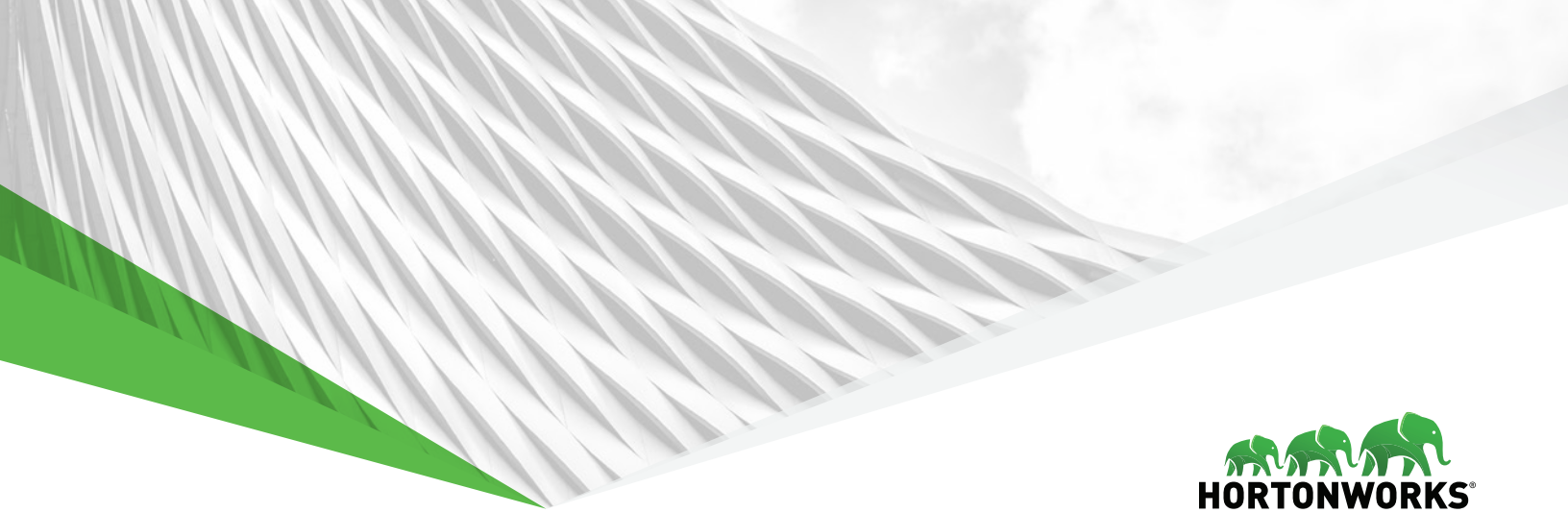

DIGITAL TRANSFORMATION AND THE NEW DATA QUALITY IMPERATIVE

Michael Ger
Richard Dobson

A HORTONWORKS WHITE PAPER
AUGUST 2018

 ©2018 Hortonworks

Contents

Digital Transformation

Today, change is transforming companies like never before. Critical business processes, once planned and rigid, are being transformed into processes that continually change based on real-time conditions. And companies who do not change do so at their own peril. According to Standard and Poor's, 88 percent of companies within the Fortune 500 in 1955 are now gone. Furthermore, according to Accenture, the failure to digitally transform is the primary reason that half of these companies have disappeared.

New Data is Key

As companies embark upon their digital transformation journeys, they quickly come to recognize *the critical role of data*. Data is increasingly seen as the new "oil" powering digital transformation, enabling real-time, learning-based decision-making across all business operations, including product development, manufacturing, supply chain, and customer experience.

CONNECTED WORLD

Perhaps no other factor has accelerated digital transformation so profoundly as the emergence of the *connected world*. Recent decades have ushered in an explosion of connected people, devices, and processes. Consider the following facts:

- **Connected people:** According to the United Nations, the number of people connected to the internet has grown 700 percent in the last five years alone.
- **Connected devices:** The number of connected devices will grow 3000 percent by 2020, according to Gartner.
- **Connected factories:** The number of connected industrial IoT devices will grow 400 percent by 2021, according to Berg Insight.
- **Connected vehicles:** PWC estimates that connected vehicles will grow by 300 percent by 2022.
- **Connected inventory:** According to McKinsey, the RFID market will expand from $12 million to $209 billion by 2021.

Growing Data Complexity

Concurrent with the world becoming more connected, *data complexity* is on the rise as well. Consider the following trends:

DATA TSUNAMI

As the connected world transforms, the traditional definition of "enterprise data" is evolving as well. Consider how new "connected world" data is ushering in a data tsunami characterized by unprecedented data volume and variety.

- **Data volume:** According to IDC[1], the size of the global datasphere (the amount of data society generates, uses, and retains) will grow from 16 ZB in 2016 to over 160 ZB by 2025, a tenfold increase.
- **Data variety:** Data variety will also explode, consisting of data from traditional *structured* RDMS-related enterprise transactional systems (PLM, ERP, CRM), as well as new *unstructured* or *semi-structured* data sources (IOT connected devices, web clickstreams, social, sensors, GPS, logs).

REAL-TIME DATA

As the world becomes more connected, it is increasingly leveraging *real-time* data as well. According to IDC[2], real-time data generation is projected to grow at 1.5 times the rate of overall data creation through 2025, amounting to almost 30 percent of all data within the datasphere.

Furthermore, as organizations embark upon their digital transformation journeys, real-time data represents the critical ingredient enabling real-time, adaptive value chain decisions and processes. As evidence, consider the following facts:

- Digital supply chains respond 25 percent faster due to real-time information (BCG).
- Sixty-four percent of consumers and 80 percent of businesses (B2B) now expect real-time communications. (Salesforce.com)
- Not surprisingly, real-time streaming analytics is projected to grow by 34.8 percent CAGR through 2021. (MarketsandMarkets)

INTERRELATED DATA

As the world becomes increasingly interconnected, so too is data becoming increasingly *interrelated*. As new data sources continue to grow, new "layers" of data are continuously being associated with data entities.

Example: *Interrelated Data Critical to Achieving Customer Centricity*

To illustrate the concept of interrelated data, consider the example of a company attempting to differentiate itself by providing new levels of *customer centricity* and personalization of the customer experience. Importantly, delivering next-generation customer centricity requires new levels of data *centricity*. As can be seen in the illustration below, achieving a 360° view of a customer requires multiple *layers* of data.

1. Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big International Data Corporation (IDC) and Seagate as part of an IDC continuous intelligence service
2. Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big International Data Corporation (IDC) and Seagate as part of an IDC continuous intelligence service

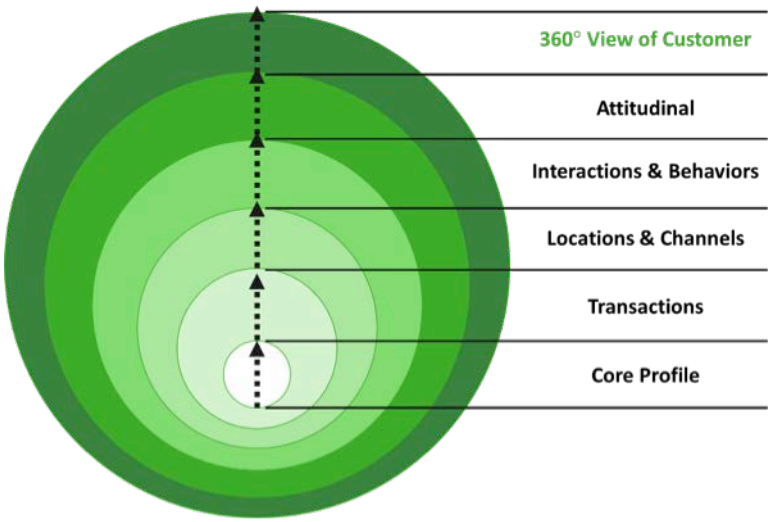

Figure 1 : Levels of customer centricity and personalization

As illustrated, each layer of data describes different customer information, and is often represented by different data types. The following table describes the layers required for a true 360° view of a customer:

LAYER	INFORMATION	DATA TYPE
Core Profile	CRM/MDM profile	RDBMS profiles
Transactions	CRM/ERP transactions	RDBMS transactions
Locations & Channels	Geo-location and device channel	GPS, IoT, etc.
Interactions & Behaviors	Web searches, considered yet abandoned purchases	Web clickstreams, logs, etc.
Attitudinal	Customer likes, dislikes, complaints	Call center, social, etc.
360° View of Customer	ALL OF ABOVE	ALL OF ABOVE

Figure 2 : Layers required for a 360° view of a customer

DATA AND GOVERNANCE SILOS—EVERYWHERE

As companies grapple with "new-age" data requirements (data volume, variety, and interrelationships), they continue to struggle with "age-old" and far more mundane challenges born from the globalization and merger and acquisition activities of the last few decades. From a systems perspective, such expansion has led to two unanticipated outcomes: First, *data silos* have increased dramatically as companies, unable to implement single-instance data solutions incorporating global needs, rapidly implement additional (and often redundant) systems to meet narrow regional requirements. Second, such decentralized data management practices have also resulted in *siloed approaches to data governance*. Data and governance silos often drive a wide range of negative business outcomes. Consider the following examples:

Example #1: Data Silos Impeding Basic Reporting
As global competition intensifies, companies are increasingly attempting to differentiate themselves via superior customer experiences. Core to this goal is fully understanding all interactions between the company and its customers (e.g., global understanding of customer purchase, return, and service histories). However, as is often the case with globally expanding companies, non-integrated regional systems render obtaining a *360-degree view of a customer* impractical, given the system fragmentation and inconsistent data across these systems. For example, is *John Doe* in one system the same as *Jon Doe* in another system? In this case, poor customer data quality (the inability to definitively identify a customer) results directly in the inability to understand customer interactions, critical to providing a superior customer experience.

Example #2: Governance Silos Impeding Data Science
Companies are increasingly collecting vast quantities of data in *data lakes* and providing this information to data scientists for a wide range of data discovery and machine learning purposes. However, while big data analytics is a relatively new phenomenon, companies often repeat the mistakes of the past by *failing to establish data governance policies within the data lake*.

For example, surveys suggest that data scientists spend up to 80% of their time finding and preparing data. Consider the scenario of two data scientists performing the same data science experiment. Each downloads data into Excel and manually corrects and manipulates the data. One approaches the correction task slightly differently than the other, resulting in inconsistent analysis results. Which is to be trusted? Time is spent correcting and validating prior analyses. Imagine the impact to data trustworthiness and resource productivity as similar analyses proliferate across the enterprise.

REGULATORY REQUIREMENTS

As data enables digital transformation across industries, organizations assume new risks (both financial and legal) in storing ever-increasing quantities of data. As a result, new *regulatory requirements*, once affecting only a few key industries, are now becoming commonplace for all. The scope of the new General Data Protection Regulation (GDPR), for example, applies to all companies in the world storing the contact information of European citizens. Per the requirement, EU citizens can, at any time, request explanations regarding the use of their personal data. A Data Protection Officer has just 30 days to respond. Consequently, organizations that are unable to trace *what* information is stored *where*—in addition to what information usage was consented to—place themselves at significant compliance risk. It's also worth noting that the worst case scenario associated with the failure to comply is a fine of up to four percent of global revenues, which is why compliance is seen as a critical C-level priority.

ON-PREMISES AND IN THE CLOUD

Adding still greater complexity to the overall landscape, data is increasingly being managed in multiple *tiers*, consisting of both *on-premises* and *cloud-based* operational models. Interestingly, according to IDC[3], although both tiering options will remain prevalent, cloud-based software growth (27%) is expected to outpace on-premises growth (5.5%) through 2021.

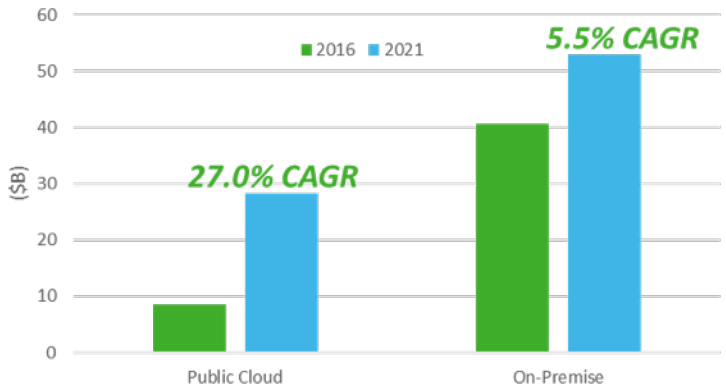

Figure 2 : Data management growth: cloud-based and on-premises

The Enterprise Data Quality Imperative

For reasons outlined in the previous section, organizations today are facing unprecedented data volume, variety, and complexity challenges. As a result, many are revisiting the subject of *data quality*.

While poor data quality has plagued companies for many decades, today's digital transformation is driving a *data quality imperative*. And the reason for this is clear: *For companies to digitally transform and successfully deliver real-time, responsive processes based on data, the underlying quality of that data is ever more critical*.

3. MARKET FORECAST
Worldwide Big Data and Analytics Software Forecast, 2017–2021
International Data Corporation (IDC)

Traditional Approaches to Data Quality

Historical approaches to managing data quality cannot deliver the order-of-magnitude improvements required for today's digitally transforming organizations. In general, traditional efforts have fallen short in several critical areas:

TOO LABOR INTENSIVE

Over the last few decades, organizations across a wide range of industries have embarked upon data quality improvement initiatives. These efforts have tended to be limited in scope due to the sheer labor *intensiveness* required to make progress. Consider typical tasks required for data quality improvement initiatives:

1. **Data schema definition:** Before data is even touched as part of a data quality initiative, a schema must be defined and agreed upon that defines the metadata, tables, and fields for the target data. This is often a lengthy process, requiring cataloging existing data and considering existing data formats. Often even more time consuming, consensus must be achieved across different business units on future data requirements and formats.

2. **Data extraction, transformation, and load:** Once target data schemas are agreed upon, extraction, transformation, and load (ETL) flows must be created, tested, and implemented. Given a potentially large number of source systems across numerous data entities, ETL processes can consume considerable staff time.

3. **Data standardization:** Perhaps the most time-consuming aspect of data quality initiatives involves standardizing the content (naming conventions, lists of values, units of measure) across data sets. While data standardization is often a manual process, tools to automate the cleansing of large data sets exist in the marketplace. However, many of these tools introduce time-consuming processes of their own—for example, rules must be manually defined to standardize data from a given input format to the standardized output format.

   The *importance of data standardization cannot be overstated*, as it is foundational to downstream reporting. Without standardized data, accurate reporting is impossible, due to non-conforming data sets and inaccurate searches, queries, and reporting aggregations.

CAN'T WITHSTAND THE DATA TSUNAMI (VOLUME AND VARIETY)

Traditional data quality approaches are unable to address the expanded scale (volume) and *scope of data* encountered in today's digitally transforming enterprises. More specifically, traditional approaches have focused only on sub-segments of the broader data constellation:

1. **Only traditional managed data types:** As noted, data variety continues to explode with the advent of "new" data sources such as IOT, web clickstreams, social, video, and images. In contrast, previous data quality efforts have typically focused purely on master data, broadly consisting of information pertaining to things (products, services), parties (customers, suppliers, employees) and locations. While master data is of critical importance, it represents only part of the story (the innermost "Core Profile" data illustrated in Figure 1, outlined above).

2. **Only data-at-rest:** Over the last few decades, data professionals have expended great effort on improving the quality of *data-at-rest*—transactional data living within OLTP databases and data warehouses. However, with the advent of IOT and other big data sources, attention must now shift to the quality of *data-in-motion*, as streaming analytics becomes foundational for taking real-time, contextual actions across value chains—a critical goal of digital transformation.

3. **Only a limited number of systems:** Due directly to the large-scale effort associated with data quality projects, such initiatives typically prioritize only a few key systems within the organization, while postponing actions on remaining systems until some elusive future date. Unfortunately, these systems often become orphaned, never receiving the data quality attention originally planned for. This results in more silos of data, with inconsistent data quality across the silos.

   And this is only exacerbated by industry's *shift to the cloud*. Companies struggling to get a handle on data quality "within their own four walls" must now contemplate strategies to manage data quality across internal and cloud instances.

TOO RELIANT ON QUICK FIXES

For the reasons outlined above, data quality initiatives tend to be fleeting. While "big bang" initiatives to improve enterprise data quality are launched with great fanfare, they often fade over time, ultimately replaced by numerous quick fixes. For example, a data quality initiative may be launched to de-duplicate and enrich data within the warehouse. This difficult and expensive exercise is often approached as a "quick-fix" initiative, with cleansing occurring only for a "snapshot" in time. As new data again accumulates over time, the quality of the data decreases until all trust is lost, and the process inevitably starts again. This results in the "trust decay" pattern outlined below.

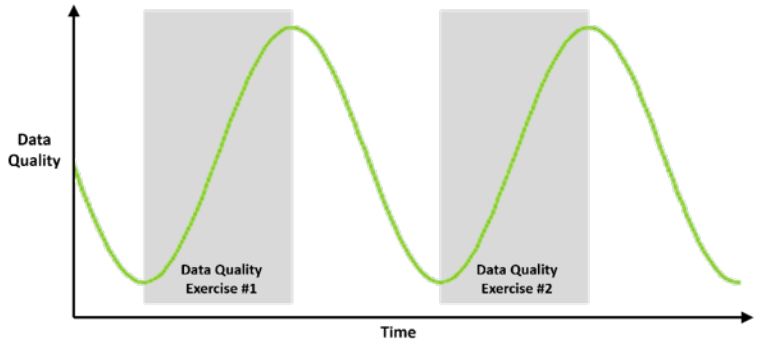

Figure 3 : Trust decay pattern

Open Source Innovation for Enterprise Data Quality Management

While traditional approaches to addressing enterprise data quality have proven insufficient for digitally transforming organizations, the good news is this: modern advances in information technology, led by open source communities, are providing significant capabilities for achieving data quality management nirvana. The following sections describe advances in three primary areas:

1. Modern data management

2. Data assurance management

3. Machine learning

MODERN DATA MANAGEMENT

Data is the new oil, whether from new or traditional data sources. Traditional data management technologies (Relational Databases, Data Warehouses, etc.) are not up to the task of managing the volume and variety of data required for next-generation digital transformations. The following sections describe open source data management innovations across the *data lifecycle* (data ingestion, storage, and processing).

MODERN DATA INGEST

While data ingest from traditional relational database systems is well understood, ingesting real-time data (IoT, sensors, web clickstreams, social, etc.) is less so. Data from these sources are ingested using various protocols, transformed as necessary, and routed for downstream processing. Data ingest challenges typically include development of the ingest flows, applying appropriate security, providing reliable flow control (given often suboptimal network connections), and guaranteeing the actual delivery of messages.

Happily, with the advent of modern, open source solutions such as Apache NiFi, data ingestion flows can be easily created via an intuitive visual user interface. In addition, capabilities are provided to aggregate, prioritize, and compress data, readying it for downstream data storage and processing requirements. More critically, capabilities have been extended to *encrypt* data, ensuring data security, and *buffer* data when system interruptions (e.g., network interruptions) occur, guaranteeing delivery of data.

MODERN DATA STORAGE

- **Universal data storage:** Technologies such as Apache Hadoop provide the ability *to store all data within a secure enterprise data lake*. Gone are analytics limitations imposed by the data silos of the past—replaced by the ability to store *all* data in a centralized, easy-to-access location. Better yet, data storage is distributed across multiple commodity storage nodes, providing extremely inexpensive "scale-out" of storage capacity.

- **Schema-less data storage:** As previously discussed, historical approaches associated with enterprise data-quality initiatives were entirely too labor intensive to be practical. A conspicuous example involved the need to predefine a data schema as a prerequisite to both extracting and storing the data to be processed. Directly attributable to this onerous requirement, data quality initiatives often either languished or stalled out completely in the early stages of a project.

  Open source Hadoop eliminates this inefficiency, as *data storage within a data lake eliminates the need to first define a data schema.* As a result, data can simply be loaded into the data lake in its raw form, and transformed only when needed for analytics. As a result, moving data into a centralized location for analytics has never been easier.

MODERN DATA PROCESSING

- **"In-place" data processing:** While traditional methods for executing data quality workloads (data wrangling, standardization, matching) can provide great benefits, they also introduced significant challenges. First, such approaches typically required extracting large amounts of data from wherever it was stored and transferring it to the data quality application for processing. This created significant administrative overhead, involving moving the data, securing the data in both locations, and keeping the two data sets synchronized. However, *what if you could perform your analysis directly where the data was stored, and not move it at all?* Hadoop not only provides a mechanism for centralized data storage, but also provides the ability to perform *data processing* workloads directly within the data lake cluster, leveraging technologies such as Apache Spark—eliminating the need to extract and move data prior to data processing.

- **Lifecycle data processing:** Not surprisingly, in the pre-connected world (pre-IOT, pre-web), data quality solutions focused on improving the quality of *data-at-rest* (transactional data sitting in relational databases), with data quality processing typically applied to batches of stored data as needed.

  However, given the rapid growth of real-time data sources and streaming analytics, data quality processing must also be applied to *data-in-motion*. Happily, modern open-source data processing engines such as Spark, in combination with Spark Streaming, provide the ability to process both data-at-rest and data-in-motion. Additionally, data can be processed in multiple modes (batch, interactive, or streaming). Taken together, these capabilities provide the ability to *process data across the complete data lifecycle*.

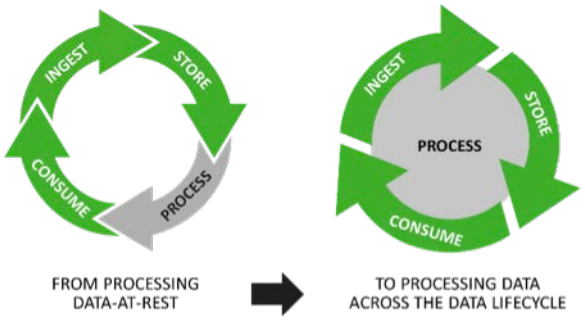

Figure 4 : Processing data across the data lifecycle

DATA ASSURANCE MANAGEMENT

While data quality initiatives have been approached for decades within organizations, more often than not they have been attempted in a piecemeal manner. Like a chain, however, the resulting data quality was only as strong as its weakest link. This was exacerbated by the fact that multiple proprietary data quality (DQ) applications were often leveraged within DQ projects, leading to complex, siloed, and expensive system deployments.

Today, modern open source solutions are increasingly being considered for end-to-end *data assurance management*. What constitutes data assurance management? The following sections outline the critical characteristics required for robust data assurance management.

DATA GOVERNANCE

Data governance concerns itself with instilling *trust* in the information provided to the enterprise. It is critical that well-governed enterprise systems provide the following capabilities:

- **Data security:** The ability to precisely control the data that people and systems have access to, and what they can edit. Such privileges should be fully auditable from a regulatory compliance perspective.
- **Data lineage:** The ability to track specific data (at the record level) across the data lifecycle, including the data's origin and how it changes over time as it moves through diverse processes, requests, and systems.
- **Review and approval:** The ability to construct auditable, rules-driven data definition and approval workflows further drives data standardization and accountability.

Given today's data volume, variety, and deployment (cloud and on-premises) complexities, managing enterprise data governance has never been more challenging. However, new *global data management* capabilities associated with open source data management applications have incorporated the ability to *centrally manage data governance across the entire data lifecycle*, ensuring the ability of enterprises to comprehensively govern data across a wide range of data assets.

	DATA LIFECYCLE			
	Ingest →	Store →	Process →	Consume
Data Security	✓	✓	✓	✓
Data Lineage	✓	✓	✓	✓
Review & Approval	✓	✓	✓	✓

Figure 5 : Centralized management of data governance across the entire data lifecycle

DATA QUALITY

Data quality describes the extent to which information can be used as a *data asset*. For example, high-quality data for a business entity must be unique, accurate, and complete. When achieved, such records are often referred to as "golden records." In pursuit of this goal, data quality systems must provide the following:

- **Data matching:** The ability to identify duplicate data, whether the data is structured, semi-structured, or unstructured. To be truly effective, matching should be provided across multiple modes (real-time, batch, and micro-batch) and data types (i.e. matching people with their transactional and attitudinal data)

- **Data merging and survivorship:** Often duplicate records may exist for a given data entity (e.g., a customer), with one particular record providing the "best" information for a given data attribute (e.g., current address). Data merging and survivorship functions provide the ability to determine which specific record provides the best data for that attribute and selects that information to "survive" in a merged golden record. Rules should be defined to either automatically create golden records or, when not possible, allow users to manually define what (if any) merge actions can occur. In addition, capabilities should exist to merge records both in bulk mode for large batches of data, and in streaming mode to support real-time data processing use cases.
- **Data discovery and profiling:** Ideally, all data should be *cataloged* and *categorized*, providing the ability to apply security and other business rules by category, enabling effective mass change capabilities on large data sets. In addition, it should be possible for profiling rules to identify specific data to be cleansed or processed. Examples of such rules include assigning metadata quality scores, specific data tags, or automatically invoking technical or business processes.
- **Data cleansing:** Provides the ability for data to be *corrected*, promoting standardization and consistency. This can be achieved either by individuals manually cleansing data, programatically via automated cleansing tools, or via a user interface and integration with third-party data providers (e.g., Post Office, D&B). To accomplish manual data cleansing at scale, it should be possible to define batch cleansing rules that can be tested on sample data sets prior to bulk updates of complete data sets.
- **Data enrichment:** Provides the ability to *add additional information* (e.g., address information, social sentiment) to a data entity (e.g., a customer record) either manually, automatically, or through integration with third-party data providers.

DATA VALIDITY

Data validity implies that data conforms to *rules* defined by organizations. Data rules may be simple (e.g., values must comply with a list of values) or vary based on more complex functions. Often, rules may vary by data *lifecycle* stage. For example, a product in a "definition" stage may have only a few mandatory attributes, but in an "approved" stage may have many more mandatory attributes (e.g., hazard notices and financial information). Data validity rules typically fall into one of the following categories:

- **Business rules:** As part of a documented data lifecycle definition, rules should be defined to assure current validity for each lifecycle stage. Business rules should be defined from a business user's point of view. For example, what constitutes a valid customer in a financial context is very different from a customer retention/marketing context.
- **Process rules:** Per data lifecycle stage, rules must be established regarding *how data is accessed* (view, update, delete) and *where data can flow* (with data lineage and audit providing the ability access and track data changes). Open source applications such as Apache NiFi provide the ability to define data processing rules that are version controlled and auditable, ensuring data processing actions are consistent and governed.
- **Batch rules:** When data is onboarded in batches, rule sets must govern the specific business and process rules to be applied in bulk. Batch rules typically also dictate how invalid data is identified and processed.

DATA RELATIONSHIPS

As previously discussed, the emergence of so many new data types (e.g., social, web clickstream) is creating new *interrelated* layers of data associated with core data entities (e.g., customer). To manage data relationships more fully, capabilities are often provided to define relationship types:

- **Linking:** Provides the ability to *directly join* data together, similar in concept to a traditional relational database foreign key relationship, where a direct reference between the data exists in either a 1-to-1 or a 1-to-N cardinality.
- **Association:** Describes the ability to create links between data based on business-defined relationship types. These types of relationships can be used to create relationships between structured, unstructured, and semi-structured data (e.g., linking customer records with emails or call center recordings) with customer records, or between data with more "loose" associations.
- **Hierarchies:** By providing a parent/child–oriented graphical user interface to traverse relationships between a business context (i.e., customer, company, geography) and other data (structured, semi-structured, or unstructured), a true 360° view of data can be achieved to provide user insights. Because the human brain naturally thinks hierarchically, it's possible for users to efficiently inspect very large volumes of data.

MACHINE LEARNING

Given the volume and real-time nature of today's data environment, traditional labour intesive attempts to address data quality concerns are simply not viable. A new approach is needed that can scale to the task at hand.

Machine learning (ML) represents a great opportunity to enable data quality management scale and efficiency. For example, rather than having staff members manually define data quality rules, algorithms can automatically derive these rules through learning based on observed data samples. Machine-learning algorithms can be applied to a broad range of tasks, including adjusting match rules to improve accuracy, learning from manual data-merging actions to create automated actions, and learning to automatically discover, categorize, assign, and process data. More specifically, ML can be used to achieve scale for the following processes:

- **ML-optimized data discovery:** This capability provides two functions. First, data must be identified and categorized from a business point of view. Second, it must derive actions that need to be taken on the data to provide business value. Machine learning can also be leveraged to discover what actions have been taken previously by systems and users, and then leverage this knowledge to take current, real-time actions.
- **ML-optimized processes:** Many organizations pre-define workflows to manage processes. However, machine-learning algorithms can provide far more adaptive controls by optimizing processes based on *previously experienced* bottlenecks (e.g., redirect approvals to more responsive channels).
- **ML-optimized linking:** Typically, linking of data is an extremely manual process, especially for softer context-specific relationships. ML algorithms can be leveraged to "sift" through data lakes, creating real-time linking across high-volume and high-velocity data.

Why Open Source?

Over the last few decades, open source communities have had a major impact on industry, with benefits being accrued across a multitude of enterprises and business use cases. Given the big data management challenges described in this white paper, we believe that data quality challenges have never been more acute, and that *open source data management for enterprise data quality makes a convincing case for itself*. Consider the following points:

1. **Economic reality:** According to IDC[4], the digital universe amounted to 16ZB in 2017, with approximately 30 percent of that data residing in enterprises. Given the huge resources required to leverage this data tsunami, the economic efficiencies associated with open source solutions will increasingly drive enterprise data platform selection decisions.

2. **Increased innovation with lower risk:** Given the importance and trajectory associated with big data analytics, machine learning, and artificial intelligence, leading enterprises are increasingly reluctant to "lock-into" proprietary technologies from commercial software vendors. For these enterprises, open source leadership is the answer, with its track record of broad, community-led innovation that no single company can sustain.

4. Data Age 2025: The Evolution of Data to Life-Critical
Don't Focus on Big Data; Focus on the Data That's Big
International Data Corporation (IDC) and Seagate as part of an IDC continuous intelligence service

Open Source Architecture for Enterprise Data Quality

As outlined in previous sections, next-generation enterprise data management solutions must provide modernized capabilities in three specific solution areas: core data management, data assurance management, and machine learning. The open source community has provided significant contributions in each of these areas. The following illustration maps Apache open source solutions to the critical solution areas mentioned above. Future white papers will provide greater technical detail for each of these areas.

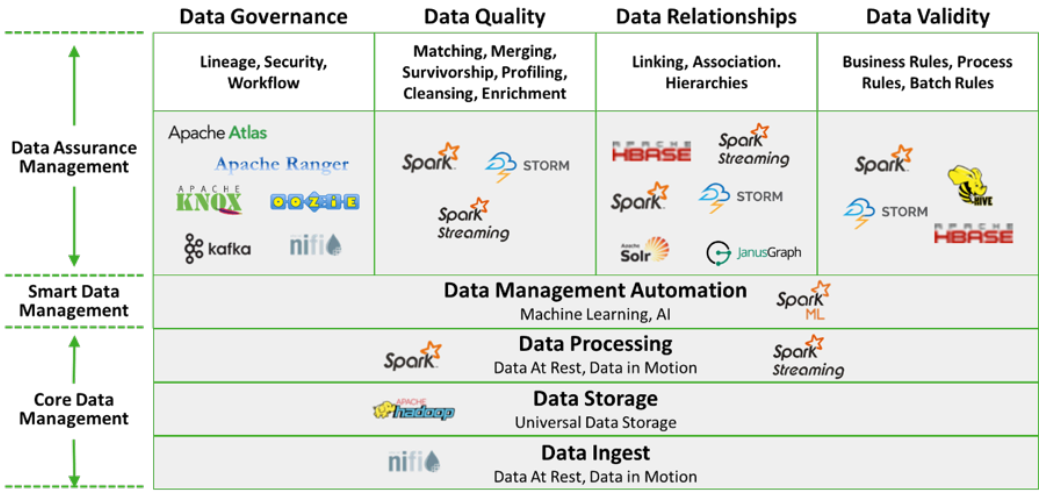

Figure 6 : Apache open source solutions

Conclusion

Companies across all industries are reimagining themselves within a digitally transformed future. Central to that future is leveraging a data tsunami resulting from newly connected consumers, products, and processes. Within this context, data quality has taken on a critical new importance. Leading companies, recognizing this, are seeking modern data management technologies to address this reality, and are increasingly looking to open source communities for innovation and answers. This white paper, the first in a series, identifies critical data management requirements and defines a high-level solution blueprint for leveraging Apache open source components to transform the data tsunami into high-quality data that can fuel digital transformation within enterprises. Future white papers will provide additional technical details on this topic.

About Hortonworks

Hortonworks is a leading provider of enterprise-grade, global data management platforms, services and solutions that deliver actionable intelligence from any type of data for over half of the Fortune 100. Hortonworks is committed to driving innovation in open source communities, providing unique value to enterprise customers. Along with its partners, Hortonworks provides technology, expertise and support so that enterprise customers can adopt a modern data architecture. For more information, visit hortonworks.com.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Contact

For further information,
visit hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

HORTONWORKS

© 2011-2018 Hortonworks Inc. All Rights Reserved.
Privacy Policy | Terms of Service

08/2018