



Executive Briefing

# **Big data in the era of heavy worldwide privacy regulations**

Mark Donsky, Senior Director, Product Management, Okera  
Nik Rouda, Principal Product Marketing Manager, Amazon

---

# Disclaimer

This presentation is intended to help organizations understand how big data platforms can be used to help comply with certain aspects of EU General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) 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 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.

---

# Today's agenda

- GDPR and CCPA overviews
- Requirements for holistic privacy readiness
- Best practices and implementation guidelines
- Case studies

# Key facts on recent privacy regulation

## General Data Protection Regulation (GDPR)

---

- Adopted on April 14, 2016 and enforceable on May 25, 2018
- Applies to all organizations that handle data from EU data subjects
- Fines of up to €20M or 4% of the prior year's turnover
- Standardizes privacy regulation across the European Union

<https://eugdpr.org/>

## California Consumer Protection Act (CCPA)

---

- Signed into law on June 28, 2018 and enforceable on January 1, 2020
- Penalties are up to \$2500 per violation or up to \$7500 per intentional violation
- Clarifications are still being made

<https://oag.ca.gov/privacy/ccpa>

# GDPR vs CCPA: key comparisons

	GDPR	CCPA
<b>Data subjects</b>	Simply refers to “EU data subjects”, some consider this to be EU residents; other consider this to be EU citizens	Applies to California residents
<b>Organizations</b>	All organizations, both public and non-profit	For-profit companies with: (1) gross revenues over \$25M, (2) Possesses the personal information of 50,000 or more consumers, households, or devices, or (3) derive at least 50% of revenue from selling consumer information
<b>Rights</b>	<ul style="list-style-type: none"><li>• The right to erasure</li><li>• The right to access their data</li><li>• The right to correct their data</li><li>• The right to restrict or object to processing of data (opt-in)</li><li>• The right to breach notification within 72 hours of detection</li></ul>	<ul style="list-style-type: none"><li>• The right to know what personal information is being collected about them</li><li>• The right to know whether their personal information is sold or disclosed and to whom</li><li>• The right to say no to the sale of personal information (opt-out)</li><li>• The right to access their personal information</li><li>• The right to equal service and price, even if they exercise their privacy rights</li></ul>

---

# Common CCPA and GDPR objectives

**The right to know:** Under both regulations, consumers and individuals are given bolstered transparency rights to access and request information regarding how their personal data is being used and processed.

**The right to say no:** Both regulations bestow individual rights to limiting the use and sale of personal data, particularly regarding the systematic sale of personal data to third parties, and for limiting analysis/processing beyond the scope of the originally stated purpose.

**The right to have data kept securely:** While differing in approach, both regulations give consumers and individuals mechanisms for ensuring their personal data is kept with reasonable security standards by the companies they interact with.

**The right to data portability:** Both regulations grant consumers rights to have their data transferred in a readily usable format between businesses, such as software services, facilitating consumer choice and helping curb the potential for lock-in.

*“Businesses need to take a more holistic and less regulation-specific approach to data management and compliance to remain competitively viable.”*

Paige Bartley, Senior Analyst, Data Management Data, AI & Analytics  
451 Research

---

# Requirements for holistic privacy readiness

**Know what data you have**

**Know how your data is  
being used**

**Implement privacy by  
design**

**Consent management and  
right to erasure**



# Requirements for holistic privacy readiness

## Know what data you have

- Create a catalog of all data assets
- Tag data sets and columns that contain personal information

## Know how your data is being used

- Auditing
- Lineage

## Implement privacy by design

- Encrypt data
- Restrict access to data with fine-grained access control
- Pseudonymization
- Anonymization

## Consent management and right to erasure

- Implement views that expose only those who have opted in, or hide those who have opted out

# Best practices and implementation guidelines

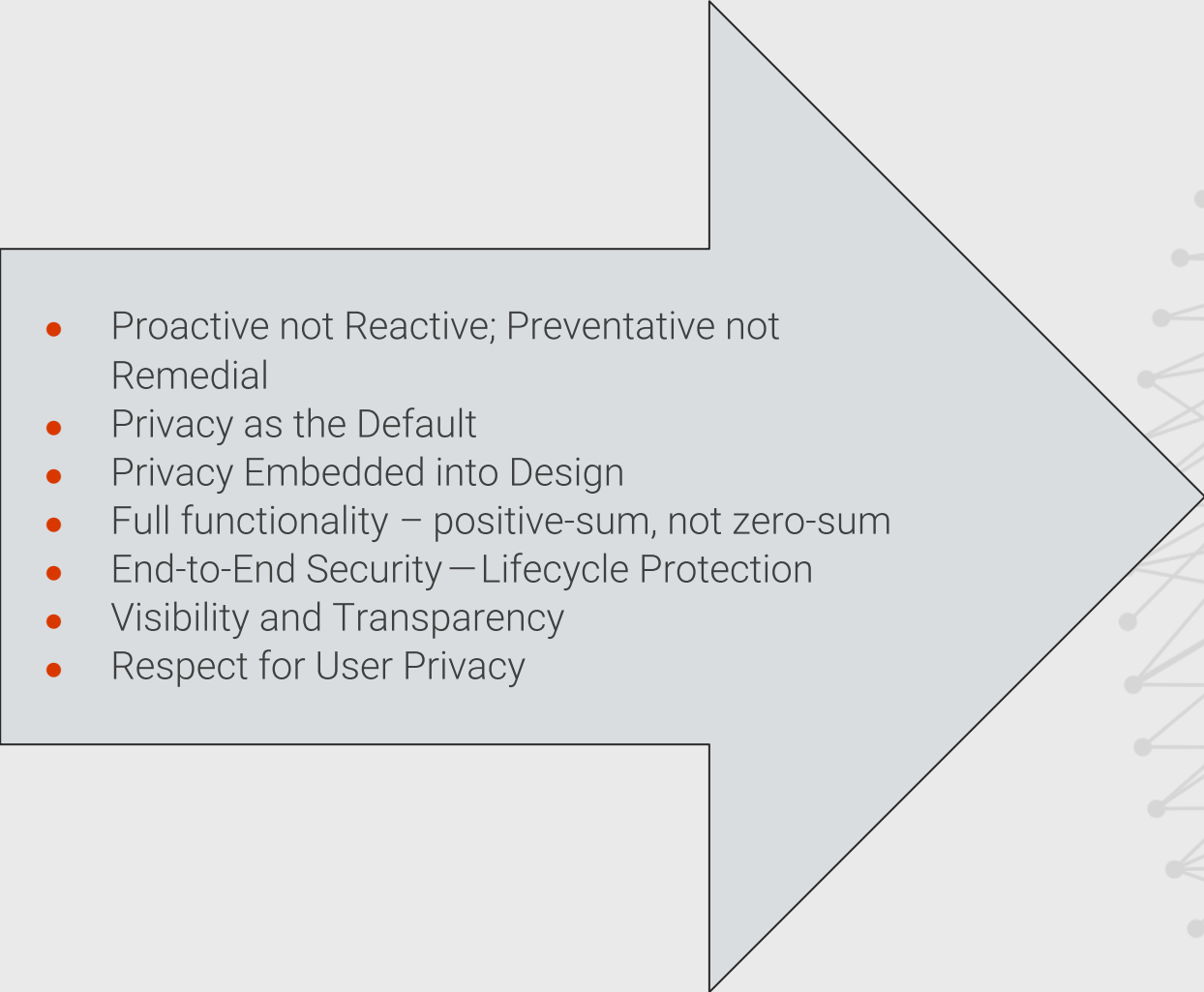


---

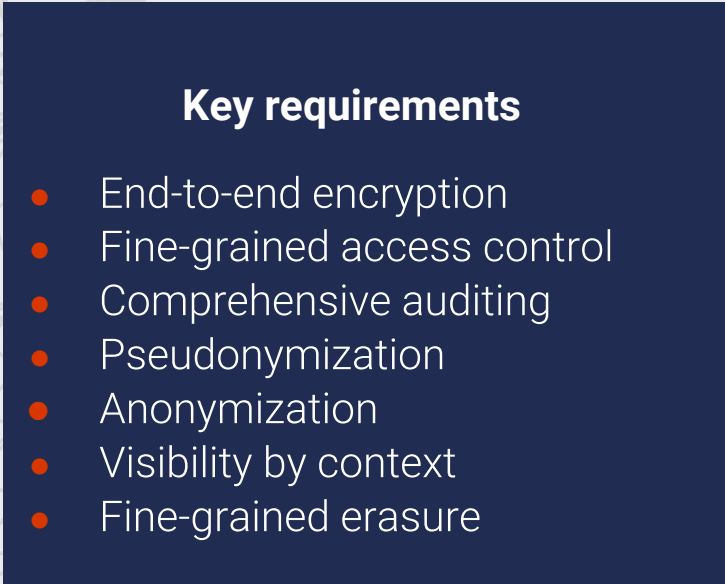
# Implementation guidelines

- Privacy by design
- Pseudonymization and anonymization
- Fine-grained access control
- Consent management and right to erasure

# Privacy by design: key principles

- 
- Proactive not Reactive; Preventative not Remedial
  - Privacy as the Default
  - Privacy Embedded into Design
  - Full functionality – positive-sum, not zero-sum
  - End-to-End Security – Lifecycle Protection
  - Visibility and Transparency
  - Respect for User Privacy

## Key requirements

- 
- End-to-end encryption
  - Fine-grained access control
  - Comprehensive auditing
  - Pseudonymization
  - Anonymization
  - Visibility by context
  - Fine-grained erasure

# Privacy by design: pseudonymization and anonymization

**Pseudonymization:** substitute identifiable data with a reversible, consistent value

**Anonymization:** destroy the the identifiable data

Pseudonymization is a good practice for privacy, but it does not guarantee anonymity

Both can be implemented with transformations of the data, either dynamically or statically.

Name	Pseudonymized	Anonymized
Clyde	qOerd	xxxxxx
Marco	Loqfh	xxxxxx
Les	Mcv	xxxxxx
Les	Mcv	xxxxxx
Marco	Loqfh	xxxxxx
Raul	BhQl	xxxxxx
Clyde	qOerd	xxxxxx

# Privacy by design: Fine-grained access control

Master table

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018 09:33:11	0234837823	238-23-9876	AZP	Sell	DE
16-Feb-2018 11:33:01	3947848494	329-44-9847	TBT	Buy	FR
16-Feb-2018 14:12:34	4848367383	123-56-2345	IDI	Sell	FR
16-Feb-2018 09:22:03	3485739384	585-11-2345	ICBD	Buy	DE
16-Feb-2018 11:55:33	3847598390	234-11-8765	FWQ	Buy	DE
16-Feb-2018 10:22:55	8765432176	344-22-9876	UAD	Buy	FR
16-Feb-2018 13:45:24	3456789012	412-22-8765	NZMA	Sell	FR

What German brokers see

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018 09:33:11	0234837823	xxxx-xx-xxxx	AZP	Sell	DE
16-Feb-2018 09:22:03	3485739384	xxxx-xx-xxxx	ICBD	Buy	DE
16-Feb-2018 11:55:33	3847598390	xxxx-xx-xxxx	FWQ	Buy	DE

Column level anonymization  
Row level filtering

# Consent management and right to erasure

## Consent Management

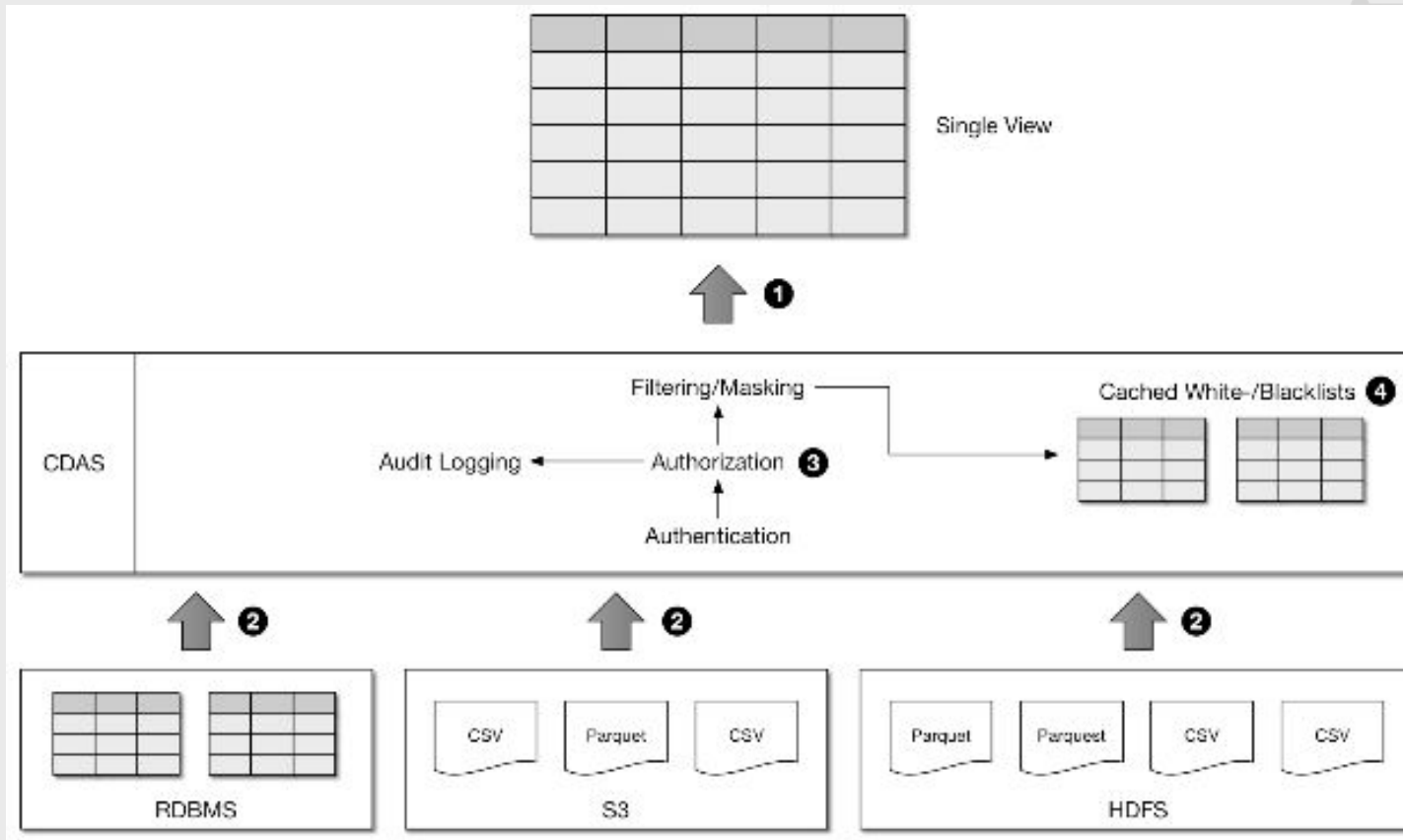
A freely given indication of the data subject's wishes by which he or she signifies agreement to the processing of his or her personal data.

## Right to erasure

Individuals have the right to have personal data erased. This is also known as the "right to be forgotten".

- GDPR requires opt-in for consent management, whereas CCPA allows opt-out
- To implement, you'll need whitelists and blacklists:
  - **Whitelists** (opt-in): A list of all record IDs of subjects that have given consent to the use of their data
  - **Blacklists** (opt-out): A list of record IDs of subjects that have opted out of the use of their data
- Implement consent management by constructing views on top of master tables that join on whitelists or blacklists
- Never provide access to master tables to data consumers

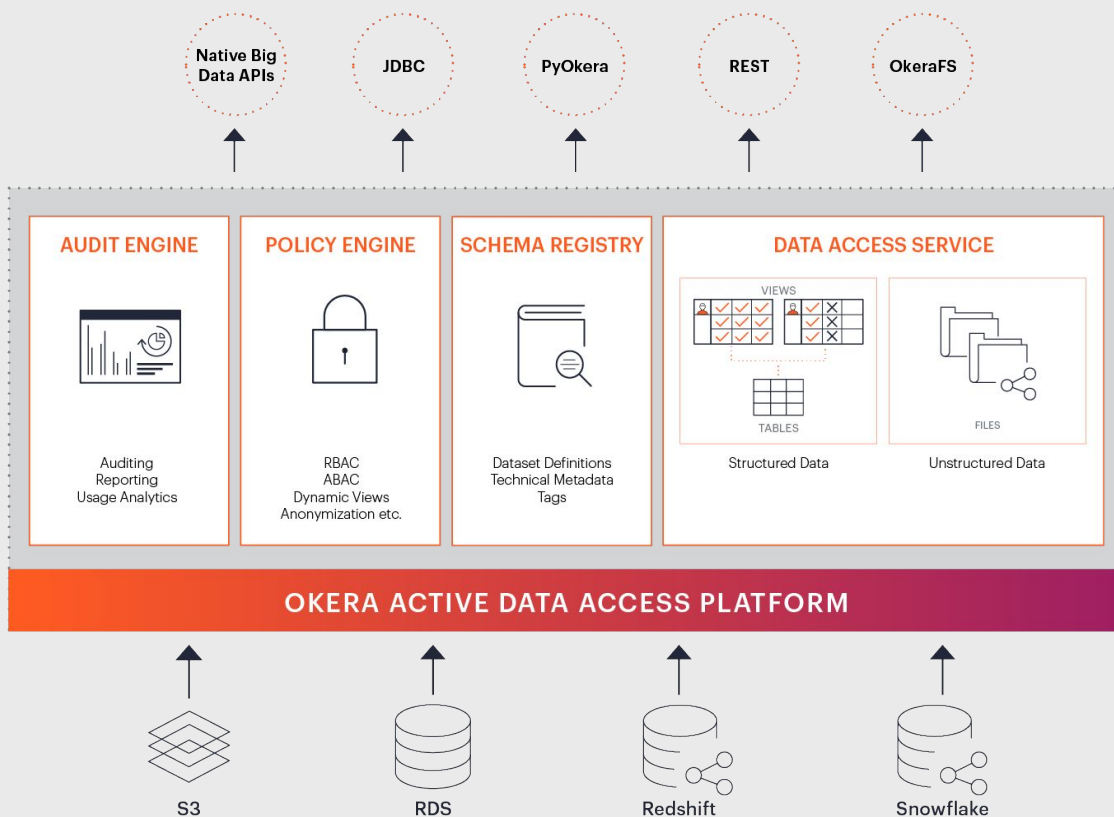
# Consent management and right to erasure in action



1. Instead of rewriting data every time that a person gives consent or opts out, the data is filtered on access using white lists and/or blacklists
2. Convert every data source into a table structure, regardless of the original format
3. Grant access to databases and datasets in a fine-grained manner
4. Note that Okera optimizes cluster resources to cache filter tables, thus making large JOINS very efficient



# Oker's Active Data Access Platform



An active management layer that makes data lakes accessible to multiple workers

- Provides **consistent, airtight protection** with fine-grained access policies
- Supports **all leading analytics tools** and data formats
- Introduces **no delay or additional overhead**

---

# Okera's holistic privacy capabilities

## **Universal policy enforcement across data formats and compute engines**

- Define policies once and enforce everywhere (Spark, SparkSQL, Python, EMR, Hive, etc.)

## **Role-based and Attribute-based access control**

- Enrich data sets with and assign access policies on business context instead of technical metadata (grant access to sensitive sales data to Charlie)

## **Full support for both pseudonymization and anonymization**

- Enrich data sets with and assign access policies on business context instead of technical metadata (show last four digits of SSN, replace email address with email@redacted.com)

## **Dynamic views for right to erasure, consent management, and easy administration**

- Simplify view administration with dynamic policies that include Hive UDFs that are evaluated just-in-time (e.g., has\_access, has\_roles)

# Okera case study: Top 5 Global Apparel Company

## Requirements

- GDPR compliance: PII anonymization, right to erasure, consent management (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

“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.”



## 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

### ENVIRONMENT

- AWS
- S3-based data lake
- Presto, Databricks, for interactive BI



Thank you!

Mark Donsky [mark@okera.com](mailto:mark@okera.com)  
Nik Rouda [nrouda@amazon.com](mailto:nrouda@amazon.com)