

About



Major University in the USA

This Major University is a groundbreaker in higher education. An eminent and esteemed research university in the world and a leader in global education, with most international students in US. It is one of largest private universities in the US.



Datagaps

Datagaps is passionate about data-driven testing automation. Our flagship solutions, ETL Validator, Data Flow & BI Validator are designed to help customers automate the testing of ETL, BI, Database, Data Lake, Flat File, XML Data Sources.

Summary

We recently had an interesting discussion with our higher university client regarding the value they were getting out of our software. As trusted advisors to our clients, we often conduct these sessions and are often surprised regarding the creative uses they come up with to solve their unique problems and the perceived value they observe.

Our relationship began two years ago when they began to migrate to a new ETL tool. They needed a way to validate that the old and new were producing the same results. They developed 1500 use cases using our automated test case generation wizards to verify the new ETL tool was producing accurate results. Where there were discrepancies, they would fix them and rerun the ETL process and the validation jobs. Most of the 1500 use cases are still used to ensure that the new ETL processes provide accurate transformations and catch issues as they creep into the processes.

The University is in the process of migrating their data warehouse to Snowflake and plans to use our tools heavily to ensure that they are able achieve **100% Data Validation.**



Our tools they use to provide



End-to-End data migration testing & ETL validations.



Validate their data analytics platforms accuracy & performance.



Complete the data journey by monitoring new data as it comes into their Snowflake Data Warehouse.

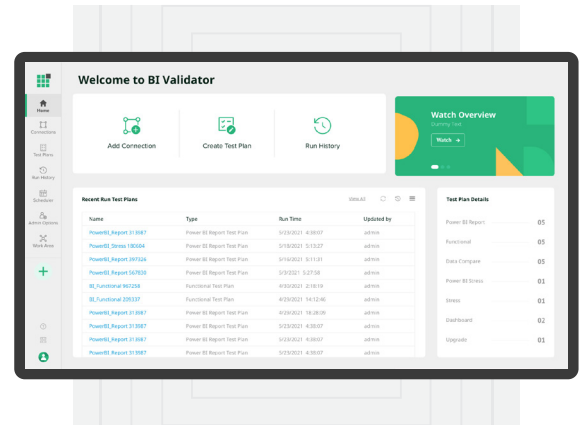


Solution

The success with ETL Validation led them to think about the other area where issues creep into the different data analytics processes. While source vs target comparison is essential during the ongoing ETL process, when that process is solidified, the problems that need to be found fall within the data analytics going against the validated data.

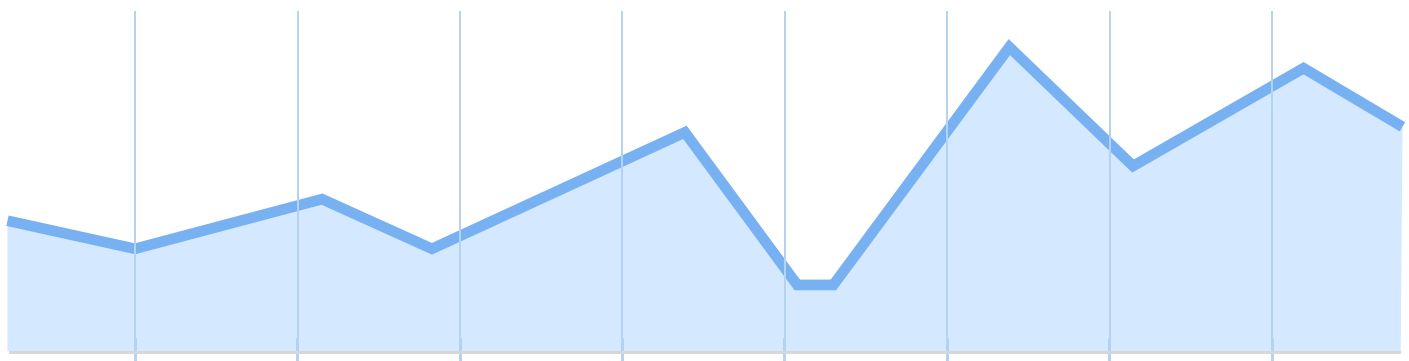
- Calculated fields can be incorrect.
- Release to release issues come up.
- Filters and parameters need to be validated as the input data is changed once to twice a month.

They use our **BI Validator** to set up comparisons between Reports and Queries, compare the visuals before and after changes to the application or releases of the software.



Progress

Noticing the success of these two key components, the University’s Data Governance team started using our tool to assess the quality of their data in their Student Information System (SIS) and the CRM system (Slate). Through our data quality scoring capability, they can see the DQ score of the overall student census data or subsets in the DQ Portal. Universities collect and publish their student census data online throughout the year. The University’s issue was that mistakes while validating the census data began to increase due to the deadlines that were fast approaching—waiting to the end to find these mistakes caused further delays and were more costly to fix. Any mistakes in the data being published can put a dent in the University’s reputation. Using the DQ Scoring allowed them to run the process anytime during the ramp-up to the deadline and find the mistakes and correct them along the way.



Data Quality Score of Overall Student Census Data



Using the DataOps Data Quality solution, they were able to cut down the resources required to validate the data by 66% while improving the quality of the data being published. The University was also able to reconcile data between their SIS and Slate systems successfully.

They described a simple example of how data errors are introduced in the first place. As with most organizations, multiple applications create that system of record. Errors get introduced when the various systems are not kept in sync. Something as simple as changing the values in the Gender field in one system but not the next creates downstream problems for your ETL and Data Analytics processes.

Key Benefits



Cost Saving

Significant cost savings as a result of automating the data validation process.



Reduce Time

Improved agility by reducing the amount of time it takes to ingest new data sources into the Warehouse.



Improved Reputation

Improved organizational reputation as much of the data analytics is published externally.



Trust

Improved trust with the stakeholders as the data quality of the Warehouse remains at a high level.

The Truth behind poor Quality of Data

Our decade of experience is captured in the model :

Input Parameters

User Name	20
Estimated Percent Data Errors	0.1%
How often do these files arrive yearly	52
Percent of data validated	10%
Average Volume Per Feed	100,000
Average Columns Per Feed	25
Minutes to fix if found early	1
Minutes to fix if found early	5

Findings

The cost fixing the data errors early during ingestion is **\$590,850**. However if the issues are found after the data reaches the DW they are 5 times more costly to fix: **\$2,954,250**. The figures are based on typical measures and minimum wage. Many people think about the amount of records that they have. We have found that the better way to think about this is based on the number of columns that might have errors. In this example there are **2,600,000,000** data elements that could have issues per year. For every **20.00%** improvement in automated corrections you save an estimated **\$118,170** cost in found early.

For a more in-depth discussion regarding how Datagaps can help you attain the same results, or if you would like the model to put your numbers in,

Please contact us on Marketing@Datagaps.com