D4 Technical Due Diligence Report

Dated	16th February 2021
Client	[redacted]
Target	[redacted](the Company)

Executive Summary

Following our review, we feel that the Company's technology is fit for purpose, can be scaled, and is supported by sufficient resourcing. There are however question marks over the degree to which the overall business proposition is scalable as it stands, given the reliance on account management and relatively limited share-of-wallet. There are also some concerns over the permanence of key staff and potential changes in the regulatory landscape that may impinge on the Company's ability to operate its core processes.

Issues

		IMPACT		
		Low	Medium	High
RISK	Low	2		4
	Medium			
	High	2		1

High risk / High impact

- The technical team is small, partly offshore and contains new hires. The CTO retains a large proportion of the technical knowledge regarding the development of the system/software, and his team are therefore somewhat dependent on him. As identified in the Early Warning Report, the CTO has no stake in the business and is already operating a separate venture. His potential departure represents the largest known risk.

Low risk / High impact

- Despite a move toward authorised social media API integration, website scraping still generates a significant proportion of the Company's source data. While the risk of scraping being banned, or heavily limited, may appear remote, it would still have a serious impact on the Company's core processes.

- In parallel to the above, there is also a risk that social media platforms may at some point restrict, monetize or close access to platform data, even if the data is anonymised. Again, this risk is judged to be low, but the landscape is rapidly shifting and the focus on data privacy may lead to platforms behaving unpredictably. For example, Apple plans to offer phone users the ability to block the upload of usage data to Facebook. While this does not impact on activity on the Facebook platform itself (which is outside Apple's control) and therefore does not affect the Company's current data sourcing directly, it may be a sign of things to come.

- The system has NOT been penetration tested at any point. Given the relative lack of sensitivity of the data, this may not be considered a risk. However, a malicious attack on the Company's infrastructure could cause serious damage and compromise data, thus company reputation.

- The relative complexity of the codebase, infrastructure and development process, coupled with the labour intensive nature of onboarding and account management, would suggest that costs might stay broadly in proportion to the size of the client base, and that economies of scale could be harder to generate.

High risk / low impact

- The Company has, by its own admission, struggled to increase client share of wallet due to the limitations of its current proposition - clients will only spend so much on determining target audiences and messages. A greater opportunity lies in managing the clients media spend, and the Company recognises that it may have to adapt or extend its service to cater for this.

- Client engagement appears to be a fairly labour intensive activity, and this is reflected in the number of account management staff. While clients, once fully onboarded, have control over their own analysis, it would seem that the Company is still required to assist on a regular basis. This helps the Company develop its service - and was highlighted as being of 'great value' - but may present issues with business scalability.

Low risk / Low impact

- The Company appears to have a good depth of technical expertise, and has increased its resource focused on data science. Much of this resource however is new - the [redacted] joined in Nov 2020 - and so there may be some doubt around the value or performance of recent hires.

- The Company has created numerous proprietary processes and algorithms, but also makes it clear that it has no trademarked or patented technology. This is due primarily to fears of revealing code as part of the patenting process. While not unusual, and indeed standard practice in Europe, for software to not be patented, it may represent a degree of risk that other companies, or indeed clients, could replicate the Company's approach, or elements thereof. Whether this would ultimately be damaging to the Company is unknown.

Areas of investigation

Architecture and infrastructure

The Company's technical infrastructure appears robust. It utilises standard enterprise-level cloud servers, provided by Google via their Western European cluster situated in Belgium.

Back-up and failover provision should be sufficient to avoid data loss or significant downtime.

Software/APIs

The Company uses a large number of external APIs to process or interrogate data. For example, many of the Company's behavioural algorithms are powered by IBM Watson. Other APIs are, on the whole, well established and operated by legitimate, recognised companies.

The Company uses several SaaS solutions for managing and monitoring the infrastructure, debugging and other maintenance procedures.

External providers include:

- Scalyr
- Github

- IBM Watson
- MediaMath
- Qa-bot
- Bugsnag
- Luminati
- Google BigData

As far as we are aware, the Company is licensed to use these services. This list does not include standard hosted frameworks.
Open Source software and programming languages used by the Company are well established and appropriate for the development of their software. All of the following frameworks are self-hosted and managed by the development team.

- SQL (data query)
- Apache Spark (server software)
- Nginx (server software)
- NodeJs (server software)
- ReactJs (frontend)
- Scala (based on Java)
- Python (data science)
- [redacted]
- R (data analysis)
- GIT (version control)
- Jenkins (automation)
- Kubernetes (deployment management)
- Argo (workflow software for Kubernetes)
- Jupyter (data science)
- Prometheus/Grafana (monitoring)
- Velero (Kubernetes backup)
- [redacted]

[redacted]

Code review

A short, but insightful review of the main, backend code, was presented by the CTO [redacted]. Coding 'good practices' seemed to be apparent. Although some 'class' and 'method' naming conventions could be considered questionable. Also, from a high-level review, the code design doesn't present much flexibility or extensibility. This said, we believe the developers are active in refactoring the codebase to improve flexibility, which is a very normal procedure in most development houses.

Code samples are included in the Appendices.

Data

All key technical processes are focussed on extracting, processing, analysing and presenting data. Key steps are as follows:

- Identification of [redacted] through analysis of wide-scale datasets, sometimes provided or suggested by clients. The proprietary [redacted] database contains the tags the Company uses to identify [redacted]. The tags are applied manually.
- Data collation from social media platforms via API, or public web scraping, targeted using [redacted] from the first step.
- Analysis of collated data using proprietary algorithms to derive trends, patterns, sentiment, key media etc.

The Company provided high-level schemas for the key databases, all of which appear to be logically structured. These are included in the Appendices.

Resourcing

The Company has a small development team, with the CTO plus three developers, all of whom appear to be situated abroad. One also appears to be operating under his own firm. The Company also has a small [redacted] team consisting of a recently appointed [redacted] and three junior data analysts.

- [redacted] - CTO
- [redacted] - Programmer - based in [redacted] with c.15 years experience. Apparently not employed by the Company.
- [redacted] - Programmer - was based in London but moved back to [redacted] in November 2020 (according to his Facebook).
- [redacted]- Data Scientist/Back-end developer - apparently situated in [redacted]. Not included in the CVs provided by the Company.
- [redacted] - [redacted] - joined November 2020.
- [redacted] - Data Scientist. Recent MSc graduate with no corporate history.
- [redacted] / [redacted] - details not provided

Technical organisation chart [removed]

Operational processes

The Company focuses on back-end uptime, and they also have more user-focused metrics, such as user engagement, but given the hybrid model of the business they have chosen to focus less on these than would a pure-play platform.

IP and licencing

The Company does not possess any patented software due to its concerns about revealing proprietary algorithms.

Current unprotected IP consists of four main elements:

- The algorithms that analyse [redacted] and web interaction data and produce [redacted] including models that have been derived using machine learning from proprietary classification/training datasets

- The [redacted]' which has been developed over 5 years by [redacted] to directly address [redacted].

- [redacted], a high-performance vector database that finds the [redacted]

- The underlying platform that pulls these together with a data ingestion pipeline into a form that clients can directly access.

Further detail on each of these elements in the Appendices.

Security

Penetration (PEN) testing has not been conducted on any parts of the system at any point. A test can be carried out quickly, so this does not represent a long-term risk, but currently there is little knowledge about the external vulnerability of their platform.

The Company has taken the following procedures to ensure system security:

- A third party solution for system authentication (auth0 which has ISO 27001 and SOC 2 certifications)
- Developer security training
- Code review of all changes
- All [redacted] systems are behind both a Google Cloud Load Balancer (+ firewall) and an internal nginx instance.
- Internal security audits of all high-risk code.

Support and continuity

The Company uses several software packages to aid the identification and resolution of technical errors. These include Qa-bot, Scalyr, Bugsnag and Grafana.

The [redacted] team handles client reports and these are handled through a dedicated Slack channel that feeds into the technical team's error handling process.

The Company has a full backup strategy for its databases and infrastructure, and disaster recovery would involve recovering from these . The whole configuration for Kubernetes Engine is stored in Github source control and can be redeployed.

Further details can be found in the Appendices.

Scalability and development

Scalability does not appear to be an issue from a technical standpoint. The Company utilises enterprise-grade infrastructure in the form of Google Cloud / Kubernetes engine, which is scalable well beyond the requirements of such a business.
The Company has highlighted the increase in cost of scraping data but this is likely to be a cost proportionate to an increase in activity.

We have been provided with a high-level summary of the development roadmap, as follows:

- Doing deeper url analysis using a [redacted] to be able to surface more relevant urls and further increase analysis quality
- Integrating data from [redacted] thus adding to the existing data we get from these platforms already.
- Further integration of other platforms such as [redacted]
- Improve quality of [redacted] algorithm (mentioned above)
- Creating an index of unique, [redacted] built and tracked by [redacted], across which we will track over-arching [redacted] (currently we track [redacted]).

This has not been placed in the context of client deliverables, so we are unable to assess whether this fits with a wider strategic plan, or might deliver commercial benefits.

The Company uses a standard Kanban process for its development workflow, using the external Clubhouse software.

Further detail on the development cycle can be found in the Appendices.

Client proposition

The overarching proposition, as it stands, is to help [redacted]. This involves considerable onboarding and ongoing engagement time from the Account/Sales team. It would appear that while the platform provides a client dashboard through which the client can manage its own analysis, there is still a requirement for the Account team to assist and guide the client.

The Company told D4 that it does not wish to target [redacted], though it has considered doing so in the past. This is partly because of the challenges of breaking into a large group like [redacted] (particularly when they already have similar in-house analytical tools), and partly because they feel they are competing with the [redacted] in terms of [redacted].

A key issue is the limitation on client share-of-wallet. Even the largest global brands spend a relatively small proportion of their [redacted] on [redacted]. The vast majority of spend is allocated to [redacted] and [redacted]. A challenge for the Company will be how it can access these budgets through its platform, and how much investment it will require to do so.

Limitations and issues

The Company suggested one key limiting factor might be scraping web content at scale. They currently use a third party solution - [redacted] - but there are limitations on capacity and implications of increased cost. The Company is looking to reduce its dependence on scraping in general via greater use of sanctioned APIs.

Company summary

The Company is a data-driven consultancy and SaaS provider that assists large corporate [redacted] in defining [redacted] and [redacted] in a more scientific manner.

The Company's key selling point is its ability to extract, process and analyse large quantities of web data, from which they can determine [redacted] - e.g, [redacted], [redacted] and [redacted].

While much of this data processing takes place out of sight, clients are given a dashboard through which they can create and manage specific activity and analytics.

The D4 investigation comprised the following:

- Provision of documents covering areas requested in the D4 Technical Due Diligence framework. These were uploaded to the D4 dataroom.
- An initial web conference to introduce D4 and cover the key areas for investigation.
- An in-depth web conference during which the client described the business proposition, and operational and technical workflow. D4 raised questions.
- A walk-through of code examples via web conference.

Appendices

Infrastructure diagram [removed]
Kubernetes architecture diagram [removed]
Data flow diagram [removed]
IP and Algorithm breakdown [removed]
Code samples [removed]
Development workflow detail [removed]
Full development roadmap [removed]
Platform documentation [removed]
Staff CVs [removed]

Disclaimer

All the information contained in this document has been collected via engagement with the Company and from publicly available sources. It is regarded as correct at the time of writing.

D4 Analysis bears no responsibility for the veracity of the information presented or for any implications of disseminating this information to the client ([redacted]).

D4 Technical Due Diligence reports are designed to identify prominent issues that could represent risk to the Client. They should not be seen as thorough investigations and information may be missed that subsequently becomes more apparent during further research.

Please note: this disclaimer is work in progress and does not represent the full scope of D4 Analysis' indemnity.

d4analysis.com

contact@d4analysis.com