bluemetrix
the data control company

BDM Validation Module

Validate the Consistency & Integrity of your data as it moves within your Data Lake

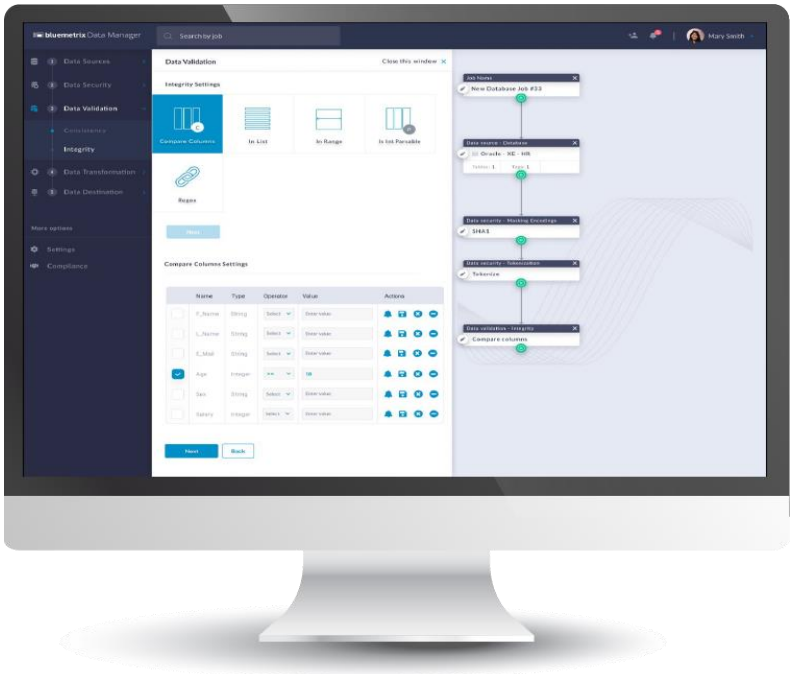
PRODUCT DESCRIPTION

BDM in its simplest mode moves data from a source to a destination, and as part of that movement process there is a module called BDM Data Validation which checks the accuracy and quality of all data as it is moved. It carries out two distinct checks on the data:

Consistency – checks the data was not corrupted in transit
Integrity – checks the value of data being moved is within expected values

These checks are carried out on the data as it is being moved between source and destination i.e. when it is in memory in the Spark cluster. This methodology is secure and optimal as it ensures that replicas of the data are not written to disk.

There is a base library of standard validations that can be applied to the different checks required i.e. Consistency (Checksum, null value, etc.) and Integrity (Numerical operators, Range Values, Limits, Regex, etc.). This library is extensible and new validation checks can be written and added as required.



KEY FEATURES

The areas of BDM that validation focus on:

- **Data Completeness**
- **Data Consistency**
- **Data Quality**
- **Data Integrity**
- **Ease of use**

KEY BENEFITS

- **Data in Transit-** Corruption of Data in transit is detected by applying Consistency checks (checksums, etc.) on the data
- **Security-** All validation checks happen in memory – there are no data copies left on disk
- **Data Integrity-** Data can be validated at an individual record level as it is being processed
- **Extensibility-** All validation algorithms are extensible and can be developed to suit the underlying data
- **Smart dashboards -** All quality data is accessible through a dashboard which will provide a snapshot of the health of the data on the cluster
- **Fast Analytics-** Increased productivity of data scientists as less time spent trying to fix data
- **Governance & Compliance-** All validation checks are recorded automatically in Atlas
- **Ease of Use** - Validation checks can be created and deployed in minutes

BUSINESS CHALLENGE

Gartner estimates that the cost of poor-quality data to organizations is in the region of US$15 million per annum. They believe this cost does not stop at the bottom line, it undermines their digital initiatives and weakens their competitive standing and leads to customer distrust.

Given the growth in data that companies are seeing, which will continue to grow exponentially as 5G and IoT Data becomes more mainstream – validating the quality of data has to be seen as a process and not an end. As data changes rapidly the definition of what is accurate may change from day to day (or hour to hour), and you need a system which will capture these changes and identify the changes that are significant for your business.

Manual checking and spot checking of data will not deliver the accuracy and the timeliness of the results that your business requires. Data validation must be automated and part of your overall data preparation process, so that you can ensure the validations are carried out when required and are as up to date as your data itself.

"Poor quality data weakens an organization's competitive standing and undermines critical business objectives. Poor data quality is also hitting organizations where it hurts – to the tune of $15 million as the average annual financial cost in 2017"

Gartner's Data Quality Market Survey

BDM SOLUTION

The BDM Data Validation module enables you to record, measure and apply data validation checks, as it moves within your data lake. Whenever data is moved into a data lake, outside of a data lake, or within a data lake. This movement is carried out through a pipeline that is created in BDM and the data can be validated at all steps of the way.

As data is validated in memory with the data moving from a source to a destination, these measurements are recorded and stored at each stage. Depending on the results of the validation check, it is possible to

- Remove records that fail the validation and continue processing the source data
- Continue processing the data with failed records, but move a copy of the record into a separate error log file
- Fail the job
- Send an alert The BDM Validation module is flexible and allows you choose and apply validations to suit your organization's needs.

TECHNICAL REQUIREMENTS

Spark V 1.x or V 2.x
Atlas V 0.8 or above

DATA VALIDATION MODULE DEALS WITH:

Data at Rest
Databases (Teradata, Oracle, DB", etc.)
Files (JSON, CSN, XML, etc.)

Data in motion
Streaming with Kafka
Spark Structured Streaming

FOR MORE INFORMATION.
To learn more about BDM and the Data Validation module
Please visit www.bluemetrix.com
| Europe +353 21 4212223 | info@bluemetrix.com