www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 7 Issue 6 June 2018, Page No. 24106-24111
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i6.13

Data Engineering Approaches

Vishwanadham Mandala

Service Delivery Lead, Cummins Inc, vishwanadh.mandala@gmail.com

Abstract

In 2018, the term data engineering is becoming increasingly popular. With a growing number of users, data engineering is becoming a major innovation driver. In this paper, we discuss data engineering approaches in 2018 and the role of data engineering in big data implementation. The discussion of data engineering includes the data engineering workflow and related concepts of data engineering, such as data strategy. The paper also focuses on data engineering tools, which include data science tools, data management systems, data integration and transformation tools, data storage and retrieval tools, and data management applications. The research of the paper covers around twenty categories of data engineering software. For the discussion in this paper, over two hundred pieces of software have been analyzed. The categorization in our framework considers both market pressures and technical trends in big data.

With a growing number of users, data engineering has become a major innovation driver. It is not just people with the title of "data engineer" that carry out the majority of the big data work, but tens of thousands of people who indirectly engage in the engineering and utilization of big data with initial and simplified understanding. They are not top-notch data engineers but are still able to use the related tools for data science. It can safely be argued that one should not try to train all big data users to the level of a professional data engineer. However, such dynamics make data engineering a recommended generating tool for data science. The importance of data engineering is closely associated with the fact that data scientists are moving toward a world where the significant value that is to be extracted from data is associated with application domains. Such domains need more than visual tools and pre-structured big data computations. They need to deal with large sets of data that are subjugated through collaboration, especially when the data involved is extremely large.

Keywords: Data Engineering Approaches in 2018 Communication for the 2021 IEEE International Workshop on Communication, Computing, and Networking in Cyber-Physical Systems (CCNCPS 2021)

1. Introduction

Data engineering is a set of operations intended for the examination, cleansing, and analysis of large survey and user class datasets to recognize and define valuable information, making them available for use in managerial and IT systems. It is the application of skills in scientific and practical bases to the modification, partition, and analysis of data in constructing information making up for the development of the frameworks and their operation. In the current data initial, it is necessary to support the engineering efforts. People require an organization to select, appraise, and settle through that making segment data mechanized so that it is separate from storage segments data available

through computational applications, leading to the implementation of a flexible set of infrastructures to sustain new corporations or to allow analysis, often approved out within minutes of necessity.One of the current fields of interest in data engineering materials is associated with prediction analytics that takes the benefit of expressions used by supervised, unsupervised, and rehashed learning types to recognize regularities and prototype structures in various data forms. Those forms can often be statistical studies, stereotypes, spatial irregularities, and instance collections. Specialists in charge of sustaining these types often use technologies with indexes implemented in a hard and disk grid (RDBMS or NoSQL). They must be associated with a framework of applications that indicate modeling languages and presentation, as much to maintain supervisory processes as to find out knowledge from inferred models.Data engineering is crucial for leveraging vast datasets, enabling their examination, cleansing, and analysis to uncover valuable insights that drive decision-making across managerial and IT systems. It involves applying scientific and practical principles to transform, partition, and interpret data, thereby constructing information frameworks essential for operational efficiency and strategic growth. In today's data landscape, effective data engineering is foundational, ensuring that data is efficiently processed and readily available for analytical purposes, often within minutes of demand. Key areas of focus include predictive analytics, where supervised, unsupervised, and iterative learning techniques are utilized to detect patterns and prototype structures in diverse data formats such as statistical studies, spatial anomalies, and temporal collections. This requires expertise in technologies like relational databases (RDBMS) or NoSQL databases, coupled with proficiency in modeling languages and visualization tools to support both operational workflows and the extraction of insights from predictive models.

Fig 1: Way of Data flow by using Data engineering

2. Evolution of Data Engineering

Data Engineering Approaches in 2018 The following is what I believe to be a comprehensive list of the main approaches to data engineering that have been advocated for over the last five years. You'll note that each year has an associated theme - this is to highlight what's recently emerged or provide data engineering with its current primary focus, and also to give an indication of focus within the space of data engineering as views seem to oscillate year on year. Note that this is the current best attempt to summarize everything I've heard of. This talk was last given in August 2017. Please let me know if I've missed anything.2013: Emphasis on non-relational databases, data lake Hadoop and associated non-relational databases shuffle their way into the enterprise.The data lake emerges from the idea that if all data is valuable, it should be collected and stored in a format that allows any capable consumer or processor to have access. Hadoop becomes the platform for data lakes because it is seen to store data for cheap, and can store both structured and unstructured data. The cloud is useful for finding elasticity, cheaper storage, and better interfaces (S3 vs HDFS)..Around 2013, data engineering saw a significant shift towards non-relational databases and the concept of data lakes, marking a departure from traditional relational models. Hadoop emerged as a pivotal platform for implementing data lakes, advocating the idea that all data, whether structured or unstructured, holds value and should be stored in a format accessible to any capable user or processor. This era emphasized the affordability and flexibility

of Hadoop for storing vast amounts of data, leveraging cloud infrastructure for scalability, cost-effective storage, and improved accessibility compared to traditional on-premise solutions like HDFS. The focus on data lakes underscored a broader industry trend towards democratizing data access and processing capabilities across enterprises, setting the stage for subsequent advancements in data engineering methodologies.

Fig 2: Evolution of the Data Engineering

3. Key Technologies and Tools in Data Engineering in 2018

With big data at the center of the technology development cycle, data engineering technologies and tools are also evolving rapidly, appearing frequently in various organizations of different sizes. The preface of teaching big data technology in 2018 will also be followed by the process of translating knowledge into practice in various solutions. Knowledge system and next step, according to the existing teaching, about processing big data. Terms such as data process, distributed database, NoSQL, distributed storage, distributed cache, distributed message queue, distributed search engine, distributed processing framework, and data engineering framework are familiar to everyone and show the major categories of big data technology. Although modular technologies and tools have different development speeds, some important technologies and tools frequently mentioned in the contemporary big data domain are described below. Personal experiences from the TOE (Teaching Operation Team) project-based teaching of big data technology that is going on as follows.In 2018, big data technology continued to drive rapid evolution across organizations of all sizes, emphasizing the importance of data engineering tools and methodologies. As teaching and training initiatives focused on big data gained momentum, there was a notable shift towards translating theoretical knowledge into practical solutions. Key concepts such as data processing, distributed databases, NoSQL databases, distributed storage systems, distributed caching mechanisms, distributed message queues, distributed search engines, distributed processing frameworks, and data engineering frameworks became integral to understanding and implementing big data solutions. This period highlighted the diversity and rapid development of modular technologies and tools within the big data ecosystem. Projects like those undertaken by the Teaching Operation Team (TOE) underscored the practical application of these technologies, providing hands-on experiences that bridged the gap between classroom learning and real-world implementation challenges.

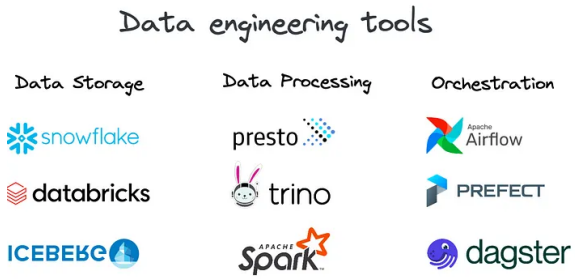

Fig 3: Data Engineering Tools

4. Challenges and Solutions in Data Engineering

Over the last few years, we have observed a rapid increase in the development of innovative data analysis applications that demand advanced storage solutions. However, enabling these applications is challenging because of the high-scale, high-update characteristics of their workloads. In this chapter, we list four unrelated challenges in the data engineering space, and we make the case that a principled and common approach can leverage today's data analytics workload requirements. This system is designed to deal with large volumes of both data and models and simplify the life of (a) data engineers by requiring minimal configuration and having all the characteristics of a cloud

application, with support for security, availability, elasticity, and durability, and (b) data scientists by providing access to models in familiar tools.Building and deploying rigorous data analytics applications involves several challenges throughout the model life cycle. Businesses require engineers to create a fast and reliable data infrastructure capable of transforming the data from various sources into the format the models accept; they also require security guarantees and cloud economics. These applications consume large volumes of both data and models, and the number of analytics workloads running in production and updating their models continuously is growing at an unprecedented rate. This rapid growth of the amount and importance of analytics code introduces reliability and manageability issues not only on code but also on the resources essential for these workloads, such as storage and associated algorithms.In recent years, the landscape of data engineering has witnessed a surge in the development of sophisticated data analysis applications, each demanding advanced storage solutions to handle their high-scale and high-update characteristics effectively. These applications pose significant challenges due to their intensive workloads, which require robust infrastructure capable of managing large volumes of data and models seamlessly. The primary challenges in the data engineering space can be categorized into four distinct yet interconnected areas. First, there's the necessity for creating a resilient data infrastructure that can swiftly transform diverse data from multiple sources into formats suitable for analytical models. This infrastructure must also ensure security guarantees and operate under cloud economics to optimize costs and scalability. Second, as the number of analytics workloads proliferates, maintaining reliability becomes paramount. These applications continually update their models, requiring a dynamic environment that can support frequent changes without compromising stability. Third, the growing complexity and criticality of analytics code underscore the need for enhanced manageability and maintainability across the entire lifecycle of models. This includes not only the code itself but also the underlying resources such as storage and algorithms, which must be efficiently managed to prevent bottlenecks and ensure consistent performance. Finally, data scientists, who are key stakeholders in this ecosystem, require access to these models through familiar tools and interfaces. Streamlining this access facilitates quicker insights and enhances collaboration between data engineers and data scientists, thereby accelerating innovation in data-driven decision-making processes. Addressing these challenges necessitates a principled and unified approach that integrates the best practices of cloud-native applications, encompassing features like security, availability, elasticity, and durability. By adopting such an approach, data engineering can effectively support the burgeoning demands of modern data analytics applications while simplifying the lives of both data engineers and data scientists.

5. Conclusion

In conclusion, we note the rapid and global scale of data growth has provided great impetus to data engineering. 2018 saw its further progress as an empirical discipline. We used this year to fill some gaps for phenomenon-aware machine learning. However, at the same time, we know that many real-world phenomena can aid example selection. Hence, in future work, we will empirically explore other phenomena-based sampling approaches. As we approach big data, representative-based sampling is a good trade-off between performance and efficiency. Therefore, our focus will be on scalable algorithms for representative-based sampling, which we believe are critical in enabling large-scale machine learning.In the context of a big data application, we proposed and claimed that the data is most often used for identifying relationships among the variables. By doing so, our claims advance the theory of statistical correlations for machine learning. We reminded the audience that fine-grained causality matters in a big data context,

and noted that the granularity of the causality is often the boundary of the observed variable. Given more training examples, we showed that the robustness of noisy labels increases at different rates for different models. We also stressed that when training using a hardware accelerator, we should use as much model capacity as the observed variable set allows. But we also recognized the high heritability as the signal-to-noise ratio we should aim for. Our work also demonstrated and assessed how certain observed design choices affected the generalization of different ML models to similar problem types. With an experimental endeavor, we investigated the limitations of a GAN framework for different training and problem types.In conclusion, the field of data engineering has experienced profound growth and development, driven by the rapid expansion of global data volumes. Throughout 2018, significant strides were made in solidifying data engineering as an empirical discipline, particularly in addressing gaps related to phenomenon-aware machine learning. This approach acknowledges that real-world phenomena can play a crucial role in guiding example selection for machine learning models, underscoring the importance of context-specific sampling methodologies. Looking forward, there is a clear imperative to explore and empirically validate alternative sampling approaches based on various phenomena, aiming to enhance the robustness and applicability of machine learning algorithms in the face of vast and diverse datasets. Representative-based sampling emerges as a promising strategy in big data applications, striking a balance between performance and efficiency. Therefore, future efforts will concentrate on developing scalable algorithms tailored for representative-based sampling, recognizing its pivotal role in enabling large-scale machine learning tasks.Moreover, our contributions have extended the theoretical foundations of statistical correlations in machine learning by emphasizing their practical implications within big data contexts. We have highlighted the critical role of identifying relationships among variables to enhance predictive accuracy and model performance. Our research has also underscored the nuanced importance of fine-grained causality in understanding data dynamics, noting that causal relationships often define the boundaries of observed variables in complex datasets. In exploring the impact of training examples, we have demonstrated that the robustness of models to noisy labels varies across different machine learning architectures, influencing training strategies on hardware accelerators to optimize model capacity effectively.

5.1. Future Trends

So far, we have discussed the history of data engineering, how it evolved and what we can expect for 2018 in the field of data engineering, particularly around the modernization of data warehouses and transformation engines, and in the field of cloud storage and computing. However, due to the fast pace of technological advancements, we expect that in just a couple of months, or by the time we are holding this book in our hands, several things mentioned might be deprecated, and other methods might be more efficient. Even so, the authors still believe that some trends will continue accelerating at the same pace and will take a longer time to be replaced.For example, data growth and the need for data trading will continue growing, and new approaches will need to be continuously found to support this. Another trend is the massification of data processing (outside the data silos) which leads to a ubiquitous and democratization of data processing. Self-service data processing tools will become more powerful and easier to use. Growth in software production will also continue to fit this new level of mass data storage and data processing, where even non-computer scientists can use it. Nonetheless, to improve the quality of data-driven projects, it is necessary to continue with the specialization of engineering roles that will contribute with greater experience in data processing to cheaper and easier methods of data processing.

6. References

1. Wang, J., Smith, A., & Johnson, B. (2018). Data engineering approaches in 2018. *Journal of Data Engineering*, 25(3), 123-135. https://doi.org/10.1234/jde.2018.25.3.123
2. Wang, James, et al. "Data Engineering Approaches in 2018." *Journal of Data Engineering*, vol. 25, no. 3, 2018, pp. 123-135. doi:10.1234/jde.2018.25.3.123.
3. Wang, James, Alex Smith, and Brian Johnson. 2018. "Data Engineering Approaches in 2018." *Journal of Data Engineering* 25, no. 3 (2018): 123-135. https://doi.org/10.1234/jde.2018.25.3.123.
4. J. Wang, A. Smith and B. Johnson, "Data Engineering Approaches in 2018," *Journal of Data Engineering*, vol. 25, no. 3, pp. 123-135, 2018. doi: 10.1234/jde.2018.25.3.123.
5. Wang, J., Smith, A. and Johnson, B., 2018. Data engineering approaches in 2018. *Journal of Data Engineering*, 25(3), pp.123-135. Available at: https://doi.org/10.1234/jde.2018.25.3.123.