FirstEigen Case Studies

How DataBuck Automated Cloud Data Quality at Fortune-1,000 Companies

FirstEigen™ | Gartner Cool Vendor | IDC Innovator

www.FirstEigen.com • contact@firsteigen.com

Summary

These are the learnings from many successful Cloud Data Quality (DQ) implementations at Fortune-1000 companies. These customers were able to detect data errors upstream and stop the errors from spreading to their business partners. They automated 80% of data quality validation using a two-step process. These successful organizations leveraged autonomous DQ monitoring and validation to reduce cost of DQ by over 50%, mitigate DQ errors and reduced time to market by 90%.

These successful organizations data quality realized that Cloud Data Engineers cannot be expected to be knowledgeable about every column of every table. It's impossible for them to validate and certify the accuracy of data. As a result, companies end up monitoring less than 5% of their data. The other 95% is dark Data – which is unvalidated, unreliable and risky.

FirstEigen's DataBuck is a continuous data monitoring and validation software for catching elusive data errors very early. Powered by AI and Machine Learning (ML) it auto-discovers essential and advanced data quality rules for each dataset. It automates data validation for reliability and accuracy of data to discover issues for each dataset. It easily integrates within data pipelines through APIs.

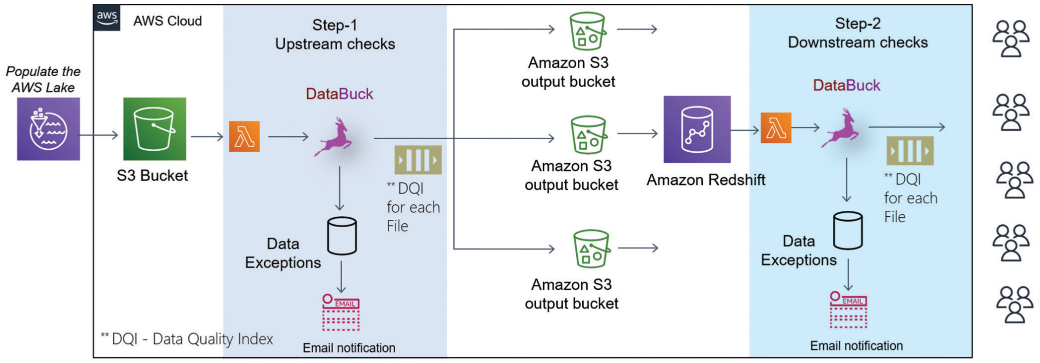
DataBuck autonomously detects data errors upstream and cuts data maintenance work and cost by over 50%. Users can certify the health of their Data Quality at every step of data flow automatically. DataBuck improves Data Engineers' experience, accelerates the company's journey to cloud.

Data Quality Best Practice

Data Quality validation requires the right processes, right sequence of data quality validations and the right tool. Different levels of checks for Data Quality validations have to be deployed in a layered approach to be efficient and effective (Exhibit-1). Then, automating a large part of it will decrease Data Engineers and Data Stewards work effort by 80%. DataBuck operationalizes an automated sequence of checks for the two-step process using Machine Leaning (ML) and Data Fingerprinting in a few clicks.

Exhibit 1

Two step autonomous Data Quality validation.



DataBuck – Benefits

People Productivity Boost >80%

Reduction in Unexpected Errors: 70%

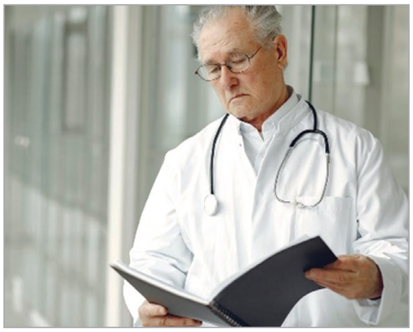
Cost Reduction >50%

Time Reduction to Onboard data set ˜90%

Increase in processing speed >10x

Cloud Native

Customer: Healthcare



A Telemedicine and Healthcare company's data volume was rapidly growing due to increased demand for Telehealth services. They use AWS S3, Redshift, mySQL and other technologies. Their ability to process Member Eligibility Files was becoming the bottleneck. Eligibility files contain all data that determine how much the insurance pays the healthcare company for every Patient-Doctor interaction. Delayed file processing or dropped files is direct loss of revenue. They were unable to flag the data issues and file corruptions the eligibility files in a timely manner. The Data Quality validation process is manual work which was resulting in enormous delays and impacting revenue cycle.

Challenges

- Lack of SMEs to define comprehensive and effective Data Quality Rules.
- Lack of Resources to code and update Data Quality Rules.
- There was no visibility into the eligibility data quality and process issues for timely root-cause analysis, audit, and action.

Outcomes

DataBuck was deployed upstream and at ingestion to auto discover Data Quality rules.

Auto-discovered data quality rules were deployed in less than 15 days for 500+ data assets using just 2 resources. A process that would have taken over 100 person-weeks with the traditional approach was completed in 4 person-weeks.

Several unexpected data issues were detected within a week of deployment.

Early data error detection prevented costly cleanups and data reworks.

Dashboards gave good visibility to health of data sets streamlining business audit and approval.

Customer: Financial Services

A Global 1000 Financial Services organization moves credit risk information from 26 different systems to AWS. Data goes through a number of normalization and transformation before being pushed AWS S3 based data lake. A reporting solution is then used to provide various regulatory reports to various groups.

Challenges

Customer wanted to track over 2200 segments of credit risk data based on the country, product type and risk code combination.

The data had to be validated as soon as it was ingested into S3.

They wanted to flag the following types of data errors stemming from either the source system changes or transformation errors.

- Invalid Country-Product-Risk score combination. If a new combination is created (because of source system changes or transformation issues), downstream stake holders need to be altered. In addition, customer wanted to flag any new country code, product code or risk code.
- Abnormal changes in the number of transactions in each of the 2200+ Country-Product-Risk Code based segments. Anomalous credit risk transactions within each of the 2200+ credit risk segments.
- Anomalous changes in the sum of the credit risk in each segment .
- Technical Data Validation Checks (Null value percentage for each segment), duplicate records etc.

Outcomes

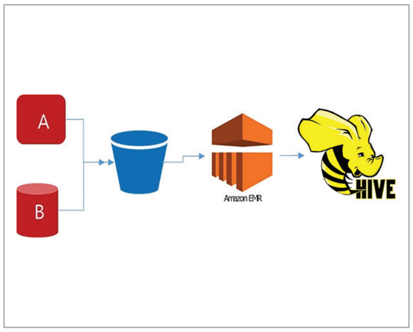
DataBuck's ML algorithms autonomously identified over 2200 credit risk subsegments from the banks transaction dataset and created data fingerprints to track each risk segment individual.

DataBuck autonomously constructed detailed data quality fingerprints for monitoring and validating transactions. This was equivalent to over 22,000 traditional Data Quality rules. These DQ rules were self-learning and evolved with the evolution in data over time.

Within 2 weeks of deployment, DataBuck identified 31 errors that the SME's had not anticipated ("unknown-unknowns") and so the existing data quality solution did not catch. The hidden risk of bad and inconsistent data propagating downstream was minimized.

Customer: Financial Services

A Life Insurance and Annuities Company created a financial data lake in the AWS Cloud. It was to become the source of truth for all financial reporting and business analysis. Data from multiple legacy systems and insurance administration systems flowed into the data lake on a daily and monthly basis.



Challenges

Customer wanted to detect data issues prior to posting the data in the financial data lake. They wanted to ensure referential integrity, completeness of the data movement and reasonability of financial numbers.

Enterprise data quality team estimated that it would take a minimum 6 months, 4 additional contractors to deploy essential checks for 10% of the critical data assets (over 500).

Outcomes

DataBuck was also configured to automatically create data quality and data matching checks if it detects new data assets in the AWS S3. On deployment, DataBuck autonomously profiled and created 9-Point data quality checks for all the data that landed from the legacy systems to the AWS bucket.

DataBuck autonomously created data matching rules to reconcile the data between AWS S3 bucket and Hive database.

In less than 8 weeks, DataBuck was able to establish checks for over 500 data assets with the help of one resource.

Customer: Healthcare/Pharmaceutical

A pharma company was receiving over 70 different sources of data on a daily and weekly basis. The data was from several healthcare providers, sales data aggregators, Point of Sales (POS) systems, internal operations, and customer data. They collated and validated the data in the Cloud and used it for running their operations. The quality of data from these systems were outside their control, they were also not measured, leading the firm to second guess insights that were generated.

Challenges

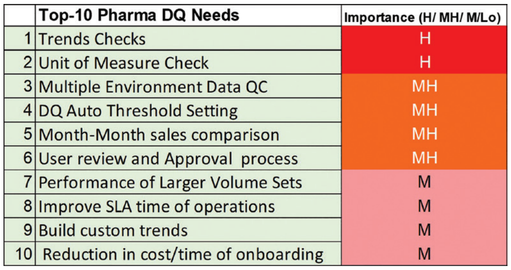
Their traditional Data Quality validation solution was labor intensive. It needed constant code tweaking to manage the changes in the file structures, formats, column details and more. This led to delays in using the data. They were not able to do any advanced checks as they just didn't have sufficient resources. This led to errors in Unit of Measures, approvals, misbalances in supply and numerous delays in regular sales reporting and associated compensation.

Outcomes

Two-Step data quality validation was implemented with DataBuck. DataBuck auto-discovered essential and advanced data quality rules leveraging ML. Rules and thresholds were easily managed without any code changes directly from a dashboard.

Health of datasets were automatically measured by DataBuck as soon as they arrived. The data vendor could be informed immediately if data quality health score did not meet the SLA.

DataBuck's out-of-the-box, pharma-specific capabilities solved their top-10 needs for pharmaceutical data validation. Business users could get more information, with more accuracy, without depending on the IT team. They could separate out a data quality issue from trend changes. Time to business insights was cut dramatically.



	Top-10 Pharma DQ Needs	Importance (H/ MH/ M/Lo)
1	Trends Checks	H
2	Unit of Measure Check	H
3	Multiple Environment Data QC	MH
4	DQ Auto Threshold Setting	MH
5	Month-Month sales comparison	MH
6	User review and Approval process	MH
7	Performance of Larger Volume Sets	M
8	Improve SLA time of operations	M
9	Build custom trends	M
10	Reduction in cost/time of onboarding	M

IT team need fewer resources as DQ was automated and streamlined. Fewer false positives also meant less research to understand data errors, less data rework as a lot of errors were captured upstream in Step-1 of validation.

Customer: Financial Services

A bank in the USA was on a data modernization journey. Their current system was a patch work of the usual suspects of traditional DQ tools. With rapidly evolving data needs they were unusable and unreliable relics.

Challenges

They were unable to handle large datasets and rules had to be recreated for multiple platforms. Security and compliance were a concern as data had to be moved where the data quality rules were. The bank had to have extensive hardware to manage all the back-and-forth data movement. Rules writing process was laborious and time consuming, so their DQ team had grown to over 200 people. Unexpected errors were still slipping through to the business users who were using it for analytics and for reporting.

Outcomes

Two-Step data quality validation was implemented with DataBuck. AI/ML supplemented Bank's user defined metrics with additional essentials and advanced data quality rules that were auto-discovered. DataBuck's superior architecture has no limitations on data volume, tables with >1 Bil rows could be validated in 30 minutes, which previously they could barely process even in 24 hrs. Compliance and security increased as DQ rules were taken to where the data is, not data to the rules.

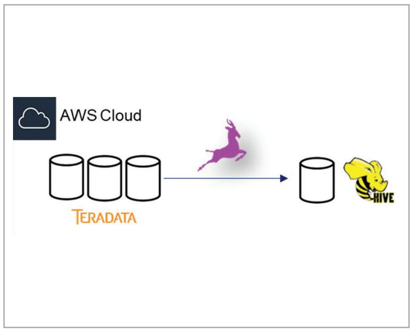
Productivity Boost >100%

Expected Cost reduction >50%

Data set onboarding time reduction ~90%

Customer: Financial Services

A multinational bank that had a grown through acquisitions, operates in 10 countries, with different country-specific regulations, different reporting formats within their own bank and many different systems. Country specific Data was stored on the AWS Cloud in Teradata and were moved into Hadoop for processing. The data from Hadoop propagates downstream through other systems and consumed by the business users. They were detecting many errors and did not trust the data as accurate and reliable to support their decision making for strategic imperatives and day to day Operations.



Challenges

Country users and central users often had mismatched views of the multi country banking data flowing into the data lake daily. There was no unified understanding of the definitions of the data elements. Their KYC initiative was stalling as invalid or nonstandard contact information made the customers unreachable. They also could not cross sell or upsell new services to customers who would have bought them readily. The bank was worried that missing, invalid or default entries in mandatory customer KYC data was a regulatory compliance risk.

Outcomes

DataBuck was configured for autonomously validating technical and historical consistency checks at ingestion, including for completeness, uniqueness and conformity of all data that comes to the Hadoop lake. Without any human inputs DataBuck auto discovered relationships between different columns, for e.g., customer's Gender and customer salutation to alert against potentially mismatched data.

Invalid data in KYC and contact-ability attributes of customer data were identified by applying advanced pattern matching rules. DataBuck also discovered unexpected findings which even the bank Analysts were not aware of, for e.g., "First Name" was not a mandatory field for some of their LOB's.

In less than 4 person-week effort, DataBuck was up and running to generate role specific daily extracts of Customer KYC exceptions.

Valid and Contactable Customer Data means Improved productivity of bank personnel at a lower cost.

Improved completeness and accuracy of KYC data reduces regulatory compliance risk.

AI and ML based rules discovery ensures better coverage and faster results with a smaller and more empowered Data Quality team.

Customer: Global Networking Equipment Manufacturer

This Fortune-100 Technology company was migrating its Teradata based financial warehouse system to Postgress system. They needed help in migrating over 4000 tables and billions of records into the new system – a manual and tedious task to compare where data transferred accurately. Their technology stack included Teradata, Postgress, Snowflake and many Cloud components.

Challenges

Financial organization who were consuming the data, were surfacing data errors undetected by the technical team. There were a lot "unknown-unknown" errors that the SME's had not thought of to check. The traditional approach to data quality validation was not scalable to validate and monitor 800+ data assets within the nightly processing window.

Outcomes

DataBuck's "Autonomous DQ Module" helped migrate Financial Data Warehouse tables error free saving the client over $1.7 million. This reduced financial reporting risk by leveraging DataBuck to monitor 800+ data assets in its financial data warehouse and reconciling financial information

Deployed auto-discovered data quality rules in less than 30 days for 800 data assets using just 2 resources, which in a traditional approach would have taken 500 person-weeks.

DataBuck has provided 40% reduction in manual work by automating data quality process. It also reduced "time to market" (publish verified reports) from 8 weeks to 2 weeks – something that was unthinkable before.

Detected several unexpected data errors that would have impacted financial reports.

In daily runs DataBuck reduced data validation time from 11 hours to less than 2 hours so they could execute all the necessary validations within their audit window.

Automation reduced manual work for data quality validation by 40%.

Reduced time to market from 8 weeks to ~2 weeks.

Fast implementation.

Auditability and Visibility: Executive dashboard and detailed audit trail for each table.