Data Wrangling: The Challenging Journey from the Wild to the Lake

Ignacio Terrizzano, Peter Schwarz, Mary Roth, John E. Colino
IBM Research
650 Harry Rd
San Jose, CA 95120
{igterriz, pschwarz, torkroth, jcolino}@us.ibm.com

ABSTRACT

Much has been written about the explosion of data, also known as the "data deluge". Similarly, much of today's research and decision making are based on the *de facto* acceptance that knowledge and insight can be gained from analyzing and contextualizing the vast (and growing) amount of "open" or "raw" data. The concept that the large number of data sources available today facilitates analyses on combinations of heterogeneous information that would not be achievable via "siloed" data maintained in warehouses is very powerful. The term *data lake* has been coined to convey the concept of a centralized repository containing virtually inexhaustible amounts of raw (or minimally curated) data that is readily made available anytime to anyone authorized to perform analytical activities. The often unstated premise of a data lake is that it relieves users from dealing with data acquisition and maintenance issues, and guarantees fast access to local, accurate and updated data without incurring development costs (in terms of time and money) typically associated with structured data warehouses. However appealing this premise, practically speaking, it is our experience, and that of our customers, that "raw" data is logistically difficult to obtain, quite challenging to interpret and describe, and tedious to maintain. Furthermore, these challenges multiply as the number of sources grows, thus increasing the need to thoroughly describe and curate the data in order to make it consumable. In this paper, we present and describe some of the challenges inherent in creating, filling, maintaining, and governing a data lake, a set of processes that collectively define the actions of *data wrangling*, and we propose that what is really needed is a *curated* data lake, where the lake contents have undergone a curation process that enable its use and deliver the promise of ad-hoc data accessibility to users beyond the enterprise IT staff.

Categories and Subject Descriptors

H.1.1 [**Systems and Information**] Value of Information.
H.3.2 [**Information and Storage Retrieval]** Information Storage.

General Terms

Management, Documentation, Design, Legal Aspects.

Keywords

Data lake, data wrangling, data curation, data integration, metadata, schema mapping, analytics sandboxes

7th Biennial Conference on Innovative Data Systems Research (CIDR '15) January 4-7, 2015, Asilomar, California, USA.

1. INTRODUCTION

We have all been inundated with facts and statistics about the data deluge that surrounds us from consumer-generated and freely available social media data, from the vast corpus of open data, and from the growing body of sensor data as we enter the era of the Internet of Things [34]. Along with the bombardment of statistics about this data deluge, there appears to be a *de facto* acceptance that there is critical new business value or scientific insight that can be gained from analyzing the zettabytes of data now at our fingertips, if only enterprise data can be freed from its silos and easily mixed with external "raw" data for self-serve, ad-hoc analysis by an audience broader than the enterprise IT staff.

Financial institutions, for example, now speak of offering personalized services, such as determining if a client is exposed to legal risks due to the contents of his or her portfolio. Such analysis requires access to internal data, external news reports and market data about the companies that make up the portfolio, as well as publicly available regulatory information. As another example, a Fortune 1000 information processing company that provides outsourcing services to manage their clients' data processing systems would also like to offer them analytic sandboxes and customized access to demographic data and economic data by geography, all of which is available from sources like the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. As yet another example, IBM Research itself has recognized that gathering a large body of contextual data and making it readily accessible to its research scientists is strategically important for innovation [12].

IBM estimates that a staggering 70% of the time spent on analytic projects is concerned with identifying, cleansing, and integrating data due to the difficulties of locating data that is scattered among many business applications, the need to reengineer and reformat it in order to make it easier to consume, and the need to regularly refresh it to keep it up-to-date [5]. This cost, along with recent trends in the growth and availability of data, have led to the concept of a capacious repository for raw data called a data lake. According to a recent definition, and as shown in Figure 1, a data lake is a set of centralized repositories containing vast amounts of raw data (either structured or unstructured), described by metadata, organized into identifiable data sets, and available on demand [5]. Data in the lake supports discovery, analytics, and reporting, usually by deploying cluster tools like Hadoop. Unlike traditional warehouses, the format of the data is not described (that is, its schema is not available) until the data is needed. By delaying the categorization of data from the point of entry to the point of use [10], analytical operations that transcend the rigid format of an adopted schema become possible. Query and search operations on the data can be performed using traditional database technologies (when structured), as well as via alternate means such as indexing and NoSQL derivatives.

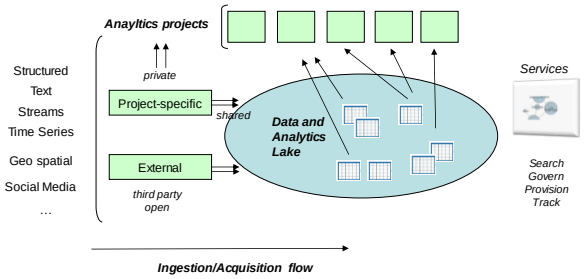

Figure 1: Data Lake Logical Architecture

While this definition of a data lake is not difficult to understand, it originates from an essential premise: the data in the lake is readily available and readily consumable by users who have less technical skill than traditional IT staff. This implies that the data lake somehow relieves such users from the well-defined but tedious and technical tasks that are required to prepare data associated with a traditional data integration platform and data warehouse architecture, such as defining standard data models, extracting and transforming data to a common data model, performing cleansing, validation, and error handling, and documenting the process [5].

This premise, however, is in stark contrast with a repository of *raw* data. As Gartner recently noted, there exist pitfalls in creating and using an enterprise-level data lake [28]. Gartner portrays a data lake as a "catch all" repository and, as such, cites problems inherent from data quality, provenance, and governance, all of which have been historically associated with traditional data warehouses. We argue here that a "raw" data lake does not enhance the agility and accessibility of data, since much of the necessary data massaging is simply postponed, potentially to a time far removed from the moment that the data was acquired. And, in addition, we believe a data lake introduces legal ramifications, from adherence to licensing terms to determination of liability and ownership of derived data. A raw data lake places the burden of such tasks squarely on the data consumer. The steps associated with a traditional data integration platform exist for a reason; data in its raw format is rarely immediately consumable for use in a specific application. For example, economic data from the U.S. Bureau of Labor Statistics [4] (BLS) represents geographic regions using its own set of codes, without which the statistical data is difficult to interpret and use in a meaningful way.

The term data curation is increasingly being used to describe the actions necessary to maintain and utilize digital data during its useful life-cycle for current and future interested users.

> *Digital curation involves selection and appraisal by creators and archivists; evolving provision of intellectual access; redundant storage; data transformations; and, for some materials, a commitment to long-term preservation. Digital curation is stewardship that provides for the reproducibility and re-use of authentic digital data and other digital assets. Development of trustworthy and durable digital repositories; principles of sound metadata creation and capture; use of open standards for file formats and data encoding; and the promotion of information management literacy are all essential to the longevity of digital resources and the success of curation efforts.* [1]

Given the challenges present when working with vast amounts of raw data, particularly upon first use, we propose that what is needed to provide self-service, agile access to data is a *curated* data lake.

In this paper, we present a number of challenges inherent in creating, filling, maintaining, and governing a curated data lake, a set of processes that collectively define the actions of *data wrangling* (see Figure 2). These are challenges not only reported by our customers, but are also challenges that we face ourselves in creating a work-in-progress data lake to be used both internally for IBM research staff as well as in client engagements.

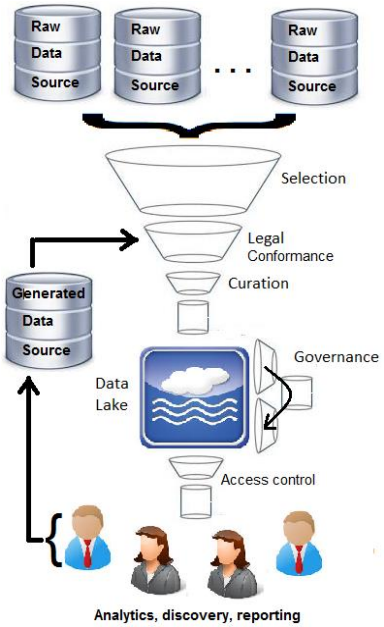

Figure 2: Data Wrangling Process Overview

We begin by describing in Section 2 the motivation for the creation of a data lake at IBM Research, as well as a high-level architecture. This allows us to draw on our experiences in order to contextualize the challenges presented starting in Section 3. Section 3 describes concerns around data procurement, focusing on data selection, obtainment, and description, paying particular attention to issues around licensing and governance. We then continue in Section 4 by describing the difficulties in readying data for use, a process called *data grooming* which encompasses data massaging and normalization. Section 5 details the concerns around usage of data in the lake, and the challenges around ensuring compliance by authorized users. Section 6 provides a brief description of data preservation, a key concern of data maintenance. Finally, we summarize and conclude in Sections 7 and 8 by introducing related work and description of future work.

2. THE IBM RESEARCH DATA LAKE

The IBM Research Accelerated Discovery Lab [12] was started in late 2012 to support multiple, independent analytic projects that may involve participants from several institutions. It is currently supporting over a dozen projects from several domains, including, for example, a project to use literature-based discovery over medical journals and patent databases to support cancer research, and a project that tracks the cost of water in different geographies

by analyzing public utility reports and news articles to define a global water index.

A key service provided by the lab is a lake of contextual data that can be used across different research projects. For example, data provided by the government agencies we have been using as examples in this paper can supply location-specific demographic data, economic data grouped by location and industry, climate data, SEC filings and the like, all of which are useful in many contexts. Important contextual information available from other sources includes worldwide patent data, medical journals, and many kinds of geo-spatial data.

Figure 3 shows a high level overview of our data lake architecture. The lake is intended to support over 500 researchers across multiple research labs. As shown in the figure, IBM researchers develop applications that run in both internal and external cloud environments and require access to data stored in the data lake. Because a firewall between the cloud environments prevents processes on the external cloud from accessing the internal cloud, we have chosen a master/slave architecture for the data lake storage, with a pipeline to transfer data on an as-needed basis from the master lake located in the internal cloud environment to the slave lake located in the external cloud.

As will be described in Section 3.1, compliance with data licensing terms and other controls on data usage is a critical and nontrivial exercise, and failure to do so can introduce significant liability for an enterprise. To assist in this task, we have developed a governance tool that tracks requests for acquisition of new data for the lake and access requests for data already present in the lake. The tool collects input from all stakeholders in the governance process recording then in a secure system-of-record, along with all relevant licenses, wrangling guidelines, usage guidelines and data-user agreements. The tool runs in the internal cloud, and is accessed via a proxy from the external cloud.

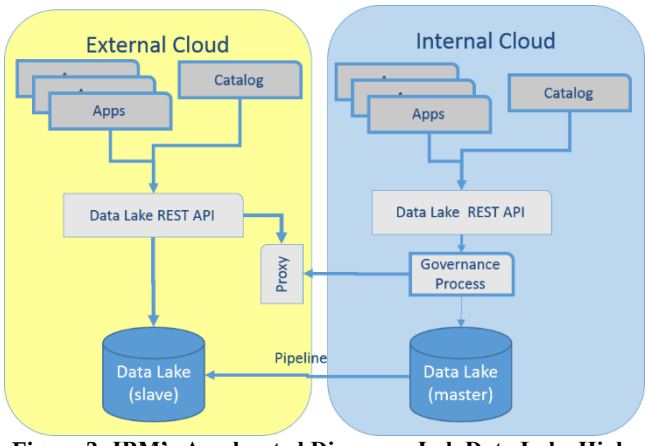

Figure 3: IBM's Accelerated Discovery Lab Data Lake High-Level Architecture

Figure 4 shows an overview of a dashboard displaying the current status of the lake. The dashboard shows a point-in-time status of data as it moves through the various steps described below. A data set represents a collection of logically related data objects (such as files or tables) that correspond to a single topic. For example, an average price data set [2] available from the Bureau of Labor and Statistics includes a set of tables for gasoline prices, food prices and household energy prices. At the time this snapshot was taken, the dashboard shows that 75 data sets have been considered in categories such as biomedical, social and economic data for inclusion in the data lake. 59 data sets are still in various steps of the process described below, and 16 have completed the process and are available for use in the lake.

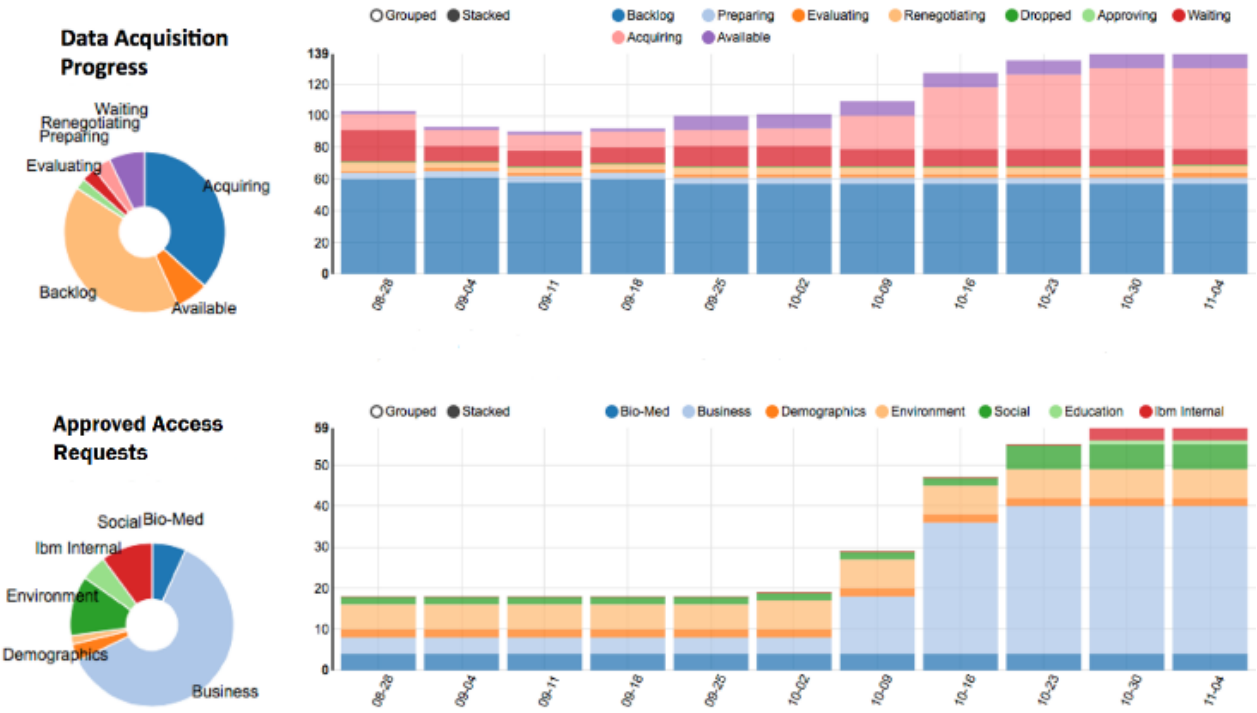

Figure 4: Data Lake Dashboard

3. PROCURING DATA

Data procurement is the first step performed by *data wranglers;* it describes the process of obtaining data and metadata and preparing them for eventual inclusion in a data lake. The potential to achieve new insights from Big Data depends in part on the ability to combine data from different domains in novel ways. In each domain, however, a plethora of data is often available, frequently from multiple sources. Given a particular domain, the data wrangler's first task is likely to be the identification of the specific sources and data sets that will be of most value to the enterprise. This can be quite challenging.

For example, basic information about the U.S. economy is available from the Bureau of Labor Statistics [4], the Bureau of Economic Analysis [3] and probably from several other sources as well. Even a single source may offer a wide variety of similar information. The National Climatic Data Center [24], for example, offers climate data as recorded by land stations, weather balloons, satellites, paleo-climatological readings, and several other options. Data may be provided at different levels of granularity (in time or space), for different time periods or locations, and in different formats. The more a data wrangler is aware of how the data will ultimately be consumed, the better he or she will be able to make good choices about which data to wrangle, but the ad-hoc nature of Big Data analysis means the wrangler must anticipate these needs, rather than react to them. Another consideration in selecting a data source centers on whether the provider supplies data in bulk, or only a few items at a time, such as in response to a narrow query. The latter type of source may provide very valuable information, but performance may preclude its use in Big Data analytics. Furthermore, patterns in the queries issued by an enterprise may reveal to the supplier information the enterprise would prefer to keep in confidence.

Beyond the utility of a data set, the wrangler must also consider the terms under which it is made available and the mechanisms needed to obtain it. These are the topics of the following two sections.

3.1 Vetting Data for Licensing and Legal Use

Once the data to be obtained has been identified and selected, the next step is to determine the terms and conditions under which it may be licensed. Often, license terms are available on a web page, but locating the license that applies to a specific data set is not always easy, and once terms are located, the typical data scientist is not qualified to understand them. To understand a license, the reader must be able to discern:

- What data is being licensed, and how or where is it being made available?
- Can the data be obtained at no cost, or is there a charge associated with access? If there is a charge, how is it applied (e.g. one-time, periodic, per data item accessed etc.)?
- What kinds of use are permitted/prohibited by the license?
- What risks are incurred by the enterprise in accepting the license?

The latter two questions are closely related, and often difficult to answer. Restrictions on the use of data abound. For example, the Terms of Use for the LinkedIn Self-Service API [22] include the following clause:

> *You … cannot use our self-service program if your Application targets current or potential paying customers of LinkedIn products or people engaging in activities related to those products—in other words, Applications used for hiring, marketing, or selling.*

By accepting such an agreement, an employee inevitably exposes the enterprise to a certain level of risk. Many of the terms therein (e.g. "potential paying customers of LinkedIn") are not precisely-defined, and while an employee may believe that their intended use of the data does not violate this license, there is always a chance that a lawsuit may be filed and a court may disagree. Furthermore, ungoverned redistribution of the data, even within the enterprise, greatly increases the likelihood that some users of the data may violate the license terms.

Other licenses constrain data use in other ways. Some licenses limit or prohibit retention of data. Many licenses require data consumers to cite the source of any data displayed by an application.

Other sources of risk arise even when the usage of the data adheres closely to the license terms. Data from third parties may contain errors, and licenses typically include a disclaimer that limits the provider's liability for damages caused by such errors. If the data is subsequently used by the enterprise in a manner that affects their customers or clients, any liability for errors must either be passed on to the customer or accepted by the enterprise as a risk. Similarly, licenses may contain terms that indemnify the provider against damages due to the accidental disclosure of Sensitive Personal Information (SPI). For certain kinds of data, notably health care data, accidental disclosure can result in very large fines.

Still another source of risk that may be incurred when an enterprise uses third-party data centers on issues of copyright. Many suppliers of data distribute or redistribute material subject to copyright and control the subsequent use of that data, either through license terms that prohibit further redistribution or by constraining the licensee to redistribute the data subject to specific terms. For example, the Creative Commons Attribution – ShareAlike 3.0 License [35], under which Wikipedia is distributed, provides free access to Wikipedia content, but specifies that material obtained under this license may only be redistributed under the same (no-cost) license. Such restrictions may be incompatible with an enterprise's business model. Furthermore, the same restriction applies to *adaptations* of the original data, or *derived works*. The Wikipedia terms state:

> *If you alter, transform, or build upon this work, you may distribute the resulting work only under the same, similar or a compatible license.*

Risk arises because it is often unclear whether a particular use of data does or does not constitute a derived work. For example, consider a process that uses text annotators to analyze copyright data licensed under such terms, and builds a knowledge graph to represent the extracted information. Is the knowledge graph a "derived work" that must be distributed free of charge? Ultimately, the answer to such questions may have to come from a court, and different jurisdictions may answer the question in different ways.

Certain special classes of data introduce additional risks. If data that contains, or might contain, Sensitive Personal Information (SPI) is placed in the data lake, controls must be in place to ensure that it is only used for legal and authorized purposes, whether the data is internal to the enterprise or acquired from third parties.

In a global enterprise, movement of data across international boundaries introduces yet more complexity Export controls may prohibit transmission of certain kinds of sensitive data, privacy laws vary from country to country, and data may be licensed under different terms in different places. For example, the SNOMED medical terminology system can be licensed free-of-charge in countries that are members of the International Health Terminology Standards Development Organisation [19] (IHTSDO), but requires a fee to be paid in other countries.

Lastly, data providers often make a distinction between research or personal use and commercial use of the data they distribute. Even many so-called "open" data sites allow their data to be used freely for research, but require a special license to be negotiated for other uses. For example, the City of Boston [6] restricts the use of their open data by businesses as follows:

> *User may use the City's Data in the form provided by the City for User's own internal business or organizational purposes and for no other purpose.*

Similarly, Yelp's [36] terms of service contain an outright prohibition on commercial use of the data in their RSS feed. As in other cases mentioned above, the lines between permitted and prohibited uses may be unclear and subject to interpretation.

What is needed to manage the various risks associated with third-party data and prevent the data lake from becoming a data swamp is a *data governance* process that brings together the many stakeholders that are affected by the decision to use such data: domain experts that can determine the data's potential value, legal advisors to digest and interpret the license terms and identify other risks, and management representatives empowered to weigh the risks and benefits and come to a decision. Assuming the benefits outweigh the risks, the end result of this process is a set of guidelines that delineate how employees are permitted to obtain and use third-party data from a particular source, expressed in clear terms that a data scientist can understand and abide by.

We distinguish two sets of guidelines that are typically needed. The first set, the *wrangling guidelines*, advises the team that will obtain the data from its source about rules they must follow to comply with the license. For example, wrangler guidelines may include technical restrictions on how a provider's web site may be accessed (e.g. "only from a specified IP address, allowing at least 2 seconds between download requests"). The wranglers may also be asked to look for and exclude certain material, such as copyright images, that fall outside the scope of the license, and must be prepared to remove any material if ordered to do so.

The second set of guidelines, *usage guidelines*, must be tailored to the specific use case(s) contemplated by the enterprise, and spell out, in context, how employees may use the data while complying with the supplier's license. Any employee wishing to obtain the data from the lake must agree to these guidelines. In most cases, permission to use the data will be granted only for a limited time, after which re-approval will be needed. Similar usage guidelines are required for data internal to the enterprise that has been contributed to the lake. In either case, controls must be in place to ensure that the data is only used for appropriate purposes.

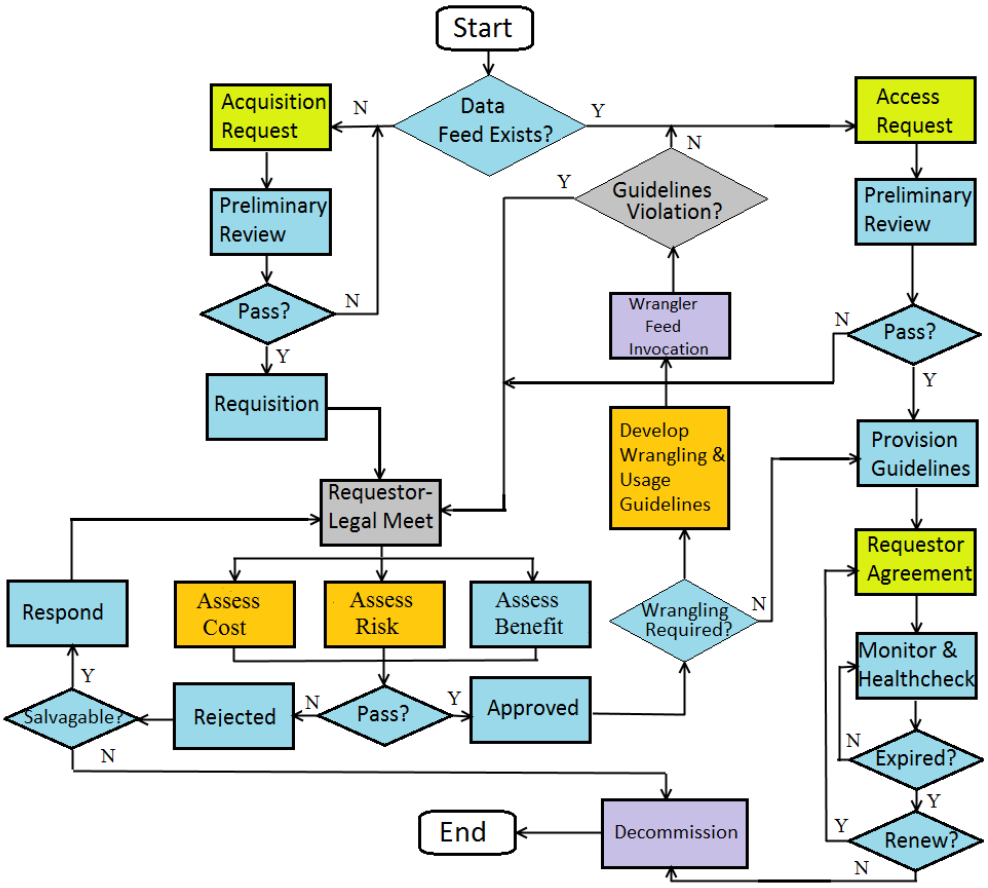

Figure 5: Sample Data Governance Process

Figure 5 depicts a high-level overview of the governance process adopted at the IBM Accelerated Discovery Lab. It illustrates a general process tailored to the two channels that require guidelines: the data acquisition channel, and the data access channel. Notice the link between the two; once wrangling and usage guidelines are in place, access requests may be processed. The process shows the lifetime of lake resident data, ending with decommission due to expiration of one of individual data access, licensing terms, or staleness (i.e., renewal not needed).

In addition to producing the wrangling and usage guidelines, the data governance process should also create a permanent record of all the information that went into the decision to use a particular data set. Given that suppliers frequently change their licensing terms, it must always be possible to ascertain exactly who agreed to what, and when.

3.2 Obtaining Data

Once data is selected and licensing terms accepted, the next challenge is transferring the data physically from the source to the data lake. As we noted above, the data sources of interest for populating a data lake are for the most part those that support bulk data download. Frequently, data sources themselves provide guidance on how to best acquire their data. Bulk data is typically delivered in files, either from a static inventory provided by the source, or through an API that dynamically constructs files in response to a query. Sometimes, the data set is so large that the only practical means of obtaining it is through the physical shipment of disk drives or tapes. This is the case, for example, if one requires a significant fraction of the National Elevation Dataset [25], a high-resolution rasterized topographical map of the United States. In most cases, however, wranglers obtain files by writing scripts that employ common tools and protocols like ftp, wget, rsync, and http that are readily available and widely understood.

In addition to the common protocols noted above, a growing number of sites support specialized protocols for transferring open data. The two most widely-used such protocols are CKAN [7] and Socrata [30]. While consumers must invest extra effort to implement these protocols, they supplement the raw data they provide with valuable metadata that might otherwise have to be collected and entered manually. We will have more to say about the importance of metadata in the next section.

Data wranglers also need to be concerned with overwhelming the source server, and in many cases data providers request adherence to procedural guidelines that aim to mitigate server overload. For example, the Security and Exchange Commission's (SEC) EDGAR [8] (Electronic Data Gathering and Retrieval) service requests that bulk ftp downloads be made between 6pm and 9am ET. For some sources such as Yelp, consumers that violate access guidelines may be subject to enterprise-wide denial of access under the licensing terms and infringement policy. Other sites, such as Wikipedia, actively manage download requests, for example by limiting the number of simultaneous connections per IP address to two.

Once the copying starts, the wrangler's job is to verify both the validity and fidelity of the download. Source providers may offer file identification checksums, file counts and sizes, or hash values for this purpose.

Of course, obtaining data is rarely a one-time event. Providers tend to frequently provide updates, causing versioning to become a concern. Furthermore, providers may provide updates in different ways. Easiest to handle are the cases where the provider either adds new files with each version, or updates the content of existing files. More complex versioning approaches entail deletion of files from one version to another, and changes in schema or file structure from one version to another. The ability to handle the different types of versioning approaches demands that the data wrangler's scripts accurately reflect the versioning strategy adopted by the provider.

3.3 Describing Data

Data alone is not useful. A data scientist searching a data lake for useful data must be able to find the data relevant to his or her needs, and once a potentially useful data set is found, he or she will want to know many things about it, e.g.:

- How is this data represented?
- Where did this data come from? (Can I trust it?)
- How old is this data?
- Can I connect this data to data I already have?

Answers to such questions require metadata of various kinds. Schematic metadata is the basic information needed to ingest and process the data, i.e., to answer the first of the four questions above. Turning again to the Bureau of Labor Statistics for an example, they distribute information about wholesale prices in the US economy as a set of related files. The schematic metadata for this data set would include information about how the data is formatted (e.g. string and/or column delimiters) and information about the schema (e.g. column names and types, foreign keys that relate the values in the various files). Unfortunately, this information is not supplied in machine-processable form. Instead, this metadata must be entered manually by the data wrangler or be re-discovered by tooling.

A second type of metadata, semantic metadata, adds meaning to data independent of its representation. It enables the data scientist to find potentially useful data sets and to answer questions like the latter three of those listed above. Placing data sets into categories and/or tagging them and their components (files, tables, columns, documents, etc.) with keywords makes searching for information easier, and information about the data set's provenance can help to resolve issues of reliability or timeliness. An advantage of obtaining data sets using "open data" protocols like CKAN is that the data returned is supplemented with a core set of important semantic metadata, including a title, description, categorization, tags, revision history, license information and more.

Semantic metadata can also help to link disparate data sets to one another. Associating elements of a data set with concepts or objects in an ontology that represents the real world can reveal connections between data sets that would not otherwise be apparent. For example, schematic metadata indicating that a numeric column actually contains postal codes, and understanding that a postal code represents a geographic region, allows the data to be plotted on a map and potentially integrated with additional geospatial data from other sources. Providing such metadata manually is a tedious process, and building better tools to do so automatically is an important area of research. [16][23].

A third type of metadata is less frequently studied and less well understood. The first user of almost any non-trivial data set discovers idiosyncrasies in the data that are crucial to understanding it and using it effectively. Continuing to use the BLS data as an example, rows representing data reported monthly use two columns to encode the year and month. Curiously, the month column contains values that range from 'M01' to 'M13'. Upon deeper

investigation, one learns that rows containing values between 'M1' and 'M12' represent data for months January through December, whereas a row containing 'M13' contains the annual average value of the statistic. A great deal of effort could be saved if subsequent users of this data set were able to consult the initial user, and become aware of this and other similar features of the data without having to rediscover them afresh.

We call this type of metadata, information about who else has used the data, what their experiences were, where they did or did not find value, and so forth, conversational metadata, and believe it to be of equal importance to the other types of metadata we have discussed [21]. The conversation that revolves around a data set among a group of data scientists bears a strong resemblance to the "buzz" that develops around a band or movie on social media, and we believe that tools for recording and searching this conversation should follow a similar paradigm. To emphasize the need for such metadata, Zeng and Qin [37] have noted that it is indispensable even if that secondary user is the same as the original one; human memory is so short that even originators must rely on their own metadata. This problem will only get worse as the amount and variety of available data increases.

4. GROOMING DATA

As we have noted, data obtained in its raw form is often not suitable for direct use by analytics. We use the term *data grooming* to describe the step-by-step process through which raw data is made consumable by analytic applications. Metadata plays a crucial role throughout this process. The first steps in the grooming process use schematic metadata to transform raw data into data that can be processed by standard data management tools. Which tools are appropriate depends on the type of data being ingested: those used for searching and manipulating genomic sequences differ from those used for geospatial data, which in turn differ from those used for tabular data.

Even focusing just on tabular data, there are a myriad of ways in which it can be represented. In some cases, the sought-after data may be embedded in PDF files or in other types of documents designed for human readability rather than processing by machine. Spreadsheets, for example, often contain artifacts like multi-column headings that do not translate directly to the abstractions of database management software. In other cases, the information may be represented in a custom format that must be converted, or at least understood, before it can be processed with conventional data management tools. Other common formats include delimited or fixed-format text files, JSON, XML and html. Even for formats like these, which were designed for automated processing, information like delimiters, field widths and data types must either be supplied or deduced, and in either case become critical aspects of the schematic metadata associated with the data set.

We also note that it is not uncommon for providers to change how their data is formatted, or to provide data for different time periods in different formats. For example, certain data collected by the National Climatic Data Center through 2011 conforms to one schema, but similar information collected from 2012 onward uses a different schema. Such changes disrupt a smoothly-running data grooming pipeline, and must be detected and accommodated.

Once the data can be ingested, normalization of certain values can facilitate further processing and enable integration with other data sets. For example, we have already noted the idiosyncratic way in which the Bureau of Labor Statistics represents dates. Integration of this data with other sources of economic data, or even something as simple as creating a graph that shows how a value (e.g. the price of gasoline) varies over time, is difficult without normalizing the dates to a standard format. Similarly, a data table must often be pivoted to permit optimal processing. Economic data from the Bureau of Economic Analysis, for example, is structured so that each year is represented by a column, with rows corresponding to specific measures, and rows containing subtotals interleaved with regular data rows. A conventional representation of this data would invert this relationship, making computation of aggregates and time series much simpler.

Throughout the grooming process, a detailed record must be kept of exactly what was done at each stage. This is particularly the case if the grooming process alters the "information content" of the data in any way. While normalization, annotation, etc. may add significant value to a data set, the consumer of the data must always be able to observe and understand the provenance of the data they rely upon.

5. PROVISIONING DATA

The previous sections have focused on getting data *into* the data lake. We now turn to the means and policies by which consumers take data *out* of the data lake, a process we refer to as *data provisioning*. It is our belief that running sophisticated analytics directly against the data lake is usually impractical. In most cases, a data scientist will want to extract a data set (or subset) from the lake and customize the manner and location in which it is stored so that the analytics can execute as efficiently as possible. However, before undertaking a possibly complex and time-consuming provisioning process, the data scientist should be able to do a preliminary exploration of the data, perhaps including simple visualizations and the like, to determine the data's utility and spot anomalies that may require further consideration.

The technical issues that arise in getting data out of the data lake are similar to those that arise with putting data into the lake, and are handled with similar tools and techniques, often in ways that are particular to the infrastructure of the enterprise. However, the point when data is taken out of the data lake represents a critical event in the data's life cycle. Once data leaves the lake, it becomes far more difficult to enforce controls over its use. A data scientist checking out data must be made aware of, and have agreed to, the usage guidelines that were prepared for his or her use case.

Unless the target user is familiar with a raw data set, uncurated data is frequently very difficult to work with. Users are required to understand its content, structure, and format prior to deploying it for gainful purpose. Additionally, as described in [12], *contextual data* is often necessary to enhance analytical practices performed on core domain data. That is, value is derived by combining pertinent domain data along with related (contextual) data from other sources. For example, a recent study on the spread of diseases analyzes DNA sample data swiped from surfaces in a city such as turnstiles, public railings, and elevator buttons to identify the microbes present at each location, but it is contextual data such as demographic data and traffic patterns that bring insight into patterns of microbes across neighborhoods, income level, and populations. In enterprise environments, open data only provides value when it can be contextualized with the enterprise's private data. But identifying and leveraging contextual data is very difficult given that providers such as data.gov, BLS (Bureau of Labor Statistics),

NOAA (National Oceanic and Atmospheric Administration) and most others typically organize their data in a hierarchy with either categorical or data-driven delineations that make sense to the applicable domain, and hence is not readily consumable unless thoroughly described via metadata.

6. PRESERVING DATA

Managing a data lake also requires attention to maintenance issues such as staleness, expiration, decommissions and renewals, as well as the logistical issues of the supporting technologies (assuring uptime access to data, sufficient storage space, etc.).

For completeness, we provide a high level description of the issues that arise around data archiving and preservation. The reason for the light treatment is that the literature is quite rich in this regard, as evidenced by the copious amounts of references located in the Research Data Curation Bibliography [2].

Data preservation has gained much momentum in recent years. In fact, scientific project proposals presented to NSF must now include a Data Management Plan; essentially a description of how data will be preserved [14].

A seminal paper [11] on scientific data preservation makes a distinction between *ephemeral* data, which can not be reproduced and must hence be preserved, and *stable* data, which is derived and therefore disposable. In non-scientific domains, such a distinction is not as simple given that issues of currency need be addressed. For example, it was widely publicized that Twitter experienced heavy soccer-related volume during this summer's World Cup, with a steady decline since [32]. While it is highly conceivable that this data will get much use as businesses wish to optimize social behaviors during sporting events, it is equally conceivable that the amount of analytics performed over this event's generated data will wane as it is replaced by information from more recent events. At what point is the data no longer necessary, if ever? The manner by which dormant data is handled becomes relevant as access to it may come in spurts. Furthermore, identifying the point in time when data is no longer necessary, either due to staleness, age, or lack of context requires setting up a preservation strategy.

7. RELATED WORK

The concept of a data lake is a natural evolution from the solid foundation of work in data federation and data integration, including Extract-Transform-Load (ETL) techniques, data cleansing, schema integration and metadata management systems [13][15][29][33] provide a historical perspective of the research challenges in these areas. All of this work contributed to the mature enterprise data integration platforms upon which many enterprises rely to build and populate data warehouses [17][18].

However, such systems require a heavy investment in IT infrastructure and skilled developers and administrators and are tailored for use with tightly controlled enterprise data. As such they restrict the flow of data into the warehouse as well as its use within the enterprise. Many recent efforts have focused on providing automation and tools to enable less skilled workers to clean, integrate and link data [20][26][31], thus enabling the flow of contextual data into an enterprise to be more fluid.

A closely related field is Digital Rights Management, which focuses on the distribution and altering of digital works [9] and the application of transformation and fair use and in copyright law [27], such as is the case for artistic mashups in audio or video recordings.

To date, however, we know of no software platform, business process to systematically define and provide provenance to support the legal and governance issues that enable curated data to flow into and out of an enterprise with the agility needed to support a new class of applications that create derived works by reusing and recombining enterprise and curated data, while still ensuring legal compliance with the potentially myriad of license restrictions associated with the source data.

8. CONCLUSION

We have shown that the creation and use of a data lake, while a simple concept, presents numerous challenges every step of the way. Even after overcoming the legal aspects of "open" data, which deal primarily with licensing and privacy issues, numerous logistical and technical challenges arise in the filling of the lake with raw data. These challenges range from such issues such as data selection, description, maintenance, and governance. We have included examples of user scenarios as well as examples of terms and conditions imposed by data providers.

The daunting nature of populating a data lake may lead some to question its purpose. However, given the vast amount of potential observations, analytics, and discoveries that are derived from cheaply homogenizing data, combined with the evolution of new software tools that take advantage of data in its raw state, not only can the data lake not be ignored, we contend that it will gain prominence in an enterprise's core operational business processes.

Further research in this area focuses on streamlining of processes around data procurement, both in terms of technical automation, and logistical optimization. Much of our immediate work concentrates on automatic data interpretation. Given the varying formats of data (tabular, csv, excel, geospatial, text, JSON, XML, proprietary, http, and many others), we investigate a manner of automated analysis and description with the goal of expediting the process of filling the lake.

Additionally, our focus also centers on the area of collaboration so as to optimize the applicability of lake resident data. Given the democratization of data that the lake provides, in addition to analytical value (whether business oriented, scientific, decision support etc.) further value can be mined from the very way that curated data is used, both within a domain and across domains. In the former, experts within a domain should be able to systematically share and leverage discoveries with colleagues. In the case of the latter, it is common for experts in one domain to experience difficulty when communicating with experts from other domains, thus highlighting the importance of both semantic and conversational metadata, as described in Section 3.3, and underlining the need for tools that facilitate data integration.

9. ACKNOWLEDGMENTS

Our thanks to Mandy Chessell and Dan Wolfson, IBM Distinguished Engineers, for their valuable insight into data lake and open data issues.

10. REFERENCES

[1] Angevaare, Inge. 2009. *Taking Care of Digital Collections and Data: 'Curation' and Organisational Choices for Research Libraries.* LIBER Quarterly: The Journal of European Research Libraries19, no. 1 (2009): 1-12.
http://liber.library.uu.nl/index.php/lq/article/view/7948

[2] Bailey, C. 2014. Research Curation Bibliography. http://digital-scholarship.org/rdcb/rdcb.htm

[3] Bureau of Economic Analysis. http://www.bea.gov

[4] Bureau of Labor Statistics. http://www.bls.gov

[5] Chessell, M., Scheepers, F., Nguyen, N., van Kessel, R., and van der Starre, R. 2014. *Governing and Managing Big Data for Analytics and Decision Makers.* IBM Redguides for Business Leaders. http://www.redbooks.ibm.com/redpapers/pdfs/redp5120.pdf

[6] http://www.cityofboston.gov/doit/databoston/data_disclaimer.asp

[7] CKAN. http://www.ckan.com

[8] EDGAR. U.S. Securities and Exchange Commission. http://www.sec.gov/edgar.shtml

[9] Feigenbaum, Joan. "Security and Privacy in Digital Rights Management, ACM CCS-9 Workshop, DRM 2002, Washington, DC, USA, November 18, 2002, Revised Papers, volume 2696 of Lecture Notes in Computer Science." *Lecture Notes in Computer Science* (2003).

[10] http://www.forbes.com/sites/ciocentral/2011/07/21/big-data- requires-a-big-new-architecture/

[11] Gray, J., Szalay, A., Thakar, A., Stoughton, C., vandenBerg, J., 2002. *Online Scientific Data Curation, Publication, and Archiving.* Technical Report MSR-TR-2002-74. http://www.sdss.jhu.edu/sx/pubs/msr-tr-2002-74.pdf

[12] Haas, L., Cefkin, M., Kieliszewski, C., Plouffe, W., Roth, Mary., *The IBM Research Accelerated Discovery Lab*. 2014 SIGMOD.

[13] Haas, Laura M., Mauricio A. Hernández, Howard Ho, Lucian Popa, and Mary Roth. "Clio grows up: from research prototype to industrial tool." In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 805-810. ACM, 2005.

[14] Halbert, Martin. "Prospects for research data management." *Research Data Management* (2013). 1: http://www.clir.org/pubs/reports/pub160/pub160.pdf

[15] Halevy, Alon, Anand Rajaraman, and Joann Ordille. "Data integration: the teenage years." In *Proceedings of the 32nd international conference on Very large data bases*, pp. 9-16. VLDB Endowment, 2006.

[16] Hassanzadeh, O., et. al. "Helix: Online Enterprise Data Analytics". Proceedings of the 20th international conference companion on the World wide web, pages 225-228, ACM, New York, New York, 2011.

[17] http://www-01.ibm.com/software/data/integration/

[18] http://www.informatica.com/ETL

[19] International Health Terminology Standards Development Organisation. http://www.ihtsdo.org/licensing

[20] Kandel, Sean, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. "Wrangler: Interactive visual specification of data transformation scripts." In*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3363-3372. ACM, 2011.

[21] Kandogan, E., Roth, M., Kieliszewski, C., Ozcan, F., Schloss, B., Schmidt, M., "Data for All: A Systems Approach to Accelerate the Path from Data to Insight." *2013 IEEE International Congress on Big Data.*

[22] LinkedIn. https://developer.linkedin.com/documents/linkedin-apis-terms-use

[23] Murthy, K., et al. "Exploiting Evidence from Unstructured Data to Enhance Master Data Management". PVLDB 5(12): 1862-1873, 2012.

[24] National Climatic Data Center. http://www.ncdc.noaa.gov/

[25] National Elevation Dataset. http://ned.usgs.gov

[26] http://openrefine.org

[27] Power, Aaron. "15 Megabytes of Fame: A Fair Use Defense for Mash-Ups as DJ Culture Reaches its Postmodern Limit." *Sw. UL Rev.* 35 (2005): 577.

[28] Rivera, J., and van der Meulen, R. 2014. *Gartner Says Beware of the Data Lake Fallacy.* Gartner Press Release. http://www.gartner.com/newsroom/id/2809117

[29] Roth, Mary, and Peter M. Schwarz. "Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources." In *VLDB*, vol. 97, pp. 25-29. 1997.

[30] Socrata. http://www.socrata.com

[31] Stonebraker, Michael, Daniel Bruckner, Ihab F. Ilyas, George Beskales, Mitch Cherniack, Stanley B. Zdonik, Alexander Pagan, and Shan Xu. "Data Curation at Scale: The Data Tamer System." In *CIDR*. 2013.

[32] https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter

[33] Vassiliadis, Panos. "A survey of Extract–transform–Load technology."*International Journal of Data Warehousing and Mining (IJDWM)* 5, no. 3 (2009): 1-27.

[34] http://wikibon.org/blog/big-data-statistics/

[35] Wikipedia. http://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative _Commons_Attribution-ShareAlike_3.0_Unported_License

[36] http://www.yelp.com/static?p=tos

[37] Zeng, Marcia L., Qin, Jian. "Metadata". New York: Neal-Schuman, 2008. ISBN: 978-1555706357