Title: Eight trends in data reliability for the next five years

### Introduction

The last five years have seen data reliability shift from a backwater infrastructure concern to top-of-mind for CEOs. In the coming years, data reliability will continue to evolve with emerging technologies and methodologies. In this blog post, we explore eight key trends that will shape the future of data reliability over the next five years.

### DataOps and the rise of data reliability as a core discipline

DataOps, the combination of data engineering and data quality practices, will gain even more prominence. As organizations scale their data processing capabilities, they will face growing challenges in managing and ensuring data reliability. This will lead to the rise of Data Reliability Engineering (DRE) as a core discipline within the DataOps framework. DRE will focus on integrating data reliability best practices, tools, and automation into the data pipeline to improve data quality and minimize human intervention. This shift will require organizations to invest in training and upskilling their workforce to develop DRE expertise.

### Machine learning for data quality assurance

Machine learning (ML) is already playing a significant role in various data processing tasks, and its impact on data reliability will only increase in the coming years. ML algorithms can help detect anomalies, identify patterns, and automatically correct inconsistencies. As ML models become more sophisticated, they will be better equipped to handle complex data types, such as images, videos, and unstructured text. This will enable organizations to improve data quality and reliability across various data sources. Furthermore, as the use of ML in data quality assurance becomes more widespread, we can expect the development of new ML techniques and tools tailored specifically for data reliability tasks.

### Data reliability becomes more cost-effective

Understanding the origin and history of data pipelines is crucial for ensuring data reliability. Data provenance and lineage tracking will become increasingly important as the amount of data produced by organizations increases exponentially. By recording the source, transformation, and movement of data throughout its lifecycle, organizations can gain better visibility into data quality issues and take corrective actions.

### The meaning of data reliability grows to encompass data privacy

Over the next five years, reliable data will mean data that is not only fresh, accurate, and complete, but privacy-respecting. Regulations like GDPR define data processors as entities that perform operations on data, on behalf of data controllers. For example, if you go to the New York Times website and it collects your email for marketing purposes, then stores that email address in Hubspot, the New York Times is the data controller and Hubspot is the data processor. Most data infrastructure tools, from Datadog to DBT to Bigeye, qualify as data processors.

Data processors have certain obligations. For example, in theory, if a user requests that the New York Times delete their email address, then the New York Times is obligated to ensure that that user’s email address is also deleted from Hubspot, Zoom, Datadog, AWS, and any other data processor it uses. Data reliability tools, which often sit on top of data warehouses, are well-positioned to implement further privacy checks on data - for example, that each piece of data collected was collected for a specific purpose and has user consent for that purpose, and does not have a DELETE request from a controller outstanding.

### Real-time techniques for data reliability

As businesses demand real-time insights for immediate decision-making, the need for data to be processed and analyzed at high speed (i.e., high data velocity) has never been greater. This has led to the rise of streaming pipelines, which continually ingest, process, and analyze data. However, as data velocity increases, so does the challenge of maintaining data quality. Real-time data streams leave little room for traditional, batch-based data quality checks. Therefore, data engineers must develop new strategies and adopt tools that can ensure data quality in real time. These could include real-time data validation, in-stream data cleansing, and machine-learning algorithms that can detect and correct anomalies on the fly.

### LLMs empower data consumers to be involved in fixing data quality issues

Large Language Models (LLMs), such as GPT-4, are poised to dramatically change how people interact with databases. Instead of manually querying databases or scrutinizing data for quality issues, users can leverage LLMs to communicate their needs in natural language. For example, a user could simply [ask the LLM for a specific data analysis or to check for data quality issues](https://www.bigeye.com/blog/can-chatgpt-identify-and-handle-data-quality-issues), and the LLM could perform the necessary database queries, data cleaning, or anomaly detection tasks. In essence, the LLM acts as an intelligent intermediary between the user and the database, simplifying the interaction and making data more accessible to non-experts. Similar to how the rise of DBT empowered data analysts and business analysts to perform data transformations, LLMs will further democratize data quality work. Rather than the finance team needing to wait for data engineers to fix a data quality issue before they are able to update their report, they will be able to do it themselves.

### Data contracts become more popular

[Data contracts](https://www.bigeye.com/blog/a-gentle-introduction-to-data-contracts), proposed by [Chad Sanderson](https://dataproducts.substack.com/p/the-rise-of-data-contracts) of Convoy and [Andrew Jones](https://medium.com/gocardless-tech/improving-data-quality-with-data-contracts-238041e35698) of GoCardless, are agreements specifying the format, content, and quality of data when it is produced. It essentially treats data like an API that must be code reviewed and versioned. Rather than all data getting dumped into something like Kafka, and then transformed/cleaned up in data warehouses, service owners choose a subset of data to expose in a structured manner.

Data contracts shift some of the responsibility for data quality towards the producers of data, i.e. application teams. They help set expectations and standards for data quality, allowing receiving systems to trust and use the incoming data. In the future, they will likely be more widely utilized.

### Out-of-the-box custom data monitoring deployments

The current generation of data monitoring tools [typically involves importing and processing all the tables/schemas in a data warehouse](https://www.bigeye.com/blog/data-observability-at-scale-monitoring-a-50000-table-data-lake-with-bigeye). The best of them, like Bigeye, then automatically generate metrics and thresholds that engineering teams can go through and enable. This is already an improvement from previous generations of observability products, which would have required teams to manually configure each metric. However, for large organizations, which might have hundreds of tables, it’s still an extremely tedious process.

In the future, we will see more turnkey solutions like Bigeye’s [BigConfig](https://www.bigeye.com/product/bigconfig), which allow you to deploy all your metrics from a single, templated config file. The templates are based on common data sources like Stripe and Hubspot and can be further customized.