International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Modern Data Integration Automation, Analysis and Business Intelligence

Sunil K[1], Revathi SA[2]

[1]*UG Student, B.E Department of Computer Science Engineering, RV College of Engineering, Bengaluru*
[2]*Assistant Professor, Department of Computer Science Engineering, RV College of Engineering, Bengaluru*

Abstract - *In recent days, the prime commodity across the world is the ever-growing data. Huge corporates keep generating data in real time with respect to its clients, customers and employees. In the years since, statistics has continued to mature as a science as have the tools and methods used to analyze data The growth of modern computing and the internet in particular has enabled the collection and analysis of data at vastly larger scales than was possible using pen, paper and tabulating machines. Analytics offers competitive value in several ways It can be used to improve customer acquisition, retention and loyalty. Identify new product opportunities and enhance existing opportunities. By improving organizational decision-making, analytics can deliver many times its cost. Considering the scale of data generated by the corporate it becomes inevitable to dedicate enormous resources, several skilled personnel, time and effort to conquer the objective of data processing, calibrating and storage in-house. The objective is to overcome the challenges corporates bring in data-pipelining technologies and directly receive processed data at the end of the data sync cycle. One sync cycle represents the continuous fetching of data generated or data changed during a specific period such as a fortnight or one month.*

Key Words: *Data Pipelines, Cloud, Data Warehouse, Data Analytics, Sync Cycle*

1. INTRODUCTION

The Major Obstacle to Analytics Data Integration is Creating this central data repository. A central data repository of records offers organization the benefits. We have to gain a big-picture view of the organization's operations and see how the parts work together instead of viewing siloed, isolated representations .We have to match records and track the same entities (customer, partner, etc) across different stages of their life cycles . We can perform analytics in an environment separated from operational systems preventing queries from interfering with operations. Creating this central data repository can be a Herculean task Every data source requires separate procedures and tools to ingest, clean and model its data This challenge has been amplified by the recent proliferation of cloud-based applications and services The appearance of web-enabled devices and sensors (i e the Internet of Things) has likewise contributed to an explosion of data (Figure 1). Since 2013, it has been a truism that 90% of the world's data was created in the last two years.

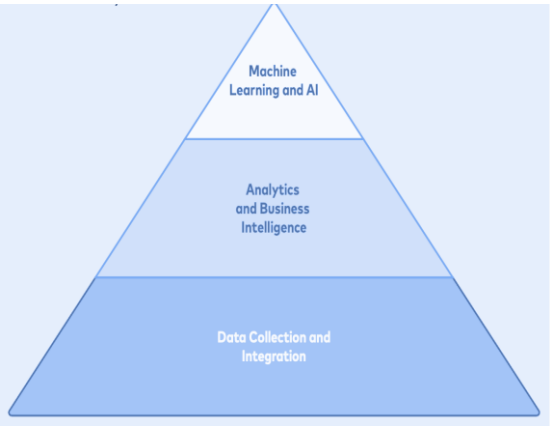

Fig - 1: Data Hierarchy of needs

The rule behind the proposed framework is an information pipeline that associates and concentrates information from source forms and pushes it into a cloud-based warehouse for the comfort of analytics and Business Intelligence (BI). Conceptually, a data pipeline is a pathway from source to destination via some transformations for various analytical applications. The first step is to connect to the source of the raw data and fetch the data load using REST API services. The system processes the load, to handle data integrity issues such as redundant data, skipped data, updated, deleted data and change of data types for a particular field based on the schema available at the source. Final step is to load the clean data into the cloud-based warehouse.

Fig - 2: Basic ELT Process

Before the data pipeline services depended on Extract-Transform-Load (ETL) as shown in Figure 2, where information was separated from the source, they were changed by the explanatory question prerequisites and stacked into the warehouse. The system mentioned was incorporated because of lack of cost-efficient remote data storage facilities at the time. Thus data was transformed so as to store data that were supposed to be queried for analytics. The transformation included consolidation of

data using certain algorithms to generate smaller amounts of data that would suffice the 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 a smaller extent of data.

Unlike ETL, the ELT approach helps reliable pipelines and reduces long-term risk. And it's the substrate of an agile analytics culture. With the onset of cost-effective cloud-based storage services, a new method came to existence. Extract Load Transform (ELT) where the data is extracted from source loaded on to an intermediate cloud facility known as data lake, where the data is cleaned and then loaded into the warehouses for analytical queries.

1.1 Objectives of Paper

Objectives of the system are set with the requirements of each module.

1. The first objective is to identify Goals of Analytics
2. The second objective is to identify Major Obstacle to Analytics Data Integration
3. The third objective is to use the Modern Approach to Data Integration to plummeting costs.
4. The fourth objective is to build a stable Data Integration With a Data Stack to collectively integrate and analyze data from a variety of sources

1.1 Organization of Paper

The paper is split into three important parts.

1. Study of existing architectures in Literature Review
2. Proposed Approach and Implementation
3. Results and Conclusion

2. Literature Review

In the cloud era, applications have become one of the predominant sources of business data apps span a huge range of operations and industries like marketing, payment processing, customer relationship management, ecommerce, engineering project management and many more They provide sophisticated services and operations, obviating the need to construct tools in-house or rely on massive outlays of labor to perform the same tasks manually. applications commonly record actions by users offering organizations highly granular pictures of their operations, from which they can deduce patterns and causal relationships Generally, the more facets of business can quantify and analyze, the more competitive we are. However voluminous data poses an overwhelming data integration challenge a typical company now uses more than 100 apps. At that scale, manual data integration is virtually impossible As we will see, many organizations still write bespoke software and create custom infrastructure to integrate data, but that approach becomes untenable when data comes from dozens of sources that generate a continuous, high-volume data throughput as shown in Figure 3.

Even at smaller scales, the workload imposed by building and maintaining complex data pipeline software can hobble analytics efforts. Heavy time commitments divert analysts, data scientists and engineers from other activities. Luckily, cloud technology offers a solution to this challenge Modern data pipeline tools, data warehouses and business intelligence platforms are cloud-based applications in their own right and they have proliferated alongside cloud technology They effectively eliminate the need to manually develop customized, in-house tools and solutions for data integration and analytics.

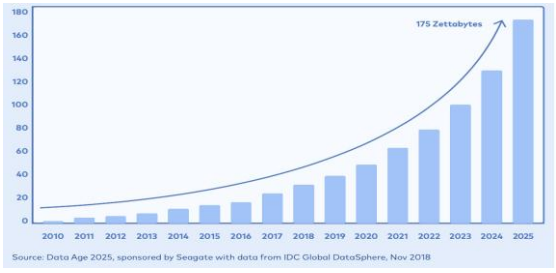

Fig - 3: Annual Size of the Global Datasphere

The concept of data pipelines is fairly recent and the advancements in this particular domain have been enhanced with recent developments in cloud infrastructure and cloud storage. These are the few innovations in the corresponding field of data pipelining.
In 2009, a research was conducted on ETL Technology [1] and it was based on the following principle. The initial software programs that encourage the first stacking and the occasional refreshment of the warehouse are usually known as Extraction-Transformation-Loading (ETL) forms. There were sure restrictions to this, the extraction of information despite everything stays a difficult issue for the most part because of the shut nature of the sources, streamlining and resumption issues and nonattendance of a benchmark is preventing future research.

Then in 2012, real-time ETL Data Warehousing was studied [2]. The aim was to achieve Real-Time Data Warehousing which is highly dependent on the choice of a process in data warehousing technology known as Extract, Transform, and Load (ETL).
In 2013, synchronous investigation [3] was progressing in the field of ELT, utilizing an information distribution center's capacity to straightforwardly import crude, natural records, concede the change and cleaning of information until required by pending reports.

Later in 2016, ETL was being embraced for a few applications, for example, in the clinical area [4]. This information must be appropriately removed, changed, and stacked into the warehouse while keeping up the integrity

of this information. It approved the accuracy of the extract, transform, and load (ETL) process, in this manner populating the Clinical research database.

At the same time Amazon's S3 [5] can come in feature since it gives bulk storage that need not be cleaned or packed to be accommodated. Amazon.com had presented the Simple Storage Service (S3), a low-price capacity utility. S3 plans to give storage as a minimal effort profoundly accessible help with a straightforward 'pay-more only as costs arise' charging model.

After one year in 2017, with the onset of Data lakes, the incorporation of Extract - Load - Transform grew faster [6]. The most straightforward expectation of information lake is to munge each datum delivered by an association to give progressively significant knowledge in better granularity.

Another year later 2018, a standardized approach was incorporated to develop a R based platform using SQL. To develop a predictable and pipe-able framework for R that influences SQL to make reproducible research on medium information an effortless reality.It hence carried the challenges of scalability based on amount of data and also algorithms were not instantaneous in case of medium data which increased latency time. Later that year, another implementation was done to aggregate scientific data for research purposes. A distributed and horizontally scalable extract-transform-load system to tackle scientific data aggregation, transformation and enhancement for retrieval and scientific data discovery [8].

In 2019, research was being done to enhance privacy for ETL processes, in particular with biomedical data [9]. To make the necessary huge datasets, data from unique sources can be coordinated into clinical and translational distribution centers. This was acknowledged on the grounds that current ETL instruments didn't bolster anonymization. In addition, basic anonymization instruments cannot be consolidated in ETL work processes at that point.

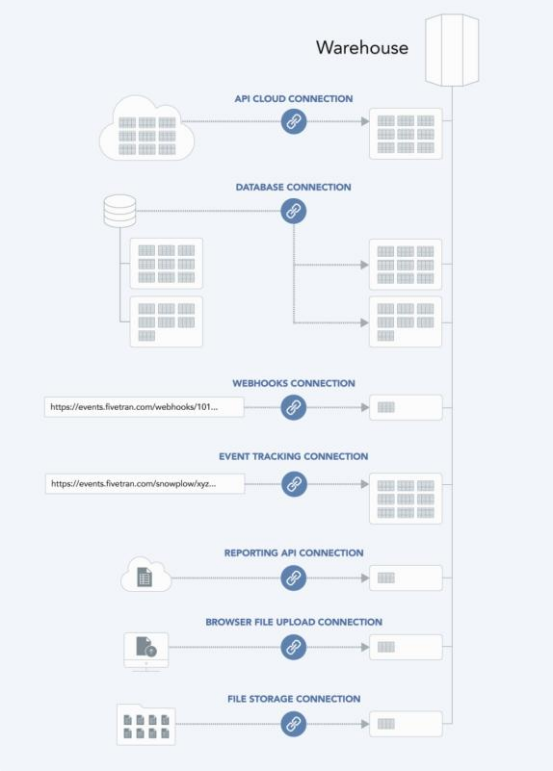

Fig - 4: A connector of each supported connector types

In the same year, another work process was being examined concerning the circulated On-Demand ETL system. DOD-ETL [10], an instrument that addresses in an imaginative way the fundamental bottleneck in BI arrangements, the Extract Transform Load process (ETL), giving it in close to real-time. The essential challenge was to deal with different information sources yet additionally give insignificant latency to respond in real-time.

Later that year, usage in the financial sector was also being explored. Extract-Transform-Load (ETL) concepts (Figure 4), big data processing methods and oriented containers clustering architecture, so as to supplant the exemplary information combination and investigation process by new idea (RDD OLAP) cubes consumed by Spark SQL or Spark Core basics [11]. But additionally, give negligible latency to respond in real-time.

3. Proposed Approach

This project is currently done keeping in mind the relative requirement of a standard ELT pipeline observed for a small-scale source. The system can be used in the real world and be of great help to corporations generating huge amounts of data on a daily basis by increasing the compute and memory used in the cloud.

3.1 ETL Workflow

The workflow that engineers and analysts must perform to produce an ETL pipeline looks like so in Figure 5. It is the industry standard among established organizations

and the acronym ETL is often used colloquially to describe data integration activities in general. ETL evolved at a time when computing power, storage and bandwidth were scarce and expensive. The technical shortcomings of ETL born of that severe resource shortage look increasingly anachronistic in the era of cloud technology.

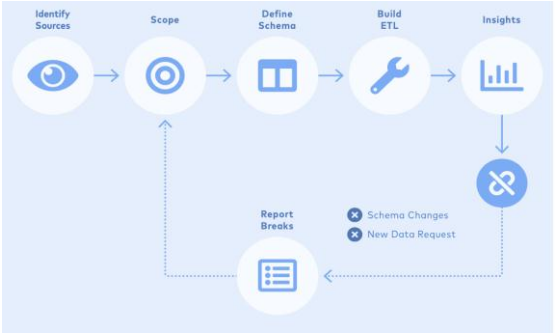

Fig - 5: Data Integration and Analytics Workflow

1. **Identify sources** – apps, event trackers or databases
2. **Scope** – determine the bounds and business goals of the report
3. **Define schemas** – model the data and determine the necessary transformations
4. **Build ETL** – write the software, specifying the details of the API endpoints to call, how to normalize the data and how to load it into the destination
5. **Surface insights** – generate reports that are digestible for key decision- makers
6. **Report breaks** – stoppages leave end users without timely data and cause downtime as a result of:
   a Schema changes upstream
   b New data requests made as analytics needs change
7. **Re-scope the project**

Conceptually, a data pipeline is a pathway from source to destination via some transformations for various analytical applications. The first step is to connect to the source of the raw data and fetch the data load using REST API services. The system processes the load to handle data integrity issues such as redundant data, skipped data, updated and deleted data, and change of data types for a particular field based on the schema available at the source. Last advance is to stack the perfect information into the cloud-based warehouse. Before, the information pipeline administrations depended on Extract-Transform-Load (ETL), where information was extricated from the source changed by the analytical query requirements and stacked into the warehouse. This system was incorporated because of lack of cost-efficient remote data storage facilities at the time.

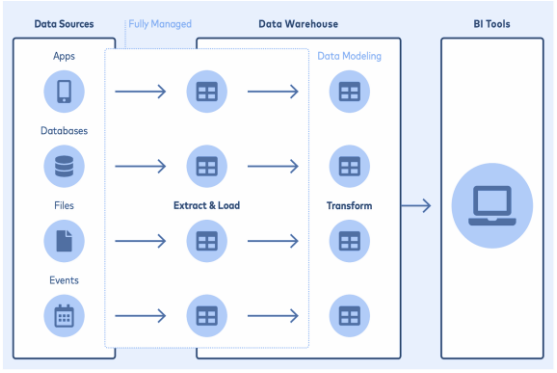

Fig - 6: Sequence of process in extract-load-transform (ELT)

Thus, data was transformed so as to store data that were supposed to be queried for analytics as shown in Fig 6. The transformation included consolidation of data using certain algorithms to generate smaller amounts of data that would suffice the 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.

4.0 Results and Discussion

In an ideal world, roughly 80% of an average data scientist's time is spent constructing data pipelines — a task for which most data scientists have limited aptitude, interest or training (Figures 7). The most obvious argument against constructing one's own ELT pipeline is the cost of building and maintaining it in terms of time, money, morale and lost opportunities.An automated data integration solution will save time, money and labor depends on organization's size and maturity as well as the particular characteristics of the data pipeline provider.

Data analysts have access to all their required data without concern for where it's stored or how it's processed—analytics just work. Until recently, the reality of analytics has been much more complicated. Expensive data storage and underpowered data warehouses meant that accessing data involved building and maintaining fragile ETL (Extract, Transform, Load) pipelines that pre-aggregated and filtered data down to a consumable size. ETL software vendors competed on how customizable and therefore specialized, their data pipelines were.

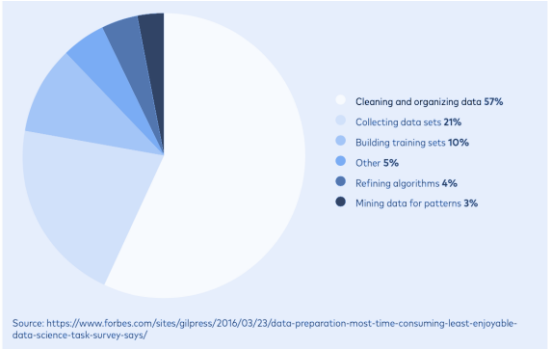

Fig - 7: Amount of work invested in Data Analysis

Technological advances now bring us closer to the analysts' ideal. Practically free cloud data storage and dramatically more powerful modern columnar cloud data warehouses make fragile ETL pipelines a relic of the past. With onset of cost-effective cloud-based storage services, a new method came to Fig. 6. difference between ETL and ELT. Existence Extract Load Transform (ELT) where the data is extracted from source and then loaded into the warehouses for analytical queries.

Modern data architecture is ELT—extract and load the raw data into the destination, then transform it post-load. This difference has many benefits including increased versatility and usability. Some of the morale from their difference are Diversion from other software engineering, data science or analytics duties — a very common irritant among new data scientists at understaffed organizations which can lead to turnover. Frustration and exhaustion from the complexity of maintaining data integrity, particularly by persons lacking the appropriate training.

Data warehouses are periodically populated with data from Operational Data Stores (ODS). The biggest hurdle to provide efficient BI latency requirements is the data processing required to make data present in a data warehouse (DW). The data fetched from the Operational Data Stores need to be processed, operations like cleaning, normalization of the data has to be performed before it is suitable for BI to analyse the data for a variety of reasons. Also, the Operational Data Stores may contain corrupted data, duplicate data that needs to be reconciled. This preprocessing is commonly known as Extract-Transform-Load (ETL). The data is first extracted from the original data source then transformed including normalization, cleansing and finally loaded into the data warehouse. Figure 4 shows the difference between ETL and ELT. ETL has improved in scalability and performance at a low level. They are not upgradable very easily and are hurlesome in nature. As a result, most BI infrastructures are exponentially experiencing an automated analysis pipeline bottleneck. ETL has to be replaced with some other strategy to provide an efficient pipeline and analysts have to spend a little time pushing his data to the warehouse.

Current ETL processes, which are specialized in processing mass data. They don't provide efficient infrastructure for incremental updates in growth of data. ETL are specialized in driving huge amounts of data to the warehouse but the critical path they follow is that they undergo a lot of complex transformations. Because the transformations are complex and difficult to understand for analysts, analysts cannot tune his data based on his own transformation after the development of initial transformations. Handling transformations and upgrading them are huge time consuming and requires a lot of understanding beforehand. ETL is clearly not for the Agile based tech driven companies. They are not compatible with the growing technology and data variability. They have serious drawbacks which drive the analyst into having a static transformation and never upgrading them. ETL applies a layered infrastructure based on DBMS access modules, file access modules, parallel read/write modules, data processing modules and to apply partitioning, parallel processing, and pipelining to ETL processes. Other disadvantages of ETL are, they are not compatible with providing sorting of data records to be only visible on the Warehouse side. Also they are inefficient in providing features for non-blocking columns, where you need to fetch certain columns as they are part of primary key data or foreign key data based on the application use case. Another important disadvantage of ETL is their inefficiency in providing automated incremental updates on the Warehouse by only capturing the change from the source and pushing it to the Warehouse.

Transforming data and uploading to the warehouse has to be a small portion of the job of an analyst and much of development work has to be concentrated into analytics. ETL are very sensitive in their transformations, if a single mistake is made initially and coming out of the mistake would be very difficult or impossible in some cases. Current ETL transformations are clearly not the right solution for massively growing data.

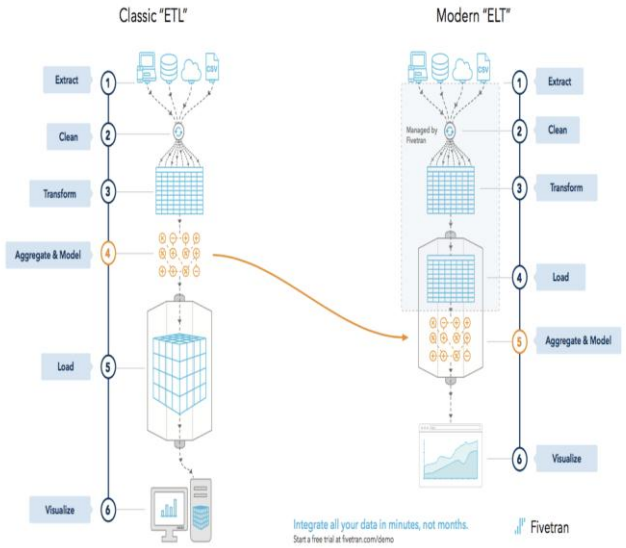

Fig - 8: Difference between ETL and ELT transformations

We need a technology which would give us flexibility over the data we pipeline and easier to configure and maintain.

Limitations of ETL Overall, the traditional ETL process has three serious and related downsides:

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.
3. Inaccessibility. More importantly, ETL is all but inaccessible to smaller organizations without dedicated data engineers. On-premise ETL imposes further infrastructure costs.

5.0 Conclusion

Technological trends know that computation, storage and bandwidth have become cheap and ubiquitous. With advances in computing, the cost of computation has plummeted over time. Likewise, in a span of about 35 years, the cost of a gigabyte has plummeted from nearly $1 million to a matter of cents. One effect of these radical cost reductions is that data warehouses can accommodate much larger volumes of data. Organizations no longer need to pre aggregate and, in the process, discard a great deal of source data. This enables analysts to perform deeper and more comprehensive analysis than ever before. Although the World Wide Web did not exist until 1991, the cost of internet transit has also decreased radically. In less than twenty years, it dropped from about $1200/ Mpbs to a matter of cents. The convergence of these three cost-reduction trends led to the cloud namely, the use of remote, decentralized, web-enabled computational resources. Cloud technology, in turn has given rise to a huge range of cloud-native applications and services.

Many organizations rely on a manual, ad hoc approach to data integration — in fact 62% use spreadsheets like Excel and Google Sheets to stitch together elements from data files and visualize data. 2 This involves downloading files, manually altering or cleaning values, producing intermediate files and similar actions.

A more sustainable approach is to maintain the silos between separate data sources while bridging the gaps between them with “federated” queries, which directly query multiple source systems and merge data on the fly. Organizations may do this with SQL query engines like Presto. The disadvantage of this federated approach is that it involves many moving parts, and its performance degrades at large scales of data. The reality is that a scalable, sustainable approach to analytics requires a systematic replicable approach to data integration — a data stack.

References

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

[2] Kamal Kakish, Theresa A Kraft, ‘ETL Evolution for Real-Time Data Warehousing’, presented at Conference: 2012 Proceedings of the Conference on Information Systems Applied Research, At New Orleans Louisiana, USA

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

[4] 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

[5] 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)

[6] 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

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

[8] 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.