Data Monitoring

Linus Rundberg Streuli

Table of contents

1 What is Data Monitoring?

- Data monitoring, in relation to data governance, is a process where data is routinely checked for issues regarding quality, integrity, security and usability.

2 Four core areas

- **Alerting**: Alerts warns of dangers, threats or problems. Monitoring systems can be set to raise alerts if data quality drops below a certain threshold or if the database is seeing suspicious behavior, for example.
- **Accounting**: Accounts are detailed descriptions of events, used for in-depth analysis of applications, infrastructure and policies.
- **Auditing**: Audits are systematic reviews of various processes in the organization to ensure that they are performing as designed.
- **Compliance**: Monitoring systems are a critical part of meeting policies, laws, standards and regulations.

3 What to Monitor

3.1 Monitoring Data Quality

- When monitoring data quality, common attributes are:
  - **Completeness**: Is there data missing? If the column is tagged as *not-null*, this should raise an alert.
  - **Accuracy**: Do the values fall inside an acceptable range?
  - **Duplication**: Is the data repeated?
  - **Conformity**: Is the data conforming to format standards?

3.1.1 Process and tools

- When setting up a data quality monitoring system, some things to consider are:
  - **Establishing a baseline**: Assess the current quality of the data, in relation to the business goals that guide the data governance program.
  - **Quality signals/Metrics**: Clearly define the metrics that are to be used for evaluating trends in data quality. These metrics are also used to set thresholds for alerts.
- One popular tool for managing and monitoring data quality is the framework Great Expectations (GX), which provides an open source “core” product as well as a commercial SaaS product, GX Cloud.
  - More info: https://greatexpectations.io/.

3.2 Monitoring Data Lineage

- Monitoring data lineage is crucial for ensuring data trustworthiness.
- Data life cycles are often complex, and well monitored data lineage can help with analyzing why dashboards are showing unexpected results, or how sensitive data is moving through the organization.
- Key attributes to track include:
  - Data transformations,
  - Technical metadata,
  - Data quality test results,
  - Reference data values, and
  - the Actor behind data transformations, such as an ETL pipeline.
- One common tool used for lineage tracking (among other things) is the open source metadata platform DataHub.
  - More info: https://datahubproject.io/

3.3 Monitoring Compliance

- Monitoring compliance means staying updated with the latest regulations and standards, as well as keeping track of the current state of compliance within the organization.
- A functioning monitoring system can also provide proof of compliance in case of an external audit.
- Implementing compliance monitoring includes:
  - understanding which regulations apply to the organization,
  - assessing which data is at risk and prioritizing, and
  - creating a monitoring plan, including roles and responsibilities.
- Is is important that employees are informed about regulations and the importance of compliance. Offer training and encourage questions.
- Create a culture of compliance and a strong relationship between the compliance person/team and the legal department.

3.4 Monitoring Data Security

- Monitoring data security is the process of analyzing user and program activity as well as network traffic, and search for suspicious patterns och unauthorized accesses or updates.
- Key areas include:
  - Security alerts and incidents,
  - Network, application and endpoint events,
  - Server logs,
  - Identity access management, and
  - Data loss.

4 Monitoring Data Governance Performance

- It is not only data and metadata itself that can - or should - be monitored.
- Monitoring the data governance program itself is an important part of evaluating efforts and demonstrating results.
- Metrics to monitor include:
  - Number of lines of business, and project teams, that have committed resources and sponsorship to data governance,
  - Status of issues that are raised, how they are handled, and the resulting impact,
  - Level of engagement and influence across the organization,
  - Value-added interactions such as training and project support, and their impact, and
  - Business value ROI from data governance investments.

4.1 Performance Management Framework

- Performance monitoring must be continuous and ongoing.
- Performance management frameworks commonly focus on these four areas:
  - **Alignment with existing governance frameworks**: Ensure that the data governance program is performing as outlined in the data governance framework.
  - **Developing performance indicators**: KPI:s for the data governance program should be well defined, relevant and informative.
  - **Reporting progress and performance**: Documenting program performance and providing reports enables informing people of the value of the program.
  - **Taking action based on performance results**: Performance results should inform decisionmaking - that’s the whole point of the program!

5 Monitoring Systems

- A monitoring system is the set of tools, technologies and processes used to capture and analyze operations on data.
- There are many data monitoring tools on the market. Some features to consider when developing, purchasing, or outsourcing a data monitoring system are:
  - **Real time analysis**: Continous monitoring and analysis is important for quick responses to changes,
  - **System alerts**: Configurable system alerts are necessary for ensuring that issues are addressed as quickly as possible,
  - **Notifications**: On top of alerts, a monitoring system needs a robust notification system with options for different channels of communication as well as enabling responses and follow-up discussion.
  - **Reporting/Analysis**: A monitoring system is a big collection of data, and needs to offer reporting abilities for presenting trends and patterns, and forecasting future events. This also include graphic visualizations such as dashboards.
  - **Customization**: Monitoring systems need to be fully customizable to meet the needs of the organization.

5.1 Independent Monitoring Systems

- It is important that monitoring systems are independent from production services, and they should not affect the performance of the systems they monitor.
- In case of a production outage or a failure event, the monitoring system need to continue monitoring - that is its whole purpose.

5.2 Passive and active systems

- Monitoring systems are either:
  - **Passive**, where tools observe data logs and process outputs under normal circumstances, or
  - **Active**, where the system is more integrated in the actual system.

6 Sources

Eryurek, et. al: Data Governance: The Definitive Guide (Chapter 8).