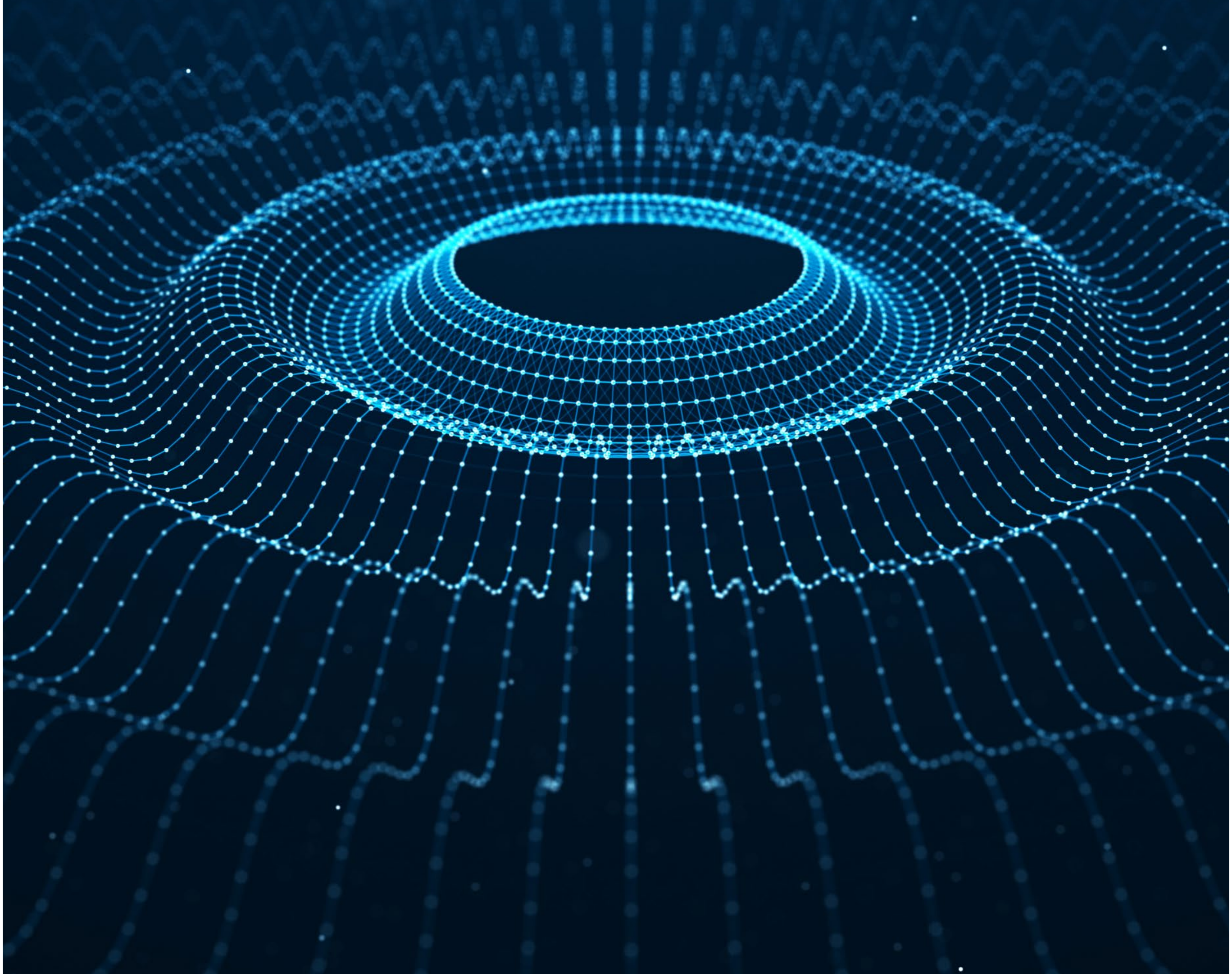
Synechron

Rethinking Data Lakes and Data Warehouses

Authored by:

Ravindran Narayanasamy

Managing Consultant - Data Management
& Engineering Practice
Synechron Business Consulting,
The Netherlands



Keeping up with technology to the likes of Google and Netflix in managing and consuming data for large financial organizations

Abstract

Poor data in terms of sourcing, management, or quality, challenges the operational efficiency of any organization. While Financial Institutions (FIs) continuously strive to build and improve data warehouses and data lakes within their finance and risk functions to gain control over data, they are typically behind the curve in terms of leveraging the recent technology trends in sourcing and managing their data compared to tech giants such as Google, Facebook, or Netflix. In this paper we make a case for stepping up the technology game for FIs in centralizing the sourcing, control, and management of data.

Introduction

Google has killed over 229 projects[4], with the longest project being almost 17 years old, and the youngest being barely 29 days old. It is either adapt or die. It is commonplace, rather the cost of running a business to scrap old technology to keep up with the times, however expensive the original investment might have been.

On that note, data lake programs have been one of the major technology investments of all large banks in the past five years. Yet, increasingly data lakes are being referred to as data swamps. With the constant changes due to regulatory requirements, evolving business and technology landscape, large datasets being duplicated into data lakes become outdated before they can reach the staging layer.

It was in 2010, James Dixon, the CEO of Pentaho, coined the word data lake with the definition given here, to the idea of centralizing all data within an organization into a Hadoop Distributed File System to harness the power of data. He even registered this in his personal blog[1].

“ ”

If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.

James Dixon,
CEO - Pentaho, 2010

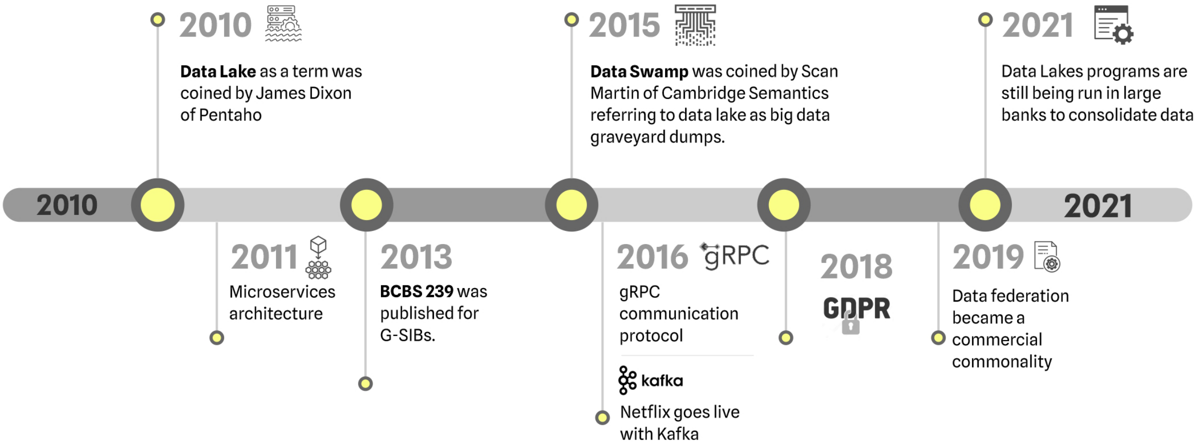
Shortcomings of **traditional data lakes**

The day-to-day problems with this model include space constraints due to data duplication especially with cloud pricing models, constant need for synchronization with the original data source, inability to keep-up with (data structure) change in upstream systems, lack of security & governance and threat to data ownership. This overall failure of the data lakes as a concept was registered early on by Sean Martin in 2015.

And since the introduction of data Lakes, we have had new regulations such as BCBS 239 (2013) and the General Data Protection Regulation (2018) which go against the data sharing model instigated by data lakes and set requirements for much higher governance and stringent control over data.

This brings us to consider data virtualization as a viable alternative to creating data lakes.

> We see customers creating big data graveyards, dumping everything into Hadoop Distributed File System (HDFS) and hoping to do something with it down the road. But then they just lose track of what's there. The main challenge is not creating a data lake but taking advantage of the opportunities it presents.
>
> **Sean Martin, Cambridge Semantics, 2015**[3]


2010
Data Lake as a term was coined by James Dixon of Pentaho
2015
Data Swamp was coined by Scan Martin of Cambridge Semantics referring to data lake as big data graveyard dumps.
2021
Data Lakes programs are still being run in large banks to consolidate data
2010
2021
2011
Microservices architecture
2013
BCBS 239 was published for G-SIBs.
2016 gRPC
gRPC communication protocol
kafka
Netflix goes live with Kafka
2018
GDPR
2019
Data federation became a commercial commonality

Data virtualization / **virtual data lake**

Data virtualization / virtual data lake is essentially an extension of data federation. Data federation is where the data stored in a heterogeneous set of autonomous data stores are made accessible to data consumers as one integrated data store by using on-demand data integration[9]. Data federation is normally used to access and merge already cleansed and conformed data in real time and neither a data lake nor a virtual data lake, replace the need for data warehouses which are typically to generate specific regulatory reports and other financial reports[7]. However, consumption of data into these data warehouses typically happen via the data lake as per current data lake strategy in most organizations.

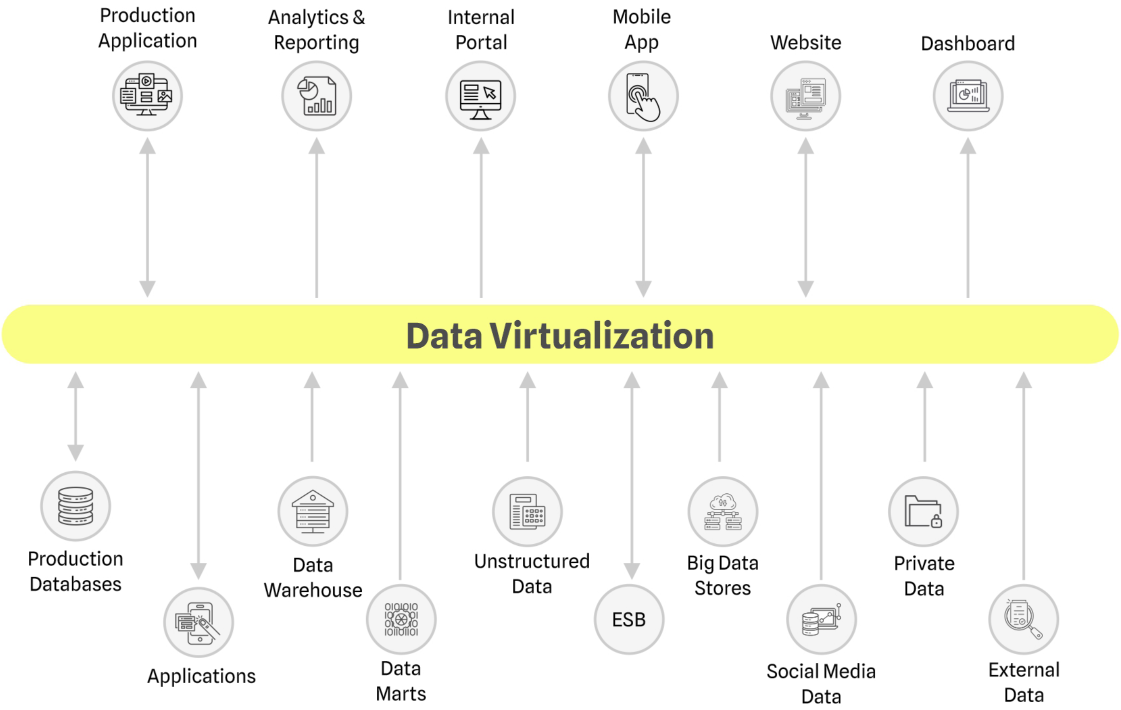


Figure 3 Data virtualization is on abstraction layer that decouples data consumers from data stores.

Data virtualization on the other hand must do the heavy lifting of accessing, profiling, cleansing, transforming, establishing control, data governance and delivering the federated data to and from any application, on-demand. It must handle all the underlying data complexity in order to provide conformed and trusted data, reusing the logic for either batch or real-time operation, whether through SQL, SOA, REST, JSON, or new acronyms yet to be specified[9].

A virtual data lake will allow the user to access data directly from the golden source, across multiple data types, location, or format without having to do any data duplication or moving of data. This will allow the resources to be allocated towards building data governance and control over data instead of throwing in effort to duplicate data in the Hadoop Distributed File System.

Data virtualization / data federation / virtual data lakes have been prevalent as early as the data lakes themselves. However, back in 2010, data lakes might have been the viable option to build a central data repository. Since 2010, Microservices was termed as a style of architecture with Netflix and Amazon being the pioneers, network and processor speeds have literally become a million times faster[10], gRPC and Kafka communication protocols have been tested and commercialized for large data transfers and long messaging queues by Netflix[12] and Uber. Taking these advancements into account, data virtualization becomes the obvious choice to replace data lakes and automatically avoid the shortcomings of a data lake.

	2010	2020
Network Speed	20 Mbps	200 Gbps
Processor Speed	515 Gigaflops	415.5 Petaflops

*A million times faster

Tangible benefits of data virtualization in terms of **data delivery**

To summarize the benefits of a virtual data lake:

- There is clearly the space saved in storage requirements of duplicated data
- Ability to enforce control & governance
- Less resources are consumed by avoiding unnecessary ETL operations
- Data is always up-to-date and accurate
- Ability to focus the resources to categorize and catalogue data
- It is easier to replace legacy systems by mapping to the established logical data model of the virtual data lake

On top of the logistical benefits listed above, data virtualization surpasses every listed business requirement of data delivery[13] as observed below, making data virtualization the obvious choice.

Business requirements of data delivery:

- **Available –** Data must be available when business users need it, regardless of whether it's new or old, simple or complex, or detailed or aggregated. It should not be hidden from business users, nor should its existence be unknown
- **Integrated –** Data from multiple systems must be combined and presented as if it comes from one system
- **Consistent –** Data must be consistent across different reports or dashboards
- **Correct –** Data must be in line with reality; incorrect data can lead to incorrect business decisions
- **Timely –** Data must be available to business users at the right time, even if this means seconds after the data was produced. Data received too late, may have no value
- **Instant –** Data must be available the instant business users ask for it, even when the data request is unique and completely ad-hoc. It should not take days before it becomes available
- **Documented –** Data must be documented and explained using descriptive metadata; without metadata it may have no business value
- **Trusted –** Business users must trust the data with which they make their decisions, and they need the data processing to be as transparent as possible
- **Actionable –** Data must be shaped, filtered, and aggregated so that only relevant data is presented and is aimed at taking business actions
- **Adaptable –** Data must be able to adapt when the business changes and when data definitions and data processing regulations change

Conclusion

We still need the conventional data lakes to do some of the heavy lifting big data operations which can only be done with Spark, HIVE and such data lake infrastructures on the Hadoop file systems, however data virtualization can help this big data be sourced and governed better as well on top of making the day-to-day operational data more accessible, with much lower time-to-market, to the business users as much as to the data scientists. A gap assessment to this end and a course correction to adapt better technology to make data accessible to one and all, can never go wrong.

Sparked your **Interest?**

At Synechron we are more than happy to explain how your organization can benefit from our expertise in data virtualization. We look forward to helping you with gradually moving to virtual data lakes from your existing data lake programs without affecting the existing processes and a balanced investment.

About the **Author**

Ravindran Narayanasamy

Managing Consultant - Data Management
& Engineering Practice
Synechron Business Consulting,
The Netherlands

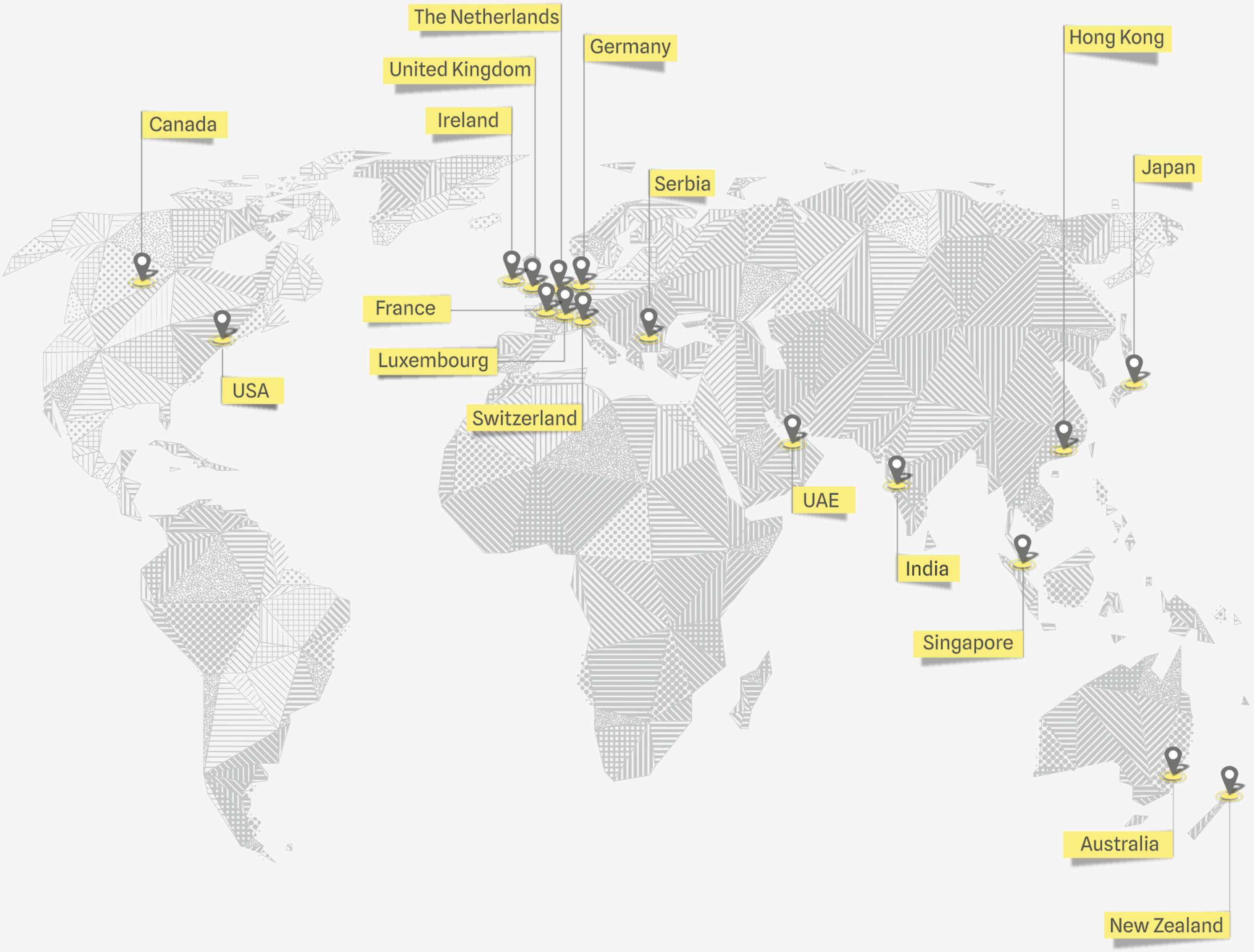
Want to learn more?

Contact Ravindran at:
ravindran.narayanasamy@synechron.com

References

1 .https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
2."NVIDIA Tesla Personal Supercomputer". Nvidia.com. Retrieved February 9, 2012.
3.https://en.wikipedia.org/wiki/Data_lake#cite_note-stein2014-7
4.https://killedbygoogle.com/
5.https://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf
6.https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/?sh=548ace071157
7.https://www.bluegranite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses
8.https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/
9.https://www.sciencedirect.com/topics/computer-science/data-federation
10.https://en.wikipedia.org/wiki/FLOPS#cite_note-66
11.https://www.gartner.com/en/documents/3500219/tableau-adds-data-federation-and-advanced-analytics-but-
12.https://netflixtechblog.com/evolution-of-the-netflix-data-pipeline-da246ca36905
13. The Business Benefits of Data Virtualization - A Whitepaper by Rick F. van der Lans - Independent Business Intelligence Analyst

Global Footprint



Synechron

www.synechron.com | info@synechron.com

Proprietary material

“This material and information is the sole property of Synechron and is intended exclusively for general information purposes. Any rights not expressly granted here are reserved by Synechron. Please note that copying, modification, disclosure of data, distribution or transmission of this material without prior permission of Synechron is strictly prohibited.”