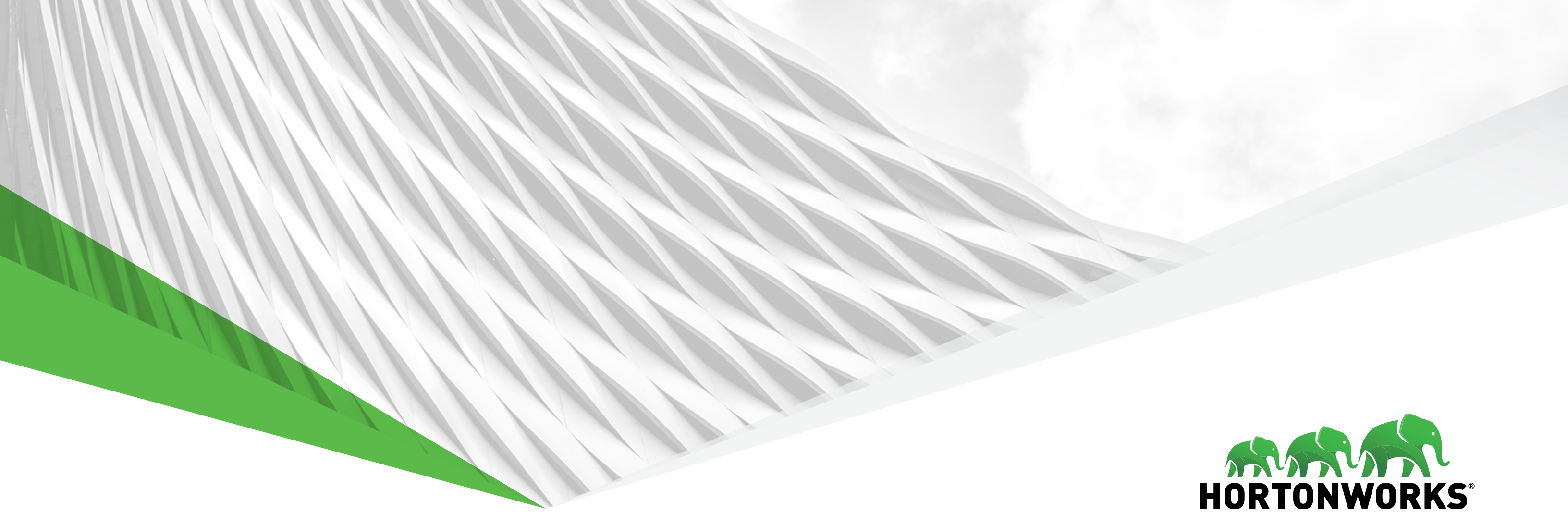


# Path to GDPR Compliance Begins with Data Governance

A HORTONWORKS WHITE PAPER

MARCH 2018

# Contents

## Overview

On December 15th, 2015, the European Parliament and the Council of the European Union reached an agreement to update the privacy standards based on today's technology and innovation. These standards are what is known as the General Data Protection Regulation or GDPR.

GDPR represents a sea change for the organizations that store or process any kind of data on EU residents. Most importantly, unlike its predecessor the Data Protection Directive 95/46/EC, GDPR is a regulation that will become law in all member states by May of 2018. GDPR also greatly extends the jurisdiction of this regulation to all the companies that process the personal data of European Union residents or subjects regardless of the company's location. Specifically GDPR will apply to all the organizations that are formed in European Union and controls and processes data, in addition to non-EU businesses that offer goods or services to EU citizens.

Companies are scrambling to comply with GDPR as the penalty for insufficient customer consent to process data or to violate the Privacy by Design concepts is stiff — fine up to 4% of annual global turnover or €20 million — whichever is greater.

## Challenges of Compliance in Big Data

Despite advances in big data storage and processing systems, regulatory compliance such as GDPR continues to pose interesting challenges to enterprises that need to manage their expanding stockpile of data from a regulatory, competitive, or privacy reasons. These challenges emanate from the factors that include:

- Data lakes are quickly becoming the centralized repositories that contain both raw as well as curated corporate and external data that enterprises need to run their businesses.
- Enterprises large and small are keeping their data longer in the data lake, as storage continues to become cheaper and processing engines become more cost effective. So not only can enterprises store all their data in one place but they are also keeping it there longer.
- The source systems that feed data into the data lake are changing continuously and new systems are constantly being added.
- There is a broader set of users in companies that are analyzing the data in the data lake. For example, in the case of a data lake, in addition to business analysts that are analyzing sales or finance data, there might be citizen data scientists and marketers that are specifically analyzing Facebook and twitter data for sentiment analysis.

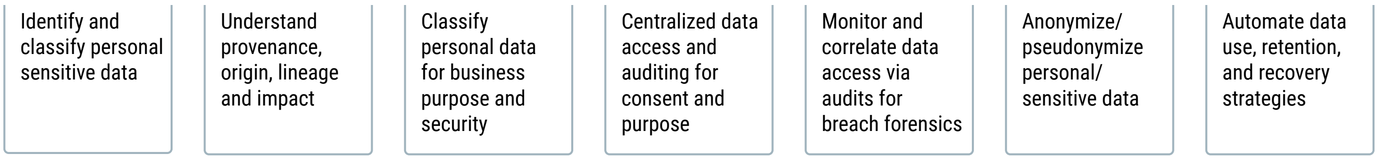

Identify and classify personal sensitive data
Understand provenance, origin, lineage and impact
Classify personal data for business purpose and security
Centralized data access and auditing for consent and purpose
Monitor and correlate data access via audits for breach forensics
Anonymize/ pseudonymize personal/ sensitive data
Automate data use, retention, and recovery strategies

*Figure 1: GDPR Considerations for Big Data*

## CYBERSECURITY AND BREACH NOTIFICATION

| BEST PRACTICES | SPECIFIC ACTIONS |
|---|---|
| • Pseudonymize and encrypt all personal data<br>• Ensure the confidentiality, integrity, availability and resilience of personal data on an ongoing basis<br>• Ensure the availability and access to personal data in a timely manner | • Deploy data masking solutions for data at rest and data in motion<br>• Invest in data replication and disaster recovery solutions<br>• Implement automated data discovery and profiling solution<br>• Engage with data protection officer and cybersecurity team |

## CONSENT

| BEST PRACTICES | SPECIFIC ACTIONS |
|---|---|
| Develop a process to acquire consent from customers that covers:<br>• Data processing & operations<br>• Parental consent for processing children's personal data<br>• Right to withdraw consent | • Establish a process to provide notice and receive consent<br>• Document and communicate the purpose for which data is being collected and processed<br>• Leverage access control to block data in cases where consent is not provided |

## PROFILING

| BEST PRACTICES | SPECIFIC ACTIONS |
|---|---|
| Place restrictions on:<br>• Processes that may be classified as profiling<br>• Automated processing of personal data<br>• Using personal data to analyze or evaluate a person including taking decisions or predicting an individual's intent or actions<br>Provide a process to inform EU subjects of their rights and to object or halt data collection/processing<br>Provide notice and access to data being collected | • Invest in interfaces that can be used by EU customers/residents to receive requests<br>• Build customer identity solution<br>• Invest in data discovery and classification solution |

## RIGHT TO BE FORGOTTEN AND DATA PORTABILITY

| BEST PRACTICES | SPECIFIC ACTIONS |
|---|---|
| Right to Erasure/Right to be Forgotten<br>• Remove data upon request<br>• Provide reasonable access to customer's own data<br>• Provide information about the objectives for which data is being collected<br>Data Portability<br>Build workflows for EU residents to:<br>• Receive their own data<br>• Request transfer of data to 3rd party/ processor<br>• Right to rectify data | • Establish a strict process to oversee automated data processing<br>• Use manual processing for analytics that involves personal data<br>• Establish consent for any manual profiling related processing |

# Role of Data Governance in Compliance

Data governance helps enterprises manage risk effectively, comply with regulations and gain competitive advantage through agile decision making via responsible use of data assets. In the context of Big Data, data governance needs to encompass all the data sources that feed data into data lakes and the entire pipeline through which data is transformed and used. Therefore, data governance needs to cover the entire big data landscape in terms of storage infrastructure, processing pipelines, and data transformation engines across the entire data lifecycle from data ingestion through retirement.

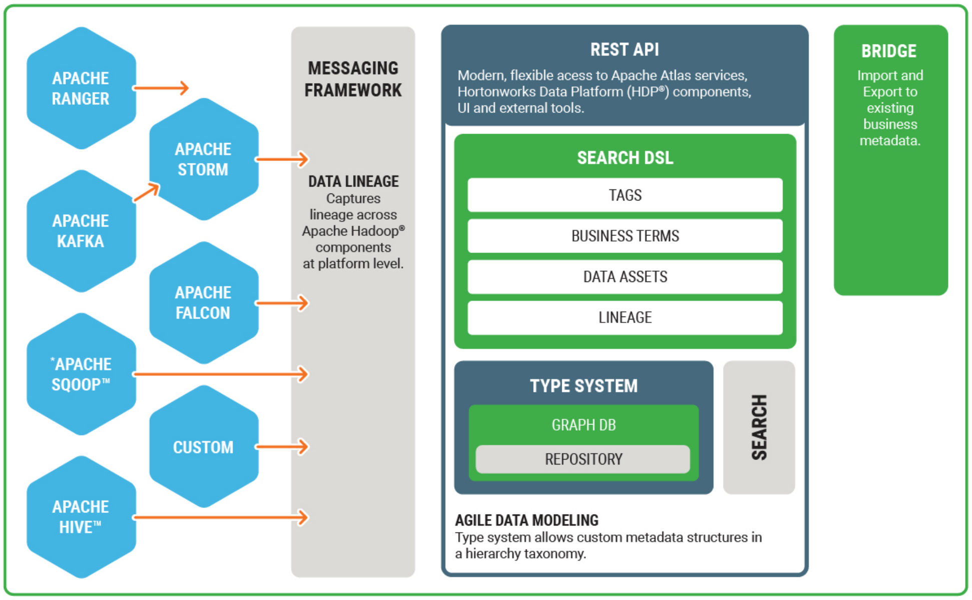
Hortonworks vision is to provide an open set of APIs, types and interchange protocols to allow all metadata repositories to share and exchange metadata. This forms the common base that can provide governance, discovery and access frameworks to automate the collection, management, and use of metadata across an enterprise. The end result is an enterprise catalog of data resources that are transparently assessed, governed and used to deliver maximum value to the enterprise.

Delivering this capability in open source is a critical part of our strategy as multiple vendors offering diverse capabilities are required to support this ecosystem. Thus, open metadata and governance technology must be freely available with an open source governance model that allows a community of organizations and practitioners to develop and evolve the base and implement it in their offerings and deployments. Apache Atlas supports an open metadata and governance compliant repository plus it provides the adapters and interchange formats to allow other metadata repositories to connect into the ecosystem.

## APACHE ATLAS

In Hortonworks Data Platform, Apache Atlas provides built-in data management and governance capabilities. Apache Atlas was incubated by Hortonworks along with a community of industry leaders in diverse industries to provide the best in class open source metadata and governance services for enterprise big data ecosystems. The vision for Apache Atlas project is to provide core metadata-driven governance services for Hadoop and enterprise data ecosystems. Key enterprise metadata and governance features in Atlas include:

- Robust Metadata Repository providing a flexible metamodel to capture technical, business, operational metadata
- Out-of-box metadata models for Apache Hive, Apache Storm, Apache Sqoop, Apache Kafka, HDFS, and HBase
- Enterprise ready real-time metadata and lineage ingestion with Hive, Sqoop, Storm/ Kafka
- Data Classification
- Metadata Catalog and Search
- Extensible APIs for custom metadata ingestion and APIs to register custom models
- Apache Ranger integration for classification based security
- Type system to easily model any type of physical, operational or conceptual metadata


APACHE RANGER
APACHE STORM
APACHE KAFKA
APACHE FALCON
*APACHE SQOOP™
CUSTOM
APACHE HIVE™
MESSAGING FRAMEWORK
DATA LINEAGE
Captures lineage across Apache Hadoop® components at platform level.
REST API
Modern, flexible acess to Apache Atlas services, Hortonworks Data Platform (HDP®) components, UI and external tools.
SEARCH DSL
TAGS
BUSINESS TERMS
DATA ASSETS
LINEAGE
TYPE SYSTEM
GRAPH DB
REPOSITORY
SEARCH
AGILE DATA MODELING
Type system allows custom metadata structures in a hierarchy taxonomy.
BRIDGE
Import and Export to existing business metadata.

*Applies to any connector that leverages Apache Sqoop including Teradata Connector

*Figure 2: Apache Atlas Architecture*

Atlas is also part of the recently launched Hortonworks DataPlane Service and the extensible services that are part of this solution. Specifically, Data Steward Studio (DSS) available as technical preview empowers businesses to understand, secure, and govern diverse data in enterprise data lakes across locations and environments so that they can confidently derive insights from it. DSS provides the following functionality.

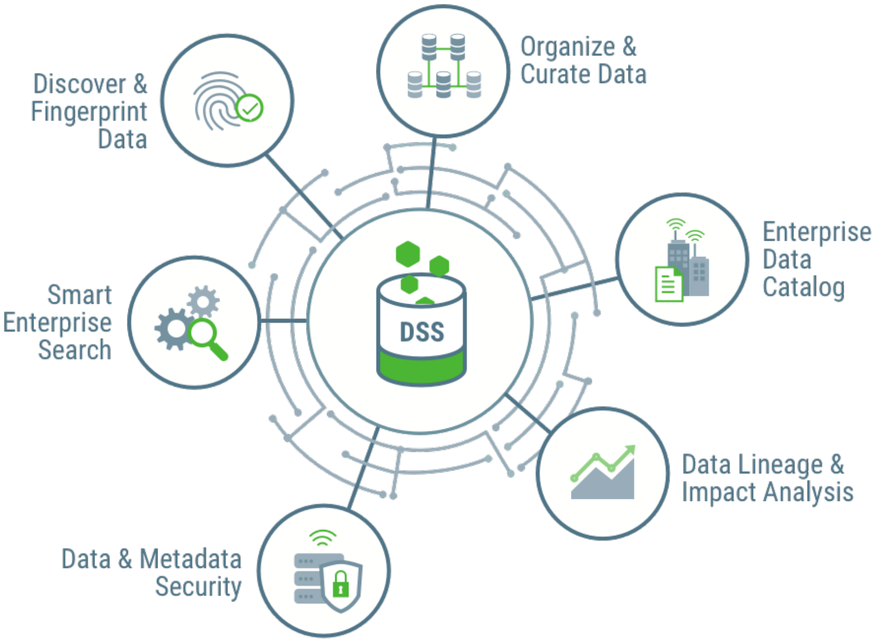


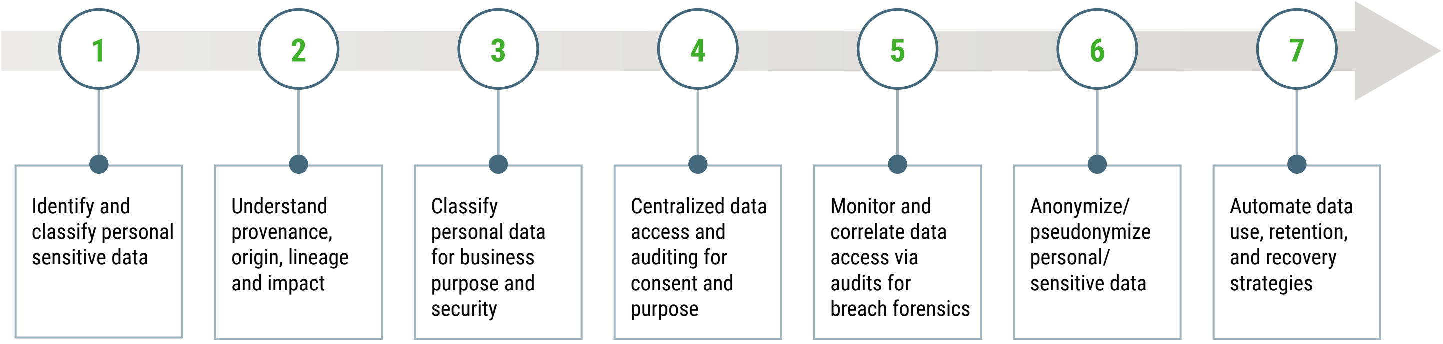
*Figure 3: Data Steward Studio*

## APACHE RANGER

Apache Ranger is Atlas's counterpart for access policy authorization in HDP. HDP utilizes Apache Ranger to provide a centralized framework to define, administer and manage security policies consistently across the Hadoop ecosystem components including HDFS, Apache Hive, Apache HBase, Apache YARN, Apache Kafka, Apache Solr, Apache Storm, Apache Knox, Apache NiFi, and Apache Atlas.

Apache Ranger also provides a comprehensive audit framework for all the services it authorizes. This framework provides rich event data along with contextual metadata. Ranger's approach to authorization is based on attribute-based access control (ABAC), which is a combination of the subject, action, resource, and environment. Using descriptive attributes such as Active Directory (AD) group, Apache Atlas-based tags or classifications, geo-location, etc., of the subjects, resources, and environment, Apache Ranger provides a modern and superior policy approach beyond simple role-based access control (RBAC). ABAC approach is consistent with recommendations outlined by NIST for ABAC in NIST 800-162. This approach enables compliance personnel and security administrators to define precise and intuitive security policies as required by GDPR at a very fine-grained level. This approach also overcomes the pitfalls of older RBAC based technologies that place a heavy burden on security administrators and lead to role proliferation and manageability issues.

## GDPR Compliance Roadmap



1 Identify and classify personal sensitive data

2 Understand provenance, origin, lineage and impact

3 Classify personal data for business purpose and security

4 Centralized data access and auditing for consent and purpose

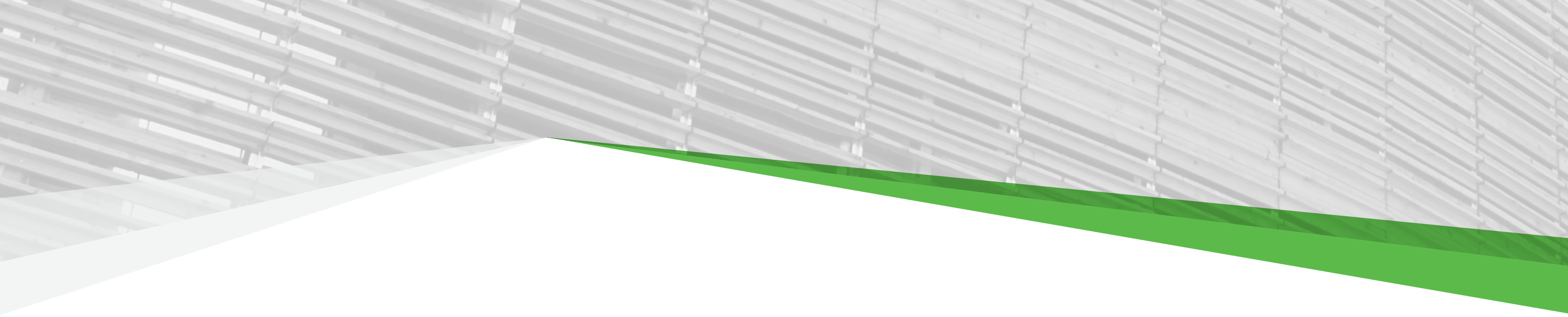
5 Monitor and correlate data access via audits for breach forensics

6 Anonymize/ pseudonymize personal/ sensitive data

7 Automate data use, retention, and recovery strategies

## How Hortonworks Solution Can Help Enterprise Comply with GDPR

| OBJECTIVE | SPECIFIC STEPS | APPLICABLE HDP COMPONENT |
|---|---|---|
| Establish an Enterprise Data Catalog & Business Glossary to define enterprise data footprint | • Identify and classify sensitive data<br>• Harvest and maintain metadata about data stores, owners, business classifications etc.<br>• Map sources that contain GDPR relevant personal data<br>• Identify locations within data sources and systems that contain high risk/ personal data | Apache Atlas |
| Track and map the movement of personal/ sensitive data through the enterprise | • Maintain a near real-time view to track data movement<br>• Understand the proliferation of sensitive data with data lineage and impact analysis | Apache Atlas |
| Implement appropriate security controls to monitor access to data across the enterprise | • Implement policy based controls to grant and monitor data access<br>• Track user activity for personal/ sensitive data to support both forensic auditing, as well as alerting for proactive breach notification<br>• Protect sensitive data through anonymization and pseudonymization using dynamic masking | Apache Ranger |



## Summary

GDPR is an issue that has visibility at the CEO and Board of Directors level due to its potential impact on company's profitability, valuation, reputation and even its ability to continue its operations. A majority of companies are unsure of the approach to take in order to comply with the upcoming regulations. It is also important to point out that for a large number of companies, development and implementation of tools, technologies, and solutions  related to GDPR compliance will continue well past May 2018.

We believe that data governance solutions combined with robust access control represents an effective starting point for organizations that are about to embark on this journey. The ability to identify and classify personal data, understand its lineage and impact, implement access and audit controls are some of the actions that can provide the foundation for organizations to comply with GDPR. In this regard open source projects such as Apache Atlas and Apache Ranger that are an integral part of Hortonworks Data Platform can play a vital role.

## 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.

This content is not intended to constitute legal advice. Readers should consult with their own legal counsel regarding compliance with GDPR and other laws and regulations applicable to their particular situation and intended use of any Hortonworks products and services.

**Contact**

For further information,
visit hortonworks.com

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

HORTONWORKS®

03/2018