ACTIAN™
a division of HCLSoftware

CHECKLIST

The Definitive Data Observability Evaluation Checklist

What to look for when evaluating
data monitoring and observability solutions

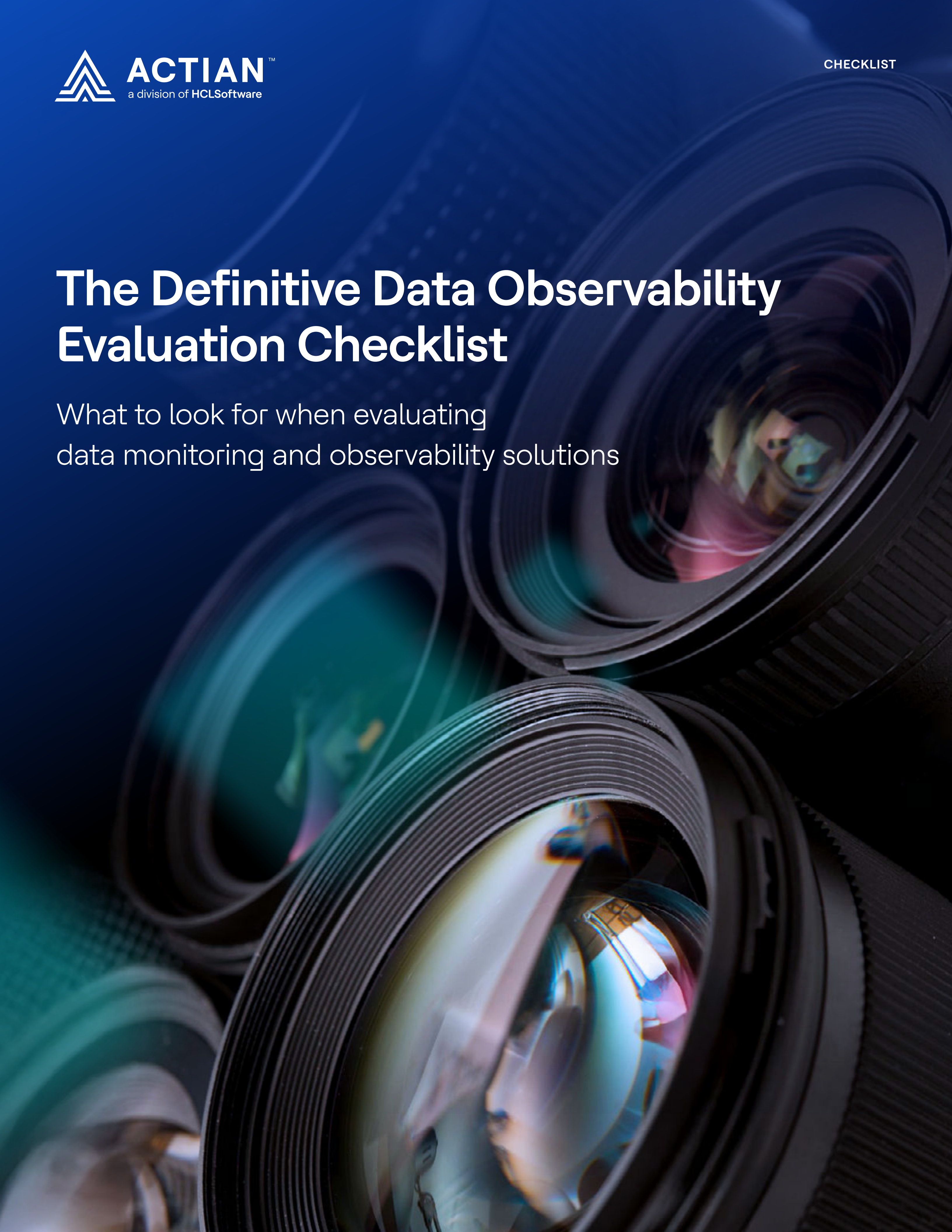


Table of Contents

Every data initiative today hinges on the reliability of the data that flows through the underlying data pipeline. Hence, monitoring key data metrics is essential for your data pipelines.

Data engineering teams across all industries and companies invest heavily in building, maintaining, and optimizing data infrastructure to building data products and ML models — yet most companies are not getting expected Returns On Investments (ROI) on these initiatives due to unreliability for underlying data.

Data observability is rapidly becoming a critical component of any data pipeline. The correct data observability platform enables data teams to detect issues as they emerge proactively. It provides capabilities for proactive introspection to diagnose of the root cause of problems before they significantly impact a business or its customers.

Armed with the right solution, data teams can now visualize and measure data health across the pipeline, where and why problems are emerging — investigating issues without added burdens like writing and executing redundant queries. Teams can even implement automation around remediations, self-healing extra using the outcomes of data observability.

This checklist covers the essential elements to consider when evaluating a data observability solution.

Integrations

One of the first and most important considerations during evaluating a data observability tool is its adaptability and supportability into the client's current and future architecture and data pipelines.

As the tools and applications across data pipelines continue to proliferate rapidly, it is crucial to identify a data observability tool that can be integrated into your ecosystem.

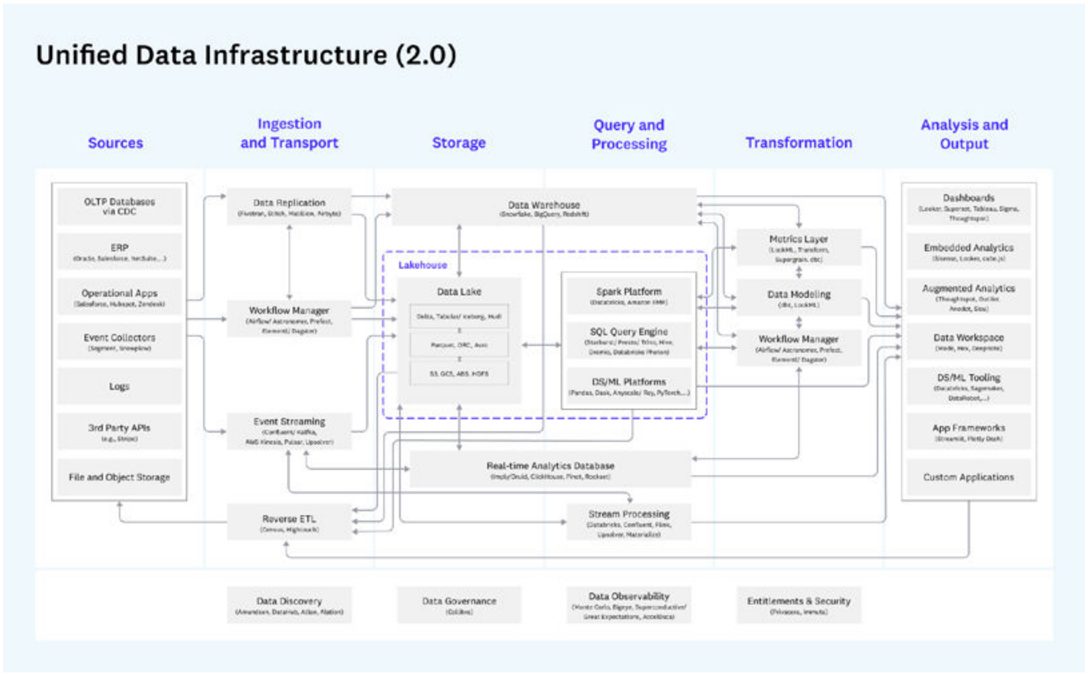

Unified Data Infrastructure (2.0)
Sources
Ingestion and Transport
Storage
Query and Processing
Transformation
Analysis and Output
OLTP Databases via CDC
ERP (Oracle, Salesforce, NetSuite,...)
Operational Apps (Salesforce, Hubspot, Zendesk)
Event Collectors (Segment, Snowplow)
Logs
3rd Party APIs (e.g., Stripe)
File and Object Storage
Data Replication (Fivetran, Stitch, Matillion, Airbyte)
Workflow Manager (Airflow/ Astronomer, Prefect, Elementl/ Dagster)
Event Streaming (Confluent/ Kafka, AWS Kinesis, Pulsar, Upsolver)
Reverse ETL (Census, Hightouch)
Data Warehouse (Snowflake, BigQuery, Redshift)
Lakehouse
Data Lake
Delta, Tabular/ Iceberg, Hudi
Parquet, ORC, Avro
S3, GCS, ABS, HDFS
Real-time Analytics Database (Imply/Druid, ClickHouse, Pinot, Rockset)
Spark Platform (Databricks, Amazon EMR)
SQL Query Engine (Starburst/ Presto/ Trino, Hive, Dremio, Databricks Photon)
DS/ML Platforms (Pandas, Dask, Anyscale/ Ray, PyTorch,...)
Stream Processing (Databricks, Confluent, Flink, Upsolver, Materialize)
Metrics Layer (LookML, Transform, Supergrain, dbt)
Data Modeling (dbt, LookML)
Workflow Manager (Airflow/ Astronomer, Prefect, Elementl/ Dagster)
Dashboards (Looker, Superset, Tableau, Sigma, Thoughtspot)
Embedded Analytics (Sisense, Looker, cube.js)
Augmented Analytics (Thoughtspot, Outlier, Anodot, Sisu)
Data Workspace (Mode, Hex, Deepnote)
DS/ML Tooling (Databricks, Sagemaker, DataRobot,...)
App Frameworks (Streamlit, Plotly Dash)
Custom Applications
Data Discovery (Amundsen, DataHub, Atlan, Alation)
Data Governance (Collibra)
Data Observability (Monte Carlo, Bigeye, Superconductive/ Great Expectations, Acceldata)
Entitlements & Security (Privacera, Immuta)

Image reference: **Emerging Data Architecture by a16Z**

At a high level, you must consider a few of the aspects below around portability in your architecture.

The source systems you will need to be supported in your ecosystem across Databases, Streams, Data Lakes, and Data Warehouses and the ease of integration with these source systems. Specifically, the Data Observability tool can monitor your system without going through a transformation process. Next comes data formats like CSV, JSON, Parquet, and Avro does the system support all the formats your data needs to be read from?

Also, consider the need to support the composability of your existing and new data pipelines.

Data Observability tools will need to handle integration with upstream systems like BI and analytics for impact analysis. They also need to support integration into your existing data catalog, so your catalog can give a 360-degree view of your data assets, including its health metrics.

Lastly, If you want the output of monitoring to drive the pipeline orchestration workflows and automate data remediation workflows, in that case, you will consider integrations into your pipeline orchestration tools like Airflow or DB spark.

ACTIAN™
a division of HCLSoftware

Integrations

Data Warehouses and Data Lakes	• Which data warehouses and Data lakes can the product monitor out of the box? • Explicitly specify your systems: Snowflake, BigQuery, Redshift, Athena, S3, GCS, Delta Lake, etc.
Streaming Sources	• Does the product monitor event streaming sources out of the box? • Explicitly specify your systems: Kafka, Redpanda, Pub-sub
Data Formats	• Does the product natively support monitoring arrays and nested structures? • For example, JSON, Parquet, JSON, CSV, Avro, Iceberg, etc.
Ease of Adding New Data Sources and Systems	• How long does it take to configure and add a new data source not supported through the platform out of the box?
Custom Transformation	• Does the tool need additional data transformation steps to read data from storage systems, like S3 or GCS, to prepare data for monitoring? • If so, does the tool have the ability to write custom transformations using a commonly used language to extract fields of interest from complex structures?
SDKs, APIs	• Does the platform provide APIs/SDKs to create new dataset connections, rules, and data quality processes, or integrate with specialty data quality tools? • Which product capabilities are accessible via REST API?
Circuit Breaker	• Does the product support controlling data pipelines via a circuit breaker pattern when data does not meet data quality or data integrity thresholds?
Data Pipeline Integrations	• Does the tool integrate with orchestration tools like Airflow and DBT to perform quality checks on data between stages and control processing based on quality scores?
Data Catalogs	• Does the platform integrate with data catalogs? Examples: Actian, Alation, Collibra
Analytical Tools	• Does the platform integrate with analytical products such as Looker, Tableau, and Power BI?

ACTIAN™ a division of HCLSoftware

Anomaly Detection

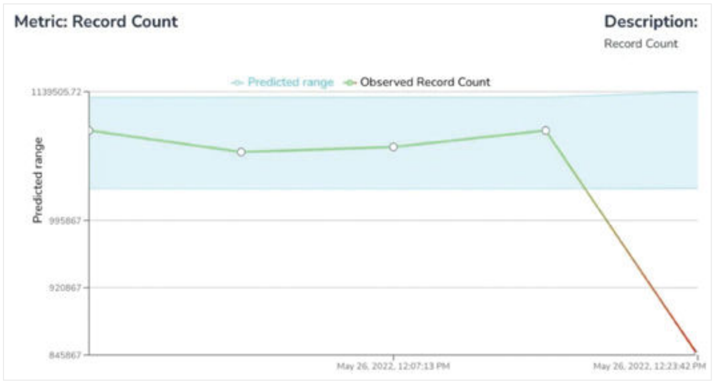
Enterprise data is transformational, hence it goes through multiple changes throughout its journey from ingestion to consumption.

Additionally, depending on the type of data, i.e., 1st party, 2nd or 3rd party, there are often anomalies like drift in values or outlier or out-of-range values. These could be true or false positives, but unless detected and investigated, these are hard to address and have a tremendous impact downstream.

At a high level, a data observability platform can help users detect anomalies in:

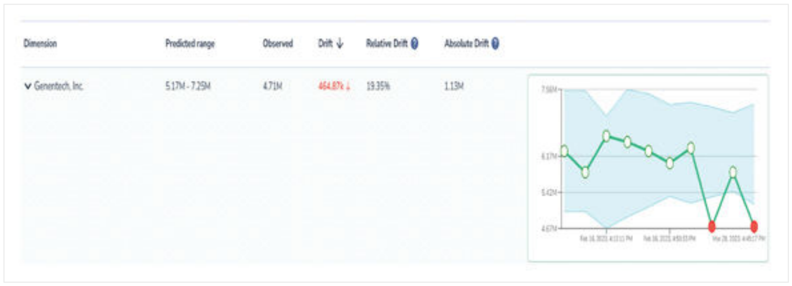
Identifying and monitoring metadata anomalies

A Data Observability tool uses historical information to automatically predict and monitor for anomalies within the behavior of key metadata metrics such as volume, schema, and freshness.



Identifying and monitor data anomalies

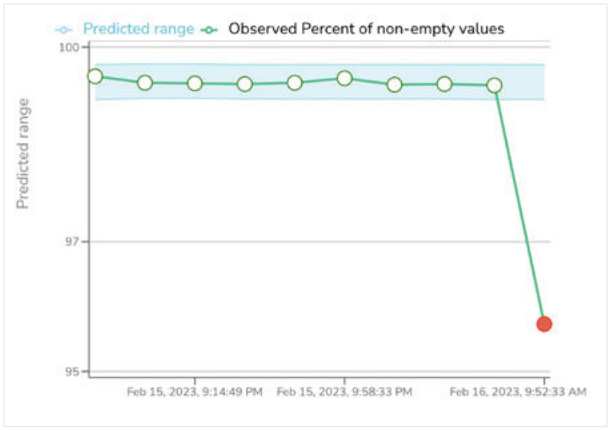
On top of metadata metrics, tools can look at the data values and monitor for anomalies based on completeness (nulls), cardinality, values distribution, range, data types, patterns, and many more characteristics.



Dimension	Predicted range	Observed	Drift ↓	Relative Drift	Absolute Drift
Genentech, Inc.	5.17M - 7.25M	4.71M	464.87k ↓	19.35%	1.13M

Identify and monitor business metrics

Data Observability tools can monitor time series metrics across any number of dimensions and analyze the data and alerts on the detected anomalies.



ACTIAN™
a division of HCLSoftware

Considerations for evaluation:

Anomaly Detection

Machine Learning	How long is a training period required each time monitoring is added to new data before the anomaly detection model can detect issues?
User Input	Can the models adapt to user feedback about false or true positives and become more accurate over time?
Anomaly Detection on Custom Metrics	Is anomaly detection available for custom business logic like calculated and derived attributes? Does this work for both cross-table and cross-field monitoring?
Bring your models	Does the tool support integrating with customers' ML models?
Customization	Do the anomaly detection models support different sensitivity levels for tables, fields, or specific metrics where more or less sensitivity is needed?

Data Quality (DQ) Reporting

Data quality refers to data accuracy, reliability, completeness, and consistency. It optimizes data to align with the desired state and meet the expectations defined in the business metadata. Data quality answers the fundamental question, "What do you get compared to what you expect?"

The outcome of Data Observability is automatically classified into Data Quality indicators and health reports. Since data observability tools are constantly scanning new and updated data for data metrics, these metrics can be used to generate and update data quality reports automatically.

A data observability tool can help automate and report data quality KPIs.

When evaluating a tool, first and foremost, understand what capabilities around DQ metrics are automatic and the degree to which the tool can take customization into account. More often than not, each company and team may have a different meaning of individual DQ KPIs based on business needs and data values.

As a part of Data Observability evaluation, users should also look into their definitions for these KPIs and map them with supportability from data Observability vendors. Below is the list of considerations specific to data quality health reporting.

Data Quality Reporting

Ease of Setup and Out of the Box (OOTB) Metrics	• Specify the data quality indicators provided out of the box. • For example, Completeness, Validity, Accuracy, Consistency, Uniqueness, and Freshness.
Data Quality Indicator: Completeness	• Can the product detect the completeness of records, for example, nulls and empty values? • Does the tool support customizations, e.g., to detect N/A as incomplete data?
Data Quality Indicator: Validity	• Can the product identify the validity of data by checking data values against expected structures and formats? Basic examples: – Emails should have [string]@[string].com – Basic US postal codes should have N=5 numbers. – Complex examples: – Make an API call to check if an address is valid.

ACTIAN™
a division of HCLSoftware

Data Quality Reporting

Data Quality Indicator: Consistency	• Can the product automate data consistency checks? • Use case 1: reconciling target/reference tables. (consistency b/w different systems) • Use case 2: time-series change detection. (consistency b/w different loads)
Data Quality Indicator: Timeliness	• Can the product detect the Freshness of data? How does the product calculate Freshness? Use cases: – Table-level Freshness (i.e., Time-series analysis on data upload that will automatically alert on delays outside expected time windows or on missing data load altogether) – Record-level Freshness (i.e., what is the percentage of outdated records (entities) and whether this percentage increases abnormally) – Entity-level Freshness (i.e. similar to record-level, except it takes into account that multiple versions of an entity may exist in the same table)
Data Quality Indicator: Uniqueness	• Can the product detect duplicate values for any attribute, even nested attributes?
Data Quality Indicator: Accuracy	• Can the product detect inaccurate values based on historical data analysis or defined rules?

ACTIAN™
a division of HCLSoftware

Data Health Reporting

Apart from reporting on the Data Quality indicators, Observability tools can also help report on the overall health of the data across the data pipelines. This includes measuring and quantifying DQ KPIs across all sources in the pipeline but also reporting on total incidents/issues reported via monitoring and their impact and resolution status.

Data Health Reporting

Data Health Dashboard	• Does the product provide easy and intuitive dashboards to show and quantify data quality KPIs and metrics across all sources?
Issue Resolution Reporting	• Can the product show how many problems have been detected, details of outstanding alerts, and the root cause of issues?
Historic Reporting	• Can the history of past results be reviewed? For how many days? • Can the user compare results from today with previous ones (7 days ago, a month ago, etc.)?

ACTIAN™
a division of **HCLSoftware**

Data Rules, Contracts, SLA.

Call it data contracts, SLAs, expectations, or rules; every system in a data pipeline has specific rules and validations that can ensure the integrity and reliability of the data that are stored, processed, and transferred. Many times ML based tools reduce the dependency on such rules as rules are rigid and hard to maintain and manage. However, even with the most advanced ML models, there are key validation on complex business rules that are best done through predefined policies.

A data observability system will not only monitor the data for anomalies but can help users interactively define such checks and then continuously monitor the data flowing through the pipeline against these rules.

These rules can be simple policies like acceptable values for a picklist based on industry code or multi- attribute policies based on business constraints.

While evaluating a data observability tool, users often consider the tool's supportability for building and managing interactive DQ rules.

Data Quality Rules

Rules Based Data Quality	Can the product provide the following capabilities: 1. OOTB field-type rules (e.g. Uniqueness, Completeness, Volume, Min, Max, Std Dev, etc.) 2. Regex (e.g., detect special characters, wildcards, prefixes, and suffixes, to identify complex text patterns or patterns that lead to the discovery of PII data such as social security or credit card information) 3. Controlled lists and lookups (e.g., for ICD codes, ISO codes, zip code or country validations, etc.)
Interactive Data Quality Rule Building	• Does the product provide profiling, pattern detection, and investigation capabilities that can lead to discovering anomalies (without any prior knowledge) and setting up the right data quality metrics? • What is the user experience and timeline between discovering data quality issues, creating new rules, testing the rules, and moving to production?
Custom and Multi-Attribute Rules	• Can the user customize the monitors? • For example, define multi-attribute metrics, calculated/derived attributes, or define custom metrics using joins and group bys. • E.g., monitor address information by tracking the completeness of four attributes (street number, street name, city, and state) together.
Segmentation	• Does the product enable users to monitor different data segments based on different data quality rules? • For example, Row Count/Volume may be different in different regions.

ACTIAN™
a division of HCLSoftware

Monitoring

Data Monitoring and observability are related concepts in managing and maintaining data. While they are often used interchangeably, they refer to slightly different concepts.

Monitoring involves collecting and analyzing data points to understand their health, performance, and behavior. It involves setting up various metrics, alerts, and dashboards to track key performance indicators (KPIs) and alerts when there are issues.

Observability, on the other hand, is a broader and more proactive approach to understanding data systems. So most data observability platforms do have monitoring capabilities.

Monitoring

Full Database	• Can the product detect schema changes, freshness, and row count/ volume changes automatically for all tables in the data warehouse?
Field Level Monitoring	• Does the product proactively monitor – Percents and/or counts of null or missing values? – Uniqueness of values and duplicates? – Freshness of records in each field? • Consistency of values between loads? • Does the product monitor min, max, average, std dev, variance, skew, and percentiles? • Can the product discover data patterns for each field and will display the frequency for each pattern?
Business Metric Monitoring	• Does the product allow the definition of custom (business) metrics? • For example, min, max, count, distinct count, sum, average, etc. on numeric fields like transaction amount.
Thresholds	• What are the different threshold types supported for a custom monitor (absolute, relative, automatic)? • Can the user change the automatic thresholds manually as needed?
Threshold Predictions	• Can the product predict future thresholds based on learnings from historical trends?
Monitoring Frequency	• How adjustable is the frequency of monitoring? Can it be set at schema, table, field, or metric level?
Monitoring Volumes	• Are data accuracy and data validation checks on all data or on samples? • For example, does the validation return a limited number of data quality checks (e.g., first 500 rows that fail to meet the data quality checks)?

Notification/Alerts

The most common outcome of data observability is notifications. A notification can be a soft notification in the UI or an alert via systems like PageDuty, Slack, SMS, etc.

Outside of integrations with alerting channels, when considering data observability tools identify your needs around API access and automation of pipeline workflows using alerts.

For example, users often need to programmatically access alerts and stop/continue the pipeline flow to implement a circuit breaker pattern.

Notification/Alerting

Alerting Channels	• What alert channels are supported? • Examples: Slack, Teams, Email, Pagerduty, Webhooks
Alert Management	• Can alerts be directed to different teams? • Can the user configure the alerting frequency to be different from the validation rule schedule? • Can users mute/snooze alerts when needed?
API Access	• Can alerts be accessed via API? • For example, to label data in a data catalog or to redirect pipelines based on data observability alerts.

Actions and Remediation

Once the monitoring system detects errors and anomalies, there could be multiple options for the next steps. The obvious one is alerting on Slack, pager duty, etc.

Key considerations in this area are: Does the tool give field-level insights and help quickly get to the source of the issue using column-level and system-level lineage?

This workflow works specifically well for unknown and new issues; however, once data engineers start seeing repeatable patterns on issues, they may want to automate the subsequent actions, i.e., create service tickets in JIRA.

Also, often companies are leveraging the outcome of data observability to constantly separate good and bad data into separate bins (buckets) and, depending on the use case sending bad data for remediation or storing separately so they can not only reduce the cost of storing, processing and transferring good data but also enables users to train ML-models on quality controlled data.

Actions and Remediation

Field Level Insights	• Is the solution able to suggest the root cause of data issues down to the field level, with no need to define manual rules for each case?
Issue Management	• Does the product detect issues, alert feedback, users involved, and resolution notes as reference for future issues? • Can the product automatically surface row-level details of data that contributed to a detected anomaly for use in debugging? Is this a preview/sample level information – or all data?
Data Binning	• Can the product persist good and bad records identified based on certain rules/scenarios into separate locations/tables? • This ensures continuous data processing as data pipelines don't have to pause good data records because of bad data records.
Lineage	• Can the product auto-generate table-level lineage within the warehouse? • Can the product auto-generate column-level lineage within the warehouse? • Does lineage extend beyond the data warehouse?

User Experience

While the feature functionality is paramount for your data observability solution, usability shouldn't have to be a tradeoff!

This is important because Data reliability is a highly distributed function across source owners, data engineering, product owners, and business users. Having a tool that's easy to adapt and collaborate across these teams will significantly impact overall ROI.

Identify this tool's key user persona in your organization and evaluate the user experience with them. For example: If the primary users are non-sql users, ensure that the tool does not implicitly assume SQL skills.

User Experience/Low-code, No-code

Usability	• Flexible, customizable dashboards that technical and non-technical stakeholders. • Interactive tool for investigation of issues. • Role based access and customizations of UI.

Time to Value

The first and foremost consideration when evaluating a tool is time to value. The time-to-value would depend on multiple factors like ease of use and set-up, time is taken by the tool (ML-based) to train its models, and start alerting, training, and on-boarding needed.

Time to Value

Time to Value	• What is the typical time to value before monitors can be set up on a system? • What is the typical time for detection? For example, how long does model training take before starting to see anomalies and outliers in the data? • If sampling is used, what monitoring frequency and what clustering techniques should be set to ensure all/most data is analyzed for data quality issues? How would this affect our time to value? • Training and onboarding time?

ACTIAN™
a division of HCLSoftware

Scale/Performance

When looking into a data observability tool, there are two critical considerations around scale and performance, the first being the tool designed to scale quickly to your current and future data needs. The elasticity of the tool should be considered for all dimensions of data, like volume, velocity, and variety. The second is the performance implications of data observability on the underlying data infrastructure. Data observability tools are metric calculation systems that require a high degree of computation to support scale. This processing will often get pushed to the underlying infrastructure, causing latency and performance degradation.

Techniques like change data capture(CDC), delta processing, and samplings can be adopted to reduce cost and improve performance; In contrast, support for CDC and delta processing is very important, and sampling can become a limitation based on business needs. For example: If users need exact records (IDs) of anomalous data for remediation, sampling techniques will not work.

Scale/Performance

Scalability	• Does the product scale elastically to the amount of data monitored, or is sampling techniques needed to limit the amount of monitored data?
Change Data Capture	• Can the product process incremental data on a daily basis (without needing to reprocess all of the data) to detect drift and other patterns?
Data Warehouse Performance	• How will the product impact the data warehouse performance that it monitors? • What techniques are used to validate the data? (e.g., via sending SQL queries down to the database)
Customization of Rules without Access to Production Database	• Is the user able to configure new metrics independently and without the need to touch the production database, for example, to create custom views?

Support SaaS, on-prem and hybrid deployments

While SaaS comes with many benefits of dynamic scale and reduced operational overhead, there are instances where a project team, line of business, or entire organization must operate in a private installation.

Look for vendors that can accommodate these needs reasonably and in the future.

Deployment

SaaS	• Is a fully managed, hosted version of the platform available?
Private Cloud	• If a private cloud is available, explain how provisioning takes place. How long do provisions and upgrades typically take?
Cloud Platforms /Vendors	• Do you support monitoring across multiple cloud availability zones or vendors?

Security

Every Data Observability vendor that gets qualified will undergo a rigorous security assessment by the InfoSec team; front-loading some security and compliance criteria will help you find quality tools that meet your needs, specifically if your company has strict policies around certifications.

Security

Certifications	• Please list all certifications. – Examples: SOC2 Type II, GDPR, CCPA, ISO 27001 standard
Security	• Do the product require the installation of an agent within our environment?
Data Privacy and Retention	• How does the product handle data privacy? • What are the data retention policies that the solution adheres to?
Multi-Tenancy and Data Residency	• Is the solution multi-tenant? • Where are the data centers located?
Single-Sign On	• Does the tool provide single sign-on?

ACTIAN™ a division of HCLSoftware

Total Cost of Ownership(TCO)

While evaluating a Data Observability tool, licensing cost might be the top-of-mind consideration. Specifically around licensing, there are different pricing metrics from vendors to consider here - usage- based pricing based on several records, tables, underlying monitored systems, or even user seats. You must consider current and future growth to calculate the licensing model.

Since most tools today partially rely on customers' underlying infrastructure, you might have to understand the additional infrastructure cost to you - consider measuring Ingress/Egress, metric and data storage, compute and processing, including warehouse queries—the best.

TCO

Licencing Model	– What is the pricing model? For example, consumption-based by volume, seats, data systems, etc.
Additional Infrastructure Costs	• Do data monitors run in the database/data warehouse (via query validation) or does the platform provide its own compute layer to create a predictable cost? • If monitoring runs inside the database, what is the increase in usage costs as a result?
Administration Costs	• What is the typical cost of maintaining, upgrading, and managing the system?
POC Cost	• Is the solution multi-tenant? • Where are the data centers located?
Single-Sign On	• What is a typical POC timeline? • How many resources are needed for a POC?

Start your data observability journey today.

Request a personalized demo and one of our data observability experts will help you:

- Understand how Actian Data Observability provides comprehensive visibility and ensures the health of your data pipelines.
- Explore the features that proactively detect issues and minimize business impact.
- Get insights into integrating data observability seamlessly with your existing data stack.

Tel +1 512 231 6000 Fax +1.512.231.6010 710 Hesters Crossing Road, Suite 250, Round Rock, TX 78681 **Actian.com**

ACTIAN™ a division of HCLSoftware