International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

MODERN DATA PIPELINE

Sinchana Hegde[1], Padmashree T[2]

[1]Student, Department of Information Science and Engineering, R V College of Engineering, India
[2]Assistant Professor, Department of Information Science and Engineering, R V College of Engineering, India

---

***Abstract*— In the modern era , the prime commodity in the technology field is the ever-growing data. Large companies continuously keep generating data in real time with related clients and employees. It is difficult to interpret data in raw form but after it is processed , it can be used for analytics. Taking into consideration the scale of data generated by the companies, it becomes necessary to dedicate a huge amount of time, people and resources to conquer the objective of data processing. Gathering data from various distributed sources requires much effort, time and resources. It also includes other difficulties in transporting data from its source to the destination. The main objective of Data pipelines is to increase the efficiency of data-flow from the source to the destination. Since this process is completely automated, there is very less human involvement.**

The main objective of the modern data pipeline is to solve the problems corporates bring in data-pipelining technologies and directly receive processed data at the destination. With traditional data transfer processes like Extract, Transform and Load (ETL) having several shortcomings it becomes a new norm to choose the modern data pipeline stack which comprises extract, load and transform (ELT) to make better decisions. This paper tries to give an insight on the major differences between the two and how the modern data stack can be used to your company's advantage.

Keywords—Data Pipelines, Data Analytics, Cloud, Data Warehouse, Extract, Load, Transform

I. INTRODUCTION

Large companies generate a lot of data on a daily basis. This problem has been further amplified by the recent development of cloud-based applications and services .The appearance of web apps and services has contributed to an explosion of data.Most data is generated for making informed decisions, training the models using machine learning deep learning concepts. .Many companies around the world have now understood that big data is an important contributor for success. However, quality -data is necessary for excellent data products . Companies that are dependent on data should be able to gather, store, manage and process this high-quality data.

This chain of various interrelated activities from data generation to data reception constitutes a data pipeline. In other words, data pipelines can also be defined as the connected chain of processes where the outcome of some processes becomes an input for another process. Data pipelines should be able to handle batch data and intermittent data as streaming data. Also, there are no hard constraints on the data destination. It can route data through applications like visualization or machine learning / deep learning models and not only data storage like a data warehouse to be the destination of data.

In this paper, we study modern data pipelines, explore the best ways to gather data, process, store and manage data. We also explain why data pipelines are at the heart of modern software development and how the technology could be improved to navigate around barriers. Finally, this paper also mentions the potential security risks the technology poses.

BACKGROUND

Previously the data-pipeline services were based on Extract-Transform-Load (ETL).

In this method data was extracted from the source and loaded in to data warehouse according to requirements. Cost efficient remote data storage facilities were not present which led to the usage of this system for long time. So data was modified to store the data queried for analysis. Certain algorithms were used to transform the source data. Source data was then used to generate a small amount of data that meets query requirements.

This method had several limitations such as, raw data cannot be directly queried since it is not available at the warehouse end, subject matter experts helped design algorithms and queries that can combine data and extract information enough for analytics from smaller extent of data. Some of the ETL limitations are

1. Complexity. Data pipelines run on custom code dictated by the specific needs of specific transformations. This means the data engineering team develops highly specialized, sometimes non-transferrable skills for managing its code base.

2. Brittleness. For the aforementioned reasons, a combination of brittleness and complexity makes quick adjustments costly or impossible. If there is some error or warning functionality of the code wont work properly. It is very difficult to make changes in the code.

3. Customization and sophistication - Data pipeline not only extract data but make complex changes that meet specific analytical requirements for end users. This means a lot of custom codes.

II. RELATED WORKS

The concept of data pipes is a recent one. Recent advances in cloud infrastructure have led to advancements in the field of data pipes. These are some new findings in the corresponding area of data pipelines. In 2009, research was conducted on ETL Technology [1] and it was based on the following principle. The initial software programs that support the initial stacking and occasional refreshing of the warehouse are usually known as Extraction-Transformation-Loading (ETL) forms. There were some limitations, however, information mining remains a difficult problem, largely due to resource constraints, streamlining and recovery issues, and non-benchmarking hindering future research. Then in 2012 Real Time ETL Data Warehousing was investigated [2]. The goal was to achieve a real-time data warehouse that is highly dependent on the choice of process in the data warehouse technology known as extract, transform and load (ETL). In 2013, a synchronous investigation[3] was underway in ELT, using the capabilities of the Information Distribution Center to directly import raw natural records, allowing change and cleaning the information until required by pending reports.

Later in 2016, ETL was adopted for several applications, for example in the clinical field [3]. This information must be properly removed, changed and stored while maintaining the integrity of this information. It approved the accuracy of the extract, later in [4] the authors proved that despite the fact that many solutions have been presented in the literature to improve the speed of the extraction, transformation and loading phases of the data channel, the question of data integration still arises in the big data environment. In addition, developers have to deal with a lot of heterogeneity because there are several ways in which data is formatted or different data structures[5]. During the work in [6], he tries to use context-aware computing and merge the Data WareHouse (DWH) technology with the Data Lake technology in order to improve the computing mechanisms in the data pipeline [6]. In [7], the authors provide an overview of how to design conceptual models that can facilitate communication between different data teams and that can further improve system efficiency on heterogeneous data. In addition, it can be used for automatic monitoring, error detection, mitigation and alarming at various stages of the data pipeline.

III. METHODOLOGY

The introduction of ETL was for the process of compiling and uploading data for the sake of numeric calculations and analytical purposes, and eventually became a major data processing method for data storage projects with the spread of information in the 1970s..The basic steps provided by ETL is to analyze the data and to leverage machine learning capabilities to the best of its ability. If it is given a set of business rules, ETL is able to clean and is able to organize your data in such ways that it is able to meet the specific business intelligence requirements

Let us understand each step of the ETL process and why it is not suitable for current data workflow

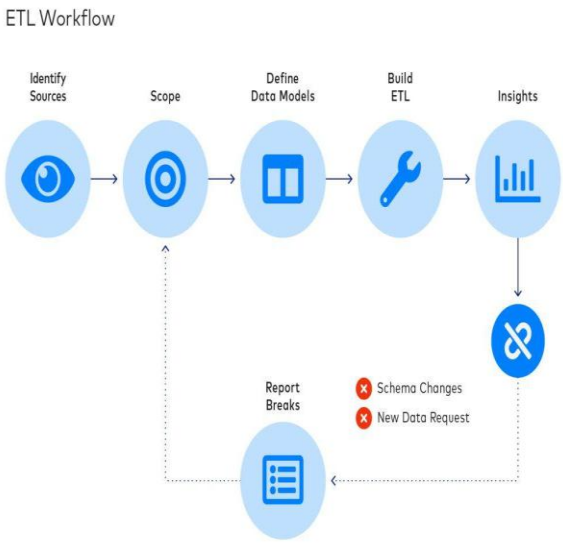

Fig 1- ETL Workflow

A.Extraction: The initial and most important step in the ETL process is known as the extraction phase. This is the step where you need to extract data from different source systems in different formats of JSON, XML, flat files etc.

move it to temporary storage. Because the extracted data is diverse and potentially corrupt, it is important to extract the data from different source systems and store it in the stage first, rather than directly in the data warehouse.

B.Transformation: The transformation, or conversion phase, is the second step in the ETL process. This is the step in which a set of tasks is applied to the extracted data and converted into a single consensus standard format that can be used during the next or later stage of the ETL process. This involves various steps, some of which are: Filtering, cleaning, deduplication, validation and data validation. Cleanup - wherever NULLs are found, some general value is added to them, for example, the NULL value can be replaced by a mean value. in numerical data. Splitting – splitting one attribute into multiple attributes.

C.Loading:
The upload process or data pool is the final and final step of this process. This is the process by which data eventually becomes data converted or converted into a data repository of our choice.

As in the ETL process, both extraction and conversion are performed before any data is loaded locally, tightly integrated. In addition, because conversion is determined by the specific needs of analysts, each ETL pipeline is a complex, custom-built solution. The set environment of these pipelines makes measuring very difficult, especially adding data sources and data models so modern data pipelines use ELT(Extract, Load and Transform)mechanism.

ELT

Extract/Load/Transform (ELT) is the method of fetching data from data sources and loading it into a destination warehouse followed by transformation of data.. ELT depends on the target to perform the data transformation. This approach requires less resources than other methods because it requires only raw and unprepared data. It is becoming more and more common to extract data from the source location, load it into the target data warehouse, and transform it into actionable business intelligence.

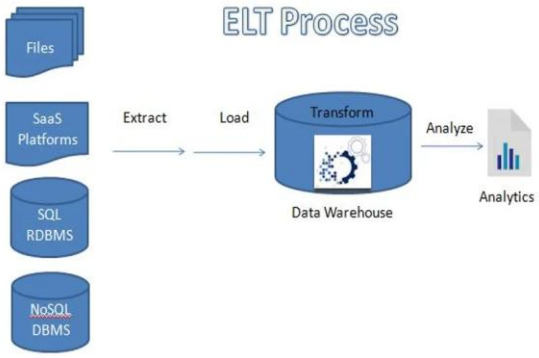

Fig 2-ELT Process

This process consists of three steps:

A.Extract — This is the step where you need to extract data from different source systems in different formats of JSON, XML, flat files. This step is the same for both ETL and ELT methods. Raw data from different applications, software are gathered in this step in ELT.

B.Load — This is where ELT evolves from its ETL.Data extracted are loaded directly to destination instead of loading it to a staging server for transformation. It reduces the cost needed to manage staging servers and also reduces the time between extract phase and load phase.

C.Transformation — This phase in ELT happens at the destination warehouse instead of transforming data in staging servers. A database or data warehouse sorts and normalizes data and transforms it according to source data. Cost of the operation will be highly reduced in ELT as transformation is happening at the destination warehouse without using any staging server for transformation.

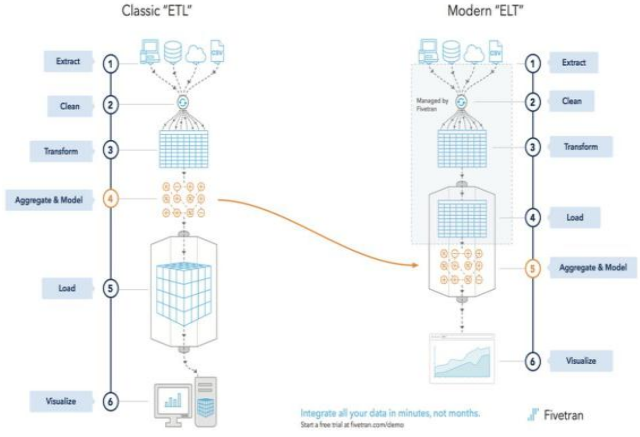

Fig-3: Difference between ETL and ELT

The most obvious difference between ETL and ELT is the difference in the order of operations. ELT copies or exports data from the source location, but instead of loading the data into the staging area for conversion, it loads the raw data directly into the target data store for conversion as needed. Both processes utilize different data stores such as databases, data warehouses, and data lakes, but each process has its strengths and weaknesses. ELTs are especially useful for large unstructured datasets because they can be loaded directly from the source. ETL is suitable for legacy on-premises data warehouses and structured data. ELT is designed for cloud scalability.

IV. RESULTS AND ANALYSIS

Detailed review of ETL and ELT have been discussed above. ETL is traditional process for collecting and transforming data into data warehouse where as ELT is modern process for collecting and transforming both structured and unstructured data into cloud based data warehouse. ELT provides support for data lakes , data marts or data lake house which was not found in ETL method. Im recent times it is necessary to handle data of any size or any type ELT is well suited for this as compared to ETL. Transformation and loading processes are more efficient in ELT than in ETL. Since ELT is cloud based it is cost efficient. whereas ETL is on-premises and requires expensive hardware. ETL is more suited with GDPR and CCPA standards. ELT has more risk of exposing private data and not complying with GDPR and CCPA standards. ELT is a modern alternative to ETL.

V. CONCLUSION

The study focuses on challenges and opportunities in implementing and managing the data pipeline. Challenges fall into three categories: infrastructure, organization, and data quality challenges. The latest cloud-based infrastructure technology provides huge amounts of data storage and computing power at low cost, storing petabytes of data in large, scalable data lakes for rapid on-demand processing. will do so. The surge in data lakes has enabled more companies to move from ETL to ELT. ELT seems to be the future of data integration and has many advantages over the old and slow process ETL. The amount of data in your business is growing exponentially, and ETL tools cannot integrate all this data into one repository for efficient processing. With increased agility and less maintenance, ELT is a cost-effective way for businesses of all sizes to take advantage of cloud-based data.

REFERENCES

[1] N. Schmidt, A. Lüder, R. Rosendahl, D. Ryashentseva, M. Foehr and J. Vollmar, "Surveying integration approaches for relevance in Cyber Physical Production Systems," 2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA), 2015, pp. 1-8, doi: 10.1109/ETFA.2015.7301518.

[2] W. Zhang et al., "A Low-Power Time-to-Digital Converter for the CMS Endcap Timing Layer (ETL) Upgrade," in IEEE Transactions on Nuclear Science, vol. 68, no. 8, pp. 1984-1992, Aug. 2021, doi: 10.1109/TNS.2021.3085564.

[3] J. Huang and C. Guo, "An MAS-based and fault-tolerant distributed ETL workflow engine," Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2012, pp. 54-58, doi: 10.1109/CSCWD.2012.6221797.

[4] B. Z. Cadersaib, Y. Ahku, N. G. Sahib-Kaudeer, M. H. -M. Khan and B. Gobin, "A Review of Skills Relevant to Enterprise Resource Planning Implementation Projects," 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2020, pp. 172-177, doi: 10.1109/ICIMCIS51567.2020.9354270.

[5] E. Begoli, T. F. Chila and W. H. Inmon, "Scenario-driven architecture assessment methodology for large data analysis systems," 2013 IEEE International Systems Conference (SysCon), 2013, pp. 51-55, doi: 10.1109/SysCon.2013.6549857.

[6] J. Sreemathy, S. Priyadharshini, K. Radha, K. Sangeerna and G. Nivetha, "Data Validation in ETL Using TALEND," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, pp. 1183-1186, doi: 10.1109/ICACCS.2019.8728420.

[7] D. Tovarňák, M. Raček and P. Velan, "Cloud Native Data Platform for Network Telemetry and Analytics," 2021 17th International Conference on Network and Service Management (CNSM), 2021, pp. 394-396, doi: 10.23919/CNSM52442.2021.9615568.

[8] A. Tiwari, N. Sharma, I. Kaushik and R. Tiwari, "Privacy Issues & Security Techniques in Big Data," 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCIS), 2019, pp. 51-56, doi: 10.1109/ICCCIS48478.2019.8974511.

[9] A. Wibowo, "Problems and available solutions on the stage of Extract, Transform, and Loading in near real-time data warehousing (a literature study)," 2015 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2015, pp. 345-350, doi: 10.1109/ISITIA.2015.7220004.

[10] J. Lu and M. Keech, "Emerging Technologies for Health Data Analytics Research: A Conceptual Architecture," 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), 2015, pp. 225-229, doi: 10.1109/DEXA.2015.58.

[11] Panos Vassiliadis, 'A Survey of Extract-Transform-Load Technology.,' July 2009 International Journal of Data Warehousing and Mining 5:1-27

[12] Kamal Kakish, Theresa A Kraft, 'ETL Evolution for RealTime Data Warehousing', presented at Conference: 2012 Proceedings of the Conference on Information Systems Applied Research, At New Orleans Louisiana, USA

[13] Florian Waa, Tobias Freudenreich, Robert Wrembel, Maik Thiele, Christian Koncilia, Pedro Furtado, 'OnDemand ELT Architecture for Right-Time BI: Extending the Vision', International Journal of Data Warehousing and Mining 9(2):21-38 · April 2013

[14] Michael J. Denney, MA,1 Dustin M. Long, PhD,2 Matthew G. Armistead, BS,1 Jamie L. Anderson, RHIT, CHTS-IM,3 and Baqiyyah N. Conway, PhD4, 'Validating the Extract, Transform, Load Process Used to Populate a Large Clinical Research Database, 'Int. J. Med. Inform., 94 (2016), pp. 271-274

[15] Valerio Persico, Antonio Montieri, Antonio Pescapè, 'On the Network Performance of Amazon S3 Cloud-Storage Service', 2016 5th IEEE International Conference on Cloud Networking (Cloudnet)

[16] Pwint Phyu Khine, Zhao Shun Wang, 'Data Lake: A New Ideology in Big Data Era', 2017 4th International Conference on Wireless Communication and Sensor Network [WCS 2017], At Wuhan, China

[17] Benjamin S. Baumer, 'A Grammar for Reproducible and Painless Extract-Transform-LoadOperations on Medium Data', arXiv:1708.07073v3 [stat.CO] 23 May 2018

[18] Ibrahim Burak Ozyurt and Jeffrey S Grethe, 'Foundry: a message-oriented, horizontally scalable ETL system for scientific data integration and enhancement', Database (Oxford). 2018; 2018: bay130.

[19] FabianPrasser, HelmutSpengler, RaffaelBild, JohannaEicher, Klaus A.Kuhn, 'Privacy-enhancing ETLprocesses for biomedical data', International Journal of Medical Informatics, Volume 126, June 2019, Pages 72- 81

[20] Gustavo V. Machado, Ítalo Cunha, Adriano C. M. Pereira, Leonardo B. Oliveira , 'DOD-ETL: distributed on-demand ETL for near real-time business intelligence ', Journal of Internet Services and Applications volume 10, Article number: 21 (2019)

[21] Noussair Fikri, Mohamed Rida, Noureddine Abghour, Khalid Moussaid & Amina El Omri, 'An adaptive and real-time based architecture for financial data integration', Journal