HPCC SYSTEMS™

LexisNexis® RISK SOLUTIONS

WHITE PAPER

Data Lake Curation and Governance with Tombolo

INTRODUCTION

Technologies like social media, e-commerce, and the Internet of Things (IoT) are fueling massive growth in the amount of data that organizations need to store and analyze.

According to a 2022 report from Statista, global data creation is projected to grow to over 180 zettabytes by 2025. This kind of dramatic growth creates problems for organizations as they work to capture, classify, and analyze the data generated by customers and their own employees and operations. Further complicating the situation is the variety of data formats which organizations now collect: structured and semi-structured data, unstructured data (emails, documents, PDFs), and binary data (images, audio, video). Finally, some of this data will include information that must remain private and protected to meet various legal and service level agreement (SLA) requirements for the use and storage of sensitive data.

Trying to manage data in real-time as it pours into an organization's data center is challenging.

Taking new data and adjusting it to meet the format requirements of existing databases uses up valuable resources, both in terms of servers and personnel. Respondents to a McKinsey 2019 Global Data Transformation Survey reported that an average of 30 percent of their total enterprise time was spent on non-value-added tasks because of poor data quality and availability. Legacy data management systems also tend to create data silos: collections of separate databases that don't communicate with each other due to mismanagement. Data silos can result in flawed data analytics as algorithms attempt to create data-driven insights without access to complete and/or correct data assets.

To keep pace with massive amounts of inbound data and avoid problems like data silos, many organizations are using Data Lake technology.

Data Lakes support extremely large, complex, and diverse datasets, and they easily accommodate new data sources such as IoT. They allow IT groups to quickly create new applications that support changing business needs by unlocking the power of complex data for all users within the organization. They also scale much more easily and cost-effectively than relational databases, and offer the huge storage and compute resources needed for data analytics. As a result, Data Lakes enable greater responsiveness for users and external customers, reduced costs, and greater scalability.

Data Lakes increase their utility and enable their own evolution by:

- Facilitating the rapid development of new data sources and analytic algorithms
- Allowing multiple users to access data and create customized applications and reports
- Supporting the easy addition of new data sources to provide a continuous value enhancement to the data available to users

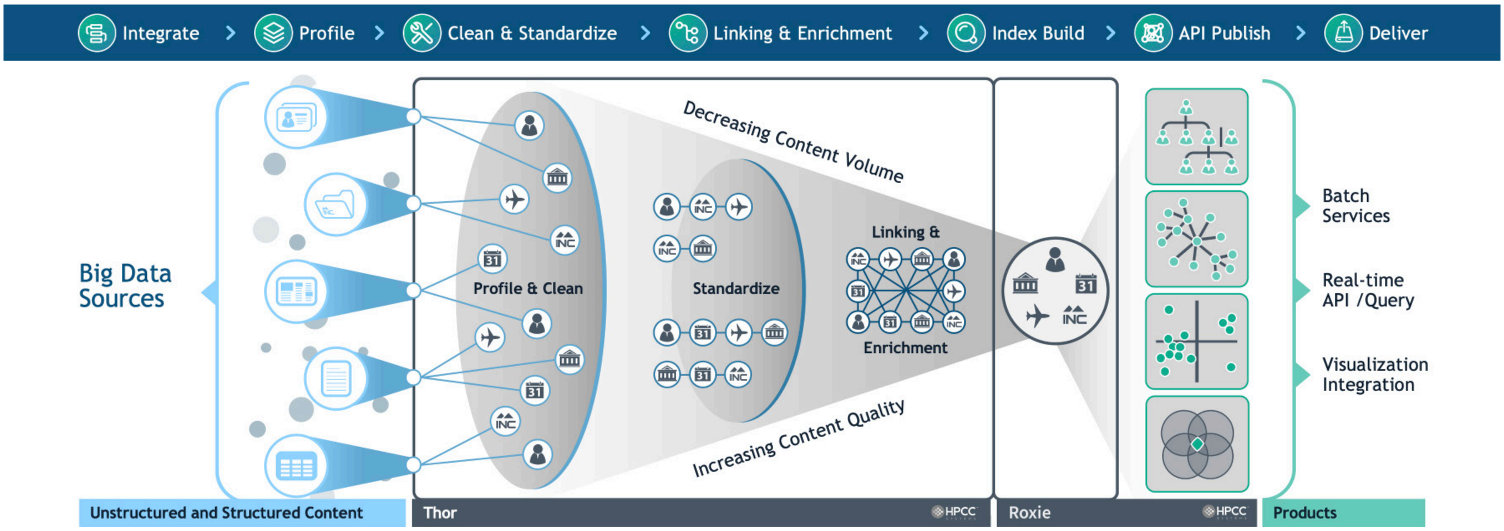


Illustration 1: The Data Lake pipeline above illustrates the process through which data is ingested by the Data Lake, transformed into formats required to conduct analysis, and finally how the results of that analysis are reported to users. For more information on Data Lake technology, please read HPCC Systems: The End-to-End Data Lake Management Solution. This whitepaper provides more detail about Data Lakes and HPCC Systems, a free, open-source, end-to-end platform for Data Lake management developed by HPCC Systems.

By storing raw data in one large, centralized repository, Data Lakes greatly reduce the possibility of data silos, but at a cost. As a Data Lake evolves, it grows in size and complexity. If not properly managed, a Data Lake can outgrow the abilities and resources of the team that manages it, negatively impacting the usefulness of an organization's data and slowing or halting the team's implementation of new analytics and applications.

Data Lakes must also securely store and monitor the use of private or sensitive data (i.e., credit card numbers, social security numbers, and medical records). As sensitive data is added to the Data Lake, the Data Lake must be able to recognize it and apply any required security and privacy protocols that keep that data safe. Furthermore, as this sensitive data is used by the Data Lake's users and applications, the Data Lake must ensure that it keeps an accurate record of where sensitive data originally came from, what steps were taken to secure that data, who has access to that data, and how they may have used that data. In other words, Data Lakes must create and maintain a record of the "lineage" for each data file. Such a file would track when the data first entered the Data Lake; every time that data file was accessed; and every time that data was read, copied, or written over.

In summary, Data Lakes must be able to support three basic functions in order for them to deliver on their potential for streamlining data management.

Data Lakes must:

1. Curate data: the ability to automatically identify and classify a data file
2. Govern sensitive data: automatically identify sensitive data files, apply any necessary usage restrictions to that data, and keep a record of who, how, and when a user or application interacts with a sensitive data file

An analogy can help illustrate the importance of curation and governance in Data Lakes. Imagine a large, world-class science, art, and history museum filled with millions of exhibits (fossils, artifacts from ancient civilizations, famous works of art, etc.), as well as the tools and equipment researchers and scientists use to prepare or study these exhibits (microscopes, brushes, chisels, chemistry equipment, reference books, etc.). Museums don't simply store all of these exhibits and tools in one large room and then tell any and all visitors to help themselves. Rather, they use curators to catalog and organize the museum's resources to help visitors locate the exhibits they're interested in seeing, or make them aware of new exhibits they might enjoy. Additionally, curators keep more valuable or sensitive exhibits and tools properly stored and secured to better control which visitors have access to them. So, while a fossilized T. Rex skeleton is publicly displayed in the museum lobby so all visitors can see it, an ancient scroll that has become brittle with age would only be available for viewing in a secure location (such as a lab or a private reading room) by a museum guest with the proper credentials (like a professor conducting research.)

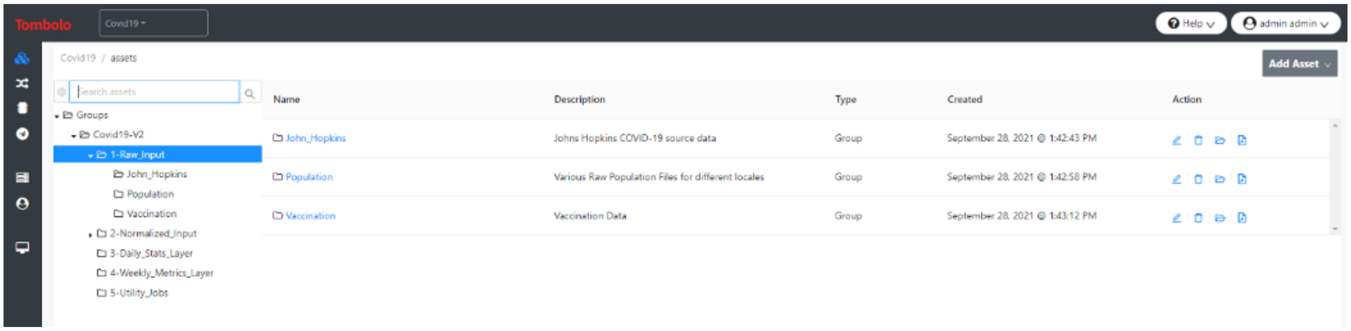
As data assets are collected and added to the Data Lake, a curation solution extracts a file's metadata so that particular file can be cataloged.

Once cataloging is completed, the curation system can identify which data assets are relevant to different analytics algorithms and/or applications. Part of the cataloging process would also include the application of governance rules that would identify sensitive data and apply rules to dictate who or what has access to that data and what they are authorized to do with that data (read, write, copy, etc.) All of the Data Lakes curation and governance capabilities would need to be automated in order for the Data Lake and its users to keep pace with the constant flood of new data being added to the Data Lake.

HPCC Systems has developed an open-source data curation and governance system to complement the powerful storage and compute capabilities of the HPCC Systems Data Lake operating system. Named Tombolo, this curation system is tightly integrated with HPCC Systems to maximize data visibility by automating the data curation and governance process. Tombolo provides the tools required to implement, document, and maintain an organizational infrastructure for new and existing data assets stored in one or more HPCC Systems Data Lakes (both cloud-based and on-premises Data Lakes.) Additionally, Tombolo can implement safeguards to govern what users and applications have access to those data assets. Tombolo conducts these curation and governance operations in an automated fashion to consistently and reliably curate huge amounts of inbound new data and ensure the continuous availability of the Data Lake. By automating the curation and governance of new data assets, Tombolo allows Data Lake users to focus on developing new applications and analytics to better leverage new data assets, and less time on formatting data, documenting changes, and other system maintenance activities that are vital to the ongoing availability of the Data Lake, but provide little value in terms of new functionality.

Tombolo gives HPCC Systems users three specific systems to assist in the management of Data Lakes:

- A Curation system that greatly enhances the transparency of the Data Lake by providing a documented roadmap of the Data Lake's assets (Datasets, Jobs, and Queries). This roadmap is updated automatically as new data are added to the Data Lake.



Covid19 / assets

- Groups
  - Covid19-V2
    - 1-Raw_Input
      - John_Hopkins
      - Population
      - Vaccination
    - 2-Normalized_Input
    - 3-Daily_Stats_Layer
    - 4-Weekly_Metrics_Layer
    - 5-Utility_Jobs

Name	Description	Type	Created	Action
John_Hopkins	Johns Hopkins COVID-19 source data	Group	September 28, 2021 @ 1:42:43 PM	
Population	Various Raw Population Files for different locales	Group	September 28, 2021 @ 1:42:58 PM	
Vaccination	Vaccination Data	Group	September 28, 2021 @ 1:43:12 PM	

Illustration 2: By curating data in the Data Lake, Tombolo tracks datasets as they evolve in the Data Lake, including keeping track of any formatting or data enrichment each data file has received.

- An Orchestration system that executes the data pipeline steps and propagates data lineage and enforces governance.

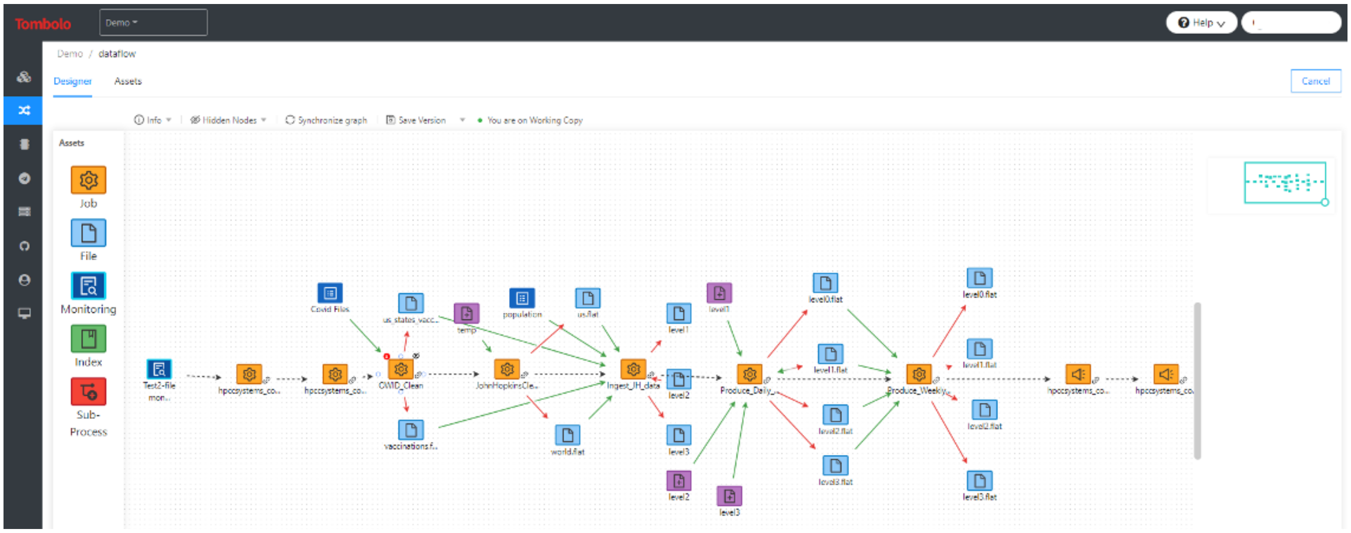
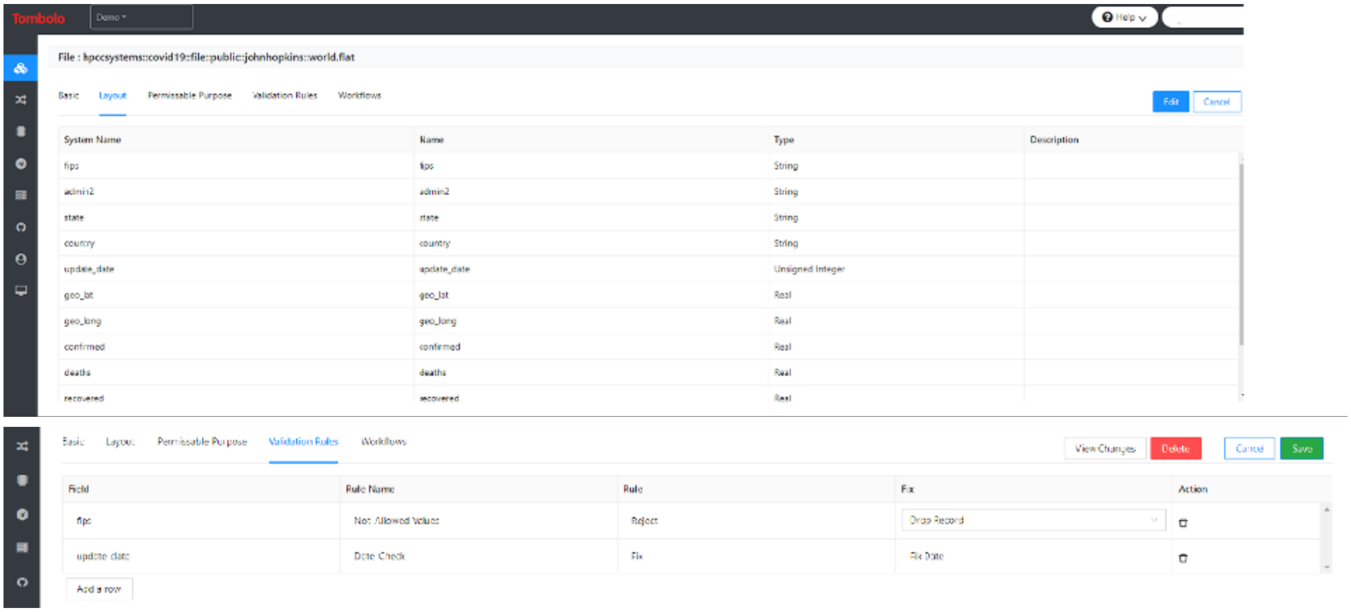


Illustration 3: Tombolo can automate the orchestration of data as it moves through the data pipeline.

- A Governance system to provide appropriate controls on the evolution and operation of the Data Lake, including the identification of potentially sensitive data that will require rules and regulations about what users and applications have access to that data.



File : hpccsystems::covid19::file::public::johnhopkins::world.flat

System Name	Name	Type	Description
fips	fips	String	
admin2	admin2	String	
state	state	String	
country	country	String	
update_date	update_date	Unsigned Integer	
geo_lat	geo_lat	Real	
geo_long	geo_long	Real	
confirmed	confirmed	Real	
deaths	deaths	Real	
recovered	recovered	Real	

Field	Rule Name	Rule	Fix	Action
fips	Not Allowed Values	Reject	Drop Record	🗑
update date	Date Check	Fix	Fix Date	🗑

Illustration 3: Tombolo's governance capabilities allow users to flag a sensitive data file for special treatment and create an audit trail covering the entire lifecycle of that data file in the Data Lake.

Tombolo is a web-based application for easy user access and features a GUI interface to facilitate the browsing of the Data Lake's assets and workflows. As the Data Lake grows in size and sophistication, Tombolo can automatically generate PDFs that document the latest changes to the data, including a list of new data assets, what users have access to specific assets, and what users are doing with those assets. This automation can also identify potential problems with the Data Lake that could impact its future availability and/or security. Tombolo can then either implement a fix on its own or flag the problem for human intervention.

TOMBOLO OVERVIEW

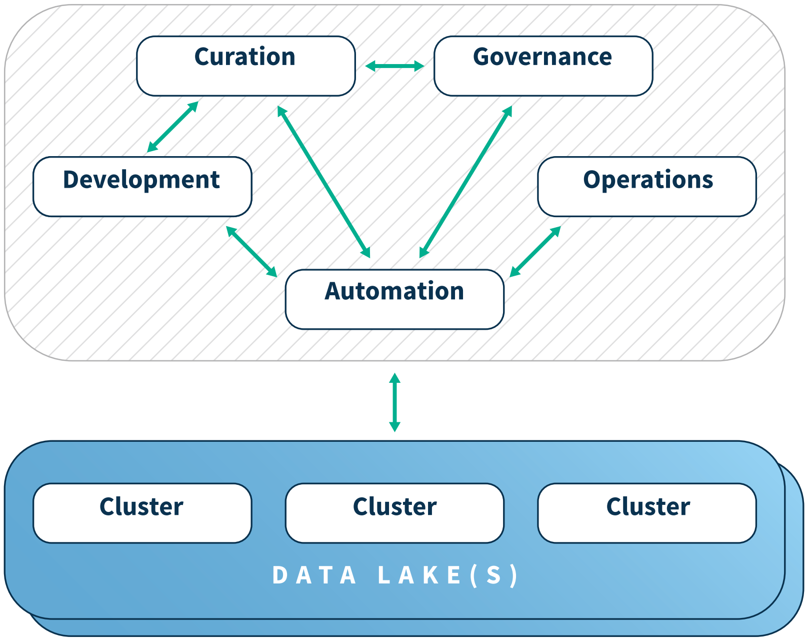


Illustration 3: Tombolo complements the processing and storage power of one or more HPCC Systems Data Lakes with an automated system for curating and governing data assets.

As of the publication of this whitepaper, the Tombolo system provides a framework for the curation and governance of data assets and is integrated with existing tools in the HPCC Systems Data Lake platform. Tombolo can organize and document all resources stored in the Data Lake, including data assets, applications, and analytics algorithms. Tombolo can also identify sensitive data assets and manage their visibility to users. The system performs all of these functions in an automated fashion to help ensure the continuous operation of the Data Lake.

ASSET CURATION WITH TOMBOLO

Once the Tombolo application is linked to one or more HPCC Systems Data Lakes, it begins to collect the metadata of every asset in the Data Lake for use in cataloging and classification.

Users can then search for assets based on specific metadata using Tombolo's query function. In addition to metadata, other characteristics of the file can be cataloged. For example, if a data asset requires verification or formatting before it can be used by an application or algorithm on the Data Lake, Tombolo can tell a user if the required verification or formatting has been completed or not. Or perhaps a particular data asset contains private or sensitive information; Tombolo can indicate if all or part of a particular asset is available for use by a specific user or application. In future releases of Tombolo, HPCC Systems plans to automate many of the processes required to prepare data for storage in the Data Lake.

Tombolo can also help automate the execution of applications stored on a Data Lake.

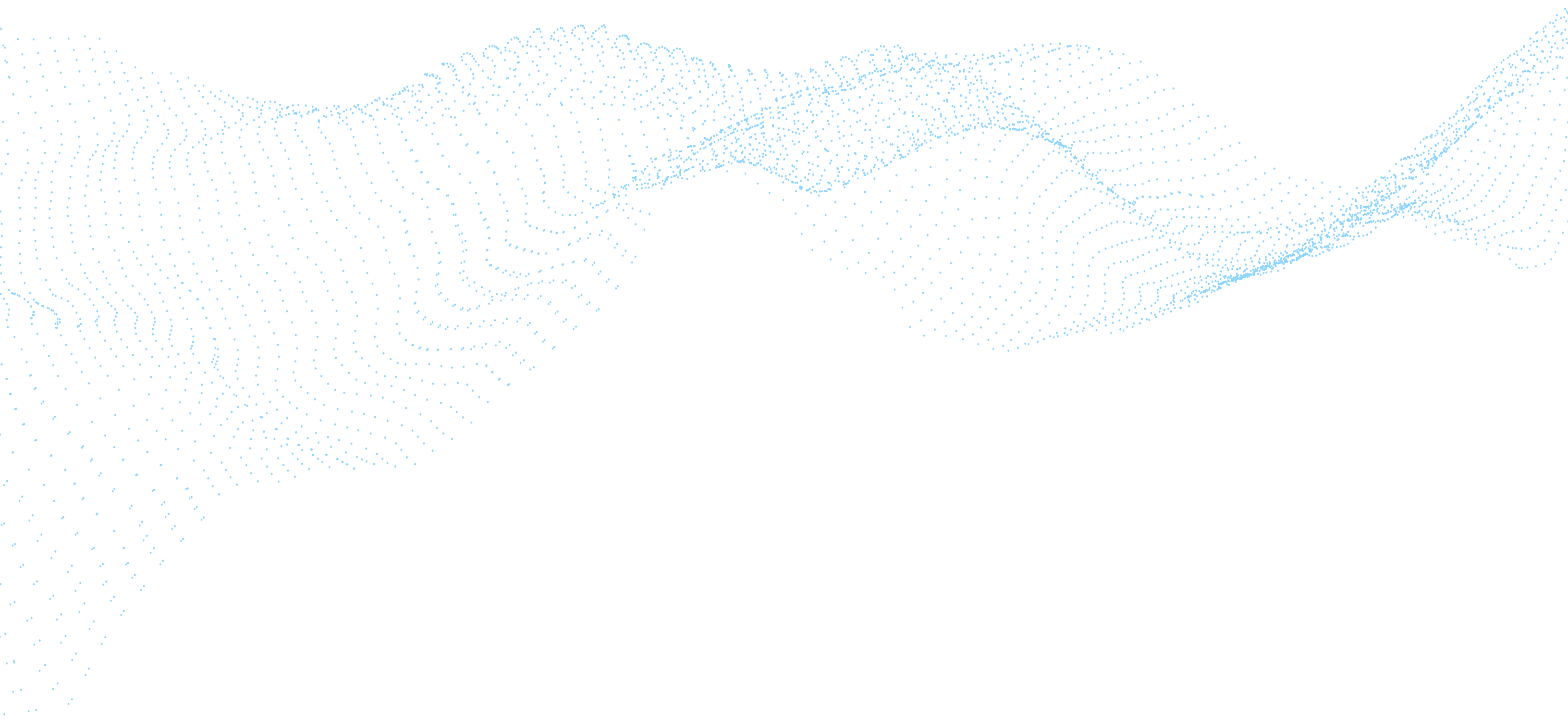
This is helpful for applications that require a change to their source code prior to every use of the application. In fact, if an application is stored and managed on a GitHub server, Tombolo can pull the latest version of the application's source code from GitHub, compile it, and then download it to the HPCC Systems Data Lake to execute. Furthermore, if the application is unable to execute or otherwise experiences problems, Tombolo can bring it to the attention of the users responsible for the application for further action.

Tombolo also identifies any business intelligence report or dashboards present in the Data Lake. Thanks to Tombolo's tight integration with HPCC, users can modify existing reports or create new ones within Tombolo as it natively supports HPCC Systems Real BI Dashboard tool, so users don't need to leave the Tombolo environment and switch to HPCC Systems to create or modify a report or dashboard.

GOVERNANCE WITH TOMBOLO

At launch, Tombolo's governance features are limited to identifying if specific data assets in an HPCC Systems Data Lake contain sensitive or restricted data. Future releases of Tombolo will provide support for expanded governance functionality, including:

- The ability to propagate access restrictions throughout the entire data development chain to help ensure sensitive data remains secure throughout its lifecycle in the Data Lake.
- Identifying potential compliance issues users may experience with sensitive data before they happen, and then flag those issues to the proper user for resolution.
- The creation of audit trails so users can automatically generate documentation to prove they are handling sensitive data in ways that comply with any regulatory or SLA requirements.



CONCLUSION

The explosive growth in data facing many organizations today is best addressed by using a Data Lake; a massive, centralized data storage and compute environment that allows all types of data assets, analytics algorithms, and applications to reside in one location.

The new Tombolo for HPCC Systems Data Lakes provides a powerful tool for users to automate the curation and governance of the data, algorithms, and applications residing on a Data Lake. This allows HPCC Systems users to spend less maintaining and formatting data, and more time developing analytics and applications to extract real business value from their data. Tombolo also provides the ability to govern who has access to sensitive data in a Data Lake, and more granular control of Tombolo's governance capabilities (including the ability to generate reports that track the usage history of a particular data file) will be available in future releases.

For more information, call 877.316.9669 or visit www.hpccsystems.com

About HPCC Systems®

HPCC Systems® from LexisNexis® Risk Solutions is a proven, comprehensive, dedicated data lake platform that makes combining different types of data easier and faster than competing platforms — even data stored in massive, mixed schema data lakes . It's also open source, free to use, and easy to learn. You can acquire, enrich, deliver and curate information faster using HPCC Systems — and the automation of Kubernetes in our cloud-native architecture makes it easy to set-up, manage and scale your data to save time and money, now and in the future. HPCC Systems offers a consistent data-centric programming language, two processing platforms and a single, complete end-to-end architecture for efficient processing. To learn more, visit us at hpccsystems.com.

About LexisNexis® Risk Solutions

LexisNexis® Risk Solutions harnesses the power of data and advanced analytics to provide insights that help businesses and governmental entities reduce risk and improve decisions to benefit people around the globe. We provide data and technology solutions for a wide range of industries including insurance, financial services, healthcare and government. Headquartered in metro Atlanta, Georgia, we have offices throughout the world and are part of RELX (LSE: REL/NYSE: RELX), a global provider of information-based analytics and decision tools for professional and business customers. For more information, please visit www.risk.lexisnexis.com and www.relx.com.