DATACTICS™

WHITEPAPER

Beyond Data Prep: Self-Service Data Quality

The drivers and benefits of a holistic, self-service data quality platform

In this paper we describe trends and technologies bringing data quality functions closer to the data and moving responsibility and control from central IT functions to data stewards and SMEs, thereby achieving greater operational efficiency and higher value data assets.

Published on 20th April 2020

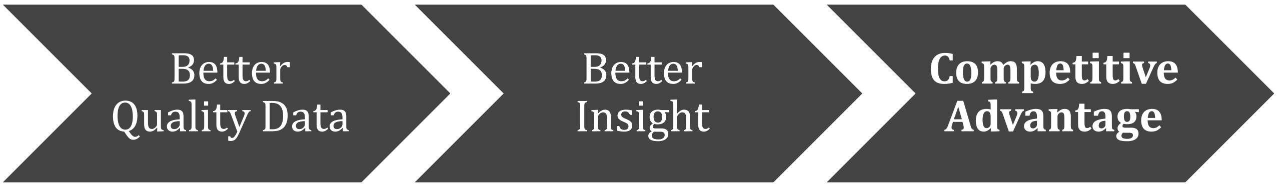
Contents

1. The Changing Landscape of Data Quality

Change

There has been increasing demand for higher and higher data quality in recent years – highly regulated sectors, such as banking have had a tsunami of financial regulations such as BCBS239, MiFID, FATCA and many more stipulating or implying exacting standards for data and data processes. Meanwhile there is a growing trend for more and more firms to become more Data and Analytics (D&A) driven, taking inspiration from Google & Facebook, to monetize their data assets. This increased focus on D&A has been accelerated by easier and lower cost access to artificial intelligence (AI), machine learning (ML) and business intelligence (BI) visualization technologies. However, in the now-waning hype of these technologies comes the pragmatic realization that unless there is a foundation of good quality reliable data, insights derived from AI and analytics may not be actionable. With AI and ML becoming more of a commodity, and a level playing field, the differentiator is in the data and the quality of the data.

> ***"unless there is a foundation of good quality reliable data, insights derived from AI and analytics may not be actionable"***



Problems

As the urgency for regulatory compliance or competitive advantage escalates, so too does the urgency for high data quality. A significant obstacle to quickly achieve high data quality is the variety of disciplines required to measure data quality, enrich data and fix data. By its nature, digital data, especially big data can require significant technical skills to manipulate and for this reason, was once the sole responsibility of IT functions within an organization. However, maintaining data also requires significant domain knowledge about the content of the data, and this domain knowledge resides with the subject matter experts (SMEs) who use the data, rather than a central IT function. Furthermore, each data set will have its own SMEs with special domain knowledge required to maintain the data, and a rapidly-growing and changing number of data sets. If a central IT department is to maintain quality of data correctly it must therefore liaise with an increasingly large number of data owners and SMEs in order to correctly implement DQ controls and remediation required. These demands create a huge drain on IT resources, and a slow-moving backlog of data quality change requests within IT that simply can’t keep up.

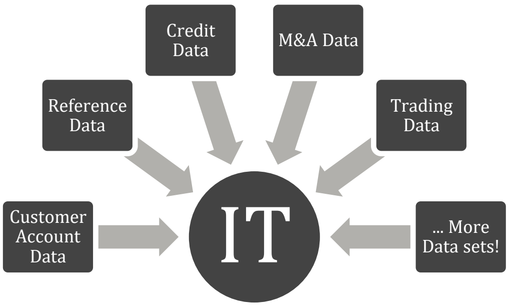


Figure 1 Growing data demands put pressure on IT resources

Given the explosion in data volumes this model clearly won’t scale and so there is now a growing trend to move data quality operations away from central IT and back into the hands of data owners. While this move can greatly accelerate data quality and data on-boarding processes, it can be difficult and expensive for data owners and SMEs to meet the technical challenges of maintaining and on-boarding data. Furthermore, unless there is common governance around data quality across all data domains there stands the risk of a ‘wild west’ scenario, where every department manages data quality differently with different processes and technology.

> ***“there is now a growing trend to move data quality operations away from central IT and back into the hands of data owners”***

Opportunity

The application of data governance policies and creation of an accountable Chief Data Officer (CDO) goes a long way to mitigate against the ‘wild west’ scenario. Data quality standards such as the Enterprise Data Management Council’s (EDMC) Data Capability Assessment Model (DCAM)[1] provide opportunities to establish consistency in data quality measurement across the board.

The drive to capitalize on data assets for competitive advantage has had the result that the CDO function is quickly moving from an operational cost centre towards a product-centric profit centre. A recent publication by Gartner (30th July 2019)[2] describes three generations of CDO: “CDO 1.0” focused on data management; “CDO 2.0” embraced analytics; “CDO 3.0” assisted digital transformation; and Gartner now predicts a fourth, “CDO 4.0” focused on monetizing data oriented products. Gartner’s research suggests that to enable this evolution, companies should strive to develop data and analytics platforms that scale across the entire company and this implies data quality platforms that scale too.

2. Key Features a Self-Service DQ Platform Should Have

To enable the evolution towards actionable insight from data, D&A platforms and processes must evolve too. At the core of this evolution is the establishment of '**self-service**' data quality – whereby data owners and SMEs have ready access to robust tools and processes, to measure and maintain data quality themselves, in accordance with data governance policies.

From a business perspective such a self-service data quality platform must be:

> ***"self-service data quality empowers data owners and SMEs to measure and maintain data quality themselves in accordance with governance policies"***

- Powerful enough to enable business users and SMEs to perform complex data operations without highly skilled technical assistance from IT
- Transparent, accountable and consistent enough to comply with firm wide data governance policies
- Agile enough to quickly onboard new data sets and changing data quality demands of end consumers such as AI and Machine learning algorithms
- Flexible and open so it integrates easily with existing data infrastructure investment without requiring changes to architecture or strategy
- Advanced to make pragmatic use of AI and machine learning to minimize manual intervention

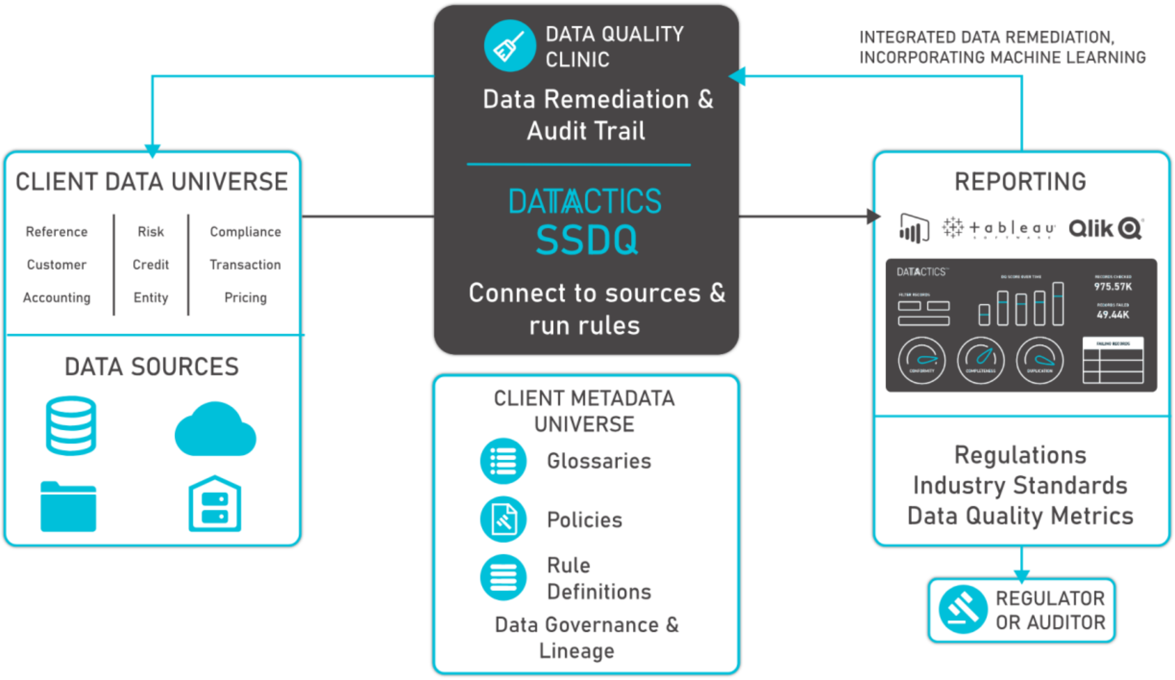


Figure 2 Self-Service Data Quality Platform integrations, outputs and remediation loop.

This goes way beyond the scope of most stand-alone data prep tools and ‘home grown’ solutions that are often used as a tactical one-off measure for a particular data problem. Furthermore, for the self-service data quality platform to truly enable actionable data across the enterprise, it will need to provide some key technical functionality *built-in*:

- ***Transparent & Continuous Data Quality Measurement***
  Not only should it be easy for business users and SMEs to implement large numbers of data domain specific data quality rules, but also those rules should be simple to audit, and easily explainable, so that ‘DQ breaks’ can be easily explored and the root cause of the break established.

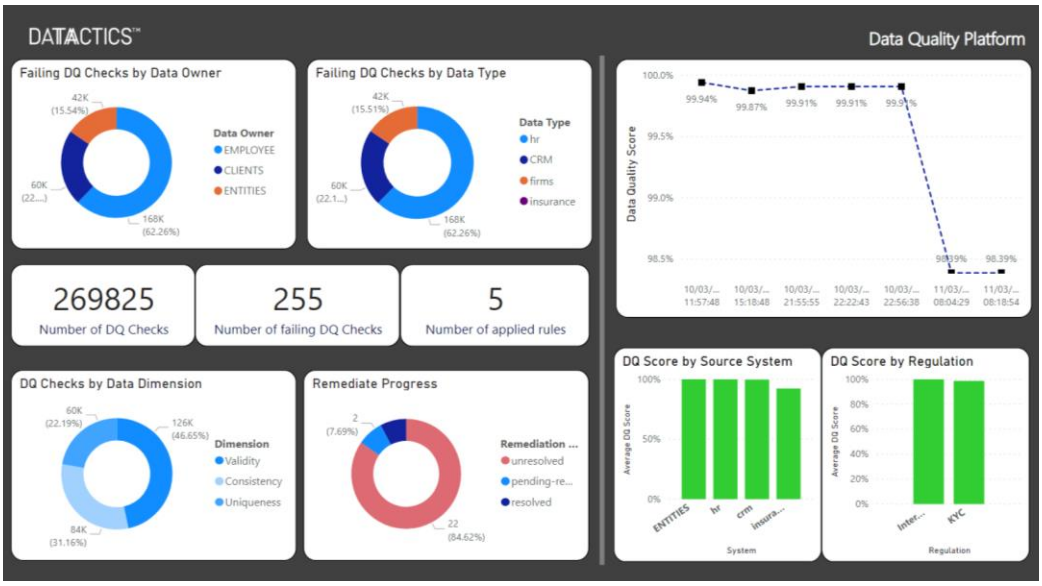
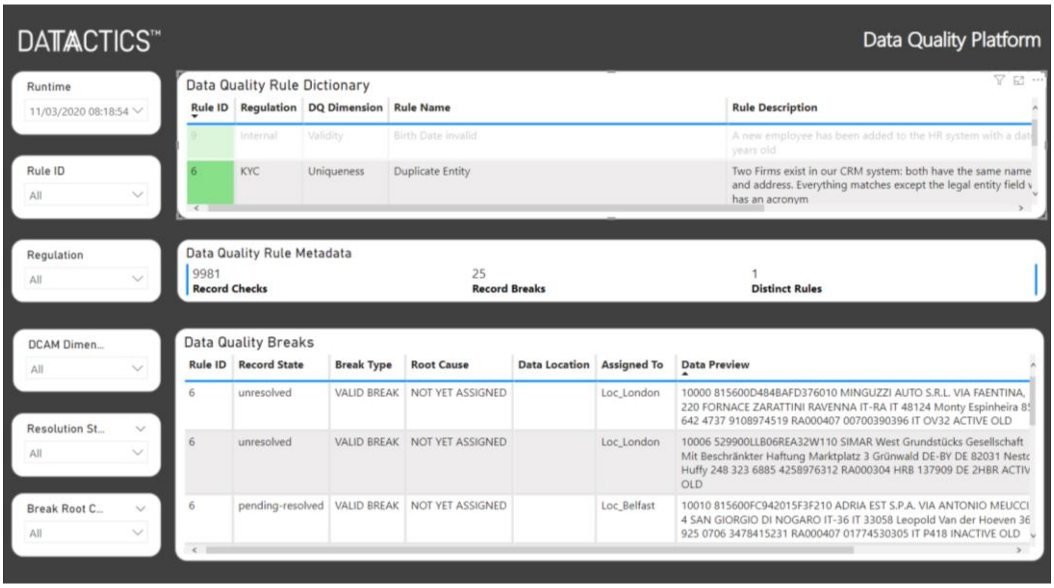


Figure 3

Example DQ Dashboard, built in PowerBI, taken from Datactics SSDQ Platform

In addition to data around the actual breaks, a DQ platform should be able to produce DQ dashboards enabling drill-down from high level statistics down to actual failing data points and publish high level statistics into data governance systems.



Runtime
11/03/2020 08:18:54

Rule ID
All

Regulation
All

DCAM Dimen...
All

Resolution St...
All

Break Root C...
All

Data Quality Rule Dictionary

Rule ID	Regulation	DQ Dimension	Rule Name	Rule Description
9	Internal	Validity	Birth Date invalid	A new employee has been added to the HR system with a dat years old
6	KYC	Uniqueness	Duplicate Entity	Two Firms exist in our CRM system: both have the same name and address. Everything matches except the legal entity field v has an acronym

Data Quality Rule Metadata

9981 **Record Checks**	25 **Record Breaks**	1 **Distinct Rules**

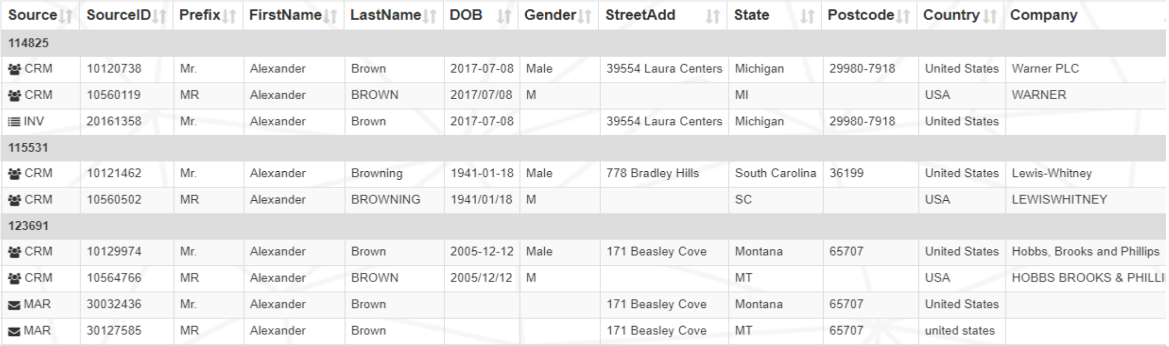
Data Quality Breaks

Rule ID	Record State	Break Type	Root Cause	Data Location	Assigned To	Data Preview
6	unresolved	VALID BREAK	NOT YET ASSIGNED		Loc_London	10000 815600D484BAFD376010 MINGUZZI AUTO S.R.L. VIA FAENTINA, 220 FORNACE ZARATTINI RAVENNA IT-RA IT 48124 Monty Espinheira 8! 642 4737 9108974519 RA000407 00700390396 IT OV32 ACTIVE OLD
6	unresolved	VALID BREAK	NOT YET ASSIGNED		Loc_London	10006 529900LLB06REA32W110 SIMAR West Grundstücks Gesellschaft Mit Beschränkter Haftung Marktplatz 3 Grünwald DE-BY DE 82031 Nestc Huffy 248 323 6885 4258976312 RA000304 HRB 137909 DE 2HBR ACTIV OLD
6	pending-resolved	VALID BREAK	NOT YET ASSIGNED		Loc_Belfast	10010 815600FC942015F3F210 ADRIA EST S.P.A. VIA ANTONIO MEUCCI 4 SAN GIORGIO DI NOGARO IT-36 IT 33058 Leopold Van der Hoeven 36 925 0706 3478415231 RA000407 01774530305 IT P418 INACTIVE OLD

Figure 4

Detail portion of data breaks

- ***Powerful Data Matching - Entity Resolution for Single View and Data Enrichment***
  Finding hidden value in data or complying with regulation very often involves joining together several disparate data sets. For example, enhancing a Legal Entity Master Database with an LEI, screening customer accounts against sanctions and PEP lists for KYC, creating a single view of client from multiple data silos for GDPR or FSCS compliance. This goes further than simple deduplication of records or SQL joins – most data sets are messy and don't have unique identifiers and so fuzzy matching of numerous string fields must be implemented to join one data set with another. Furthermore, efficient clustering algorithms are required to sniff out similar records from other disparate data sets in order to provide a single consolidated view across all silos.

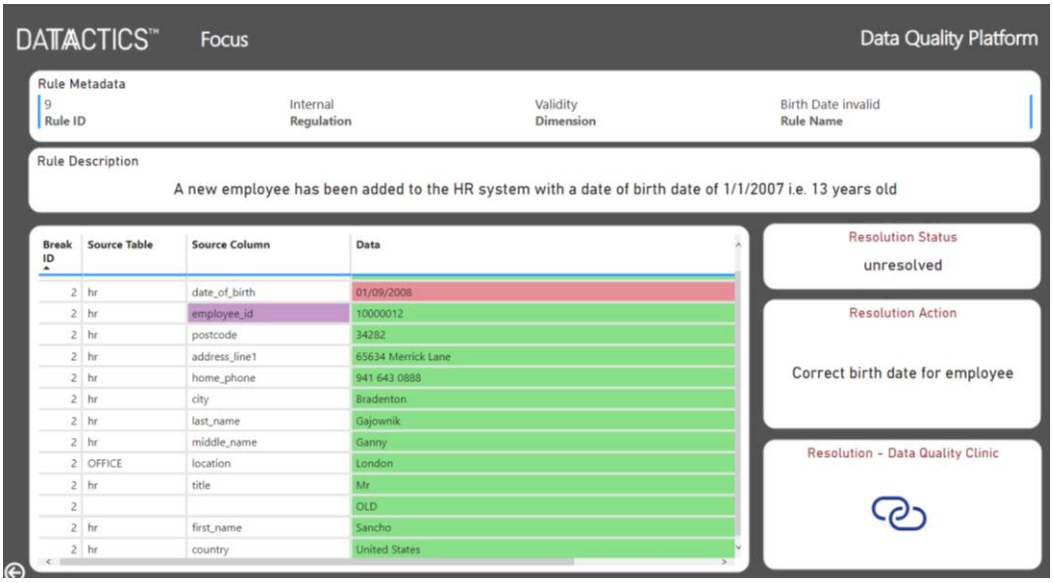


Source	SourceID	Prefix	FirstName	LastName	DOB	Gender	StreetAdd	State	Postcode	Country	Company
114825											
CRM	10120738	Mr.	Alexander	Brown	2017-07-08	Male	39554 Laura Centers	Michigan	29980-7918	United States	Warner PLC
CRM	10560119	MR	Alexander	BROWN	2017/07/08	M		MI		USA	WARNER
INV	20161358	Mr.	Alexander	Brown	2017-07-08		39554 Laura Centers	Michigan	29980-7918	United States	
115531											
CRM	10121462	Mr.	Alexander	Browning	1941-01-18	Male	778 Bradley Hills	South Carolina	36199	United States	Lewis-Whitney
CRM	10560502	MR	Alexander	BROWNING	1941/01/18	M		SC		USA	LEWISWHITNEY
123691											
CRM	10129974	Mr.	Alexander	Brown	2005-12-12	Male	171 Beasley Cove	Montana	65707	United States	Hobbs, Brooks and Phillips
CRM	10564766	MR	Alexander	BROWN	2005/12/12	M		MT		USA	HOBBS BROOKS & PHILLI
MAR	30032436	Mr.	Alexander	Brown			171 Beasley Cove	Montana	65707	United States	
MAR	30127585	MR	Alexander	Brown			171 Beasley Cove	MT	65707	united states	

Figure 5 Clustering algorithms find similar data across multiple data sets

- ***Integrated Data Remediation Incorporating Machine Learning***
  It's not enough just to flag up broken data, you also need a process and technology for fixing the breaks. Data quality platforms should have this built in so that after data quality measurement, broken data can be quarantined, data owners alerted and breaks automatically assigned to the relevant SMEs for remediation. Interestingly, the manual remediation process lends itself very well to machine learning. The process of manually remediating data captures domain specific knowledge about the data – information that can be readily used by machine learning algorithms to streamline the resolution of similar breaks in the future and thus greatly reduce the overall time and effort spent on manual remediation.

> ***"The process of manually remediating data captures domain specific knowledge about the data – information that can be readily used by machine learning algorithms to streamline the resolution of similar breaks in the future"***



Rule Metadata

9	Internal	Validity	Birth Date invalid
Rule ID	Regulation	Dimension	Rule Name

Rule Description

A new employee has been added to the HR system with a date of birth date of 1/1/2007 i.e. 13 years old

Break ID	Source Table	Source Column	Data
2	hr	date_of_birth	01/09/2008
2	hr	employee_id	10000012
2	hr	postcode	34282
2	hr	address_line1	65634 Merrick Lane
2	hr	home_phone	941 643 0888
2	hr	city	Bradenton
2	hr	last_name	Gajownik
2	hr	middle_name	Ganny
2	OFFICE	location	London
2	hr	title	Mr
2			OLD
2	hr	first_name	Sancho
2	hr	country	United States

Resolution Status

unresolved

Resolution Action

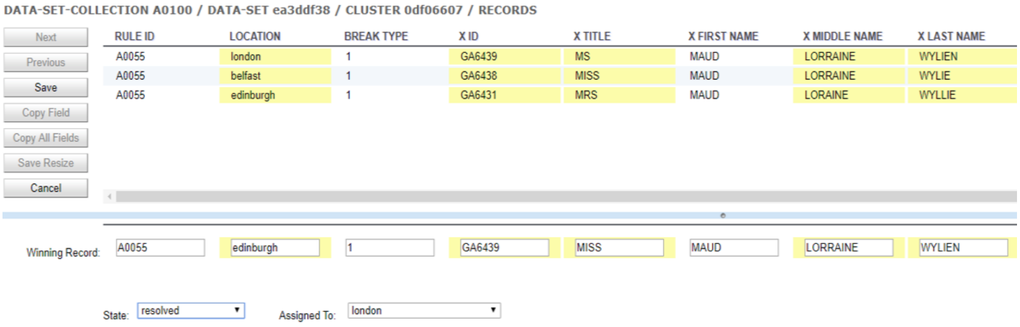
Correct birth date for employee

Resolution - Data Quality Clinic

Figure 6

Specific break highlighted with link through to resolution tool.

- ***Data Access Controls Across Teams and Datasets***
  Almost any medium to large sized organization will have various forms of sensitive data, and policies for sharing that data within the organization e.g. 'Chinese walls' between one department and another. In order to enable integration across teams and disparate silos of data, granular access controls are required – especially inside the data remediation technology where sensitive data may be displayed to users. Data access permissions should be set automatically where possible (e.g. inheriting Active Directory attributes) and enforced when displaying data, for example by row- and field-level access control, and using data masking or obfuscation where appropriate.



DATA-SET-COLLECTION A0100 / DATA-SET ea3ddf38 / CLUSTER 0df06607 / RECORDS

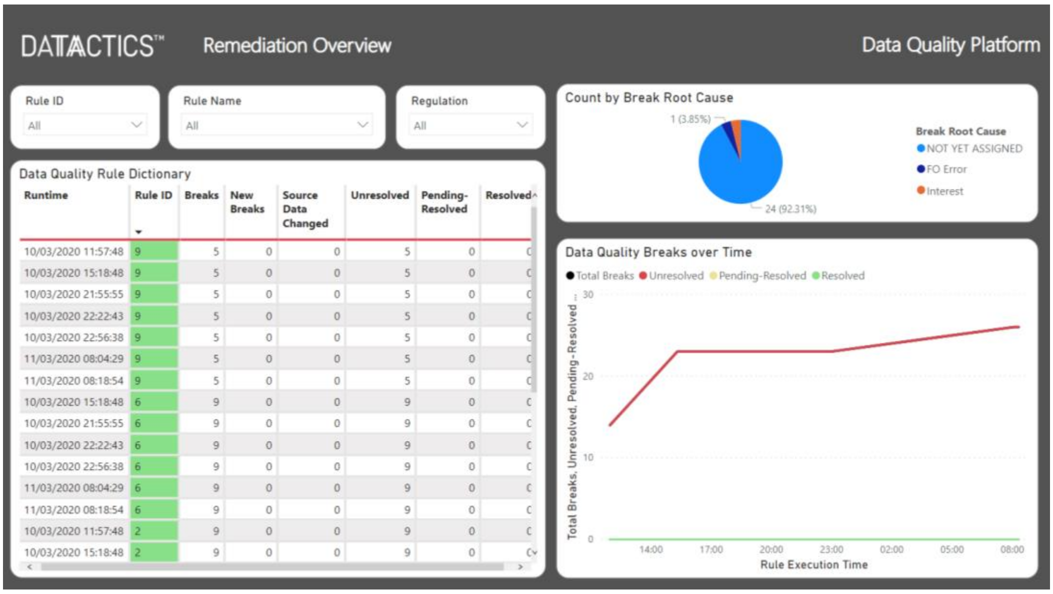
	RULE ID	LOCATION	BREAK TYPE	X ID	X TITLE	X FIRST NAME	X MIDDLE NAME	X LAST NAME
	A0055	london	1	GA6439	MS	MAUD	LORRAINE	WYLIEN
	A0055	belfast	1	GA6438	MISS	MAUD	LORRAINE	WYLIE
	A0055	edinburgh	1	GA6431	MRS	MAUD	LORAINE	WYLLIE
Winning Record:	A0055	edinburgh	1	GA6439	MISS	MAUD	LORRAINE	WYLIEN

State: resolved ▾ Assigned To: london ▾

Figure 7

Example manual remediation screen in self-service data quality

- ***Audit Trails, Assigning and Tracking Performance***
  Providing business users with tools to fix data could cause additional headaches when it comes to being able to understand who did what, when, why and whether or not it was the right thing to do. It stands to reason, therefore, that any remediation tool should have built-in capability to do just that with the associated performance of data break remediation measured, tracked and managed.



Data Quality Rule Dictionary

Runtime	Rule ID	Breaks	New Breaks	Source Data Changed	Unresolved	Pending-Resolved	Resolved
10/03/2020 11:57:48	9	5	0	0	5	0	0
10/03/2020 15:18:48	9	5	0	0	5	0	0
10/03/2020 21:55:55	9	5	0	0	5	0	0
10/03/2020 22:22:43	9	5	0	0	5	0	0
10/03/2020 22:56:38	9	5	0	0	5	0	0
11/03/2020 08:04:29	9	5	0	0	5	0	0
11/03/2020 08:18:54	9	5	0	0	5	0	0
10/03/2020 15:18:48	6	9	0	0	9	0	0
10/03/2020 21:55:55	6	9	0	0	9	0	0
10/03/2020 22:22:43	6	9	0	0	9	0	0
10/03/2020 22:56:38	6	9	0	0	9	0	0
11/03/2020 08:04:29	6	9	0	0	9	0	0
11/03/2020 08:18:54	6	9	0	0	9	0	0
10/03/2020 11:57:48	2	9	0	0	9	0	0
10/03/2020 15:18:48	2	9	0	0	9	0	0

Figure 8

Tracking and measuring DQ break remediation activity

- ***AI Ready***
  There's no doubt that one of the biggest drivers of data quality is AI. AI data scientists can spend up to 80% of their time just preparing input data for machine learning algorithms, which is a huge waste of their expertise. A self-service data quality platform can address many of the data quality issues by providing ready access to tools and processes that can ensure a base level of quality and identify anomalies in data that may skew machine learning models. Furthermore the same self-service data quality tools can assist data scientists to generate metadata that can be used to inform machine learning models – such 'Feature Engineering' can be of real value when the data set is largely textual as it can generate numerical indicators which are more readily consumed by ML algorithms.

> ***"AI data scientists can spend up to 80% of their time just preparing input data for machine learning algorithms, which is a huge waste of their expertise"***

3. Conclusion

The move towards a self-service oriented model for data quality is a logical way to keep up with the expanding volumes and varieties of data. This trend is confirmed in a recent Forrester research[3] report (Oct 31st 2018) which saw a dramatic shift in data quality operations from IT to business lines in the last 2 years. However, data platforms need to be architected carefully to support the self-service model in accordance with data governance to avoid the 'wild west' scenario. Organizations that successfully embrace self-service data quality and implement a self-service data quality platform are more likely to benefit from actionable data, resulting in deeper insight, in more effortless compliance, and in significant competitive advantage.

4. References

1. EDMC's DCAM: https://edmcouncil.org/page/aboutdcamreview
2. Gartner 30th July 2019 – CDO 4.0: https://www.gartner.com/en/newsroom/press-releases/2019-07-30-gartner-research-board-identifies-the-chief-data-officer-4point0
3. Forrester Analytics Global Business Technographics Data And Analytics Survey, 2018 and 2017

5. About Datactics

Datactics helps banks ensure compliance with financial regulations, aids AML & KYC functions, and eliminates roadblocks common in data management. We specialise in class-leading, self-service data quality and fuzzy matching software solutions, designed to empower business users who know the data to visualise and fix the data.

6. General Enquiries:

Website: www.datactics.com

Email address: info@datactics.com
Phone number: +44 (0)28 9023 3900

Datactics Ltd.
1 Lanyon Quay
BT1 3LG
Belfast
+44 (0)28 9023 3900
datactics.com

DATACTICS™