PERSISTENT

WHITEPAPER

DATA LAKES

DISCOVERING, GOVERNING AND TRANSFORMING RAW DATA INTO STRATEGIC DATA ASSETS

Fernando Velez,
Chief Data Technologist,
Corporate CTO Office

Sunil Agrawal,
Chief Architect,
Corporate CTO Office

www.persistent.com

© 2017 Persistent Systems Ltd. All rights reserved.

1. Introduction

Early Business Intelligence (BI) tools allowed for simple reports and dashboards that provided business users a glimpse about the current state of their business. Over time, BI tools evolved to perform statistical analysis and trend forecasting to aid in business planning and optimization. Today, BI tools are transitioning from the world of information (e.g., what happened) to prescribing actions to optimize the business (e.g., what actions lead to most desirable results, and recommending these actions).

In this race from descriptive to prescriptive analytics [1], new business demands from the business include analyzing data effectively to come up with business insights that can be actionable either by humans or by systems. The goal is to go beyond the simple "what is happening" and "why it has happened" questions to those such as "what's in the future" and "here are the recommended actions."

This paper is not about predictive or prescriptive analytics, but about how to best collect, manage and prepare the data inputs to any kind of analysis. Indeed, if we want to enable these transformations in BI tools, the key is to provide as large a pool of data as possible that may be relevant for analysis, within and outside the organization, both for immediate analysis or in the future; and to open it to all possible consumption workloads geared towards extracting business value, both traditional BI and machine learning / statistical workloads.

This is exactly what ***data lakes*** provide. Data lakes is an emerging concept geared to unleash the power of data relevant for analysis within and outside an organization. In this paper, we present Persistent's point of view on data lakes. Through specific industry use cases, we discuss how data lakes can be used to discover, govern and explore data, so that it can be turned into a strategic asset through advanced analytics.

Let's illustrate the data lake at work with an example in a typical organization.

> The organization has adopted a data lake for enabling and enhancing CRM. The data lake contains all possible data about customers: their initial interactions with the organization as prospects, their contracts when they became customers, their transactions as customers and their support questions. Marketing campaigns are being built as an application on top of the data lake, based on purchase patterns, and how recent and how frequent have their purchases been. At this point, let's assume the marketing organization gets a report from the BI team showing that the most recent campaign was not successful in terms of reach and response. The business user analyzes the data in the report to find out why, and to see what can be done to make it successful going forward. He brings in a data scientist. The data scientist believes that having a different perspective may help: if she could now bring in recent data from social sites from people talking about the organization's products, she might figure out what was the problem with the campaign. Once she brings the social data into the lake, she can now explore it, gather a filtered dataset with relevant posts, improve the quality of the dataset to be able to match the products talked about in the social sites to those products being promoted in the marketing campaign, ask new questions and eventually apply machine learning algorithms to support a hypothesis about the data: for instance, that the problem of the campaign had to do with the fact it did not reach the people who were the decision-makers. If she is able to obtain that new insight and tell the story with a clear visualization[1], she can now take it to the marketing department for them to decide which new action to take.

[1] For instance, comparing the percentage of opened e-mails from a previous successful campaign in the geographical area of interest with the open e-mail percentage from the new campaign on the same geographical area (which are too low).

© 2017 Persistent Systems Ltd. All rights reserved.

As shown in Figure 1 below, data from various sources reflecting all possible perspectives on one or several business process are first ingested into the data lake without any global model structuring the data. Data discovery, transformation and governance allow to prepare and consume data at later point in time.

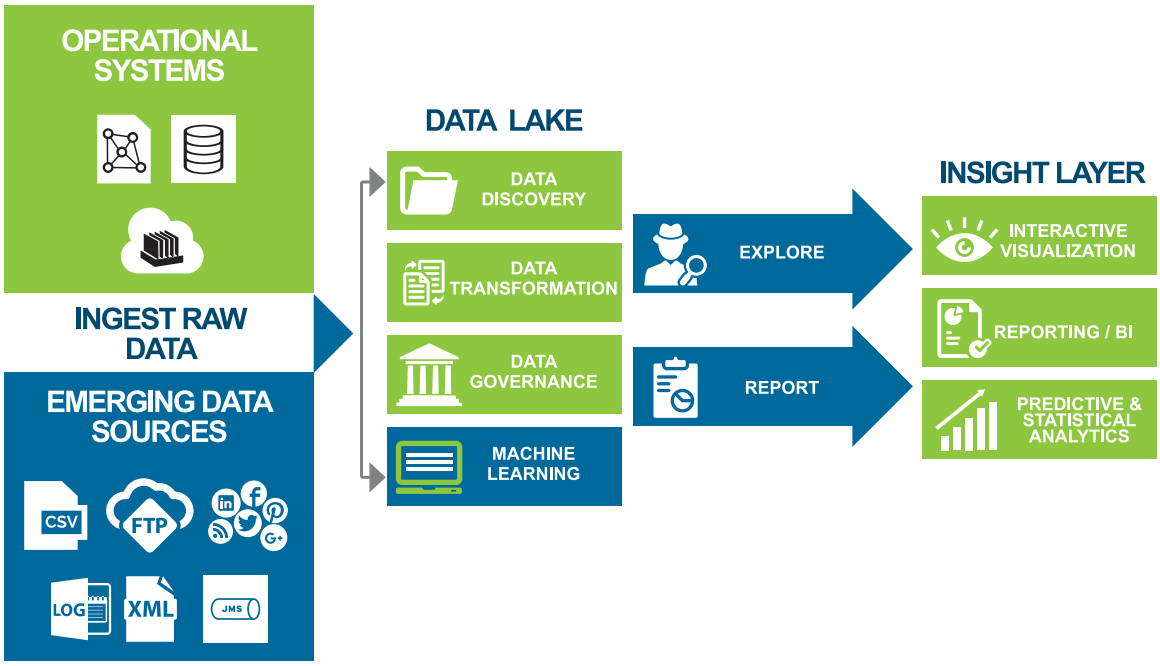

Figure 1: Enabling real business insights with data lakes

At consumption time, when business needs are identified, the user, a data scientist or an advanced analyst, is directed to the data lake to get answers for questions related to this business need. He is confronted to a lake that looks to him more like an ocean of data, and no a priori understanding of its contents. He needs an easy way to discover the subset of data necessary for his analysis. The data discovery layer makes it possible for him to enter several keywords and get initial answers. He then needs to determine the origin of the data found to assign a level of trust. This is provided by the data lake's governance layer, which also secures it from unauthorized access. He or other data scientists can further examine these (now trusted) raw datasets to apply statistical and/or machine learning algorithms to support or reject hypothesis on the data. Once the relevance of the data is ascertained, its quality may then be verified and enhanced, and transformations (filters, joins, aggregations) can be further applied for appropriate consumption. Business analysts can then use visualization tools to build reports and analyze trends. The result may be new insights that can either be available as predictions and forecasts, or as alerts and notifications to users as per business requirements.

We structure this paper as follows. section 2 shows how data lakes can help to overcome the challenges created by data silos that are commonly faced in enterprises especially when deriving insights at the organizational level.

While some enterprises have passing familiarity with the term “data lake”, they are not able to fully comprehend the meaning of the term beyond the idea that a data lake is akin to a big data repository, nor are they able to differentiate between a “data lake” and an enterprise data warehouse. We address in section 3 these topics, in particular:

1. We describe precisely what is a data lake and why it is becoming a must in today's world where data is exploding beyond control
2. We explain the difference between data lakes and data warehouses, and we suggest how organizations can get the best of both worlds in their transformation journey

© 2017 Persistent Systems Ltd. All rights reserved.

In the remaining of this paper, we answer the "What, Why and How" questions about data lake implementations within an organization. While some organizations have realized the need for a data lake, they might lack maturity and knowledge of how to implement it. For them, the real challenge is how to quickly move their data to a data lake and get insights out of it. They are looking for guidance in terms of architecture blue prints, components and processes in order to implement a successful data lake. We highlight in section 4 the key components of a data lake:

- Flexible Data Ingestion
- Data Discovery
- Data Governance (data transformations to improve quality, lineage, auditing, security)
- Data Exploration and Visualization
- Data Storage
- Infrastructure and Operations Management

In section 5 we present specific industry use cases taking advantage of the full power of a data lake. In section 6, based on Persistent's experiences in implementing data lakes for various enterprises, we describe the key business and technical challenges that require prior planning. In section 7, we highlight six best practices to overcome these challenges in data lake implementations. Finally, section 8 summarizes the document.

2. Data Silos in the Enterprises

Driven by the explosion of available data and the technology to harness it, managers at many levels of the organization are now consuming data and analytics based on data in unprecedented ways. The rise of data-driven decision making is real, but it is becoming increasingly evident that one of the biggest impediments to effective data-driven decision making is in organizational silos [2].

Indeed, the reality of large enterprises is that competitive pressure or need for specialization drive business department leaders to focus first and foremost on their individual goals. They create their own set of requirements for reporting and analysis of the data they own, for which the IT department ends up creating a different data *mart*, a smaller data warehouse specific to a given function. All too often, department leaders prefer to manage separate release and feature cycles from other departments, they choose to work in isolation and fail to conform to a corporate data model (which may not even exist). They end up cobbling together piecemeal solutions that are not well integrated with other departments' data. This gives birth to *data silos*. Other causes for data silos include mergers and acquisitions.

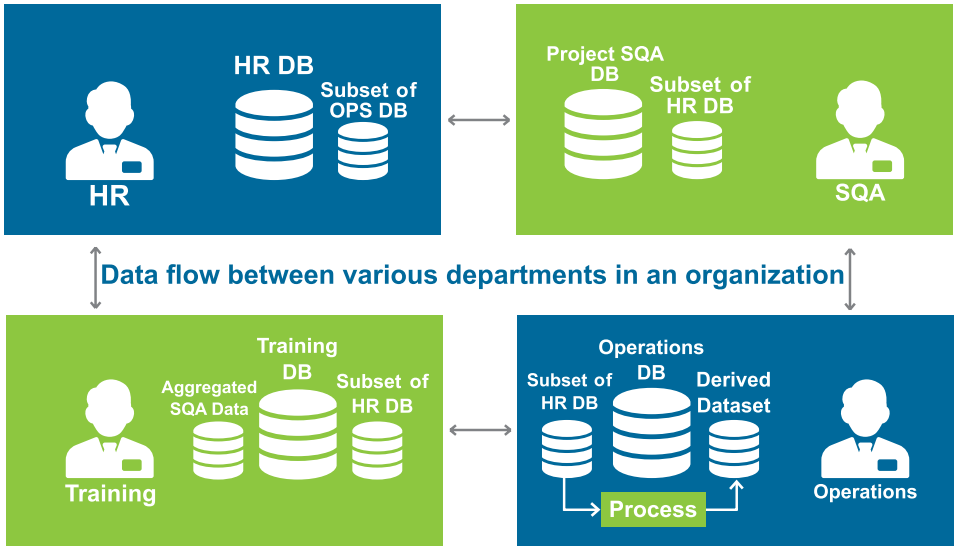

Figure 2: Data Silos in the organizations: Four different departments using copies of each other's data and deriving new datasets based on the copies.

© 2017 Persistent Systems Ltd. All rights reserved.

The response has been for the IT department to integrate these data silos. At the highest level, the corporate level, this means deploying an enterprise **data warehouse** (EDW).

Some forward-looking organizations have even anticipated the creation of a corporate data warehouse model out of all transactional databases from each business functional area. From this model, they either embark into delivering an all-encompassing data warehouse implementation (from where data marts can be derived), or they deliver data marts that conform to the corporate model, thereby avoiding silos.

In our experience, however, these organizations are more the exception than the rule. We have seen that most organizations fall at the lower end of the spectrum: they approach data integration as a tactical activity, through a series of one-off projects that fail to recognize that there will always be more data integration projects.

While data warehouses have become the classic answer to remove silos, they have challenges of their own, notably:

- They take a long time to implement, and the ROI is unknown until it is available in production to end users.
- They are inflexible: they are only designed to answer a fixed set of questions.
- They are brittle: the target data model is frequently out of date before BI has even started, and it takes time to bring them up to date with the changed source models.
- Due to large volumes, performance is impacted. Also storage of large volumes with live data becomes very costly. Hence they need to archive data and then throw from archive after certain retention period.
- Traditional data warehouses are typically not well equipped to ingest, store and analyze unstructured data from internal applications (e.g., weblogs), or external data (social data, documents and images) from cloud and mobile applications.

Data lakes is an emerging approach to extracting and placing all data relevant for analytics in a single repository. All data means both big and “small” data, as well as data internal to the organization and external to it. Data lakes are an alternative to data warehouses to put an end to data silos. Before we introduce more formally the concept, as well as our technical architecture for the concept, we want to examine the main actors in the organization as well as their needs. We have both data producers, who want to control the data (for good reasons, but which tend to fuel this data silo problem), and data consumers, who want to solve business problems with new insights they can gather through analyzing data, of which the internal part is being owned by the producers.

Needs of Data Producers

The Data Producer is the business head who owns the data specific to the business function. As part of their job, they are generally worried about the following:

- Regulatory Compliance: producers need to ensure that business policies and rules are followed while storing and accessing data from the data lake in order to comply with laws and regulations.
- Data usage visibility: Data producers would like to know who used the data and when.
- Privacy and Access Control: Data producers would like to control the access of data and privacy information such as Social Security Number (SSN), Date of birth (DOB) and credit card information.
- Minimal IT overhead: Data producers are also worried about IT overhead in terms of infrastructure and resources to manage large and complex data warehouses.
- Manage Cost: data producers are also worried about overall cost of implementing a data warehouse or a data lake. They need to manage this cost based on a perceived ROI.

Needs of Data Consumers

Here data scientists and business analysts are referred as data consumers. They are the ones who discover, explore, visualize and ultimately get business value in the form of insights to executives.

© 2017 Persistent Systems Ltd. All rights reserved.

Some of the requirements of data consumers include:

- Quickly find relevant data through self-service: they should be able to discover and explore data sets lying in the data lake which are relevant to their current research with minimal dependency on the IT organization.
- Validated data sets: consumers should be able to find and access validated, curated data sets; as a minimum, they should be able to determine the quality of datasets and their provenance to provide confidence over insights found using these datasets.
- Support a variety of analytics workloads: consumers should be able to perform any kind of analytics. Business analysts should be able to formulate ad-hoc, interactive queries, aggregations, and easily build visualizations and dashboards. Data scientists need to perform data mining and applying machine learning algorithms.
- Support scenarios where reports need to be delivered in real-time.
- Access to historical data to allow looking at the trends and compare data values across time windows.

Need a Data Custodian to Handle Impedance Mismatch

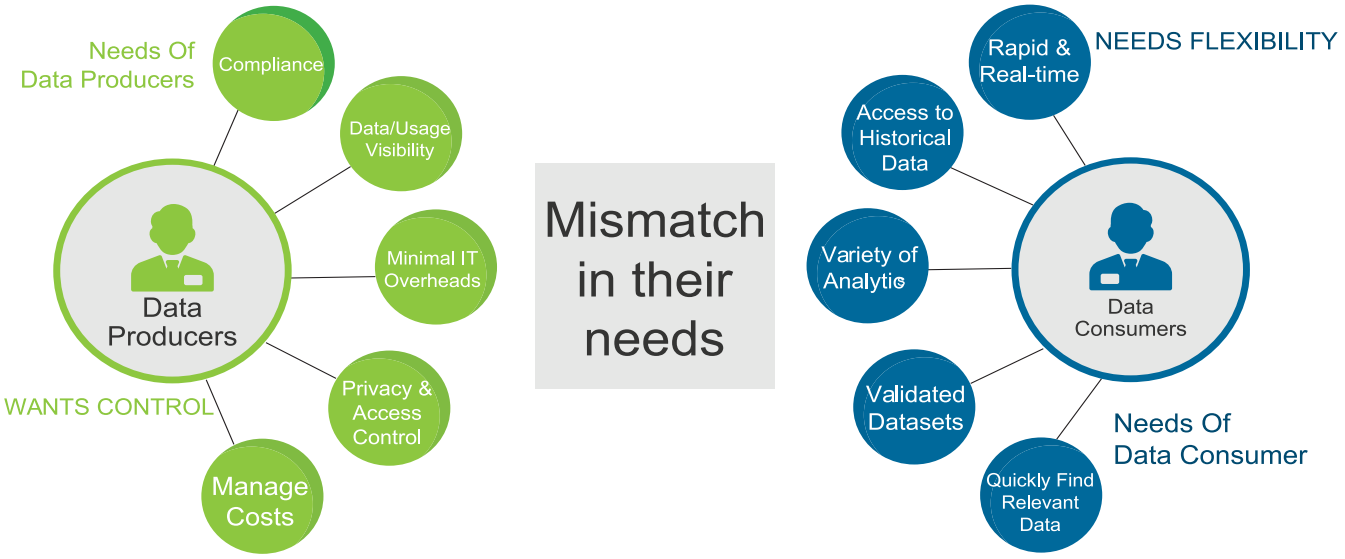

Figure 3: Mismatch between needs of producers and consumer (for control and flexibility, respectively).

As we can see in Figure 3, the needs for the producers and consumers, summarized in the respective bubbles, can be quite at odds. We call this an impedance mismatch. The team implementing the data lake tries to solve this mismatch by becoming the intermediate layer – Data Custodians. In order to handle the impedance mismatch between data producers and data consumers, data custodians need to provide for the following:

- A flexible data ingestion framework which not only allows to read any kind of data (unstructured and structured) into the data repository, but also doing so in an automated fashion, where it can be curated or transformed for further processing.
- A low cost multi-structured data storage layer so that storing historical data for any number of years is not an issue.
- Data governance tools to validate and cleanse data, to capture and manage metadata at different levels to make data searchable, and to provide auditability
- Data integration tools to allow combining data from different datasets and providing users with a unified view of these data.
- Set up data discovery tools to allow consumers to search relevant data through metadata tags.
- Set up security tools to control access of data and encrypt it wherever required.
- Set up tools for multi-type analytics workloads, as described above.
- Based on access, job scheduling and its priority management is also required.

© 2017 Persistent Systems Ltd. All rights reserved.

3. Data Lakes

Data Lake Definition

A Data lake is a central **data repository** that can store **multi-structured** (i.e., structured, semi-structured and unstructured) data in **native format**. It then uses a variety of processing tools to **discover** and to **govern** the data, which include improving its overall quality, making it **consumable**; finally, it allows tools and exposes APIs for consumers to explore and extract business value through several types of consumer workloads.

Data lakes are typically built using Hadoop, given the potentially very large data volumes involved. One of the primary motivations of a data lake is to provide as large a pool of data as possible without losing any data that may be relevant for analysis, now or in the future. Data lakes thus store all the data deemed relevant for analysis, and ingests in raw form, without conforming to any data model.

Data custodians can defer the intensive data preparation process within the lake until clear business needs are identified. Governance tools allow IT and data scientists to discover trusted, raw data and explore it through data mining workloads that may provide initial insights that help in determining such business needs. When that occurs, the relevant data in the lake can then be validated, cleansed and made to conform to a structured schema, for consumers to ultimately derive the full business value from the data.

Data Warehouses Vs. Data Lakes

Data lakes are far more flexible alternatives to data warehouses to gather all the data of an organization that is relevant for analysis (and put an end to data silos), especially as they become more outward looking and increase their exposure to cloud and mobile applications.

Indeed, data warehouses are designed with a very different purpose in mind, which is to make sure that a family of queries that are to be asked of the data return trusted results with high performance. As such, they require upfront modeling, transformation of the source data to the target model, and cleansing of the data, so that only good quality data conforming to the model can be loaded. As mentioned in section 7 below, for most regulatory industries, not having an EDW is not an option, as regulations mandate the need of quick and real-time access to structured data through a carefully designed set of queries.

The table below summarizes the main differences between enterprise data warehouses and data lakes.

Data Warehouses	Data Lakes
Upfront modeling and conformance of the ingested data to the model (this is also called “schema-on-write”).	No upfront modeling, the ingested data is stored in raw form. You can put off modeling until you use the data (“schema-on-read”).
The data model reflects a list of requirements that include a fixed collection of questions that are to be asked of the data.	There is no initial requirement containing a fixed collection of questions. Questions will arrive after the data is loaded.
Some of the available data may not be brought to the EDW, as it may not fit in the model. Semi-structured data and unstructured data are generally not brought in.	All data relevant to the organization may be brought, be it structured (e.g., database tables), semi-structured (e.g., web logs) or unstructured (e.g., text documents).
EDWs incur in long implementation cycles before analysis can start.	Data preparation and analysis can start as soon as the data is collected.

© 2017 Persistent Systems Ltd. All rights reserved.

Data Warehouses	Data Lakes
Data is validated and cleansed eagerly, at load time, so that the answers to questions can be trusted.	The ingested data can be validated and cleansed at a later point in time. Trusted data will only result after a well-managed governance process.
Holistic view of data for analysis is available once implementation is ready.	Holistic view of data with required context for analysis is much more difficult to obtain.
Changing the data model may be time consuming given all the business processes that get tied to it.	No global model structures the data. The existing partial models and queries can be reconfigured on the fly with more agility.
Data is easy to consume, typically through BI tools sending statements written in standard declarative query languages (SQL, MDX).	Data is consumed through declarative query languages (HiveQL, not completely up to standards), as well as programs written in scripting languages (Pig), and procedural languages. The latter support data mining/ machine working workloads for a new type of consumer, the data scientist.

In summary, data warehouses and data lakes are tools designed and optimized for different purposes. They can co-exist in an organization's landscape. Data from the warehouse can be fed into the lake; and conversely, data from the lake can be a source for the warehouse, or a data mart (a portion of a data warehouse specializing on a business process). We will revisit this point below.

Discovering and Governing Data in the Lake

Business users will need flexibility to search and explore the ingested data sets on the fly, on their own terms. Metadata tagging is the means to identify, organize and make sense of the raw data ingested in the lake. It can be performed both by custodians and consumers and by data lake automated processes. Once the data is tagged, users can start searching datasets by entering keywords that refer to tags. Importantly, tagging has an important role in managing unstructured data such as documents in the data lake, in the areas of capturing document semantics through tags and making documents searchable.

While there has been lot of buzz and proof-of-concepts done around big data technologies, the main reason why acceptance in production environments has been slow is the lack of data governance process and tools, which are used to:

- **Improve and monitor the quality of data.** If the quality of data is too low (i.e., if it is incorrect, inconsistent, inaccurate, incomplete or duplicated), little or no value can be derived from it. Besides, in some regulated industries, data must be of good quality for compliance. Data lakes should have the means for data custodians to assess the level of quality of data, as well as the rules to clean it up. In addition, these rules may be applied to selected, **curated** datasets currently being consumed by business users and being refreshed regularly, to monitor the quality of the data over time.
- **Provide trust and traceability of data:** The data lake tools should make sure that lineage metadata is automatically captured to track both the original data source of data as it is ingested in the lake, as well as any data transformation applied to it subsequently. Without traceability through lineage tools the data lake may become useless, as it may contain huge quantities of data that can't ultimately be trusted because its origin is uncertain. This has been identified as one of the main business challenges in section 7 below.
- **Enforce security:** Assure data producers that data inside the data lake will be accessed by only authorized users.
- **Provide auditability:** Any access to data should be recorded in order to satisfy compliance audits.

© 2017 Persistent Systems Ltd. All rights reserved.

In short, data governance is the means by which a data custodian can balance control requested by data producers and flexibility requested by consumers in the data lake. We will provide more detail on each of these topics in our architecture blueprint below.

Implementation of data governance in the data lake depends entirely on the culture of the enterprises. Some may already have very strict policies and control mechanisms put in place to access data and, for them, it is easier to replicate these same mechanisms when implementing the data lake. In enterprises where this is not the case, they need to start by defining the rules and policies for access control, auditing and tracking data.

4. Architecture Blue Print of Data Lake Platform

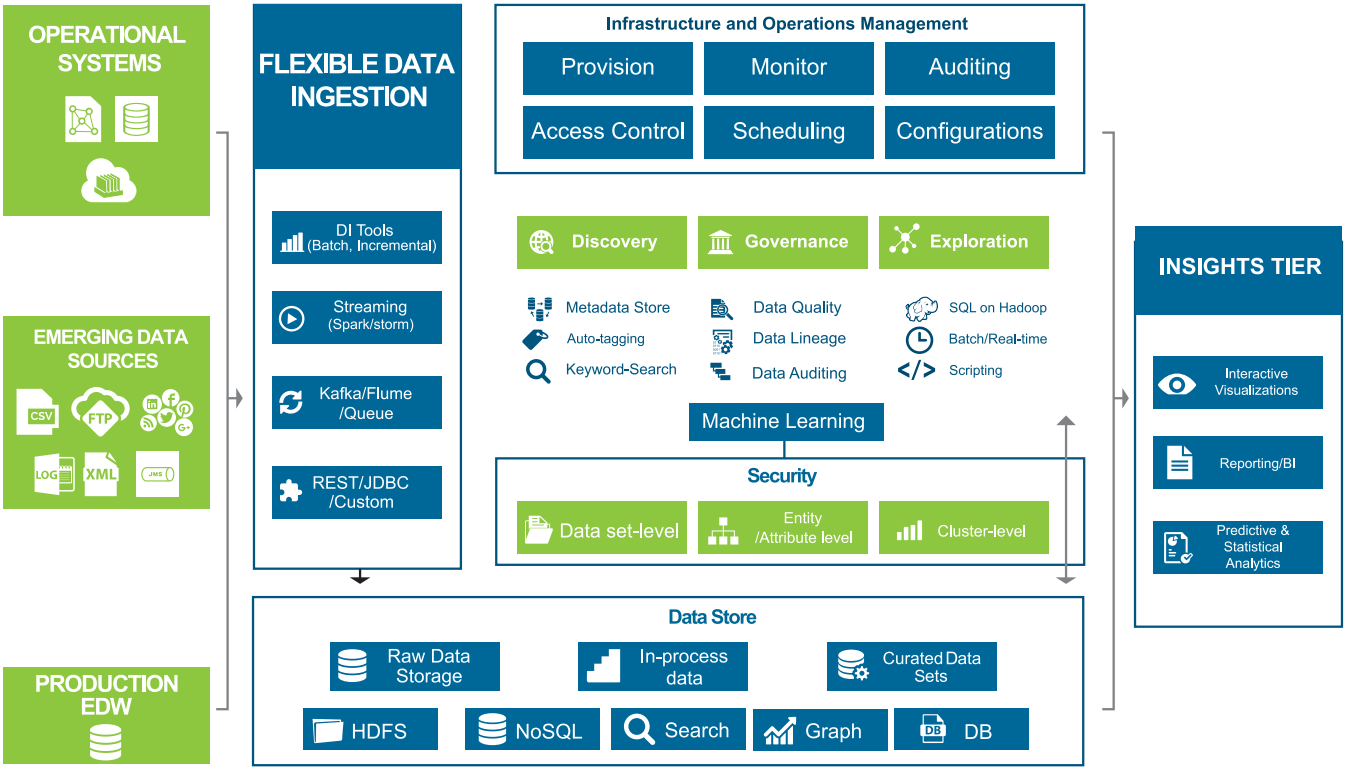

Figure 4: Architecture Blue Print of a Data Lake

Flexible Data Ingestion

The data ingestion layer contains connectors to extract data from a variety of sources and load it into the lake. It provides a simple and unified interface to

- Configure connectors to on premise and cloud sources, whether they are databases, files of any format, streaming sources, or applications accessed through APIs
- Control the ingestion modes: bulk/batch, or continuous flows (either from streaming sources or change data capture facilities from databases)
- Visually create data collection flows that allow specifying the parts to ingest from the source (e.g., schemas or datasets from an RDBMS) and the place to store the incoming data
- Schedule the collection flows getting data from bulk/batch connectors, to control the time of ingestion

www.persistent.com

© 2017 Persistent Systems Ltd. All rights reserved.

The data ingestion layer allows for tracking and troubleshooting the data ingestion flows. It also creates descriptive metadata on data onboarding, such as source provenance and onboarding timestamp, automatically. See [3] for a review of the tools available for implementing ingestion in the market.

Data Discovery

In a data lake, by definition, data analysis queries arrive after load time. As we saw in the section above, this is in sharp contrast to data warehousing, where the target data model is shaped to respond to the queries known in advance.

For this reason, post-data ingestion large data collections need a data understanding stage before one can even start data preparation or analysis. Metadata *tagging* (or simply tagging) is the way to express this understanding, through identifying, organizing and making sense of the raw data ingested in the lake. In increasing order of complexity, tagging encompasses

- The traditional schema information (table/dataset name, description and attribute metadata)
- The information about the data values through profiling.
    - This may include the identification of higher level "business types" on columns or sets of columns, e.g., people names, addresses, cities, countries, and the like.
    - On text documents, this translates into the ability to scan documents and parse out concepts based on the words and the context in which those words appear. These will be used as tags to attach to the document and for internal indexing of document data.
- The relationships/links between attributes of different datasets may be discovered or indicated as tags explicitly by the user, and
- Higher level, business-specific tagging and synonyms between tags, which allow for a shared convergence of meaning among users.

Tagging can be performed both by custodians and consumers and by a data lake automated discovery process. In practice, discovery tools work by taking a sample of the datasets they are requested to inspect which needs to be indicated as part of this process; profiling and relationship discovery is carried out on this sample. The data lake matures as a result of the user interaction and feedback through this tagging process. New datasets embodying findings that are produced by business analysts as the data lake gets used also have to be tagged, so that these findings can be reused and integrated into new ones.

Importantly, tagging enables dataset discovery. Users can start searching datasets by entering keywords that refer to tags, for example, `'Movies', 'Orders', 'Complaints'`. It can also be a mixture of tags and data values (`'Movies in India'` or `'Complaints logged by Customer Peter'`). Finally, users can formulate search queries where results are datasets containing columns constrained by data type or business type (e.g., `City: ('New York' or 'Paris'))`.

In reference [4] a review of the libraries available for detecting metadata is presented, as well as some ready-to-use tools that can augment data and schema discovery.

Data Governance

Now we discuss each of the mechanisms introduced in the previous section.

Data Quality

As discussed before, data quality is a necessary condition for consumers to get business value out of the lake. The process and tools for filtering and purifying the quality of data in the lake is basically the same than in a data warehouse. The glossary in appendix 1, which we will be using below, serves as a crash introduction to the technology and vocabulary of data quality.

www.persistent.com

© 2017 Persistent Systems Ltd. All rights reserved.

Contrary to data warehouses, in a data lake, data scientists and business users could in principle consume bad quality data, and get incorrect or misleading insights. This motivates the following choices:

- The liberal choice: data consumers should be able to tell by themselves the quality of the datasets they discover before exploration (through profiling or inspecting quality metadata).
- The cautious choice: business analysts should only be able to discover curated datasets. Data scientists are technically skilled to assess and improve the quality of datasets by themselves.
- The traditional choice: the platform for most business users is not the data lake, but a data mart or warehouse fed upstream with data from the lake[2]. This choice may be motivated by the experience of an organization with a given ETL or BI tool, and by the performance requirements of target queries. Data in the lake is consumed by data scientists and some power business analysts to identify some new insights on a collection of related datasets.

The data custodian (and eventually the data scientists) can assess the quality of the data by using profiling tools or machine learning algorithms. She can then meet with data producers to agree on standardization of data elements, as well as the minimal business validation rules that need to be put in place for the data to be deemed consumable.

Once this is done, the data custodian can define (i) the transformation flows needed to cleanse the data for it to meet agreed data standardizations and the necessary integrity constraints and business rules, (ii) define the validation rules that make sure that the data is clean, as well as (iii) the data matching and consolidation rules to remove duplicates and improve data accuracy in the lake. The above transposes a traditional data quality process, which applies to applications working with transactional data in the lake. However, clickstream analysis applications, fraud detection applications or applications based on sensor or social data in the lake may call for different data quality processes. We illustrate the point with two examples: (i) as mentioned in the retail banking use case of section 5, entity matching from social site data with customer master data may involve extracting textual information from social postings that can be matched to customer master data attributes; and (ii) environmental monitoring applications relying on sensor data would rather prefer data quality to be about detecting anomalous readings, rather than cleansing away data that does not conform to strict integrity constraints.

Curated datasets may get permanently refreshed from their original source, so it is appropriate to apply their corresponding validation rules to these datasets, and to monitor the results of validation rules over time.

There are two broad options for toolsets addressing quality of data in a lake. The first is through native Hadoop technology, through MapReduce, HiveQL or Pig programs or, more recently, with Apache Spark. The second is through the use of ETL tools that know how to work efficiently with Hadoop data, notably by pushdown optimizations, i.e., generating and executing the Hadoop code equivalent to the dataflow transformations authored in their graphical UIs. All leading ETL tools provide pushdown facilities by now.

Data Lineage

Data Lineage is a process by which the lifecycle of data is managed to track its origin and its journey from origin to destination, and visualized through appropriate tools.

Data lineage tools are essential to help build trust with data consumers by allowing them to visualize and certify the original data source and the time at which the data was ingested in the lake, as well as around the transformation and rules applied to data when it goes through a data analytics pipeline. By visualizing the data lineage, business users can determine who and when transformations were built, e.g., for cleansing purposes by IT personnel or by data scientists, or for getting new insights, by data scientists or by business analysts. This will allow business users, for instance, to identify and understand the derivation of aggregated fields in a report or a dashboard. They will be also able to reproduce data points shown in the data lineage path. It also helps to debug step-by-step the data pipeline.

Data lineage visualization should show users all the hops the data has taken before generating the final output. It should display the queries run, table, columns used, or any formula/rules applied. This representation could be shown as nodes (data hops) and processes (transformation or formulas), thus maintaining and displaying the dependencies between datasets belonging to the

[2] The data mart can be populated with curated datasets from the lake; or alternatively, the data quality processes are applied to selected, non-curated datasets by ETL engines in their way to feed the data mart.

© 2017 Persistent Systems Ltd. All rights reserved.

same derivation chain. As explained in the section above, tags are generalizing metadata information such as table names, column names, data types, and profiles. Hence, tags should also be part of derivation chain.

Reference [5] provides a review of the data lineage tools available on the open source Hadoop world or commercial tools working on top of Hadoop data.

Data Auditing

Data auditing is the process of recording access and transformation of data for business fraud risk and compliance requirements. Data auditing needs to track changes of key elements of datasets and capture "who / when / how" information about changes to these elements. A good auditing example is vehicle title information, where governments typically mandate the storing of the history of the vehicle title changes along with information about when, by whom, how and possibly why was the title changed.

Why is data auditing a requirement for data lakes? Well, transactional databases don't generally store the history of changes, let alone extra auditing information. This takes place in traditional data warehouses. Auditing data requires its share of storage, however, so after 6 months or a year it is a common practice to move it offline. From an auditing perspective, this period of time is small. As data in the lake is held for much longer periods of time, and as data lakes are perfect candidate data sources for a data warehouse, it makes sense that data auditing becomes a requirement for data lakes.

Data auditing also keeps track of access control data in terms of how many times an unauthorized user tried to access data. It is also useful to audit the logs recording denial of service events.

Data auditing may be implemented in two ways: either by explicitly copying previous versions of dataset data elements before making changes, as in the traditional data warehouse slow changing dimensions [7], or by making a separate note of what changes have been made, through DBMS mechanisms such as triggers or specific CDC features [8], or auditing DBMS extensions [9]. Data auditing tools available on the open source Hadoop platform, summarized in [5], follow the first pattern.

Security

It is important for enterprises adopting data lakes to protect sensitive information which is proprietary or personally identifiable information. There are also compliance requirements such as HIPAA, SOX, PCI, etc. and need to adhere to corporate security policies. The issue gets magnified as the data lake in many cases is multi-tenant, which hosts the data for multiple customers or business units – distributed, multi-tenant world.

Security should be implemented at all layers of the lake starting from Ingestion, Storage, Discovery and Consumption through analytics. The basic requirement is to restrict the access to data to trusted users and services only. The following are the key features available for data lake security.

1. **Network Perimeter Security** protects the lake from access from outside the organization and is a must for any data lake.
2. **Authentication** will verify user's identity and ensure they really are who they say they are.
3. **Access Control** is the next step to secure data by defining which dataset can be accessed by the users or services.
4. Sensitive data in the cluster should be secured at rest as well as in motion. We need to use proper **Data Protection** techniques which will protect data in the cluster from unauthorized visibility. Encryption and data masking are required to ensure secure access to sensitive data.

Reference [6] provides a review of the APIs, protocols and tools used to enforce security which could be applicable in data lake implementations.

Data Exploration

The preamble for data exploration, as highlighted above, is dataset discovery. Once the user has identified the right dataset to work with, she can start exploring it. For the data lake to be successful, the data exploration facility needs to provide the following key features:

© 2017 Persistent Systems Ltd. All rights reserved.

Flexible Access

The following are the key topics for flexible access from the data exploration point of view that a data lake design team needs to consider.

- Interfaces: Consumers have different skill sets and would prefer to access the data stored in the data lake using their favorite mechanisms and/or tools (e.g. SQL, scripting, programming, tools with friendly GUIs such as dashboards). Don't try and fight this. All these different interfaces must be supported by a data lake.
- Workloads: Different consumers put different kind of load on the system. For example, data scientists do ad hoc or statistical analysis to understand the data (whether values are binary, categorical, ordinal, binomial, etc.), and identify the key attributes from these data sets, or support a thesis about the data they see to derive value from it. Business users interact with the system by browsing, filtering, profiling, aggregating, etc. These workloads need to be supported irrespective of the format the data is in, as the ingestion process accepts all data from all possible sources in raw format.

Self-service

Data Consumers should be able to explore the data with minimal dependency on the IT organization. As mentioned above, they need to be able to discover curated datasets, as well as inspect their lineage metadata, without involving IT directly. Data consumers also have preferences of how they'd like to consume the results of their analysis. Many prefers in the form of graphical reports, whereas some users would want to consume it programmatically, through APIs.

Team work

Given that the data lake contains vast amounts of data in raw form coming from multiple sources, it might take a significant amount of time to find the correlation between data sets previously unknown. Data lakes need to provide a collaborative team environment so that analysis and findings of one user (or group) can be shared with other users (or groups), to avoid duplicate effort and improve the overall business outcome.

In [10] the reader will find a presentation of some of the tools and technologies available to support a powerful data lake exploration environment.

Data Storage

IIn the data lake, data storage is the key component of the architecture. Storage design not only has to meet the demand of data growth but also ability to provide fast access to the data, which depends on the type of exploration workloads we examined in the previous section.

As mentioned earlier, to eliminate data silos, we need to build a reliable data lake which can store data currently being handled by traditional sources but also massive volume coming from unstructured data sources. The data lake, while primarily holding raw data, also stores in-process as well as processed, curated data.

Data lake storage doesn't have any constraint in terms of format of data. In fact, data can be stored in compressed formats such as GZIP and snappy to improve space and network bandwidth utilization. Multiple formats and compression techniques can be used in the same data lake to best support specific data and query requirements. In general, data lake storage should be scalable, low cost and should support multi-protocol access.

While it is certainly difficult to architect a perfect storage system for the data lake, here are a few desired attributes:

(I) Scalability: No degradation in performance irrespective of data volume, i.e., throughput and response time must be scalable.

(ii) Fast access for data exploration workloads: For example, data stored on file system in raw format needs to be indexed to support text search queries. Similarly, for key based lookups, data could be stored in key-value stores like some of the NoSQL databases.

(iii) High availability: Widely distributed system, across data centers and/or geographies, along with a fault-tolerant design that makes the system highly available.

© 2017 Persistent Systems Ltd. All rights reserved.

(iv) Ease of integration: With an enterprise ecosystem of public, private, or virtual private cloud, with enough protection.

(v) Elasticity: Flexibility to add or remove capacity based on need.

(vi) Durability and affordability, in terms of recurring management and storage cost.

For storage, there is not a single storage option which fits all requirements. Hadoop Distributed File System (HDFS) is the leading data storage technology which is mature now for large scale persistent data deployments. There are other NoSQL options like Cassandra, HBase (in the Hadoop product family), and MongoDB which are also available for key value storage.

Enterprises can either build on premise or cloud. For cloud services like Amazon S3, Google Cloud Storage, and Microsoft Blob store can be used.

Infrastructure and Operations Management

Most of the Hadoop distributions come with the monitoring tools. In many cases thought, we may need to add additional features on top of the existing monitoring tools using REST APIs. Some of these additional features could be specific for the data lake requirements like curation of data and effectiveness of regulatory compliance. Apart from tools available with distribution, platform, third-party monitoring tools like *datadog* can also be used.

Most of the Hadoop distributions come with the web UI for installing / configuring the setup. However, automation should be done for configuration and deployment including handling pre-requisites.

5. Industry Use Cases

Healthcare Industry Use Cases

Enough of the technical perspective, we now look at some industry use cases.

The health care industry has recently gone through a data revolution where the main actors are sharing data in unprecedented ways [11]: health care service providers have digitized and shared patient medical records, insurance companies are storing claims data in medical databases, and pharmaceutical companies have started sharing their research and clinical trial data in the public domain. This has led to tremendous growth in term of data for medical systems. The insight collected from these various sources can be useful for patient care and management. However, very frequently, healthcare providers do not have a sufficiently integrated view of data about patients, their ailments and related therapies, as well as their claims to make the right decisions at the right time.

Let's discuss a use case to understand how the data lake concept can help the Healthcare industry overcome the challenges inherent in integrating all these different sources of data.

Cohort Builder Solution

The cohort builder application should enable the identification of a population of patients who match a user defined criteria for a cohort. Cohorts can be used for clinical trials, disease management and targeted therapies.

In order to build a cohort, data needs to collect from various data sources:

- Electronic Medical records
- Patient Management System
- Clinical Trials
- Laboratory data
- Financial Systems
- Social Media

The data from these source systems need to be ingested in a data lake repository using standards such as Health Level 7 (HL7). During ingestion, the right healthcare taxonomy should be used to tag different datasets and their columns with the same tag for similar datasets and fields. Tagging can also be done for PII relevant information such as SSN and DOB. The IT administrator then can create tag based policy to hide PII information from being available for all users.

© 2017 Persistent Systems Ltd. All rights reserved.

As a next step, data discovery tools can be used to search the data within lake. These tools typically work with samples of datasets in the lake. In this example, a data scientist or a business analyst can use a discovery tool to search which are the datasets holding available patient relevant information. Once these are found, it is important to make sure their origin can be trusted. The user should be able to check the lineage of the data (data source of provenance) and discard those that don't qualify. A similar discovery process may apply to disease information, and may be even be targeted to search for a particular ailment: for instance, allergies. Then, relationship links can be found between these two types of datasets, e.g., which patient datasets are related to which datasets containing allergy data, and find out how these are these linked: for instance, through a common patient id or a patient name. When datasets come from different sources, patient names would typically need to be used, but lack of standards for people names may deliver a low percentage of matches. Enhancements to standardize patient names will ultimately be needed.

The criteria for the cohort may also need other information such as clinical trials, or laboratory data. At the end of this discovery process a relevant data model of interconnected data is identified. However, the model needs to work with the complete selected datasets from the lake, and needs to be enhanced to include data quality enhancements such as name cleansing and standardization. Once the data has been processed, a new data mart within the data lake can be exposed to a wider user base of business analysts and/or data scientists. To validate the data sets, besides the lineage of the data (data provenance), analysts should be able to check any transformation applied to the data and any intermediate datasets generated before the data mart.

The business analysts can then do their analysis, which may include running specialized algorithms to further filter the patient population to determine the final cohort. This cohort along with its lineage information can be presented to a final board to get approval to run the clinical trials.

Clinical Quality Measure Solution

Another application of this data lake for hospitals is to meet reporting requirements for clinical quality measures (CQM). Hospitals should be able to provide timely performance snapshots of CQM down to patient and physician level, through reports where any issue in regulatory compliance in CQM can be found out.

Let's illustrate this with an example where the CQM measure to track for the hospital is “Percentage of patients ages 18 years and older, with a BMI documented during the current encounter or during the previous six months AND with a BMI outside of normal parameters, for which a follow-up plan has been documented during the encounter or during the previous six months of the current encounter”. This particular CQM would require data lake consumers to discover Patient and Encounter datasets. Other CQMs may require similar data sets to be able to report the compliance. If no such datasets are discovered, they need to be included in the data lake ingestion pipeline. Otherwise, analysts need to create a new data model with these datasets and then provide the requested CQM report. As in the previous use case, joining and filtering transformations, eventually preceded by cleansing transformations to make the patient data joinable with the encounter data, will be needed to feed the model. In addition, from a compliance perspective, business analysts should be able to validate the data provenance and produce a data lineage report for regulatory compliance.

This CQM report should include a chart with the aggregated measure to track, as well as a facility to drill into patient data and lab tests to be conducted by physicians on next visit of the patients. It will also be helpful in giving wellness score of the patient, for which specific calculation transforms may be needed.

Retail Banking Industry Use Case

In retail banking, customers are offered different products; checking account, savings account, credit card, loans etc. While these are not only different products, they are different departments within Banks. Data for these customers is frequently stored separately in silos. And hence customer analysis is generally done in isolation, without looking at the complete set of products owned by customers. This is not only inefficient but usually very frustrating to the customer.

Apart from looking at the consolidated view of a customer, there are many challenges which are faced by financial banks due to the lack of agility of traditional data warehouses, for instance, when adding new data sources and easily correlating with existing data. In order to share a common data model for different product teams, some product team typically ends up compromising their needs. In addition, importantly, they incur heavy costs to keep the data for many years due to regulations. Building a data lake in retail banking, not only would solve the above challenges, but help with the following additional use cases:

© 2017 Persistent Systems Ltd. All rights reserved.

- Know Your Customer (customer 360)
    - Find out customer's spending and saving patterns to suggest specific products and services.
    - Find out customer's overdraft/loan patterns (get external data on credit scoring).
    - Create customer's overall profile and segment them.
- Build solutions on the data lake and expose to retail Customers
    - Help customers categorize their spending (Entertainment, Utilities, Insurance, Investments, Healthcare (Beauty), Restaurants/Food, Shopping) for Monthly/ Quarterly/ Yearly expenses.
    - Value Add: Machine learning to auto categorize spending.
- For corporate customers, if data from various sources can be combined and analyzed, correlations between different types of expenses can be discovered which could ultimately help in optimizing expenses. In addition, organizations can get a global view of their categorized expenses. This will help them reorganize the efforts and resources of the organization where expenses are optimal.
- Build solutions to better understand customer complaints which are not only being funneled through the bank support web site, call center and other bank channels but also through postings on social media platforms. Banks would like to be able to
    - Ingest the complaint messages from selected social platforms,
    - Classify the complaints against the bank's products,
    - Extract sentiments of these classified complaint messages,
    - Match complaints authors with actual bank customers, and then
    - Mash this information with the formal, on premise complaint data from the bank's support organization to provide insights on the combined data through analytics.

One of the main issues is thus resolving social site complaint authors, of which only IDs or handles (which may or may not include names) are known, to customer identity master data attributes (including names, products purchased, geographic information, etc.) which may also be made part of the data lake.

6. Challenges of implementing the Data Lake

Implementing the data lake is faced with many challenges which all stakeholders including business users and technical IT team should be aware about.

Business Challenges

Leaving the data lake as the IT department's project will probably lead to failure. The main challenge does not lie in filling a lake with data (that's the easy part), but getting value out of data in the lake, which ultimately remains the responsibility of the data lake consumers. To meet this challenge, organizations should focus on the following two points:

(i) A thoughtful approach to governance in the data lake involving business users and data scientists. On this point, it may be appropriate to quote Gartner [13]: “Data lakes therefore carry substantial risks. The most important is the inability to determine data quality or the lineage of findings by other analysts or users that have found value, previously, in using the same data in the lake. By its definition, a data lake accepts any data, without oversight or governance. Without descriptive metadata and a mechanism to maintain it, the data lake risks turning into a data swamp. And without metadata, every subsequent use of data means analysts start from scratch”.

(ii) A strategic understanding of the type of business use cases that insights generated from the data lake can solve, and which can't be solved with the current analytics systems. It may be enough with identifying a first use cases upfront, at least to convince executive management to get sponsorship to start a data lake project. After all, data lakes are about flexibility: not all use cases nor questions have to be known upfront.

© 2017 Persistent Systems Ltd. All rights reserved.

Most of the business challenges have to do with the cleanliness of data in the lake for consumption:

- The IT team may get involved every time to prepare data and improve its quality for consumption. This results in long time to extract value, becoming dependent on the IT team's bandwidth.
- With a large number of data sets within data lake, it may be hard to find which ones are of higher value for a given business problem.
- Providing guarantees that data is accurate for regulatory compliance, and for which data lineage and data audits have to be captured and shown accurately.
- In previous sections, much has been discussed about differences between EDW and data lakes but, for most regulatory industries, not having an EDW is not an option, as regulations mandate the need of quick and real-time access to structured data through a carefully designed set of queries. In such cases, business and IT teams need to augment the EDW with the data lake, which helps with providing insights on many analytic use cases where models and queries may not have been prepared upfront.

Technology Challenges

- The data lake by definition lacks structure for data and it may not be easy to classify and categorize data without use of appropriate tools. The same goes for data lineage, data quality and security. In addition to tools, proper processes need to be in place for these topics.
- The speed of evolution of Hadoop ecosystem is remarkably swift. Almost every week there are major updates of many tools and technology and possible new ones coming every month. It becomes difficult for the architecture and developer community to keep pace with this speed of evolution and make appropriate implementation suggestions to customers.
- When data lakes are implemented to ingest and manage data from software defined things (IOT), this will demand a large set of connectors and protocol capabilities by the data ingestion tool. Not only is there a new variety of data sources, but within the same type of connector there may be different formats/schemas to deal with, which are due to version differences between the devices at the other end of the connectors.
- While governance tools and metadata management tools are employed, the operational team still needs to ensure and comply with different regulatory requirements for respective domains, as well as privacy regulatory requirements, which is an added challenge.

7. Six Best practices for implementing the Data Lake

We expect readers to regard data lakes as a technology solution to the problem of getting a timelier, more complete and more flexible set of insights through analyzing data assets. As it can be noticed from the previous challenges section, there are various business and operational related problems which need to be resolved first. Based on our implementation experience, best practices are needed while implementing the data lake, and which we summarize in this chapter.

1. Train and Prepare Business Users & Data Scientist for the data lake

Business users will use the data lake to get new insights and aggregated views of the business using business intelligence visualization tools. They would also like to co-relate data sets in an ad-hoc manner along with ability to ask “what-if” questions on their own. At the executive level, business users would like to get 360 degree views of customers and partners, from Marketing to Support including their own organization, in single view.

- Empower your business users so that they become self-sufficient on how to use metadata to discover the data in the lake, find about its origin, assess its quality level through inspecting quality metadata and/or profiling, and explore it with the tools they can use. The use of self-service data preparation tools [12] could also be evaluated for business users. This needs to be decided depending on organizational factors, skillsets and resources.
- Associate your business users to the process of giving feedback for quality of content, and creating business-specific metadata tags (including specifying synonyms) which will help in the data lake maturing process.
- Make sure that as new datasets are generated by business users and/or data scientists, these are appropriately tagged by their authors so that insights can be located and reused.
- Ask business users to provide a ranking for value of a dataset: some kind of well agreed tag and corresponding possible values, to indicate whether it was useful, and for which goal or initiative.

© 2017 Persistent Systems Ltd. All rights reserved.

2. Data Architect should build a flexible data lake platform

- Data Architects should choose technologies which have higher user adoption and, most importantly, those that are easier to build and develop using the tools provided.
- Look at big data architecture patterns for implementing the data lake[3]. Don't hesitate to have an enterprise data warehouse co-exist with the data lake (in fact most of the time, they will co-exist given they are solutions to different needs, especially in large enterprises).
- Tailor the architecture to your specific industry, ensuring that capabilities that are standard and necessary for your domain are inherent part of the design.
- Build a design that is guided by disposable components integrated through service APIs.

3. ETL Architects must manage the data ingested to support governance

In the data lake world, the ETL architect still has the role of extracting and loading the data. The transformations are done after storing the data in the HDFS layer. While there are tools which provide these capabilities, ETL architects need to manage the data ingested to support data governance.

The ETL architect hence needs to follow the following practices, to this end:

- Identify the data owner (data producer) of the ingested data. Get from this person the requirements for creation, usage, privacy, regulatory, and encryption business rules that apply to the incoming data, and document them. Document the context, lineage, and frequency of the incoming data.
- Identify the data steward(s) charged with driving the agreement on data standards, validation and business rules monitoring the health of datasets from specific data sources.
- Make sure the ingestion framework has a metadata management capability to provide metadata such as source provenance, onboarding timestamp and structural metadata extraction automatically.
- Create a catalog for datasets ingested in the data lake. Organize and categorize data in the catalog as it arrives so that user can search for it, entering values for properties such as onboarding time, source provenance, subject area, schema attributes, and the like.
- Continuously measure data quality as it resides in the data lake, especially those datasets that have been curated for consumption and for which there are data pipelines refreshing the dataset with updates taking place at the data source.

4. Define and Document Governance Attributes of the Data Lake

The data lake Administrator needs to ensure that the right governance tools are installed in the data lake environment. The administrator not only needs to ensure data is protected and secured, but also is being tracked and tagged for search.

- Make sure governance tools record data lineage as it flows and processed through the system. Create a rule to prevent data from entering the data lake if its provenance is unknown.
- Classify the security level (public, internal, sensitive, and restricted) of the incoming data. Limit access to sensitive data to those with the proper authorization in the data lake.
- Data may need to be either masked or tokenized, and protected with proper access controls. An integrated management platform can perform masking (where data from a field can't be visualized) and tokenization (changing parts of the data to something innocuous). This type of platform ensures you have a policy-based mechanism, like access control lists, that you can enforce to make sure the data is protected appropriately.
- Automate capturing of column profiles, and column correlation.
- As raw data arrives in the lake, profiling will uncover a certain amount of errors, as well as non-compliance to formats and standards established by the data steward and required for end-user analysis. This means you need to curate this data before data scientists start looking at it for some meaningful discovery, or make sure that your data scientists are able to handle this task. As the business value of the data becomes clear (which may be at ingestion time, or at a later point in time), create workflow orchestrations to curate this data in HDFS before loading into Hive or NoSQL databases for analysis.

[3] A good example is lambda architectures, for supporting big data applications with fault tolerance, linear scale out capabilities, and low latency queries with updates that can come at any velocity: both batch and real time streams of new or updated data.

© 2017 Persistent Systems Ltd. All rights reserved.

5. Allow Data Scientists to Discover and Explore freely and apply Machine Learning on entire data sets.

Data scientists perform the role of discoverer and explorer in the data lake. They discover new insights by using various dataset discovery, profiling and exploratory tools. They look at the datasets and try building models on top of them so that they can perform exploratory ad hoc analysis and machine learning algorithms to come up with some hypothesis. The goal is to provide initial insights that help in determining a real business need.

- Make sure the data scientist has an environment where flexibility can be provided with lesser amount of governance to interact with data in real-time, and where the he/she can build scripts in some scripting language (e.g., Python or R), and run it on a cluster to get a response in minutes or hours, rather than days.
- Depending on the industry, data quality assessment may need to go beyond basic profiling. Machine learning scripts are increasingly being written by data scientists used to detect errors in the data that are unlikely to be found in a timely fashion through normal consumption and where the cost of latent errors is high[4].
- If the data for his analysis is too dirty, she may need the help of the IT developer to clean it up, or she may do that herself. Self-service data preparation tools may come in handy, although data scientists are technically savvy and may do without them.
- Let data scientists also decide on exploration or visualization tools.

6. Support the Data Lake with Robust Monitoring and Operational Tools

The data lake operations manager is responsible for managing and monitoring the data lake cluster in all environments.

- Create and enforce policy-based data retention. Space may become an issue over time, even in Hadoop; as a result, there has to be a process in place to determine how long data should be preserved in the repository, and how and where to archive it.
- Monitor data quality for those data sets that have been curated and for which there is a data pipeline bringing in regular updates from the sources.
- Monitor to assess the effectiveness of regulatory compliance (for instance, on PII).
- SLA management is very important in the enterprises. Make sure it is defined as per the cluster and available environment requirements. You can apply different storage mechanisms based on SLA requirements.
- Monitor and measure the clusters effectively even if it means additional resources, as it could lead to disaster without it.

8. Conclusion/Summary

Driven by the explosion of available data and effective technology to manage it, enterprises are now consuming data and analytics based on data in unprecedented ways. The rise of data driven decision making is real. And it's spectacular.

Data lakes is an emerging approach to extracting and placing all data relevant for analytics in a single repository. All data means data internal to the organization and external to it, both big and “small”. Data lakes are an alternative to data warehouses to put an end to data silos in an organization, which is one of the biggest impediments to effective data-driven decision making. Their biggest advantage is flexibility: by ingesting and storing data in native format, a far larger pool of data will be available for analysis, even if clear business needs are not initially identified. Data scientists can explore the data and, through data mining workloads, provide initial insights that help in determining such business needs. The availability of technology and tools is enabling IT departments to deploy data lakes in the organization, which in turn is helping solve the mismatch between the needs of data producers and those of consumers as described in this white paper.

We have seen that to implement a data lake, data governance is one of the top priority for organizations. Data lake governance is based on the organization's ability to create and manage metadata at the different levels we described throughout this paper. Data scientists are a technically savvy consumer audience who can contribute significantly to the creation of such metadata but, in order to capitalize on the opportunity that the lake presents, both data producers and especially the wider audience of business user consumers need to come on board.

[4] A good example are constraints that could be learned from the data because they represent deviations from the norm (as opposed to logical constraints, which require relatively deep knowledge about the governing processes and cannot be derived from the data).

© 2017 Persistent Systems Ltd. All rights reserved.

Meeting their needs starts by making these users self-sufficient on how to use metadata to discover the data in the lake, find about its origin, determine its quality level, and explore it with the tools they can use. It then continues with associating them to the process of validating the content from the quality point of view, identifying synonyms and creating business-specific metadata tags which will help the data lake maturing process.

We hope this white paper has given you the tools and confidence needed to not just dip in a toe, but to jump all in to the data lake, as the ripple you create will have a far reaching impact on your organization.

Appendix 1 - Technology and vocabulary of data quality

- **Data Profiling** refers to the inspection of data elements as well as the relationship of data elements among them. It delivers statistics about the data, which provide insight into the meaning and quality of data, and helps in identifying data quality issues.
- **Data Standardization** refers to the process of reaching agreement on common data definitions, representation and structure of all data elements and values, as well as formatting of values into consistent layouts.
- **Data Validation** is the execution of the set of rules that make sure that data is clean. These rules return Boolean values indicating whether the data row or element verifies the criteria, or not. Some of these rules are those defined or implied by data standardization, but business rules may be more involved than that (e.g., "imported products need to have tariff data").
- **Data Cleansing** is the process of fixing of non-valid data to make it valid, or clean(er). Cleansing is generally about applying transformations to data values to meet the agreed data standardization, integrity constraints or other business rules.
- **Data Matching** and **consolidation** is the task of identifying, matching and merging records that correspond to the same real world entities. This typically happens when bringing data from different data sources together, but may occur within the same data source. This task reduces data duplication and improves data accuracy in the lake.
- **Data Enrichment:** The enhancement of the value of internally held data by appending related attributes from external sources (for example, consumer demographic attributes and geographic descriptors).

References

[1] Business Analytics: The Next Frontier for Decision Sciences. Evans, James R., Lindner, Carl H. (March 2012), Decision Line. 43 (2). http://connection.ebscohost.com/c/articles/74427370/business-analytics-next-frontier-decision-sciences

[2] https://www.fr.capgemini.com/resource-file-access/resource/pdf/The_Deciding_Factor__Big_Data___Decision_Making.pdf

[3] https://www.linkedin.com/pulse/how-flexible-should-data-ingestion-layer-vishal-toshniwal

[4] https://www.linkedin.com/pulse/tags-first-read-fast-step-discover-explore-enrich-your-agrawal

[5] https://www.linkedin.com/pulse/effective-data-governance-key-controlling-trusting-quality-agrawal

[6] https://www.linkedin.com/pulse/let-data-flow-securely-through-lake-akshay-bogawat

[7] http://www.kimballgroup.com/2008/08/slowly-changing-dimensions/

[8] http://www.dwh-club.com/dwh-bi-articles/change-data-capture-methods.html

[9] http://tdan.com/database-auditing-capabilities-for-compliance-and-security/8135

[10] https://www.linkedin.com/pulse/data-exploration-finding-treasure-lake-chandraprakash-jain

[11] http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

[12] Improving Data Preparation for Business Analytics, David Stodder, TDWI research report, July 2016, https://tdwi.org/Webcasts/2016/07/Improving-Data-Preparation-for-Business-Analytics.aspx

[13] http://www.gartner.com/newsroom/id/2809117

© 2017 Persistent Systems Ltd. All rights reserved.

About Persistent Systems

Persistent Systems (BSE & NSE: PERSISTENT) builds software that drives our customers' business; enterprises and software product companies with software at the core of their digital transformation. For more information, please visit: www.persistent.com

India
Persistent Systems Limited
Bhageerath, 402,
Senapati Bapat Road
Pune 411016.
Tel: +91 (20) 6703 0000
Fax: +91 (20) 6703 0009

USA
Persistent Systems, Inc.
2055 Laurelwood Road, Suite 210
Santa Clara, CA 95054
Tel: +1 (408) 216 7010
Fax: +1 (408) 451 9177
Email: info@persistent.com

DISCLAIMER: "The trademarks or trade names mentioned in this paper are property of their respective owners and are included for reference only and do not imply a connection or relationship between Persistent Systems and these companies."

© 2017 Persistent Systems Ltd. All rights reserved.