Summary Insights: Data Acquisition and Cleansing with AI

RFG Perspective: Most enterprises are still struggling with implementing automated techniques for data acquisition and cleansing. In fact, the maturity level of in-house data catalogs and dictionaries leaves a lot to be desired. Most companies lack the capability to rapidly acquire and cleanse data within their business unit – and have an even greater challenge identifying data sources in other units within the enterprise and then gaining access and usage of the data. Until Business and IT executives correct this situation, organizations will strain to execute DataOps projects in a timely fashion.

Business Executives base their decisions on the data with which they are presented. Having the right data with good quality is key to ensuring productive business outcomes. But as the volume of data expands from terabytes to petabytes throughout enterprise data domains, there are increasing quality challenges given the velocity, veracity, volume, and variety of data inflows. If these data sources are to be effectively incorporated into DataOps projects designed to support effective business analytics, IT organizations will require automated methods of acquiring and cleansing the data.

The RFG100 Panel's Key Takeaways on Data Cleansing Options

On Sep. 30, 2020 RFG facilitated a video conference on "Data Acquisition and Cleansing with AI." The panelists on the call were:

- Charles Elkan, Managing Director, Goldman Sachs
- Terri Sage, CTO, 1010Data
- Ted Dunning, CTO and Data Fabric Technologist, HPE
- Shawn Butler, Director of Solutions Architecture, Fishtech Group
- Jean Bozman, President, Cloud Architect Advisors

A real-time poll of the RFG100 attendees showed several top takeaways: (See full poll results in the Appendix at the end of this document, below).

- More than 60% of respondents have human-readable data catalogs while less than 40% have machine-readable ones
- More than 45% said they had human-readable data dictionaries; only half that amount had machine-readable ones.
- More than 60% of respondents said they needed to request access to data sources from other divisions or business units while 46% said they needed permission to copy or use the data.
- Less than 25% said they had the ability to validate usability of the data through machine-readable means.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*

We believe the lack of availability of machine-readable catalogs and data dictionaries and the inability to automatically utilize corporate data residing in other business units will greatly impair DataOps efforts – and for many firms make DataOps a non-starter.

This report unpacks the content of the RFG100 discussion, which included input from top IT executives from the banking and financial industries.

The Need for Machine-Readable Data

In today's world speed is an imperative or one can miss new market opportunities. While there are DevOps and DevSecOps teams working to spew out new applications in rapid order, DataOps – the data side equivalent – can only move as fast as the organization's ability to locate, acquire, cleanse, and incorporate the data into the mix. That requires machine readable metadata and data sources – primarily, data catalogs and data dictionaries. Unfortunately, as indicated by the RFG poll results, most institutions do not have that capability.

One participant stated his firm use a software program that describes data in a machine-readable format. Importantly, it allows search engines to access data across business units. This approach to ingesting and tagging data should be standard practice, but unfortunately, it is uncommon.

Business units prefer to protect their data rather than sharing it with other lines of business (LOBs). This creates "data silos" that prevent unified views of corporate data for analytical purposes. While it is true that building and preserving "data silos" may just be a holdover policy from prior years, it can also be a compatibility challenge – e.g., format, lineage, currency (time and/or monies), relevancy, comparability, etc. Business and IT executives need to focus on breaking down these data silos and work to break down inter-unit barriers so that there is seamless sharing of knowledge, access and usage.

Data Cleansing

Dirty data is the bane of the industry. One reason for dirty data is lack of motivation to keep data current. This dirty-data situation can be driven by a "not-my-job" syndrome or by a lack of incentive to clean up data when it pertains to data that is no longer important to a business unit. A lack of motivation to keep things current can lead to the ongoing maintenance of dirty data that really should be discovered and managed more carefully. Best practice is to create an incentive for all data owners to ensure data is updated, kept current, and stored in right place in timely fashion.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; Contact: inquiry@rfgonline.com

Teams spend 80 percent of their time cleaning data and only 20 percent analyzing it. However, we already know that data cleansing is least costly when the data is corrected early in the cycle. When the dirty data persists, the operational cost of working with that data goes up 10x for each additional phase of the data life-cycle. The same phenomenon is seen with code-bugs – IT organizations must address them earlier, or it will cost more later on to address the problem. In particular, the lack of clean data in healthcare is a major challenge and errors can be life-threatening.

In addition to the above-mentioned data errors, RFG 100 participants pointed out five other data usage challenges:

- Sync problems
- Software bugs in data processing
- Obfuscation by users
- Validity challenge
- Cross-field validation

It was suggested that data scientists can clean data. Data scientists often use Microsoft's Excel Power BI to eliminate duplicates, one participant noted. After the data is cleaned, in most cases there is the need for normalization and filling in missing data, which today is still an art. Some firms have found that they can use AI/ML for finding and addressing bad and missing data. The use of AI/ML cuts operational costs and time significantly and allows for early integration into the DataOps process.

Data Discovery and Generating Metadata

For most enterprises data has been created and inherited from decades of transactional processing and stored in a variety of systems throughout the company. Although the very oldest data has likely been archived, there still are aging databases and aging formats in which older data is being stored and accessible online that still may be of value to new applications.

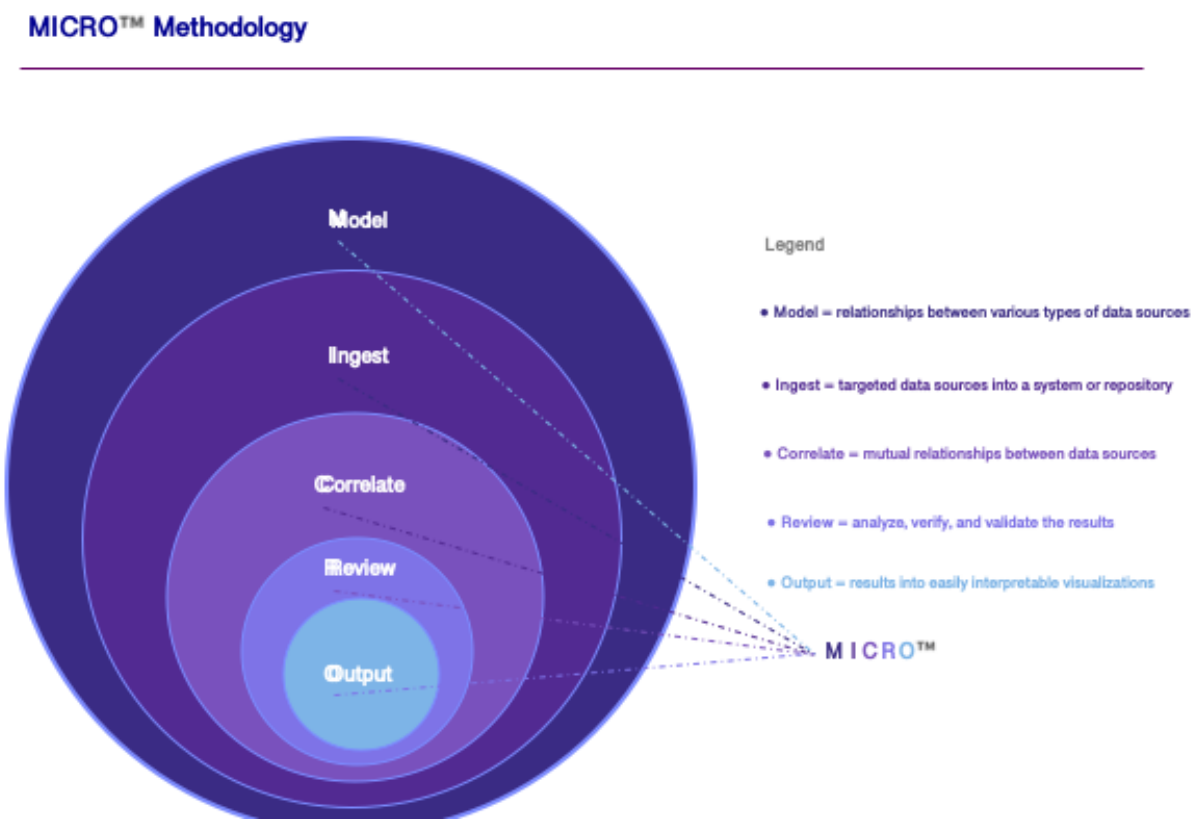
Data discovery must be performed in order to "take inventory" of all corporate data regardless of age – and artificial intelligence (AI) and software automation tools should be used where it is possible to do so. There are systems and storage vendors that can provide such tools – as can cloud services providers – for a fee. Alternatively, customers can work with systems integrators (SIs) and third-party consultants to locate and catalog all the datasets and data.

Metadata must be generated and made available to all in the enterprise so that, as mentioned above, searches of business unit data warehouses and data lakes by

other business units can occur seamlessly and efficiently. As we described in the RFG June 2020 report: *DataOps: Companion to DevSecOps for Reimagining Applications,*" generating metadata is a key ingredient to speeding up advanced analytics for DataOps projects – also known as "sprints" assigned to focused, and time-limited, team projects.

Correlating and Measuring Data Sources for Risks and Revenue Impact

Quite often enterprises that are aggregating data neglect the consideration of data sources when aggregating data. These data sources tend to exist in data silos within and between business and operational units, information and network security, and development and security operations. Business and IT executives should consider correlating the data sources and measuring them in order to reduce the risks and revenue impact. This is attainable through the following "MICRO" methodology that Building Noble Solutions developed and was shared by Shawn Butler, one of the panelists:



This method allows for monetization of the data while addressing security and financial, legal and brand risk. It is suggested that enterprises focus on applications that represent 70-80 percent of revenues.

How Much Data is Enough? More is Better

The more data used the better, as data accuracy is improved. That is, redundancy improves quality. Even "old or aged" data – data created by applications that have since been enhanced and that may be outdated (but current) may still be usable. In some cases, it may not be the data that is useful but, rather, the metadata from the old data that is valuable as the new datasets may lack that associated information.

Linking multiple sources can create a difficult situation in which PII (personal identification data, as in the E.U.'s GDPR and California's CCPA data-privacy regulations) considerations did not exist when the data was segregated. As this can become a major privacy issue, someone within the enterprise must be responsible to ensure data segregation is enforced in cases where it is required.

A best practice is to keep data warehouses where they are, rather than copying them over, panelists suggested. In other words, identify them, find them, and then correlate and use them no matter how dirty the stored data may be. Data cleansing is not always needed as long as the problem can be addressed by utilizing multiple data sources. In many use cases the users may not care about the bad data fields.

SUMMARY

Now that enterprises have sufficient compute power, memory and storage capabilities to handle vast quantities of data, customers can achieve better results by utilizing the aggregate of applicable data that is available to them. AI/ML tools will enable to firms to cleanse the data rapidly and enhance the quality of the data. Using these new approaches to discover and identify previously unavailable data, all lines of business within an enterprise will actually be able to "know what they know" rather than suffer from lack of access to and knowledge of data that they are unable to utilize.

RFG POV: Corporate data today is an underutilized (and often hoarded) asset that should be unleashed across the enterprise and accessible in machine-readable form by all lines of business. Eliminating data silos will make data "agile" and available for new applications and DataOps. Business and IT executives will need to address these data quality requirements if they hope to deliver new customer-facing applications that take holistic views of customers and prospects, both now and in the past.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; Contact: inquiry@rfgonline.com

Additional relevant research and consulting services are available. Interested readers should contact Client Services to arrange further discussion or interview with Mr. Cal Braunstein, CEO and Executive Director of Research. Jean S. Bozman, President of Cloud Architects Advisors LLC, co-authored this report.

APPENDIX 1: RFG100 Survey Results

Questions to the RFG100 attendees, followed by the poll results (shown as percentages of total):

- 1. Does your organization have internal data catalogs and/or data dictionaries available to you to assist you with data acquisition and sourcing?
  - Data catalogs – human readable 62%
  - Data catalogs – machine readable 38%
  - Data dictionary – human readable 46%
  - Data dictionary – machine readable 23%
  - None 23%
- If needed data resides in another division/business unit, do you have automatic access to the actual data source and automatic copy and usage rights? Who is responsible for keeping your copies updated?
  - Must request access to the data source 62%
  - Automatic access to the data source 15%
  - Automatic copy and usage rights globally 8%
  - Must obtain permission to copy and/or use data 46%
  - My organization is responsible for getting updates 15%
  - Original division controls sending updates 23%
  - Mutual agreement determines who does updates 51%
  -
- If the needed data source resides in another division/business unit, do you have a way of interrogating the metadata to determine items such as currency, frequency of update, quality, format, location, etc.?
  - Yes, by human 46%
  - Yes, machine readable 23%
  - No 31%

---

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; Contact: inquiry@rfgonline.com

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; Contact: inquiry@rfgonline.com