Fusion Alliance

A RENEWED CRISIS: THE CALL TO DEVELOP DATA ASSETS RESPONSIBLY

Author: Andrew Bittermann
Twitter: @BittermannBI

Introduction

With the emergence of powerful, next-generation business intelligence tools aimed at enabling self-service, the threat to the traditional data asset development and governance control paradigms is facing a renewed crisis.

How does an organization not just promote self-service, but truly encourage the development of new, powerful data assets by those who know the data the best? More to the point, how do they do this while also ensuring the proliferation of these assets occurs in such a way as to not create chaos?

Shadow IT is the idea of non-IT resources creating or implementing non-IT sanctioned assets or processes. Because shadow IT has emerged as an increasing problem for organizations who cannot keep up with the needed pace for information needs outside of IT, how will IT respond to shadow IT veterans now being retooled with substantially more powerful weapons -- while still encouraging the healthy behaviors that lead to faster and more valuable data asset development? How does an organization develop data assets responsibly in this environment?

The solution lies in striking the right balance between the conflicting forces of increased governance and the rapid expansion of the democratization of data. With an increased appetite for information and the emergence of

Key Takeaways

- *New, powerful tools make it increasingly likely that shadow IT is creating unsanctioned data assets at a greater rate than ever before*
- *The exponential growth of the amount of data available combined with the growing insatiability of the business for new information add fuel to the fire*
- *The need to control this explosion of data is more critical than ever*
- *The strain on existing governance controls and structures demands that organizations re-examine current processes*
- *Information needs must be recognized and embraced in the context of a well-governed environment to prevent a further strain on the relationship between business and IT*

What's the solution?

Establish an effective workflow for the review and promotion of data assets.

Doing so will be the key to bringing non-sanctioned, yet valuable data assets into the fold and will better position companies to achieve the appropriate balance.

new technologies that are changing the way companies and individuals interact with their data, solving this "old" problem is of more importance than ever.

Let's Get Specific. What's New About the Old Problem?

Shadow IT has become a growing problem at a growing rate. Excel has been a huge enabler of Shadow IT in: data preparation, visual and non-visual data analysis, and perhaps most concerning, the creation of separate persistent data sources (sometimes referred to as spreadmarts) that become unofficial, trusted data sources that reside outside of IT.

The growth of data and information has accelerated the demand for knowledge workers to get the data they require when they need it and in a format that is usable for their consumption. Not a new problem, but now there's a new chapter being written.

Enter the plethora of self-service data tools, particularly in the BI space. Self-service greatly reduces the demand on IT for performing repetitive tasks and enables users to personalize how they view and access their data. Tools like Tableau have been very effective in providing rich visualizations and helping users quickly see their data, usually differently than ever before. This has simply added fuel to the fire of demand for self-service data preparation.

Data preparation tools traditionally required more technical skills and toolsets focused on ETL (extract, transform, load), which traditionally lived in IT, but new tools are emerging that enable less-technical business users to accomplish these same tasks. In the process, shadow IT has become better equipped, but also more "dangerous" than ever.

As an example, when Alteryx emerged as the primary contender in this space, business users could start to accomplish work that would traditionally been out of reach for even the savviest Excel veteran. Flattening XML, geo-spatial custom polygon prep, blending, cleansing, etc., can all be performed without knowing any traditional IT development.

In addition to providing business-friendly abilities to transform and prepare data, business users who understand fundamental statistics can now create predictive models, including various regression models, neural networks and Naïve Bayes classifiers. For those who understand how to use them, they can create some very powerful models without writing a single line of R (a programming language used for data science).

Shadow IT just got a ballistic upgrade, and with it has come the need, now more pronounced than ever, to start understanding what data governance and responsible data asset development look like in this new era.

So What Needs to Change?

Mindset

First and foremost is recognition of the problem. Companies need to take a fair and honest look at the data management practices and acknowledge where pockets of shadow IT exist.

While Excel is obviously ubiquitous, understanding who has licenses for some of the newer tools, like Alteryx, may be a good place to start. When pockets are identified, ask some fundamental questions. How are these tools being used? What information is the business craving? What assets have they developed that could benefit a larger group or even the enterprise?

> "SHADOW IT (INFORMATION TECHNOLOGY) JUST GOT A BALLISTIC UPGRADE.
>
> WITH IT HAS COME THE NEED, NOW MORE PRONOUNCED THAN EVER, TO START UNDERSTANDING WHAT DATA GOVERNANCE AND RESPONSIBLE DATA ASSET DEVELOPMENT LOOK LIKE IN THIS NEW ERA."

The historical mindset that could be characterized as "scolding" those who created these one-off information stores needs to be turned on its head. Process, information requirements, tooling, and potentially infrastructure and architecture can then be re-examined based on the reality of what the business needs, thereby creating a self-service machine that can keep pace.

Information Demand Management
When you understand these fundamental drivers that led to the development of these assets, you will want to take a critical look at the demand management process for information needs throughout the organization.

When we talk about data asset management in the context of this discussion, it's important to understand the lifecycle of a data asset originating outside of IT. Data assets can only be created or derived from other datasets to which the end user already has access. While traditionally this was more focused on internal datasets, blending with external data, such as market, weather or social, is now more common.

Regardless, the first step in the data asset management lifecycle is to acquire the data. Data from multiple sources is blended and prepped for consumption, which typically includes steps to validate, cleanse and optimize data based on the consumption need. The data asset development lifecycle looks something like this:

- Intake – How are new requests for information captured? Once captured, how are they reviewed? Grouped or consolidated? Prioritized?
- Design – How will new datasets be rationalized against existing sets? How will common dimensions be conformed? How does the consumption architecture affect the homogeneity of data sets being created?
- Curation – How will the data be cleansed and groomed based on the consumer's requirements? Will different "quality" or certification levels of the data be needed?
- Output – How will data be delivered? A semantic layer that can be consumed by visualization tools? A more modern data marketplace where customers (end users) can shop for the data they need?
- Understanding – How will metadata (technical and business) be managed and made available for consumers for these sets? How is the business glossary populated and managed?
- Access – Who will have access to various delivered assets? Will control require row- or column-level security, and if so, what's the most efficient and secure way to implement those controls?

Let's be Clear on Governance
The term governance gets thrown around a lot, so let's be specific. While IT tends to have more controls around how data sources are formed and exposed to support governed access, governance itself and the development of a mature governance program, must be seated in the business.

For example, while implementing the actual controls to provide row-level security or perspectives will ultimately fall on IT, the true owners of what should be considered trusted data, its business definition, quality rules, and understanding of lineage, are the business data stewards who govern the data. In this sense, IT becomes the stewards responsible for ensuring those business-driven governance requirements are met.

There are books and papers on this subject, so for the context of this paper, the key takeaway is this: mature data asset development provides an efficient way for democratized data and data assets to be brought into the governance fold in such a way that does not hinder the free development of those assets. How is this accomplished? Workflow.

Workflow – The Bridge Between Democracy and Governance
We just talked about bringing the creation and persistence of data assets into the controlled IT fold. But allowing the business and knowledge workers to quickly and to freely blend, experiment and discover the most effective fit-for-purpose data sets for their information needs takes the burden off of IT to try to figure out what the business needs.

Let knowledge workers develop the sets, submit those of value to the intake process and, ultimately, then be persisted in a manner consistent with governance requirements, quality rules and access architectures that make them suitable and consumable by a larger group.

The underlying key here is workflow. How requests are processed, prioritized, reviewed, and either approved or rejected, is the critical gatekeeper that prevents democratization from turning into the aforementioned chaos. Workflow is the bridge-builder between democratized asset development and IT-implemented controls.

Management of the backlog of requested data inherently will require a gated process for intake along with defined processes for reviewing and promoting assets for department or enterprise consumption. Not every

"COMMUNICATION TO SUPPORT THE REQUIRED ORGANIZATIONAL CHANGE MANAGEMENT IS CRITICAL."

one-off dataset is an asset – the validity of the data produced must be verified to examine the lineage of the data and any transformation logic (e.g., joins, calculations) that was used in its creation. Furthermore, the usefulness of the data must be scored or assessed in some objective way to determine if it should be published and to whom. The workflow process should also address access and security requirements.

So what does workflow mean? Workflow is the process established by which data assets are submitted for enterprise or departmental consideration as governed assets and persisted according to IT's self-service standards. Identifying the roles, process (inputs, outputs, gates) and relevant governance structure is fundamental to get a meaningful workflow in place.

Furthermore, communication to support the required organizational change management is critical, a point that cannot be emphasized enough. Shadow IT is called that for a reason – to get those datasets and those who create them to willingly step in the light is a culture shift that should be managed as such. Communication is absolutely paramount in this regard.

Preparing Data Assets in a Big Data World

Enter the preparation of data in non-traditional sources. With the advent of the data lake, the overall reference architecture for most companies has changed. There is now a new "marshalling" or staging sector that allows companies to land vast amounts of data – structured, unstructured, semi-structured, what some have labeled collectively as "multi-structured" or "n-structured" – in a single region for retrieval at a later time.

This data may later be consumed in its raw form, other data may be slightly curated to apply additional structure or transformation, while other data will be groomed into highly structured and validated fit-for-purpose, more traditional structures.

Podium Data has developed a useful metaphor when speaking of these three levels of data. "Bronze" refers to the raw data ingested with no curation, cleansing or transformations. "Silver" refers to data that has been groomed in some way to make it analytics-ready. "Gold" refers to data that has been highly curated, schematized and transformed suitable to be loaded into a more traditional data mart or enterprise data warehouse (EDW) on top of a more traditional relational database management system.

Podium, like Microsoft and others, has adopted a "marketplace" paradigm when talking about the development of data assets for consumption. The concept is basically the creation of datasets persisted to a common portal where consumers can "shop" for the data they need. Podium provides it's "Prepare" functionality to schematize and transform data residing in Hadoop for a marketplace type of consumption.

AtScale is another Hadoop-based platform for the preparation of data. It enables the design of semantic models, those that are meaningful to the business, for consumption by tools like Tableau. Unlike traditional OLAP semantic modeling tools, a separate copy of the data is not persisted in an instantiated cube. Rather, AtScale embraces OLAP more as a conceptual metaphor. For example, when Tableau interacts with a model created in AtScale on top of Hadoop, the behind-the-scenes VizQL (Tableau's proprietary query language) is translated in real time to SQL on Hadoop, making the storage of the data in a separate instance unnecessary.

Alteryx, mentioned earlier, is also a powerful tool for extracting data from Hadoop, as an example, manipulating it, then pushing it back into Hadoop for consumption.

These tools and platforms exist to meet a fundamental need for the creation of data assets in a big data world. Rapid ingestion, profiling and creation of metadata are critical even for Bronze-level data. Tools and platforms will continue to be created to meet the need of curating this data to meet Silver- and Gold-level consumption needs.

Where to Start

Establishing or refining your demand management process is important, but a great way to jumpstart the population of your information needs backlog can be as simple as creating an inventory of current one-off data stores and sources throughout your department or organization.

While existing IT-generated data assets may not be assets at all, these one-off sources are clearly valued; they serve a real need and have value to those utilizing them.

Furthermore, based on the inventory of these stores, and after completing a more formal collection of information needs, current IT-generated assets should be rationalized. The cost to maintain them, particularly if they're not actually being used, makes them liabilities, not assets.

This basic exercise will shift IT's focus to start providing higher-value data and information for the business, while potentially driving down cost by retiring the production of lower-value reports or marts. In short, inventory one-off non-sanctioned sources, catalog information needs, retire low-value IT information assets, and kick-start your improved demand management process by focusing on providing the highest-value assets.

Leveraging Toolsets to Implement Governance

It is worthy to note that many self-service tools have a server component to their overall architecture that is used to implement governance controls. Both row-level security (RLS) and column-level security (sometimes referred to as perspectives) can be put in place, and implementations of that security can be accomplished many times in more than one way.

Many of these tools can leverage existing group-level permissions and security that exist in your ecosystem today. Work with a consulting services partner or the vendors themselves to understand recommended best practices in configuring the tools you have selected in your environment.

Conclusion

Creating an environment that encourages the creation of self-service and democratized data asset development by the business is important, but, when unchecked, results in the proliferation of potentially redundant or conflicting data sources, none of which are under IT's purview.

Responsible development and management of these assets requires a balance achieved through the appropriate workflows and oversight necessary to ensure assets are brought into the fold of normal governance for sharing across the enterprise.

While not their only responsibility, the growth in the number of chief data officers nationally and worldwide certainly is symptomatic of the need for responsible and intentional data management, including governance. So, at a basic level, there is a fundamental need for these management practices to be put in place in the first place, and this Brave New World of self-service and proliferation of enabling tools simply makes this need more acute.

Developing a mature view of data management is fundamental. What needs to change is the acknowledgment that rapid data asset development can be fundamentally positive if governed well, which may require new or updated governance processes to ensure that success.

abittermann@fusionalliance.com
linkedin.com/abittermann/
@BittermannBI