Supercharge Your Data Knowledge With Agile Data Intelligence

Authors:

Timm Grosser

Jacqueline Bloemen

BARC

Table of Contents

Supercharge Your Data Knowledge With Agile Data Intelligence

Timm Grosser, Senior Analyst Data & Analytics

Jacqueline Bloemen, Senior Analyst Data & Analytics

May 2022

RESEARCH NOTE

Analyst Opinion and Takeaways

Agile data intelligence is about connecting metadata within a metadata fabric

Agile data intelligence encompasses the company's ability to extract value from metadata. It provides the framework for using data efficiently and reducing complexity in data processes. Agile data intelligence is based on 3 core areas called 'smarts'. These take over the necessary stitching work within the metadata fabric. They lend a data intelligence platform the agility required for it to be operated effectively and efficiently. This is key, given that metadata is difficult and time-consuming to integrate and link.

Data knowledge is underutilized – start taking metadata seriously

Fast reaction times and resource bottlenecks force us to become more efficient and to automate in order to master complexity. If data & analytics is a success factor for future business, then agile data intelligence ought to be the prerequisite for its efficient implementation. However, the reality is different. According to a BARC survey, 87 percent of respondents state they could significantly improve their corporate results if they optimized the way they handle data (Leverage Your Data, 2020, n=412). Metadata is the key to the solution. We have been working with metadata on a daily basis for years, but we are still in the early stages of deriving direct benefit from the linkage of disparate metadata. But this is essential to efficient data usage today and in the future.

Metadata use needs smart technologies

Automation requires smart technologies that use metadata as a knowledge base to gain new insights. Data intelligence platforms

help to simplify the difficult and time-consuming process of metadata integration, linking, enrichment and analysis through automation. They reduce the complexity of distributed (meta) data and system landscapes and help to make the use of metadata feasible. According to the BARC survey How to Rule Your Data World, users see the greatest benefit of technology in data transparency, in being able to find data quicker and more easily and, above all, in saving time.

Orion's approach for integrating and linking metadata

Orion Enterprise Information Intelligent Graph supports the core capabilities of agile data intelligence. Benefits from metadata are particularly promoted through the automated collection and evaluation of operational metadata, with the objective of making causes and effects in data activities transparent in a timely manner. The approach of building the metadata model from the bottom up and the ability to automatically link a wide variety of metadata types with each other is strongly reasonable. If the automation proves itself in practice, new possibilities for the use of metadata will be created. Overall, the need to get more information about data is growing, or rather it is one of the biggest demands on the data landscape today. Orion's roadmap is promising and it remains to be seen to what extent it will be able to directly operationalize metadata and directly benefit from it.

No benefits from metadata without conscious promotion of a data culture
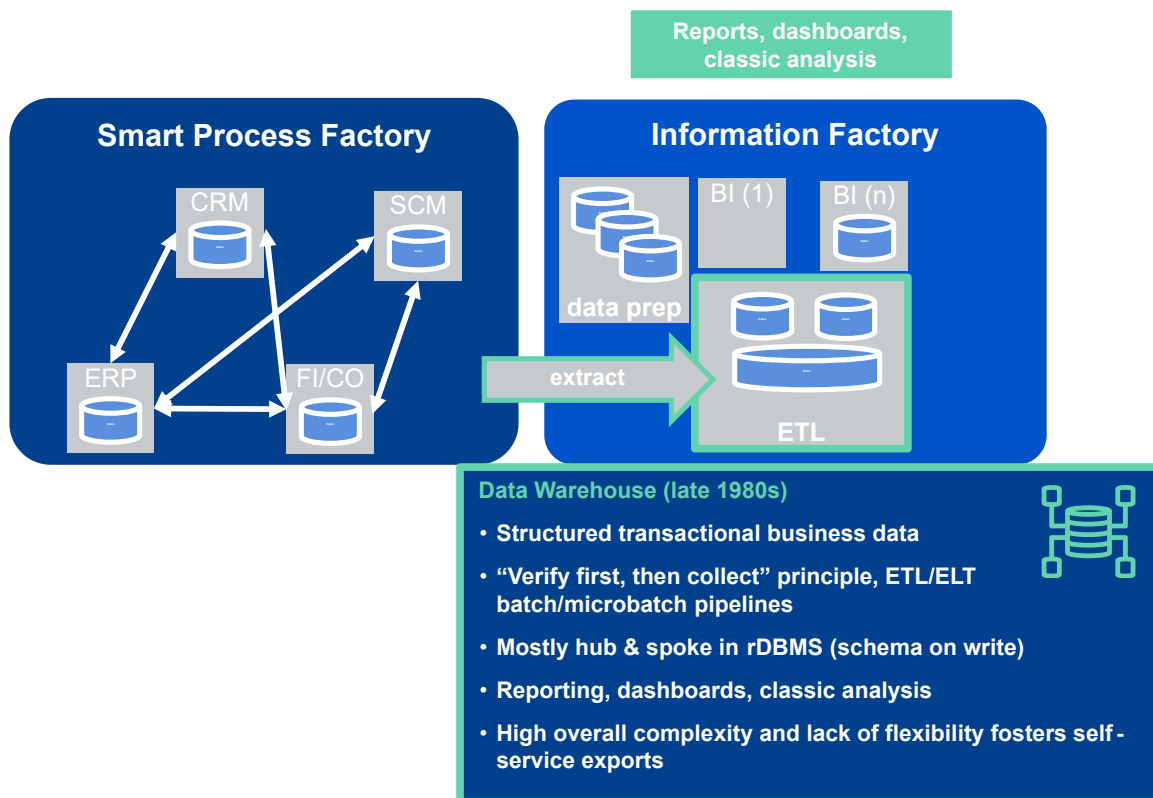
There is still homework to do for companies looking to make successful use of metadata management tools. The use of metadata requires the willingness of the entire company to share knowledge. This demands a data culture that creates this very mindset and encourages the creation, collection and analysis of metadata. This is done through transparency of the value of metadata, as well as the promotion of knowledge sharing, for example, through incentives, processes and capabilities. To create acceptance, the collection and use of metadata must not be seen as an additional effort.

What Is a Data Fabric and Why Do You Need One?

The concept of the data warehouse, which emerged in the late 1980s, endured for a long time. It was considered the undisputed best practice for making corporate data available for business intelligence applications. It follows the principle of first verifying and preparing the structured, transactional business data from different sources and then combining them in an integrated data model.

However, the concept is getting old now. Many consider the data warehouse approach to be only moderately successful. Users complain of insufficient benefits, quality, efficiency and effectiveness. In many companies, this has led to a proliferation of self-service BI solutions. While this can be quite successful locally, it leads to further fragmentation of the data landscape in the medium term. The number of data silos grows, complexity and administrative effort increase, and data inconsistencies jeopardize the meaningfulness and reliability of the information provided.

> Data silos characterize data & analytics landscapes.

Smart Process Factory

CRM

SCM

ERP

FI/CO

Information Factory

BI (1)

BI (n)

data prep

extract

ETL

Reports, dashboards, classic analysis

Data Warehouse (late 1980s)

• Structured transactional business data

• "Verify first, then collect" principle, ETL/ELT batch/microbatch pipelines

• Mostly hub & spoke in rDBMS (schema on write)

• Reporting, dashboards, classic analysis

• High overall complexity and lack of flexibility fosters self-service exports

Figure 1: Data warehouse concept of the late 1980s

Especially in times of rapidly changing markets, decision support systems should promote the fastest possible growth in knowledge. Advanced analytics and new ways of handling data also create new requirements that go beyond traditional concepts. Over time, new concepts have been introduced and new technologies needed. In turn, new types of data have had to be supported, for which the mostly relational data warehouse is not very well suited. Also, explorative analytics using data discovery, data preparation and data science is not adequately supported by classic BI tools and processes. Against this backdrop, the concept of the data lake emerged around 2010.
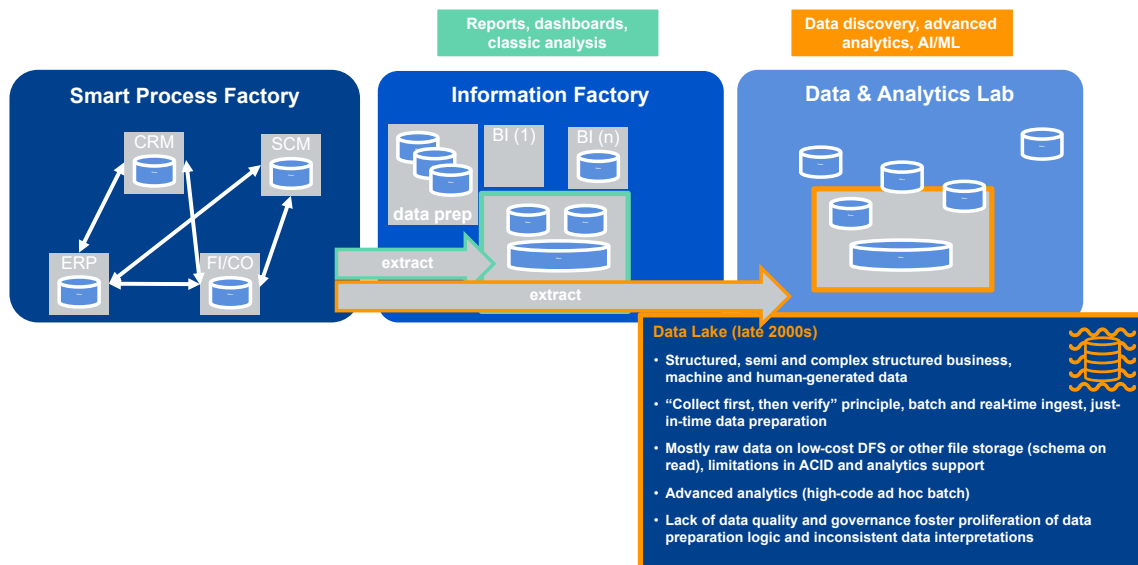
Figure 2: Data lake concept of the early 2010s

However, this concept also turned out to have its weaknesses. Although raw data could be provided quite quickly, its use in the context of analytical use cases involved a great deal of data preparation. For some use cases, pre-processed data warehouse data was an option, but it was not connected to the data lake. Agile working in the data lake increasingly led to an expansion of the data silos stored in it. The data lake evolved into the much-cited 'data swamp'.

As a result, it became increasingly clear that both concepts - data warehouse and data lake - each had their own strengths as well as weaknesses. So why not bring together the strengths of both concepts in a combined environment? This is how the data lakehouse came to be, which is now considered by many to be current best practice.

However, all three approaches – data warehouse, data lake and data lakehouse – have a common weakness because they follow a similar basic idea. They physically move data to one location in order to store and process it there. In principle, they try to solve the issue of data silos in source systems with more data silos. This makes less and less sense in the increasingly distributed data landscapes of digital enterprises, and is clearly driven by the rapidly growing proliferation of multi/hybrid cloud and edge applications. In the face of growing data volumes, large-scale data movement is not a viable concept. Cloud providers' pricing models also make it commercially unattractive to extract extensive amounts of data from the platform. The discourse about the „right" data architecture is thus entering the next round.

A future-proof data architecture must make it possible to jointly use physically distributed data assets. There must be no constraint on storing all data together physically in one place. At the

> Trying to overcome data silos by building MORE data silos is like rearranging the deck chairs on the Titanic.

same time, it must be possible to define a logically holistic view of it in order to ensure that the data can be used across the board.

Therefore, the data architecture of the future must meet the following requirements:

- Reduced complexity and extent of data processes while increasing flexibility and agility

- Simplified access to and use of data, especially for business users such as data producers and data consumers (including power users and data analysts)

This is where data fabric comes in. Data fabric is a concept and architecture principle that enables better utilization of disparate data sources independent of usage and deployment type, and regardless of location. It helps organizations leverage their data & analytics assets in a more efficient, effective and simplified manner to better support decision-making, improve business performance and create new innovations.

A data fabric enables the use of distributed data pipelines designed according to different patterns. These include classic data pipelines from the information factory (ETL processes), interactively and iteratively developed data pipelines from the data & analytics lab, and the near or real-time, event-driven data pipelines of the smart process factory. Furthermore, data fabric assets include not only database tables and other data objects, but also further data products, such as reports, dashboards and advanced analytics models. This makes a data fabric federated and multi-modal.
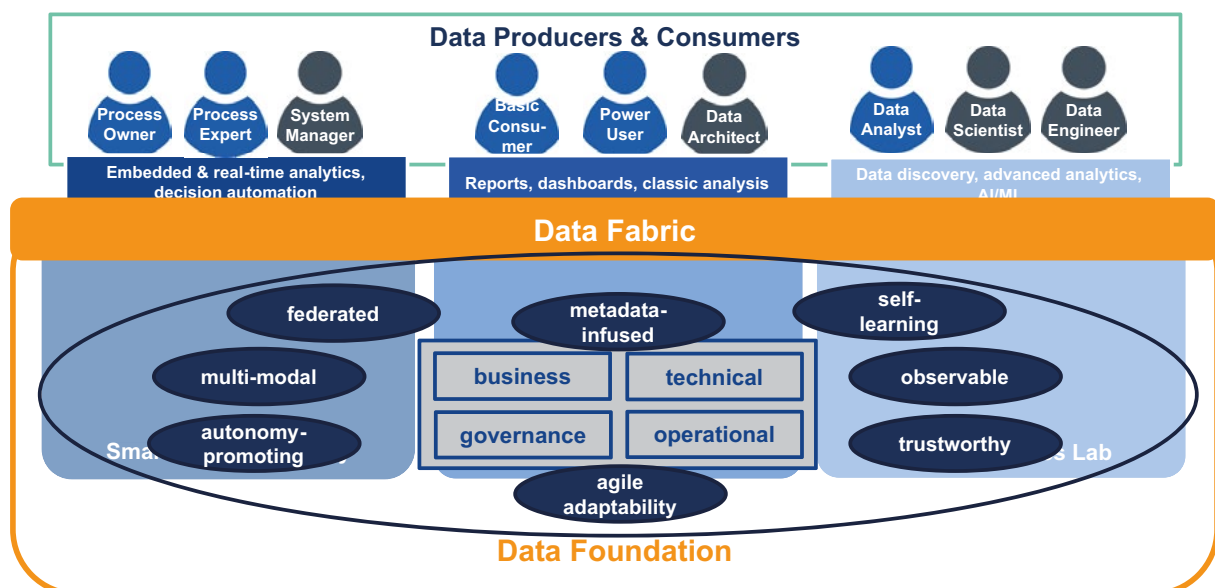

Figure 3: Characteristics of a data fabric

Practically speaking, a data fabric is the sum total of the data management components required for this purpose. At first glance, this is nothing new. But what is different about the data fabric is that it is holistic, enabling the use of distributed data & analytics assets without having to physically run or transfer them to a single location. In other words, it provides the ability to virtualize not only data objects, but also data pipelines and analytics components. At the same time, it provides business and technical users with centralized access to these assets and encapsulates the complexity of the actual physical implementation. To meet these requirements, a data fabric is primarily dependent on the availability of business and technical metadata, but also operational and governance metadata. These are the knowledge base of a data fabric, from which further typical characteristics such as autonomy-promotion, self-learning, observability and trustworthiness are drawn.

A data fabric facilitates optimized delivery, management, consumption and governance of data & analytics that is augmented, automated, smart and AI-driven. This makes it a central enabler of data democracy and cross-functional collaboration in the data-driven enterprise.

Agile Data Intelligence at the Core of Data Fabric

The biggest challenge for data consumers today is to find, understand, trust and use the data that is relevant to their questions. According to the BARC Survey Leverage Your Data, 65 percent of respondents agree that their analysts spend a lot of time searching for the right data and analyses and repeating work that has already been done, which hampers their productivity. Consequently, 59 percent consider providing more information about data as the topmost priority in improving this situation while 56 percent believe that creating a business glossary that facilitates cross-departmental understanding of data is an important approach to solving the problem. However, only 39 percent and 27 percent respectively are taking these steps. So why are companies treading water here?

Our survey concludes that it is mainly a lack of available documentation and explicit knowledge that stands in the way of achieving these goals. Compiling this documentation from scratch is a manual and time-consuming activity. The few available data experts with business domain expertise have neither the time nor the priority to perform this task according to 55 percent of our survey participants. So, we need to find solutions that can compensate for the lack of human resources: solutions that learn from existing explicit knowledge and generate extended

> **Data fabric is a concept to simplify the use of distributed data & analytics assets without the need to physically integrate data.**

> **Inefficiency in data handling is mainly caused by a lack of available documentation and explicit knowledge.**

> **WANTED - solutions that automatically learn from existing explicit knowledge and generate extended knowledge from it.**

knowledge from it. Such explicit knowledge is available extensively in the organization, namely in the form of metadata. However, metadata is rarely collected consistently and brought to life with the help of analytics and machine learning. This process is increasingly referred to as data intelligence, a term coined by our analyst colleagues at IDC. IDC defines data intelligence as "leveraging business, technical, relational, and operational metadata to provide transparency of data profiles, classification, quality, location, context, and lineage, providing people, processes, and technology with trustworthy, reliable data."[1]

So, data intelligence sounds like a solution to our dilemma. But the problem with extracting, collecting and connecting all this metadata to service these needs is, much like the manual documentation of expert knowledge, it is a difficult and time-consuming task. Without a smart and automated approach, it is doomed to fail. This is where 'agile data intelligence' comes in. Similar to how data fabric helps stitch together the organization's business data landscape, agile data intelligence is the corresponding fabric for metadata.

Agile data intelligence is based on three core areas. These cover the necessary stitching work within the metadata fabric and lend a data intelligence platform the necessary agility to be operated effectively and efficiently.

> **Corporate knowledge is in the metadata.**

> **Leveraging metadata is a difficult and time-consuming task.**

> **Agile data intelligence helps to connect metadata like a metadata fabric.**

Smart and automated extraction, collection and cross-linking of siloed, disparate metadata sources

Metadata can be found in countless structured and unstructured sources. For example, well-structured metadata can be found in databases and data access engines, data integration platforms and analytics tools. But metadata is also hidden in program code. That must be extracted from the detailed logic of the code. The extraction of metadata from textual sources is also challenging. In the second step, metadata from a wide variety of sources must be linked together, which is often quite tricky due to the lack of uniform keys.

Smart support of human and machine-driven metadata curation and crowdsourcing

The next stage towards data intelligence is the enrichment of the collected metadata with further machine-derived information or expert knowledge. This includes, for example, results from data profiling and data quality trend analysis, data protection and privacy policies, or feedback from business and technical users.

---

1 https://blogs.idc.com/2019/11/25/defining-data-intelligence-intelligence-about-data-not-from-data/

Such enrichments can be performed manually or preferably driven by advanced analytics.

Smart and flexible metadata analysis to provide augmented and automated support for the needs of diverse user roles

Metadata is only truly valuable when brought to life. At best, metadata analytics is driven out of the core of the agile data intelligence platform and automated. The needs of different personas need to be met. Data producers and data consumers need support in developing, operating and using data products. Data facilitators need insight into who is using what data for what purpose, and what the overarching obstacles are in the data landscape.

Orion Enterprise Information Intelligent Graph

Orion's Enterprise Information Intelligent Graph covers the core requirements of agile data intelligence and offers interesting functions to evolve a data culture. At its core are functions for automatically collecting, storing, preparing, enriching and evaluating metadata.

The goal of the solution is a real, living, breathing map of the data and IT landscape with the help of metadata, in particular through the automated linking of a wide variety of metadata. This serves to gain new insights and, in addition, the ability to draw direct benefits from metadata.

> The goal is to generate benefits from linking different types of metadata.

Particular attention is paid to operational metadata, which the tools can use to monitor and analyze data activities and make the effects on data processes transparent. For example, any risks in data for BI reporting can be derived from an accumulation of problem reports for an ERP system, or the effects of security breaches in source systems can be made transparent for subsequent applications.

The relatively young product (released in 2017) today supports over 20 customers mainly for compliance and migration use cases. With the support of more data & analytics use cases, the better support of business information consumers should be further advanced. Users of the tool are therefore not only IT staff and data stewards, but also business users such as business analysts, data scientists and even chief data officers.

> The product has evolved from its core competence in data lineage.

The tool can be run in the cloud (AWS, Azure, GCP), making it easily accessible to users. It relies on the reusability of already

used tools for collaboration, issue management, rule management, data quality, and on top of that, it offers integration options with the existing data landscape via open APIs. It therefore brings with it the ability to integrate well into data & analytics architectures and users' workflows.

The Secret Sauce of Orion is in Metadata Integration and Linking

Orion takes a bottom-up approach to building the metadata repository. This means that metadata from a variety of sources is integrated automatically and the metadata can be automatically linked to each other.

The platform provides 60+ connectors that help to generate or extract metadata to data & analytics artifacts and to collect operational metadata to data processing processes and data usage. What is special is the ability to generate metadata from programming code or from application version states as found in version control systems such as GIT. This makes direct access to production systems optional. Log files and Enterprise Scheduler are interesting metadata sources, providing additional information about assets and their relations.

Figure 4: Overview of most relevant metadata connectors

Orion can be categorized as a kind of metadata crawler that attempts to extract as much information as possible in the form of metadata from a wide variety of sources.

After scanning the metadata, Orion tries to automatically link a wide variety of metadata types, such as business, technical, operational or organizational metadata, among other things

with the objective of being able to derive new insights.

With the objective of being able to map the company reality as accurately as possible, the metadata can be kept up to date via near-time mechanisms. For this purpose, there are dedicated change data capture functions that help to process only changed metadata and handle mass amounts of metadata more effectively.

The strength of the solution lies in the automation of integration and linking of metadata. This is mainly based on a knowledge database with rules for linking as well as the ability to identify metadata via functions such as matching or pattern analysis. This enables, for example, the extraction and interpretation of metadata by scanning annotations in applications or code snippets.

Intelligent Data Curation Support

Orion supports the enrichment of metadata in various ways. With loading, for example, additional metadata can be collected through data profiling. The statistics provide valuable insights into the status of the dataset and help monitor data quality and regulatory compliance. For this purpose, Orion provides the Amazon AWSlabs Deequ Profiler out-of-the-box but is also open for integration with other data profiling tools.

Both humans and machines bring structure to the metadata through clustering, categorization, classification and labeling. Given the high complexity of metadata maintenance, this process should be automated as much as possible to save valuable human resources. For example, machine learning mechanisms help to cluster, categorize and label data & analytics assets. In essence, Orion uses pattern recognition for identification and its rule database for the assignment logic. Cluster rules, classification rules, tagging rules and business rules can also be individually designed or extended. Orion thus provides a certain intelligence that can save time and effort and help to reduce manual effort.

Users can edit, rate or make recommendations on business metadata through the UI, supported by approval workflows.

Orion sees high potential benefits in the evaluation of operational metadata

Value comes from the analysis of metadata. Orion covers the spectrum from search to reporting and metadata analysis, both role and context-specific. The user gets started via a web-based portal page, which can be customized with the help of small dashboard tiles for metadata analysis. This helps to display essen-

tial information optimized for the user. Thus, data stewards can stay informed about the curation status of their data & analytics objects, and IT experts can keep tabs on the use of the data sources they are responsible for in subsequent applications.

Search functionality is extensive and, using a scoring mechanism, delivers qualified result lists that can be further refined with filters. For more efficient evaluation of hit lists, Orion uses parameters such as ratings, certifications, frequency of use and criticality to sort results according to relevance to the search query.

Orion offers a number of pre-canned reports for the evaluation of metadata to support many kinds of users (e.g., ratings, user activities, ingestion jobs, issue overviews). In addition to classic lineage analysis and graph analysis, additional algorithms are used to find similarities and for clustering of similar datasets, which is frequently used to identify redundancies and data bloat.

Figure 5: Analysis of issues created in context of applications

Code analysis functionality also helps users to better assess and evaluate SQL code and provide insights into the quality of the code. Through close technology partnerships, such as with Snowflake, it is also possible to evaluate SQL code on an application-specific basis. Orion draws the knowledge for the assessment from its rules database.

Metadata Delivery and Workflows to Support Data Governance Processes

Companies need much more than metadata for data-driven action. Data governance processes, in particular data stewardship processes, make a significant contribution to success. Orion actively supports these with special stewardship dashboards (e.g., outlier detection, similarity reports, inventory object counts

etc.), data quality monitoring and analyses (data quality trend analyses, quality scores), data governance workflows and the monitoring of compliance with policy rules (e.g., policy rule violation reports). Thanks to the integrated workflow engine, even more complex workflows can be freely designed as release processes with status adjustment. It is possible to design a data shopping-style release process for data, which is not directly supported by Orion's solution.
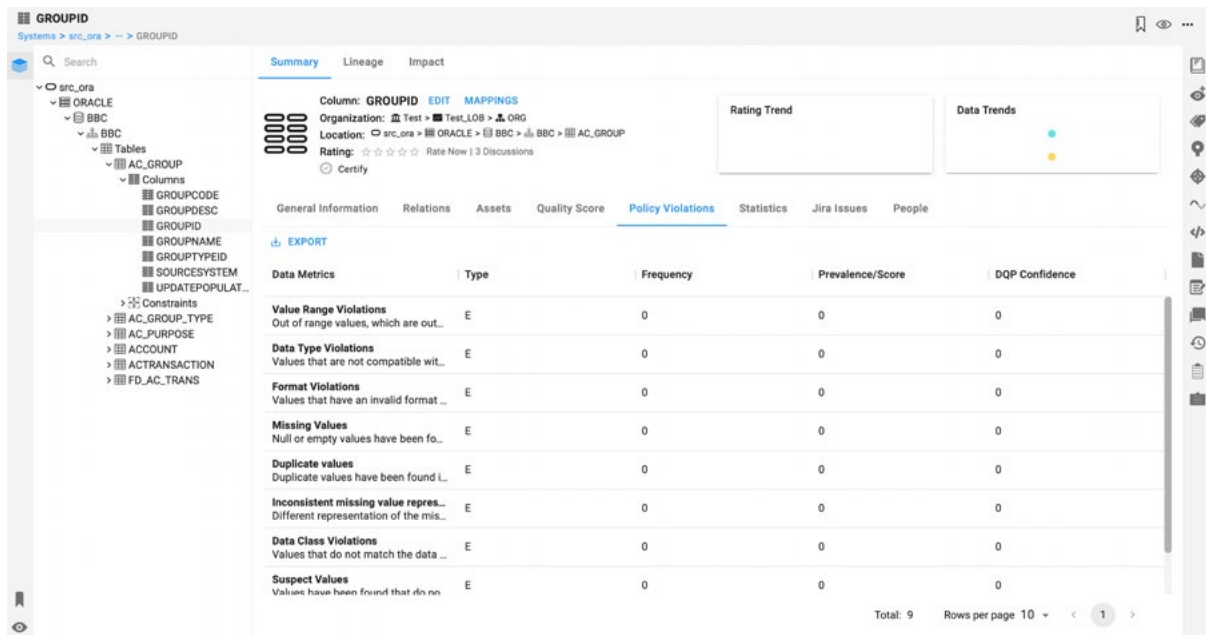

Figure 6: Analysis of policy violations

Integrations with issue management software such as Jira and ServiceNow give data stewards greater control. Integration with collaboration tools such as Teams and Slack, on the other hand, enables the promotion of metadata exchange not only for data stewards but also between other users.

Orion

Orion was founded in 2017 with a mission to disrupt the Information management space with a platform to provide the lowest Total Cost of Ownership and Fastest Time to Value.

Orion's customers include Global 5000 companies in banking insurance, retail, healthcare and information technology who are challenged by a variety of complexities; regulatory, cloud migration, transformational and silo-based constraints.

Orion weaves the most comprehensive metadata knowledge graph in near real-time to create a self defined data fabric in days/weeks vs. months or years.

Orion leverages AI/ML to automate the ingestion of more than 60+ technologies into a format that is vendor/technology agnostic and automatically stitching that together to create a self-defined Data Fabric of your ecosystem with Zero Impact to Production (ZIP).

Orion is headquartered in the San Francisco Bay Area. It has global offices in other US cities, Sweden, Germany, Estonia and India and Reseller partners in Japan, China and Singapore.

www.oriongovernance.com

info@oriongovernance.com

BARC – Business Application Research Center

BARC is a leading enterprise software industry analyst and consulting firm delivering information to more than 1,000 customers each year.

For over twenty years, BARC analysts have combined market, product and implementation expertise to advise companies and evaluate BI, Data Management, ECM, CRM and ERP products. A long-running program of market analysis and product comparison studies forms the basis of BARC's comprehensive knowledge of all the leading software vendors and products, as well as the latest market developments and trends.

BARC events provide a focused overview of leading software solutions, trendsetting developments and current requirements, as well as market developments in particular enterprise application categories.

Germany

BARC GmbH

Berliner Platz 7

D-97080 Würzburg

+49 931 880651-0

www.barc.de

Austria

Meldemannstraße 18 / 01.14

A-1200 Wien

+43 660 6366870

Switzerland

BARC Schweiz GmbH

Täfernstraße 22a

CH-5405 Baden-Dättwil

+41 56 470 94 34

Rest of the World

+44 1536 772 451

www.barc-research.com