CENTRE FOR FINANCIAL REGULATION AND INNOVATION

White Paper

Creating a Financial Data Lake for Academic Fintech Research.

Daniel Broby[1]* | **Huckleberry Hopper**[2]†

[1]Strathclyde Business School, Glasgow, Scotland

Correspondence
Daniel Broby,
Strathclyde Business School, Stenhouse Wing, 199 Catherderal Street, Glasgow G4 0QU
Email: daniel.broby@strath.ac.uk

Funding information
The University of Strathclyde is a leading international technological university that has made *Fintech* one of its strategic clusters. The creation of a data loch (data lake) is an integral part of this initiative.

This paper presents the case for a Financial Technology (Fintech) data lake. Fintech is impacting business models and its concepts require testing. The definition of Fintech is imprecise, but it is characterized by the use of technology as applied to digital financial transformation. The software and programming driving it is evolving and should be evaluated before being introduced into financial markets. Its development impacts "client money" and this can be risky unless supervised. Fortunately, such experimentation can be done in a controlled way using a regulatory sandbox. This allows Fintech concepts to be checked for reliability and robustness, using consenting live accounts (which receive a special regulatory exception). We propose a less risky supplementary approach, namely the testing of concepts on real but "blinded" financial big data files stored in a data lake. In this way, back-testing, out of sample experiments and forward performance checks can be done without the risk of losing money. We investigate how to implement such a data lake in order to do this.

KEYWORDS
Fintech, Strategy, Business models, Innovation, Financial Services, Disruption, Artificial Intelligence, Taught finance, Fintech Scotland, Data lake, Data warehouse.

* Director, Centre for Financial Regulation and Innovation
2† Researcher, Centre for Financial Regulation and Innovation

1 | INTRODUCTION

A data lake is a research tool. It differs from a data warehouse in its design pattern. It is a central repository of data. It is, however, more than a storage repository. It holds raw data in its native format employing flat file architecture. This is superior to traditional data warehouses where the limiting design factor is the way load patterns are extracted. In effect, the problem is that the data is prepared for specific use and therefore narrowed down for purpose. This limits its usage and functionality. This paper addresses this limitation, exploring instead how data lakes can be used to research academic and commercial Financial Technology (Fintech) concepts.

Our insights are not new. Bharadwaj et al (2013) argued that a data lake is perfect for experimentation on Fintech concepts. This is because, as they explain, data lakes are a hybrid data management solution. As such, they facilitate real-time analytics and the processing of extremely large data volumes. Used in an academic context, data lakes can provide the framework for machine learning and hypothesis testing. Our contribution is in suggesting that a financial data lake can be used to do this in a controlled fashion, in much the same way as a "regulatory sandbox".[1]

We argue that the development of a financial data lake is important to a Fintech ecosystem because it can help address the dynamic and evolving nature of innovation. A disciplined approach is required in order to addresses the challenges that emerge in tandem with rapid change. Figure 1 illustrates how this can be achieved with a data lake located alongside a shared physical space. A data lake can speed up big data ingestion, accelerating analysis and allowing for greater data diversity. It can be used to back-test and perform out of sample experiments.

A data lake concentrates various types of data in one place. They can be used to create a cheap repository for multiple users. Clearly, however, the data needs to be analysed. One consequence of the Fintech revolution is that there is a shortage of the required people with data analytical skills to do this, as noted in Karkkainen et al (2017). This is the result of the rapid growth in big data and the resulting increased need for advanced analysis, especially in view of the continuously innovating business models of the sort highlighted by Stein and Morrison (2014). Not all research, development and testing can be done in house by Fintech firms, meaning that there is a commercial demand for such data repositories.

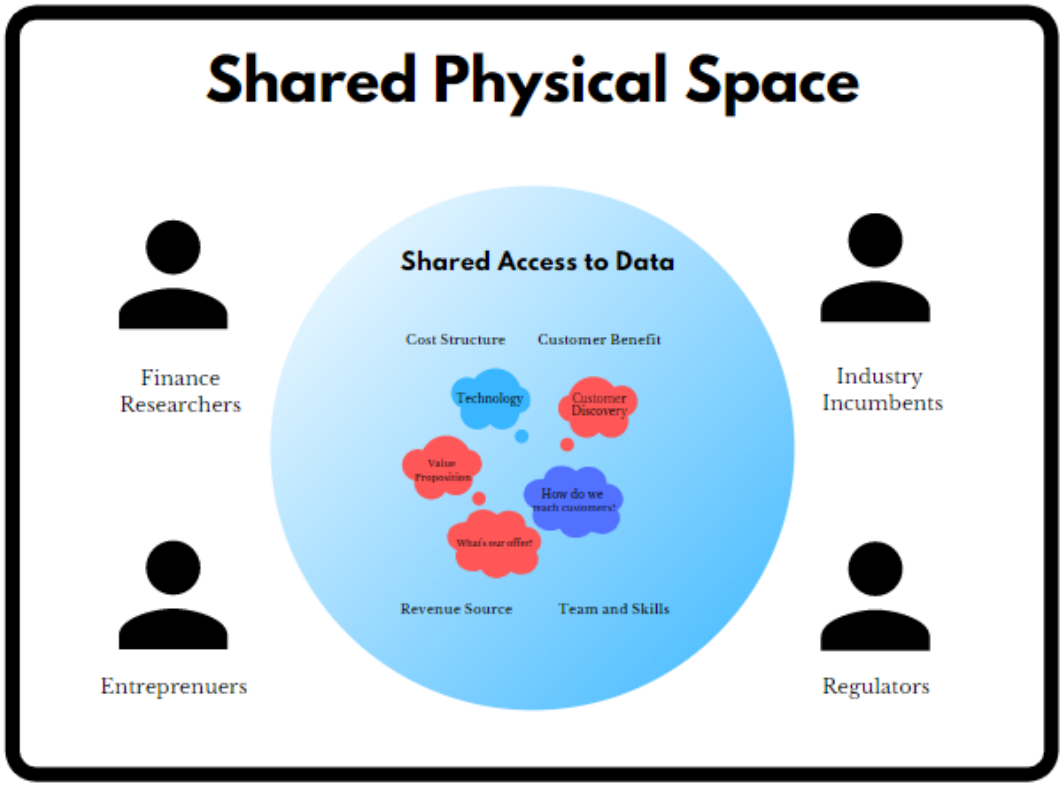
In order to manage financial innovation, companies and academic institutions must consider technology and strategy together. This means testing their concepts in a controlled way. Reflecting this belief, Bharadwaj et al, (2017) argued that it no longer sufficient to think of technology strategy as a subordinate process. It should, instead, be an aligned component of business strategy. We therefore suggest, that too manage financial innovation in a responsible way, it is necessary to test the robustness of new approaches.

What is clear is that a data lake facilitates data mining. Thomas and Sycara (1999), however, point to validation and the importance of simplicity when doing this. The research environment will be similar to that of a sandbox, in which users can apply ideas, validate business models, and conduct the rapid analysis to answer pressing questions in a big-data environment.

In summary, by its very nature, finance is a data-heavy industry. To ensure validation, in an innovation laboratory setting, requires a flexible and accurate data repository for user to draw from. As of now, such a large and easily accessible repository of data does not exist for financial research in an academic setting. In response to this, we propose the creation of a data lake for financial research and innovation is an essential component for industry interaction and innovation. We therefore propose a blueprint for building a data lake ecosystem for Fintech related academic innovation and its commercialization.

[1] A regulatory sandbox is a framework that overseas small scale, live testing of innovations in a supervised fashion.

FIGURE 1 Fintech Innovation Laboratory



2 | THE KEY DIFFERENCES BETWEEN A DATA LAKE AND A DATA WAREHOUSE

There are a number of defining differences between a data lake and a data warehouse, many of which are well documented by academics such as (Russom, 2017). That said, the importance of a design structure can and should be understood in the context of its relationship to other database management systems and the necessary consequences of these structures.

The first key difference is the data structure itself. In a data lake the ingested data is stored in its raw, or close to raw form. In a data warehouse the data is first transformed and then loaded. This traditional data warehouse pattern tends to follow an Extract Transform Load (ETL) structure. This can be best understood as data transformation and processing prior to storing the data for later use.

The data lake should be a repository of unstructured data. If commonly held assumptions are correct then then as much as 90 percent of its content may be unstructured. As such the issues of meta-data and schema definition and/ormark-up will be critical to its success. From a technical perspective it should use high performance computers not only for its large storage capacity but also to serve up access to this data. The HPC functionality needs to support 'classical' modelling as well as local versions of BERT / PyTorch-Transformers would allow users to research many of

the questions around pre-trained models/transfer learning in the context of FinTech. These typically require access to powerful GPU (or even TPU) capacity. Many current solutions tend to use services-on-the-Cloud, but there may be reasons that Fintech clients would be much more at ease with the language modelling also happening within a contained/secure environment such as the data lake.

Data lakes, on the other hand, follow a Extract Load Transform (ELT) pattern. The resulting difference stems from how the data is processed, by whom, and when. With the former, the resulting data tends to be siloed, classified, and designed for one particular use. Transformation, however, takes resources and time. Thereby reducing the flexibility of multiple source integration and storage costs.

In an innovation laboratory, flexibility of analysis for advanced users is one important property that must be maintained. The ELT pattern allows for users with adequate analytical capabilities to explore large and integrative data-sets unconstrained by the prior transformation of the data. This, therefore, increases the scope and data in which innovators and researchers can work with, allowing for more complete analysis of problems that would otherwise be unavailable.

The second major difference between a data lake and a data warehouse is in the cost of storing data over time. In traditional data warehouses the data is inherently optimized and placed in a data silo for the rapid and often repeated analysis. For business users conducting repeated and structured analysis of data-sets this is often superior. In the context of the Innovation laboratory, however, the data won't necessarily be queried and extracted in rapid, structured manner. Rather, the purpose of the data is exploratory rather than routine. With this end purpose in mind, the resulting data lake is therefore superior because the Innovation laboratory data lake will follow an Extract Load Transform (ELT) pattern. Put another way, the amount of storage and transformation is shaped by usage. As a result, the processing costs and therefore the storage costs of maintaining such a data lake are considerably lower (Terrizano et al, 2015; Lock, 2017). When combined with efficient storage and effective distributed computing, this enables a centralized data source for advanced users that can be extracted from nearly any time period. The increased time and flexibility components enable previously unavailable research question to be investigated as needed.

A key aspects of the proposed data lake is the flexibility and accessibility of data. With a data lake the intended users can and should be able to access the data lake securely and flexibly. Such a flexible structure simultaneously (and seemingly paradoxically) expands and restricts the users of the data-sets. The first two major differences change the user base fundamentally. The Fintech Innovation Laboratory is intended as an interactive space for companies, researchers, regulators, and entrepreneurs to conduct fundamentally diverse and exploratory research. With low storage costs, data explorability, and extraction protocols the data lake design pattern allows such flexibility. However, such an expansion of the data is often accompanied by the increased need for the skills needed to query the data lake. Unstructured or loosely structured data requires transformation on the back end, which changes the nature of the users who are trying to access such data under question. This particular challenge will be examined more thoroughly in later sections of this white paper.

3 | AN INNOVATION LABORATORY FOR FINANCIAL TECHNOLOGY

Our proposed data lake would be the centrepiece of an innovation laboratory. As a concept, this was first defined by (Lewis and Moultrie, 2005) as a "dedicated research facility intended to support innovation and research projects independent of daily task". Such a facility will allow for what is defined by (Argyris and Schon, 1978) as "double loop learning" and "curiosity exploration". this is performed outside of operational execution. Double loop learning is shaped by the nature of the questions being asked in the research setting. In many ways double-loop learning can be

thought of as asking questions that fundamentally challenge or change the nature of the underlying structures of a theory, research question, or business model. This is exactly what industry requires of academic led Fintech research.

Today, there are many noted users of innovation laboratories. A Fintech version would be used primarily in the financial industry and the consulting industry. To enable the successful implementation of an innovation laboratory there are a number of key factors that facilitate the exploration and analysis of double-loop problems. For example, the IBM Accelerated Discovery Laboratory is a research environment intended to facilitate interdisciplinary research projects. This enables new predictive capabilities based on the analytics.

As (Haas et al, 2017) explained "one of the affordances of the Accelerated Discovery Lab is a workspace that provides a flexible work environment for individuals and groups. The space is outfitted to facilitate creativity and collaboration through access to simple, yet effective tools such as whiteboards and displays that can be moved and configured for the needs of those using the space."

An innovation laboratory's structure is designed to be supportive of its research cases. There are not many innovation laboratories that allow for industry and academic interaction and innovation – fewer yet that do so for financial research. In response to this, we propose a Fintech innovation laboratory supported by a data lake. The strength and value of implementing an innovation laboratory emerges in this way; as a result of creating an environment where researchers, industry participants, academics, and entrepreneurs may conduct double-loop exploration in the creation of innovative business models and novel analytical inquiries.

More specific to the application of Fintech, such a laboratory requires a number of unique infrastructure pieces, including of course a data lake. Finance, inherently, is a field of study that generate large amounts of data from a variety of sources with unique formats. Such data types include pdfs, emails, xls formats, etc. Consequently, most innovation relating to finance is tightly related to data and the insights that can be derived from its analysis, hence the need for a data lake.

In order to build a successful Fintech innovation laboratory with a heavy reliance on data there needs to be a wealth of data that can be easily accessed and used. This data, particularly that containing financial information, must necessarily be anonymous for use in the controlled testing of innovative financial ideas. Beyond that consideration, the innovation laboratory must be flexible enough to handle and sustain a variety of different types of analysis, ranging from textual analysis through artificial intelligence to large data-set pricing mechanisms.

3.1 | The data lake as an infrastructure component for financial research

At its core, a data lake is a design architecture that enables large amounts of data to be stored in their near raw format cheaply and securely. By its very design, a data lake can ingest a number of different file types - both structured RDBMS data and unstructured data (pdfs, emails, etc). As an emerging design architecture, its applications are well noted across a variety of industries with use-cases in air-traffic control, research, and others (Boci et al, 2015; Haas et al, 2014).

In the case of the innovation laboratory the underlying purpose of the data lake is to provide safe access to all different types of financial data to the researchers to conduct market feasibility, product viability, scenario analysis, and other flexible queries without impacting individual data or exposing client money to risk.

To ensure its usability, a data lake requires an disciplined approach in the collection and exploitation of metadata. This includes the establishment of a sample population of financial data and a catalog describing data sources ingested. For example, Financial Services Markup Language (FSML), based on the Standard Generalized Markup Language (SGML), can be used. This allows users to define the financial information items that constitute a document.

It is also important, as the proposed data lake is based on financial information, that standards are in place for the

identification of sensitive content, such as personal information, and definition of governance processes to ensure such content is used appropriately. The experience required includes a knowledge of a range analytic architectures, traditional warehousing, distributed computing, Hadoop, NoSQL and virtualization. [2]

3.2 | The data lake as a driver of consulting revenue

A data lake, such as the one proposed in this paper, can be used as a centre piece of a consulting strategy, driving revenue for the academic institution that hosts it. As part of the Strategic Fintech concept Broby, (2019) suggested that the way to structure such an offering is to:

- Validate the customer demand.
- Test robustness and security protocols.
- Develop an implementation strategy.
- Develop a diffusion approach for acceptance.

In order for the consulting concept to be financially rewarding, the process process must necessarily include the various actors across industry and academia. As such, the data lake itself should be of sufficient size and relevance to be useful to the various stakeholders. The financial data included in it should, as a result, include extensive banking and insurance records as well as transactions. The identifiers, although blinded, should break down customer type, such as age, sex and socio-economic status.

One of the attractions of a data lake from a consulting perspective is the associated computing processing power, storage and dark fibre. Corporations do not always have access to such infrastructure and if they do they don't want it used for testing at the expense of operations. on the data lake relies needs to be of sufficient robustness for companies to data mine the resource. Its users would be, after all, individuals and companies that do not have the necessary hardware to do that, or are unwilling to use their own.

4 | CHALLENGES OF IMPLEMENTING DATA LAKE TECHNOLOGY

When considering the challenges of using the data lake design pattern in financial research one must consider both the traditional challenges of managing the application of a data lake as well as the particulars of managing sensitive financial information that could be used to identify users.

As Terrizzano et al, (2015) aptly puts it, "while this definition of a data lake is not difficult to understand, it originates from an essential premise: the data in the lake is readily available and readily consumable by users who have less technical skills than traditional IT staff." Such a premise determines the context of the addressable challenges of implementing a data lake in financial research.

4.1 | The Conceptual Architecture of a data lake

The term 'lake' is the consequence of the conceptual storage of data in nearly any format for use in a variety of manners. Extending the analogy, data is stored on the data lake, and then only loaded and transformed when it is needed. In its essence a data lake is a design pattern where all relevant data, both structured and unstructured, can be

[2] Research undertaken on the data lake is most likely to be done using Agile and Waterfall methodologies.

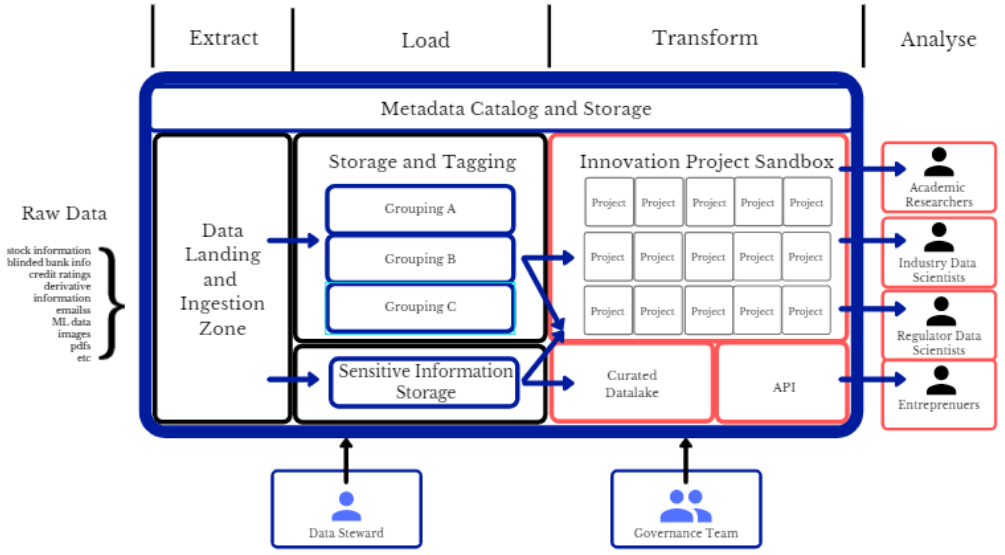
deposited for later use by users with granted access. The conceptual framework to understand the flow and structure of the data is through zones.

The primary benefits of a data lake, with adequate transformation capabilities and a sufficient metadata backbone, is that its data can be ingested rapidly, stored inexpensively, and structured when it is needed. This allows for a variety of types of advanced analytics and as such can be used as a sandbox.

With the data lake architecture, there are three general zones for data to flow through. It must be able to (1) ingest a large volume of different financial data types from a number of banking and insurance sources for storage and cataloging. It must also be able to (2) organize and maintain such a repository of client level data for later transformation by teams with differing levels of analytical sophistication and needs. Finally, the data lake should be able to (3) extract and transform transaction and payments data as needed by researchers and users while maintaining the original data lake, academic standards of conduct, and any legal requirements related to information sourcing.

In order to secure the data lake for shared access, it must also have a governance and meta-database management policy, as detailed in the section below. Without this, it risks rapidly becoming unmanageable 'data-swamp'. In an data lake governance report, (IBM, 2018) suggest that there are four major building blocks of a governed data lake, the extraction, the loading, the transformation and the analysis. Adopting this framework, our proposed data lake structure, for a Fintech innovation laboratory, is shown below in Figure 2:

FIGURE 2 The building blocks of a Fintech data lake.



The key components of this configuration are as follows:

1. Data Landing and Ingestion Zone: This is where a variety of different data formats are stored temporarily prior to their metadata tagging and organization.
2. Data Steward: This is a data lake administrator who monitors metadata, data quality, and the data tagging compo-

nents of the data lake. This user is intended to ensure the structural quality of the data lake in conjunction with the Governance Team.

3. Governance Team: Due to the heterogeneous nature of the data-types, data-sources, and user capabilities, a governance team is required to ensure both the data quality and the legal requirements of managing the data lake - across ingestion to the administration of the data to project specific teams. As is well known in governance circles, this team will be primarily concerned with data authentication, encryption issues, access control and audibility, and regulatory requirements of managing data lake.
4. Innovation Project Sandbox: The key function of the data lake intended for users with advanced capabilities in both analysis and data transformation. It requires an applied meta-database management system, these users would be able to effectively make their own project copies directly from the data lake for their own analysis.
5. Sensitive Information Protection Zone: This is required due to the sensitivity of certain financial information. In the Fintech data lake, this is addressed through blinded bank data. A subsection with increased governance will ensure the balance of user safety and ease of access.
6. Storage and Tagging Zone: This is the primary repository of the data lake. Within this repository, the data steward, the governance team, and the supporting platform catalog, record metadata, and store data for later use.
7. Arrows: The arrows are indicate the flow of information.

Such a configuration aims to provided shared access to a large and centralized of data for users of differing levels sophistication. Specific to managing a data lake in a university setting, such a data lake must necessarily prioritize data governance and integrity over speed in its first iteration. That said, the manifestation of this data lake for financial innovation will necessarily require a team of individuals to enable its core functionality - and therefore figure 2 should be considered as a framework rather than a detailed list of specifications.

To justify such a dedicated approach and the resultant investment, the value and challenges of implementation must be examined. In this respect, the value of a data lake is best understood by comparison to its relationship with traditional data warehouse architectures.

5 | DATA MANAGEMENT PLAN

A data management plan should be prepared specifically for working with large amounts of financial data. This should address the problem of deductive identification. This problem results from the fact that financial information is inherently sensitive. It is therefore critical to ensure the data lake provides the ability to research, test, and implement innovative ideas as well as business models without unnecessarily risking the financial and other data of customers. This means blinded data is a must. In essence, the problem with blinded financial data is that it is possible to identify where it came from with some investigative work. For example, university staff have a public pay scale and as such the salary line, age, sec and location used together could identify the blinded record.

The data lake should be managed in accordance with RCUK Common Principles on Data Policy [3]. The data lake should be run with a team working on the Research Data Management and a University Library Service, which is also a response to the drive to make research data openly accessible.

The data lake will generate the following new data and resources. The data will be stored in different formats: text-delimited for numerical data. The data will be analysed using Excel, MATLAB, R and Maple. Bayesian Network (BN) modelling will be conducted in GENIE, Python and R.

[3] see the University of Strathclyde's (UoS) Research Data Policy and Code of Practice on Research Ethics

Quality issues in the financial data will be addressed at the time of data collection and data entry. We will use several validation techniques to ensure that the data recorded reflect the actual facts, responses and observations or events. The data lake will require clear research guidelines and frameworks to ensure consistency, for instance for note-taking for observation.

5.1 | Meta-database management of unstructured and structured data

One of the primary benefits of data lakes is its ability to rapidly ingest and cheaply store information. In tandem, however, is the related challenge of managing the accompanying metadata. While metadata in traditional data warehouse architectures is well-established, extracting both structured and unstructured from a variety of sources presents its unique challenges. Due to the purpose of the rapid ingestion, poorly managed data lakes can quickly become 'data-swamps' - unstructured repositories of information that are impossible to use. Without a strong metadata management system data ingested into a data lake is hardly usable (Hai et al, 2016; Quix et al, 2016; Russom, 2017)).

To address this challenge in data lake management it is essential to have strong data governance and a flexible process to systematically address issues as they emerge. In this regard, we advocate for a data steward and governance team combined with the agile approach to solve future unforeseen data lake problems (Hagstrom et al, 2017).

There are also a number of regulatory challenges. Given that the storage costs in this particular data lake are low it becomes possible to source data from a variety of sources in a variety of formats. Flexibility of sourcing consequently increases the complexity of managing the legal requirements and adherence to the terms of accessing such data. With recent developments in data protection and the must be backed by compliance team.

5.2 | Blinding and encrypting the data

The information and research data in the proposed data lake will need to be be saved on an encrypted and password protected computers, and ultimately stored on a password-protected server. The server should automatically back up every 15 minutes on a University backup system, which should be replicated to at least two buildings to avoid potential loss of data during accidents (e.g. fire). The servers should also have enough capacity to store all expected amount of data generated by researchers.

Subject information (i.e., personal identifiers) will need to be stored separately from their data, so the data will be stored anonymously. Some data will have alias identity, for example, for banking data, to preserve participant anonymity while enabling real world money patterns to be observed.

Access to the data depository should be automatically monitored by the data storage system; the log file will be stored in a folder that can be accessed only by the project leader and senior system administrator. Access to data files will be restricted to project researchers, and will be formalised by a data management agreement.

All background IP tested on the data lake shall remain the property of the party contributing same. Formal disclosure of IP deemed appropriate for commercial exploitation should be made.

All data will be also need to be archived, for example, using the UK Data Service (for electronic data). Written data/hard copies of documents that are no longer part of an ongoing study will be stored in a secured warehouse used by the University. Materials in the data lake will have to be anonymized or de-identified as appropriate, converted to searchable formats, and stored. Following project completion, and following research publications, these data will become publicly available.

5.3 | Data user profiles

We suggest that there will be four primary users of the innovation laboratory's data lake, namely (1) Entrepreneurs and Innovators; (2) Industry incumbents (3) Academic Researchers and (4) Regulators.

In order to be successful as an industry collaboration, such a laboratory requires shared access to a wide variety of financial data-sets in a centralized and secure location. We suggest the essential infrastructural component that enables the functioning of the Fintech innovation laboratory is the data lake. A successful implementation of a data lake in a Fintech innovation laboratory will allow companies to conduct market feasibility, product viability and scenario analysis without impacting individual data or exposing client money at risk.

The remainder of the paper will explore the concept of the data lake and the associated benefits as well as challenges associated with implementing the data lake for financial research.

5.4 | Usability of unstructured data for non-technical users

There is a challenge in ensuring users have the ability to structure the data. In the context of the innovation laboratory, there may then be a skill gap between users with advanced analytical capabilities (such as data analysts) and those with less advanced analytical capabilities (such as the entrepreneur or entry-level business analyst). A second issue within this context is that of making sense of the data under analysis. Data is only useful if its context and applicability is understood. Conducting analysis for the sake of analysis is unlikely to yield meaningful and disruptive insight.

In a university context, there is work being done to address such a skill gap for both hard skill development and critical thinking. This means incorporating blends of traditionally taught financial skills with unique fintech components and ICT hard skill components. Others have proposed specific training models for users of datalakes (Nargesian et al, 2019). Such an expansion of this logic would further close the skills gap needed for users to realize the value of unstructured data from a data lake.

6 | CONCLUSION

In this paper we introduced the application of a data lake as applied to the Fintech Innovation Laboratory. We have shown that applying a data lake in a university setting has the capability to enable a shared innovation environment between incumbent companies, entrepreneurs, regulators, and finance researchers. The successful implementation of the data lake in Fintech research can and should facilitate innovative research and disruptive business models that would otherwise be unavailable in the marketplace.

We highlight the commercial benefits of such a collaboration, The data lake facilitates a consultancy model, backed by validation by independent (academic) third party experts. This can drive revenue in a setting where universities wish to leverage their assets.

It is our belief that the role of academia in finance can and should evolve in tandem with that of industry. The scope and impact of Fintech has yet to be fully realized and In many ways the future of finance is going to be difficult to predict. This is especially true the farther out we try to predict and how specific our predictions are. What we do know is that data is going to continue to expand exponentially and enhance our ability to drive insight from that data. The Fintech innovation laboratory increase our abilities of extraction, innovation, and therefore, strategic management of financial research.

references

Arner, D.W., Barberis, J. and Buckey, R.P., 2016. FinTech, RegTech, and the reconceptualization of financial regulation. Nw. J. Int'l L. and Bus., 37, p.371.

Argyris, C. and Schon, D. (1978) Organisational learning . Addison Wesley, Reading, MA.

Boci, E., Ph, D. and Thistlethwaite, S. (2015) 'A novel big data architecture in support of ADS-B data analytic Hadoop Distribution Software', 2015 Integrated Communication, Navigation and Surveillance Conference (ICNS). IEEE, pp. C1-1-C1-8.

Broby, D. (2019) 'Strategic Fintech', Centre for Financial Regulation and Innovation, pp. 1–13.

Carr, A. (2017) Optimizing data lakes for financial services.

Christensen, C. M., Raynor, M. E. and Mcdonald, R. (2019) 'What Is Disruptive Innovation.?', 2015, pp. 1–17.

Cito Research (2014) 'Putting the Data Lake to Work A Guide to Best Practices', p. pp 1-12.

Gomber, P. et al. (2017) 'On the Fintech Revolution.: Interpreting the Forces of Innovation , Disruption and Transformation in Financial Services', (3), pp. 1–35.

Haas, L. et al. (2014) 'The IBM Research Accelerated Discovery Lab', (April 2016).

Hagstroem, M., Roggendorf, M., Saleh, T. and Sharma, J. (2017). A smarter way to jump into data lakes. [online]

Hai, R., Geisler, S. and Quix, C. (2016) 'Constance.: An Intelligent Data Lake System Constance.: An Intelligent Data Lake System', (November 2017).

Ibm.com. (2018). Governed Datalake for Business Insights.

Karkkainen, T., Panos, G., Broby, D. and Bracciali, A. (2017). On the Educational Curriculum in Finance and Technology. [online] Pure.strath.ac.uk.

Lewis, M. and Moultrie, J. (2019) 'The Organizational Innovation Laboratory', Creativity and Innovation Management, (March 2005).

Lock, M. (2017) 'Angling for insight in today's data, October 2017', (October).

Nargesian, F., Miller, J. and Pu, K. Q. (2019) 'Data Lake Management.: Challenges and Opportunities', pp. 2–5.

Bharadwaj, A., Sawy, O., Pavlou P., and Venkatraman, N., (2013) 'Digital Business Strategy: Towards a next generation of insights', MIS Quarterly, 37(2), pp. 471–482.

Peat, J., Kelly, O. and Broby, D. (2017) 'Fintech: hype or reality?', University of Strathclyde International Public Policy Institue, pp. 1–15.

Quix, C., Hai, R. and Vatov, I. (2016) 'GEMMS: A Generic and Extensible Metadata Management System for Data Lakes', pp. 129–136.

Russom, P. (2017) 'SAS Best Practices Report: Data Lakes', Q Research, pp. 1–42.

Stein, B. and Morrison, A., 2014. The enterprise data lake: Better integration and deeper analytics. PwC Technology Forecast: Rethinking integration, 1(1-9), p.18.

Terrizzano, I. et al. (2015) 'Data Wrangling: The Challenging Journey from the Wild to the Lake'.

Thomas, J.D. and Sycara, K., 1999, July. The importance of simplicity and validation in genetic programming for data mining in financial data. In Proceedings of the joint AAAI-1999 and GECCO-1999 Workshop on Data Mining with Evolutionary Algorithms.

Walker, C. (2015) 'Personal Data Lake With Data Gravity Pull', (October).