BRIDGING DATA SILOS USING BIG DATA INTEGRATION

Jayesh Patel

Senior Member

ABSTRACT

With cloud computing, cheap storage and technology advancements, an enterprise uses multiple applications to operate business functions. Applications are not limited to just transactions, customer service, sales, finance but they also include security, application logs, marketing, engineering, operations, HR and many more. Each business vertical uses multiple applications which generate a huge amount of data. On top of that, social media, IoT sensors, SaaS solutions, and mobile applications record exponential growth in data volume. In almost all enterprises, data silos exist through these applications. These applications can produce structured, semi-structured, or unstructured data at different velocity and in different volume. Having all data sources integrated and generating timely insights helps in overall decision making. With recent development in Big Data Integration, data silos can be managed better and it can generate tremendous value for enterprises. Big data integration offers flexibility, speed and scalability for integrating large data sources. It also offers tools to generate analytical insights which can help stakeholders to make effective decisions. This paper presents the overview on data silos, challenges with data silos and how big data integration can help to stun them.

KEYWORDS

Data Silo, Big Data, Data Pipelines, Integration, Data Lake, Hadoop

1. INTRODUCTION

A data silo is a segregated group of data stored in multiple enterprise applications. Most applications store raw and processed data in various ways. Each application has its own features and tools to allow business users an easy access to processed data using dashboards and reports. Most of these dashboard and reports are application specific. As a result, teams in various business units will have limited visibility of the data in the enterprise and they will only see a partial picture. They will not be able to extract the full value of data from various enterprise applications which sources data silos. Data silos restrict sharing information and collaboration among teams. It leads to poor decision making and negatively impacts profitability [16][17].

With evolving APIs in enterprise applications and the emergence of new technologies, frameworks, and platforms, there are various opportunities to integrate these silos. Integrating thousands of disparate data sources is now easier than ever before due to Big Data Integrations and tools. This paper will focus on how big data integration can help integrating disparate data sources.

2. RELATED WORK

Data integration has been around since last few decades and it has been evolving. Data silos are often managed differently by enterprises based on their priorities and challenges. Data Integration is the key to bridge data silos and generate true value of data [15].

As applications can be heterogeneous, it is challenging to integrate them. To integrate heterogeneous sources, data integration mitigates risks and offers flexibility. In paper [21], multiple approaches were discussed to integrate heterogeneous data warehouses by an example of a practical system.

The paper [14] compares traditional data warehouse toolkits with big data integration. It provides details on methodology, architecture, ETL challenges, processing and storage for data warehouse and big data lake. It summarizes characteristics of data warehouse and Big data followed by a proposed model to process big data.

A seminar on big data integration [2] discussed techniques such as record linkage, schema mapping and data fusion. It also summarized how big data integration relieves challenges with Big data.

Big data integration is reviewed in paper [11] by presenting issues with traditional data integrations and relevant work on big data integration. Challenges of Big data is also discussed with techniques to resolve them in a big data environment.

A book on Big Data Analytics [1] discusses strategies and roadmap to integrate big data analytics in enterprises. It presents how big data tools and techniques can help to develop big data applications to manage data silos.

3. DATA SILOS AND CHALLENGES

3.1. Data Silos

In the age of artificial intelligence, data science, machine learning, and predictive analytics, generating insights from the right data at the right time is critical. These days, there is a strong wave to adopt artificial intelligence and machine learning to be data-driven and to gain a competitive advantage.

The biggest hurdle to attain success in this endeavor is data integration and preparation [18]. Due to the increasing number of business applications in the corporate world, data sit in hundreds or thousands of applications or servers that result in data silos. It becomes worse at the times of mergers and acquisitions. On the other side, it is not practical to give everyone in the company to give access to all applications. Even if that is true, it will take so much time to integrate required datasets without a proper strategy. It takes days, months or even years to manually acquire and integrate data from data silos. Data silos are formed due to various reasons not limited to structural, vendor contracts, and political [16].

3.2. Challenges with Data Silos

According to a survey by the American Management Association [8], 83% of executives told that their organizations have silos and 97% think silos have a negative effect. The simplest reason is they impact overall decision making and impacts each business vertical.

It is obvious that data silos restrict visibility across different verticals. Data enrichment is not possible with data silos and that's why it impacts negatively on informed decision making. Additionally, data silos can represent the same problem or situation differently. For example, an active subscriber in a SaaS (software as a service) company may mean different to finance team than the marketing team. For finance team, if a subscriber is paying a subscription or using a promotion, he/she will be considered active. For marketing, active status depends on various

activities on the platform like login, activity, etc. That leads to inefficiency and additional work to determine which source is accurate.

3.3. Bridging Data Silos

As data silos are formed from multiple applications and processes, data reside at various places-cloud, on-premise, applications servers, flat-files, databases, etc. A key thing here is to find value in data and define which data silos should create maximum value if they are integrated. One way you can overtake data silos is to strategize sources of data silos to enhance collaborations and communications among multiple departments and teams. From separating different processes or applications to unifying them will break down data silos. However, it requires major efforts and a change in the overall culture of the organization [16][17].

Another way to solve this problem is to integrate these data silos using integrations techniques and tools. Integrating these data silos is a costly and time-consuming process. There have been multiple frameworks and tools for data integrations but we will focus on big data integration due to its long term benefits.

Let's understand the need for integrating data silos in the corporate world with a simple analogy. In a household kitchen, you will find sugar in a container, coffee in another container and creamer in a different container. We consolidate all three in different proportions to make a delicious coffee. Similarly, multiple applications in enterprises form data silos for specific operations. When executives and investors look at a company as a whole, a clear and better view of the overall picture will go a long way. Integrating data silos in an effective way can resolve this critical challenge.

4. BIG DATA INTEGRATION

Enterprise applications generate a variable volume of data real-time, near-real-time or not-real-time. Data can be structured, semi-structured or unstructured [5][14]. Almost 80% of data from enterprise applications is either unstructured or semi-structured [9]. Volume can be low or high but there is no clear bar to classify volume as low or high [11]. Applications can generate static or dynamic data structures and they are heterogeneous in most cases [14]. Most of these demonstrate the characteristics of Big Data, often known as the seven V´s: Volume, Variety, Velocity, Veracity, Value, Variability, and Viability [4][14].

Traditional data integration techniques work well for structured data [21]. Traditional data integration can handle some of the characteristics of Big Data but it fell short on handling semi-structured and unstructured data at scale [2][3]. With advancements in big data technologies and infrastructure, it becomes much easier to integrate a variety of data sources at scale [11]. Distributed processing and distributed messaging systems made integrations possible at high scale. Map Reduce and Hadoop used to be a good option to integrate data silos [14]. Spark and Hadoop can be better to integrate a variety of data sources after recent development in Spark [6]. Kafka and other distribution messaging technology made real-time data integration possible [12][13]. Depending on the needs, data can be processed in batch, or real-time using big data tools.

To integrate data silos, you can either extract raw data from each application or pull processed data from sources based on the requirements and nature of applications. Following general steps are identified to integrate data silos based on several use cases:

1) Data Discovery: Identify data silos to be integrated. It is a data discovery and justification phase. Evaluate and determine what integrating data silos will benefit the entire corporation. Set the clear expectation, benchmark, and constraints for each data silo. This phase will give you some idea of priorities on which data silos should be handled first.

2) Data Extraction: Determine what to extract from sources. Data can be from relational databases, cloud, IoT sensors, flat files, application logs, images, large documents, and so on [1][19]. Data extraction should not put more burden on data sources and should not impact other production processes. "Schema on read" can be beneficial for semi-structured and unstructured data and big data integration can handle it easily [10].

3) Data Export: Export data to temporary storage or messaging system. Kafka is a really powerful message broker for high-throughput real-time event data and can be really effective in integrating data silos [12][20]. Other alternatives should be evaluated based on needs.

4) Data Loading: Load data to the cloud, Hadoop Distributed File System (HDFS) or any big data platform real-time or in batches [6]. Data can be loaded as-is in raw format. Based on requirements, data can be transformed and stored as well.

5) Data Processing and Analytics: Consolidate data sources based on business needs. This is a step to process unstructured or semi-structured data to structured data so that it can be used in analytics. It includes aggregating data, generating key performance metrics and analytical data models [7].

5. BIG DATA LAKE

Data Lake is an excellent choice for data consolidation and for integrating data silos. It is an integral part of Big Data Integration. It offers several benefits over the traditional data warehouse. Data lake supports the strategy to develop the universal culture of scalable data-driven insights. They are highly flexible, agile and reusable. As they are designed for low-cost storage, they are commonly used to store historical data from various sources [15].

When you have all data you need at one place, you can use it for all purposes. Instead of storing only raw data, data lake should also store processed data, metrics and aggregations to take full advantage. In order to avoid extra processing and time, common aggregations and metrics should be kept ready in the data lake. That way it can be an enterprise data platform serving as a single source of truth.

As the journey to the big data lake is not short, be stick to goals identified during data discovery. This journey will definitely open doors for the new opportunities and will help enterprises to subjugate data silos.

6. CONCLUSIONS

Today data are being generated in hundreds or thousands from enterprise applications at an unprecedented scale. As these applications are used by different business verticals and teams, data sharing is not easy even within the specific business vertical. As a result, multiple data silos are formed in enterprises. Data can generate enormous value if integrated, analyzed and presented correctly. This paper presented some challenges with data silos and how big data integration can help to tackle them. It discussed the use of Big Data Lake to integrate data silos and how it can serve multiple needs through a single platform. As big data integration is evolving day by day, new frameworks, platforms, and techniques are expected to make integrations easier in the future.

Additionally, data governance for big data lake and data access control on enterprise data have a huge scope for development in the future.

REFERENCES

[1] David Loshin, "Big Data Analytics", Elsevier, 2013

[2] Xin Luna Dong, Divesh Srivastava, 2013. "Big Data Integration", ICDE conference 2013

[3] Sachchidanand Singh, Nirmala Singh, 2012. "Big Data Analytics", International Conference on Communication, Information & Computing Technology (ICCICT), Oct 19-20, 2012.

[4] P. Bedi, V. Jindal, and A. Gautam, "Beginning with Big Data Simplified," 2014.

[5] D. L. W.H. Inmon, Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault. Amsterdam, Boston: Elsevier, 2014.

[6] J. G. Shanahan and L. Dai, "Large Scale Distributed Data Science Using Apache Spark," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 2323–2324.

[7] A White Paper, 2013. "Aggregation and analytics on Big Data using the Hadoop eco- system"

[8] Comfort LK. Risk, Security, and Disaster Management. Annual Review of Political Science. 2005;8:335–356.

[9] eWEEK. (2019). Managing Massive Unstructured Data Troves: 10 Best Practices. [online] Available at: http://www.eweek.com/storage/slideshows/managing-massiveunstructured-data-troves-10-best-practices#sthash.KAbEigHX.dpuf [Accessed 11 May 2019].

[10] Soumysen, Ranak Ghosh, Debanjali, NabenduChaki, 2012. "Integrating XML Data into Multiple ROLAP Data Warehouse Schemas", International Journal of Software Engineering and Application (USEA), Vol 3, No.1, Jan 2012.

[11] B.arputhamary and L.arockiam. "A Review on Big Data Integration" IJCA Proceedings on International Conference on Advanced Computing and Communication Techniques for High Performance Applications ICACCTHPA 2014(5):21-26, February 2015.

[12] J. Kreps, N. Narkhede, and J. J. Rao, ''Kafka: A distributed messaging system for log processing,'' in Proc. NetDB, 2011, pp. 1–7.

[13] J. Liao, X. Zhuang, R. Fan and X. Peng, "Toward a General Distributed Messaging Framework for Online Transaction Processing Applications," in IEEE Access, vol. 5, pp. 18166-18178, 2017.

[14] Salinas, Sonia Ordonez and Alba C.N. Lemus. (2017) "Data Warehouse and Big Data integration" Int. Journal of Comp. Sci. and Inf. Tech. 9(2): 1-17.

[15] Analytics Magazine, 03-Nov-2016. "Data Lakes: The biggest big data challenges," [Online]. Available at: http://analytics-magazine.org/data-lakes-biggest-big-data-challenges/. [Accessed: 11-May-2019].

[16] Alienor. "What Is a Data Silo and Why Is It Bad for Your Organization?" Plixer. July 31, 2018. Accessed May 11, 2019. https://www.plixer.com/blog/network-security/data-silo-what-is-it-why-is-it-bad/.

[17] "4 Best Ways To Breakdown Data Silos [Problems and Solutions]." Status Guides. February 26, 2019. Accessed May 11, 2019. https://status.net/articles/data-silos-information-silos/.

[18] G. Press, "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," Forbes, 23-Mar-2016. [Online]. Available: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#6d5fd596f637. [Accessed: 11-May-2019].

[19] AmitkumarManekar, and Dr. G. Pradeepinib (2015,May), "A Review On Cloud Based Data Analysis". International Journal on Computer Network And Communications (IJCNC) May 2015,Vol.1 No.1

[20] L. Duggan, J. Dowzard, J. Katupitiya, and K. C. Chan, “A Rapid Deployment Big Data Computing Platform for Cloud Robotics,” International journal of Computer Networks & Communications, vol. 9, no. 6, pp. 77–88, 2017.

[21] Torlone, Riccardo. (2008). Two approaches to the integration of heterogeneous data warehouses. Distributed and Parallel Databases. 23. 69-97. 10.1007/s10619-007-7022-z.

Authors

Jayesh Patel completed his Bachelors of Engineering in Information Technology in 2001 and MBA in Information Systems in 2007. He currently work for Rockstar Games as Senior Data Engineer, focusing on developing data-driven decision-making processes on Big Data Platform. He has successfully built machine learning pipelines and architected big data analytics solutions over the past several years. Additionally, he is a senior member for IEEE.