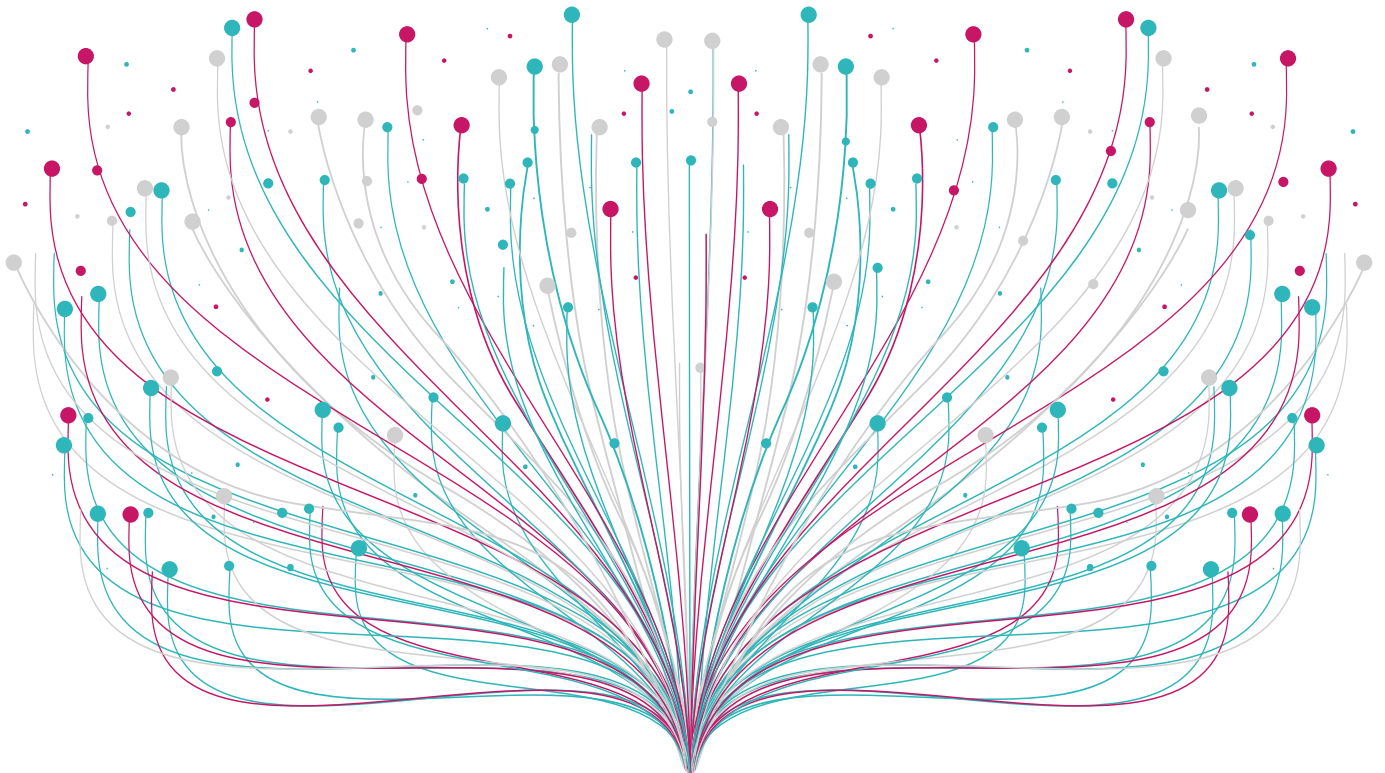


# OPTIMISING DATA LAKES FOR FINANCIAL SERVICES

## FOUR KEY QUESTIONS AND A WORD OF WARNING

By using a data lake, you can potentially do more with your company's data than ever before. You can gather insights by combining previously disparate data sets, optimise your operations, and build new products. However, how you design the architecture and implementation can significantly impact the results. In this white paper, we propose a number of ways to tackle such challenges and optimise the data lake to ensure it fulfils its desired function.

/ A WHITE PAPER BY ANDREW CARR



**SCOTT LOGIC**

ALTOGETHER SMARTER

## / CONTENTS

Data lakes: not just one-size-fits-all	3
What is a data lake?	4-5
Q1 Should I transform my data, or leave it raw?	6-7
Q2 Should I apply more filters to my data?	8-9
Q3 Should I switch from batch to streaming?	10-11
Q4 Should I change the way I label my data?	12-13
And a warning... Who has access to your data?	14-15
Contact us	16

# DATA LAKES: NOT JUST ONE-SIZE-FITS-ALL

MANY FINANCIAL SERVICES INSTITUTIONS HAVE ALREADY SEIZED ON THE POTENTIAL OF DATA LAKES. BUSINESSES ACROSS THE WORLD ARE ENHANCING THEIR IT SYSTEMS TO GET A BETTER UNDERSTANDING OF THEIR MARKETS AND CLIENTS. AS A RESULT, DATA LAKES HAVE EMERGED AS A VIABLE WAY TO POOL DATA AND GET A MORE DETAILED VIEW OF EVERYTHING THAT IS HAPPENING ACROSS THE COMPANY.

In financial services in particular, there is an additional driver. A decade after the financial crisis, businesses are more aware of the value of getting a holistic perspective on risk. Having a more complete view of all the data available in various company systems could be the difference between being aware of issues and being blindsided by them.

**ON THE SURFACE, DATA LAKES SEEM TO PROVIDE MANY OPPORTUNITIES. THEY PROVIDE NEW OPTIONS FOR SELF SERVICE REPORTING, SATISFYING REGULATORY REQUIREMENTS, GATHERING A SINGLE SOURCE OF REFERENCE DATA, AND OPENING UP DATA FOR NEW PURPOSES.**

For example, more established data warehouses require stored data to be transformed into a very specific format, such as a star schema for super fast retrieval of very specific data. A data lake allows all the data to be stored in a raw format. This gives users the freedom to transform it later, once they've found a valuable or innovative use. Some of these uses only appear when you analyse the raw information in all its glorious detail.

Data lakes also provide a useful repository for a whole range of data from different sources, which can be used to provide more accurate risk reporting and insights. By combining data previously held in various silos, companies can get a more unified view of their customer interactions and operations. With the emergence of frameworks such as Hadoop, the cost of storing huge amounts of data on commodity hardware has further sweetened the deal for firms who see the benefits of having a large repository of data.

However, any smart business knows that there's no such thing as a silver bullet. While many financial services firms are attempting to capitalise on the potential opportunities of data lakes, many run into issues, ranging from the disappointing to the disastrous.

As ever, it's not technology that ensures success, but what firms decide to do with it. So, if your data lake isn't meeting your expectations, here are a few questions you could ask yourself to identify the root causes and optimise performance.

# WHAT IS A DATA LAKE?

DATA LAKES ARE FAMED FOR THEIR FLEXIBILITY. IN ESSENCE, A DATA LAKE IS A PLACE WHERE YOU CAN DEPOSIT RAW DATA FROM A WIDE RANGE OF SOURCES, AND LEAVE IT THERE UNTIL YOU'VE DECIDED ON THE BEST USE FOR IT.

Data warehouses are populated using an "Extract Transform Load" (ETL) approach, meaning that information is adapted for a particular use before being stored. In contrast, data is commonly loaded into a data lake before being transformed, following an "Extract Load Transform" (ELT) pattern.

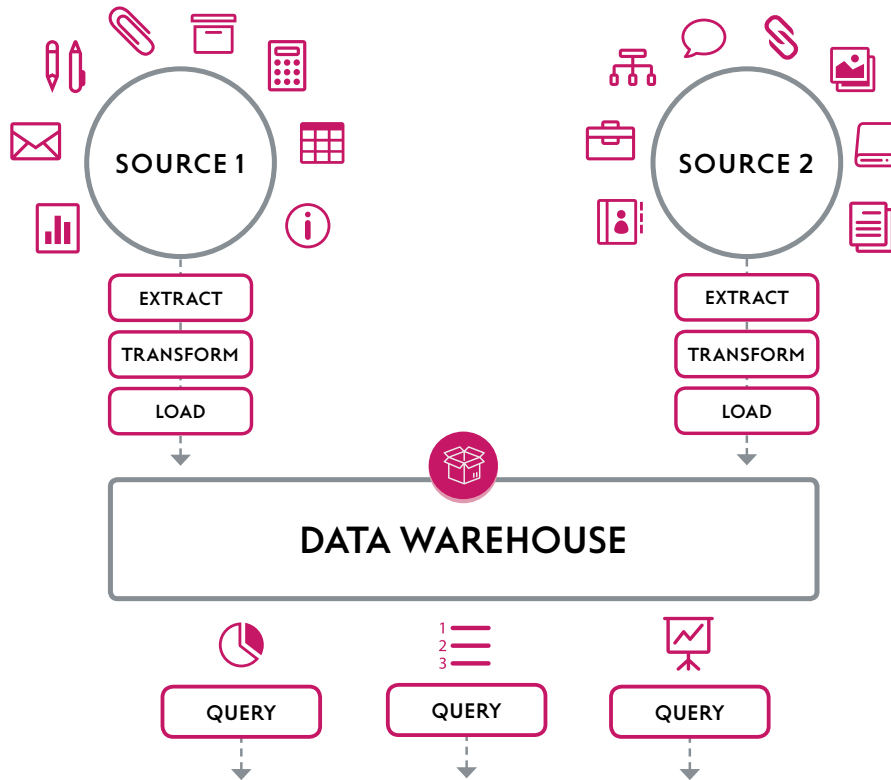
ELT can have its advantages. For example, imagine that a regulator enquires about which trades were carried out on a particular stock, to investigate a potential market manipulation event. They may require information on what each trader typed into their system, to understand how those trades were ordered.

Being able to examine multiple systems means that a company can see a client's full journey from enquiry to conversion, linking phone calls, instant messages and web traffic. This information - coupled with precise timestamps - can be invaluable in refining a sales conversion process, and can be so much easier to examine using raw data.

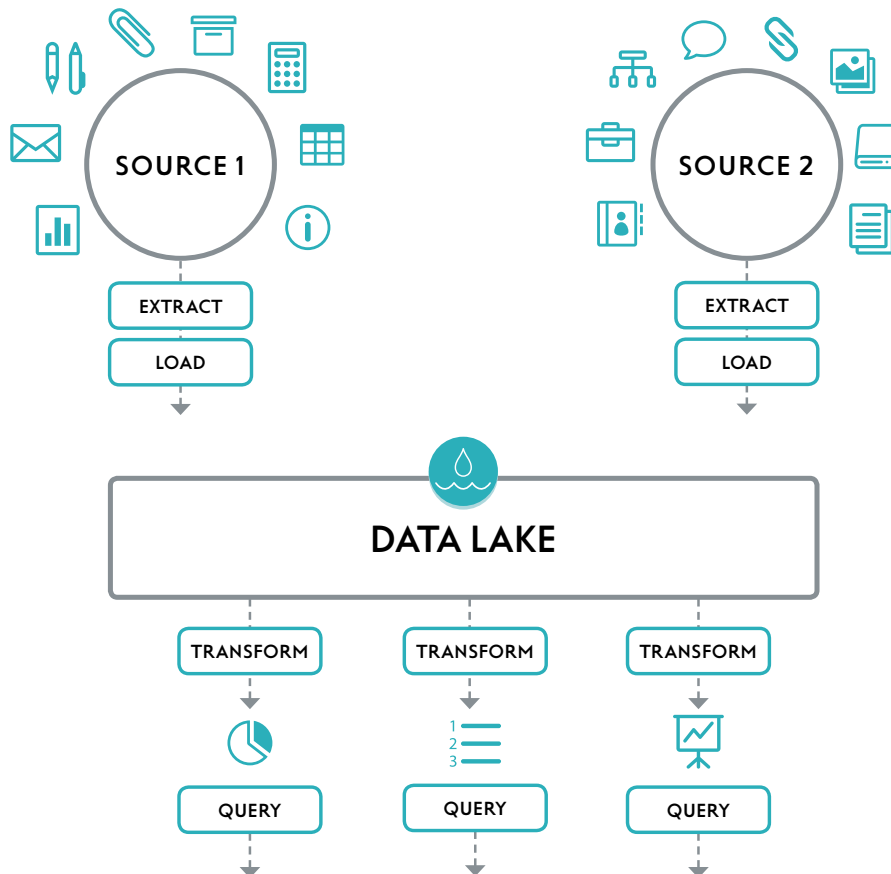
These are questions which may only be adequately addressed by examining the full raw data. Certain information may not be included in optimised data stored in the data warehouse, or an Online Transaction Processing system (OLTP) such as the optimised storage in the original trading system.

**THE ONLY WAY TO ENSURE ALL REQUIRED DATA IS STORED FOR ALL POSSIBLE FUTURE USES IS TO KEEP THE DATA RAW, POSSIBLY INCLUDING LOGS, OR TRANSFORM THE DATA WITHOUT LOSING ANY DATA AT ALL.**

## Extract Transform Load pattern, used in data warehouses



## Extract Load Transform pattern, often used in data lakes



Q1

# SHOULD I TRANSFORM MY DATA, OR LEAVE IT RAW?

A WELL-PUBLICISED BENEFIT OF DATA LAKES IS THE ABILITY TO STORE RAW DATA FOR A VARIETY OF FUTURE USES. BUT IS THAT REALLY WHAT YOUR BUSINESS NEEDS?

In theory, a compelling use for your data lake could be to use it as a giant toy box, depositing vast quantities of diverse data into one place for future study. But whether you should *actually* do that completely depends on what you want to use the data lake for.

Some firms may want to build a data lake containing raw data if they are carrying out exploratory work on the data, perhaps with a view to creating an unoptimised prototype of a new system. However, if you have a defined purpose for your data lake, not transforming your data can often be counter-productive.

Firstly, the data would not be optimised appropriately for that use, and the data lake might contain conflicting, duplicated or inaccurate data.

Secondly, if you don't transform the data, it will carry the idiosyncrasies of its source system. If you are drawing data from multiple sources, that can exacerbate the problem.


Thirdly, the data would still need to be transformed when retrieved, slowing down the development process even further. There is also the danger of generating inconsistent results from the same source data, if multiple developers write different functions to transform it.

Transforming data before storing appears to fly in the face of one of the main benefits of a data lake. However, doing so can prevent business issues that stem from the analysis of misleading or incorrect data.

Let's take a common example, in which a company has multiple Customer Relationship Management (CRM) systems and multiple ways clients can interact. The firm could store all the client interactions from these multiple CRM systems, and then analyse them to spot cross-selling opportunities. In these cases, there are many common transformations that are required to get the data into the right state.

Are the values consistent, and does a specific value have the same meaning across the data lake? Is duplicated data preventing the linking of data from separate systems? Is the client named consistently in each system?

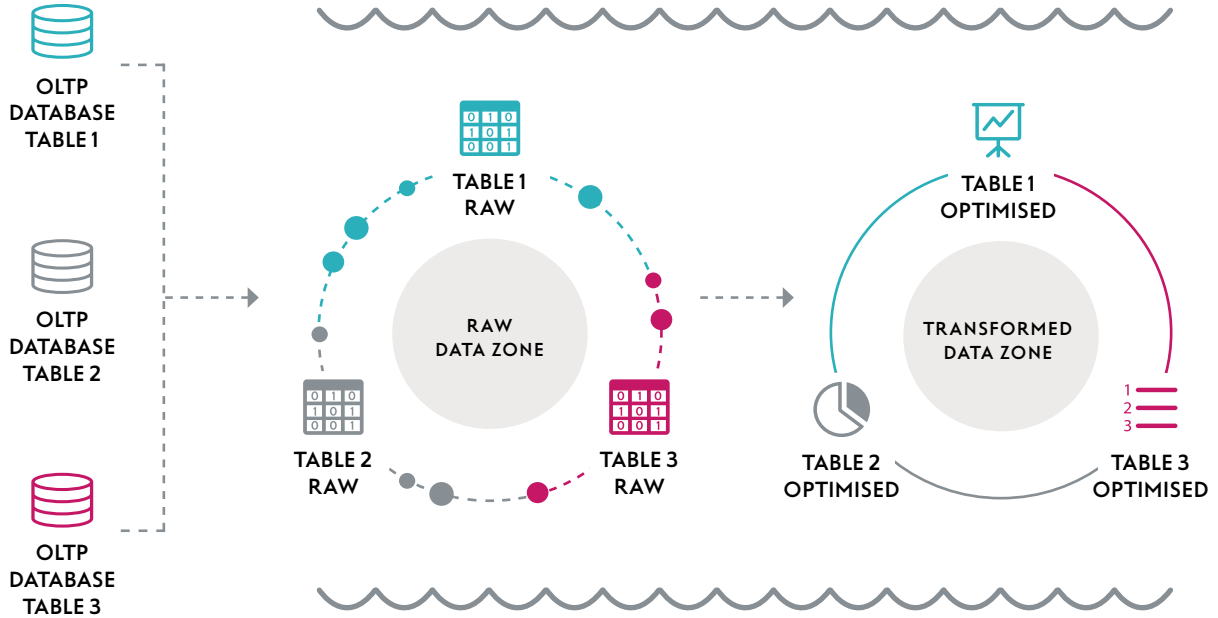
It's easy to underestimate how much work these transformations can be. In fact, it can often become the most time-consuming task on a project when teams are using source data from multiple systems.



Transforming data before storing appears to fly in the face of one of the main benefits of a data lake

# DATA LAKE

SOURCE SYSTEMS



In these cases, you may wish to consider adopting an ETL tool for your data lake which has feature rich transformations. These tools can save a significant amount of time and effort, and a lot of them are evolving specifically to be optimised towards getting data into a data lake.

When you're using optimised data but still have a compelling driver for keeping your raw data, it is worth considering a third option. This involves separating raw data and transformed optimised data into different zones. An approach like this allows analysts and applications to draw on each set of data for their own purposes. However, it does also involve storing twice as much data, and comes with the added complexity of keeping both sets of data within the same data lake.

## / SUMMARY

Knowing when to transform your data is tricky, but crucial. Transform too early and you risk losing subtle information held within the raw data. Transform too late, and you place the responsibility on the consuming processes, or data scientists who may not understand the format precisely. What you decide to do should largely depend on what you plan to do with your data lake. However, you may want to consider the third option, which is to store the data in both "raw" and "refined" zones. This can be a solid solution, but tread with care, as it features the advantages and disadvantages of both approaches.

## Q2

# SHOULD I APPLY MORE FILTERS TO MY DATA?

WHEN VARIOUS SOURCES OF DATA ARE DEPOSITED IN THE SAME PLACE, THE POOR-QUALITY DATA CAN POLLUTE THE ENTIRE POOL.

It's hard for any organisation to guarantee universally high-data quality when it's dumping the contents of many systems into a data lake.

After all, the data will have been impacted by many changes in requirements, the organic growth of bespoke systems, and an IT landscape with lots of duplication.

One side effect of depositing multiple datasets in a data lake is that some parts of the data are "brought back from the dead". Some of these may be corrupted or out-of-date columns of data. Others may be data located in corners of systems that are no longer tapped for operational use.

**IN A DATA LAKE, THERE MAY NOT BE RESTRICTIONS IN PLACE TO STOP THIS UNHELPFUL DATA RISING TO THE SURFACE AND POLLUTING THE ENTIRE POOL.**

If your data lake is being used to feed operational processes such as self-service reporting, or if it's being opened up to clients, you should be aware that you run the risk of systems drawing out inappropriate data for operational use if you don't sift the information first.

Once again, the fact that a data lake allows users to deposit data without sifting through it doesn't necessarily mean this is your best option. If your data lake is facing issues with data quality and duplication, it's always worth stepping in closer, filtering or throwing out bad data, and imposing extra oversight into what data sources and which tables and columns flow into the data lake in the first place.

The advent of feature-rich ETL tools now enable almost anyone to place data in a data lake, so it is useful to consider whether you need to put controls in place to limit who has the ability to do this.

In the previous point, we discussed the use of "zones" featuring different types of data. If you are interested in setting up a "refined zone" for transformed data, remember that this should only contain "high-quality data". Therefore, there should be checks to ensure that only data with the right format and quality should be able to pass into this zone.

Data lakes are often misused by organisations that are not fully aware of the potential and limits of the data they are collecting. Having insight into the quality and nature of your data will dictate what uses your data lake can sustain.





## / SUMMARY

Should you dump all your data in a data lake and work out what is worth using later? Should you only transfer data which has been quality-checked? As ever, which approach you take depends on your requirements. If you see your data lake as a place to mine for data science discovery projects, you may want to do the former. If you want software teams to draw on that data to build applications, you may opt for the latter. As mentioned in the previous point, “doubling up” the data by storing it in both raw and transformed formats could give you the best of both worlds, although it doesn’t remove any of the down-sides of either approach.

### Q3

# SHOULD I SWITCH FROM BATCH TO STREAMING?

MANY FINANCIAL INSTITUTIONS PROCESS THEIR DATA IN BATCHES. BUT COULD STREAMING OFFER A MORE VALUABLE SOLUTION?

It is common to see financial services firms transfer their data using batch processing. This makes sense, as the approach mimics the routine of the finance industry. Between trading windows at the exchanges, information is passed along the pipeline of teams, almost mimicking the old way that tickets were collected and passed on from traders. However, there's a strong and growing case for institutions to consider streaming instead.

While a streaming model might lower operational costs when it's in place, businesses will have to shoulder the initial costs of developing the streaming architecture. A firm's willingness to do this will entirely depend on the use they have for the data they are transferring, and the purpose of the data lake being fed.

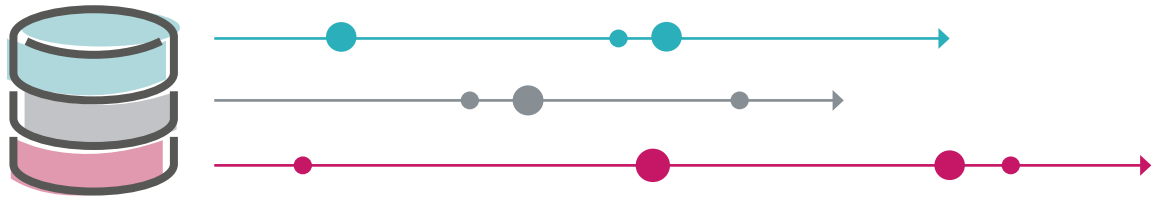
A batch processing approach may be acceptable for a firm using its data lake as a non-time-sensitive repository. If the use is more crucial, or relies on real-time (or at least relatively up-to-date) information, then streaming will often emerge as the better option.

A company may launch a batch job during a break in trading, such as the end of the trading day. If a batch job takes an hour, and it fails, it commonly has to be run again in full until it works. As the volumes of data being processed increase, batch processes take longer to complete.

Furthermore, trading is global, and fast-paced. Some batch jobs need to complete within a certain time window, and failure to do so can have an impact, such as preventing trading on certain products or markets when trading resumes. There is also the issue of running a batch job while your system is required to be "operational" for other tasks.

Your data lake may be fed by an Online Transaction Processing (OLTP) system with the relevant source data, but if the process of transfer is a timed batch job, it may not have the most up-to-date information. This can have consequences, particularly if your data lake is used to provide transaction information to clients, or give the organisation a "Single Version of Truth" of the data.

It certainly appears that streaming is an approach whose time has come



Using your data lake as a centralised reference data source - or a "Single Version of Truth" - can come with its own set of challenges. You are essentially taking a data lake - which is usually a non-critical system - and making it a dependency of some of your critical systems. This impacts on engineering, support, design, and maintenance, and is often twisting the data lake to become a key component in your critical systems data flow. It is possible to do this, but it does have wider ramifications. A common requirement for this approach is to move to streaming over batch, so the reference data provided to other systems is as up to date as possible.

In several cases, streaming can be a more effective - and often more cost-effective - alternative once up and running. Streaming systems are usually a simpler implementation than batch, and can often handle states automatically, as opposed to the typically-manual state transition management required by batch. Instead of a timed transfer of information, a streaming system can pass on the data as and when required. When failures occur, a streaming system can pass or fail single events rather than require that the entire batch job re-starts.

Innovations in stream processing - led by technologies such as Apache Kafka, Samza, Apex, Spark Streaming, Storm and Apache Ni-Fi - mean a streaming solution is easier to develop, deploy and maintain than ever before.

## / SUMMARY

Many companies have traditionally opted for batch processing, as it mirrors their older physical processes. However, streaming has matured to the extent that it can be a cheaper and simpler approach. It is easier to run and maintain, and supports more use cases such as using the data lake as a central reference data store for a business. It certainly appears that streaming is an approach whose time has come.



## Q4

# SHOULD I CHANGE THE WAY I LABEL MY DATA?

WHEN FIRMS OPEN UP THEIR DATA LAKE, IT CAN BE HELPFUL TO CLASSIFY THE INFORMATION TO SPEED UP THE SEARCH PROCESS. BUT IS THAT ALWAYS THE BEST APPROACH?

Data lakes can be valuable, enabling a broader group of users to gather and harness data for various business cases. However, by nature, they are often a dense and wild forest.

The financial services industry is complex, and the information stored in raw datasets can often be misleading. Some firms have responded with automated approaches - such as machine learning techniques - that classify and tag datasets, building up a more complete "Data Dictionary". Others carry out the process manually, reducing the risk of incorrect tagging. Some do not tag the data at all.

However, it is important to note upfront that there is a limit to what even the most advanced classification technologies can achieve, and that mis-categorisation can still occur. Relying on automated tagging of multiple data sources can be problematic, and they should generally be regarded as "educated guesses".

Depending on the expertise of the data scientist or user that is exploring the data, the Data Dictionary may need to be extremely explicit in its definitions. And even then, there can be misunderstandings around how the data should be applied or linked.

Also, remember that a data lake is often a repository, rather than a carefully-cultivated source of current data. Data duplication and quality issues can often throw off the analysis of datasets. Users may start linking bad data with out-of-date data, or joining together incompatible or untransformed information. Relying on incorrect data can stop an intriguing concept from making the leap to becoming a trustworthy platform or product.

**IF YOU'RE CONSIDERING LESS PROBLEMATIC OPTIONS FOR YOUR DATA LAKE, YOU MAY WISH TO FALL BACK ON A MORE GRADUAL APPROACH TO RELEASING DATA MORE WIDELY.**

Instead of opening up huge volumes of unreliably-tagged data - or attempting to tag vast amounts of data yourself - start with a small group of systems that have most potential, and build up your dictionary as you go.



Your data could also be arranged into distinct “zones”, split up based on the quality stored within them. These would range from a “landing zone” for raw data that has just been transferred into the data lake, to a “refined zone” containing data that is clean and properly structured. While this may be a strong approach for some use cases, the added storage and zoning requirements mean that it’s not always preferable to keeping the data raw-only or transformed-only.

While having extra eyes on your data could lead to interesting innovations, it’s important to make sure your data lake isn’t just a “swamp” of unreliable information.

### / SUMMARY

Labelling can be key to a useful data lake. Businesses are beginning to examine and refine best practices, and their approaches can be enhanced by evolving machine learning tools that automatically classify data. However, we would sound a note of caution. While these new tools can help, they aren’t the only answer, and a certain amount of trial and error and manual tagging/classification should be applied as well.

# AND A WARNING...

## WHO HAS ACCESS TO YOUR DATA?

YOUR DATA LAKE WILL GIVE YOU THE OPPORTUNITY TO VIEW DATA FROM A VARIETY OF DIFFERENT SYSTEMS IN ONE PLACE. BUT HOW DO YOU CONTROL WHO SEES WHAT?

While data lakes can be home to a dizzying variety of data, it can be catastrophic for a financial services firm to leave them open to everyone.

Firms have data stored in a variety of different places. In some cases, the sources of this data are OLTP systems which tend to restrict data access based on carefully-defined rules within the system itself. When this data is transferred to the data lake, these restrictions are often lost. This means that the walls that are legally required to manage access to information are broken down. Breaching the "Chinese walls" that prevent improper trading and conflicts of interest can have devastating and costly consequences.

Given the potential cost of a major regulatory breach, we advise that businesses consider access control at the earliest possible point. In fact, the best approach is to have access control mapped out and put in place while designing the data lake itself. However, it is not uncommon for financial services firms to put off imposing access controls until a later date. This could be a real problem if the data lake is then enhanced to feed Open APIs or features any form of client access.

As the potential capacity of data lakes grows, it is also important to consider any emergent access control issues that may not have been envisaged when the data lake was initially designed. For example, if new types of data are deposited in the data lake, it may require a refinement of the access control method.

**NEW TECHNOLOGIES SUCH AS APACHE RANGER FOR HADOOP CAN BE CONFIGURED TO HELP INSTITUTIONS CONTROL AND MONITOR ACCESS TO DATA.**

In addition to defining which users can access which data, it is vital that firms consider the policies they apply to data, and how it is encrypted, masked and walled off to avoid breaches.





 **WE WOULD ADVISE OPTING FOR ONE OF THREE APPROACHES.**

---

**If you plan to place untransformed data in your data lake, you should heavily restrict which areas each data scientist gets access to**

---

**You may also decide to transform it all, in which case you can apply the controls from the source systems**

---

**...or apply zones for data, ranging from raw to “refined”. In this case, you apply the techniques for both raw and refined data**

---

All of this may - on the surface - appear to fly in the face of one of the main drivers for using a data lake. But if you are pooling data in one place, you are also removing the safety net of controls that would prevent users from viewing data they shouldn't be able to see. It is essential that every data scientist that has access to your data is aware of the Chinese Walls, where they are, and where their limits should be.

**/ SUMMARY**

It is so tempting to leave access control to the end. However, with the use of data lakes for business-critical applications really starting to take off, access control is now an area of real focus. The earlier access control is considered, the easier it is to apply controls both at the design level (by potentially having separate zones) and at the technology level (by using any of the new technologies). Remember: the cost of getting access control wrong can be very high, especially in cases of data breaches, or linking data which isn't allowed to be linked.



## LET US HELP YOU ASK THE RIGHT QUESTIONS

Most leading financial institutions are setting up data lakes to help them solve operational issues and pave the way for new innovations. Scott Logic has helped many firms to think through their options, and make the right decisions to optimise performance.

If you need help tackling a complex technology problem, or assessing how to deploy or optimise data lakes for your organisation, our experienced and pragmatic consultants are available to deliver advice and build solutions that are guaranteed to be a help, not a headache.

**TO ARRANGE A CONSULTATION CONTACT  
ANDREW CARR ON:**

+44 333 101 0020

[acarr@scottlogic.com](mailto:acarr@scottlogic.com)

**SCOTT LOGIC** / ALTOGETHER SMARTER

3rd Floor, 1 St James' Gate  
Newcastle upon Tyne  
NE1 4AD

+44 333 101 0020

[scottlogic.com](http://scottlogic.com)