Title: Data Pipeline Monitoring vs. Data Quality Monitoring: What's the Difference and Why It Matters

### Introduction

While data monitoring often conflates the health of the data pipelines with the health of the data itself, in fact, they should be considered two separate disciplines: data pipeline monitoring and data quality monitoring. In this post, we'll delve into the key differences between the two and why it is essential to have both in place for a seamless and accurate data-driven process.

### Data Pipeline Monitoring: Ensuring Smooth Data Flow

Data pipeline monitoring focuses on the jobs and tables that move the data, e.g. Snowflake and Airflow. The main aspects of data pipeline monitoring include freshness (when each table was last updated), volume (how many rows are being moved), and job run durations. Data pipeline monitoring is typically the responsibility of data engineering or data platform teams.

By monitoring the data pipeline, you can ensure that your ETL (Extract, Transform, Load) processes are running smoothly and that the data is flowing seamlessly between different stages of the pipeline. This helps avoid bottlenecks and ensures that your data is up-to-date and ready for analysis.

### Data Quality Monitoring: Assessing the Contents of the Data

Data quality monitoring, on the other hand, focuses on the contents of the data. This includes aspects such as freshness (how old the values are), completeness (rate of nulls, blanks, etc.), duplication, and format compliance. Data quality monitoring is often the responsibility of data science and analytics teams, who need to ensure that the data they are working with is accurate and reliable.

By implementing data quality monitoring, you can identify issues such as null values, duplicates, and outliers that may affect the accuracy of your data-driven insights. This helps ensure that your ML models and analytics are based on high-quality data, ultimately leading to better decision-making.

### The Importance of Both Data Pipeline and Data Quality Monitoring

While data pipeline monitoring and data quality monitoring can be done with two separate systems, to truly understand the behavior of your pipeline, it is vital to correlate information from both. For instance, if you notice that a table has been refreshed later than usual with a larger number of rows, and you also find a significant number of duplicated IDs, this could indicate an issue with an ETL job. In this case, combining data pipeline monitoring (freshness and volume) with data quality monitoring (duplicates) can help you identify and resolve the problem.

### Prioritizing Data Pipeline Monitoring Before Data Quality Monitoring

It is generally recommended to prioritize data pipeline monitoring before data quality monitoring. This is because, if the data isn't flowing smoothly through the pipeline, there's no point in worrying about data quality. Once the data engineering team has ensured the smooth operation of the data pipeline, they can hand over the responsibility of data quality monitoring to the data science and analytics

teams. This division of labor allows each team to focus on their area of expertise and ensures that both aspects of data management are adequately addressed.

### The Role of Analytics Engineers in Pipeline and Quality Monitoring

With the rise of tools like DBT, the role of analytics engineer has evolved as a mix of data analyst and data engineer. Analytics engineers both have an understanding of how the data is consumed in dashboards and statistical models and write SQL to perform data transformations. They can serve as a valuable bridge in the sort of correlation work mentioned above.

### In Practice: The Intersection of Data Pipeline and Data Quality Monitoring

In reality, the division between data pipeline monitoring and data quality monitoring is not always clear-cut. However, having a strong understanding of the two concepts and their respective responsibilities can help organizations make informed decisions about which aspects of their data management processes need attention.

### Data pipeline monitoring and data quality monitoring in Bigeye

#### Metadata metrics

[Metadata metrics](https://www.bigeye.com/blog/introducing-metadata-metrics-instant-data-observability-for-your-entire-data-warehouse) are freshness and volume metrics that are enabled all data sources as soon as they are connected to Bigeye. Since they are extracted from your data warehouse query logs, they are cheap to compute. Since they mostly track the success or failure of pipeline jobs, Metadata metrics roughly correspond to data pipeline monitoring

#### Column-level metrics

Column-level metrics are metrics like averages, sums, and percentage nulls that can be enabled at a column level. Since they monitor the actual contents of the data, checking for nulls, duplicates, and format compliance, column-level metrics correspond to data quality monitoring.

### Conclusion

In summary, by investing in both data pipeline monitoring and data quality monitoring, organizations can maximize the value of their data and drive more informed decision-making across the enterprise. With Bigeye, you get a solution that does both.