Metadata Blending: Automating Data Lineage Views

Dr. Geoffrey Malafsky

CEO, Technik Interlytics LLC,

Chief Scientist, The Bloor Group

20181106

Business Intelligence (BI) cannot fulfill its namesake (i.e. intelligence) without validated information about the data used for reports and visualizations. Simply pushing data of any kind and any origin into a dashboard yields attractive graphs of no particular value to decision makers and analysts. Without trust in the data, which comes from knowledge and validation of all of its characteristics, it is just as meaningful to make business graphs from weekly winning lottery numbers. The intelligence arises from gleaning actionable insights.

To know the data, we need to manage its full lifecycle. This covers the birth to grave range of processing and handling. Doing so benefits three main business areas:

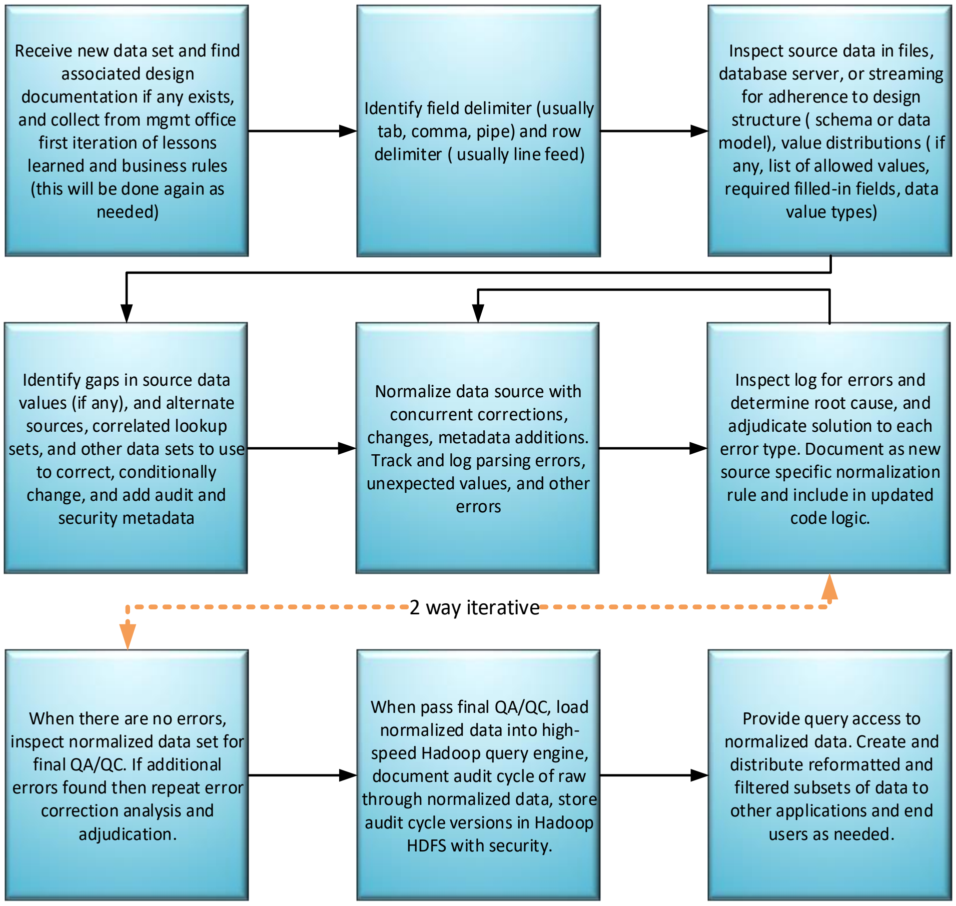
1. Analytics: usable granularity from many corporate systems merged into a single coherent data set offers greater insights for quality control, monetization, decision making, new product planning, and customer experience management.
2. Efficiency: data commonly exists in multiple systems and storage locations but with enough minor variations that it is really no longer the same. Better management can reduce duplication, eliminate conflicts, streamline processing, and enable automation such as with data pipelines. Important outcomes are stronger Quality Assurance (QA) and Quality Control (QC) with overall shortened timelines to produce BI results.
3. Laws: a spate of new laws require traceable auditing of data used for regulatory compliance and dealing with privacy and healthcare data (e.g. GDPR). Metadata is specifically described in laws as required evidence.

The main stages of a data lifecycle include the following activities A-E.

A. collecting it in a variety of formats and transfer methods
B. moving it into corporate databases
C. coordinating and normalizing variations
D. merging different sets
E. deploying reports and analytical products.

For each stage, we need to know characteristics such as: who supplied it; when it was received; what format it is in; how it was packaged; where it was placed; who managed the process; what business rules were used; errors that occurred; which versions used; degree of trustworthiness; pedigree; and other custom categories important to each organization.

However, this becomes more complicated when we realize the lifecycle does not proceed in an orderly linear fashion through the stages. Rather, as each stage progresses, we usually discover that some data is missing or wrong forcing us to go back and get new data or new specifications (e.g. schemas). This is highlighted in a real project flow chart from my Big Data consulting in large organization financial audit processing.


Receive new data set and find associated design documentation if any exists, and collect from mgmt office first iteration of lessons learned and business rules (this will be done again as needed)
Identify field delimiter (usually tab, comma, pipe) and row delimiter (usually line feed)
Inspect source data in files, database server, or streaming for adherence to design structure (schema or data model), value distributions (if any, list of allowed values, required filled-in fields, data value types)
Identify gaps in source data values (if any), and alternate sources, correlated lookup sets, and other data sets to use to correct, conditionally change, and add audit and security metadata
Normalize data source with concurrent corrections, changes, metadata additions. Track and log parsing errors, unexpected values, and other errors
Inspect log for errors and determine root cause, and adjudicate solution to each error type. Document as new source specific normalization rule and include in updated code logic.
2 way iterative
When there are no errors, inspect normalized data set for final QA/QC. If additional errors found then repeat error correction analysis and adjudication.
When pass final QA/QC, load normalized data into high-speed Hadoop query engine, document audit cycle of raw through normalized data, store audit cycle versions in Hadoop HDFS with security.
Provide query access to normalized data. Create and distribute reformatted and filtered subsets of data to other applications and end users as needed.

Notice that significant data quality issues and processing errors can occur that cause iteratively redoing steps. Indeed, the lifecycle stages can easily be performed like A-B-A-B-C-A-B-C-D-E-C-D-A and so on. Each time this occurs, the tracking information (as metadata) must be updated. In real environments, this rigorous content management rarely is done due to time and personnel shortages, and the pressure to deliver results. The outcome is missing or inaccurate documentation and an increasing level of unknowns about the data's processing and handling. Compounding this problem is the data spread over multiple processing systems (e.g. ETL engines) managed by different groups each with their own business rules.

Collating and blending all of this information to make trustworthy, enterprise scale BI products has been laborious and time consuming. Each tool's information must be extracted, correlated and mapped to others, and a final end to end map and timeline of processing made. Doing so one time is difficult. Doing so to keep up with all the changes across all the groups is overwhelming.

Fortunately, we have many new technologies coming from ubiquitous networks, inexpensive large storage, inexpensive and compact computing power, and a large number of free packaged software libraries providing sophisticated processing functions. Additionally, there are new business models for outsourced hosting, management, and processing with flexible prices per scaling. We can exploit these to unify disparate data at large volumes with faster lifecycles and lower Total Cost of Ownership (TCO). This includes Big Data, Cloud, Artificial Intelligence (AI), and X-as-a-Service providers (e.g. X= S for Software, X=P for Platform).

So, we are poised to use this confluence of new technologies and service providers to make the difficult problem easy to solve. Use Big Data with AI in the Cloud with secure networking and we can collect everything, organize it, search rapidly, and get answers in seconds with minimal work. Yes, same promised land but different technology acronyms from the multiple versions proffered over the past 20 years.

To see if this time is really our win, we have to explore a bit deeper. To start with, we have to define the characteristic information and how we represent it. This is the realm of metadata. The usual definition of metadata is 'a set of data that describes and gives information about other data'[1] or more succinctly 'data about data'. While correct, this is not a very practical definition. Metadata is information that augments something else, usually what we consider the data object. The distinction between metadata and data depends on context and governance. We can include fields in either a data model or a metadata schema. It is up to our use case. So, a data model with instance records about books can have the core data set include a field for published date, or not. If not, we can construct a metadata schema to describe each data record (which does not tell us the published date) and have the published date in the metadata record linked to the corresponding data record.

Metadata is extensively used for objects like documents, images, and audio files stored on computers. In fact, there are layers of metadata in these cases since each operating system maintains metadata as well. There are formalized schemas defined by industry standards such as: Learning Object Metadata (LOM)[2] , Dublin Core[3], and Data Documentation Initiative[4].

[1] https://en.oxforddictionaries.com/definition/metadata, accessed 20181106

[2] IEEE 1484.12.1-2002 - IEEE Standard for Learning Object Metadata, https://standards.ieee.org/standard/1484_12_1-2002.html

[3] Dublin Core Metadata Initiative, http://dublincore.org/

[4] Data Documentation Initiative, https://www.ddialliance.org/

However, here we are concerned with the data lifecycle of processing and handling so while all of this metadata is valuable, we are concentrating on the metadata describing how the data was moved and transformed. We want to know all the changes and locations the data experienced from when it first arose in the organization through using it for BI products. This is its lineage and includes myriad dispersed information on ingest, transformations, mappings, and schema variations.

Fortunately, this is what much of the new technology does well. We want to exploit the technology to eliminate the bulk of the manually intensive and complicated work needed to gather, map, and validate metadata from many integration tools, databases, and storage servers. We can connect systems even in hybrid cloud architectures and automatically extract their processing and file handling metadata. We can use Machine Learning to automatically refine mapping fields despite semantic inconsistencies. We can use cluster-based search engines to provide rapid results from very large amounts of distributed storage.

Doing so will satisfy all three of our original business needs: analytics; efficiency; and legal. You should gain significant tangible business benefits in the short-term making this one of the easier business cases to propose to your financial managers. With your new capability and insights, business goals long desired but too difficult to attain will become feasible. These include key endeavors such as enabling organizational transformation, protecting investments during modernization, demonstrating the importance of strong BI, and inter-group collaboration.