Executive Briefing

GDPR: Getting your data ready for heavy, new EU privacy regulations

Steve Ross, Director, Product Management, Cloudera
Mark Donsky, Senior Director, Product Management, Okera
Disclaimer

GDPR is a complex and detailed regulation.

There's no single method or solution that will make all organizations compliant.

This presentation is intended to help organizations understand how Big Data platforms such as Cloudera software and services can be used to help comply with certain aspects of EU General Data Protection Regulation (GDPR) requirements. Applicability of any of these capabilities depends on each organization’s own requirements specific to their business. Every organization should determine its own needs with regard to GDPR and then evaluate solutions for suitability to those needs. The information contained in this presentation is not intended to be and should not be construed to be legal advice. Organizations must not rely on the information herein and they should obtain legal advice from their own legal counsel or other professional legal services provider.
General Data Protection Regulation (GDPR)

- Enforced from May 25, 2018
- Substantial penalties
- Obligations of the organization
- Personal Data
- Rights of the consumer
- Applicable worldwide

- Right to be forgotten/erasure
- Right to access information
- Right to data portability
- Right for processing to be restricted

- Heavy fines for violations
- Up to 20M Euros or 4% of the annual global turnover for the preceding financial year

- Any organization with any users in the EU needs to be compliant.
- Includes companies based outside the EU, processing personal data from EU residents in connection with the offering any goods or services or monitoring user behavior.
- Includes data processor and data controller

- Across people, process and technology
- Impacts how personal data is collected and used

© Cloudera, Inc. All rights reserved.
Examples of personal data
Seven Key Principles of GDPR

- **Lawfulness, fairness, transparency**
  How do I track personal data?

- **Accountability**
  How can I demonstrate compliance? How do I report breaches in 72 hrs?

- **Storage limitation**
  How do I erase individual data records upon request when the file systems are immutable?

- **Integrity and confidentiality**
  How do I apply IT controls to prevent unauthorized access?

- **Purpose limitation**
  How do I track consent while using data science tool choice?

- **Accuracy**
  What is a low-overhead way to fix data?

- **Data minimization**
  How can I anonymize personal data? How do I prevent unlawful data transfers?
The path to GDPR compliance includes...

- **Foundational Activities**
  - **Stakeholder awareness**
    - Make sure stakeholders are fully aware of the GDPR and the impact it will have on the organisation. Ensure that key messages are understood by the Board and senior management - and gain buy-in for remediation and change programmes.
  - **Readiness assessment**
    - Conduct a Readiness assessment to understand how near or far away your organisation is from relevant new requirements of the GDPR and the potential effort required.
  - **Data inventories and mapping**
    - Compile an inventory of the personal data that is collected, who it is shared with and what controls govern its use. Authorities will expect this information to be made readily available.
  - **Governance**
    - Use the GDPR to assess your holistic approach to privacy - do you have a Data Protection Officer? Who is ultimately accountable? How are you going to bring together different areas of the business to manage privacy risks on an ongoing basis?

- **Typical Remediation Considerations**
  - **Legal and compliance**
    - Review approaches to capturing consent. Re-draft privacy notices and determine how compliance will be demonstrated.
  - **Technology**
    - Deploy technology and processes to bring about a Privacy By Design culture, and create robust breach management procedures, consider masking and encryption.
  - **Data**
    - Ensure the organisation has the right data governance practices to respond efficiently to the new rights afforded to individuals, such as the rights to erasure and portability.

*Source: Cognizant*

*Where are you along this path?*
How big data solutions can help accelerate GDPR compliance
A single place to secure and govern for GDPR compliance

For data that is stored on a single platform, there is a single place to secure and govern for GDPR compliance, across all analytic workloads.
The Seven Principles of GDPR

1. Integrity and Confidentiality
2. Accountability
3. Lawfulness, fairness and transparency
4. Purpose limitation
5. Data Minimisation
6. Accuracy
7. Storage Limitation
How big data solutions can help

<table>
<thead>
<tr>
<th>GDPR Principle</th>
<th>Typical Challenges</th>
<th>Big data solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrity and Confidentiality</td>
<td>Applying industry standard IT security controls to prevent unauthorised access.</td>
<td>Comprehensive encryption and key management. Strong authentication, fine-grained authorization.</td>
</tr>
<tr>
<td>Accountability</td>
<td>Demonstrating compliance. Detecting and analyzing breaches within 72 hours.</td>
<td>Comprehensive audit trail. Breach detection (cybersecurity solutions).</td>
</tr>
<tr>
<td>Lawfulness, fairness and transparency</td>
<td>Implementing a way to keep track of personal data.</td>
<td>Track classification and lineage of personal data elements.</td>
</tr>
</tbody>
</table>
Fine-grained access control and data masking

Apache Sentry and Okera ODAP for column-level permissions and views with masking, row filtering, dynamic masking, and tokenization
### Comprehensive Compliance Requirements

**Compliance Must Haves — common to most regulations**

<table>
<thead>
<tr>
<th>Capability</th>
<th>Watch out for solutions with gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ <strong>Encryption:</strong> All sensitive data encrypted at rest and in motion</td>
<td>✗ Data un-encrypted on local disks. Key management missing safeguards. 3rd party tools needed to plug this critical gap.</td>
</tr>
<tr>
<td>✓ <strong>Authorization:</strong> access limited to the data required by job role</td>
<td>✗ Set access controls separately/repeatedly in each framework and storage silo. Very difficult to ensure correct permissions</td>
</tr>
<tr>
<td>✓ <strong>Audit:</strong> All user actions must be logged (for forensics use)</td>
<td>✗ Audit can be disabled / Audit drops under load</td>
</tr>
<tr>
<td>✓ <strong>Visibility:</strong> All data must be classified according to sensitivity</td>
<td>✗ Limited business metadata, coarse-grained lineage</td>
</tr>
<tr>
<td>✓ <strong>Erasure:</strong> Individual records can be updated or deleted (GDPR)</td>
<td>✗ Restricted to batch deletions and other scaling challenges</td>
</tr>
</tbody>
</table>

Solutions can only go to production if they meet top compliance “must haves”
### How big data solutions can help

<table>
<thead>
<tr>
<th>GDPR Principle</th>
<th>Typical Challenges</th>
<th>Big data solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose limitation</td>
<td>Track consent and data usage while allowing data scientists to mine it using tools of choice.</td>
<td>Data protection officer (DPO) can audit exactly how data was used. Data scientists work with flexibility while keep data governed.</td>
</tr>
<tr>
<td>Data Minimization</td>
<td>Removing or anonymising data where possible. Preventing unlawful data transfers outside the EU while still enabling outsourcing.</td>
<td>Classification tags can indicate allowed purpose. Redacted views limit what certain people can access.</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Finding a low-overhead way to fix data.</td>
<td>Fast updates of individual records.</td>
</tr>
<tr>
<td>Storage Limitation</td>
<td>Deleting individual personal data records in Hadoop and Cloud storage, since those file systems are immutable.</td>
<td>Erasure of individual records. HDFS and Cloud storage options.</td>
</tr>
</tbody>
</table>
The right to erasure - challenges

1. **Existing data architectures** may spread personal data across many objects
2. **Self-service** generates derived datasets also subject to GDPR
3. **Volume and variety** means any solution needs to scale
4. **Storage capabilities** limit erasure options
   - HDFS and cloud object stores are “immutable”
Erasing individual records on HDFS and cloud storage

- Concentrate personal data in a small number of “lookup tables”
- Replace personal data in most locations with anonymized or pseudonymised data
- Instead of deleting records upon request, add them to a “to be deleted” list
- Execute a periodic batch job to remove “to be deleted” records by rewriting entire files/tables

Issue: The re-write step could render the cluster unusable for a period of time
A survey of solutions that can help
Governance: auditing, lineage, metadata

Capabilities:
- Inescapable, detailed audit – enabling forensics
- Full tracking of personal data
- Lineage tracking and visualization
Enterprise-grade encryption & key management

Encryption
- ALL data at rest: HDFS, HBase, metadata databases, temp files, ingest paths
- ALL data on the wire

Key Management
- Automated key replication & backup
- HSM backed key protection

Redaction
- Sensitive data in logs

Legend
- Metadata Store
- Encrypted Data
- Encryption Key

Manager
Navigator
Impala
Hive
Sentry
HDFS
Kudu & HBase
Navigator Key Trustee

Log Files
Ingest Paths, Temp/Spill files
HSM (optional)
Erasing individual records on Kudu

```sql
delete from customers
where name = 'Cole Moynihan' 
and id = 2398654;
```

Deleted successfully

Query History

- a minute ago: `select * from customers`
- 2 minutes ago: `select * from retail`
- an hour ago: `select * from anupam limit 100`
Laptop vs Centralized Data Science

“Laptop Data Science”
Typical Big Data Environment
Data scientists pull data to their laptops so they can run their own tools

• Copy personal data to laptops
• Fails GDPR compliance audit
• Potential data breach

Centralized Data Science

• Personal data remains governed
• Purpose limitation enforced

Data Science Workbench
Discovery, Access Control, and Governance

Simplified access, minimal complexity
- Active schema registry
- Multi-tool, multi-data, multi-cloud
- Collaborative workspaces

Scalable protection
- Fine-grained access control
- Tokenization & anonymization

Greater visibility
- Rich audit trail
Okera’s Product Overview
Okera case study: Top 5 Global Apparel Company

Requirements

- GDPR compliance: PII anonymization, right to be forgotten (join-based filtering)
- No way to do per user authentication – group based at best and wouldn’t scale
- Need to support existing jobs, and tools in functional and performant environment
- Future proof, more tools and patterns coming, simplify architecture

What we delivered

- In production in time for GDPR (May 2018)
- Half of pipelines with sensitive data read through Okera
- 20PB scanned / year
- Single catalog, single copy of data
- Per user auditing / usage understanding

“We need to be a technology-based company if we want to survive, especially since shoppers don’t go to the mall like they used to.”

ENVIRONMENT

- AWS
- S3-based data lake
- Presto, Databricks, for interactive BI
Cloudera case study: large EU telecom

- Modified data flow and schemas for GDPR
- Data classification and curation using Cloudera Navigator.
- Data anonymization on ingest - via Spark ETL pipeline & lookup table in Kudu
  - Anonymized data can be widely used
  - Personal data confined to lookup table - access tied to consent
- Consent management tracking - via Kudu
- Access control at column and row-level ensures consent for personal data use is enforced
<table>
<thead>
<tr>
<th>GDPR Principle</th>
<th>Typical Challenges</th>
<th>Big data solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Integrity and Confidentiality</strong></td>
<td>Applying industry standard IT security controls to prevent unauthorised access.</td>
<td>Comprehensive encryption and key management. Strong authentication, fine-grained authorization.</td>
</tr>
<tr>
<td><strong>Accountability</strong></td>
<td>Demonstrating compliance. Detecting and analyzing breaches within 72 hours.</td>
<td>Comprehensive audit trail. Breach detection (cybersecurity solutions).</td>
</tr>
<tr>
<td><strong>Lawfulness, fairness and transparency</strong></td>
<td>Implementing a way to keep track of personal data.</td>
<td>Track classification and lineage of personal data elements.</td>
</tr>
<tr>
<td><strong>Purpose limitation</strong></td>
<td>Track consent and data usage while allowing data scientists to mine it using tools of choice.</td>
<td>Data protection officer (DPO) can audit exactly how data was used. Data scientists work with flexibility while keep data governed.</td>
</tr>
<tr>
<td><strong>Data Minimization</strong></td>
<td>Removing or anonymising data where possible. Preventing unlawful data transfers outside the EU while still enabling outsourcing.</td>
<td>Classification tags can indicate allowed purpose. Redacted views limit what certain people can access.</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Finding a low-overhead way to fix data.</td>
<td>Fast updates of individual records.</td>
</tr>
<tr>
<td><strong>Storage Limitation</strong></td>
<td>Deleting individual personal data records in Hadoop and Cloud storage, since those file systems are immutable.</td>
<td>Erasure of individual records. HDFS and Cloud storage options.</td>
</tr>
</tbody>
</table>
Questions?
Thank you

Steve Ross, sross@cloudera.com
Mark Donsky, mark@okera.com
Pillars of a comprehensive compliance solution

GDPR principles: Integrity, Confidentiality, Accountability, Lawfulness, Fairness, Transparency

**Perimeter**
Guarding access to the cluster itself

- Technical concepts: Authentication, Network isolation

**Access**
Defining what users and applications can do with data

- Technical concepts: Permissions, Authorization

**Visibility**
Reporting on where data came from and how it’s being used

- Technical concepts: Auditing, Lineage

**Data**
Protecting data in the cluster from unauthorized visibility

- Technical concepts: Encryption, tokenization, data masking
# How big data solutions can help

<table>
<thead>
<tr>
<th>GDPR Principle</th>
<th>Typical Challenges</th>
<th>Cloudera Enterprise Capabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrity and Confidentiality</td>
<td>Applying industry standard IT security controls to prevent unauthorised access.</td>
<td>Navigator Encrypt and Key Trustee: strong encryption</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Apache Sentry, Okera ODAP: Fine-grained authorization</td>
</tr>
<tr>
<td>Accountability</td>
<td>Demonstrating compliance. Detecting and analyzing breaches within 72 hours.</td>
<td>Cloudera Navigator, Okera ODAP: Comprehensive, inescapable audit trail</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cloudera Cybersecurity solutions</td>
</tr>
<tr>
<td>Lawfulness, fairness and transparency</td>
<td>Implementing a way to keep track of personal data.</td>
<td>Cloudera Navigator, Okera ODAP: Classify/tag and track</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lineage of personal data elements</td>
</tr>
</tbody>
</table>
Pillars of a comprehensive compliance solution
GDPR principles: Integrity, Confidentiality, Accountability, Lawfulness, Fairness, Transparency

Perimeter
Guarding access to the cluster itself

Technical concepts:
Authentication
Network isolation

Cloudera Manager

Access
Defining what users and applications can do with data

Technical concepts:
Permissions
Authorization

Apache Sentry
Okera ODAP

Visibility
Reporting on where data came from and how it’s being used

Technical concepts:
Auditing
Lineage

Cloudera Navigator

Data
Protecting data in the cluster from unauthorized visibility

Technical concepts:
Encryption, tokenization, data masking

Navigator Encrypt & Key
Trustee
Okera ODAP
## How big data solutions can help

<table>
<thead>
<tr>
<th>GDPR Principle</th>
<th>Typical Challenges</th>
<th>Big data solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose limitation</strong></td>
<td>Track consent and data usage while allowing data scientists to mine it using tools of choice.</td>
<td>Cloudera Navigator: DPO can audit exactly how data was used. Cloudera Data Science Workbench: keep data governed</td>
</tr>
<tr>
<td><strong>Data Minimization</strong></td>
<td>Removing or anonymising data where possible. Preventing unlawful data transfers outside the EU while still enabling outsourcing.</td>
<td>Cloudera Navigator: tags can indicate allowed purpose, time limit. Apache Sentry: Redacted views</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Finding a low-overhead way to fix data.</td>
<td>Apache Kudu: Fast updates of individual records</td>
</tr>
<tr>
<td><strong>Storage Limitation</strong></td>
<td>Deleting individual personal data records in Hadoop and Cloud storage, since those file systems are immutable.</td>
<td>Apache Kudu: Fast erasure of individual records. HDFS and Cloud storage options</td>
</tr>
</tbody>
</table>
Enterprise-wide breach detection

Detect advanced threat leveraging machine learning models

Investigate
Search across Apache Spot’s user, endpoint, and network open data models for full context and accelerated investigation

Respond
Use Apache Spot open data models and Cloudera Navigator to see if the threat is widespread
Cloudera’s own GDPR compliance example

Cloudera created a personal data system of record and consent catalog in our internal EDH using Apache Kudu

Before:
- over 4 million rows of personal data across 3 external systems with duplicate records

After:
- a system of record with 2.2 million unique accounts
- a single source of truth for consent
- a process to exercise individual data protection rights