**Data Observability**

Data Observability is turning point for all your data operations going forward, Data Quality frameworks and Data Governance strategies that are nice to one have ideology and now actionable with advances in this category.

It is crucial to comprehend the wealth and health of the data in your system, and it can manage data health across various IT technologies and across the data lifetime.

Nearly 20% of a Data Engineers time spent on investigating and debugging Data Quality problems.

Data powers Digital Services, products decisions your company depends on.

What happens if the data is wrong?

Does the right people use the right data for their right applications?

Bad data waste time, resources, trust in data and results in lost of missed revenue.

AS part of the Observability there are 3 main things that we need to focus on

Diagram, schematic

Description automatically generated with medium confidence

Data observability an end-to-end approach to enable teams to deliver more reliable and trustworthy data.

* Identify, Trouble shoot and resolve data issues
* Knowing there is a problem before users identify
* Who to talk for problems, upstream and downstream effects of changes

When it comes to Pipelines

* Is pipeline executing?
* Pipeline is not moving correctly
* How fast the data is moving
* Data Quality with in the pipeline – no record row, a value was expected

There are 5 pillars to accomplish this

* Freshness
* Distribution
* Volume
* Schema
* Lineage

Diagram

Description automatically generated with low confidence

Freshness

Freshness measures the frequency of updates to your data tables as well as how recent they are. Freshness is especially important when making decisions because outdated information is essentially synonymous with time and money wasted.

Distribution

The distribution of your data's potential values tells you whether your data falls inside a range that is acceptable. Data distribution enables you to determine whether your tables can be trusted based on the data you have.

Volume

Your data tables' size provides insight into their completeness and the state of your data sources. You should be alert if the number of rows decreases from 50 million to 3 million.

Schema

Changes in the way your data is organised, or schema, are typically indicators of broken data. It's essential to keep track of who and when adjusts these tables if you want to know how well your data environment is functioning.

Lineage

Where did it go wrong is always the first question asked when data is incorrect. Data lineage reveals whose teams are producing the data and who is accessing it, as well as whose upstream sources and downstream users were impacted. As a single source of truth for all users, good lineage also collects data-related data (also called as metadata) that relates to governance, business, and technical rules for data tables.

Observability tools will take a metadata approach, where the data is coming from, where it lands and detects the issues at this place Data Governance plays a key role.

Icon

Description automatically generated Monitoring: A Dashboard that provide an operational view of the pipeline system.

Data monitoring involves checking the accuracy and completeness of the information entered into the database. To make sure the entered data will achieve the objectives of the registry or clinical trial, the data can also be evaluated.

Icon

Description automatically generatedAlerting: A monitoring system's ability to detect and inform the operators of significant events that indicate a serious change in state is known as alerting.

Shape

Description automatically generated with low confidenceTracking: set to interact specific events

A trace depicts how a request or action travels across application from beginning to end.

Graphical user interface, text, icon

Description automatically generatedComparison: A comparative observational study's goal is to ascertain whether measurements between tables from various schemas differ.

Shape

Description automatically generated with low confidenceAnalysis: Automate detection issues that adopts your data and provide a brief overview of Activity monitor, DB locks, throughputs, and Resource utilization across the database.

![Shape

Description automatically generated with low confidence]()Logging: Logs are a great source of visibility if you want to know when a problem happened or which actions or patterns are associated with it. Because they offer a thorough record of all events and mistakes that occur during the lifecycle of software resources, logs are a pillar of observability.

Shape

Description automatically generated with low confidenceSLAs Tracking: Ability to measure Data Quality in Data pipeline against pre – defined standards. Service-level agreements (SLAs) are a technique that many businesses employ to specify and gauge the quality of service that a particular vendor, item, or internal team will provide—as well as possible remedies if they don't.

The most important part is monitoring which requires a lot of efforts, below are the couple of activities that will be taken care as part of monitoring

If Data is at REST,

* Is the data is arriving on time?
* Is data being updated frequently as needed as to be…
* Is this expected Volume of data, is it complete?
* Is it available to use?
* Are there any schema changes?

If the data is in Motion

* Monitor the pipeline performance like state, Duration, Delay, Retries
* What operations are transforming the data set before reaches the warehouse
* Are there any unexpected changes in execution of metadata

Success Parameters: Under what condition the run was successful.

Column profile:

* What is the expected Range for column data
* What is the expected schema
* How Unique is the column
* Tracking Trends with Means, min, max, Skewness, Kurtosis and Z-Score

Rows:

* Values in the row level are in expected format
* Are the values length expected to be
* Is there enough information useful to the end user
* Check for any Business rules

A picture containing application

Description automatically generated

With these above rules Data Observability aims to bring Stability and Reliability to Data products