WHITE PAPER

FirstEigen

New Paradigm in Big Data Quality Testing- Self Learning Algorithms

Summary

Gartner reports that 40% of data initiatives fail due to poor quality of data and affects overall labor productivity by ~20% [1]. When multi-dimensional data flows at a high volume and velocity, in different formats, from multiple sources and through multiple platforms, measuring data quality using conventional approaches is a nightmare. The conventional data validation tools and scripting approaches are architecturally limited and unable to handle massive scale of Big Data volume and meet processing speed requirements. They are also ill equipped to handle the sheer number of variables that need to be quality tested. A new paradigm is needed to handle data quality of Big Data.

This whitepaper outlines the next evolution in ensuring data quality of Big Data. FirstEigen has implemented this advanced paradigm in its data quality solution, DataBuck. The Big Data Quality solution must at the minimum satisfy these:

- ✓ The underlying engine must have the horsepower to handle massive data volumes, which even the Big Data edition of major ETL's do not
- ✓ Users must be able to set up data quality tests with the fewest clicks (need only 3 clicks out-of-the-box with DataBuck)
- ✓ Manually setting rules for 100's of Tables with 1000's of columns is unrealistic. The gamut of data quality rules specific to the dataset must be autonomously learned using self-learning algorithms. DataBuck is the first tool designed for this approach
- ✓ Results of quality indicators must be translatable into relevant metrics for different stakeholders, like Executives, Team Leaders and Data Quality Owners

DataBuck

Data Quality Testing and Monitoring Framework

The most common challenges to data quality in big data applications are ensuring that incremental data flowing in is:

- ✓ Reasonable: Incoming dataset volume and measurement are reasonable and consistent with its expected fingerprint
- ✓ Non-Duplicate: Data does not have duplicate records
- ✓ Complete: Expected non-null columns do not have null values
- ✓ Accurate: All string columns and numerical fields have values consistent with its expected fingerprint
- ✓ Consistent: Dataset does not have anomalous records

In order to achieve the above objectives in a systematic way, DataBuck measures six key data quality indicators to assess the incremental data quality of the all tables or selected group of tables in big data applications. If any of the indicators are outside of expectations an alert would be sent automatically. Each variable or column has specific expectations (or its fingerprint) which is self-learned by the tool autonomously.

Indicator Name	Type	Definition
Dataset Reasonability Indicator	Reasonability	Aggregate macro-level measures are tracked, like record count and many other variables, and compared against its expected fingerprint
Record Anomaly Detector	Anomaly	Anomalies identified based on either the current dataset fingerprint or expected fingerprint
Primary Keys Duplicate Indicator	Duplicates	Identify duplicates based on the primary keys
All Keys Duplicate Indicator	Duplicates	Identify duplicates based on all the fields present or user defined fields
Null Count Indicator	Completeness	Compare the number of null values of each non-null fields and/or user defined fields with its expected fingerprint
String Column Quality Indicator	Quality	Compare the string type columns with its expected fingerprint
Numerical Column Quality Indicator	Quality	Compare the Numerical type columns with its expected fingerprint. Deviations are an indicator of potential quality failure

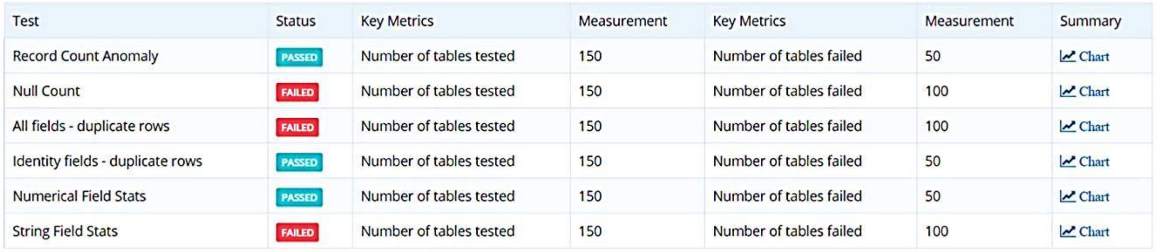
To ensure the quality of the big data applications, organizations need to adopt a two-prong testing and monitoring strategy as defined below:

Initial Data testing: Prior to initiating the ongoing data quality indicators, organizations should measure the data quality indicators for all the existing data. This will help establish the baseline for the six data quality indicators discussed above. In addition, this will help identify potential quality issues of the existing data set.
Continuous Monitoring: Once the data quality baseline is established, key indicators should be measured and monitored for all subsequent data loads on an ongoing basis to identify issues and understand key trends.

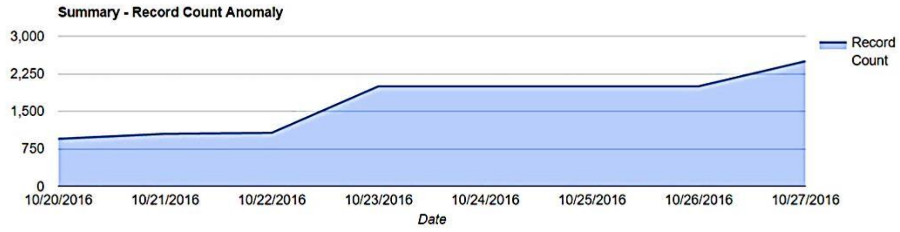
Data Quality Dashboard Framework

1. *Data Pipeline Level*

Data quality has to be tracked at multiple levels. The highest instance being at the data pipeline or schema-level.
Exhibit-1: Data pipeline level view of errors



Test	Status	Key Metrics	Measurement	Key Metrics	Measurement	Summary
Record Count Anomaly	PASSED	Number of tables tested	150	Number of tables failed	50	Chart
Null Count	FAILED	Number of tables tested	150	Number of tables failed	100	Chart
All fields - duplicate rows	FAILED	Number of tables tested	150	Number of tables failed	100	Chart
Identity fields - duplicate rows	PASSED	Number of tables tested	150	Number of tables failed	50	Chart
Numerical Field Stats	PASSED	Number of tables tested	150	Number of tables failed	50	Chart
String Field Stats	FAILED	Number of tables tested	150	Number of tables failed	100	Chart



2. *Input Source Level*

The errors then have to be drilled down to the Table-level. Which of the hundreds of incoming Tables have errors in them and of what type?

Exhibit-2: Input source level view of errors

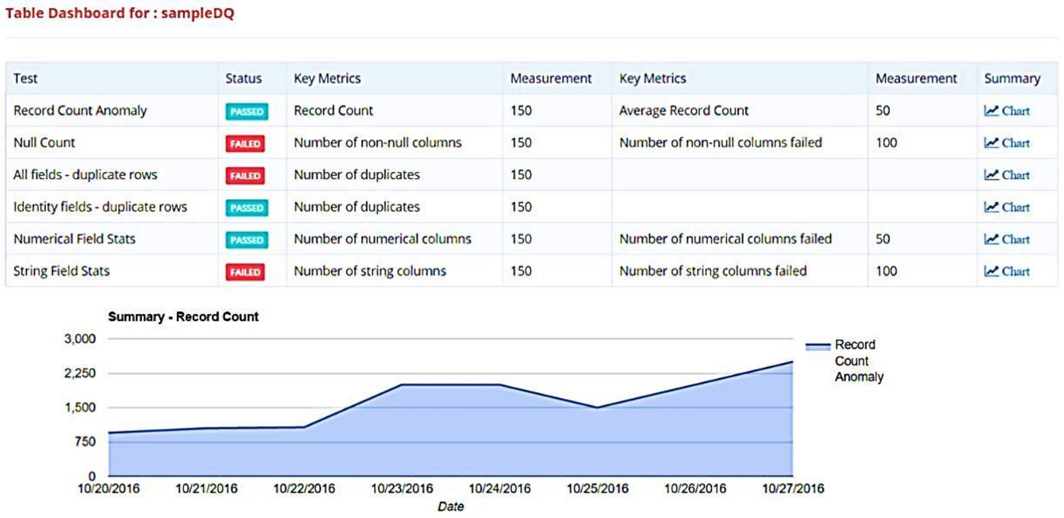


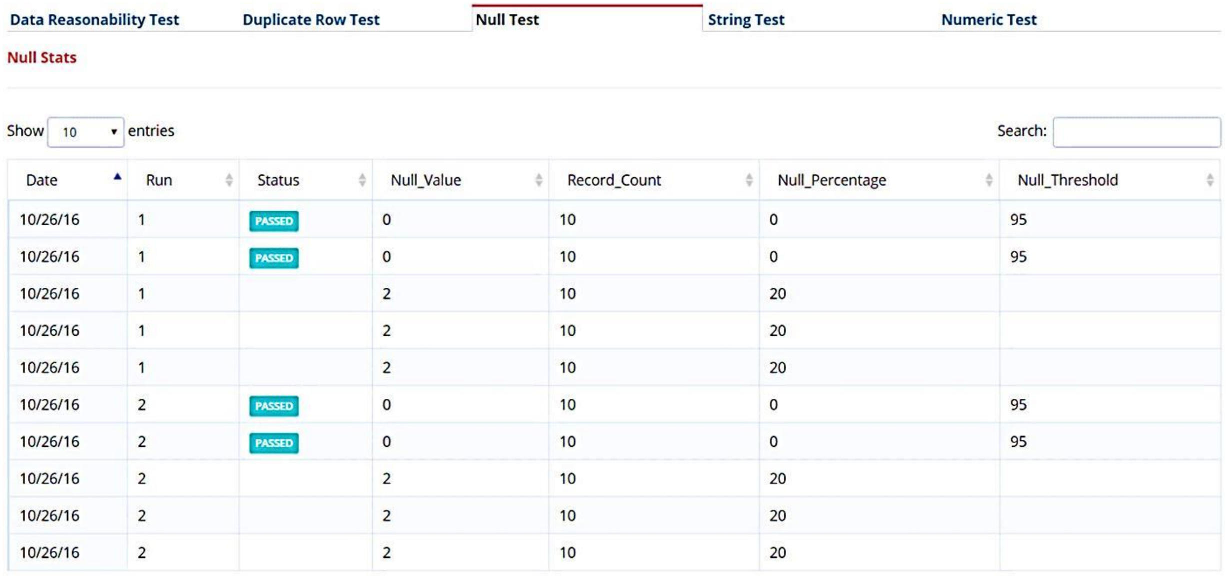
Table Dashboard for : sampleDQ

Test	Status	Key Metrics	Measurement	Key Metrics	Measurement	Summary
Record Count Anomaly	PASSED	Record Count	150	Average Record Count	50	Chart
Null Count	FAILED	Number of non-null columns	150	Number of non-null columns failed	100	Chart
All fields - duplicate rows	FAILED	Number of duplicates	150			Chart
Identity fields - duplicate rows	PASSED	Number of duplicates	150			Chart
Numerical Field Stats	PASSED	Number of numerical columns	150	Number of numerical columns failed	50	Chart
String Field Stats	FAILED	Number of string columns	150	Number of string columns failed	100	Chart

3. *Input Column Level*

In order for the Data Quality Engineers to be able to fix the issues they need to be able to drill further down to the exact column which has errors. This will also help them understand the source of errors, for a permanent fix.

Exhibit-3: Input column level view of errors



Date	Run	Status	Null_Value	Record_Count	Null_Percentage	Null_Threshold
10/26/16	1	PASSED	0	10	0	95
10/26/16	1	PASSED	0	10	0	95
10/26/16	1		2	10	20	
10/26/16	1		2	10	20	
10/26/16	1		2	10	20	
10/26/16	2	PASSED	0	10	0	95
10/26/16	2	PASSED	0	10	0	95
10/26/16	2		2	10	20	
10/26/16	2		2	10	20	
10/26/16	2		2	10	20	

Key Features of an Exceptional Big Data Quality Solution

1. *Ease of Use- Minimal manual intervention*

In a typical big data application, number of tables and the columns can be onerous. Big Data pipes often carry 100's of Tables, with 1000's of columns per Table. Analyzing individual columns and coding individualized rules for those 100,000's of columns is overwhelming, non-trackable over time and can easily introduce errors. In many big data scenarios, users often do not have quality requirements/expectations of incoming data – which makes the rules writing for data quality even more cumbersome. A well-designed data quality tool should auto discover the tables, the columns, the applicable rules and define the quality indicators without significant involvement from the users.

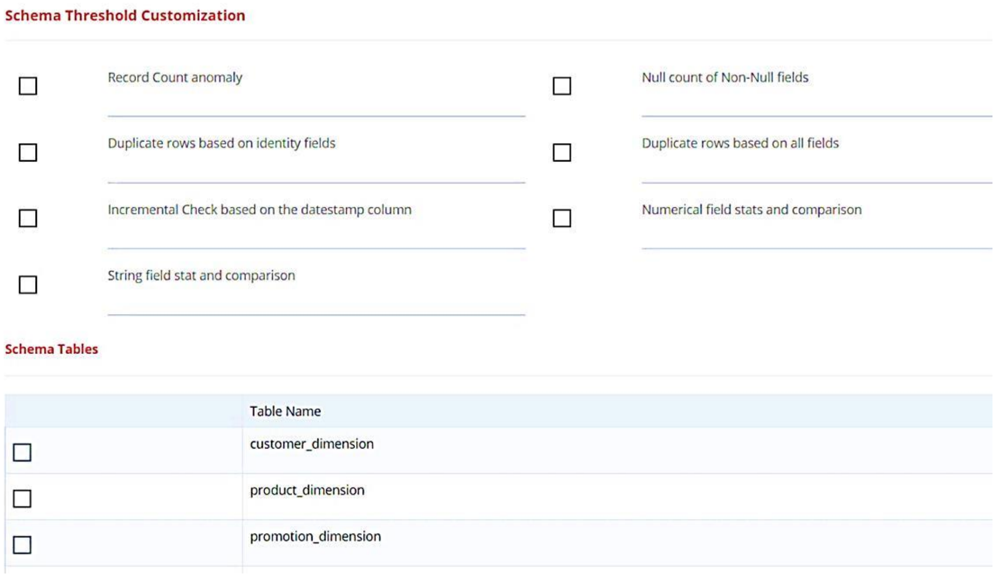
2. *Ease of Customization*

Invariably, some tables and columns will be more significant than others for a user. They should be able to apply additional data quality rules for individual tables wherever they want to override the system generated rules.

3. *User Notification*

With 100's of tables and 1,000's of columns, it will be very difficult to monitor the results of the data quality tests. User should be notified of the high priority exceptions, as chosen by them, via email alerts.

Exhibit-4: Big Data quality tool with autonomous rule-discovery, which can also be customized as needed at Schema level and the Table level



Schema Threshold Customization

□ Record Count anomaly

□ Duplicate rows based on identity fields

□ Incremental Check based on the datestamp column

□ String field stat and comparison

□ Null count of Non-Null fields

□ Duplicate rows based on all fields

□ Numerical field stats and comparison

Schema Tables

	Table Name
□	customer_dimension
□	product_dimension
□	promotion_dimension

4. *Results Visualization*

A well-designed solution should provide role-based dashboard that has relevant information for the executive and the data analyst. It should give quick overview of quality status and areas with most quality issues. The data should ideally be portable to other commercial visualization tools as well.

5. ***Speed of Processing***

   The underlying technology has to be scalable to handle large volumes of data. It's not a nice-to-have, it's a must. Nearly all traditional tools used today experience a "time out" when used for even modest data volumes; because they are built for a time when data volumes, velocity or variety were just regular. The typical tools, including the latest Big Data Edition of major ETL's, can now ingest data in the Big Data format (HDFS) but still the underlying processing engine is at least 10x slower than a tool built on a Big Data Spark platform, like DataBuck is. This was shown in a recently published benchmarking study[2].

6. ***Ability to Isolate Bad Data***

   The data quality tool should isolate bad data so the user can track the root causes and apply mass corrections to the data.

7. ***Auditability***

   When the data is needed for regulatory compliance, data quality tool should allow easy auditing of data quality tests that were applied on the data.

8. ***Referential Integrity Indicators***

   This last key requirement is the ability to detect orphan records based on primary-foreign key relationship. As complexity increases due to larger data volume and myriad of inter-relationships between them, lot of unwanted data gets left behind, unpurged. Keep constantly pruning it, else end up like a library (imagine a large one, like the Library of Congress) with misplaces books and books removed from the shelves but never put back. That would be a recipe for a junk yard very soon.

A tool with these key features will be a good start in the right direction to get a handle on your Big Data. Fix these upfront, else you risk losing the entire Big Data initiative as trust erodes away in the data and your team is working overtime to fix the ever snowballing errors.

References:

[1] Measuring the Business Value of Data Quality, Ted Friedman, Michael Smith, Gartner Inc., 2011

[2] Hadoop Integration Benchmark- Product Profile and Evaluation: Talend and Informatica, By William McKnight and Jake Dolezal, October 2015, MCG Global Services; http://www.talend.com/blog/2015/12/15/when-it-comes-to-big-data-%E2%80%93-speed-matters