# Risk areas

## Hosting and infrastructure

### Executive summary

TalkWalker's choice of infrastructure vendor is a limitation.

### Detail

TalkWalker uses Hetzner as an infrastructure provider. For a product like TalkWalker, Hetzner seems on the lower end, regarding apparent scale and experience.

The supplied document describing production issues and notifications supports this impression, with blown fuses, power outages, and routing issues in 2017.

TalkWalker is likely to be one of the largest clients Hetzner has, assuming investment-friendly levels of business growth. It's likely the choice of infrastructure provider becomes an increasing liability.

## Architecture re-establishment and relocation

### Executive summary

TalkWalker's ability to restore or move their production is likely to be limited.

### Detail

TalkWalker has made a conscious choice to maintain the entire client base in a single instance of their platform. This has distinct advantages, such as high levels of control over functionality across all clients.

One interesting consequence of this - in conjunction with their apparently long experience with their hosting provider - is that they are likely to have relatively little experience with establishing and deploying their software, as opposed to maintaining and updating it.

Assuming correct backups and restore procedures, the response to a broader failure (such as widespread hardware failure, security compromise, etc.) is to re-establish the systems. It is not clear how prepared TalkWalker is for such a contingency.

The conventional wisdom is to assume restores from backup fails. An experienced operations team is aware of this and periodically tests backups and restore procedures; it may be TalkWalker does this, or the CISO being recruited will institute this, but it remains a risk factor.

Data backups appear to be stored in the same physical location where production lives.

In some areas, the TalkWalker technical architecture describes backups as what seems more appropriate to call failover or high-availability mechanisms. These are good to have, but they should not serve as backups, especially when sharing a physical infrastructure with the systems they are serving backups for.

Connected to their choice of infrastructure provider, should TalkWalker need to move away from Hetzner, the resulting project is likely to be non-trivial in scope and size.

## The platform is in technical adolescence

### Executive summary

The architecture is still evolving, and so is the technical team's understanding of it. As a useful metaphor, the TalkWalker platform is in technical adolescence.

### Detail

TalkWalker's production issues log paints a picture of a technical team which is still learning how their architecture performs (and importantly, how it fails) under scale and load.

There is abundant evidence that the technical team is capable. This observation is more a matter of correctly identifying the project and products stage in the maturity life-cycle.

It is a well-designed architecture, but it is clear that it is not immune to teething problems. The chances of a client's usage patterns affecting availability for other clients is not low.

It's likely that the type of end user who sits in front of TalkWalker and uses it on a daily basis probably isn't too concerned with the difference between two and four nines. However, when the architecture is multi-tenant, one hour of downtime is multiplied by the entire client base and is proportionally more dramatic.

Downtime is one potential failure condition; data loss is another. We cannot currently say that the risk of a data loss scenario is non-existent, given the (otherwise quite natural) frequency of data source or data handling adjustments described in the production issues.

To be clear, this is not a reflection on the technical team; any sufficiently complex system will pass through stages of understanding and hardening. To mitigate this factor, we remember that the nature of the data in question here is individually not significant. For example, losing a few tweets does not have the same consequences as a health information system or trades in a brokerage losing a few medical journals.

## GDPR and data-oriented legislation

### Executive summary

Current and upcoming legislation concerning privacy and data protection represent a compliance risk.

### Detail

TalkWalker is aware of the impending GDPR, and in our discussions, their response is twofold.

First, they are going to hire a CISO. The quality of their technical organization suggests such a hire would make a real positive difference.

Secondly, our understanding is that TalkWalker does not see their competitors as being any more in compliance than they are. This isn't reassuring, even if it is a very practical stance.

Even though it is easy to forget, the GDPR is not the only piece of legislation which might impact TalkWalker. As an example, the ECJ upholds citizens' "a right to be forgotten." A complete understanding of the legal impact is not in scope here, but from a technical perspective, the implementation of any form of "forget this person" functionality is not going to be cheap or easy.

We have no opinions on the scale of this risk, only its existence.

## Rights on media types

### Executive summary

TalkWalker's core is a large dataset in which clients operate. Whether this source data will or will not require a license is unclear. It is also unclear how much of TalkWalker's current licensing efforts cover it, and how it might scale in the future.

### Detail

While quite a lot of the sources TalkWalker harvests are performed via approved APIs subject to formally defined agreements, they do acquire what appears to be a non-trivial amount of data from the internet, e.g., blogs and web forums.

TalkWalker has touched briefly on this concern, stating they respect websites' robots.txt (instructions to web crawlers suggesting what is allowed) but this is separate from usage rights on the indexed material. For example, an indexed blog might be placed under a Creative Commons license prohibiting commercial derivatives.

Whether Elasticsearch indexes constitute derivative works is a legal matter beyond our scope. Nevertheless, it isn't clear whether the handling of media source content licenses beyond formal commercial agreements is at the correct level.

TalkWalker mentioned having a person working on these matters by coordinating with online commercial entities. The risk here is that the volume of this type of work is in our opinion not guaranteed to remain near current levels.

# Opportunities

## A homogenizing architecture may have other applications

### Executive summary

The TalkWalker architecture specializes in homogenizing very diverse data. This might lend itself to monetizing beyond social media.

### Detail

The architecture TalkWalker have produced might be seen more generally as a system specializing in the acquisition, rationalization, and exposition of data sourced from heterogeneous third party data sources. From this perspective, the focus on social media analytics and monitoring is only a special case.

It is worth considering if the same or a similar architecture can be applied to other domains. In which case, the cost of developing, maintaining, and extending the architecture is spread across a greater business spectrum, and the R&D cost is amortized more broadly.

In particular, the TalkWalker architecture capability to handle time-series creates business value in data which otherwise was not apparently so valuable, and therefore worthy of stockpiling.