

Data Intelligence: Empowering the Citizen Analyst with Democratized Data

Prepared by:

David Loshin
President
Knowledge Integrity, Inc.
(301) 754-6350
loshin@knowledge-integrity.com

Sponsored by:



Introduction: Data Democratization

The nexus of three key technological advances – exploding data production, a lowered barrier to entry for big data computation, and the development of application libraries encapsulating advanced analytics, machine learning, and artificial intelligence – has established a fertile environment for data innovation. Streamlined data access and analysis capabilities have enabled the evolution of the “citizen analyst” – a business-oriented problem-solver with enough technical knowledge to understand how to apply analytical techniques to collections of massive data sets to identify emergent business opportunities.

The priorities of the citizen analyst are straightforward: find the right data assets to support rapid prototyping of reports and analyses, develop visual presentations, and develop the appropriate data stories to advise senior management. And while the fundamental aspects of information management such as ingesting, integrating, storing and providing access to increasing numbers of data assets are meant to support the citizen analyst in developing their reports and analyses, some key practical data management issues contribute to a growing need for enterprise data governance, including:

- **Increasing data volumes** that challenge the traditional enterprise’s ability to store, manage, and ultimately, find data;
- **Increased data variety**, balancing structured, semi-structured, and unstructured data, as well as data originating from a widening array of external sources;
- **Reducing the IT bottleneck** that creates barriers to data accessibility;
- **Desire for self-service** to free the data consumers from strict predefined data transformations and organizations
- **Hybrid on-premises/cloud environments** that complicate data integration and preparation; as well as
- **Privacy and data protection laws** from many countries that influence the ways that data assets may be accessed and used.

Empowering the citizen analyst demands *data democratization*, in which shared enterprise data assets are made available to a broad set of data consumer communities in a governed way. The objectives of governed data democratization include:

- Raising **data awareness** among the different data consumer communities to increase awareness of the data assets that can be used for reporting and analysis,
- Improving **data literacy** so that individuals will understand how the different data assets can be used,
- Supporting **observance of data policies** to support regulatory compliance, and
- Simplifying **data accessibility** and use to support citizen analysts’ needs.

Effective data democratization (and consequently, enabling citizen data analysts) is dependent on accumulating, documenting and publishing information about the data assets available for use across the enterprise data landscape. **Data intelligence** is produced by coordinated processes to survey the data landscape to collect, collate, and publish this critical information, namely:

- **Reconnaissance:** Understanding the data environment and the corresponding business contexts and collecting as much information as possible;
- **Surveillance:** Monitoring the environment for changes to data sources;
- **Logistics and Planning:** Mapping the collected information production flows and mapping how data moves across the enterprise;
- **Impact Assessment:** Using what you have learned to assess how external changes impact the environment;
- **Synthesis:** Empowering data consumers by providing a holistic perspective associated with specific business terms;
- **Sustainability:** Embracing automation to always provide up-to-date and correct intelligence; and
- **Auditability:** Providing oversight and being able to explain what you have learned and why.

This paper explores each of these different data intelligence processes and then provides a check list for evaluating data intelligence products.

Reconnaissance

“A preliminary survey to gain information.”¹

Data warehousing is the typical organizational approach to sharing data for reporting and analytics. Yet the processes for making data available through the data warehouse are relatively constrained: after the data warehouse’s dimensional model has been architected, that model is generally set in stone, and changes are made with great difficulty. Data sets are extracted from selected systems sources, and a predefined set of data standardizations and transformations are applied before the data sets are loaded into the data warehouse.

Modern environments for reporting and analytics are less constrained, as organizations augment the data warehouses with data lakes – real or virtual repositories intended to collect data sets in their original format and making those data sets available to different users and allowing them to consume that data in ways specific to their own needs. Yet this poses a few new challenges. First, with a plethora of data objects strewn about the organization, one must determine which enterprise data assets are suitable for sharing. The second, and more important issue is that there is a need for a clear understanding of the data assets that are available, including information about the structure, content, and holistic characteristics of shared data.

This implies that there is a need for *reconnaissance*: understanding and accumulating information about the enterprise data environment so that data asset metadata can be documented and made available for data consumers. There are three aspects associated with reconnaissance that support collecting and documenting data asset information:

¹ <https://www.merriam-webster.com/dictionary/reconnaissance>

- **Objective assessment:** An objective assessment uses automated tools such as data profiling for scanning through structured and semi-structured data objects to surface information about structured data asset structural metadata as well as provide an objective validation of the data asset's quality.
- **Direct engagement:** This aspect involves directly engaging the data consumers in the environment to solicit their needs, collect and log business terms and their definitions, and harmonize those definitions.
- **Capturing learned knowledge:** This final aspect combines what can be learned into a shared environment that documents discovered metadata, business terms and their definitions, and data quality rules. In turn, the data stewards can determine which business terms (and definitions) are associated with physical data assets.

These reconnaissance activities are the methods for gaining information about the enterprise data landscape and capturing that information so that others in the enterprise can search and review the structural details about shared data assets.

The essential value of reconnaissance is data transparency. For example, a large insurance organization that had embarked on a master data management (MDM) project was surprised to learn that there were thousands of undocumented data assets that not only existed in different nooks and crannies of the enterprise environment, but that some of these undocumented data assets had been incorporated into production processes. Surfacing these hidden gems raised awareness about additional corporate data dependencies and highlighted the need for metadata harmonization prior to implementing an MDM product.

Surveillance

*"Close watch kept over someone or something."*²

Having performed our preliminary assessment of the enterprise data landscape, it is important to continually be aware of the potential for changes to the ecosystem. The extended information enterprise may span multiple environments, including on-premises, multiple clouds, SaaS and PaaS systems, edge computing nodes, Internet of Things (IoT) devices, etc., and many of the corresponding data sources are subject to change at any time. *Surveillance* is meant to convey the ability to automatically monitor the different sources, continually inquire about critical metadata characteristics such as:

- Structural metadata and the potential for changes in the set of attributes or their data types and sizes,
- Holistic data object attributes such as data owner, size of the data source, or date of last refresh,
- Streaming characteristics, such as data message structure, rate of data production, and streaming volume.

² <https://www.merriam-webster.com/dictionary/surveillance>

Continuous monitoring allows you to compare what is learned to what is already known and alert the data users if there are significant changes that might require attention. Surveillance can be implemented by layering automated monitoring on top of a proactive metadata management framework. An automated capability can point to all the data sources, scan the environment on regular periods, document the current state of the data sources, and trigger actions should there be any significant changes.

Any product that implements this capability requires broad connectivity to a variety of data resources, including text files, semi-structured files (such as XML or JSON files), both relational and NoSQL databases, documents, and data streams. It would also need to provide more intelligent connectivity that can scan code, ETL specifications, scripting languages, etc. to infer modifications to data resources. Having this type of connectivity allows you to reverse engineer data mappings from within the environment, while continuous scanning allows you to document different versions of enterprise metadata that can be monitored for changes.

The essential value of surveillance is transparency. For example, a large financial institution had automated processes in place to extract data from a set of sources, move that data to a staging area, and then load the staged data into a data warehouse. At some point, a source table interface changed, leading to the failure of data extraction. However, because no surveillance monitoring was in place, it took a month before anyone realized that the data warehouse loading process had been failing for weeks. Instituting surveillance would have alerted a data steward to the source modification and triggered a process to review the data production pipeline.

Logistics for Planning

“The detailed organization and implementation of a complex operation.”³

Increasing the breadth of data sources that are brought into the extended information enterprise is a boon for data analysts, as it enables more innovation in creating downstream information products. Yet as the number of data sources increases, so does the complexity of information production flows within the enterprise. This suggests the need for not only understanding how data sets are ingested, integrated, and made available across the organization but also how to implement new integrations and incorporate them in an efficient manner as new data sources become available.

One might think that documenting a static view of the data production flows provides enough insight to support efficient process orchestration. Yet due to the increasingly dynamic nature of data ingestion, a static snapshot of the environment will be insufficient as more data sources are added to the mix. One can attempt to manually oversee these integrations, but implementing new integrations manually will become unsustainable as the enterprise data landscape expands. This suggests the need for two *logistics* capabilities:

³ <https://en.oxforddictionaries.com/definition/logistics>

- Data lineage: Producing a representation of the current state of the lineage across the enterprise, and
- Data pipeline creation: Automating the creation of new integrations.

There are many tools that can produce a representation of enterprise data lineage. Some survey the data integration processes and produce what is essentially a static representation of the data lineage. One of the challenges of this approach is its static nature – while the snapshot of the different processing streams is accurate as of the point in time that the survey was performed, in a dynamic environment there are always going to be modifications that are not captured in each snapshot.

An alternative approach materializes the data lineage on demand through reviewing continuously scanned data mappings. By capturing how data elements are mapped from source to each intermediate target, your metadata repository can be used to generate lineage on demand to get the most accurate and up-to-date view. This approach not only provides a reverse-engineered view of the data, it also helps automating the creation of new data integrations. If you want to forward engineer a new integration, you can view the existing information flows, determine whether any existing information flows can be adapted to model a new integration flow, and then automatically generate the code.

The essential value of logistics is control. For example, an e-commerce company is interested in ingesting numerous externally produced data streams to feed an adaptive analytics recommendation engine. As new data sources are surfaced, the data engineers want to rapidly adapt existing data pipeline to the new sources to speed the integration. Automated data pipeline creation enables agility in ingesting new data streams.

Impact Assessment

“The ability to identify the consequences of changes made to an environment.”

Systems are designed to meet specified requirements. Yet business process requirements are not static; external priorities change, there are modifications to laws and regulations, and business motivations are adjusted due to market pressures. But if business process requirements change, how does this affect the operations of existing systems? The question is more concrete from the information management perspective: if there are data dependencies within an enterprise, and the underlying data sources change, how does one determine what downstream information flows, operational systems, business processes, and reports and analyses are affected?

In the past, there were basically two methods to determining impacts. The first was to manually review the environment to attempt to flag potential issues, but this approach is not only difficult and tedious, but also prone to error. The second approach was to allow the change to be introduced and then see where the system broke – not optimal in terms of business continuity, but at least it highlighted the areas of impact!

Since neither of these approaches are satisfactory, there is a need for automated processes for *impact assessment*. Impact assessment is a process that can analyze the data lineage and trace the dependencies of source data to their various touch points, thereby determining how changes to the environment impact dependent systems. Impact assessment provides intelligence about what applications need to be reviewed and potentially modified to accommodate the change to a data source. Impact assessment builds upon the intelligence accumulated through reconnaissance, surveillance, and logistics to support the acute needs of system maintenance planning and execution.

The essential value of impact assessment is preparedness. As an example, a healthcare provider network constantly monitors legislative initiatives to determine when new regulations and laws are introduced that might affect internal system processing. As new policies are introduced, the organization's use of impact assessment identifies the systems that are affected by the introduction of a new area of compliance affecting protection of patients' personal data.

Synthesis

*"The combining of often diverse conceptions into a coherent whole."*⁴

System developers often work in a vacuum. They are tasked with building a system that executes a particular task, satisfies one aspect of a business process, or solves a specific problem. A byproduct of this siloed development is that people often reinvent the wheel – creating the same or similar reference data sets, developing slightly different data models to represent the same types of entities, and applying the same or similar business rules to their own views of the data.

The challenge appears when data analysts attempt to combine data sets from different sources and find that the variances among what they thought were the same data element concepts skew their reports and analytics. One critical task of a data governance program is to identify where these variations are present across the organization, seek to understand whether the differences are relevant, harmonize where possible, and differentiate where it is not possible.

This suggests the need for synthesis of different intelligence data points that allow you to construct a holistic perspective of enterprise information concepts – which concepts are used, how they are used, and how they can be aligned. Synthesis embraces the determination of all process touch points associated with a specific business term, finding all the policies and business rules that are alike, identifying data assets that are associated with defined reference domains, and surfacing collections technical asset metadata all in one view.

Synthesis provides the data intelligence required to develop a holistic perspective blending enterprise metadata, data lineage, and corresponding integrations. And having this holistic mapping of the data origination points, data process flows, data touch points, and the corresponding semantics and rules for intended use, one can institute a set of controls to validate compliance with data validity rules and data quality expectations. Leverage the data lineage to

⁴ <https://www.merriam-webster.com/dictionary/synthesis>

identify those locations in the process flows where the business processes depend on well-defined, high-quality information. Inserting controls in those locations provides a means for continuously monitoring and reporting the quality of data across the enterprise.

The essential value of synthesis is harmonization. As an example, a manufacturing company embarked on a process of replacing a collection of ERP systems with a single ERP application. This initiative required the consolidation of key enterprise entity types (such as customer, dealer, vendor, etc.) into unified master data indexes. Yet with historically siloed development, the variances among the different underlying models made it difficult to effectively consolidate replicated information into a unified view. Synthesis would enable the manufacturer to accumulate the references to different instance data for all of the entity types and highlight similarities and differences among the core data models, enabling data stewards to harmonize business term definitions and corresponding data element specifications as a prelude to entity integration.

Sustainability

“The ability to be maintained at a certain rate or level.”⁵

Clearly, data intelligence provides critical information that supports data democratization. Yet the value of information degrades as it ages, suggesting that you need to make sure that the data intelligence process always provides the most up-to-date and correct knowledge about the environment. *Sustainability* focuses on the ability to not only implement processes for data intelligence, but to also orchestrate the execution of data intelligence procedures to maintain the timeliness and integrity of the enterprise data intelligence mappings over time. And as the scope of the extended information enterprise expands, the critical idea thing is that this cannot be done manually – you will need an ability to automate as much as possible.

Remove the human factor and automate the discovery and presentation of data intelligence. Adopt technologies that will automatically crawl through the enterprise to identify data assets, continuously survey the data landscape, provide up-to-date maps of the data lineage, and infer knowledge to help harmonize the distribution and use of business terminology. Continuously monitoring the enterprise not only provides a series of versioned snapshots of information use, it provides the framework for incorporating machine learning and artificial intelligence algorithms to help with sensitive data discovery and to guide downstream data consumers with recommendations of data assets that can benefit their reporting and analysis needs.

The essential value of sustainability is time savings. As an example, a financial institution embarked on the task of accumulating and organizing the metadata associated with the different databases in production. A back-of-the-envelope calculation determined manually doing this analysis would take two staff members three years to complete, over which time older systems would be retired and new ones introduced. The manual approach would never provider a complete and consistent view

⁵ <https://www.google.com/search?q=define+sustainability>

of the organizational metadata. Instead, orchestrating automated procedures to accumulate and infer metadata would more effectively produce a consistent metadata view.

Auditability

“The ability to provide a systematic review or assessment of something.”⁶

Data democratization at the enterprise level depends on harmonization and alignment across the organization. There are many details requiring oversight to ensure that data assets can be accessed, integrated, and used for the variety of downstream purposes. Reference data, shared domains, business terms, definitions, business rules, data quality expectations, and data policies: there is always going to be a need for data governance and oversight. In any intelligence operation, governance and oversight provide the guarantee the trustworthiness of what you have learned. This suggests the need for *auditability*: the ability to explain what you have learned, how you came to those conclusions, and the enforcement of the policies for making sure that there is accountability.

From one perspective, auditability encompasses the ability to systematically ensure conformance to a set of predefined standards and report the levels of compliance with those standards. Practically, instituting auditability relies on the combination of the intelligence artifacts delivered by our other data intelligence processes, such as metadata standards and data quality rules discovered during the reconnaissance process that can be validated at different stages of data flow pipelines represented by the mappings used to produce the data lineage. Deploying data controls that automatically verify conformance to standards enables quantification of data quality and usability.

From another perspective, auditability reflects the use of well-defined processes and trusted tools for producing your data intelligence. Regular reviews of your data intelligence procedures and the tools they use will provide a level of confidence in the knowledge those procedures deliver.

The essential value of auditability is confidence. As an example, an external review of a government agency identified some organizational risks associated with poor data quality. And while introducing some data quality practices were intended to address the root causes of those data quality issues, there was still a need to produce an auditable report demonstrating that proper data controls had been put in place. Once auditability was established, the external reviewers could be engaged to reevaluate the identified risks.

⁶ Adapted from <https://en.oxforddictionaries.com/definition/audit>

Considerations

This paper has examined a broad swath of processes for data intelligence to support a growing pool of citizen analysts to support the necessary operational data governance needs, including:

- **Reconnaissance** for metadata assessment and discovery;
- **Surveillance** to monitor for changes to identified data sources;
- **Logistics and planning** to map data pipelines and data lineage;
- **Impact assessment** to determine how changes impact the environment;
- **Synthesis** to provide a holistic business glossary, business terminology guide, and improve data literacy;
- **Sustainability** to ensure that the data intelligence is up-to-date and correct; and
- **Auditability** for oversight and governance.

Together, these processes and practices support an overall framework for expanding accessibility to enterprise data assets in a governed manner. Enterprise data intelligence activities combine expertise and well-defined processes with the right set of tools and technologies to support operationalization. When considering implementing a data intelligence program, look for products that support these types of capabilities:

- **Reference data management** for capturing and harmonizing shared reference data domains,
- **Data profiling** for data assessment, metadata discovery, and data validation,
- **Data quality management** for data validation and assurance,
- **Data mapping management** to capture the data flows, reconstruct data pipelines, and visualize data lineage,
- **Data lineage** to support impact analysis,
- **Data pipeline automation** to help develop and implement new data pipelines,
- **Data cataloging** to capture object metadata for identified data assets,
- **Data discovery** facilitated via a shared environment allowing data consumers to understand the use of data from a wide array of sources.

Finding the right set of technologies for data intelligence can help automate the discovery and assessment of enterprise data assets. Gaining an accurate perspective of the corporate metadata landscape reduces friction in data accessibility and utility, improves overall quality, and accelerates digital transformation as more individuals become adept at reporting and data analysis.

About the Author

David Loshin, president of Knowledge Integrity, Inc. (www.knowledge-integrity.com), is a recognized thought leader, TDWI affiliate analyst, and expert consultant in the areas of data management and business intelligence. David is a prolific author on topics related to business intelligence best practices. He has written numerous books and papers on data management, including *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* and *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is a frequently invited speaker at conferences, online seminars, and sponsored websites and channels including [TechTarget](#). His best-selling book, *Master Data Management*, has been endorsed by many data management industry leaders.

David is also the Program Director for the Master of Information Management program at the University of Maryland College of Information Studies.

David can be reached at loshin@knowledge-integrity.com.

About the Sponsor

erwin has always been the most trusted name in data modeling, but we became a private, stand-alone company in March 2016. Since then, we've acquired enterprise architecture, business process modeling/management, data harvesting, metadata management and data governance technologies.

These acquisitions, along with significant R&D, have culminated in the only data governance software platform with integrated capabilities for enterprise modeling, data cataloging and data literacy. The erwin **EDGE** creates an “**enterprise data governance experience**” so all stakeholders can discover, understand, govern and socialize data assets both at rest and in motion.

Whatever the role – data scientist, data steward, ETL developer, enterprise architect, business analyst, compliance officer, CDO, CEO – we're all “data people” who can improve how our organizations operate. And from our perspective, data governance drives everything – especially data intelligence for risk management, agile innovation and business transformation.

Please visit www.erwin.com for more information.