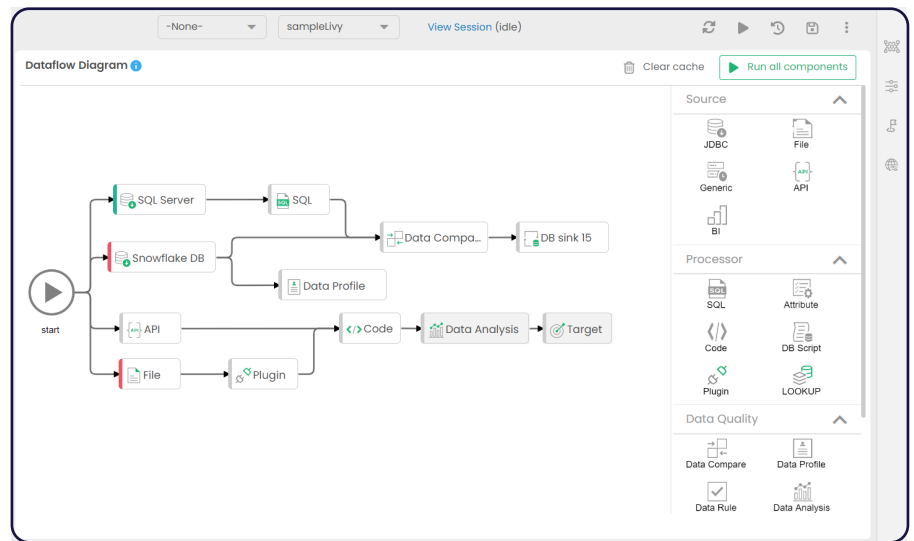


"Datagaps' Dataflow has improved the data quality at every stage in our pipeline and reduced the downtime at migrations, data structure updates and daily operations" - Data Architect at NYU

Dataflow, a module within DataOps Suite to efficiently handle and validate Big Data ETL/ELT Pipelines. Built on patented architecture, it leverages Apache Spark as the underlying engine. Data Engineers, Tests, Analysts can easily and quickly be deployed into data intensive operations. With in-product as well as integrated reporting and notification systems, management can save tremendous amount operational and overhead costs.



60% Reduction in Migration Testing Time

35% Reduction in Migration Time

40% Reduction in Data Quality Testing Time

30% Reduction in Total Cost of Operation

Popular Supported Sources and Integration





Feature List

- Zero Code Deployment
- Container-Based Security
- Enterprise Collaboration
- SQL, Python, Scala and Plugin Support
- Automated Reporting System
- Data Compare Node
- Metrics Comparison Node
- Metadata Comparison and Validation Nodes
- Data Migration Wizard
- Scheduling Capabilities
- API, Flat File and BI Connector
- Email and Slack Notification
- CI/CD Enablement
- Data Profiling Node
- Data Rules Node
- Data Pipelines and Conditional Processing

Data Compare

The Data Compare Node is an easy to deploy dataset / transformation comparison node that is setup in less than 5 clicks. With column-wise mismatches and thresholding parameters, every use-case of comparison can be validated easily.

The suite comes with a Migration Wizard Tool that automates the creation of comparison test cases for easy migration testing.

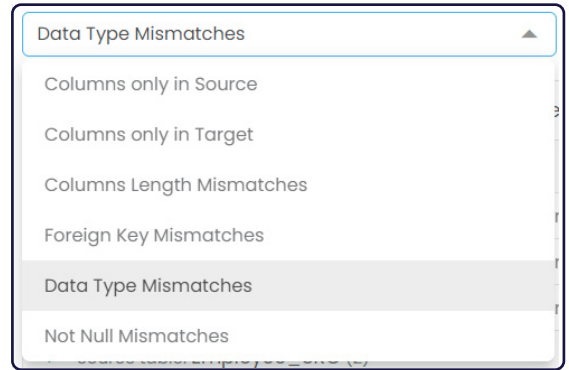
The screenshot displays the Data Compare tool interface. At the top, it shows 'Dataset A : 18484 Records (SQL)' and 'Dataset B : 18489 Records (Attribute)'. The 'Run' button is highlighted. Below the datasets, there are two donut charts: one for 'SQL' and one for 'Attribute'. The 'SQL' chart shows 18479 Matched (99.9%) and 5 Mismatched (0.02%). The 'Attribute' chart shows 18479 Matched (99.9%), 5 Mismatched (0.02%), and 5 Only In B (0.02704%). The 'Status' is 'Failed'. Below the charts, there are tabs for 'Matched Rows', 'Mismatched Rows', 'Column-wise mismatches', 'Only In A', 'Only In B', 'Duplicates In A', 'Duplicates In B', and 'Stkrows to Display' (set to 50). The 'Column-wise mismatches' tab is active, showing a table with columns 'Column Names', 'CustomerKey', and 'EmailAddress'. The table lists two rows of mismatches for the 'EmailAddress' column.

Column Names	CustomerKey	EmailAddress
EmailAddress 2	11000	A : jon24@adventure-works.com B : jon24@adventure-works.org
CommuteDistance 1	11033	A : jaime41@adventure-works.com B : jaime41@adventure-works.edu



Easy Metadata Compare

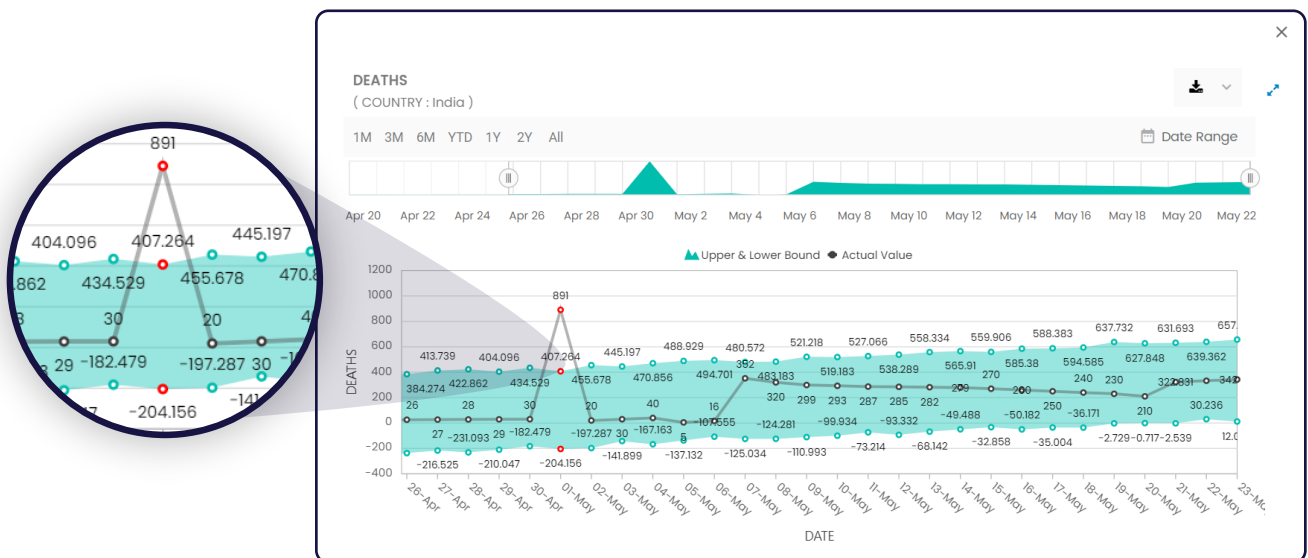
With a single node, 2 compete schemas' metadata can be compared instantly with detailed reporting on lengths, type, foreign key mismatch, non-null and exclusive column difference.



Source Schema ...	Target Schema ...	Target Table Na...	Source Column ...	Target Column N...	Source Data Type	Target Data Type
▼ Source table: Customer_SRC (3)						
dbo	dbo	Customer_SRC	YearlyIncome	YearlyIncome	float	money
dbo	dbo	Customer_SRC	NumberCarsOwned	NumberCarsOwned	decimal	tinyint
dbo	dbo	Customer_SRC	Name	Name	nvarchar	varchar
▼ Source table: Employee_SRC (2)						
dbo	dbo	Employee_SRC	Title	Title	varchar	nvarchar
dbo	dbo	Employee_SRC	FirstName	FirstName	varchar	nvarchar

Anomaly Detection

Using ML and Prediction techniques, metrics can be tracked, and users can be notified in case of statistically significant anomalies are found the dataset.





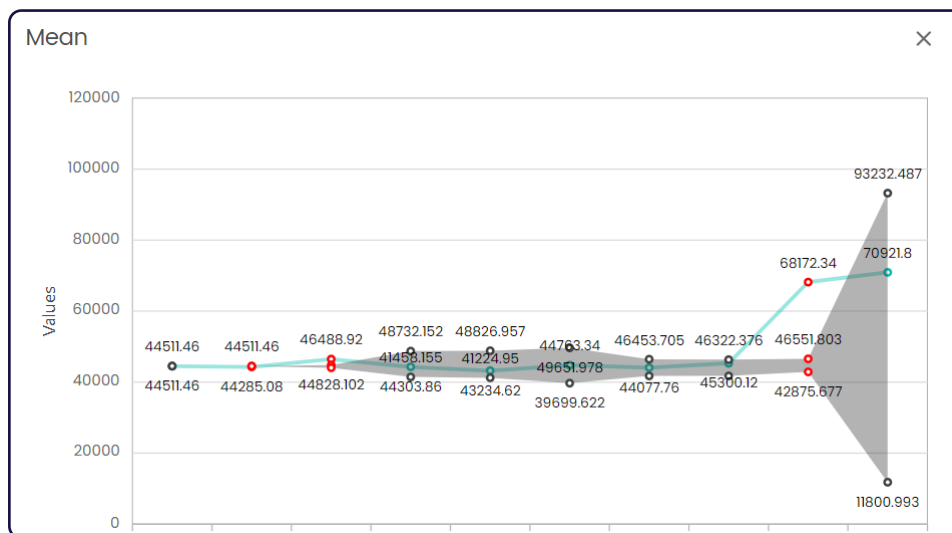
Profiling

Within Dataflows, a Profiling node can be applied to a dataset to track various aggregates, frequency and patterns within the data.

The aggregations help point out anomalies such as sharp changes in mean, lengths, min-max, decimal points and such. Frequency Analysis is used to keep track of the nature of distribution in the datasets. Pattern analysis keep track of the string patterns and # of digits in a particular dataset.

Aggregate	Profile Value	Average (from previous batch runs)	Trend
Row Count	50	50	
Null(%)	0.0	0	N/A
Null count	0	0	N/A
Distinct(%)	100.0	100	
Distinct count	50	50	
Constancy	0.02	0.02	
Min Value	N/A	N/A	N/A

Frequency Analysis		Pattern Analysis	
Value	Count	Pattern	Count
AB89294	1 (2%)	xx99999	50 (100%)
AB64352	1 (2%)		



Request a Demo