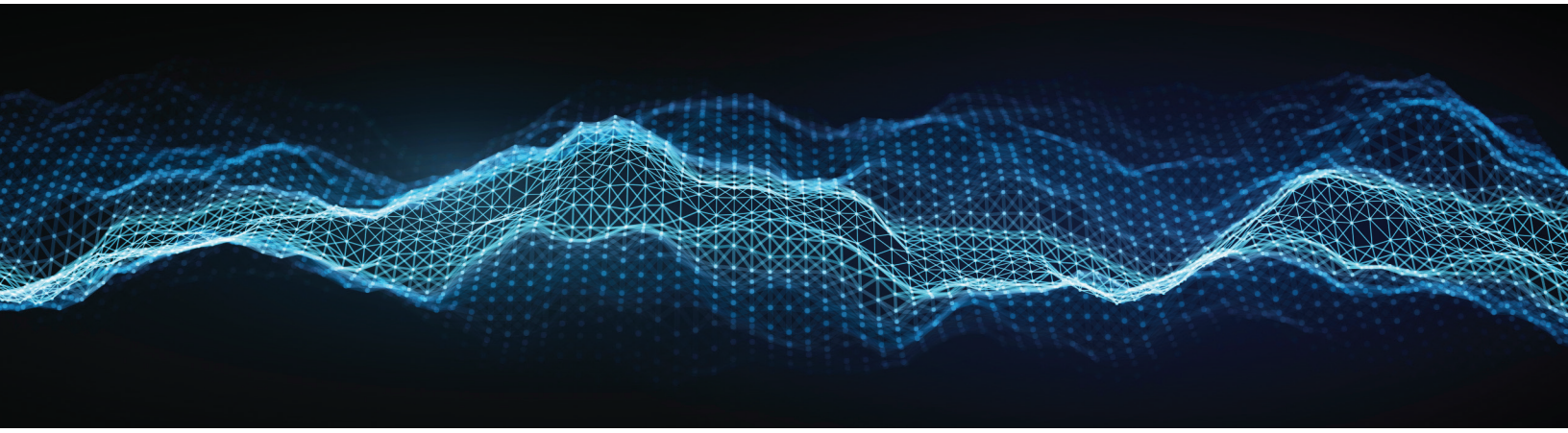
FirstEigen White Paper



13 Essential Data Validation Checks for Trustworthy Data in the Cloud and Lake

FirstEigen™ | Gartner Cool Vendor | IDC Innovator

www.FirstEigen.com • contact@firsteigen.com

Trustworthy Data in the Cloud and Lake: Best Practices

When data is moved into and out of a Data Lake and/or the Cloud, the IT group and the consumers of data are faced with some common questions:

- *"How can we be sure we can trust the data in the Cloud/Lake?"*
- *"What is a quick, rigorous and efficient process that can engender great trust in the data?"*

The only way to create trust is to validate every piece of data every time it moves in and out of a repository.

Challenge

Every repository has a different set of rules that holds the data together. Each of the 1,000's of tables and files within each repository has uniquely different data validation rules. Making it very hard to identify, create and maintain 100,000's of rules for even medium sized repositories.

Exhibit 1

The framework to validate data at ingestion or at any location data moves.

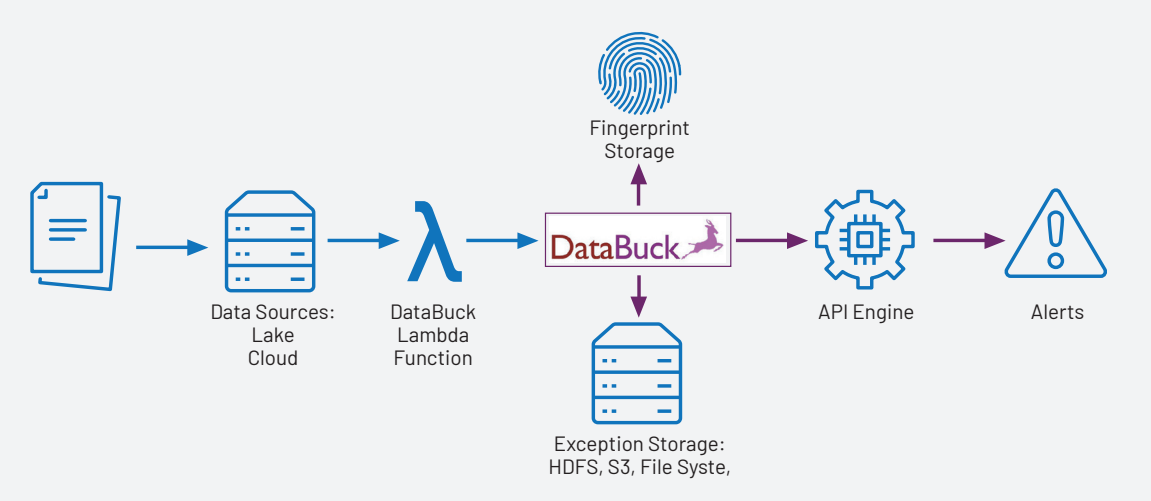


Table 1 and 2 are the collection of the best practices questions that Lake and Cloud users ask. The most important Data Quality validation checks are dynamic checks and are very hard to code (or not possible) and maintain using traditional approaches. **When these questions are operationalized it translates to over 1,000 data validation rules per data set.** FirstEigen's DataBuck addresses each of these questions easily with out-of-the-box capabilities, without needing any coding or configuration. It reduces a process that usually takes 9-12 months to just 1-2 weeks.

Splitting the overall Data Quality checks into these two levels of checks dramatically increases the reliability of data.

1. Lake-Level Checks

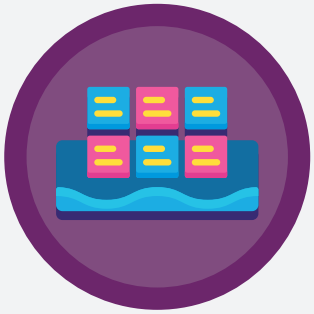
These are basic sanity checks on incoming data without any SME involvement. Errors like duplicates or incomplete dataset or lack of conformity or validity should be quickly caught and the source contacted to send better quality data. It'll be too late if these errors are discovered weeks or months later at the time of data consumption. If left unfixed, they propagate throughout the organization costing 10x more to fix the damage. Using the traditional approach, it would take years to onboard essential DQ checks for a Lake. With next-gen DQ tool, like DataBuck, DQ program on a Lake can be in production in 2 weeks.

2. Application-Level Checks (at Point of Consumption)

This validation is business-specific or use case-specific. It will need SME involvement. The SME should not waste their time on creating basic Lake-level checks. They should be working with reasonably clean data to create more advanced checks like, anomalies or validating inter-column relationships, etc.

Exhibit 2

Data Quality validation must occur at two levels.



Lake-Level Validation

Ensure basic sanity in the lake.
Completeness
Uniqueness
Conformity

Application-Level Validation (at consumption)

Business Specific
Application Specific
Deep SME Involvement
Consistency
Accuracy
Etc.

Table 1: Operational Best Practice for Lake/Cloud DQ Validation

(*ODTA- Operationalizing Difficulty with Traditional Approaches and Tools)

#	Operational needs for DQ Validation in Lake/Cloud (Success Criteria)	ODTA
1	**Automated data quality health checks upon file arrival in an AWS S3 bucket.** A "listening bot" will be invoked when a new file is created in S3. 1. The listening bot should schedule data quality checks that should be run and order the actual execution of the DQ checks on the file. i. The file will have metadata that relates the file to a logical table. The DQ checks should be defined for the logical table instead of the individual file. ii. The listening bot should check if the logical table already exists. • If the logical table is new, essential DQ checks should be automatically created. • If the logical table already exists the defined checks for that table should be executed. iii. Order the execution of the data quality checks on the file in the DQ validation software. 2. Must queue and execute the ordered DQ checks on the file. 3. Notify AWS SNS when the DQ checks on the file are finished. 4. A Lambda Function is triggered upon the notification in AWS SNS i. The Lambda Function must pick the result from DQ validation software and push it into the regular monitoring flow for the Data Lake. 5. The processing in the asynchronous pattern should be done using a Hadoop cluster. Only the dashboard and configuration part of DQ validation software should be part of the cloud compute installation.	M
2	**No coding required.** DQ validation software must be able to automatically generate a set of tests with minimum manual intervention.	H
3	**Tailoring and filtering of pre-defined tests.** Tailoring and filtering of pre-defined tests.	H
4	**Reusable templates.** Ability to define a template that could be reused on tables and files with the same layout. Template must be created the first time it's seen by DQ validation software. If another file with a different name, and the same meta data, arrives in the same bucket (folder) software should be able to use the previously created template and rules to validate the data.	H
5	**DQ validation software must have the ability to be invoked via API's.**	H
6	**Results in DQ validation software must be available through API's.** Able to retrieve DQ validation results through API's.	L
7	**Checks in results can be scripted through API's.** The ability to script and configure checks in DataBuck through API's.	L
8	**Able to process Apache Parquet files stored in AWS S3 through API's.** The ability to process the data as it will be available in the CUSTOMER Data Lake.	L
9	**Able to process a Hive table based on a set of files Parquet files stored in AWS S3.** The ability to process a logical dataset using Hive that is based on multiple files in the data lake.	M
10	**Workload Management.** Ability to queue and control the processing of jobs.	M

Table 2: Data Quality Functional Best Practice for Lake/Cloud

(*ODTA- Operationalizing Difficulty with Traditional Approaches and Tools)

#	Functional Needs for DQ Validation in Lake/Cloud	ODTA
1	**Validate completeness of data transfer between Cloud Data Lake (eg., S3) and Data Warehouse (eg., Redshift). Example of rules:** 1. Schema-level Fingerprinting 2. Table-level Fingerprinting 3. Cell-level Fingerprinting	M
2	**DQ rules in the Completeness Dimension. Example of rules:** 1. Check percentage of Null values in a column 2. Check if the number of records for every micro segment is in line with its historical expectations	M H
3	**Data Quality rules in the Uniqueness Dimensions. Example of rules:** 1. Check for duplicates on a logical business key 2. Check against a "master data" table/file	L
4	**Data Quality rules in the Timeliness Dimensions. Example of rules:** 1. Check for a continuous timeline on a logical business key (valid_from to valid_to columns on multiple rows has no gaps) 2. Check for overlaps in the timeline on a logical business key (valid_from to valid_to columns on multiple rows has overlaps)	VH
5	**Data Quality rules in the Validity Dimensions. Example of rules:** 1. Check if all values (in an attribute) are dates 2. Check if all values (in an attribute) are timestamps 3. Check if all values (in an attribute) are numeric values 4. Check if all values (in an attribute) are integer values	L
6	**Identify records with uncommon inter column relationships.** Check for multi-column 'If… And If… Then…' relationships. (Dynamic DQ Rule) • Eg., IF Col_A = TERM, AND IF Col_B = Loan, AND IF Col_E = 1 or 5 or 10, THEN Col_F = DISCOUNT_Loan	H
7	**Anomalous relationships between date columns (Dynamic DQ Rule)** • Eg., Claim_Date > Plan_Join_Date, or Current_Date – Birth_Date < a reasonable value, etc.	H
8	**Orphan and "New Born"- Identify new values in columns that have not been seen historically (Dynamic DQ Rule)** • Eg., for a bank transaction check counterparty type or Fed counterparty type and the US LCR counter party type.	H
9	**Data Quality rules in the Accuracy Dimensions. Example of rules:** 1. Range check on Dates 2. Range check on Numeric values 3. Check against a predefined list of values 4. Check against a reference data set (eg., Currency codes, etc.) 5. Check against a regex pattern	L
10	**Flag anomalous individual records in a batch** • Eg., for a loan data set flag loan records where the base rate, spread-rate and the loan amount are anomalous compared ro other loans in the same combination of profit center, loan type and loan duration (Dynamic DQ Rule)	H
11	**Anomalous individual records w.r.t. history- consistency with historical patterns for this time of the year (Dynamic DQ Rule)** • Eg., a participants sudden change in Retirement_Savings or Salary compared to its own past data in previous client file • Eg., in a trading data set dynamically check if a product suddenly sees new transactions with new counter parties. Some products or counter parties may have been dormant and show sudden activity.	VH
12	**Identify anomalies of "cumulative" values or column aggregate values within a micro segment in comparison to historical behavior (Dynamic DQ Rule)** • Eg., for a micro segment of the data set check if the Total_Savings_Contribution or Total_number_of_Records is in line with historical patterns	VH
13	**Data Quality rules in the Consistency Dimensions. Example of rules:** 1. Check values against another dataset	M

FirstEigen's DataBuck is recognized by Gartner and IDC as the most innovative data validation software for the Lake and the Cloud. By leveraging AI/ML it's >10x more effective in catching unexpected data errors in comparison to traditional DQ tools. DataBuck's out-of-the-box 9-Point Data Quality checks needs no coding, minimal human intervention, and is operationalized in just a few clicks. It increases the scalability of discovering and applying essential data quality checks to 1,000's of tables by auto-discovering relationships and patterns, auto updating the rules, and continuously monitoring new incoming data. It reduces man-years of labor to days.

For further information contact@FirstEigen.com

FirstEigen™ | Gartner Cool Vendor | IDC Innovator

www.FirstEigen.com