly/Articles/Accelerating-the-Data-Science-Ecosystem-135500.aspx.

17. Meng, Xiao-Li. "Data Science: An Artificial Ecosystem ¬∑ Harvard Data Science Review." Harvard Data Science Review, HDSR, 28 June 2019, hdsr.mitpress.mit.edu/pub/jhy4g6eg.
18. Kegl, Balazs. "The Data Science Ecosystem." The Data Science Ecosystem, Paris-Saclay Center for Data Science, www.tmci.mines-paristech.fr/wp-content/uploads/2015/02/Data-science-a-revolution-in-progress.pdf.
19. Kelleher, John D., and Brendan Tierney. Data Science. The MIT Press, 2018.