



MMC Ventures
Data observability
– the rise of the
data guardians

Data observability – the rise of the data guardians

AI and analytics are penetrating every sector and business function, driving efficiencies and creating new revenue opportunities. However, a bottleneck to realising the benefits of these data-driven applications is bad data quality; garbage in, garbage out. Bad data is becoming progressively harder to guard against as data volumes skyrocket and data pipelines grow ever more complex. Gartner now estimates that bad data quality causes a business to lose, on average, \$12.9m a year. As a result of this growing pain point, a new product category has emerged to protect companies against bad data: data observability.

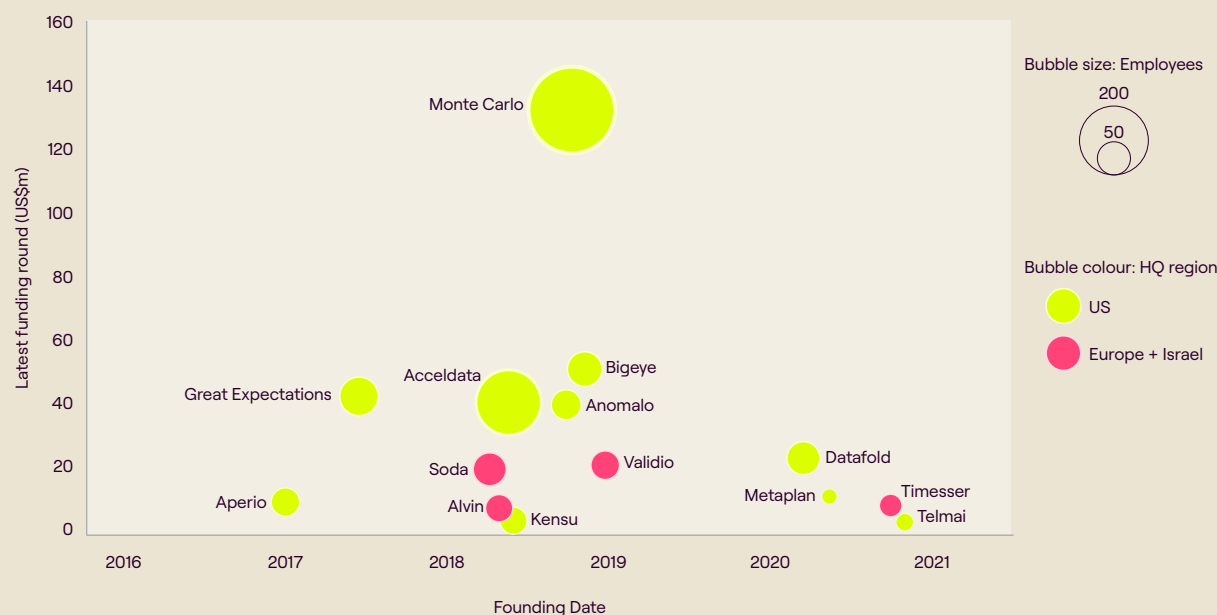
In this note, we outline what you need to know about data observability, why you should care and some of the market dynamics at play, sharing our learnings from over 40 interviews with vendors, data practitioners and industry experts. Overall, we believe this category will continue to rapidly expand, driven by the growing importance of data quality and the complexity of the problem. We also expect consolidation over the long-term given the proliferation of data tools. For now however, we are still in the rapid innovation phase, with data observability solutions starting to become a key component of the modern data stack.

The north star for data observability solutions is to reduce the time to detect and resolve data quality issues. By using machine learning, data observability solutions gain an understanding of what is considered normal activity for a business’s data, and send alerts if a deviation occurs. We at MMC Ventures have identified 26 start-ups and open-source projects in this space, all with a unique approach. These solutions have been met with much enthusiasm from data teams, and have also had a warm reception from investors raising over \$370m in their latest financing rounds.

Bad data quality causes a business to lose, on average, \$12.9m a year

SOURCE: GARTNER

Data observability landscape



Note: Employee count from LinkedIn (20/11/2022)
Source: Company press releases, Crunchbase, LinkedIn

MMC Ventures is one of the most active early-stage tech investors in Europe, focused on Series A. During the past two decades, we've built a successful track record of supporting high-growth technology companies. We distinguish ourselves through our commitment of going deeper – on the technologies we invest in, and the partnerships we build with founders.

We conduct in-house research, providing us with a differentiated understanding of emerging technologies and sector dynamics to identify the areas and themes that have the potential to create the next multi-billion European success stories.

For many years we've been focused on AI and its potential to shape a wide range of sectors. Off the back of our research, we have built a large portfolio of incredible entrepreneurs leveraging technology to do amazing things, including the likes of Peak AI, Synthesia, and Signal AI. In our journey with these companies, we also saw the shift from monolithic data infrastructure architectures to best of breed solutions. Here we have built a portfolio we are proud of, which includes Quix, Snowplow, Ably, MindsDB, Tyk and Cloudsmith.

Although we have seen substantial innovation in AI and data infrastructure, we are still in the early phases of adoption. One challenge to further adoption of these technologies is the quality of the data that powers these applications. As a result, in recent years, there has been a proliferation of solutions that help prevent, detect and resolve data quality issues. This category has been dubbed data observability. Given we see these solutions as key enablers for data-driven businesses to thrive and for continued adoption of AI, in this note we do a deep dive on the space to understand how they work, how they are differentiated, and where they are heading.

Author:

Oyvind Bjerke, Research Manager

Oyvind's role is to lead and coordinate MMC's research activities, supporting the investment team as they go deeper into the technologies we invest in.

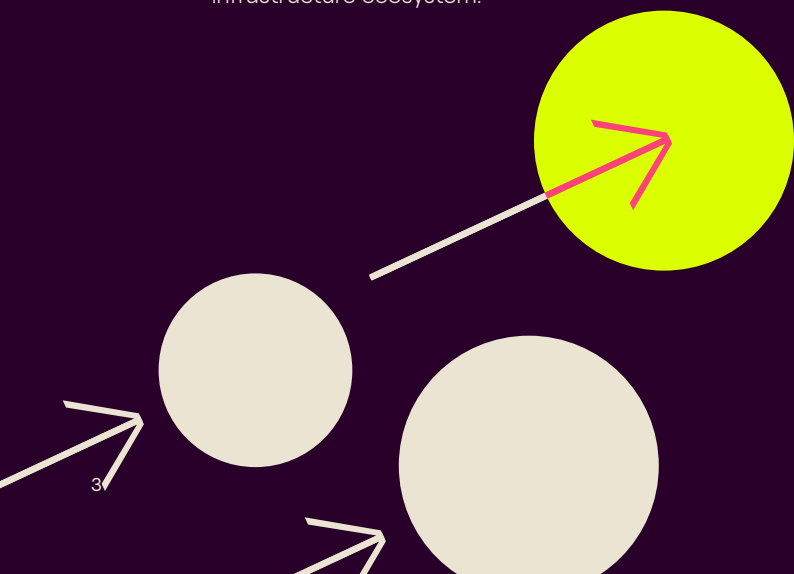
Contributor:

Nitish Malhotra, Investment Associate

Nitish is an Associate in the Investment team. He loves to dig into enterprise software topics, with a keen interest in the vibrant cloud, DevOps and data infrastructure ecosystem.

If you're an entrepreneur building something in this area please reach out, we love learning about new ideas!

EMAIL: oyvind@mmc.vc or nitish@mmc.vc



Contents

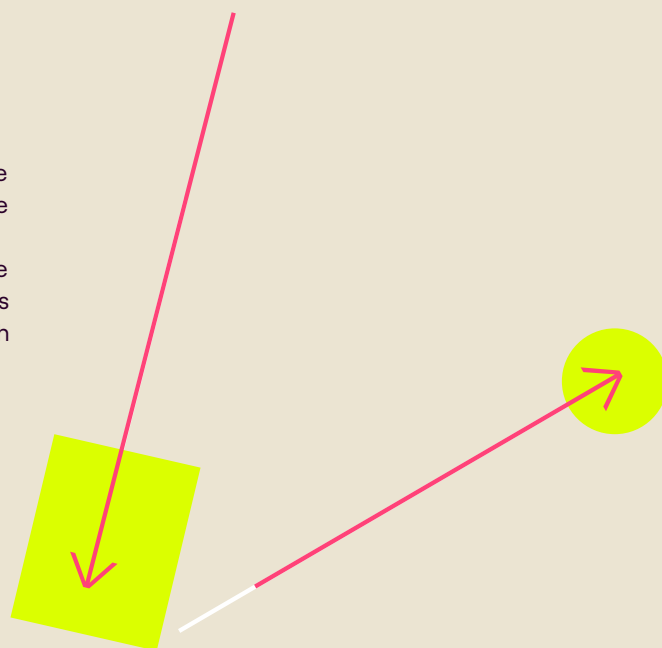
Executive summary	5
1. Setting the scene	10
The old guard – data quality solutions are not new	10
Big data – mo’ data mo’ problems	11
Big tech – pioneering the next generation of data quality tools	11
Next generation data quality tools go to the market – data quality for all	12
Category and vendor definitions – finding their identity	14
2. Core concepts	15
Integrations – setting up the perimeter	16
What is monitored – the usual suspects	16
Anomaly detection – sounding the alarm	17
Root-cause-analysis – tracking down the culprit	18
3. Differentiation and further development	19
Metrics – depth of tests	20
Integrations – shift left	20
Root-cause-analysis – granularity is key	22
User friendliness – data quality for all	23
Scalability – going big without going broke	25
Prevention capabilities – pre-empting bad data	26
4. Market dynamics	28
Pricing and pricing power – differentiation is the key	28
Disruption and consolidation – a landscape poised for change	30
5. Appendix: Data observability in operational technology	34

Executive summary

Here we highlight some key takeaways from the report and outline how data observability tools came to be, how they work, how they differentiate, and what we think lies ahead.

What you need to know about the data observability space:

- **Large greenfield opportunity** – The data observability category is still at its nascent stages, but we expect massive growth going forward as every sector will increase its adoption of analytics and AI where data quality is paramount.
- **Still a ton of innovation to come** – Leading vendors are coming up with new functionality every quarter, as well as building out their suite of integrations. A particularly exciting area here is ‘shift left’ developments, which involve technologies that catch and prevent bad data earlier in the data lifecycle, before it can cause harm downstream.
- **Consolidation on the horizon** – We believe we will see substantial consolidation of data solutions over the next five years. Specifically, we see a natural category merge between data observability and data catalogues due to overlapping capabilities and potential synergies. This potentially means some significant exits in the coming years, but vendors also need to build out their products with half an eye on where they might fit in the landscape as the market heads towards a consolidation phase.
- **Disruption from larger tech players** – There are worries that players within the modern data stack, such as dbt and Snowflake, could infringe on the data observability space as they are providing some similar capabilities. In the medium term, we think they may continue to expand these capabilities, but will remain confined to their area of the stack. In contrast, data observability will have an advantage of covering the stack more broadly. However, the wildcards are the established software observability companies, such as Datadog, who many believe their next logical step is to enter the data observability space.
- **The operational technology market opportunity** – We see a significant opportunity for data observability in the operational technology (OT) market (e.g. energy, manufacturing sectors). This is due to the massive OT market becoming increasingly data-driven. Most of the IT data observability players we discuss in this report are not at all focused on the OT opportunity today, and we do not expect them to do so over the next few years. We therefore believe this will remain a separate category likely to see its own winners emerge.



There is a growing dependency on (high quality) data across every sector.

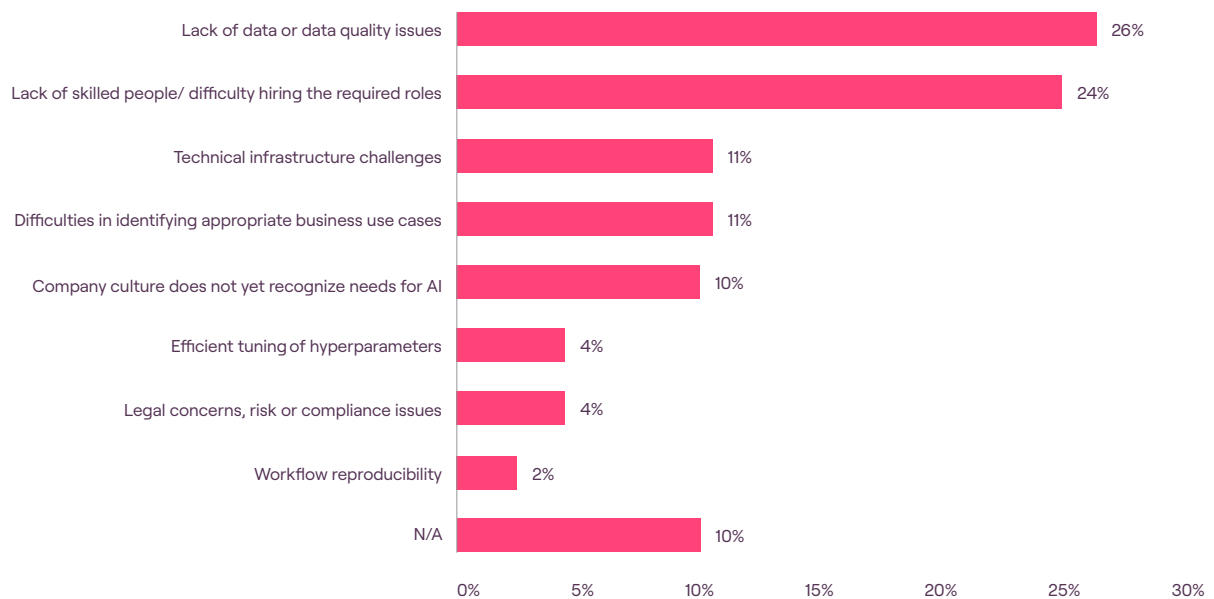
Data-driven applications are increasingly empowering organisations in a variety of ways.

For instance, BI dashboards give employees insights into business performance, improving decisions, while machine learning models can improve inventory management by predicting demand to optimise stock levels. This growing dependency on data applications also means the consequences of bad data have increased.

For instance, a change to a table upstream can accidentally break a BI dashboard, leading to wrong decisions, or an incorrect format in an address can cause inventory to be sent to the wrong warehouse, leading to dead stock. In a recent public example, Unity estimated it would lose c\$110m due in part to the ingestion of bad data into its models.¹ As shown in the survey below, a lack of data and data quality issues is the main bottleneck for further AI adoption.

AI adoption survey – Lack of data or data quality issues is the main bottleneck

Question: What is the main bottleneck holding back further AI adoption? (select one)



Note: Respondents were those considered AI mature – those who already used AI in analysis or production

Source: The quest for high-quality data – O'Reilly (oreilly.com)

¹Unity loses \$110m due in part to bad data

We are seeing the rise of a next generation of data quality tools

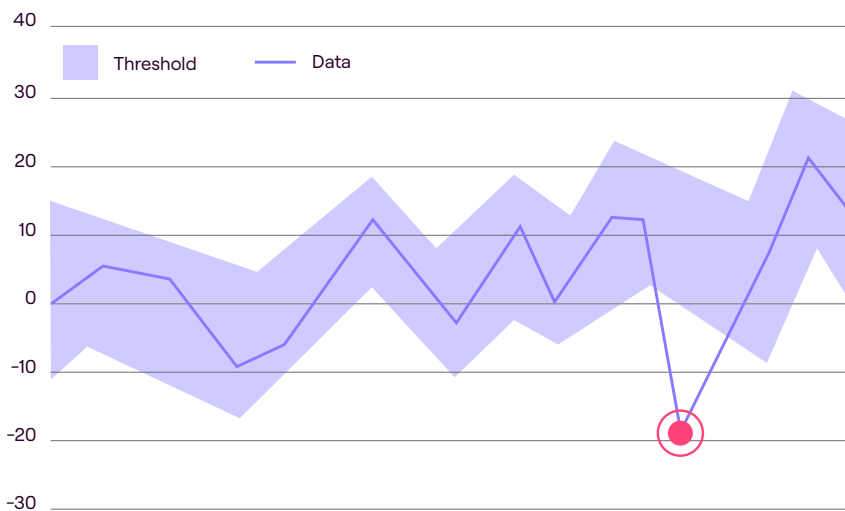
To combat bad data, companies have historically relied on traditional data quality tools. These tools were either purchased from legacy vendors such as Informatica or Talend, or created internally by the companies themselves. Although initially effective, these solutions failed to scale when companies began to embrace Hadoop and Cloud computing, where data volumes became orders of magnitude larger. As such, new open-source projects and start-ups emerged to address this issue, leveraging automation and machine learning (ML) to tackle larger data volumes.

Two of the most well-known solutions are Great Expectations (GX) and Monte Carlo, founded in 2017 and 2019, respectively. These tools saw rapid adoption, which investors recognised, and already by 2022 Monte Carlo reached a \$1.6bn valuation. Monte Carlo's founder, Barr Moses, coined the term data observability for this new category, inspired by the software observability companies from the DevOps world, such as Datadog. In total we have now identified 26 solutions in the space. Looking to the future, we expect substantial continued long-term growth for the data observability category. This will not only be driven by demand from current businesses that are in need of improving their data quality, but from the large number of companies that will inevitably have to become more data-driven to remain competitive.

The basics of how data observability solutions work

- **Integrations** – Data observability solutions typically collect data from the data warehouse, but will also collect data from many other places in the data stack (eg BI tools, traditional databases) depending on the solution. They also integrate with multiple other tools in the stack such as Slack in order to send alerts to data teams when issues are detected.
- **Defining bad data** – There are many forms of bad data. Some of the most common are: data not updating on time, too few or too many table rows, and issues related to schema changes.
- **Catching bad data** – Solutions will typically use anomaly detection to catch bad data. Machine learning models will forecast what they expect the data to look like based on historical data, taking into account factors such as seasonality. If the actual data goes outside of the forecasted thresholds, an anomaly alert is sent to the data team. For instance, if a table normally has 10 million rows, and this suddenly drops to 1 million, there is a good chance something has gone wrong and an alert will be sent.

Anomaly detection – red dot represents an anomaly as data went outside of forecasted threshold



Source: Datafold

● **Root-cause-analysis (RCA)** – There are also several root-cause-analysis tools on the market that help track down the origins of the bad data. The two most common ones are:

● **Data lineage** – Provides visibility on how data flows through the data stack. This allows users to see both where the issue originated upstream e.g. data ingestion, and what could be impacted further downstream, e.g. BI dashboards or machine learning models.

● **Segment analysis** – Pinpoints where in the underlying data the anomaly occurred. (It can also significantly enhance the discovery of data quality issues.)

Emerging winners are mainly differentiating through...

● **Root-cause-analysis** – Many solutions have an RCA tool, where they tend to have either data lineage or segment analysis. The RCA tool is often a main deciding factor for customers, and therefore several vendors aspire to eventually have both. In our view, segment analysis is the more differentiated as it is less common.

● **Integrations** – Vendors are constantly increasing their number of integrations, allowing their solutions to be more useful and suitable to a broader set of customers. A particular point of differentiation here is how far upstream solutions can validate data (i.e. shift left) as they want to catch issues as early as possible.

● **Prevention tools** – These are tools that help reduce bad data occurring in the first place by pushing data quality into an earlier phase of the data lifecycle (i.e. shift left tools). A few vendors have recently begun to introduce these types of tools, e.g. 'data contracts' and 'Data Diffs'. These tools have generally been well received, and we think prevention tools in general have a lot of promise and we would expect more to emerge.

● **Scalability** – The ability to handle large volumes of data efficiently while minimising costs is a key differentiator for large enterprise customers. Vendors with good scalability capabilities have been able to displace competitors on this basis.

● **User friendliness** – User-friendliness largely comes down to how easy the solution is to deploy, learn and use. This has been an integral part of Monte Carlo's success, as customers can more quickly get value from the solution and it can be adopted more widely across the business.

● **Depth of testing** – This is the number of metrics that can be tested. This is unlikely to be the main deciding factor in purchasing decisions, but something that is often considered.

Price scales with consumption and vendors believe they have solid pricing power

● **Pricing model** – There has been a general shift towards a consumption-based pricing model (away from platform fees), usually based on the number of tables monitored or data processed (which tends to increase over time).

● **Price** – A few vendors that cater to smaller companies charge as little as \$3.6k per annum for a starting package. Vendors focused on more mature companies typically charge between c\$100k–300k per annum (c20–25% of Data Warehouse cost, n = 3). We have not yet heard of a contract north of \$1m, but it's clear how this will happen based on scaling to larger datasets.

● **Pricing power** – Vendors catering to smaller customers tend to be product-led and compete on price. Most vendors however focus on functionality, integrations and user-friendliness. Founders remain confident they can maintain pricing power as long they can offer superior value given the importance of data quality. Additionally, they believe they can weather an economic downturn, not only due to data quality being critical, but because they are cost savers as they reduce labour costs. Some can also help reduce the data warehouse bill.

This remains an emerging category with potential disruption and expected consolidation

● **Disruption from the larger tech players in the stack** – There are worries that players within the data stack, such as dbt, Snowflake, Fivetran and Airflow, could disrupt start-ups in the space. dbt has already taken market share in data testing and Airflow and Snowflake now offer lineage. Some founders are less concerned, as they believe these larger vendors will stay confined to their area of the stack, whereas data observability tools will have an advantage as they cover the stack more broadly. We generally believe the more sophisticated tools with broad integrations and differentiated functionality are currently not under much threat from these moves.

● **Data observability and data catalogue consolidation** – We think that these two categories will begin to merge as products increasingly overlap and enterprises demand broader suites from fewer vendors. Collibra's acquisition of OwIDQ was an early example of this.² We believe deep data quality vendors (with good scalability, segment analysis and shift-left capabilities) will be best-positioned here, as both potential targets or acquires, as they are the most differentiated from data catalogues. However, we think a large-scale consolidation at this stage would be premature given both categories are still undergoing rapid innovation and consolidation would hinder that process.

² Collibra acquires OwIDQ

There are similar challenges in the industrial market, but surprisingly little overlap

Enterprises in markets such as manufacturing, energy and chemicals are experiencing their own explosion of data and data quality issues with the proliferation of machine sensors and AI systems. We have seen how bad data can create plant downtime, with automotive manufacturers losing more than \$2m on average on a single hour of lost production.³ This space, often referred to as 'operational technology' (OT) rather than IT, has a massive TAM, but so far we have only identified two pureplay OT data observability companies in the space.

Although OT and IT data observability solutions have many similarities conceptually, there are many practical differences: they operate in different stacks, with different data, different use cases and different customers. Speaking to founders in the IT data observability space, they appeared to have no intentions of entering the OT space any time soon, and most were unaware of its existence. This means competition will remain limited for some time to come.

The main challenge for OT data observability vendors is the current lack of digital maturity of customers, both in terms of assets and culture, which has meant these industries have generally not been quite ready to embrace data observability solutions. However, we see a clear trajectory of this changing, and we therefore think this is an exciting area for new entrants and a space to watch.



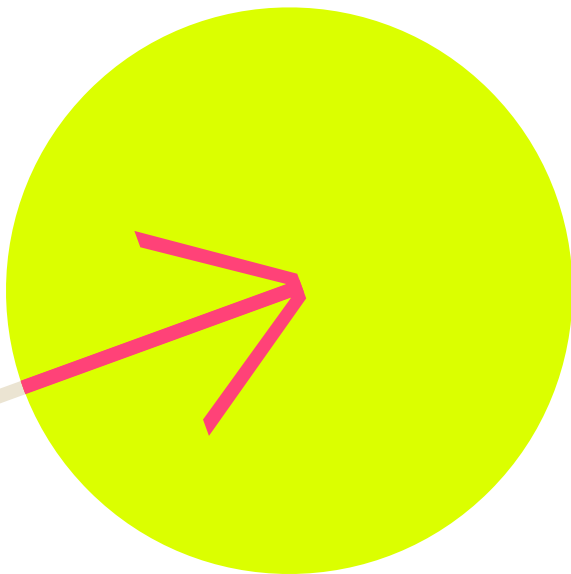
We've seen how bad data can create plant downtime, with automotive manufacturers losing more than \$2m on average on single hour of lost production.

³The True Cost of Downtime 2022 – Senseye Industry Insights' (n= 56)

1. Setting the scene

The old guard – data quality solutions are not new

For as long as there has been data, there has also been bad data. As IT technology evolved over the decades, data use cases and volumes expanded, creating ever more data quality problems. These problems demanded solutions, and by the 1990s various types of tools were on the market.⁴ By the dawn of the new millennium, larger tech companies saw the growing importance of offering their customers data quality tools, sparking a wave of acquisitions, including Ascential's acquisition of Vality for \$94m in 2002, who themselves were acquired by IBM in 2005.⁵ ⁶ Then in 2006, Informatica bought the data quality company Similarity for \$55m, as they felt threatened by industry consolidation, particularly by IBM.⁷ ⁸ Additionally, new vendors emerged, such as Talend, which launched Talend Data Quality in 2008.⁹ These vendors, along other familiar names such as SAP, Oracle and Precisely, are the ones seen today in Gartner's Magic Quadrant for data quality solutions.¹⁰



⁴ Survey of data quality tools from the 1990s

⁵ Ascential acquires Vality for \$94m

⁶ IBM acquires Ascential

⁷ Informatica acquires Similarity for \$55m

⁸ Informatica feels threatened by industry consolidation

⁹ Talend launches Talend Data Quality

¹⁰ Gartner's Magic Quadrant for data quality solutions

Big data – mo’ data mo’ problems

“Companies that were generating 10TB of data a year, are now generating 10TB of data a day.”

Rohit Choudhary, Founder and CEO, Acceldata, Making Data Simple (April 2022)

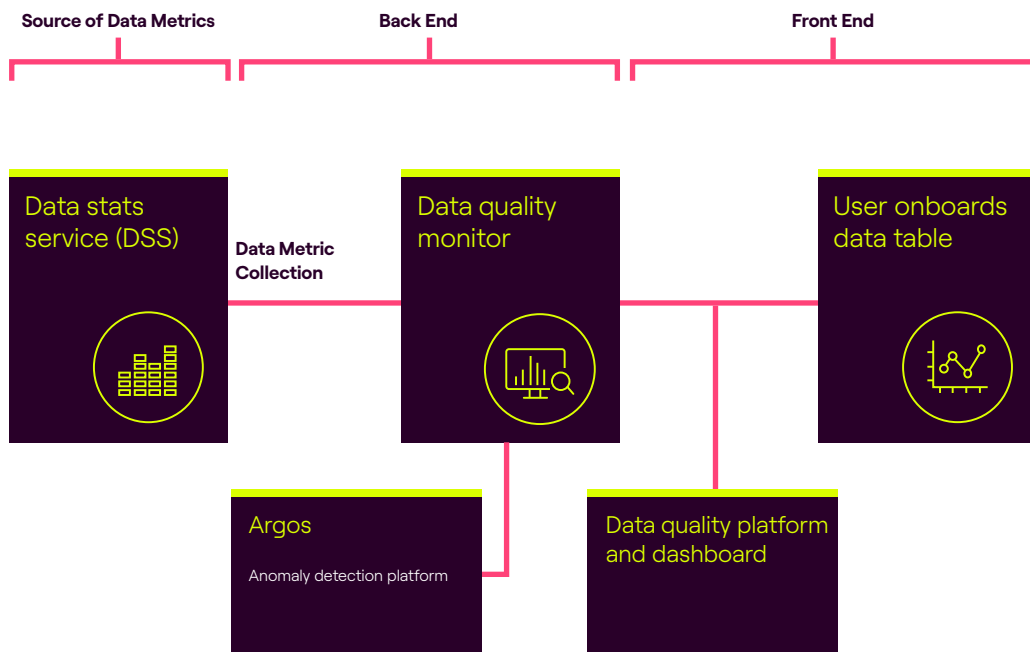
Traditional data quality tools worked well enough for their time. They allowed manual testing for things such as data duplicates and missing values, often as one-off tests before the data was consumed. However, as companies started embracing Big Data and technologies like Hadoop, and later cloud computing, data volumes exploded. Traditional data quality tools were no longer fit for purpose as they lacked scalability and automation. Companies would often instead create their own, ad-hoc data quality tests, but this was highly time-consuming, and typically only caught 20% of the data quality issues.¹¹ According to a survey, data professionals spend 40% of their time on data quality.¹² It also usually takes over four hours to detect data issues and an average of nine hours to resolve, meaning there is plenty of time for bad data to do harm. Therefore, a lot of time has been spent dealing with issues once the bad data has already affected data users. As a result, data engineers were often stuck firefighting data quality problems rather than moving the business forward by designing new data systems and implementing new tools.

Big tech – pioneering the next generation of data quality tools

Prominent tech players such as Uber, LinkedIn and Airbnb were some of the earlier companies to face these challenges, given their vast data volumes and the real-time nature of some of their products.^{13 14 15} As such, they began to develop their own tools and architecture to fend off bad data.

For instance, Uber launched its platform Argos in 2014, which used anomaly detection to automatically identify data quality issues.¹⁶ Over the years, Uber continued to iterate on its data quality tools. In 2020, Uber launched its Data Quality Monitor, which “automatically locates the most destructive anomalies and alerts data table owners to check the source, but without flagging so many errors that owners become overwhelmed.” By investing in sophisticated data quality platforms, big tech companies reduced bottle necks in the data stack and substantially improved efficiencies, allowing them to break free from the shackles of bad data and continue to scale.

Uber’s Data Quality Anomaly Detection Architecture



Source: Uber

¹¹ Testing only catches 20% of data issues

¹² Survey: The State of Data Quality, 2022'

¹³ Uber blog, 'Monitoring Data Quality at Scale with Statistical Modeling'

¹⁴ LinkedIn blog, 'Towards data quality management at LinkedIn'

¹⁵ Airbnb blog, 'Data Quality at Airbnb'

¹⁶ Uber blog, 'Identifying Outages with Argos, Uber Engineering's Real-Time Monitoring and Root-Cause Exploration Tool'

Next generation data quality tools go to the market – data quality for all

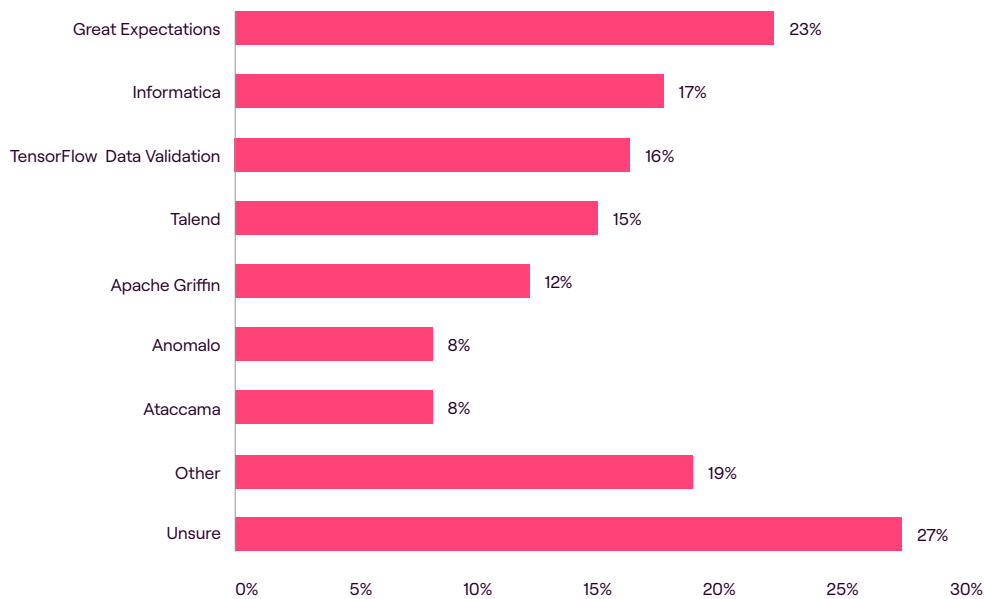
Although the larger tech pioneers were able to adapt by building their own sophisticated data quality solutions, the vast majority of companies did not. They often lacked the budget and know-how, given they were not primarily data businesses. As such, open-source projects and start-ups began emerging in the late 2010s to address the new paradigm of data quality problems.

Great Expectations – open-source takes the lead

One of the earliest tools to tackle this issue was Great Expectations (GX).¹⁷ Abe Gong and James Campbell founded GX (formally known as Superconductive) in 2017, and in early 2018 launched GX as an open-source project. GX provided a framework for asserting expectations for what the data should look like with a flexible declarative language and automated testing at batch time. It quickly became the most popular open-source data quality platform and in a 2021 survey of data practitioners, 23% said their organisation used GX.¹⁸ In February 2022, the company raised \$40m in a Series B which it is using to develop its first commercial product.

Data Quality Survey – Great Expectations leading market share

Question: What tools does your organisation use to improve data quality? (check all that apply)



Source: 2022 State of Data Engineering

Note: n = 372 respondents, 39% Data Engineers, 13% Data Architects. Survey ran between 24 June to 23 August, 2021.

¹⁷ Introducing Great Expectations

¹⁸ '2022 State of Data Engineering'

Bigeye – bringing big tech solutions to the general market

Kyle Kirwan and Egor Gryaznov were early employees at Uber, where they were part of the team that built tools to manage the massive scope and scale of Uber's data. This included data catalogues, data lineage, freshness tracking, quality testing and incident management. In 2019, Kirwan and Gryaznov founded Bigeye (previously known as Toro) with the aim of making the tools they had at Uber available to the general market. When choosing a focus area, the founders spoke to data engineers from various companies and noticed that data quality was one of the most significant pain points, and chose this as their focus. In April 2021, Bigeye raised \$17m in a Series A, with backing from Olivier Pomel, founder of the highly successful software observability company Datadog. In September 2021, Bigeye raised \$45m in a Series B.

Monte Carlo – the loudest voice

Monte Carlo was founded in 2019 by Barr Moses and Lior Gavish. In many ways Monte Carlo is now considered the market leader in the commercial space. Although we are not aware of any public disclosures on its revenue size and users, it has raised the most money by a significant margin, with \$236m raised in a 20-month period, and a \$1.6bn valuation in its latest round in May 2022¹⁹. The company also boasted a 100% customer retention during 2021 and between its Series C and D it doubled revenues every quarter.²⁰ Moses has also been highly influential in shaping the narrative of the category, and is credited with coining the term 'data observability'.²¹ Monte Carlo's success has not only been driven by being the most vocal vendor in the space, but also for providing a highly user-friendly product compared to many of its competitors.

All in all, we have identified 26 start-ups, scale-ups and open-source projects in the data observability space - each with their own unique approach.

Data observability start-ups

COMPANY/SOLUTIONS	FOUNDING YEAR	DESCRIPTION	LATEST FINANCING			EMPLOYEES (REGISTERED ON LINKEDIN)	
			Type	Date	Amount (US\$m)	Number	Growth (YoY)
Monte Carlo	2019	Data Observability	Series D	2022	135.0	216	159%
Acceldata	2018	Data Observability	Series B	2021	35.0	169	49%
Great Expectations	2017	Data Quality	Series B	2022	40.0	60	50%
Bigeye	2019	Data Observability	Series C	2021	45.0	59	51%
Soda	2018	Data Quality	Series A	2021	13.8	55	77%
Anomalo	2018	Data Quality	Series A	2021	33.0	44	266%
Validio	2019	Data Quality	Seed	2022	15.0	37	105%
Datafold	2020	Data Observability	Series A	2021	20.0	36	100%
Kensu	2018	Data Observability	Seed	2022	4.2	35	34%
Alvin	2018	Data Lineage	Seed	2022	6.0	32	255%
Aperio	2017	Data Quality*	Series A	2020	8.5	28	27%
Lightup Data	2019	Data Quality	n/a	n/a	n/a	25	78%
Sifflet	2021	Data Observability	Seed	2021	n/a	23	109%
Timeseer	2020	Data Quality*	Seed	2022	6.0	16	60%
Telmai	2020	Data Quality	Seed	2021	2.8	13	18%
Metaplane	2020	Data Observability	n/a	2023	8.4	11	120%
Elementary	2021	Data Observability	n/a	n/a	n/a	6	200%
re_data	2020	Data Observability	Pre-seed	2021	n/a	5	25%
Cito	2021	Data Observability	n/a	n/a	n/a	2	n/a
Apache Griffin	2016	Data Quality	n/a	n/a	n/a	n/a	n/a
Deequ	2019	Data Quality	n/a	n/a	n/a	n/a	n/a
Databand	2018	Data Observability	Acquisition (IBM)	2022	n/a	n/a	n/a
Datakin	2019	Data Lineage	Acquisition (Astronomer)	2022	n/a	n/a	n/a
Holoclean	2019	n/a	Acquisition (Apple)	2020	n/a	n/a	n/a
OwIDQ	2017	Data Observability	Acquisition (Collibra)	2021	n/a	n/a	n/a
TOTAL					373	872	

KEY US Europe + Israel

**used for operational technology*
Note: Employee count from LinkedIn (20/11/2022)
Source: Company accounts, Crunchbase, LinkedIn

¹⁹ Monte Carlo Series D

²⁰ Monte Carlo 100% customer retention

²¹ Moses coins the term Data Observability

Category and vendor definitions – finding their identity


The term data observability draws inspiration from the software observability platforms used in DevOps. Datadog is a leader in this space, which was founded in 2010 and is now a listed company with \$1bn in revenue in 2021, growing 70% YoY.

Software observability platforms give visibility into a company's IT infrastructure to track the health of its systems, send alerts if something is wrong and allow users to drill down into granular details when needed. Data observability companies are therefore not reinventing the wheel, but rather leveraging many of the same principles of tracking, monitoring and triage, but for DataOps rather than DevOps. Many Data observability companies will refer to themselves as 'the Datadog for data'.

Data observability is now generally what the category is referred to and has also become a recognised category by Gartner.²² According to Moses, data observability is "an organisation's ability to fully understand the health of the data in their systems."²³ Furthermore, she describes data observability solutions as using "automated monitoring, automated root cause analysis, Data Lineage and data health insights to detect, resolve, and prevent data anomalies".

Although the term 'data observability' has seen widespread adoption, there is still an abundance of terms used to describe different tools and concepts in the space: data quality, Data Lineage, data reliability, data trust, data testing, pipeline testing, data health, data integrity, data monitoring, metadata observability, data quality engineering, data integrity, pipeline monitoring and deep data observability.

Speaking with the founders, it was challenging to find consensus of what all the different terms meant. Many companies have embraced the term data observability for their product, likely due to the wide mindshare of the term. Others resisted its definition, wishing to signal differentiation in their product offering.



Manu Bansal
Co-Founder of Data Quality platform Lightup Data

Data Quality \neq Data Observability
Data Quality \neq Data Observability
Data Quality \neq Data Observability

Source: LinkedIn

For this report, we refer to the whole category as data observability (sorry, Manu!), but split it into two sub-categories: data quality and Data Lineage. Data quality tools for our purposes generally refers to tools that check that data is accurate, complete, consistent, and timely.²⁴ Data lineage tools on the other hand show how the data flows throughout the stack, which helps identify where in the stack the bad data originated and what it could impact. Platforms that have both capabilities we refer to as data observability. This taxonomy, albeit not perfect, is helpful for our purposes.

²² Gartner's definition of data observability

²³ Barr Moses' definition of data observability

²⁴ DAMA UK's definition of data quality

2. Core concepts

In this section, the core concepts of data observability are outlined to illustrate how they work in practice:

Integrations

where in the stack data observability tools sit



What is monitored

the data quality issues which are being checked for



Anomaly detection

how bad data is caught



Root-cause-analysis

the tools used to find where the bad data quality originated

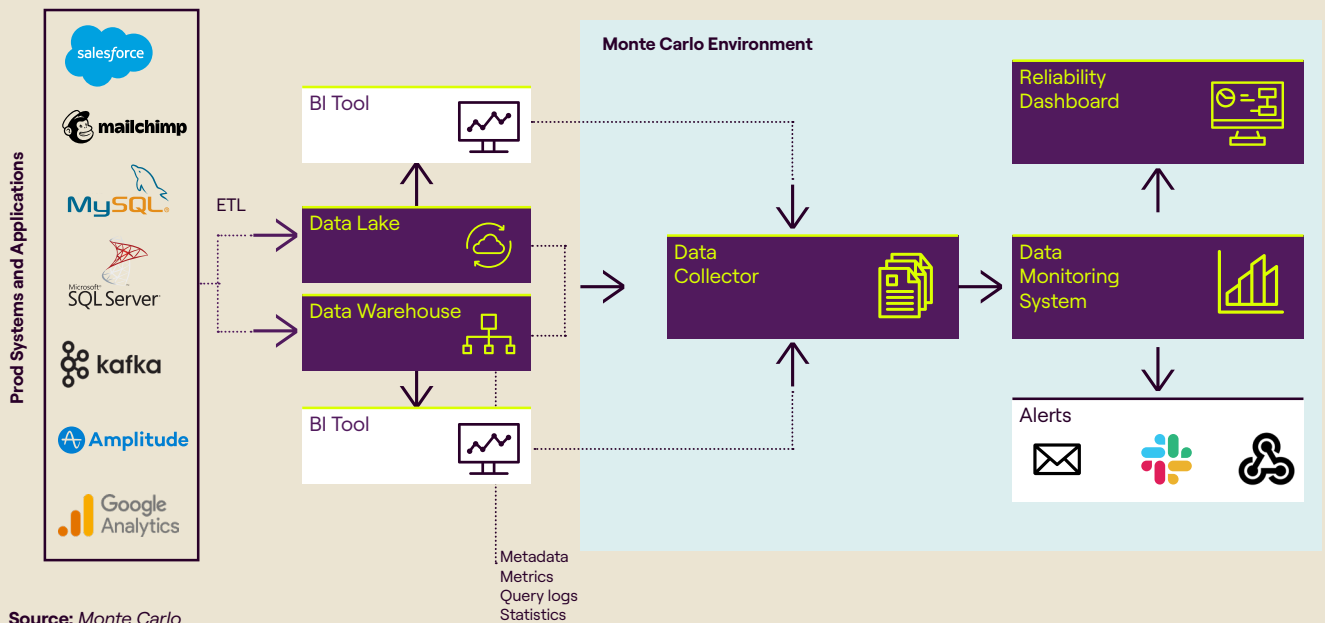


Setting up the perimeter

Integrations – setting up the perimeter

Data observability solutions typically collect data from data warehouses (e.g. Snowflake), though depending on the solution, they will also connect to data lakes (e.g. Databricks), traditional databases (e.g. MySQL), streaming data (e.g. Kafka), object storage (e.g. S3) and BI tools (e.g. Looker). The solutions also typically integrate with alerting channels (e.g. Slack), transformation tools (e.g. dbt) and data catalogues (e.g. Alation). An example of Monte Carlo's integration is illustrated below.

Monte Carlo's integration



Source: Monte Carlo

What is monitored – the usual suspects

The most common data quality issues are freshness, volume and schema changes. As such, these issues are what is most commonly checked for, and some solutions will check tables for these data issues out-of-the-box.

- **Freshness** – checks that data is refreshing at a rate that aligns with expectations.

Example of issue: Data that is supposed to update daily has not been updated for 48 hours.

- **Volume** – checks that the volume of data is in line with expectations.

Example of issue: The number of rows in a table is 500 instead of the 50,000 expected.

- **Schema** – checks that the data is organised as expected.

Example of issue: A column is missing from a table.

For more important tables, additional metric tests can be added, as shown below.

METRIC	DEFINITION	SUPPORTED FIELD TYPES
% Null	Percentage of rows that have a NULL value	All
% Unique	Data Observability	All
% Zero	Data Quality	Numeric
% Negative	Data Observability	Numeric
Min	Data Quality	Numeric

Source: Monte Carlo

Anomaly detection – sounding the alarm

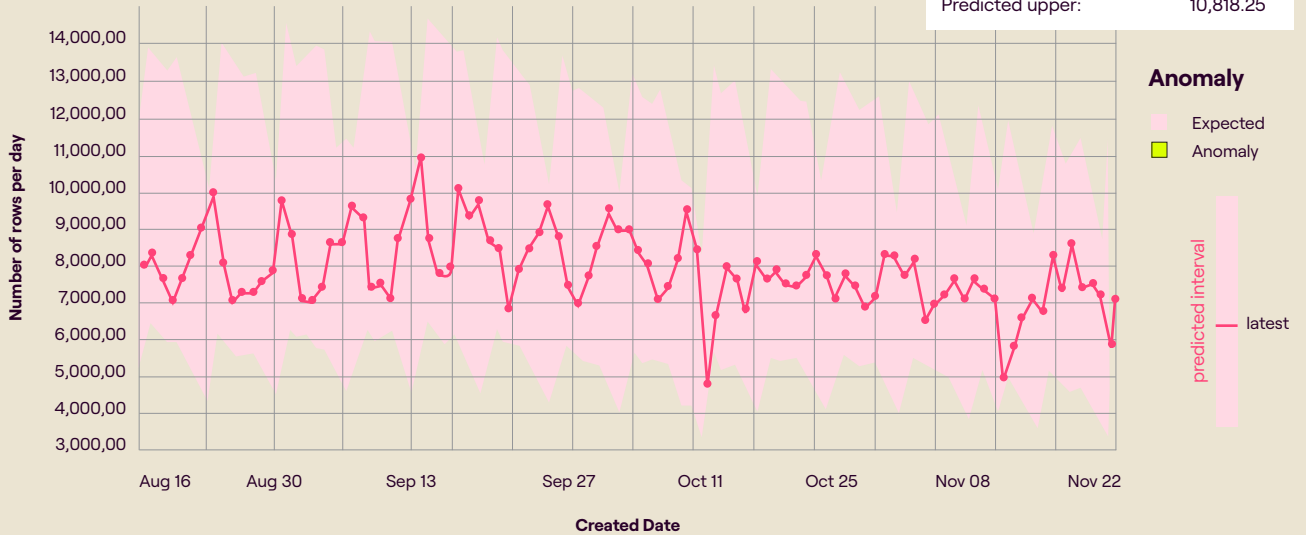
Once installed, data observability solutions typically spend a few weeks observing the company’s data to get a baseline for what is considered normal activity. Once the baseline is established, the tool sends an alert if an anomaly is detected.

One of the main challenges with anomaly detection is accuracy. Parameters must be tight enough to detect problematic outliers, but not too tight to overwhelm data engineers with false positives. To minimise these issues, an array of considerations need to be factored in. For instance, the system needs to understand to what extent a deviation is considered an anomaly. This might be different for different users and use cases. Seasonality may also need to be factored into the baseline of what is considered normal behaviour.

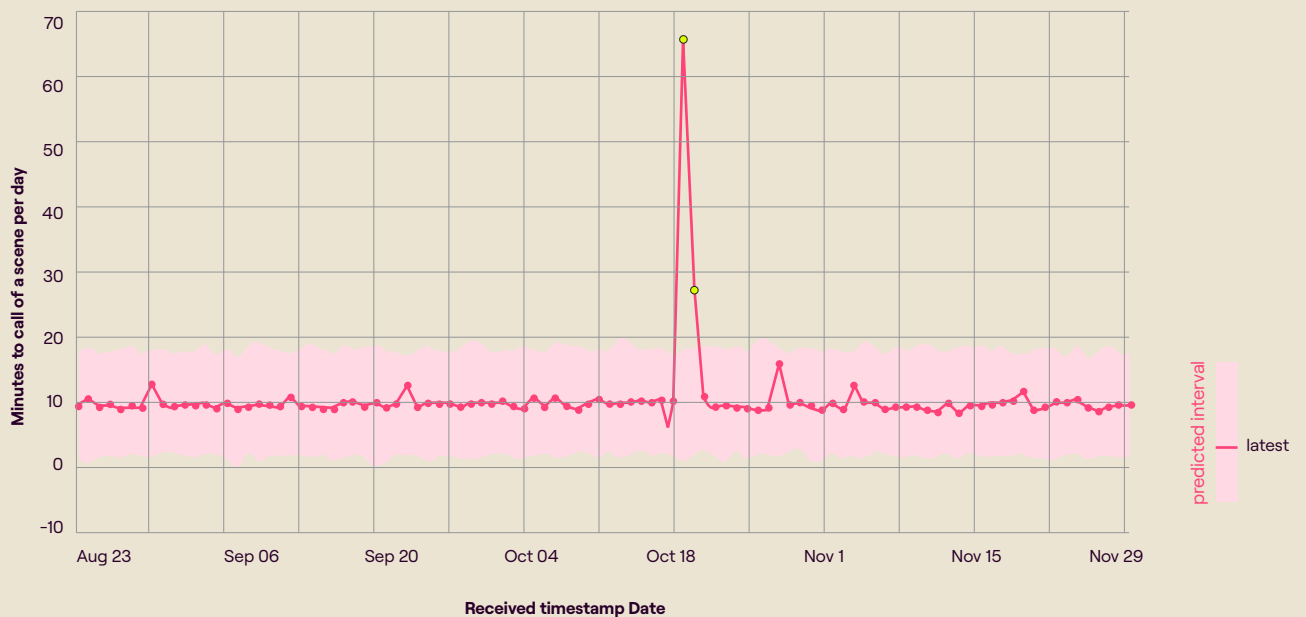
In the below figure, the shaded area represents the forecasted interval range of how many rows of data is expected per day. These thresholds are forecast based on historical data. As we can see, the actual data falls within the expected interval, and therefore no anomaly is raised. In the next figure, we see the average time taken for a fire department to reach the call of a scene. On October 19th, there was a significant unexpected spike. Therefore, an anomaly was flagged, and a message would have been sent to the relevant people, often via an alerting channel such as Slack or via email.

Anomalo’s row count monitor – no anomalies detected

Created: Monday November 23, 2020
Number of rows per day: 7,065.00
Predicted lower: 5,222.86
Predicted upper: 10,818.25



Anomalo’s time monitor – anomalies detected



Source: Anomalo

Root-cause-analysis – tracking down the culprit

Once an anomaly has been detected, the next step is to figure out where the issue originated. There are various root-cause-analysis (RCA) tools on the market to facilitate this, where the most common ones are data lineage and segment analysis.

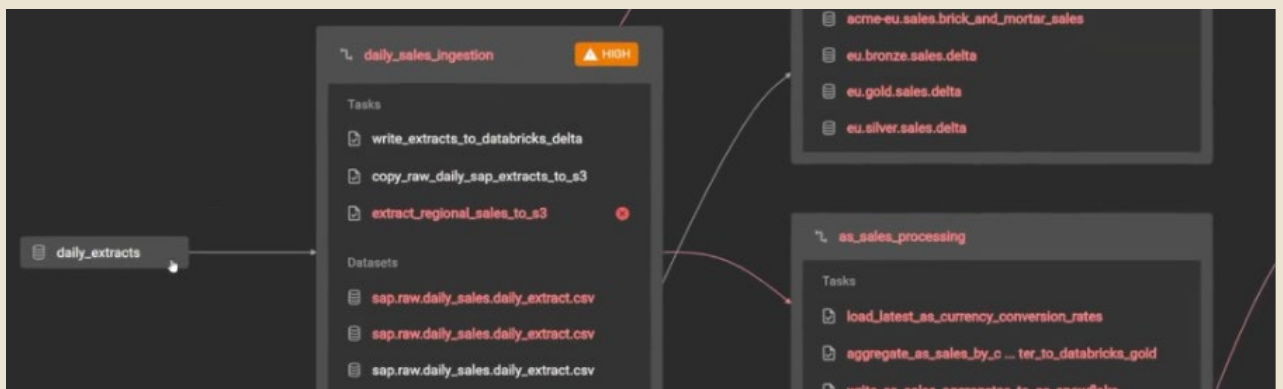
Data lineage

Data lineage provides an overview of how the data flows throughout the stack. Lineage is generally presented as a visual, providing a view of the data pipelines. This is often one of the favourite features for users, as they no longer feel like they are flying blind. When something breaks, data teams can now easily see where the issue originated upstream. It also allows them to see the potential impact downstream, informing them if something critical downstream broke as a result. Additionally, it enables

data engineers to know beforehand what might break downstream if they make a change upstream. For example, if a column is deleted, they will know which BI dashboards will be affected.

The figure below shows a snippet of lineage, where the red text indicates an issue. In the centre, we can see that the 'daily_sales_ingestion' process failed due to the 'extract_regional_sales_to_s3' task failing, impacting the below datasets. To the right we can see the subsequent impact this has on other processes, such as 'as_sales_processing' failing. This one issue had a very wide impact, which would have traditionally resulted in multiple people trying to solve many separate issues. However, with lineage it is possible to pinpoint the root cause of the problem quickly, allowing the business to solve the issue much faster, saving time and minimising business disruption.

Databand's data lineage visualisation tool – finding the root cause



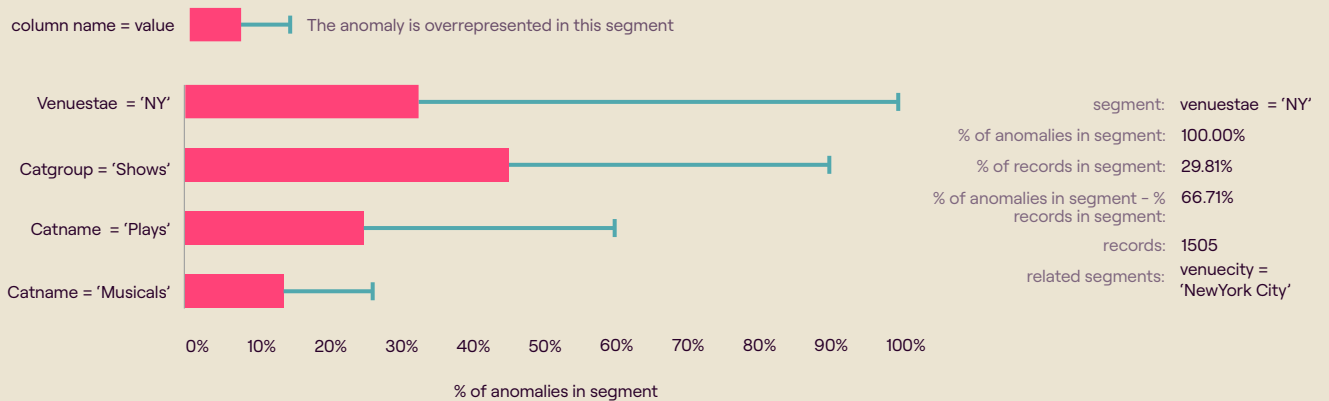
Source: Databand

Segment analysis

Segment analysis tools focus on finding issues in the underlying data. For instance, if there is an unusual amount of 0s, it can track down at a granular level to see which data segment is affected. This is illustrated below, where an anomaly has been detected due to 30% of column values for ticket numbers are 0s – an unusually high amount compared to what is expected. When the tool further

inspects the data, it finds that 100% of the anomalous rows are in the venue state New York. Therefore, the user can quickly see where the data issue occurred so that further action can be taken. We also note that segment analysis tools are not only useful in the RCA process after the anomaly has been detected, but that they can also enhance the ability of the solution to detect anomalies in the first place.

Anomalo's segment analysis tool – spotting the anomalous segments



Source: Anomalo

3. Differentiation and further development

In this section, we highlight the key areas where companies differentiate themselves, as well as areas they are focusing on for further development.

● Depth of tests

the number of metrics that can be tested

● Breadth of integrations

the extent to which the stack is covered

● Remediation capabilities

the granularity of RCA tools and the potential for resolution tools

● User-friendliness

the ease of use for both technical and non-technical data users

● Scalability

the ability to handle large data volumes while minimising costs

● Prevention capabilities

tools that help reduce the occurrence of bad data (shift left)

Data contracts – a new method of ensuring data quality across more stakeholders

Data Diff – a new technology to test how changes to code will impact data assets


Metrics – depth of tests

A differentiator for several vendors is the number of metrics that can be tested for. GX and Bigeye were commonly mentioned for the extensive depth of their tests, with Bigeye boasting over 60 prebuilt metrics. We have encountered customers who have had either GX or Bigeye sitting alongside Monte Carlo, as Monte Carlo is more limited in this aspect. Although the depth of tests is important for some customers, from our discussions, this usually was not the main deciding factor for why they bought a solution.

Integrations – shift left

Vendors are increasingly expanding their list of integrations to make their solutions suitable to a broader set of customers. This includes integrating alerting interfaces such as Slack and data catalogues such as Alation where data quality issues can be surfaced. Some of the more recent announcements have been dbt integrations, where for instance Anomalo now provides data quality for dbt metrics.²⁵

Most solutions are currently collecting data from data warehouses and their next-door neighbours in the stack. However, many vendors want to broaden this to get true end-to-end coverage. In particular, they want to go further upstream closer to the data source, often referred to as ‘shift left’, which is the idea of pushing data quality to earlier in the lifecycle. The benefit is that, if an issue occurs further upstream, it can be caught sooner before it creates problems downstream.



“We have the downstream BI integrations, so Looker, Tableau, Mode, Sigma, [and] Metabase. We support [...] the transactional databases, like MySQL and Postgres, and increasingly many OLAP databases like ClickHouse. Well, that’s where we stop. And honestly, that’s where everyone in our category stops today. I’m not very happy with that, because this is just the level one of monitoring. When you check out an observability tool in two years, or in five years, it’s going to be completely different it’s going to be [...] fully end-to-end. I think that is not only important, but really critical, because data is ultimately not produced from your data warehouse.”

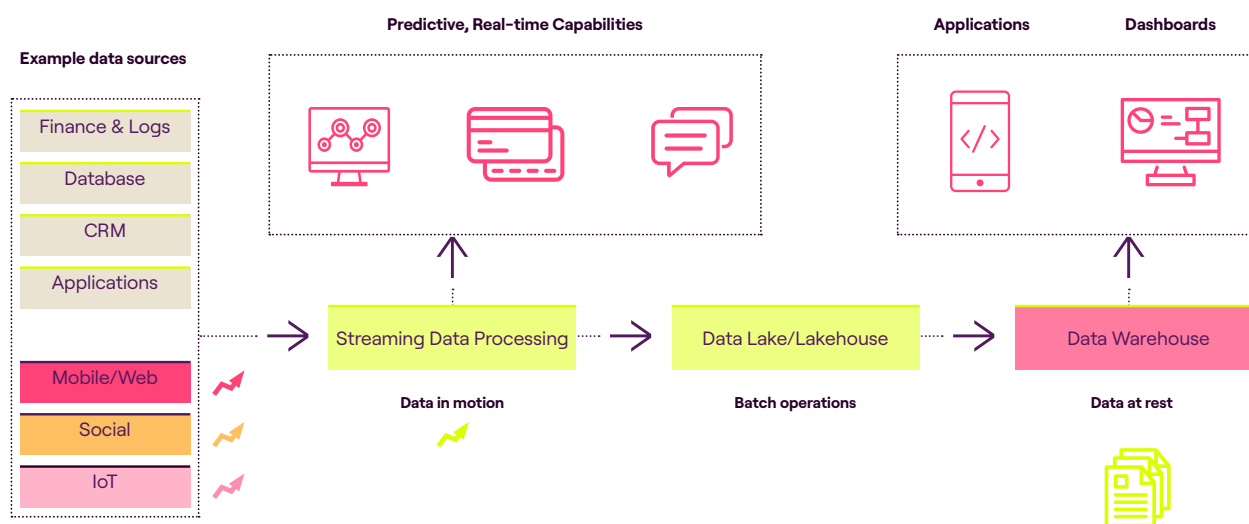
Metaplane CEO, Kevin Hu, The Data Stack Show, (June 2022)

²⁵ Anomalo blog, ‘Anomalo Partners With dbt Labs to Bring Data Quality to Key Business Metrics’

One of the areas many data observability companies aspire to extend coverage to is data streaming technologies such as Kafka, given their growing popularity. This will allow businesses to validate data before it goes further downstream to data lakes, data warehouses, or before it is used in real time operations such as fraud detection and ride-sharing apps. However, validating streaming data is considered more technically challenging than traditional batch data, though companies such as Acceldata and Validio already have this capability. They also have auto-remediation tools that automatically address data issues without the need for immediate manual intervention.²⁶ In the future, we would expect more data observability companies to support streaming data to remain competitive.

Data stack – vendors want to ‘shift left’ from the data warehouse

Example of data streaming



Source: Validio

Some founders also want to be able to test CRM data sources such as Salesforce. However, here there are technical challenges to overcome. For instance, Salesforce is API-based, making it harder to query for testing purposes. Potential developments here will also depend on integrations and innovations in other parts of the stack. For instance, Salesforce and Snowflake recently announced a real-time data sharing partnership.²⁷ One founder suggested that these types of developments might be a step towards solving the issue of checking data quality at the source, as this could potentially be done from the warehouse.

²⁶ Acceldata Auto Remediation

²⁷ Salesforce blog, 'Salesforce and Snowflake Expand Partnership with Real-Time Data Sharing'

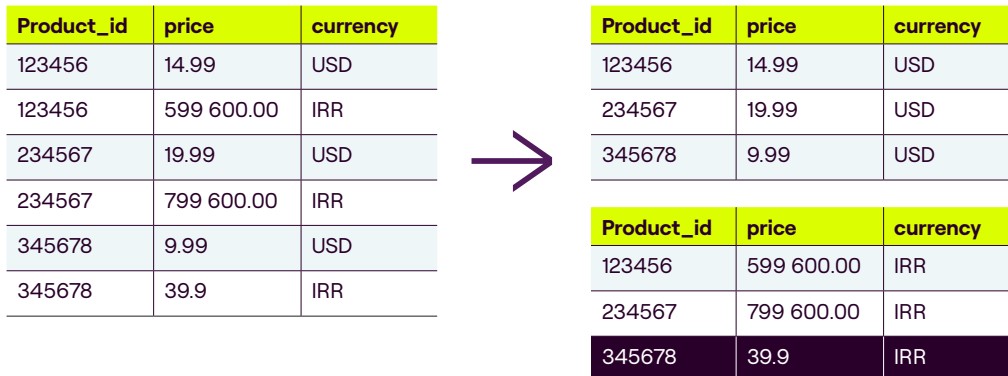
Root-cause-analysis – granularity is key

Some customers explained that the critical factor when deciding between different solutions was the RCA tool. It mainly came down to a choice between segment analysis or data lineage, as solutions tend to have one or the other (though other RCA tools do exist such as Lightup’s Time Correlation Analysis). Customers we spoke to have come down on either side of this decision, depending on what they needed most. The choice also relied on what other tools they had in their stack. For instance, some chose segment analysis as they already had data lineage from their data catalogue tool. Some customers also bought two data observability tools to get both capabilities.

Both RCA tools are not created equal however. Differentiators in lineage tools include how broad the coverage is across the stack and to what extent the tool can surface relevant and granular information (e.g. table vs column lineage) without overloading users with too much detail.²⁸ Given the strong demand for lineage tools, many vendors who did not previously have lineage plan to add this in the future, with Bigeye announcing its lineage tool in November 2022.²⁹

Segment analysis tools also have differentiation. For instance, Validio’s ‘Dynamic Segment Analysis’ runs all initial tests on sub-segments of the data instead of just running them on the aggregated data. This allows it to find data quality problems that otherwise would have been hidden. This is illustrated in the figure below, where the price of a product is denominated in both US dollars (USD) and Iranian Rial (IRR). As seen on the left-hand side, there is a large variation in price due to the order of magnitude differences in currency exchange rates. This large variation means it is difficult to spot an outlier in the aggregated data. However, if the data is segmented into separate currencies, it is much easier to catch. This is shown to the right, where one can clearly see that product_id 345678 for Iranian Rial is an outlier, where some 0’s are missing.

Validio’s Dynamic Segment Analysis – catching difficult to spot data outliers



The diagram illustrates the process of segmenting data to identify outliers. On the left, an aggregated table shows prices for product_id 123456, 234567, and 345678 in both USD and IRR. The IRR prices are significantly higher (599,600.00) than the USD prices (14.99, 19.99, 9.99). A large purple arrow points to the right, where the data is segmented into two separate tables. The top table shows the USD prices, and the bottom table shows the IRR prices. In the bottom table, the price for product_id 345678 (39.9) is highlighted in a dark purple box, indicating it is an outlier.

Product_id	price	currency
123456	14.99	USD
123456	599 600.00	IRR
234567	19.99	USD
234567	799 600.00	IRR
345678	9.99	USD
345678	39.9	IRR

Product_id	price	currency
123456	14.99	USD
234567	19.99	USD
345678	9.99	USD

Product_id	price	currency
123456	599 600.00	IRR
234567	799 600.00	IRR
345678	39.9	IRR

Source: Validio

Several companies want to go beyond RCA and provide resolution capabilities. This involves assisting in finding solutions to the data issue after the bad data has been identified. There are already some capabilities here, with Validio for example providing the option to filter out bad data points so that a data stream is ready to be used in downstream use cases. However, from our customer discussions, most vendor resolution capabilities were still fairly basic, with suggestions such as ‘make sure your table is connected’, reminiscent of the infamous Clippy.³⁰

Designing a resolution tool is difficult because, even though RCA tools can identify where the bad data occurred, it does not say why it occurred. Without knowing the ‘why’ it is challenging to suggest a solution. Additionally, problems can be very specific to the individual customer. We do imagine there is some potential for resolution tools given there is likely a large set of common problems that customers face with similar solutions. Therefore, if one has a large bank of solutions to various problems, a system could be built leveraging AI to map a customer’s problem to this bank, which could find the appropriate solution. That said, we have little insight into whether this will be useful in practice.


²⁸ Metaplane blog, ‘Column Level Lineage’

²⁹ Bigeye announces data lineage

³⁰ The infamous Clippy

User-friendliness – data quality for all

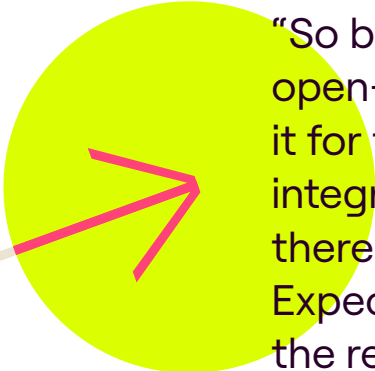
When speaking to customers, we noticed that user-friendliness was a key driver of their purchasing decision. User-friendliness includes how easy it is to deploy, learn and use. As such, several products have focused heavily on these areas. Monte Carlo, Bigeye and Metaplane all boast minimal deployment time, with Metaplane stating that deployment, on average, takes under 30 minutes, but can be done in under 10.³¹



“We looked at offerings from Databand, Alteryx, Acceldata, Amazon (Deequ), Soda, Informatica, and Great Expectations. The place where Monte Carlo blew the other tools out of the water was the turn-key nature of getting started. With no configuration (beyond the initial connection) we were able to detect some serious revenue impacting events in our systems. For most of the other products we looked at, we would have needed a lot of engineering-heavy time to get that same kind of coverage and peace of mind.”

Brandon Beidel, Director of Product Management, Red Ventures (January 2022)

For all its success, probably the largest weakness of GX is its user friendliness, as hardcoding often is needed. Speaking to users, the lack of user-friendliness became a significant hurdle when implementing it across their company, especially when training new people on how to use GX. As such, there has been a migration towards other, more user-friendly tools. Of the data practitioners who did not have a budget for data quality tools, several had switched from using GX to dbt's data tests, given it was much easier to adopt. GX founders are acutely aware of the issue and are working on making their solution more user-friendly.



“So back then Great Expectations was the main open-source software package and we mainly used it for the built in tests that it had, combined with the integration with our tech stack. However, these days there are more providers of these tests and Great Expectations is very hard to implement compared to the rest.”

Joost Boonzajer Flaes, Data Engineering Manager, IKEA (December 2022)

³¹Metaplane blog, busting myths on how long it takes to implement data observability

Given the popularity of dbt, with 16k organisations having adopted it, their testing tools too have seen widespread use.³² As a result, we have seen another evolution take place, where there are now open-source data observability projects such as Elementary and re_data building directly on top of dbt to provide additional functionality. Here users can install a data observability package in their dbt project. Developing a data observability platform on top of dbt has its challenges due to the constraints of the dbt environment, but the upside is that they do not need to integrate with multiple data warehouses. For Elementary, the goal was to make a seamless user experience for dbt users by allowing them to remain in their natural habitat.



“As a design concept, you can think of Apple, where they give you a similar experience between your iPhone and your Mac. If people work in dbt all day, we should give them in Elementary an experience that is as similar as possible, maintaining the same interfaces. And to use Elementary and to get value from the product, you don’t know need to know anything that you wouldn’t know from getting value from dbt. [...] We kept strictly to the skillset dbt users had. If you want, and if you know python, you can use Elementary as a platform to add additional capabilities to your dbt project.”

Maayan Salom, Co-Founder, Elementary Data, Open Source Startup Podcast (November 2022)

Vendors are not just making their solutions user-friendly for current users, but are also focusing on democratising access to data quality tools to non-technical users. This includes low/no-code interfaces and personalising data quality by surfacing the relevant issues to the relevant users. For instance, if a business user is interested in finance data for a report, they can be made aware of any issues with that specific data set. By democratising data quality tools, it relieves pressure on often overworked data engineer middlemen, removing bottlenecks and reducing friction. Several vendors, such as Monte Carlo, Soda and Anomalo, have come a long way in this regard, and other vendors are also making this a priority in current development.



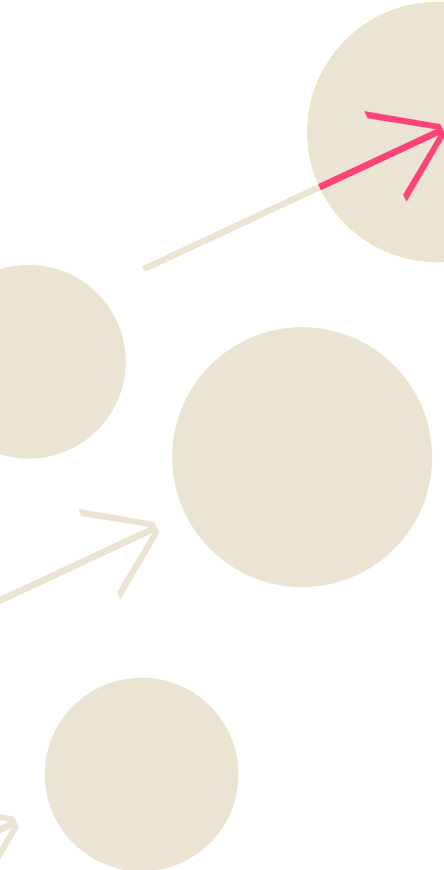
“Today, most of our Monte Carlo users are actually our business team, which has been a huge benefit to our data group. It has allowed the business to get more confidence that data is showing up on time, in the right shape, in the right volume without having to constantly run back and grab an engineer for 30 minutes to debug every instance of a perceived issue.”

Brandon Beidel, Director of Product Management, Red Ventures (January 2022)

³² 16,000 organizations using dbt

Scalability – going big without going broke

A solution's ability to scale is dependent on how much data it can test while simultaneously minimising costs. In addition to the fees paid to the vendor, there are other hidden costs when using a data observability tool. For instance, when an incident occurs and data teams are trying to find the root cause, there can be a significant amount of pushdown querying on the data warehouse, increasing the data warehouse bill. We imagine this cost varies enormously, but for one customer, this hidden cost was higher than the cost of the data observability tool itself. The cost also varies between solutions, and vendors such as Lightup and Validio note that this is an area where they differentiate themselves. Lightup highlighted that its pushdown costs are minimal as its architecture has been designed so that they do not need to do full table scans. A few vendors have also mentioned that they were able to displace competitors based on scalability for very large enterprises.



“We have proven success with large enterprises that want to run data quality checks on more than 1 billion rows of data per day and Lightup has been able to support such workloads without a hitch. Our platform succeeds at this by issuing push-down queries that are aware of the time-partitioning and time-indexing structure of the data model, which is why those checks scale so well without creating any noticeable load or cost impact on the underlying data warehouse or lakehouse.”

Manu Bansal, CEO and Co-Founder, Lightup (December 2022)

“Our system is built for both batch and streaming data, which gives us the capability to select whether the data needs to be ingested in order to be validated (and treated as a stream), or whether the existing warehouse capabilities should be used via push-down. In this way, we can optimize for whichever approach is the most cost effective for any data set and type of validation”

Patrik Liu Tran, CEO and Co-Founder, Validio (January 2023)

Prevention capabilities – pre-empting bad data

In more recent times, vendors have started to provide features that help prevent data issues from occurring in the first place. These are considered shift left tools as they push data quality to an earlier phase of the data lifecycle. Two features that have recently gotten a lot of attention are data contracts and Data Diff.

Data contracts – pushing data quality upstream

Although there are various definitions of a data contract, a non-controversial description is that these are formal agreements between data stakeholders, usually between data producers (e.g. software engineers) and data consumers (e.g. data scientists). They make stipulations on the data, defining what it is used for or what it should look like, such as schema and value ranges.

In some ways, data contracts are similar to data tests but have the benefit of notifying all the relevant stakeholders, including data consumers and data producers, when a contract is violated. This is contrary to what often happens

now, as the data producer will make changes to suit their own operational needs, unaware of the downstream impact. Therefore, armed with this new knowledge, the data producer may decide not to make the change if it is unimportant, or renegotiate the contract, and in general be more mindful of downstream consequences. Data contracts can also have the benefit of being a form of documentation, which gives context to data stakeholders and protects organisations from institutional knowledge being concentrated in a few individuals.

Data contracts are still at a very nascent stage however and come in many different forms, though we have already seen adoption by pioneers in the space such as GoCardless and Convoy.^{33, 34} Additionally, a few data quality companies such as GX and Soda are also providing data contract tooling. For instance, Soda launched its 'Agreements' in June 2022, as depicted below.³⁵ Agreements allow data contracts to be written in a highly user-friendly declarative language, intended for a broad set of data users.

Creating a data contract (Agreements) in Soda – user-friendly language for broad adoption

```
1. Select a Data Source 2. Write Checks 3. Identify Stakeholders 4. Set Notifications 5. Set a Scan

Define the SodaCL checks that define an expected state of "good quality" for data in this data source.

1 checks for retail_orders:
2   - schema:
3     warn:
4       when schema changes: any
5
6     fail:
7       when required column missing: [order_id, customer_id, product_id]
8       when wrong column type:
9         order_date: date
10        ship_date: date
11
12   - duplicate_count(order_id) = 0:
13     name: No duplicate orders
14
15   - missing_count(shipping_address) = 0:
16     name: Shipping address should always be provided
17
18   - invalid_count(payment_method) = 0:
19     valid values: ["Debit Card", "Paypal", "Cash"]

scan_definitions/retail_default_scan/sales_dashboard.yml
```

Source: Soda

Although we have generally seen a positive reception to data contracts, there has still been some debate on their overall usefulness.³⁶ Some argue that the bureaucracy they introduce outweighs the benefits. For instance, the upfront negotiations and re-negotiations between data stakeholders could potentially be a time sink. Additionally, data contracts could act as straightjackets for software engineers, impeding product development. We think the key to getting value from data contracts is that they

are used in the right context. For instance, a strict data contract could be warranted if specific data is driving a critical production pricing model. However, if the data is just being used for some exploratory work, then a data contract might be more hassle than it is worth. Although it is still early days, we see data contracts as a valuable tool to fight bad data, and believe we will see more widespread adoption and more vendors providing capabilities here.

³³ GoCardless blog, 'Improving Data Quality with Data Contracts'

³⁴ Convoy blog, 'The Rise of Data Contracts'

³⁵ Soda blog, 'Soda Cloud Previews Self-Serve Data Quality Checks & Agreements for Data Consumers'

³⁶ Mode blog, 'Fine, let's talk about data contracts'

Data Diff – validating the change

Data Diff is a tool provided by Datafold which helps detect potential data quality issues before they enter production.³⁷ It has seen an overwhelming amount of positive feedback from customers and experts that we spoke to, and even some reluctant praise from other founders. Data Diff’s primary ability is that it allows users to automatically test how changes to code will impact data assets in production. Therefore, it removes the worry that a change to the code will potentially break something downstream. Previously, this would have been done by writing manual tests, which was time consuming and slowed the business down.³⁸ Thus, Data Diff saves data teams significant amounts of time and helps the business move faster.

The below figure shows Data Diff’s pull request impact analysis report. In this example we see the impact of putting new code into production, where 2.9% of rows would be lost, 13,447 of 55,540 rows would change, and 4.0% of values would be different. This allows the data engineer to easily see if the potential code change will have the impact they expected.

Data Diffs pull request impact analysis report

DATA DIFF	DATADIFF-DEMO.DBT.DIM_BUSINESSES	
Environment	Production	Staging
Branch	master	dim_business_refactor
Rows	57,189	55,540 (-2.9%)
Columns	7	7 ✓
Exclusive PKs	1,648 (2.9%)	(0%) ✓
Differing rows	13,447/55,540	
Differing Values	4.0%	

Source: Datafold

Data Diff’s secondary capability is that it can validate data when it is replicated between data stores by checking that they match. For instance, if there is a discrepancy when replicating data from Postgres into Snowflake, Data Diff can give the exact rows of where the discrepancies occurred. This is a useful capability as replications are not always perfect and it allows users to know the extent of data deterioration. Depending on the severity of the issue, they can for instance retrigger the replication process or manually fix the issue. This is different from regular data quality testing which is about data meeting expectations based on historical data.

When speaking to customers who were planning to purchase Datafold, they were specifically buying it for this capability. One customer even had a deep data quality tool and lineage tool, but felt Datafold was an essential addition. Given Data Diff is Datafold’s flagship product, with not too many overlapping features with other data observability solutions, it does not see itself as a direct competitor. We could imagine however that other vendors are considering building this tool, though from our conversations with industry experts, it will not be easy to copy. We do note that Bigeye’s Deltas tool has some similar capabilities, though as far as we understand, it is less powerful as it provides less granularity.³⁹

³⁷ Datafold’s Data Diff

³⁸ Datafold blog, Data Diff use case

³⁹ Bigeye’s Deltas

4. Marketing dynamics

In this section we look at the forces shaping the market dynamics in the space. This includes:

● Pricing and pricing power

pricing models and how companies can maintain pricing power

Pricing and pricing power – differentiation is the key

In terms of pricing models, some platforms such as Monte Carlo still have a platform fee, but generally we have seen a shift towards a consumption-based approach, typically charging for the amount of data or tables monitored. As far as we understand, Monte Carlo is also experimenting with a table monitoring pricing model. We also see companies shifting away from the per seat pricing element as they want to democratise access and encourage anyone within the organisation to use the product. Given platforms are increasingly becoming low/no code, a business analyst can therefore check the data quality of a field they are interested in, without paying for an additional seat.

In terms of cost, there is a broad spectrum here. On the one hand, there are vendors that cater to smaller companies, such as Metaplane, charging as little as \$3.6k per annum for a starting package.⁴⁰ ⁴¹ On the other hand, the deep data quality vendors – that focus on scalability,

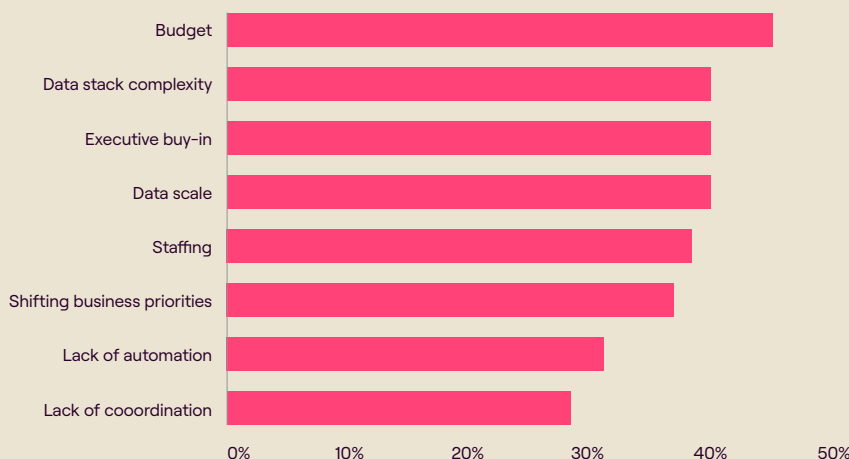
● Disruption and consolidation

- Why we think data observability tools are here to stay
- Thoughts on larger tech players entering the space
- The data catalogue vs data observability debate
- Potential consolidation within the category
- Final thoughts on the future of the category

segment analysis and shift left capabilities – tend to cater to more mature customers. These vendors usually charge between \$100k–300k per annum, with some larger enterprises paying just under \$1m. To put this in context, we asked a few customers how much they spent on data observability compared to their data warehouse, where the answer was c.20%–25%.

As indicated in the survey below, budget is still perceived as the biggest obstacle to improving data quality. This partly explains why open-source solutions such as GX and dbt test have been so popular. From our conversations with customers, we noticed that the willingness to spend on data observability depended on a few factors, such as the company’s digital maturity and how critical data quality was for its operations. Additionally, those who had benefitted from Covid lockdowns had large budgets to spend, with some even having bought two solutions. Those who were currently struggling and seeing budget cuts were often using open-source tools instead.

The biggest obstacle to improving data quality is budget



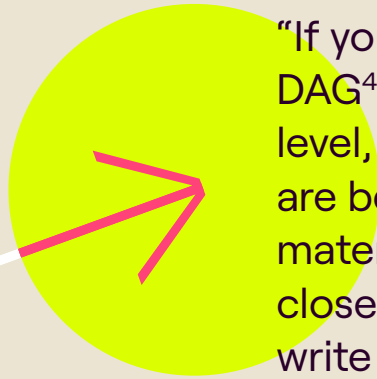
Survey question: Using the choices below, please rank in order the biggest obstacle to improving data quality at your organisation
Note: Top 3 Ranked Responses, n = 300 (data engineers)

Source: Survey: The State Of Data Quality, 2022

⁴⁰ Metaplane blog, busting myth that data observability needs to be expensive

⁴¹ Metaplane pricing

In terms of pricing resilience, we note that many data observability vendors position themselves as cost savers. The reasoning is that it would be much more costly to have people checking the data rather than the solutions. Additionally, some solutions can also help reduce costs directly by identifying wasteful activities in the data warehouse that can be eliminated. Therefore, many vendors feel they can easily justify their cost and that their products will be resilient in the face of an economic downturn.



“If you look at a huge pipeline or a huge dbt DAG⁴², if you would only analyse it at a rudimentary level, you could say that, ‘well, all of these tables are being used every day cause the entire DAG is materialised every day’. But if you then look even closer, you can see that all the queries are all just [...] write queries and there is nothing else happening. That can reveal huge opportunities for pruning data and thus saving costs. So for things like BigQuery, where it costs \$5 per terabyte scanned [...] you can show the hard cash you would save by pruning those and that is perfectly fine because there is one dashboard connected to this model but it was last viewed a month ago so it’s probably okay.”

Martin Sahlen, co-founder, Alvin, Data Engineering Podcast (October 2022)

Speaking to customers, they anecdotally confirmed that their data observability solutions were indeed cost savers. One customer using one of the more expensive solutions said it would be three times more costly to hire data engineers to do the same job. Additionally, dealing with data issues was the least fun part of the job, hence he would be worried about employee satisfaction and churn. He also noted that the solution caught things his engineers would have never caught. He did not either have an option to use open-source tools as they lacked key capabilities. Even though the price was the main pain point of the product, which had raised questions from his CEO, he could still comfortably justify the costs given the benefits, and was even able to expand his data observability budget.

From speaking with the vendors, only a few saw themselves competing on price. One founder said he was not worried about pricing competition at all, as he was confident that quality would win in the long run. Overall, our sense is that, for many customers, data quality is so critical that data observability platforms can maintain pricing power relatively well as long as they can maintain differentiation. If differentiation does begin to wane, pricing power too would likely diminish, especially given that switching costs appear fairly low because of how much focus there is on making the platforms easy to deploy, learn and use. Nevertheless, several founders argued that they could maintain their differentiation as their solutions were architected fundamentally differently than competitors, hence other vendors would have to start from scratch to catch up.

⁴² DAG definition

Disruption and consolidation – a landscape poised for change

Data observability tools are here to stay

In our discussions with practitioners and founders, no one could imagine a scenario where data observability tools would no longer be needed. We agree with this sentiment as the alternative is to design a stack where bad data cannot occur in the first place, which we think is unrealistic. Building both a perfect system and expecting humans not to make mistakes within that system has historically been extremely difficult to achieve. Additionally, change is constant, not only in terms of data, but also in terms of pipelines, software, employees and use cases, and change often leads to unintended and unforeseeable issues that need to be caught. Data observability is similar to cybersecurity in many ways, as businesses will need these technologies to always be on guard to protect the business against threats, be it bad actors or bad data.

Musings on disruption from the larger players

Although data observability capabilities are expected to remain, there are varying opinions on whether it remains as a separate category or becomes part of other larger platforms. A few suggested ETL/ELT tools (e.g. Fivetran) could infringe on the data observability space, given their unique access to the source data. Many believe that the next logical step is for the software (IT) observability companies (e.g. Datadog, New Relic) to enter the space. One expert argued that software observability companies would have an easier go-to-market path as they could provide data observability as a cross/up-sell, which would be met with little resistance given that the relevant part of IT budgets is already fairly sizeable. Some also believe dbt will be a big disrupter to the space, given that they are already providing testing and are taking market share away from players such as GX.

In terms of which specific data observability feature could see disruption, data lineage is a common topic. Several experts have argued that, given more and more platforms are providing lineage, this will soon be a commoditised feature. Here people have pointed to Snowflake, who is now also starting to provide their own lineage. Kevin Hu, the co-founder of Metaplane, argues that lineage from Snowflake would be much more powerful, as they have much better coverage and accuracy on their data warehouse compared to anyone else. He sees this as a potential challenge to current lineage tools, but also sees benefits as this unlocks capabilities for their own tools to become much more powerful. For instance, they can combine Snowflake's lineage of the warehouse within their own lineage tool that covers a broader area of the stack, providing an overall more robust lineage tool for Metaplane.

This argument rests on the assumption that Snowflake will not start to provide lineage beyond its warehouse. This is a general argument against more prominent tech players in the stack disrupting data observability, as the belief is that the likes of dbt and Fivetran will remain largely confined to their own area of the stack, and hence not disrupting data observability tools that cover the stack more broadly. While we see this as a reasonable argument in many cases, it does not hold true in all cases. For instance, Astronomer, the creator of Airflow, acquired the data lineage tool from Datakin in March 2022, which covers the stack more broadly.⁴³ Overall, it is hard to say whether data lineage becomes commoditised, as there is still differentiation in quality between different tools. But we think providing lineage itself will soon become a standard offering, so just providing a standard lineage tool will not differentiate the solution.

⁴³ Astronomer acquires Datakin

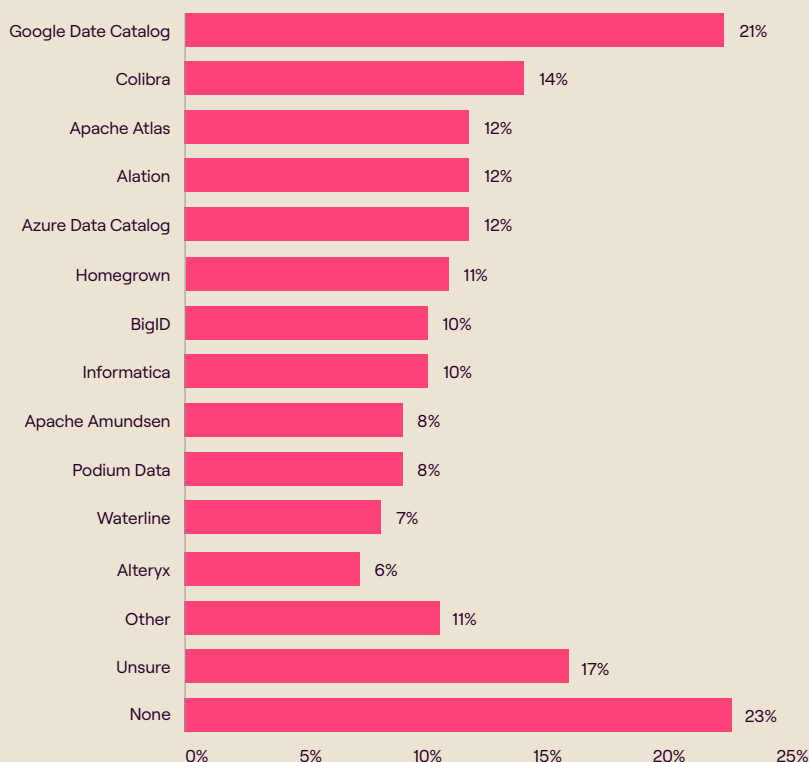
The data catalogue vs data observability debate

There has been much debate about the potential convergence of data observability and data catalogues. Data catalogues have been around for decades, providing an inventory of a company's data, enabling visibility and access. Similar to data quality tools, they have also gone through generational shifts and rebrands, adapting to the

evolution in data infrastructure and increasingly leveraging automation.⁴⁴ There is now an extensive list of vendors, which includes prominent names like Collibra, who raised \$250m in a Series G in November 2021 at a \$5.25bn valuation.⁴⁵ Alation is another well-known player who raised \$123m in a Series E in November 2022 with a valuation of over \$1.7bn.

Data catalogue market share

Question: What Data Catalogue and Data Discovery tools are in use in your organisation? (check all that apply)



Source: 2022 State of Data Engineering, n = 372

“A lot has been said about what data catalogues can do and not do. I think catalogues are an interesting concept, but they are just the beginning of the journey. But I personally believe that catalogues have literally failed the data and analytics industry for the last 40 years.”

Rohit Choudhary, Founder and CEO, Acceldata, Making Data Simple (April 2022)

For a few years now, there have been some tensions between data observability and data catalogues, with blogs such as 'Data Catalogs Are Dead' from Barr Moses.⁴⁶ One of the arguments has been that data catalogues have been failing data users and would be better served by data observability. However, as shown in the image below (taken

from a more amicable article from Monte Carlo), there are still many necessary capabilities that data catalogues have, which data observability does not, and therefore a replacement of data catalogues at present would be premature.⁴⁷

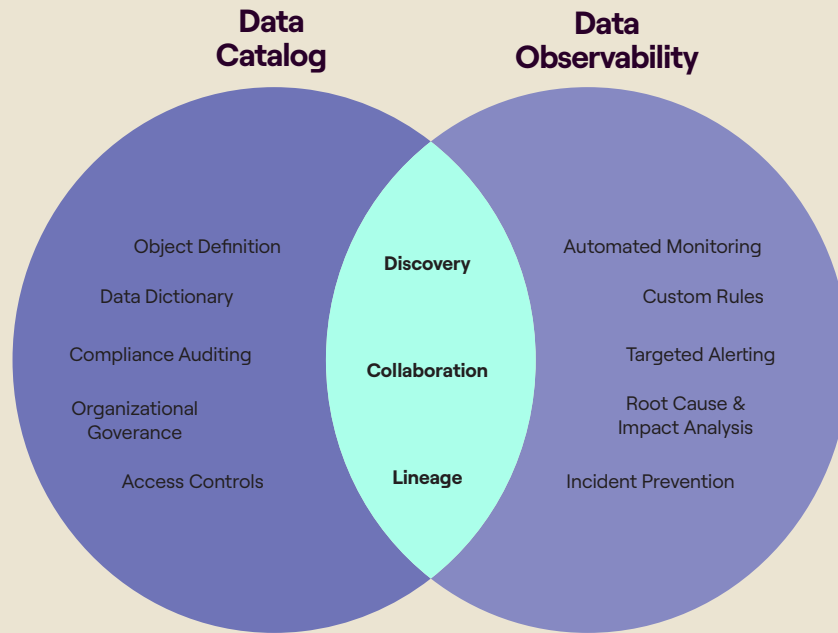
⁴⁴ Castor blog, 'Data Catalog Benchmark for Mid-Market Companies'

⁴⁵ Catalog of Catalogs

⁴⁶ Monte Carlo blog, 'Data Catalogs Are Dead; Long Live Data Discovery'

⁴⁷ Monte Carlo blog, 'Data Observability First, Data Catalog Second. Here's Why'

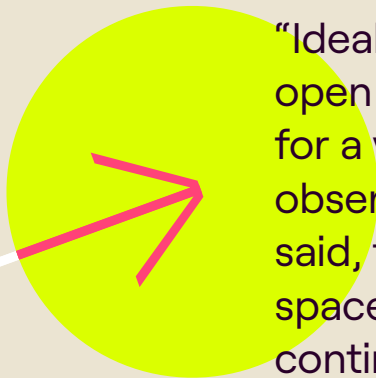
Data catalogue and data observability overlap



Source: Monte Carlo

However, data catalogues and observability share some key features, such as discovery, which refers to the ability to find data sets and understand their context. Additionally, they also share lineage. Because of these

overlapping capabilities and customers generally wanting fewer data platforms, many believe these two categories will eventually merge.



“Ideally, if you have all your metadata in one open platform, you should be able to leverage it for a variety of use cases (like data cataloguing, observability, lineage and more).[...] That being said, today, there’s a ton of innovation that these spaces need independently. My sense is that we’ll continue to see fragmentation in 2022 before we see consolidation in the years to come.”

Prukalpa Sankar, Co-Founder, Atlan, humans of data (January 2022)

Most industry insiders believe that a major convergence will not happen in the near term, as both categories are still in a rapid innovation phase. Consolidation, therefore at this stage would be counterproductive as it could put a damper on innovation. However, we could see some partial consolidation in the near term, driven partly by macro-economic pressures and crowding of the space. Already in February 2021, we saw Collibra acquiring OwIDQ, with data observability now being a key product offering

with 80 customers as of June 2022.⁴⁸ ⁴⁹(Collibra brands itself as ‘data intelligence’, which covers a broad set of data solutions, including data catalogue, data governance, data privacy and data observability). Alation, on the other hand, has so far chosen to partner with data observability companies instead, where data observability solution can integrate into the data catalogue, allowing data quality issues to be surfaced in Alation’s interface.⁵⁰

⁴⁸ Collibra acquires OwIDQ

⁴⁹ Collibra has 80 data observability customers as of June 2022

⁵⁰ Alation’s data observability partnerships

In terms of winners and losers from a potential consolidation, we think this will benefit the deep data quality vendors that have good scalability, segment analysis and shift left capabilities. This is because they are the most differentiated from data catalogues. Either the deep data quality tools could acquire a small next-generation catalogue themselves, or they can be acquired by one of the large catalogues on very favourable terms. An eventual consolidation between these categories could also put pressure on solutions such as Monte Carlo, as Monte Carlo, in some ways, is a hybrid between data quality and data catalogue, doing a bit of both. If data catalogues merge with deep data quality tools, then they could surpass Monte Carlo on both fronts. This assumes, however, that Monte Carlo does not catch up in data quality capabilities. And regardless, we would still see Monte Carlo as being relevant for mid-sized companies who do not need deep functionality in both domains.

Consolidation within the data observability category

Speaking with practitioners who have multiple data observability solutions, they were generally under some pressure from the finance department to switch to a single vendor. Although they were currently unwilling to compromise, they would be more than happy to switch to a single vendor if that vendor could provide all the needed functionality. This would not only keep the finance department happy but also bring simplicity. Therefore, consolidation between vendors could offer a significant competitive advantage (assuming this is technically feasible), as customers would not need to mix and match as they currently do, allowing these vendors to take market share. We could imagine companies focused more on lineage could merge with those focused on data quality, as well as smaller players with unique technologies being acquired by the larger players with greater reach. However, we do not expect a lot of consolidation in the near term, given most of the founders we spoke to were confident in their cash runway. But once cash piles begin to dwindle, some may be forced into an earlier sale or merger.

Final thoughts

There are many potential scenarios of how the category unfolds. Looking to the past, several of the original data quality tools were eventually swallowed up by larger players such as Informatica and IBM. A few did remain independent and are still in business today, though they are largely unheard of. We do believe data observability will remain independent in the near term, but we think consolidation in some form is inevitable once the industry has matured, given enterprise customers do, in our experience, generally want fewer products.

Several industry experts we spoke to also believed there will be a significant amount of reaggregation of data tools, in general, going forward, given there is too much fragmentation at the moment. Some vendors, such as re_data, are attempting to address this problem by acting as an aggregator and control centre for different data tools.⁵¹ This is similar to what we have seen in cybersecurity threat detection tools and are currently observing in cloud hosting orchestration. Control centres could simplify operations, but it still means having to deal with multiple vendors.

We do see a strong argument for data observability eventually merging with data catalogues, and do believe this will happen to a large extent. However, it is hard to imagine the ramifications if large players such as Datadog decide to make acquisitions in the space. Already we have seen IBM acquire Databand in July 2022, though founders seem less concerned about this move, possibly due to IBM's legacy status.⁵² Perhaps we end up in a scenario where various types of larger platforms offer data observability as a cross-sell.

Due to the various dynamics at play, including future innovations in other areas of the stack, we have little certainty of how the market will look in five years. Though we do believe data observability in some form will remain a key solution within the data stack for the foreseeable future, given we are increasingly moving to a data-driven world where data quality will be a vital factor for success.

⁵¹re_data, 'Your data trust dashboard'

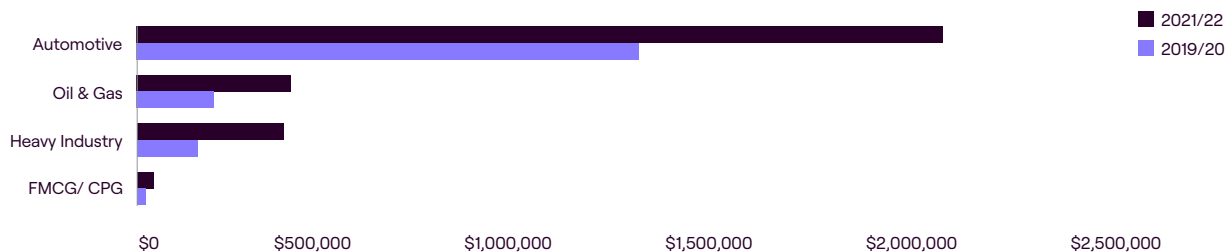
⁵²IBM acquires Databand

5. Appendix: Data observability in operational technology

The main body of this report has focused on data observability in relation to IT. But there is also another class of data quality solutions which focuses on operational technology (OT) for the massive asset-intensive industries. This includes industries such as chemical manufacturers, wind farms and oil refineries, which are increasingly using sensors to monitor their physical assets. These sensors have enabled asset-intensive industries to become increasingly data-driven, furthering automation and optimisation while allowing assets to communicate with one another. This evolution is known as Industry 4.0.⁵³ Because OT is increasingly relying on data, data quality too has also become increasingly important, as bad data can have a host of negative consequences.

One of these negative consequences is unplanned downtime. For instance, if a sensor becomes defective and is transmitting stale data, this can result in a warning not being flagged, such as temperatures getting too high, leading to malfunctions and causing a production halt. Unplanned downtime can be very costly to a business, not only due to the lost revenue from the lost output, but costs are also incurred in rectifying the issue, and penalties may need to be paid for not meeting contractual obligations. There are also indirect costs, such as loss of reputation with customers. A study by Senseye found that unplanned downtime costs Fortune Global 500 companies \$129m per facility on average and, in total, nearly \$1.5tn annually – 11% of total revenues.

Cost of one hour of unplanned downtime for a plant (US\$)



Source: *The True Cost of Downtime 2022 – Senseye Industry Insights (n= 56)*

In addition to unplanned downtime costs, poor data quality also has other negative consequences for a business, which include:

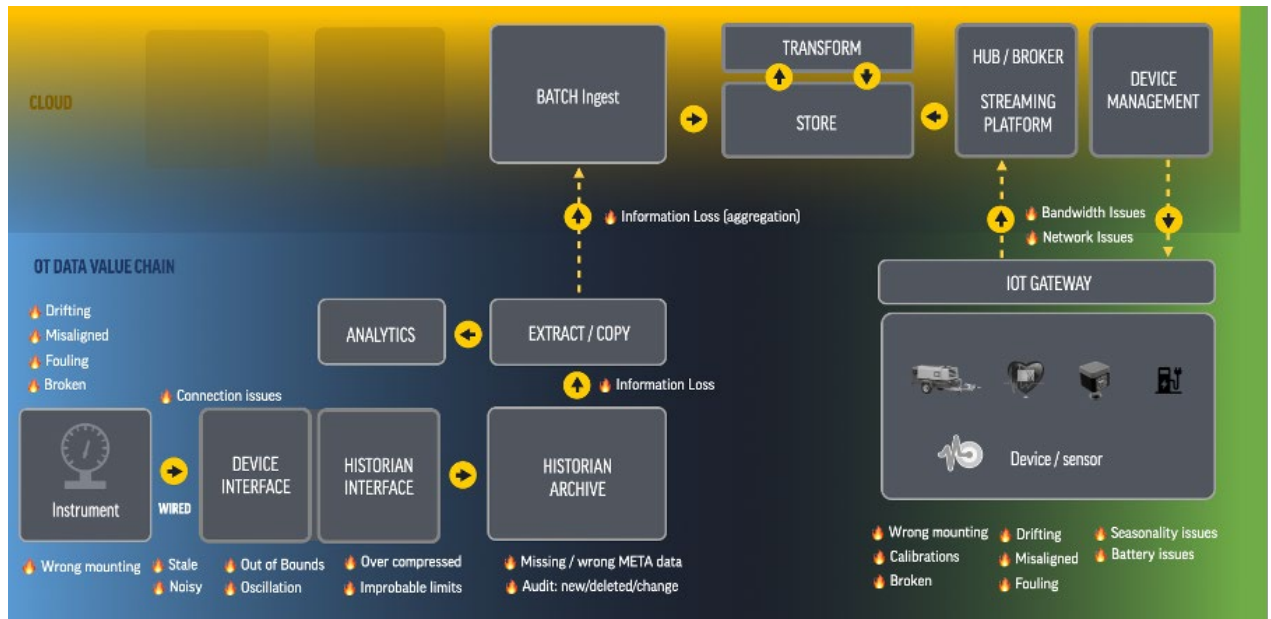
- Reducing the value from data analytics, which limits the ability to optimise the plant to make it run more efficiently.
- Reducing the reliability of safety systems.
- An increase in false positives, which results in extra cost due to unnecessary maintenance inspections.
- Operators losing trust in their systems, where for instance, too many false positives can create alert fatigue, causing operators to ignore alerts on real issues.
- Higher emission costs, as companies will often overreport emissions to be on the safe side to avoid regulatory fines.

Similar to data engineers in IT, OT engineers have often ended up building their own ad-hoc tools to address data quality issues. However, these tools tend to take a long time to build, often do not have good useability, and coverage tends to be limited. As shown in the figure below, OT typically comprises of complex systems with many points of failure, meaning it is difficult to build a complete end-to-end data quality tool in-house.

⁵³ *Industry 4.0 explainer*

IoT Value Chain – Where data goes bad in Operational Technology

Note: flame indicates where a potential data issue may occur



Source: Timeseer

To address data quality issues in OT, a few start-ups have emerged to provide OT data observability. A leader in this space is Aperio, which was founded in 2017, and a newer rival is Timeseer, which was founded in 2020. Additionally, larger OT data management platforms have also built their own data quality tools, such as Cognite and OSIsoft. Although data quality is a key focus for Cognite, OSIsoft's tools are fairly limited from our understanding, and have therefore partnered with Aperio, which can integrate with their platform. Other vendors, like MMC portfolio company Senseye (acquired in 2022 by Siemens), focus on anomaly detection directly from machine sensors and provide workflow capabilities around case management and predictive maintenance.

Aperio estimates that investing in their software and just reducing just a few outages a year provides an ROI of 20-40x.⁵⁴ Speaking with OT practitioners from different fields, they universally agreed that improving data quality would unlock significant benefits. Depending on the industry, potential benefits vary. For some, reducing outages was the priority. For others, accurate reporting was the key use case. Although all agreed that improving data quality was important, not everyone saw data observability solutions as a silver bullet.

Speaking to an oil and gas OT engineer, he said that in theory, for non-real-time use cases such as predictive maintenance and production optimisation, deploying a data quality solution could provide a lot of potential benefits. His worry, however, was that when data issues are discovered, such as a degrading sensor, there might be little incentive for the operators to fix the issue. The operators, who make decisions in real-time, often have very intimate knowledge of their systems and are present 24/7. They would, therefore, often see it as unimportant if a sensor is a bit off and see it as a waste of time and money to fix it. As a result, the underlying data quality issues would not get solved. So although he could see energy sector companies buying data quality solutions, getting value from them might be more challenging, which could result in product churn.

Another issue that can limit the adoption of data observability tools is the digital maturity of the plant. Typically, when designing and building industrial assets, the aim is often that they should last for decades. Therefore, many of the current assets in plants today were not designed for the modern age of Industry 4.0. As a result, many plants would need to be retrofitted with sensors and have other upgrades to become digitally mature enough to get to a stage where a data solution would be relevant. One expert estimated under half the industry would be digitally mature enough to get any benefit out of a data observability solution. This means that, although there is a large TAM, many plants are not yet ready.

⁵⁴ Aperio ROI 20x-40x

Overall, we still believe OT data quality solutions will be increasingly needed as asset-intensive industries become more and more data-driven. The larger uncertainty is the speed of adoption, which will be driven by how long it takes for both plants and culture to become digitally mature. Once these organisations make progress here, we'd expect data quality solutions to become commonplace. For instance, in the oil and gas example above, we would expect operators to more readily embrace data quality solutions over time, since they would be held accountable for changing failing sensors in order to improve overall business efficiency. What will drive an increase in digital maturity will not only be competition, but also regulation, which has been responsible for much of the digital initiatives in the OT space (e.g. reporting requirements). Some also point to a generational shift happening in asset-intensive industries, especially in oil and gas, which should bring a gradual increase in cultural acceptance of digital transformation.

The other question will be whether it is the point solutions or the larger data management platforms that will dominate the OT data observability space. This will, in part, depend on to what extent the larger platforms want to prioritise building sophisticated data quality tools, or if they are happy for point solutions to solve this problem instead and allow them to integrate into their platforms.

In terms of potential competition from the IT data observability vendors, we have not seen any evidence for this yet. Founders of the IT data observability companies we spoke to were generally unaware of the OT use case. We do not imagine either that it would be easy to expand into the OT vertical. Although there are many similarities conceptually, there are many practical differences: they operate in different stacks, with different data, different use cases and different customers. We expect in the medium term that the IT data observability players will be busy expanding their capabilities, improving integration and useability rather than expanding into the OT domain.

Overall, given the limited competition, the large TAM and the growing dependence on data, we think OT data observability is an exciting category with a lot of potential, where the main challenge will be timing.

