Challenges and Opportunities to Business Analytics and Integration Due to Edge and 5G Technologies

Sathish Kumar Sampath
Boston University

Abstract— For organizations to succeed in their respective domains, they need to define a Business Strategy that outlines the Key Performance Indicators that are measurable. Organizations, then, define a Data Strategy that takes all data into consideration and provides analytics that eventually maps to the Key Performance Indicators that are defined.

The recent advancements in technologies based on 5G and Edge deployments that are producing Big Data are primarily driven by the motivation of providing a highly interactive and localized customer experience. Organizations that are deploying solutions in Edge locations therefore need quicker analytics to react faster and therefore they need a data strategy that can meet their demands.

This research paper aims to provide an overview of the various Data Integration and Business Analytics strategies that organizations have implemented as part of their overall Data Strategy. Additionally, this paper discusses the challenges faced by organizations deploying Big Data solutions due to the traditional Data Strategies and proposes a concept to meet the needs of organizations that need near real-time analytics by leveraging Analytics tools and Artificial Intelligence.

Keywords—Big Data, 5G, Data Strategy, Data Integration, Business Analytics, Edge, Artificial Intelligence

I. Introduction

Data is at the core of strategy and decision-making for organizations to grow and succeed. For Organizations to execute their Business Strategy, a strong Data Strategy is required that can turn data into value and map to the overall Key Performance Indicators defined in Business Strategy. As Organizations continuously challenge their assumptions and their previous performances, they come out with aggressive Business Strategies and this in turn requires an innovative Data Strategy that can deliver results with a quicker turnaround time.

[6] A successful and comprehensive Data Strategy, it requires a good understanding of all the data sources and their appropriate structures. Bringing data together from different source systems, ensuring the data is in the right format, and keeping the integrity of the data as it gets transformed are some of the key objectives for any Data strategy. Therefore, for any organization to be successful, a comprehensive Data Strategy is required that addresses the challenges due to these various heterogeneous systems by providing a strategy for various translation layers to bring uniformity without affecting the integrity and additionally provide a process for cleaning the data before consolidation. After ensuring the data is consolidated, it then needs to be loaded into an appropriate Business Analytics tool for the purposes of Analytics and decision-making.

The entire process of Data Integration could range anywhere from 30 mins to 12 hours (or even more) depending on the total number of source systems, the size of the data, and the associated transformations. Therefore, a big challenge for the successful execution of Data Strategy is ensuring that the Data Integration timeline is as aggressively low as possible, at the same time, cost-effective to ensure it can deliver results that matter. Furthermore, with Edge technologies and 5G solutions going mainstream, there is a bigger need to react quickly based on analytics results, and this, in turn, adds more pressure for an effective data strategy with quicker turnaround times. According to Gartner estimates, by 2025, 75% of the Enterprise data will be handled in Edge sites and therefore an effective Data Strategy is required to handle the data generated out of the Edge locations.

[8] This research paper argues that the current methodologies of Data Integration and Analytics, and the recent evolutions in Edge Analytics have limitations in meeting the demands of the next generation technologies such as Smart Cities, the Internet of Things, and others that rely on Edge and 5G technologies. This paper additionally proposes a solution that can effectively meet the demands of technologies.

II. Background

[1] Research has shown that when Organizations either have a single data center or multiple data centers producing homogeneous data, there is little need to transform data before analyzing the same for decision-making purposes.

When little to no transformations were required, Data Integration from multiple homogeneous solutions typically happened within 2 hours– 4 hours on average for about 1TB

979-8-3503-2759-5/23/$31.00 ©2023 IEEE
DOI 10.1109/CSCE60160.2023.00264

of data from each data center. A small POC performed revealed that when the transformation involves 2 fields within a particular table from 2 source systems that contain 1TB of data each, and then when an integration operation is performed, the time it takes is close to 2 hours. Note that this POC only considers transformation and integration, it does not count loading into the database. Therefore, the overall time it takes from Data Integration to Analytics takes about 5-7 hours. Note that the use case that is taken into consideration is a very simplistic one. A realistic use case will involve additional transformations and therefore the numbers here are proportional to the transformations and integrations.

[1] After Cloud solutions became prominent and started gaining popularity, Organizations started to deploy either private cloud, hybrid cloud, or public cloud solutions. This multi-deployment model meant that data was getting generated from various systems in either a structured or unstructured format and to make sense of all the incoming data, it was important to have a comprehensive data strategy with the ability to capture changes for effective decision-making purposes. Due to the complexity associated with consolidating data from various source systems, the Data Integration activity along with Business Analytics does take a longer time to execute before it can be up for analytics. A POC done on this and along with additional research indicates that it takes anywhere from 12-24 hours, depending on source systems, the structure of data, the amount of data, and the necessary transformations.

[3] The next wave of technological advancements such as the Internet of Things, Smart Cities, Gaming Technologies, and others that are currently gaining a lot of prominence is dependent on Edge solutions. The need for organizations to build and adapt to these technological advancements is primarily driven by the idea of providing a real-time and localized customer experience. This would mean that organizations need a stronger data strategy in place to integrate all the disparate data coming from different systems in different geographies and a corresponding Analytics tool to analyze these in a real-time manner for effective decision-making purposes.

A. Technology Background

1) Data Integration Methodologies

[5] The two popular methodologies associated with the Data Integration pipelines are ETL (Extract, Transform, and Load) and ELT (Extract, Load, and Transform).

While ETL is the most popular as it accounts for the transformation and cleaning up of data before using the same for analytics purposes, there are limitations associated with it as well. The primary and most concerning limitation is that the entire process is a time-consuming and costly one. When there are multiple sources with disparate formats, the complexity further increases as Transformations need to be completed for data originating from all sources and these cause further delays.

With ELT, the Load process happens before the Transform process. Due to this approach, data in different shapes and formats are first loaded either in a warehouse or data lake before using it appropriately for Analytics purposes. This method significantly reduces the overall time since the transformation is done on a need basis. However, the biggest limitation of this process is the security concern. Data that is stored in the warehouse has the possibility of containing sensitive information and if not handled appropriately can result in leaks.

Another methodology that is gaining popularity and addresses the security issue of ELT is the EtLT model. This is the Extract(E), Transform(t), Load(L), Transform(T) model. The Transform process after the Extract process is an intermediate one and is typically used to do a lighter clean-up of data and ensure the security concerns are addressed before the Load operation is performed.

While both ELT and EtLT methodologies were introduced to deal with Big Data and the need for quicker responses, they still lack the flexibility and faster response times that are typically required by organizations for real-time decision-making.

2) Business Analytics Tools

Business Analytics tools are deployed at a Centralized location. After all the data is consolidated, they are loaded into Business Analytics and Intelligence tools for comprehensive analytics, that will further be used for decision-making purposes. Every organization has certain Key Performance Indicators (KPIs), and the goal of a data strategy will be to ensure that these KPIs are defined and tracked appropriately in the Business Analytics tools.

These Business Analytics tools need a complete view of all the data for producing a complete analytical view of the performance of the Organization and therefore it is important that the data strategy has a solid data integration strategy.

III. Data Strategy for Edge Deployments

As mentioned in the previous sections, many organizations are focused on enhancing the customer experience by providing localized and quicker turnaround times to customer requests. Before reviewing the Data Strategy that Organizations are implementing for Edge Solutions, here is a quick review of 5G and Edge technologies, and the challenges of implementing an effective Data Strategy.

A. 5G and Edge Technologies

[4]5G technology is built on the core concept of ensuring lower latency, faster response, and higher bandwidth. The main characteristics of 5G are that it needs to handle massive

amounts of data, needs to have a stringent Quality of Service, supports heterogeneous environments, and allows interoperability.

[4] Edge technologies are focused on deploying solutions as close as possible to customers. This is necessary to ensure that critical applications and infrastructure can stay as close as possible to users, therefore, allowing for a faster response time and provide localized experience. Figure 1 is a model of Edge Computing. The Edge sites in the figure indicate the deployment of critical infrastructure deployed in a geographical location to cater to the needs of the users or customers at that location. Edge sites then synchronize with the infrastructure or application in the cloud appropriately.

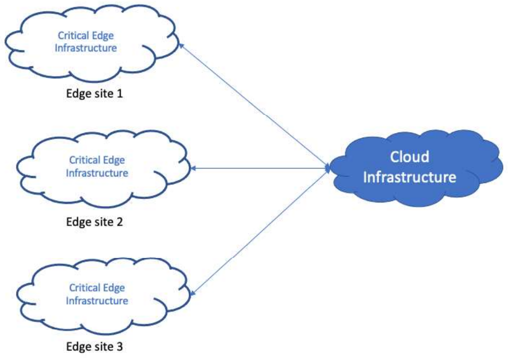
Critical Edge Infrastructure
Edge site 1
Critical Edge Infrastructure
Edge site 2
Critical Edge Infrastructure
Edge site 3
Cloud Infrastructure

Fig.1: Edge deployments

Therefore, both 5G and Edge technologies, either combined or separately will be the enablers for various technologies producing Big Data such as the Internet of Things, Smart Cars, Gaming solutions, Connected Devices, and others.

For both 5G and Edge technologies to be successful, organizations need a near real-time analysis of the performance of their entire solution and the ability to make faster decisions based on incoming data from these deployments. Therefore, it is critical to ensure Data Integration and the corresponding Analytics is as close as possible to the deployment to reduce the turnaround time and can provide analytics in near real-time.

B. Limitations with current Data Integration and Analytics solutions in meeting Edge Deployment requirements

The primary purpose of organizations deploying solutions that rely on Edge and 5G technologies is to provide a highly interactive, localized customer experience. Therefore, they will also expect real time analytics of their solution. The current set of Data Integration and Analytics solutions has certain limitations in addressing this requirement due to the following factors:

- Analytics solutions are deployed at a centralized location, typically in a cloud and it needs a consolidated view of the data to make decisions. Consolidation of data can happen only after the data goes through either ETL or ELT or EtLT model and this cannot happen on a real-time basis.
- Edge locations are expected to produce terabytes of big data and consolidating all of these is a time-consuming activity.
- When new Edge locations are deployed, these can further complicate an already existing Data Model with additional big data. Therefore, a strong Data Strategy needs to be in place to handle these large amounts of data.
- A strong data governance model needs to be in place to ensure that there is no stale data sitting in any of the locations because of either transformations or mis outs.

To put all these limitations into perspective and review what this is, we shall take the previous POC numbers and extrapolate this further. When a single edge site generates 1TB of data in a day and if there are 5 edge sites in place, this results in 5TB of data that needs to be transported into a single data warehouse. This whole operation is expected to take at least 2-3 days considering it is a 1Gbps Ethernet connection. Transformations are then expected to take place which again can take 1-2 days, depending on the complexity. Therefore, for any meaningful analytics, it is expected to be at least 4-5 days.

C. Recent developments to address Challenges

1) Streaming Analytics

[4] Streaming Analytics is a process where data is continuously collected from the source systems and they are streamed to a location, that is typically a data warehouse. Streaming Analytics works based on real-time queries that pull data from the source and after data reaches a warehouse, they will be processed appropriately and consolidated through Transform methodologies.

Streaming Analytics aims to resolve the lagging issue that traditional solutions as data is collected in real-time as it is generated. However, it should not be confused with Real-time Analytics as Analytics is expected to happen only after consolidation. Consolidation of data will happen only after all the source data is collected in a location and transformed appropriately.

Therefore, taking the above example into consideration, this option is expected to reduce the whole turnaround time by 1 day or two since the data is streamed as it is collected.

2) Analytics at the Edge

[4] To ensure that Data is processed and analyzed either at the Edge or near the Edge, Data should be stored at the Edge location, instead of sending it to a centralized data warehouse. An Analytics solution also needs to be deployed near the Edge that can act on the data at the localized location (Fig 2). Therefore, the typical ETL or ELT pipeline is not expected to

add value in this deployment as multiple TBs of data are expected to be generated and any centralized solution is not expected to be scalable to handle this large volume of data. However, multiple EtL pipelines are expected to be deployed near the Edge that can serve as an input to the Analytics model. Taking the above example and putting into this option's perspective, the results for Analytics at the Edge is expected to be available within 1-2 days.

This solution however raises an important question as to whether the big picture will be missed when data is processed at a local site and therefore having a consolidated and centralized view will not be possible. The answer to this lies in how organizations define their data strategy. Depending on the organization's goals and Key Performance Indicators (KPI), a data strategy needs to be drafted and this needs to consider if Edge Analytics is required or having a consolidated Data and Business Analytics will be good enough. If Edge Analytics is required, it is important to have multiple EtL pipelines with a view of the type of data for each of the Edge deployments.

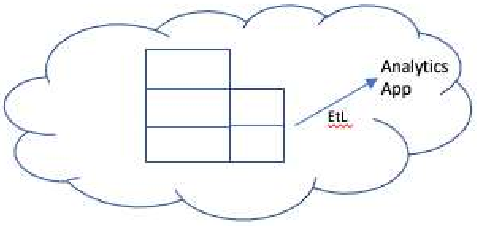

Fig 2: Edge site with local Analytics

IV. Proposal for Analytics tools to leverage Artificial Intelligence

The whole Data Strategy is defined with the goal of measuring an organization's metrics that are part of the overall Business strategy. Therefore, a Data Strategy needs to take into consideration consolidating data from various sources and landing on one target that can further be used to feed into an Analytics solution deployed in a cloud or a centralized location. However, with Big Data challenges that originate TBs of data from each edge site, it is practically a huge cost-impacting solution to consolidate all the data into one centralized repository.

This research paper proposes an Interconnected Analytics Solution (Fig 3). The proposal is that a Business Analytics application will be deployed either near the Edge or within the Edge that performs Real-time (or near Real time) Analytics and then an interconnected network of Edge Analytics applications to the Cloud Analytics application. The outcome of Analytics from each of the BA applications at the edge site will be transferred to the Cloud Analytics application and these will be consolidated for Overall Analytics at an Organizational level. The output of this activity will then be mapped to the Organization's KPIs to review the performance of the organization.

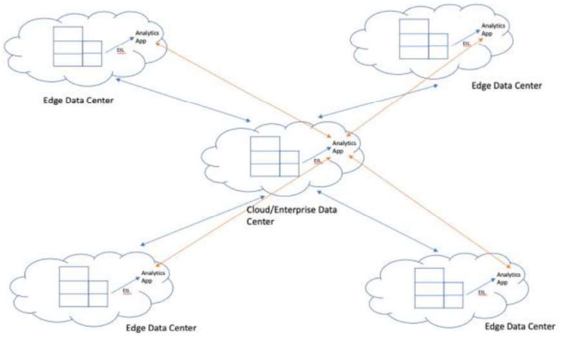

Fig 3: Interconnected Analytics Solution

Additionally, for an Edge Analytics application to perform Real-Time analytics, it would still need to have a better perspective of data from other Edge locations or from the central Cloud location. Artificial Intelligence (AI) will be leveraged to identify if an Edge Analytics application has all the information and if it does not, data will either be fabricated or augmented based on already existing processed data from the Cloud Analytics solution. Models shall be built and improved upon to ensure that the overall Analytics solution becomes intelligent over a period.

With AI, Data and Business Analytics solutions deployed at Edge locations has the potential to become smarter by augmenting data and visualizing a bigger picture while analyzing for a particular edge location. Here is a detailed view of the proposal:

- Organizations need to have a Data Strategy that will encompass separate strategies for each of the Edge locations.
- Each Edge site is expected to have a separate EtL pipeline feeding into the Analytics application that is either deployed in the same Edge site or in a cluster of Edge locations.
- The Analytics application is expected to be either in a mesh model with other Analytics applications (at other Edge sites) or the Enterprise Data Center.
- With Artificial Intelligence, all Analytics applications in the cluster shall share intelligence and build models that will give a perspective of business both from the Edge location and from the overall Organizational level.
- With this model, Real-time analytics can be performed intelligently and effectively. Additionally, the performance of the overall Organization shall be measured as well.

1) Benefits of using Artificial Intelligence(AI)

Artificial Intelligence (AI) can bring in multiple benefits to this approach:

- With AI, a certain pattern can be identified based on past metrics, and this can be used to fabricate or predict the complete picture of the organization in an Edge Data Center.
- [3]AI has proven to be a very successful approach in analyzing big data and especially, when it comes to machine learning, advanced algorithms, and improved computing power and storage.

2) *Advantages of this approach:*

This approach is expected to have multiple benefits:

- Organizations will have the ability to get an Edge site-specific view in real-time and derive an organizational view from the central Analytics solution.
- The entire data generated from a specific Edge site does not need to be consolidated at a central location. It can instead stay at an Edge site, and this can reduce costs.
- The Data Integration pipeline can be local at the Edge site. Transformations can be confined to the data that is gathered at the Edge site. This can therefore reduce the overall turnaround time from days to hours.
- Security concerns can be addressed as the EtL pipeline is expected to handle all sensitive information and the data does not leave the Edge site.
- Since all data will be processed at the Edge, the chances of leaving out data from analytics is minimal. There will be no stale data.
- Analytics data that is exchanged between the various Analytics solutions will be in a uniform pattern. Therefore, no transformation will be expected.
- This solution is expected to be scalable. Organizations will continue to add new Edge sites or deploy solutions in new geographies. In this case, the Analytics solution can be deployed at the local site and can be hooked to the central site.

3) *Reduction in TurnAround time and Costs*

The driving point behind this proposal is to have a reliable data strategy that can effectively reduce costs, provide analytics within a shorter timeframe, and additionally provide options for localized analytics. Based on the results of the POC that was performed on individual deployment models, here is the inference on how this proposal is beneficial to organizations:

- When analytics is performed locally at the Edge, organizations can potentially eliminate the task of migrating data to a centralized data warehouse. For a 1TB of data from a particular Edge center, this proposal will bring down the turnaround time from 2 days to 6 hours.
- This proposal additionally brings down the costs of moving the entire data to a particular central repository as data is now handled and stored at the Edge.
- With AI, Analytics solutions deployed in multiple sites can exchange the outcome of the results and therefore a comprehensive Analytics solution can be deployed with lower costs.

V. Conclusion

Business Analytics and Data Integration methodologies have undergone massive changes with changes and advancements in technologies and deployments. When Cloud solutions started gaining prominence and widespread adoption, ETL and ELT models of Data Integration started becoming popular since they ensured data can be captured from any kind of cloud deployment (private, public, or hybrid) and with the ability to capture changes in real-time. Additionally, improvements were made within Business Analytics to support widespread data integration technologies and provide a comprehensive solution to customers. These models also supported the expansion into new regions and new deployment models. However, Organizations have always struggled with getting real-time analytics for quicker turnaround to issues.

With Big Data, Edge, and 5G technologies going mainstream, traditional methodologies are expected to be under a lot of pressure to cope with the demands of organizations as they are expected to provide faster responses to customers. Therefore, vendors of Data Integration and Analytics are expected to adopt to newer ways to meet the demands. The proposal of deploying multiple Analytics applications and interconnecting, and then leveraging Artificial Intelligence, that is described in this research paper shall be considered by Vendors as they look to meet customer demands.

References

[1] N. Tatbul, "Streaming data integration: Challenges and opportunities," *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, Long Beach, CA, USA, 2010, pp. 155-158, doi: 10.1109/ICDEW.2010.5452751.

[2] Yang Bao, Shi Wei Deng, Wang Qun Lin, Research of Data Cleaning Methods Based on Dependency Rules

[3] Yanqing Duan, John S Edwards, Yogesh K Dwivedi "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda"

[4] N. Hassan, K. -L. A. Yau and C. Wu, "Edge Computing in 5G: A Review," in *IEEE Access*, vol. 7, pp. 127276-127289, 2019, doi: 10.1109/ACCESS.2019.2938534.

[5] Maurizio Lenzerini, "Data integration: a theoretical perspective,"

[6] Anhai Doan, Alon Halevy, Zachary Ives, "Principles of Data Integration"

[7] M. Mohammadi, A. Al-Fuqaha, S. Sorour and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2923-2960, Fourthquarter 2018, doi: 10.1109/COMST.2018.2844341.

[8] Sabuzima Nayak, Ripon Patgiri, Lilapati Waikhom, Arif Ahmed, “A review on edge analytics: Issues, Challenges, opportunities, promises, future directions and applications