Technology Trends

Datalakes

Enterprise Architecture, Chief Technology Officer Branch

Version 0.1
Date 2019-7-12

Shared Services Canada
Services partagés Canada

Canada

Table of Contents

Business Brief **3**
Technical Brief **3**
Industry Use **5**
Canadian Government Use **5**
Implications for Shared Services Canada (SSC) **6**
Value Proposition 6
Challenges 8
Considerations 10
References **12**

Business Brief

In an ever-increasing hyperconnected world, corporations and businesses are struggling to deal with the responsibilities of storage, management and quick availability of raw data. To break these data challenges down further:

- Data comes in many different structures.
    - Unstructured
    - Semi-Structured
    - Structured
- Data comes from many disparate sources.
    - Enterprise Applications
    - Raw Files
    - Operation and Security Logs
    - Financial Transactions
    - Internet of Things (IoT) Devices & Network Sensors
    - Websites
    - Scientific research.
- Data sources are often geographically distributed to multiple locations
    - Datacenters
    - Remote Offices
    - Mobile Devices

In an effort to resolve these data challenges, a new way of managing data was created which drove data oriented companies to invent a new data storage mechanism called a Data Lake.

Data Lakes are characterized as:

- Collect everything.
    - A Data Lake contains all data; raw sources over extended periods of time as well as any processed data.
- Dive in anywhere.
    - A Data Lake enables users across multiple business units to refine, explore and enrich data on their terms.
- Flexible access.
    - A Data Lake enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines."

Data Lakes are essentially a technology platform for holding data. Their value to the business is only realized when applying data science skills to the lake.

To summarize, usecases for Data Lakes are still being discovered. Cloud providers are making it easier to procure Data Lakes and today Data Lakes are primarily used by Research Institutions, Financial Services, Telecom, Media, Retail, Manufacturing, Healthcare, Pharma, Oi I& Gas and Governments.

Technical Brief

The most popular implementation of a Data Lake is through the open source platform called Apache Hadoop. Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. Hadoop was originally created by researchers at Google as a storage method to handle the indexing of websites on the Internet; At that time it was called the Google File System.

A Data Lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions."

Data can flow into the Data Lake by either batch processing or real-time processing of streaming data. Additionally, data itself is no longer restrained by initial schema decisions and can be exploited more freely by the enterprise. Rising above this repository is a set of capabilities that allow IT to provide Data and Analytics as a Service (DAaaS), in a supply-demand model. IT takes the role of the data provider (supplier), while business users (data scientists, business analysts) are consumers.

The DAaaS model enables users to self-serve their data and analytic needs. Users browse the lake's data catalog (a Datapedia) to find and select the available data and fill a metaphorical "shopping cart" (effectively an analytics sandbox) with data to work with. Once access is provisioned, users can use the analytics tools of their choice to develop models and gain insights. Subsequently, users can publish analytical models or push refined or transformed data back into the Data Lake to share with the larger community.

Although provisioning an analytic sandbox is a primary use, the Data Lake also has other applications. For example, the Data Lake can also be used to ingest raw data, curate the data, and apply Export-Transform-Load (ETL). This data can then be loaded to an Enterprise Data Warehouse. To take advantage of the flexibility provided by the Data Lake, organizations need to customize and configure the Data Lake to their specific requirements and domains.

Industry Use

There are a variety of ways Data Lakes are being used in the industry:

- **Ingestion of semi-structured and unstructured data sources (aka big data**) such as equipment readings, telemetry data, logs, streaming data, and so forth. A Data Lake is a great solution for storing IoT (Internet of Things) type of data which has traditionally been more difficult to store, and can support near real-time analysis. Optionally, you can also add structured data (i.e., extracted from a relational data source) to a Data Lake if your objective is a single repository of all data to be available via the lake.
- **Experimental analysis** of data before its value or purpose has been fully defined. Agility is important for every business these days, so a Data Lake can play an important role in "proof of value" type of situations because of the "ELT" approach discussed above.
- **Advanced analytics support**. A Data Lake is useful for data scientists and analysts to provision and experiment with data.
- **Archival and historical data storage.** Sometimes data is used infrequently, but does need to be available for analysis. A Data Lake strategy can be very valuable to support an active archive strategy.
- **Distributed processing** capabilities associated with a logical data warehouse.

How TD Bank Made Its Data Lake More Usable
https://www.datanami.com/2017/10/03/td-bank-made-data-lake-usable/
Toronto-Dominion Bank (TD Bank) is one of the largest banks in North America, with 85,000 employees, more than 2,400 locations between Canada and the United States, and assets nearing $1 trillion. In 2014, the company decided to standardize how it warehouses data for various business intelligence and regulatory reporting functions. The company purchased a Hadoop distribution and set off to build a large cluster that could function as a centralized lake to store data originating from a variety of departments.

Canadian Government Use

In 2019, the Treasury Board of Canada Secretariat (TBS), partnered with Shared Services Canada and other departments, to identify a business lead to develop a Data Lake (a repository of raw data) service strategy so that the GC can take advantage of big data and market innovation to foster better analytics and promote horizontal data-sharing.[i]

Big data is the technology that stores and processes data and information in datasets that are so large or complex that traditional data processing applications can't analyze them. Big data can make available almost limitless amounts of information, improving data-driven decision-making and expanding open data initiatives. Business intelligence involves creating, aggregating, analyzing and visualizing data to inform and facilitate business management and strategy. TBS, working with departments, will lead the development of requirements for an enterprise analytics platform.[ii]

Data Lake development in the GC is a more recent initiative. This is mainly due to the GC focussing resources on the implementation of cloud initiatives. However, there are some GC departments engaged in developing Data Lake environments in tandem to cloud initiatives.

Notably, the Employment and Social Development Canada (ESDC) is preparing the installment of multiple Data Lakes in order to enable a Data Lake Ecosystem and Data Analytics and Machine Learning toolset. This will enable ESDC to share information horizontally both effectively and safely, while enabling a wide variety of data analytics capabilities. ESDC aims to maintain current data and analytics capabilities up-to-date while exploring new ones to mitigate gaps and continuously evolve our services to meet client's needs.[iii]

Implications for Shared Services Canada (SSC)

Value Proposition

There are three common value propositions for pursuing Data Lakes. 1) It can provide an easy and accessible way to obtain data faster; 2) It can create a singular inflow point of data to help connect and merge information silos in an organization; and 3) It can provide an experimental environment for experienced data scientists to enable new analytical insights.

Data Lakes can provide data to consumers more quickly by offering data in a more raw and easily accessible form. Data is stored in its native form with little to no processing, it is optimized to store vast amounts of data in their native formats. By allowing the data to remain in its native format, a much timelier stream of data is available for unlimited queries and analysis. A Data Lake can help data consumers bypass strict data retrieval and data structured applications such as a data warehouse and/or data mart. This has the effect of improving a business' data flexibility. Some companies have in fact used Data Lakes to replace existing warehousing environments where implementing a new data warehouse is more cost prohibitive. A Data Lake can contain unrefined data, this is helpful when either a business data structure is unknown, or when a data consumer requires access to the data quickly.

A Data Lake is not a single source of truth. A Data Lake is a central location in which data converges from all data sources and is stored, regardless of the data formatting. As a singular point for the inflow of data, sections of a business can pool their information together in the Data Lake and increase the sharing of information with other parts of the organization. In this way everyone in the organization has access to the data. A Data Lake can increase the horizontal data sharing within an organization by creating this singular data inflow point. Using a variety of storage and processing tools analysts can extract data value quickly in order to inform key business decisions.

A Data Lake is optimized for exploration and provides an experimental environment for experienced data scientists to uncover new insights from data. Analysts can overlay context on the data to extract value. All organizations want to increase analytics and operational agility. The Data Lake architectural approach can store large volumes of data, this can be a way in which cross-cutting teams can pool their data in a central location and by complementing their systems of record with systems of insight. Data Lakes present the most potential benefits for experienced and competant data scientists. Having structured, unstructured and semistructured data, usually in the same data set, can contain business, predictive, and prescriptive insights previously not possible from a structured platform as observed in data warehouses and data marts.

Challenges

Although Data Lake technology has many benefits for organizations dealing with big data it has its own challenges. For example:

Data Governance and Semantic Issues

The biggest challenge for Data Lakes is to resolve assorted data governance requirements in a single centralized data platform. Data Lakes fail mostly when they lack governance, self-disciplined users, and a rational data flow. Often, Data Lake implementations are focused on storing data instead of managing the data. Data Lakes are not optimized for semantic enforcement or consistency. They are made for semantic flexibility, to allow anyone to provide context to data if they have the skills to do so.

Putting data in the same place does not remove it's ambiguity or meaning. Data Lakes provide unconstrained, "no compromises" storage model environment without the data governance assurances common to data warehouses or data marts. Proper meta data is essential for a Data Lake, without appropriate meta data the Data Lake will not work as intended. It is beneficial to think of meta data as the fish finder in the Data Lake.

Lack of Quality and Trust in Data

Data quality and trust in the data is a perennial issue for many organizations. Although data discovery tools can apply Machine Learning across related datasets from multiple data sources to identify anomalies (incorrect values, missing values, duplicates and outdated data), quality and trustworthiness of data continue to be an issue for Data Lakes who can easily become data dumping grounds. Some data is more accurate than others. This can present a real problem for anyone using multiple data sets and making decisions based upon analysis conducted with data of varying degrees of quality.

Data Swamps, Performance, and Flexibility Challenges

Data stored in Data Lakes can sometimes become *muddy* when good data is mixed with bad data. Data Lake infrastructure is meant to store and process large amounts of data, usually in massive data files. A Data Lake is not optimized for a high number of users or diverse and simultaneous workloads due to intensive query tasks. This can result in performance degradation and failures are common when running extractions, transformations, and loading tasks all at the same time. On-premises Data Lakes face other performance challenges in that they have a static configuration.

Data Hoarding and Storage Capacity

Data stored in Data Lakes may actually never be used in production and stay unused indefinitely in the Data Lake. By storing massive amounts of historical data, the infinite Data Lake may skew analysis with data that is no longer relevant to the priorities of the business. In keeping the historical data the metadata describing it must be understood as well. This decreases the performance of the Data Lake by increasing the overall workload of employees to clean the datasets no longer in use for analysis.

Storing increasingly massive amounts of data for an unlimited time will also lead to scalability and cost challenges. Scalability challenges are less of a risk in public cloud environments, but cost remains a factor. On-premises Data Lakes are more susceptible to cost challenges. This is because their cluster nodes require all three dimensions of computing (storage, memory and processing). Organizations of all kinds generate massive amounts of data (including meta data) and it is increasing exponentially.

The storage capacity of all this data (and future data) will be an ongoing challenge and one that will require constant management. While Data Lakes can and will be stored on the cloud, SSC as cloud broker for the GC will need to provide the appropriate infrastructure and scalability to clients.

Advanced Users Required

Data Lakes are not a platform to be explored by everyone. Data Lakes present an unrefined view of data that usually only the most highly skilled analysts are able to explore and engage in data refinement independent of any other formal system-of-record such as a data warehouse.

Not just anyone in an organization is data-literate enough to derive value from large amounts of raw or uncurated data. The reality is only a handful of staff are skilled enough to navigate a Data Lake. Since Data Lakes store raw data their business value is entirely determined by the skills of Data Lake users. These skills are often lacking in an organization.

Data Security

Data in a Data Lake lacks standard security protection with a relational database management system or an enterprise database. In practice, this means that the data is unencrypted and lacks access control. Security is not just a binary solution. We have varying degrees of security (unclassified, secret, top secret, etc.) and all of which require different approaches. This will inevitably present challenges with the successful use of data from Data Lakes.To combat this, organizations will have to embrace a new security framework to be compatable with Data Lakes and Data Scientists.

Considerations

Shared Services Canada (SSC) has an excellent opportunity to capitalize on its mandate of providing data storage service to GC's other departments. SSC, as the GC's Service Provider, could potentially a centralized GC Data Lake and allow GC Data Scientists access to this central data using a single unified Data Lake interface. However, this is a project which should be implemented after cloud has been adopted and enterprise data centers have been migrated to in order to provide adequate infrastructure and scaling.

Data Lakes should not be confused for conventional databases although they both store information. A Data Lake will always underperform when tasked with the jobs of a conventional database. To combat this, SSC must create data architectures that define the proper application of Data Lakes. Too often, Data Lakes suffer from lack of foresight on what they're supposed to achieve. Creating a Data Lake becomes the goal rather than achieving a strategic objective. Shared Services Canada (SSC) should consider designing Data Lake infrastructure around Service-Level Agreements (SLA) to keep Data Lake efforts on track. This includes ensuring that SSC has established clear goals for Data Lakes prior to deployment.

SSC should also consider building an expert special group focussed on advanced analytics and experimental data trend discovery in Data Lakes. While the fundamental assumption behind the Data Lake concept is that everyone accessing a Data Lake is moderately to highly skilled at data manipulation and analysis, the reality is most are not. SSC should consider significant investment in training employees necessary skills, such as Data Science, Artificial Intelligence, Machine Learning, or Data Engineering.

SSC should be cognisant that there are significant overinflated expectations revolving around Data Lakes. Inflated expectations lead to vague and ambiguous use cases and increased chances of catastrophic failures. As a Service Provider, SSC must be strict in establishing clear goals for Data Lake provision efforts before deployment. SSC, should be wary of attempts to replace strategy development with infrastructure. A Data Lake can be a technology component that supports a data and analytics strategy, but it cannot replace that strategy.

SSC should be concerned with the provision and running of the infrastructure, the departments themselves are responsible for the data they put in the Data Lake. However, as a Service Provider, SSC should monitor the Data Lake with regards to data governance, data lifecycle for data hygiene, and what is happening in the Data Lake overall. Depending on technology, SSC will need to be very clear on how to monitor activities in the Data Lakes it provides to the GC.

SSC should consider a Data Lake implementation project as a way to introduce or reinvigorate a data management program by positioning data management capabilities as a prerequisite for a

successful Data Lake. Data will need to be qualified before it hits the data lake, this can and should be done in a system of record first. In this way the data can be organizedto fit into the Data Lake implementation.

SSC should create policies on how data is managed and cleaned in the Data Lake. Automated data governance technologies should be added to support advanced analytics. Standardizing on a specific type of governance tool is an issue which must be resolved. Additionally, planning for effective metadata management, considering metadata discovery, cataloguing and enterprise metadata management applied to Data Lake implementation is vital. Rigorous application of data discipline and data hygiene is needed. To combat this, SSC should use data management tools and create policies on how data is managed and cleaned in the Data Lake. The majority of Data Lake analysts will prefer to work with clean, enriched, and trusted data. However, data quality is relative to the task at hand. Lowquality data may be acceptable for low-impact analysis or distant forecasting, but unacceptable for tactical or high-impact analysis. SSC assessments should take this into account.

Design Data Lakes with the elements necessary to deliver reliable analytical results to a variety of
data consumers. The goal is to increase cross-business usage in order to deliver advanced analytical insights. Build Data Lakes for specific business units or analytics applications, rather than try to implement some vague notion of a single enterprise Data Lake. However, alternative architectures, like data hubs, are often better fits for sharing data within an organization.

References

https://www.dataversity.net/data-lakes-101-overview/#

https://www.youtube.com/watch?v=v3yv88h68GY

https://www.pcmag.com/article/347020/data-lakes-explained

https://hortonworks.com/blog/enterprise-hadoop-journey-data-lake/

https://en.wikipedia.org/wiki/Google_File_System

https://www.sqlchick.com/entries/2016/7/31/data-lake-use-cases-and-planning

https://www.sqlchick.com/entries/2016/7/31/data-lake-use-cases-and-planning

https://dataconomy.com/2018/07/six-reasons-to-think-twice-about-your-data-lake-strategy/

https://www.arcadiadata.com/blog/the-top-six-reasons-data-lakes-have-failed-to-live-up-to-expectations/

https://aws.amazon.com/solutions/data-lake-solution/

---

[i] Treasury Board of Canada Secretariat. (March 29th, 2019). *Digital Operations Strategic Plan: 2018-2022*. Government of Canada. Treasury Board of Canada Secretariat. Retrieved 26-May-2019 from: https://www.canada.ca/en/government/system/digital-government/digital-operations-strategic-plan-2018-2022.html

[ii] Ibid.

[iii] Brisson, Yannick, and Craig, Sheila. (November, 2018). *ESDC Data Lake – Implementation Strategy and Roadmap* Update. Government of Canada. Employment and Social Development Canada – Data and Analytics Services. Presentation. Last Modified on 2019-04-26 15:45. Retrieved 07-May-2019 from GCDocs: https://gcdocs.gc.ca/ssc-spc/llisapi.dll?func=ll&objaction=overview&objid=36624914