Challenges and Benefits of Deploying Big Data Storage Solution

Jabrane Kachaoui
Laboratory of Information Technologies and Modeling
Hassan II University, Faculty of Science Ben M'Sik
Casablanca, Morocco
jabrane2005@gmail.com

Abdessamad Belangour
Laboratory of Information Technologies and Modeling
Hassan II University, Faculty of Science Ben M'Sik
Casablanca, Morocco
belangour@gmail.com

ABSTRACT

Since data is at the heart of information systems, new technologies and approaches dealing with storing, processing and analyzing data have proliferated. Data Warehouses are among the most known approaches that tackle data storing and processing. However, they reached their limits in dealing with large quantities of data as those of Big Data. Consequently, a new concept which is an evolution of Data Warehouse known as "Data Lake" is emerging. This paper presents a detailed analysis that compares Data Lake and Data Warehouse key concepts. It sheds lights on the aspects and characteristics for the sake of revealing similarities and differences. It also emphasizes the complementary of the two technologies by showing the most appropriate use case of each of them.

CCS CONCEPTS

• Computing methodologies~Distributed computing methodologies

KEYWORDS

Data Lake, Data Warehouse, Hadoop, NoSQL, Big Data, Distributed databases, Repository, Data Mart, Ad Hoc

ACM Reference format:

Jabrane Kachaoui and Abdessamad Belangour 2019. Challenges and Benefits of Deploying Big Data Storage Solution. In *Proceedings of ACM SMC conference (SMC'19). ACM, Kenitra, Morocco, 5pages.https://doi.org/10.1145/3314074.3314097*

1 INTRODUCTION

The first Business Intelligence (BI) tools enabled creating simple reports and dashboards that give users insight into the current state of the business. Over time, BI tools have evolved to make statistical analysis to help planning and optimization. Today, BI tools are moving from the information world to the prescription of actions to optimize business (for example, what actions lead to the most desirable results, and recommend those actions). In this migration from descriptive analysis to prescriptive analytics, new business demands include analyzing the data effectively to obtain business ideas that can be beneficial [10].

The goal is to go beyond simple questions "what is happening" and "why did it happen" to ones like "what happens in the future" and "here are the recommended actions". This paper is not meant to explain predictive or prescriptive analysis, but about the best way to collect, manage, and prepare data for any type of analysis. Indeed, if we want to perform these transformations by BI tools, the solution is to provide a warehouse or repository of data as wide as possible that may be relevant for analysis, inside and outside the organization, for immediate or future analysis. And that's exactly what a Data Lake provides.

The temptation is often strong to relate the data lake to a classic data warehouse, but the differences between the two are important, and this on several levels. The data lake aims to absorb raw data streams and make them usable by transforming them to satisfy different analysis needs. This is ultimately extremely classic and does not bring anything new to what the "ETL – Data Warehouse – Data Mart" could do. This new approach, however, is different in that it allows you to load the data and then transform it to make it workable. Initiatives around data are very often limited by the difficulties inherent in the collection and ingestion phases in the systems. On this point, the fact of being able to load the data on a platform in an almost crude state, and to quickly iterate to use them is an undeniable advantage. We often talk about an ELT (Extract-Load-Transform) rather than an ETL (Extract-Transform-Load) approach we were used to. Where a data warehouse pushes the data of their origins towards their consumers according to a relatively fixed way where each Data Mart is supposed to satisfy a need, one has here a much greater flexibility. It is indeed up to each consumer to materialize his need and to extract the different source data and then combine them to make sense of them.

Many people think that Data Lakes are just reincarnation of the Data Warehouse. Others have focused on how much better a Data Lake is, while others are standing on the shoreline screaming, "Don't go in! It's not a lake, it's a swamp!» However, is this true behind the new Data Lakes?

To answer to all these questions, in this paper, we will define these strategies, explain the differences, and show that "Data Warehouse

SMC'2019, March 28–29, 2019, KENITRA, Morocco
 ISBN: 978-1-4503-6129-3
https://doi.org/10.1145/3314074.3314097

vs. Data Lake" is no longer the question. The two technologies go hand in hand for a better way of data management.
This paper proceeds in the following manner. Section 2 provides a brief background of previous research about Data Warehouse and Data Lake. Section 3 investigates similarities and differences between the two concepts. Moreover, summarizes the paper and its contributions. Section 4 discusses a new way to deal with data.

2 DATA WAREHOUSE AND DATA LAKE

This section provides a brief background on Data Warehouse and Data Lake concepts.

2.1 Data Warehouse

The Data Warehouse is a database dedicated to the storage of all the data used for decision-making and decision analysis [13]. The Data Warehouse is exclusively reserved for this purpose. It is fed from the production databases using ETL tools (Ex-tract Transform Load).
The Data Warehouse is not just a copy of the production data. It is organized and structured. Father of this concept, Bill Immon in his book "Building the Data Warehouse" (John Wiley and Son 1996) describes it as follows:
"Subject oriented, integrated, nonvolatile, time varying collection of data in support of management decisions."

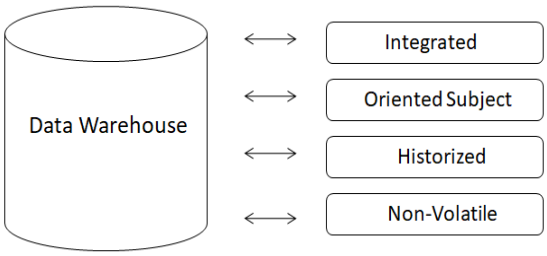

Figure 1: Data Warehouse characteristics

- Oriented Subject: In Data Warehouse, data is organized by theme. Theme-specific data, such as sales, will be returned from the various OLTP production bases and consolidated.
- Integrated: The data come from heterogeneous sources using different types of format. They are integrated before being used.
- Non-volatile: Data do not disappear and do not change over the treatment over time (Read-Only).
- Historized: Non-volatile data is also time stamped. It is possible to visualize the evolution over time of a given value. The degree of detail of archiving is relative to the nature of the data. Not all data deserve being archived.

2.2 Data Lake

Data Lake is fundamentally different from Data warehouse or Data Mart [3] even if all these approaches have the same aim which is storing data [6]. A Data Lake is a data repository for storing a very large amount of raw data in their native format for an indefinite period of time. This storage method makes it easier to coexist between the different schemas and structural forms of data, usually blobs of objects or files [7].

When accessing Data Lakes, users determine:

- Data types and sources.
- Quantity of data.
- Time of getting data.
- Types of analytics.

Are all of these possible in a data warehouse? Probably not. Moreover, even if they were possible, achieving them in a period of time that business users would find acceptable is unlikely – especially in today's rapidly changing environments. Beyond that, one particular schema almost certainly will not fit every business need. To wit, the data may ultimately arrive in a way that renders it virtually useless for the employee's evolving purposes [6].

3 LITERATURE RELATED WORK DISCUSSION

This section presents a literature review and investigates similarities and differences between Data warehouse and Data Lake

3.1 Literature review

Various studies have been conducted on determining the relevant criteria for evaluating and selecting big data storage concepts. For example, Brian Stein and Alan Morrison [1] state that data storage evaluation criterion includes size, low cost, fidelity and ease of accessibility, by believing that Date Lake can be 10 to 100 times less expensive to deploy than conventional data warehousing. P.Tyagi and H.Demirkan [6] believe that Data Lake has a lot of features that added value in terms of flexible data model, connecting external data, Federation and virtualization by choosing which data to copy and which to leave in the source of truth. C. Campbell [7] highlights five key differentiators of a data lake and how they contrast with the data, these five criteria are storage, data, users, scalability and timeliness.

3.2 Comparison Criteria

Based on a comprehensive review of the related literature, twenty one most important criteria are identified for evaluating and selecting big data storage selection.

- **Data:** Data Warehouses are composed of data extracted from transactional systems and only deals with structured data. Non-traditional data sources such as web server logs, sensor data, social networking activities, text and images are largely ignored. New uses of these types of data continue to be found, but consuming and storing them can be expensive and difficult.
  To the contrary, in a Data Lake ALL these types of non-traditional data are included and kept independent of source and structure. They are maintained in their raw forms and transformed whenever we are ready to use them [7].

- **Scalability:** The attempt of scaling within EDW is not easy task because of its rigid and predefined nature. This scaling is done in a way that is compatible with the 3V (Volume, Velocity, Variety) of Big Data. While some EDW implementations are capable of handling volume, velocity is often a technical challenge that required costly solutions (custom hardware, specialized databases, etc.). Variety is also difficult since each piece of content requires new flows and correlative processes to be described in advance. In addition, every new data analysis or use often requires a new Data Mart. In many cases, this scaling takes place outside the EDW environment, in spreadsheets, or in memory tools because the EDW infrastructure is unable to manage it [4].
- **Quality:** Current data warehouse users are often suspicious of data in terms of the origin and the way to use it. Owing to the inaccuracy of such kind of data, users may not trust them to the extent that they worry about the lack of their effectiveness. Therefore, the common response to this is to find out an appropriate way to avoid the Data Warehouse for the sake of getting the data themselves directly from internal or external sources. This invariable leads to multiple conflicting instances of the same data.

  The quality of data requires a clear and easy understanding on the part of the users. The lack of comprehension of the quality of the data makes users lose confidence and creates conflicts. This can also happen in data lake environment.

  The best solution, then, is to find out explicit and clear insight to the quality of data so that users can make a decision whether it is appropriate for the intended purpose [4].
- **Flexibility:** Data Warehouse seems to be ineffective as a source of data for many different reasons. First, the users of this approach face some difficulties to immediate access to obtain the data they need whenever they want it. Second, users find themselves unable to choose tools for their data analysis. Last but not least, data warehouse as a source of data is limited to one type of data to the extent that provides the users with a single local data rather that global ones.This flexibility must be one of the main guiding principles of a new solution that Data Lake offers [5].
- **Timeliness:** Introducing new content to EDW can be a time-consuming and heavy process. When users need immediate access to data, even short processing delays can be frustrating and lead users to avoid using the appropriate processes in favor of getting the data quickly themselves. Users also may waste valuable time and resources to extract the data from operational systems, store and manage it themselves, and then analyze it. Even data professionals fight to find and manipulate the appropriate data [5].

  The Data Lake approach solves this problem through the use of performance-enhancing techniques, such as in-memory storage [4].
- **Search-ability:** With many current EDW solutions, users cannot rapidly and easily search for and find the data they need when they need it. Inability to find data also limits the users' ability to leverage and build on existing data analyses. Advanced analytics users require a data storage solution based on an IT "push" model (not driven by specific analytics projects). Unlike existing solutions, which are specific to one or a small family of use cases, what is needed is a storage solution that enables multiple, varied use cases across the enterprise. What is needed is a simplified landscape where the data is consolidated and the formats are known.

  Data Lake supports multiple reporting tools in a self-serve capacity, allows rapid ingestion of new datasets without extensive modeling, and scales large datasets while delivering performance. It supports advanced analytics, like machine learning and text analytics, and allows users to cleanse and process the data iteratively and to track data for compliance. Users should be able to easily search and explore structured, unstructured, internal, and external data from multiple sources in one secure place. The quality and metadata associated with this information are collected in a way that allows for continuous enhancement and improvement [4].
- **Metadata:** Metadata allows users to find their data in Data Lake and to derive value from it. When ingesting data into a data lake, it is essential to automate the application of metadata. It exits three types of metadata in order to have a more complete picture of data.
  - Technical: helps to capture the type of data (text, JSON…) and the structure (fields with their types) of each data set.
  - Operational: helps to capture provenance and target location of data, size and number of records.
  - Business: helps to capture all information to the user like descriptions, tags masking rules… [14].
- **Data Governance:** Without the structure and controls to manage and maintain the quality, consistency, and compliance of data, a data lake can rapidly devolve into a data swamp. One of the data lake's principle advantages, the common store and access of data, is also one of its weaknesses. Historically, organizations have used data distribution and technical barriers as a substitute for policy controls and enforcement. Bringing this data together in the lake makes the need for policies and data sharing agreements acute. It also exposes other types of suboptimal behavior, such as information hiding and excessive manual manipulation. Furthermore, every time some new data is brought into the environment, it has the potential to combine with the data already present. This in turn creates the need for new policies and new quality checks. Without a robust governance environment that covers the information in the lake as well as its sources and use cases, it is impossible to leverage the data. Information is hidden, or incomplete, or of unsuitable quality because of missing governance [4].
- **Accessibility:** Even if the data might be available, its value is limited if users are unable to find or understand the data. Again, this is the outcome of having highly automated data governance. Big data is a gigantic expansion of the 3Vs of data. For this to function properly, there must be a robust infrastructure that is easily scaled to deal with the volume, a solid platform for managing the performance of the data and the velocity of change, and comprehensive governance for tracking the attributes that

explode with data variety: quality, lineage, sharing, policies and responsibilities for the data. With-out these three things, the data will not be accessible, because you will either not be able to find what you need, access it fast enough, or be able to find understand and exploit it [4].

- **Heterogeneous Tooling:** In Data Warehouse, Data is accessible such tools as SQL and standardized BI tools. In a data lake, data is accessible throughout programs created by developers, SQL type systems and other methods [2].
- **Maintenance and structural evolution:** The complexity of changing the structure is a major complaint about Data Warehouse. It needs redefine modeling in each specific requirement. However, Data Lake, allows new and unlimited associations. All data are automatically integrated [9].
- **Data security:** Data Warehouse technologies have been around for decades, while Big Data technologies are relatively new. Thus, the ability to secure data in a Data Warehouse is much more mature than in a Data Lake. Furthermore, the security tools used in Data Warehouse can't scale to keep up with big data, or just won't work at all [14]. The data loaded in a Data Lake is without any supervision and this can lead to compliance risks. [11].
- **Size:** With Data Lake based on Hadoop, data volumes are petabytes scale while Data Warehouse does not exceed Teraoctects [1]. According to Azure Data Warehouse, the maximum size of the database is 240 TB compressed on disk, this space is independent of tempdb or log space. Therefore, this space is dedicated to permanent tables. The columnstore cluster compression is estimated at 5X. This compression allows the database to reach a volume of about 1Pb when all tables are clustered columnstore (the default table type) [15].
- **Storage:** Data Among the advantages of Big Data technologies like Hadoop is the storage cost that is almost low compared to the Data warehouse. This is because Hadoop is an open source software so the license and community support is free. In addition, Hadoop is designed to be installed on basic hardware at a low price (the cost of deploying a data lake solution is 10 to 100 times less than the data warehouse) [1].
- **Ad hoc analysis:** Supporting Ad hoc analysis is something that operational systems are usually not able to do without a negative effect on performance. By providing parallelism, the Business Data Lake architecture overcomes these constraints and allows for ad hoc analysis as needed [15].
- **Loyalty:** Because Hadoop-based Data lakes preserve data in their original form and record changes to data and contextual semantics throughout the data lifecycle, compliance and internal auditing is easier. Maintain even after transformations, aggregations and updates [1].
- **Multiple sources:** Assembling data from multiple sources takes a long time to obtain this data across a wide range of different analytical platforms. The Business Data Lake approach provides a unique analytical environment in which data from multiple systems is ingested.
- **Data objectivity:** In Data Warehouse, data represents a certain vision and priorities of the company. The data is selected and processed. However, In Data Lake, no values associated with prior data. The data remains unchanged: preservation of information fidelity [2].
- **Intelligence and modeling:** In data warehouse, Modeling is necessary upstream (ETL, structuring). Intelligence is before the request. This case is ideal if the needs are known in advance. In Data Lake, there is no modeling before ingesting data; identification of data is done by metadata. Intelligence is being realized by rendering algorithms. It is Ideal when data is not clearly identified.
- **Deployment experience:** Data Warehouse exists since decades, companies have been able to overcome the problems they had before and have a clear understanding on the term deployment. However, Data Lake is a new technology, companies are discovering Hadoop for the first time and undertake the development of their Data Lake without a real experience, right skill set and knowledge, including Java, Hadoop, and other NoSQL data sources, they are faced with problems of deployment slowness and implementation deficiency. Then, objectives become obsolete [18].
- **Users:** Data warehouse has been around for decades so they are accessible to all users regardless of their technical abilities. On another hand, Data Lake is a relatively new concept, it requires extensive competences to be effective, and that's why it is better for data scientists. [17].

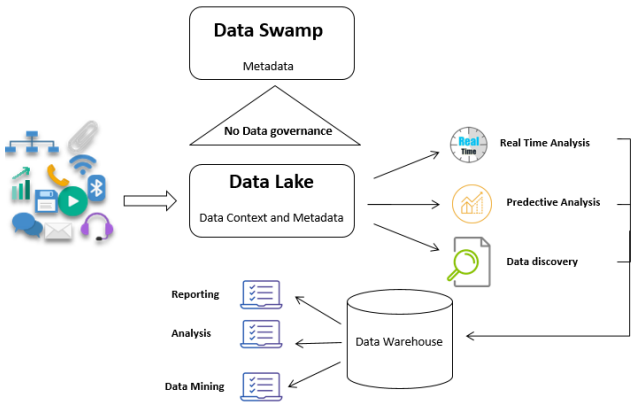

Figure 2: Complementarity of Data Lake and Data Warehouse architecture

4 DISCUSSION

It is not possible to consider in the current state that Data Lake can definitively replace the Data Warehouse. Reliable, standardized and accessible reporting requirements to a large population of users are well represented in business and cannot be handled satisfactorily by another architecture than Data Warehouse. Nevertheless, the setting up of a Data Lake, which collects and stores all data, offers new possibilities of analysis. Data are less guided and are accessible to a limited user (data scientists). However, they can be more responsive and more open to anticipating situations and valuing the company's capital [8].

Complementing an existing enterprise data warehouse (DW) with a data lake can be a complex but smart and benefit step for companies in order to enjoy a reduction in storage costs. It gives more flexibility and speed in terms of data processing and capturing unstructured, semi-structured and streaming data, and frees up bandwidth in data warehouse for business intelligence analytics [14].

5 CONCLUSION AND FUTURE WORK

Deciding to implement a Big Data storage solution provides different approaches to data analysis and usage. Which concepts to use and when depends upon some planning ahead of time [16]. The speed of technological innovation is still evolving a lot. We are probably living today the same revolution that we experienced 30 years ago with the democratization of the relational database. However, that does not mean that a data lake will replace data warehouse. Contrariwise, combining both technologies will work best.

Our future works will aim to explore the Data Lake population, the usage, the data gravity influence, the role and the tools around but also how the real world is adopting and implementing this information architecture evolution:

• Data Lake challenges:
  - Data Governance.
  - Metadata Management and Enhancement.
  - Data gravity influence.

• Role and impact regarding the data lake implementations:
  - The Architecture impacts.
  - The Interaction between Data Lake and the decision support Systems
  - The IT infrastructure role into the data lake solution
  - Data gravity influence

REFERENCES

[1] Stein, Brian and M.Alan, "The enterprise data lake: Better integration and deeper analytics," PWC Technology Forecast, 2014..
[2] Teradata and Hortonworks, "Putting the Data Lake to Work: A Guide to Best Practices," CITO, 2014.
[3] FDew Atkins, "Agile Business Intelligence Data Lake Architecture," 2014.
[4] Collibra and Knowledgent, "Driving Data Agility with the Data Lake," 2016.
[5] Knowledgent, "How to Design a Successful Data Lake," 2014.
[6] P.Tyagi and H.Demirkan, "Data Lakes: The biggest big data challenges Why data lakes are an important piece of the overall big data strategy," 2017.
[7] C.Campbell, "Top Five Differences between Data Lakes and Data Warehouses," 2015.
[8] M.MONESTIER, "Le Data Lake est-il le nouveau Data Warehouse ? ," 2017.
[9] O. Boussaid, " Gestion des données massives," 2017.
[10] F.Velez, S.Agrawal, "Data Lake: Discovering, Governing and Transforming Raw Data into Strategic Data Assets," 2016.
[11] P.Russom, "Emerging Best Practices for Data Lakes," 2016.
[12] Cambridge Semantics, "Anzo Smart Data Lake™ Enterprise Graph-Based Data Discovery, Analytics and Governance," 2015.
[13] C.Paschalidi, "Data Governance: A conceptual framework in order to prevent your Data Lake from becoming a Data Swamp," 2015.
[14] Zaloni, "Why Your Data Warehouse Needs a Data Lake and How to Make Them Work To-gether ," 2011.
[15] Microsoft, "SQL Data Warehouse capacity limits," 2018.
[16] M.Knight, "Data Warehouse vs. Data Lake Technology: Different Approaches to Man-aging Data," 2017.
[17] M.Peterman, "Data Warehouse vs Data Lake: The Key Differences Every Organization Should Know," 2017.
[18] A.Nadkarni, "Data Analytics Infrastructure and the Essential Data Lake: A Global Study,"2017.