Metadata Management as a Strategic Data Imperative

Gal Ziton, Octopai
Dan Yarmoluk, Atek Access Technologies

The data management landscape is highly complex, with Big Data and NoSQL technologies creating a variety of languages and skills that most organizations do not have internally. Pig, Hive, Hadoop, Spark, Scala, MongoDB, and HBase, as well as new emerging environments, were created due to a lack of skills amongst organizations, and yet data orchestration becomes increasingly complex. As we move to a Big Data, algorithmic, machine learning, and artificial intelligence world, most companies do not have the means to extract meaningful insights in their traditional data warehouse due to siloed proprietary technologies. The emergence of an easy-to-use, centralized metadata governance solution like Octopai is essential to move efficiently throughout the data journey and data value chain. Metadata management is the underlying skill that enables organizations to focus on discovering new patterns, predicting future events, and simulating different scenarios that is easy, efficient, and transparent, but must be treated with the foundational and strategic importance of the modern data enterprise. Evidence-based, data-driven insights and digital transformation are imperative for the next-generation organizations. In order to harness this change, glean new insights, and make decisions faster, companies must look at their data operational processes and look for efficiencies. Metadata management has strategic, efficiency, financial, and regulatory implications that must be addressed to break data value bottlenecks.

Data is growing at an astounding rate. Humanity is moving to a universe of sensors, or "resolution revolution", where more granular data is being produced at a rate than can be truly analyzed and acted upon. As we move into a world of IoT, Industrial IoT, social media, people analytics, and ultimately to automation, machine learning, and artificial intelligence, the hot topics seem to focus on the end of the value chain of data science rather than the bulk of the process in data management. The hype in the "what could be" for artificial intelligence supersedes the effective and tangible roadmap to get there. The industry must consider alleviating the bottlenecks of data management and leverage metadata management as a strategy baked into product management and innovation at the organizational level to unlock the value of data science.

This paper outlines the foundations of big data, new programming skills and changing environments, time required for data preparation and value in managing metadata management as a system or strategy directly correlated to success in business analytics, BI and data science. Metadata management will be defined as the factory in productionizing data science and the defining factor in yielding meaningful insights.

What is Big Data?

The term Big Data means different things to different people. What we call Big Data is the data measured in Terabytes, and not Gigabytes. Well, how much space is one Gigabyte? We need to understand how storage is measured.

A bit, the smallest storage space, is equivalent to 1=0, and off =0. A byte, is 8 – 28 = 256 and 16 bits = double-byte = 256 X 256 = 65,536 bits. This goes forward to storage measurements using the international system of units, every 1,000 bytes (10^3) is given a distinct name, being Kilobyte (10^3); Megabyte (10^6); Gigabyte (10^9); Terabyte (10^{12}); Petabyte (10^{15}), Exabyte (10^{18}), Zettabyte (10^{21}) and Yottabyte (10^{24}).

The industry has seen that traditional databases struggle to keep up at the Terabyte level. However, what do these numbers mean? In a post from Business Insider in 2013, the company announced that Facebook has about 1.15 billion users and each user having over 200 pictures.

That amounts to about 230,000,000,000 imagines, and if each image is about 3MB, then the picture database size at Facebook is 690,000,000,000,000,000 bytes or 690 Petabytes or PB. That same post also indicated that this size of the image storage increases by 350 million photos each day. Over a decade ago, it was published that Google processed 24 Petabytes (compare that to the 50 petabytes of all of the written works of mankind from the beginning of recorded history in all languages) of data per day, all suggesting the volume categorization of Big Data.

In addition, the velocity and processing issue is pushing new highs, and in 2014 ACI information group announced that every minute:

- Facebook users share nearly 2.5 million pieces of content
- Twitter users tweet nearly 300,000 times
- Instagram users post nearly 220,000 photos
- YouTube users upload 72 hours of new video content
- Apple users download nearly 50,000 apps
- Email users send over 200 million messages
- Amazon generates over $800,000 in online sales

We are dealing with data growth, and the 4 “V”s or characteristics of Big Data are

- Velocity – Data grows very fast, such as 24 PB per day (Google)
- Variety – Data is in all kinds of formats, both structured and unstructured
- Volume – Data size is humongous, such as 1.5 PB (Facebook);
- Veracity – Data is not clean nor reliable

We need to be able to process humongous amount of information from all kinds of sources at a fast pace. Companies that are struggling to store, manage, and analyze the data they do have nonetheless take advantage of the emergence of data science. The dilemma facing many organizations is how to store these types of data with traditional computer architectures, disk storage, and relational databases. The answer is that you can't, and a totally different approach to storing and processing the information is required. Welcome to the world of Big Data, parallel processing, and NoSQL.

With the inability to scale-up to meet the needs of ever-growing data, data growth required a scale-out approach due to coordination overhead, communication delay, remote file access, remote procedure calls, sharing disks, sharing memory, etc – all issues that any distributed system must deal with. A distributed system is multiple computers that work together cohesively with parallel programming support, distributed (or parallel) operating system, distributed file system, and distributed fault tolerance.

For example, Hadoop is a reliable, fault-tolerant, high performance distributed parallel programming framework for large scale data written in Java. Cloudera and Hortonworks have Hadoop environments to have a UI framework, SDK, workflow, scheduling, metadata, data integration, scheduling, metadata, languages, compliers, read/write access, and coordination. Hue, Oozie, Pig, Hive, Hbase, Flume, Scoop, and Zookeeper are evolving into other preferred environments and languages.

Interestingly enough, Pig was created at Yahoo as an easier programming language to accommodate the limited skills within their company, using Hadoop to focus more on analyzing large data sets and spend less time having to write mapper and reducer programs. Like actual pigs, who eat almost anything, the Pig programming language is designed to handle any kind of data. The takeaway is that it is very difficult to automate or keep this orchestration of Big Data moving and flowing in the right direction, namely to the business or organization to answer

questions. It must therefore be deduced that managing Big Data orchestration is critical to gleaning insights. Gaining insights or exploring patterns with unsupervised learning, machine learning, and deep learning is the next frontier, but we must constantly flow data to leverage those techniques.

The players that leverage Big Data – Google, Facebook, Apple, and Amazon – are well-known. They have so much data, they can provide hyper-customized deals and have a deeper understanding of their customers and business. They have the skill sets to harness and deploy productionized data environments for near real-time changes in this ever-changing marketplace. In many cases, they have only scraped the surface with structured data at this rate, but will continue to make advances as they automate missing data and impute with success. However, if this is the case, the rest must look to data preparation not as a major bottleneck, but as a strategic and fundamental imperative in feeding Big Data, data science, and the business.

Data preparation as most time-consuming task

Data scientists spend 60% of their time on cleaning and organizing data and 19% of their time on collecting data sets, meaning data scientists spend around 80% of their time preparing and cleaning their data for analysis – what I refer to as the "black hole". The number is about the same for their least enjoyable parts of data science.

The black hole is a massive amount of time met with an equal proportion of disdain by those doing that work. And again, most of the love is given to the shiny presentation layer, the end-results of analytics or visualization. However, we must be able to serve the data to the business more quickly under an increasingly stressed data organization with Big Data volume and velocity.

The answer lies in metadata. David Lyle, VP of Business Transformative Services at Informatica, wrote that, "The difference between success and failure is proportional to the investment an organization makes in its metadata management system." He goes on to state that to automate the process, reuse the cleansed data or production through careful preparation and annotation, and semantically define all data to make it useable and trustworthy. These small, repeating patterns further bog down the machine towards information production. If we harness the power of metadata, we construct the factory with metadata as the solid foundation that all information tools and data discovery can stand upon.

Metadata management as core competency to data-driven organizations

With Big Data increasing and the time-consuming task of data preparation, the industry is looking for better ways to improve efficiency in translating data into meaningful information. With that said, when analyzing the data journey and the constituents data serves, metadata is the common recurring theme that enables an organization that replaces “data wrangling” with data discovery. One has to look at the opportunity costs of a limited time and value data scientists or data analysts spending his or her time on data preparation. Data scientists or computer programmers are considered some of the highest paid workers, yet we've allocated their time to tasks that truly do not affect the bottom line. The bulk of their time, as we have seen, are spent on potentially getting to those bottom line questions. BI groups put in lots of valuable time and effort to manually discover, find, and understand metadata while the business loses precious time to market.

Metadata management must be a core competency, a place of innovation and of strategic importance. Organizations must regularly audit their metadata and data preparation processes in order to compete. In addition, data governance and stewardship are now considered vital. Those who see governance and security as places of innovation will lead the next-generation of organizations.

Enabling visibility and control of metadata that is scattered across the enterprise data landscape is key to success. That requires tools and management processes that reduce weeks of manual processes to minutes and enable the business to move faster, allowing self-service capabilities for exploration. First, the process should allow to share and reuse the data and knowledge about the data. Metadata scanning can automatically gather from a wide variety of sources, including ETL, databases and reporting tools. Second, metadata should be stored and managed in a central repository to enable share and reuse data sets, data definitions, metadata, and master data. Third, metadata understanding, with smart algorithms, should model and index all types to enable the user to quickly locate and understand cross connections across connections.

Fourth, a central tool, namely a smart search engine using hundreds of crawlers to enable searches of all metadata to present results in seconds, should be in place. Finally, a visual mapping tool is needed to create full lineage of the data journey as it flows through multi-vendor systems to assist in the data context to continually improve the management process.

David Stodder’s recommendations in Improving Data Preparation for Business Analytics in TDWI’s Q3 2016 Best Practices Report serves as a constructive guide on the overarching themes to create an environment for more productive users. Make shortening time to achieving business insight a data preparation improvement priority. Focus on reducing the time data preparation

takes by using innovative technologies and methods to ensure higher levels of repeatability through shared data catalogs, glossaries, and metadata repositories.

While all the above-mentioned reasons are good reasons to productionize metadata strategies with strong tools and thought leadership, the General Data Protection Regulation (GDPR) will come into effect in May of 2018. This means that all businesses and organizations that handle European customer, citizen, or employee data must comply with the guidelines imposed by GDPR. The regulation will require consent management, data breach notifications, the potential appointment of a data protection officer, privacy by default processes, privacy impact assessments, and an understanding of the rights of data owners.

Consent management will require that the process data under the GDPR be governed by the consent of the owner, and metadata enables the registration and administration of consent. Data breach notifications require metadata to provide information on the creation date of the information, the location and name of the database comprised, and when the data breach took place. The metadata repository will be the main source of information for processes and measures for protection of personal information. Privacy, by default, will require that only information that meets specific criteria for the overall purpose of the business be collected, and metadata can assist in ensuring specific data processing are performed according to specific guidelines. Organizations will be required to perform Privacy Impact Assessments (PIA). Documentation will be greatly enhanced and accessible for PIAs with metadata and its governance. Lastly, the rights of data owners will increasingly become an issue and the time to start for 2018 is now to mitigate any business risk. Rising to the level of metadata management and process internally will allow one to deal with the complexity of compliance. As anything, if you harness change and that complexity, it can become of source of innovation and market leadership.

Over the years of discussions with various companies across various verticals, listed below is illustrative of the feedback from thought leaders as well as power users of metadata:

Claudia Imhoff, Ph.D., President Intelligent Solutions, Inc. says, "*A centralized metadata solution is needed in order to be a successful hunter-gatherer in the new complex world of data analytics.*"

Bala Venkatraman, BI Expert states a metadata strategy and tool "transformed metadata from a nice-to-have to a must have," and continues, "*Democratization of the data landscape yields an overall increased efficiency for any BI team*."

Revital Mor, Head of BI for Harel Insurance, states that a metadata strategy and tool is an "*easy way for mapping data flow from end to end. It's an important and necessary for any organization that values data accuracy*".

Consider some typical scenarios and tangible outcomes below as a summation and case for action:

Business Challenge #1
A business user discovered inaccurate information on revenue reports and loses precious time waiting for the BI group to respond.

BI Challenge #1
The BI group must reduce the amount of time and resources required to locate the source of inaccurate reports and correct the problem while maintaining accuracy.

How BI worked prior to metadata tools and strategic imperative #1
In order to identify the source of the problem, BI groups had to manually trace the data to discover and understand all the hops that the data went through to land in a specific report. They had to analyze which database tables and ETL processes were involved in and affected by the particular inaccuracy and correlate the metadata to the other reports that were affected – a very expensive, error-prone and time-consuming process.

Metadata tools and strategic solution #1
Enabling visibility and control of metadata, it can assist the information team to discover all related ETL processes and tables use to produce a report in a matter of seconds. It can compare multiple reports to show specific metadata items, fix the problem with all the metadata centralized in one place with a search that is simple, accurate and effective.

Value to the organization #1
Cut the metadata search and tracing costs by more than 50% with a faster time-to-market than initially estimated with little to no business disruptions or production issues when fixing the problem.

Business Challenge #2
New regulation calling for companies for change HR headcount reporting from monthly to daily must be implemented accurately and quickly, without business disruptions.

BI Challenge #2
The BI group must quickly and accurately locate the formula calculating the headcount in reports, database tables, ETL processes, views and stored procedures – without additional manpower. The BI group must also understand all possible impacts of making the change.

How BI worked prior to Metadata tools and strategic imperative #2
In order to locate the specific calculation, BI had to manually map out the entire BI landscape. With metadata scattered all over the place, this is a very time consuming and therefore costly process, which can require multiple development and design cycles before completing the change.

Metadata tools and strategic solution #1
Enable BI to make the change easily in a fraction of the time with cross-platform metadata search engine. Easily and quickly map out the entire BI landscape with a detailed visual map showing where the calculation is being made in every system; discover the exact ETL processes and tables that make up a report; show all objects related to the calculation instances throughout the entire landscape on one screen; and implement the change quickly and accurately.

Value to the organization #2
Cut the project costs by more than 50%; complete project faster than originally estimated; elimination of overtime hours or additional staff; zero business disruptions or production issues.

Wiriten By:

Gal Ziton is CTO & Co-Founder of Octopai

&

Daniel Yarmoluk is Director of IoT and Data Science at ATEK's All Things Data Podcast and VertiAI
