Gartner

COOL VENDOR 2019

De-Identifying Data in Snowflake and Amazon Redshift

Harold Byun

VP Products

baffle

Introduction

- Overview of Data Analytics Trends and the Move to Cloud Data Lakes
- Key Data Privacy Challenges
- Methods for Data De-Identification
- Architecture Models to Support a De-Identified Data Pipeline
- Live Demo of De-Identification and Data Processing
- A Glimpse Into Privacy Preserving and Advanced Data Analytics
- Q&A

Questions throughout – use the chat panel

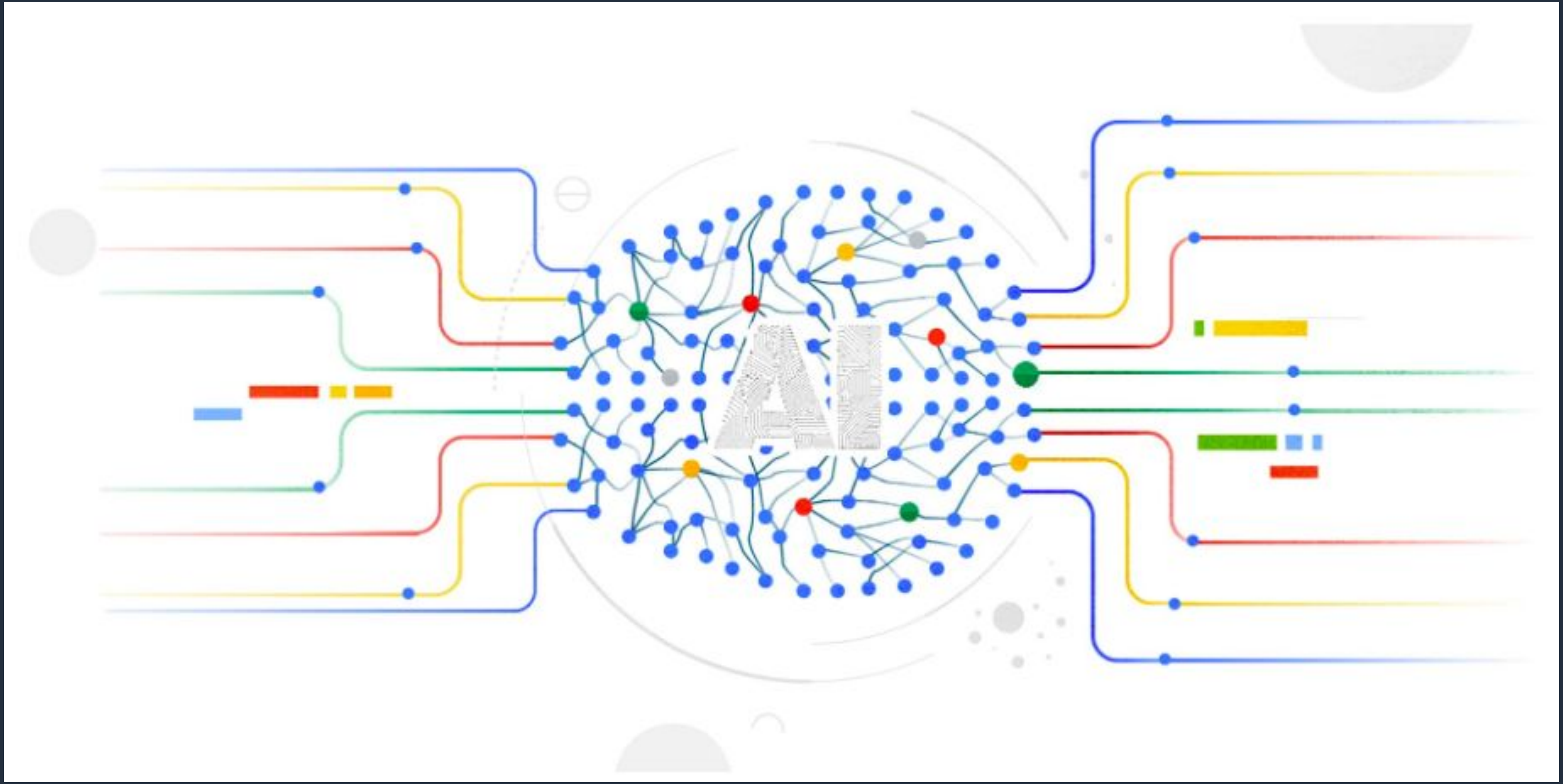
Email info@baffle.io, harold@baffle.io

Speaker Bio

Harold Byun is VP of Products at Baffle, an end-to-end data-centric protection company. His career has focused on data containment and security technologies including data loss prevention and activity monitoring, cloud access security broker, and mobile data containment capabilities. He holds several data security related patents.

Overview of Data Analytics Trends and the Move to Cloud Data Lakes

AI and Big Data are a Big Deal

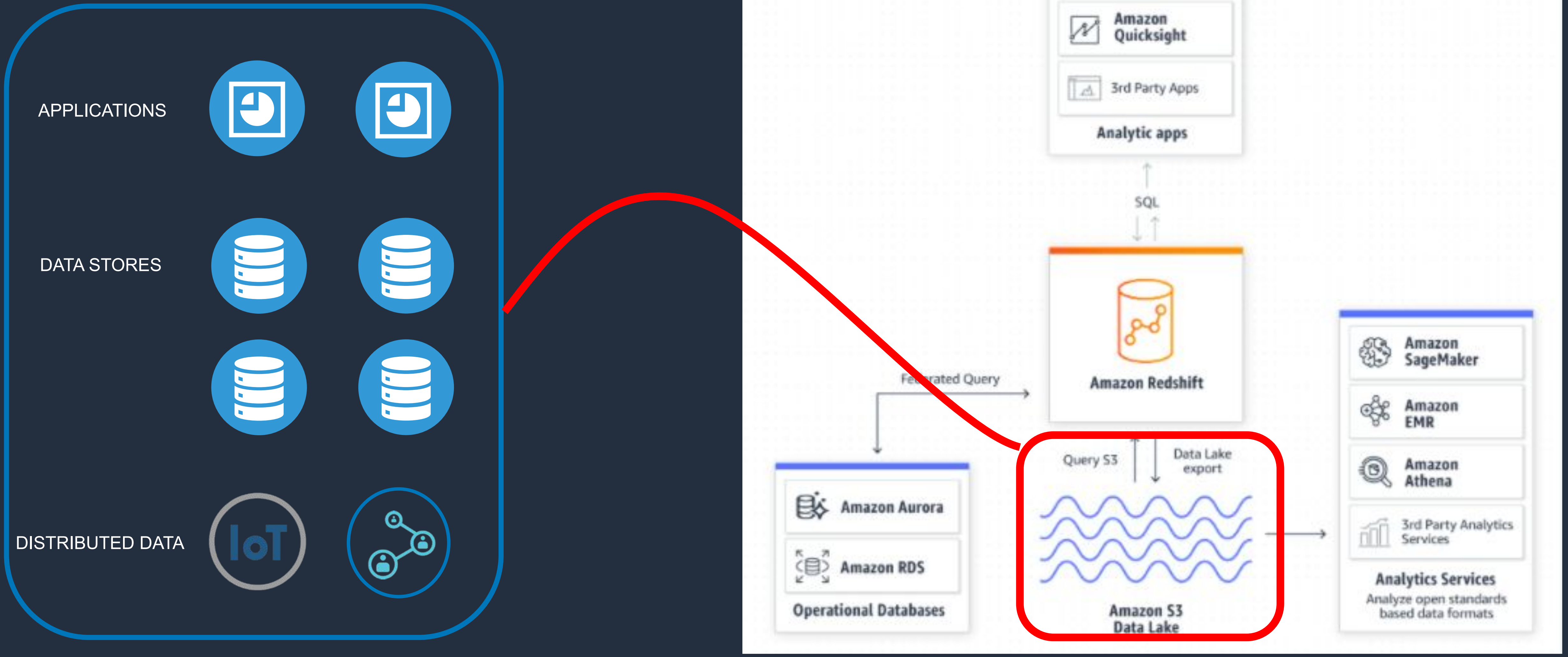


Trends Impacting Cloud Data Analytics and Data Lakes

1. By the end of 2024, 75% of organizations will shift from piloting to operationalizing artificial intelligence (AI), driving a 5 times increase in streaming data and analytics infrastructures. (Gartner)
2. By 2022, 35% of large organizations will be either sellers or buyers of data via formal online data marketplaces, up from 25% in 2020 (Gartner)
3. Existing on-premise big data environments remain static and are running out of room
4. A significant move to leverage cloud-based data lakes for analytics and AI/ML
5. Continued inadvertent exposure of data in aggregated environments

Moving to Cloud-based Data Lakes

ENTERPRISE – CURRENT STATE



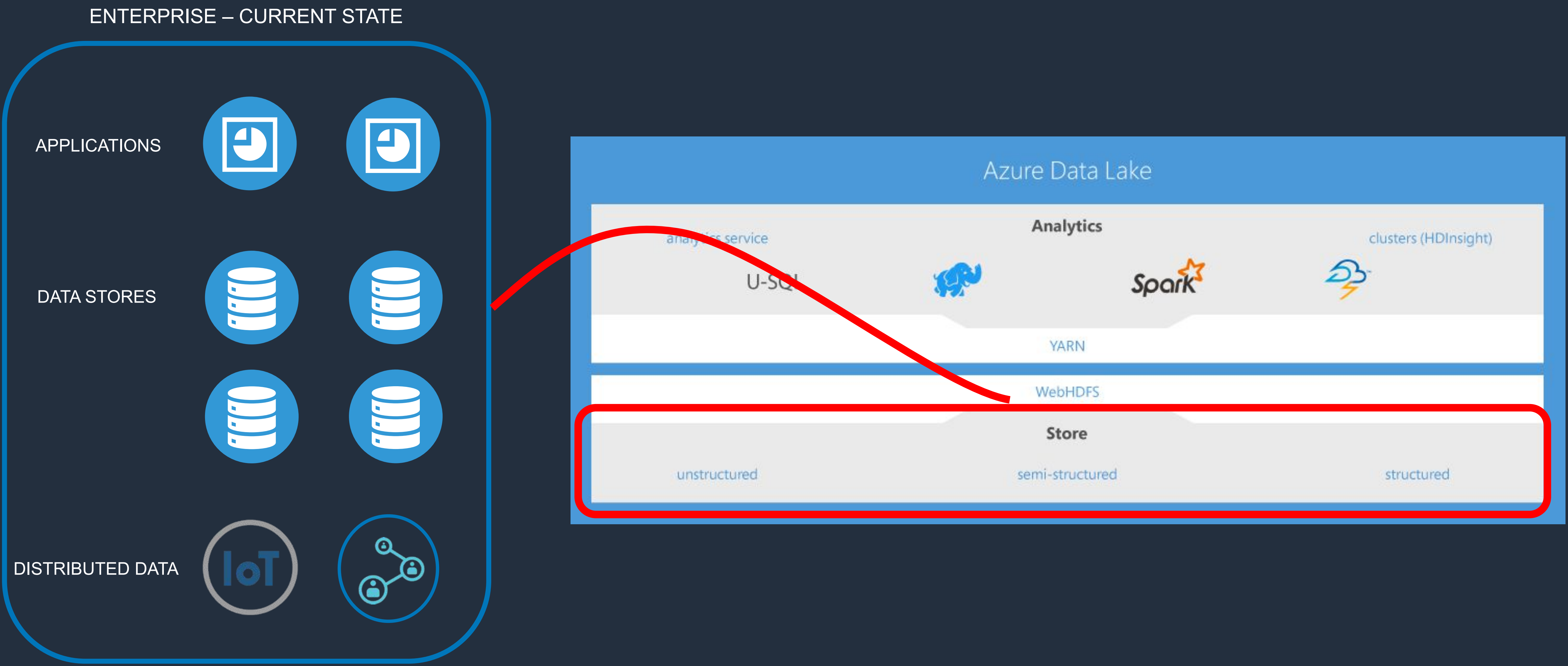
APPLICATIONS

DATA STORES

DISTRIBUTED DATA

Confidential & Proprietary

Moving to Cloud-based Data Lakes



Key Data Privacy Challenges

Continued Data Exposure or Leakage

1

2

3

Data breaches continue unabated

Data loss and leakage is the #1 cloud security concern (2019 Cloud Security Report)

Third party risk and data sharing

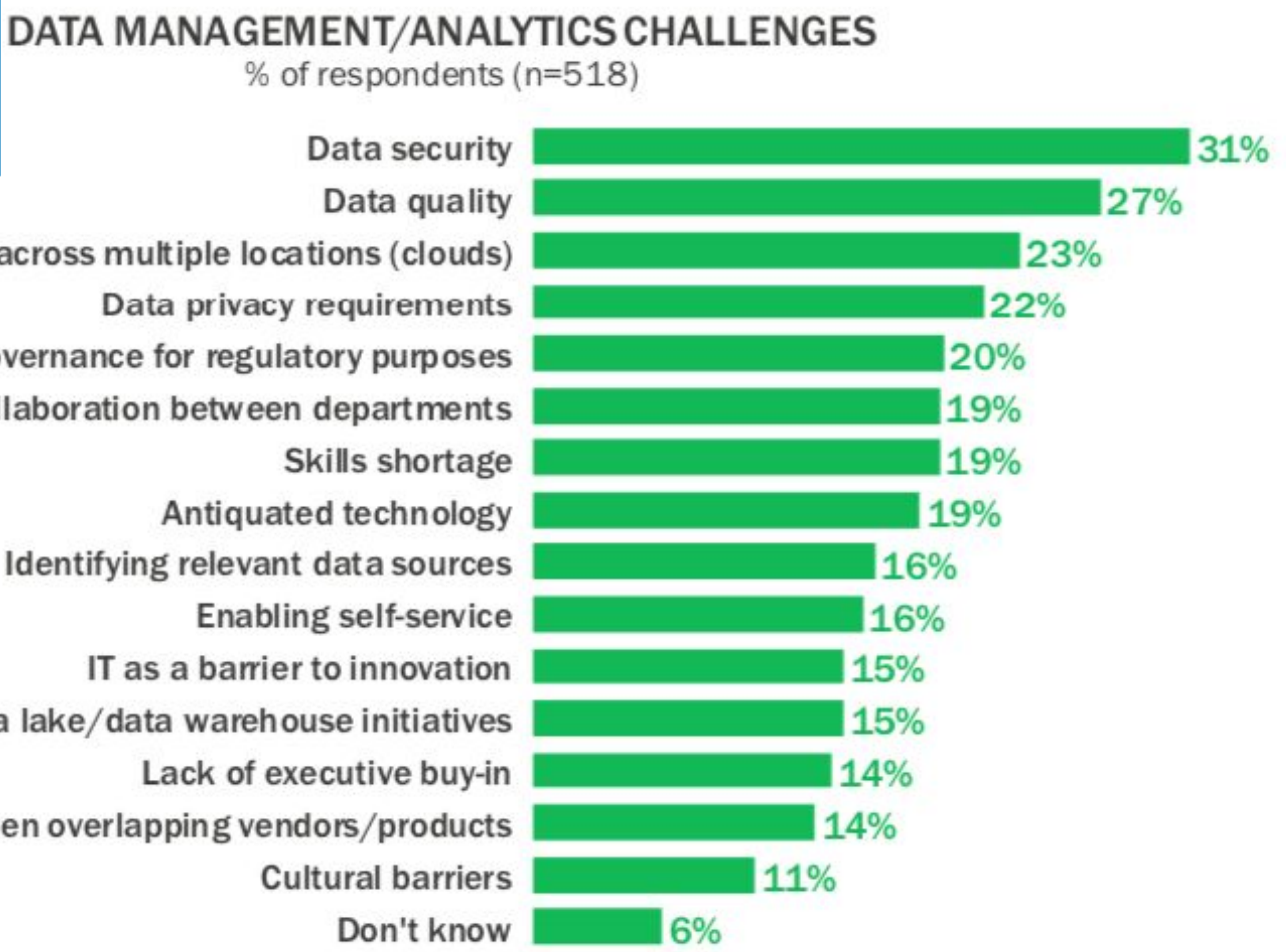
~60% of CISOs have reported data leakage via a third party in 2018. (Ponemon Institute)

Cloud storage data leaks continue

Over 1 billion records leaked and an estimated 11% of cloud storage left open to public

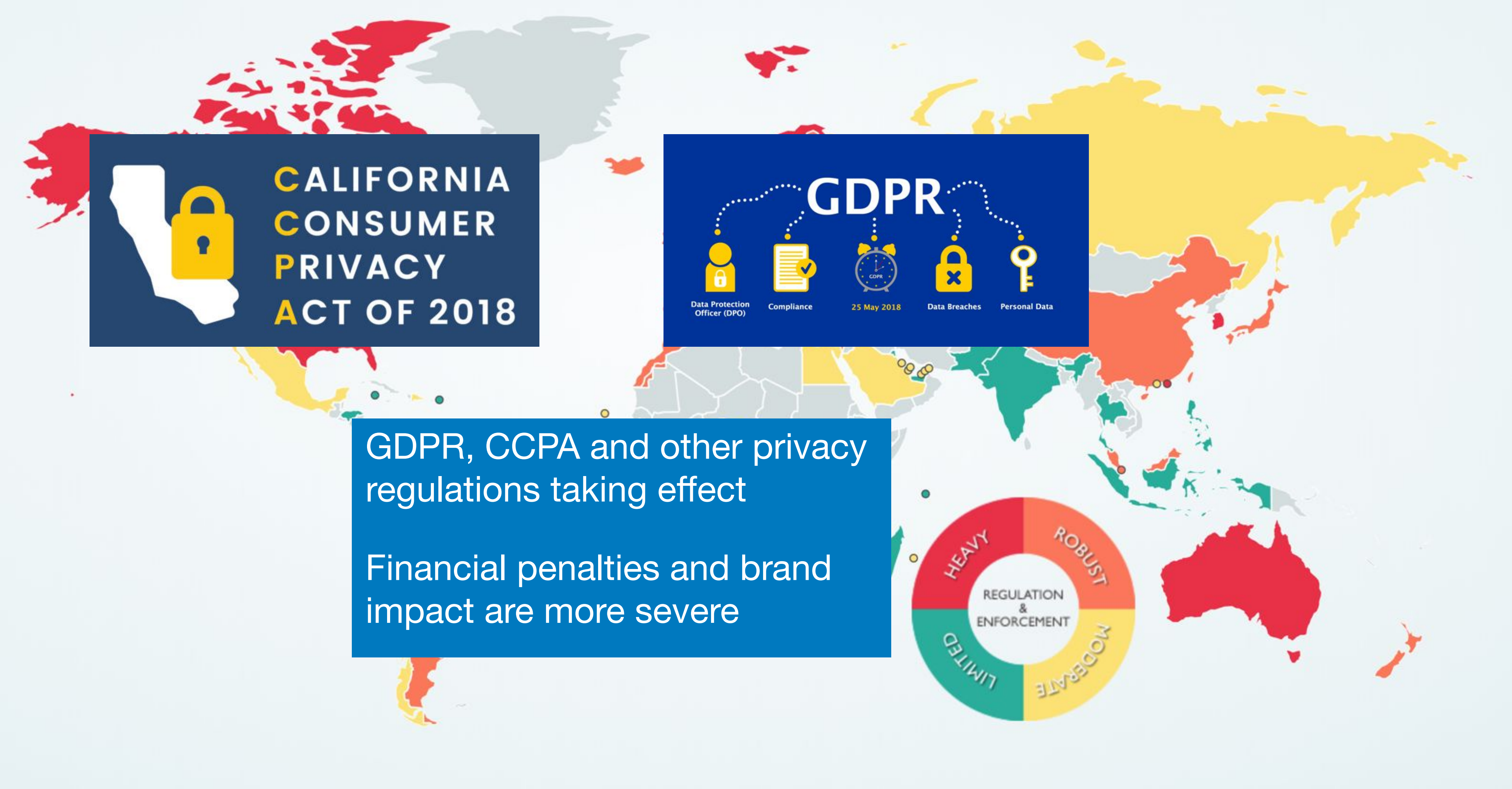
Data Analytics Challenges

Q: What are the biggest data management/analytics challenges faced by your organization?



Source: 451 Research's Voice of the Enterprise: Data & Analytics, 1H 2019

Privacy Around the World



GDPR, CCPA and other privacy regulations taking effect

Financial penalties and brand impact are more severe

Source: https://www.dlapiperdataprotection.com/index.html?t=about&c=AO

Data Privacy Enforced

16 JUL 2020 NEWS

Walmart Sued Under CCPA After Data Breach

Source: 451 Research's Voice of the Enterprise: Data & Analytics, 1H 2019

Data Privacy Resources

Gartner Report on Privacy Preserving Analytics

Gartner

Licensed for Distribution

By Bart Willemsen, Jie Zhang, Nader Henein

Learn how Gartner can help you succeed

Overview

Key Findings

[illegible]

Recommendations

[illegible]

Analysis

[illegible]

What You Need to Know

[illegible]

Data is the heart of all business intelligence (BI) and analytics activities, yet all personal data brings privacy risk with it — a risk that must be treated to ensure that value drawn from insights can actually be used.

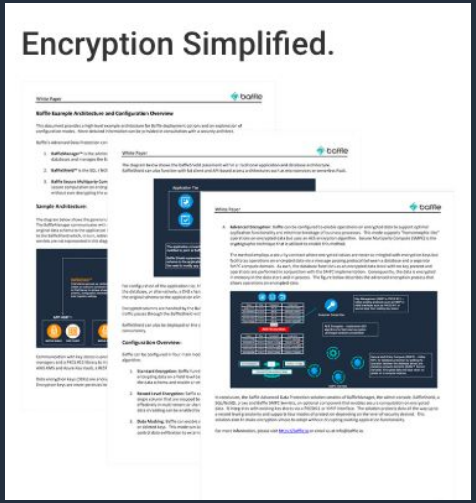
CCPA Compliance Simplified

baffle

CCPA Compliance Simplified

Methods you can use to accelerate and simplify compliance with the California Consumer Privacy Act (CCPA)

Encryption Simplified White Paper



Encryption Simplified.

White Paper

baffle

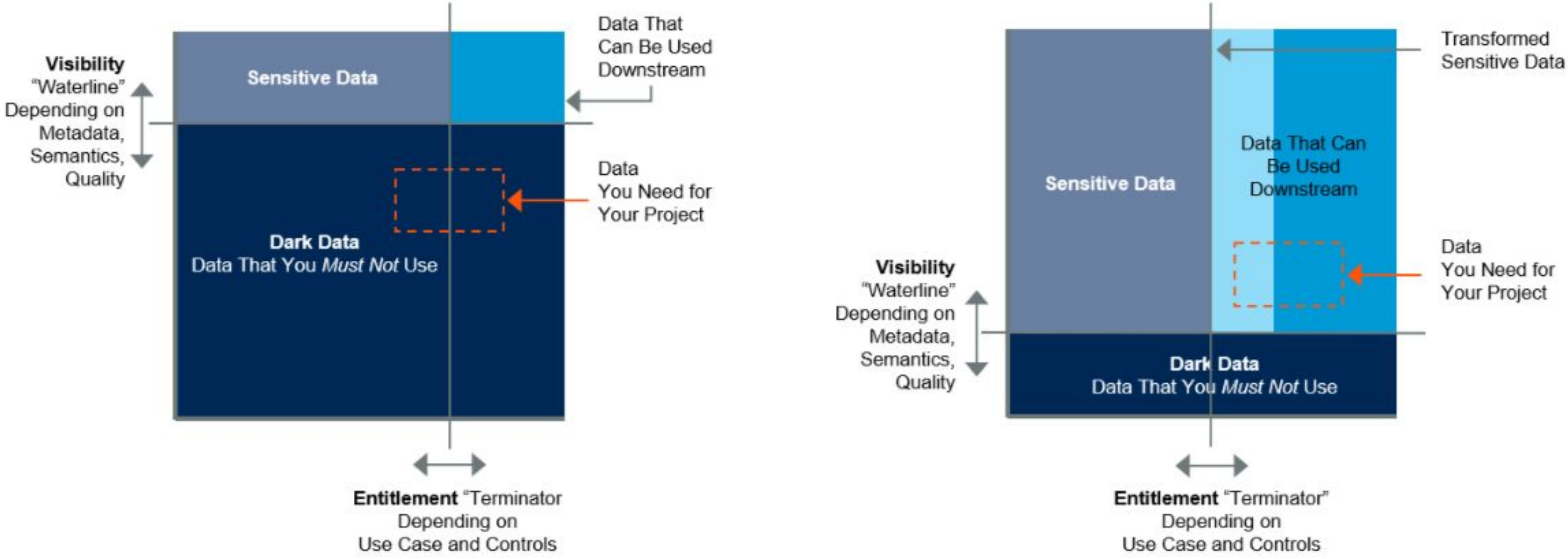
Baffle Example Architecture and Configuration Overview

Sample Architecture:

Configuration Overview:

Privacy? So What, You’re Going to Collect Data Anyway

Continued Data Exposure or Leakage



Source: Gartner
ID: 464663

Source: Gartner, “Securing the Data and Advanced Analytics Pipeline”, 27 Jan 2020

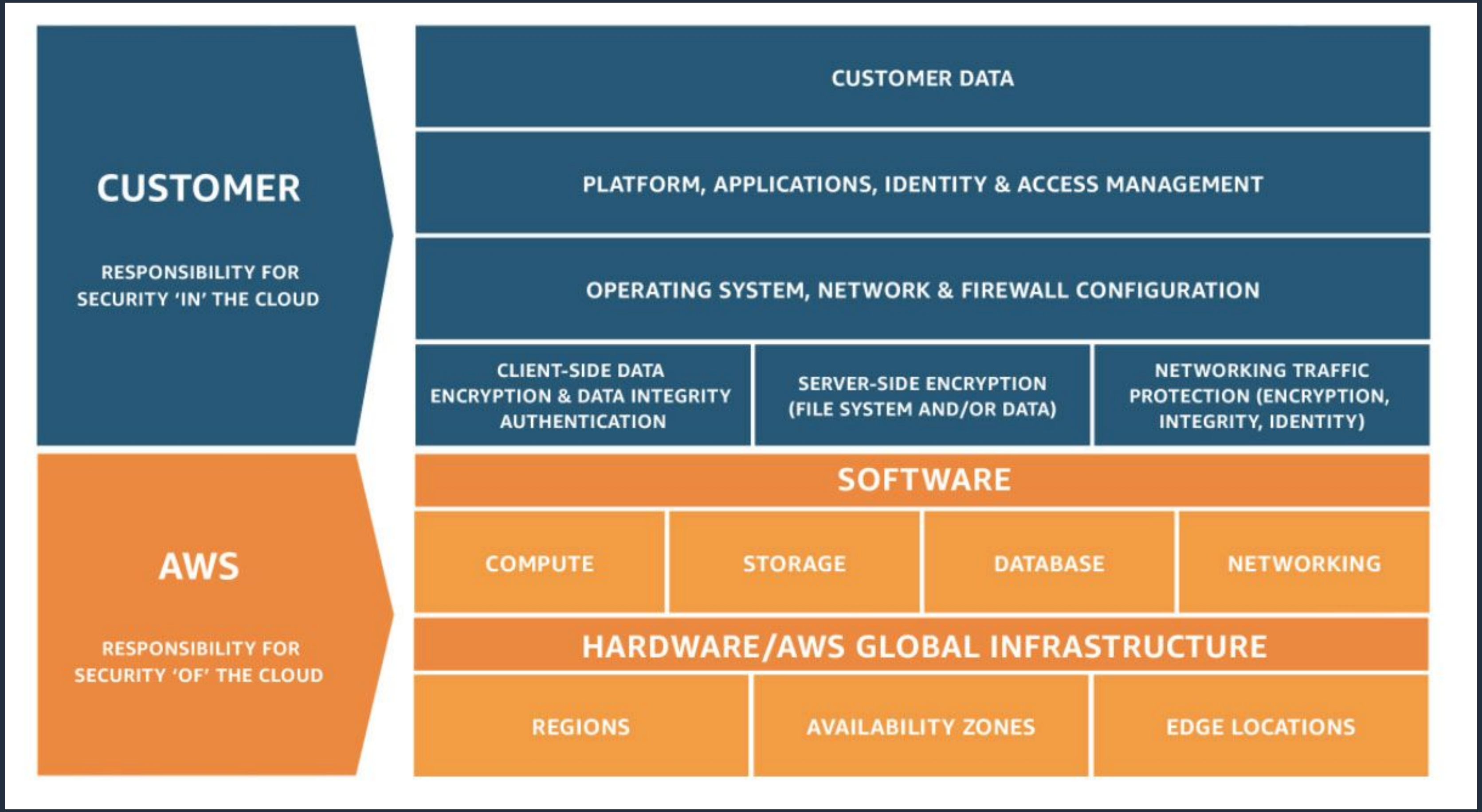
Methods for Data De-Identification

Infrastructure vs. Data

Customer responsibility “Security in the Cloud”

AWS responsibility “Security of the Cloud”

AWS is responsible for protecting the infrastructure that runs all of the services offered in the AWS Cloud.



CUSTOMER

RESPONSIBILITY FOR SECURITY ‘IN’ THE CLOUD

CUSTOMER DATA

PLATFORM, APPLICATIONS, IDENTITY & ACCESS MANAGEMENT

OPERATING SYSTEM, NETWORK & FIREWALL CONFIGURATION

CLIENT-SIDE DATA ENCRYPTION & DATA INTEGRITY AUTHENTICATION

SERVER-SIDE ENCRYPTION (FILE SYSTEM AND/OR DATA)

NETWORKING TRAFFIC PROTECTION (ENCRYPTION, INTEGRITY, IDENTITY)

AWS

RESPONSIBILITY FOR SECURITY ‘OF’ THE CLOUD

SOFTWARE

COMPUTE

STORAGE

DATABASE

NETWORKING

HARDWARE/AWS GLOBAL INFRASTRUCTURE

REGIONS

AVAILABILITY ZONES

EDGE LOCATIONS

Existing Infrastructure Control Methods

NOTE: This is not an exhaustive list

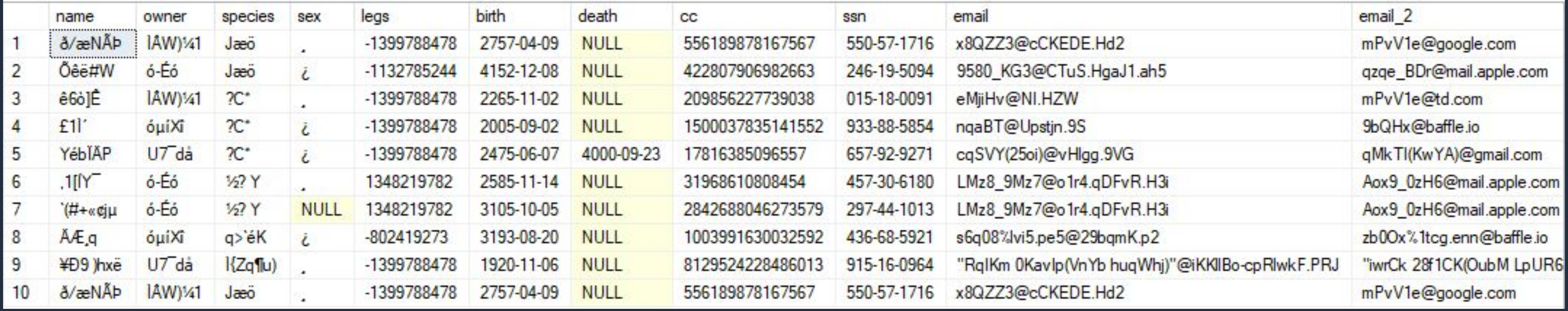
AWS	Azure
Block S3 public access	Azure AD integration for authorization to Azure Blob Storage
Bucket ACLs	Azure AD, roles and secure access signatures (SAS)
IAM Roles for controlling access from instances	Secure Access Signatures – SAS allows for a URI with resource and query parameters to restrict access and authorization to storage resources. Can be established as a service or user delegation
Monitoring and Logging: - Policy-based discovery for open principal access ”*” - ListBucket assessments - Access monitoring with CloudWatch, CloudTrail - Discovery via Macie	Monitoring and Logging: - Advanced Threat Protection - Access monitoring via Azure Monitor
Encryption at-rest: - SSE S3 – Server-side encryption with AWS Managed Keys - SSE-KMS – Server-side encryption with customer keys stored in AWS KMS - SSE-C – Server-side encryption with customer provided keys - Client-Side Encryption – Data is encrypted before upload using client encryption	Encryption at-rest: - Enabled by default for all blobs - Microsoft-managed keys – blob encryption using a Microsoft key store - Azure Key Vault – Customer-managed keys to encrypt blob storage and Azure files - Customer-provided keys – customer owned key store used to encrypt blobs
HTTPS / TLS – Encryption in-transit	HTTPS / TLS – Encryption in-transit
VPC Endpoints – Establishes S3 connectivity via VPC to prevent traffic from traversing the public internet	Azure Private Endpoints – Enables connectivity via VPC to prevent traffic from traversing the public internet

Common Methods for De-Identification

Supported Data Protection Modes	Description
Data Encryption	Table or column-based encryption using randomized, deterministic AES-CTR encryption or FPE
Secure Data Tokenization (TOK)	Uses deterministic AES encryption to generate a deterministic encrypted transform for a given value. Can be applied to support JOINs and foreign key constraints to preserve referential integrity. Does NOT use code book method
Format Preserving Encryption (FPE)	Supports encryption where the cipher text output has the same form of the input. Preserves length of the data type. Can be applied to support JOINs and foreign key constraints to preserve referential integrity. Does NOT use code book method. Cannot be used in conjunction with RLE or Advanced Encryption. Baffle uses NIST approved FF1 and FF3-1 algorithms for FPE
Data Masking	Supports a library of masking formats that protects data at the presentation layer to prevent users from viewing data in the clear. Masking can be applied using static alphanumeric characters, randomly generated data values, and/or partially mask data values. Masking can be applied to both clear text and/or encrypted data
Role-based Data Masking	Supports role or group-based policies in conjunction with data masking policies to restrict viewing of data based on group membership or other attribution.
Advanced Encryption (SMPC)	Support for privacy preserving analytics and secure data sharing on encrypted table or columnar data using randomized AES and secure multiparty compute (SMPC). This encryption mode facilitates operations and analytics on encrypted data across multiple parties without revealing data to other participating parties.

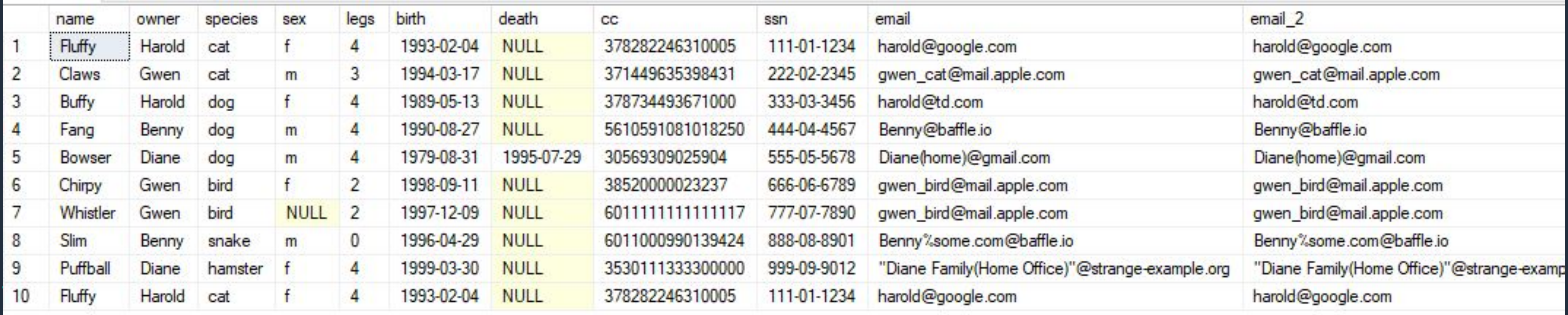
Objects Encryption vs. Data-Centric Encryption

ENCRYPTED DATA



	name	owner	species	sex	legs	birth	death	cc	ssn	email	email_2
1	ð/æNĂÞ	ÌÅW)¼1	Jæö	¸	-1399788478	2757-04-09	NULL	556189878167567	550-57-1716	x8QZZ3@cCKEDE.Hd2	mPvV1e@google.com
2	Õêë#W	ó-Éó	Jæö	¿	-1132785244	4152-12-08	NULL	422807906982663	246-19-5094	9580_KG3@CTuS.HgaJ1.ah5	qzqe_BDr@mail.apple.com
3	ê6ò]Ê	ÌÅW)¼1	?C*	¸	-1399788478	2265-11-02	NULL	209856227739038	015-18-0091	eMjiHv@NI.HZW	mPvV1e@td.com
4	£1Ì´	óµíXî	?C*	¿	-1399788478	2005-09-02	NULL	1500037835141552	933-88-5854	nqaBT@Upstjn.9S	9bQHx@baffle.io
5	YébÏĂP	U7¯då	?C*	¿	-1399788478	2475-06-07	4000-09-23	17816385096557	657-92-9271	cqSVY(25oi)@vHlgg.9VG	qMkTI(KwYA)@gmail.com
6	.1[ÍY¯	ó-Éó	½? Y	¸	1348219782	2585-11-14	NULL	31968610808454	457-30-6180	LMz8_9Mz7@o1r4.qDFvR.H3i	Aox9_0zH6@mail.apple.com
7	`(#+«¢jµ	ó-Éó	½? Y	NULL	1348219782	3105-10-05	NULL	2842688046273579	297-44-1013	LMz8_9Mz7@o1r4.qDFvR.H3i	Aox9_0zH6@mail.apple.com
8	ĂÆ¸q	óµíXî	q>`éK	¿	-802419273	3193-08-20	NULL	1003991630032592	436-68-5921	s6q08%Jvi5.pe5@29bqmK.p2	zb0Ox%1tcg.enn@baffle.io
9	¥Ð9)hxë	U7¯då	Ì{Zq¶u)	¸	-1399788478	1920-11-06	NULL	8129524228486013	915-16-0964	"RqIKm 0KavIp(VnYb huqWhj)"@iKKlIBo-cpRlwkF.PRJ	"iwrCk 28f1CK(OubM LpUR6
10	ð/æNĂÞ	ÌÅW)¼1	Jæö	¸	-1399788478	2757-04-09	NULL	556189878167567	550-57-1716	x8QZZ3@cCKEDE.Hd2	mPvV1e@google.com

CLEAR TEXT DATA



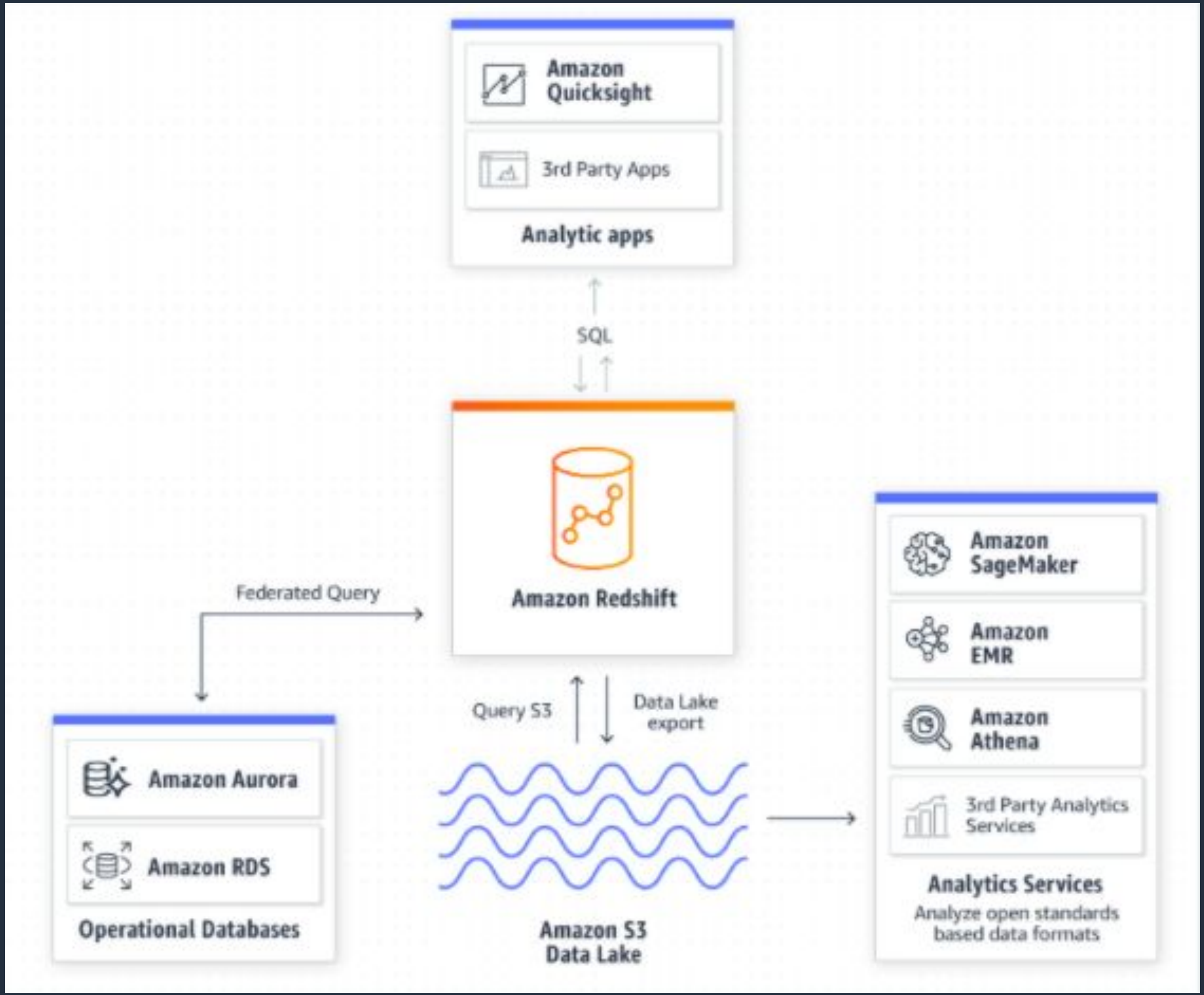
	name	owner	species	sex	legs	birth	death	cc	ssn	email	email_2
1	Fluffy	Harold	cat	f	4	1993-02-04	NULL	378282246310005	111-01-1234	harold@google.com	harold@google.com
2	Claws	Gwen	cat	m	3	1994-03-17	NULL	371449635398431	222-02-2345	gwen_cat@mail.apple.com	gwen_cat@mail.apple.com
3	Buffy	Harold	dog	f	4	1989-05-13	NULL	378734493671000	333-03-3456	harold@td.com	harold@td.com
4	Fang	Benny	dog	m	4	1990-08-27	NULL	5610591081018250	444-04-4567	Benny@baffle.io	Benny@baffle.io
5	Bowser	Diane	dog	m	4	1979-08-31	1995-07-29	30569309025904	555-05-5678	Diane(home)@gmail.com	Diane(home)@gmail.com
6	Chirpy	Gwen	bird	f	2	1998-09-11	NULL	38520000023237	666-06-6789	gwen_bird@mail.apple.com	gwen_bird@mail.apple.com
7	Whistler	Gwen	bird	NULL	2	1997-12-09	NULL	6011111111111117	777-07-7890	gwen_bird@mail.apple.com	gwen_bird@mail.apple.com
8	Slim	Benny	snake	m	0	1996-04-29	NULL	6011000990139424	888-08-8901	Benny%some.com@baffle.io	Benny%some.com@baffle.io
9	Puffball	Diane	hamster	f	4	1999-03-30	NULL	3530111333300000	999-09-9012	"Diane Family(Home Office)"@strange-example.org	"Diane Family(Home Office)"@strange-examp
10	Fluffy	Harold	cat	f	4	1993-02-04	NULL	378282246310005	111-01-1234	harold@google.com	harold@google.com

Key Benefits

- De-identify, tokenize or encrypt data INSIDE objects and files
- Safe harbor from accidental data leaks from key privacy and compliance regulations
- Accelerate cloud-based data analytics programs by addressing key security and privacy concerns

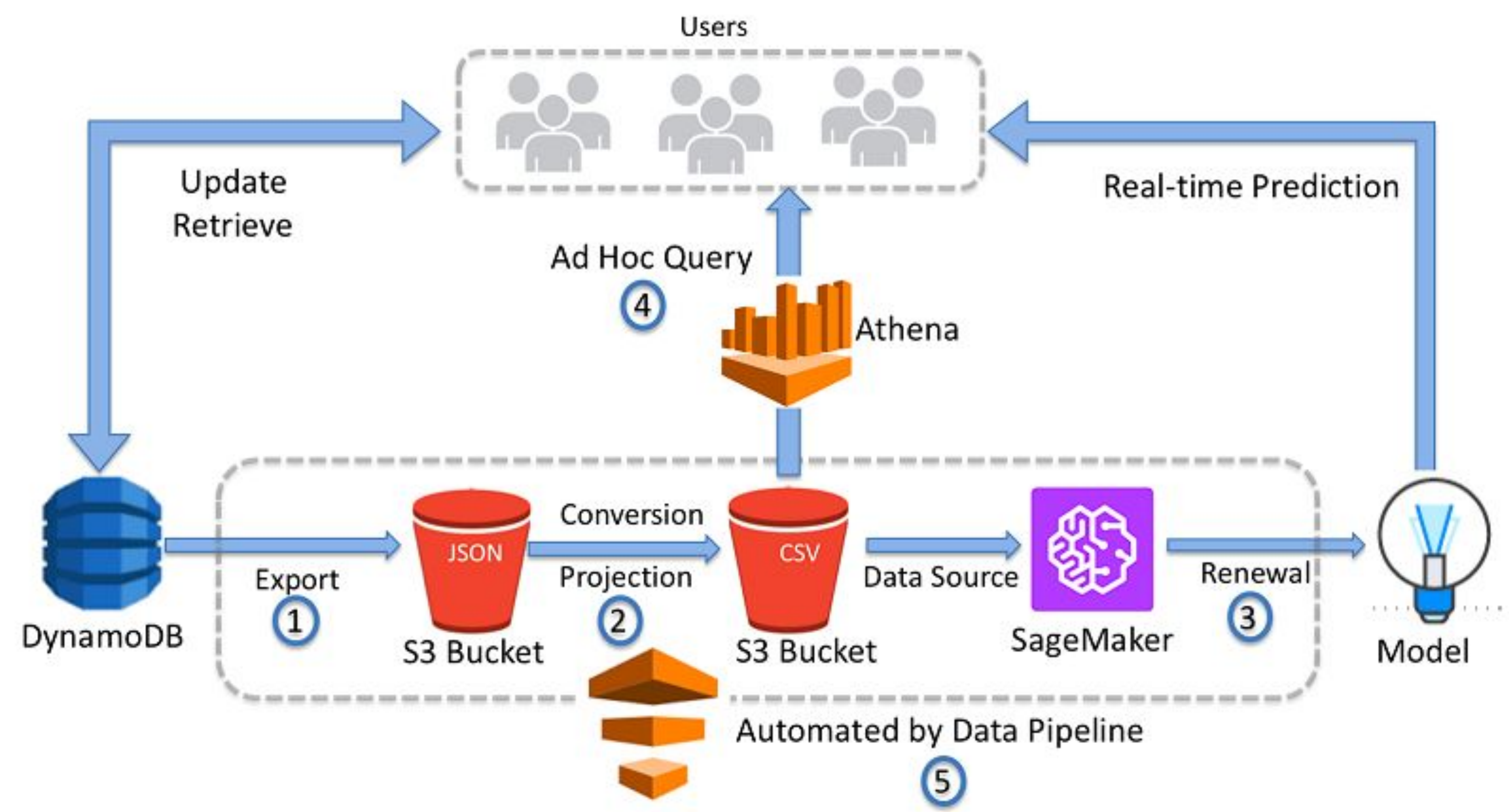
Architecture Models for a De-Identified Data Pipeline

Data Pipeline Architecture

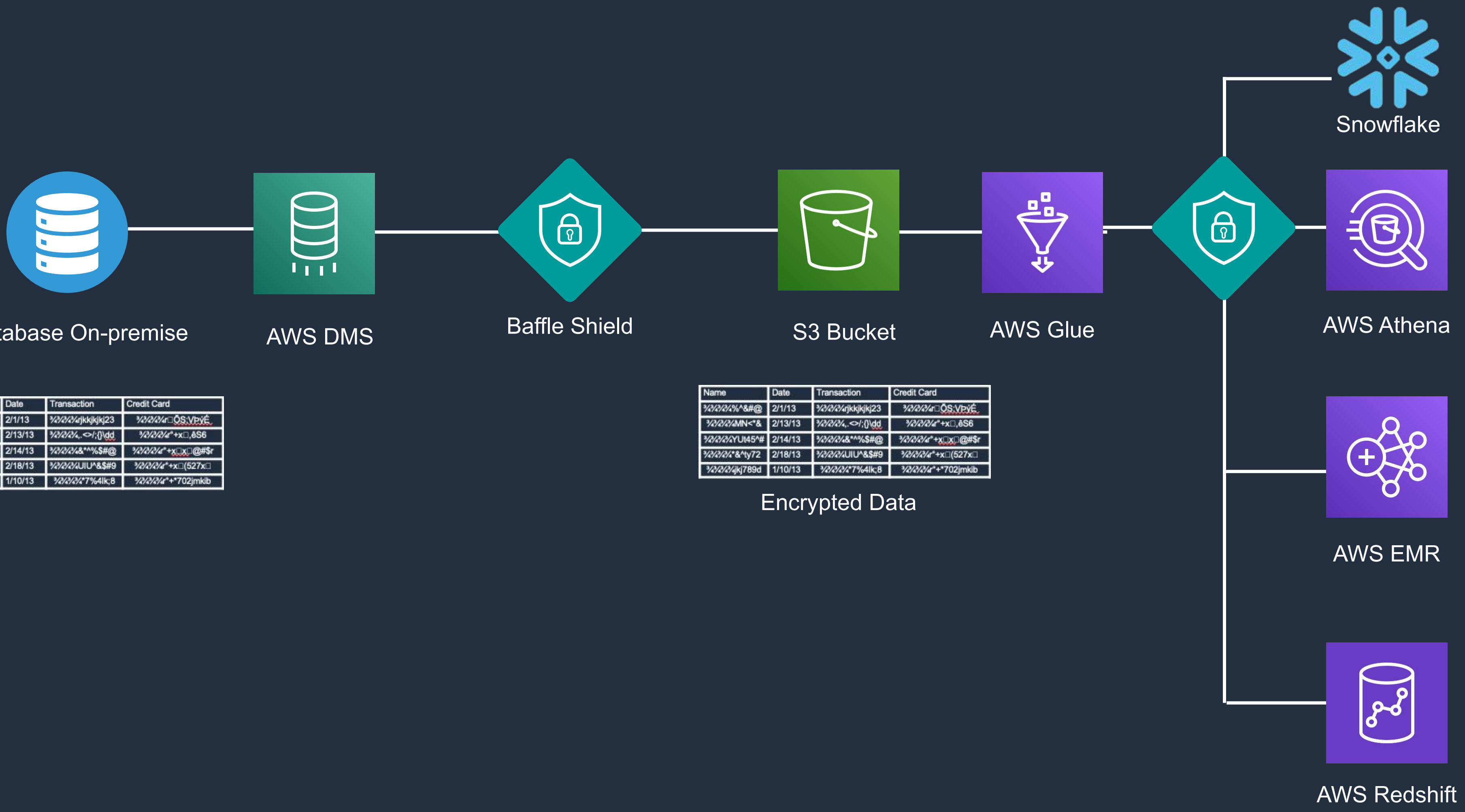


Confidential & Proprietary

Data Pipeline Example



Example of a De-Identified Pipeline



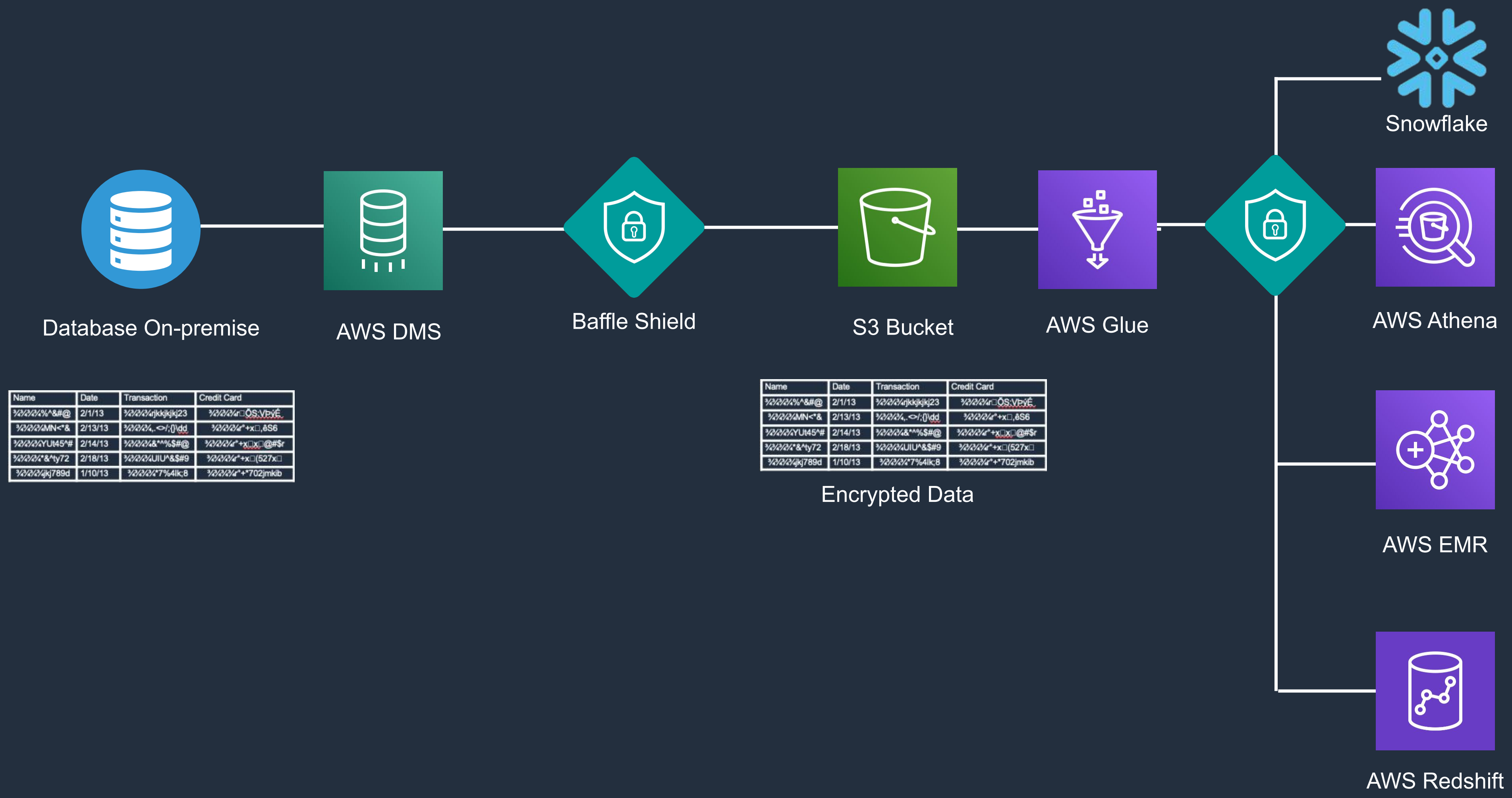
Name	Date	Transaction	Credit Card
¾¾¾¾%^&#@	2/1/13	¾¾¾¾rjkkjkjkj23	¾¾¾¾r□ÕS;VPyÉ
¾¾¾¾MN<*&	2/13/13	¾¾¾¾,.<>/;{}\dd	¾¾¾¾r"+x□,êS6
¾¾¾¾YUt45^#	2/14/13	¾¾¾¾&*^%$#@	¾¾¾¾r"+x□x□@#$r
¾¾¾¾*&^ty72	2/18/13	¾¾¾¾UIU^&$#9	¾¾¾¾r"+x□(527x□
¾¾¾¾jkj789d	1/10/13	¾¾¾¾*7%4lk;8	¾¾¾¾r"+*702jmkib

Name	Date	Transaction	Credit Card
¾¾¾¾%^&#@	2/1/13	¾¾¾¾rjkkjkjkj23	¾¾¾¾r□ÕS;VPyÉ
¾¾¾¾MN<*&	2/13/13	¾¾¾¾,.<>/;{}\dd	¾¾¾¾r"+x□,êS6
¾¾¾¾YUt45^#	2/14/13	¾¾¾¾&*^%$#@	¾¾¾¾r"+x□x□@#$r
¾¾¾¾*&^ty72	2/18/13	¾¾¾¾UIU^&$#9	¾¾¾¾r"+x□(527x□
¾¾¾¾jkj789d	1/10/13	¾¾¾¾*7%4lk;8	¾¾¾¾r"+*702jmkib

Encrypted Data

Live Demo

Example of a De-Identified Pipeline



Database On-premise

AWS DMS

Baffle Shield

S3 Bucket

AWS Glue

Snowflake

AWS Athena

AWS EMR

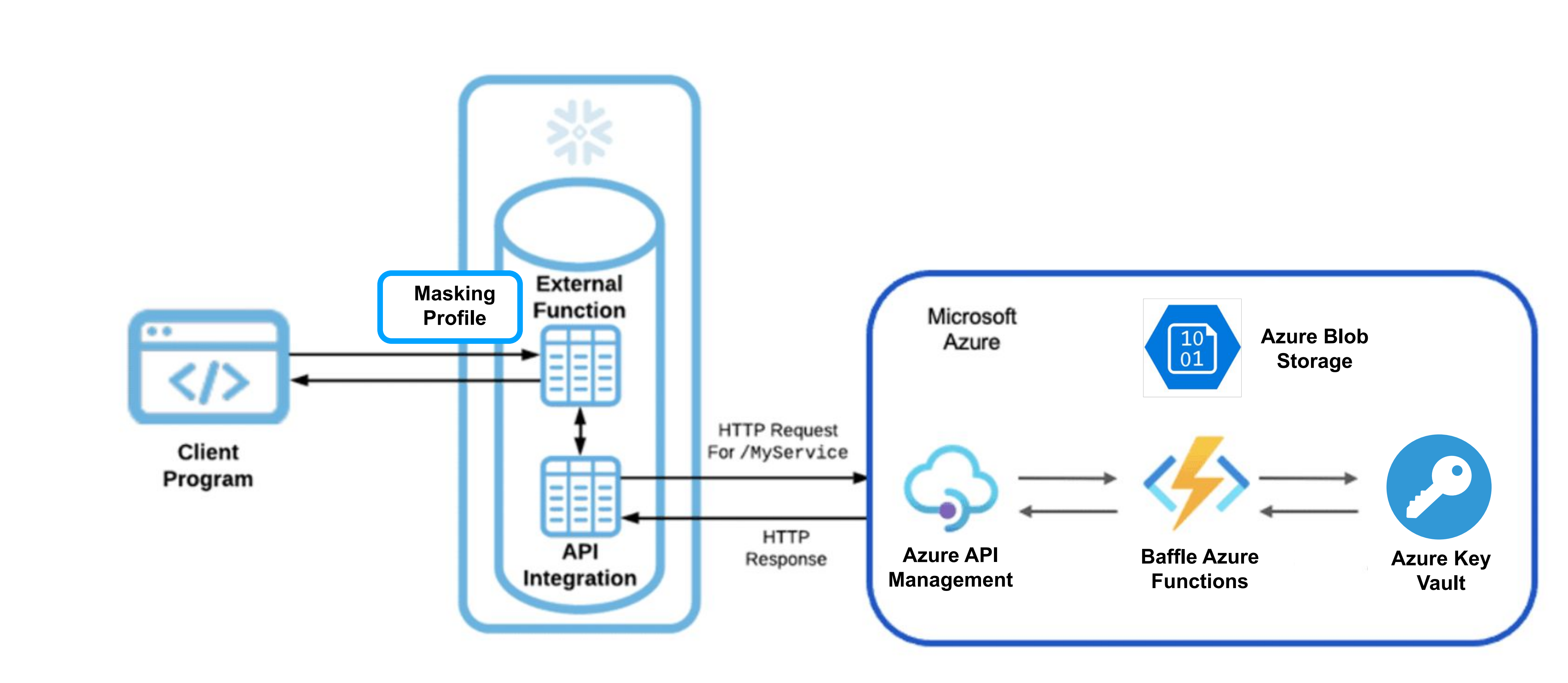
AWS Redshift

Name	Date	Transaction	Credit Card
¾¾¾¾%^&#@	2/1/13	¾¾¾¾rjkkjkjkj23	¾¾¾¾r□ÕS:VPyÉ
¾¾¾¾MN<*&	2/13/13	¾¾¾¾,.<>/;{}\dd	¾¾¾¾r"+x□,êS6
¾¾¾¾YUt45^#	2/14/13	¾¾¾¾&*^%$#@	¾¾¾¾r"+x□x□@#$r
¾¾¾¾*&^ty72	2/18/13	¾¾¾¾UIU^&$#9	¾¾¾¾r"+x□(527x□
¾¾¾¾jkj789d	1/10/13	¾¾¾¾*7%4lk;8	¾¾¾¾r"+*702jmkib

Name	Date	Transaction	Credit Card
¾¾¾¾%^&#@	2/1/13	¾¾¾¾rjkkjkjkj23	¾¾¾¾r□ÕS:VPyÉ
¾¾¾¾MN<*&	2/13/13	¾¾¾¾,.<>/;{}\dd	¾¾¾¾r"+x□,êS6
¾¾¾¾YUt45^#	2/14/13	¾¾¾¾&*^%$#@	¾¾¾¾r"+x□x□@#$r
¾¾¾¾*&^ty72	2/18/13	¾¾¾¾UIU^&$#9	¾¾¾¾r"+x□(527x□
¾¾¾¾jkj789d	1/10/13	¾¾¾¾*7%4lk;8	¾¾¾¾r"+*702jmkib

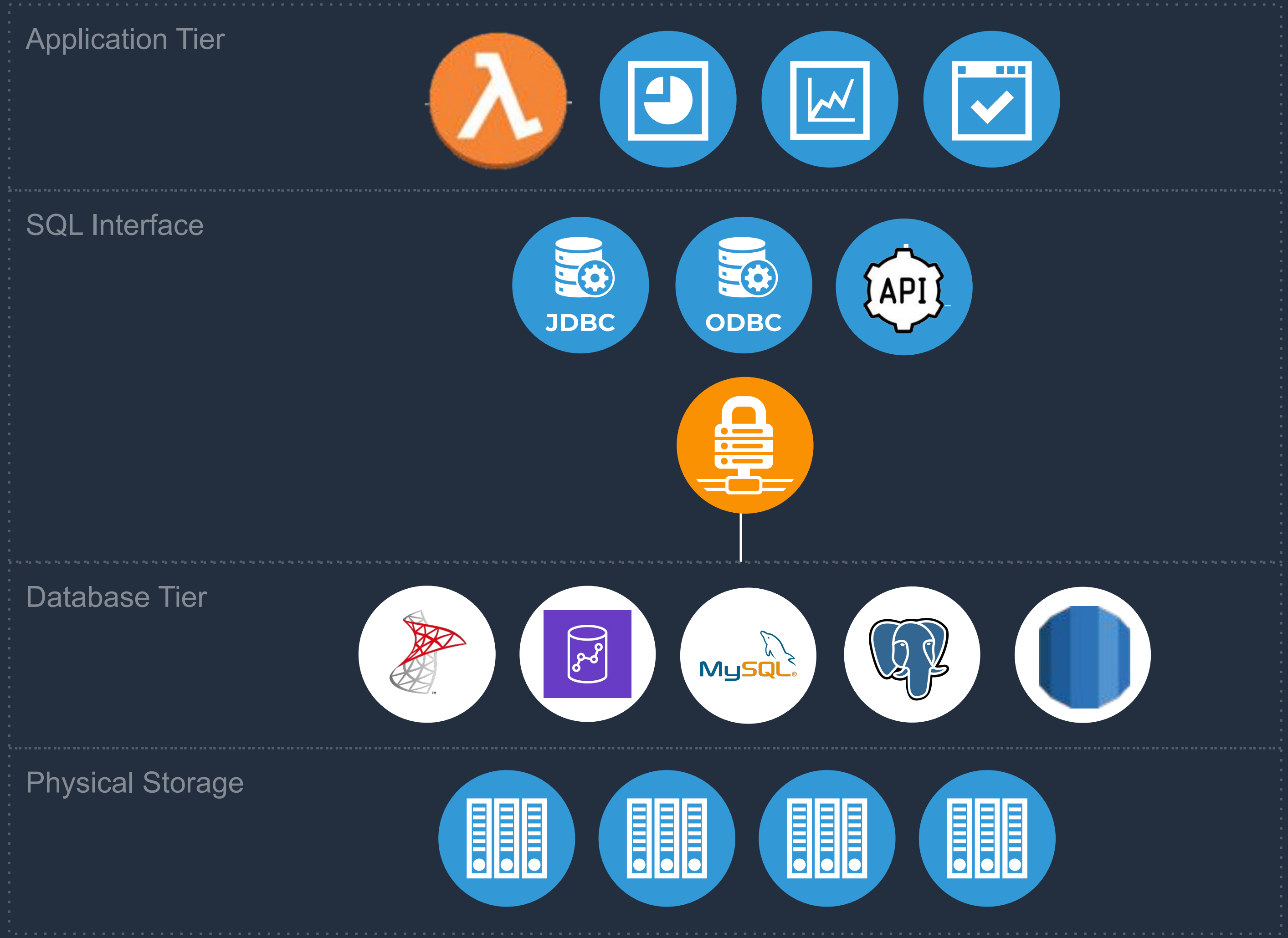
Encrypted Data

Baffle / Snowflake Integration



Baffle’s Data Protection Service Architecture

Make data breaches irrelevant



Baffle Manager

- Cloud-based management console for all data encryption and key management across the enterprise
- Comprehensive compliance and audit reporting
- Provides protection for applications, business intelligence tools, containers and serverless code

Baffle Shield

- Restricts access and decryption to calling application
- Enables data access monitoring to track anomalies
- No changes to the application required
- Supports a variety of databases including Amazon RDS

Baffle Secure Multiparty Compute (SMPC)

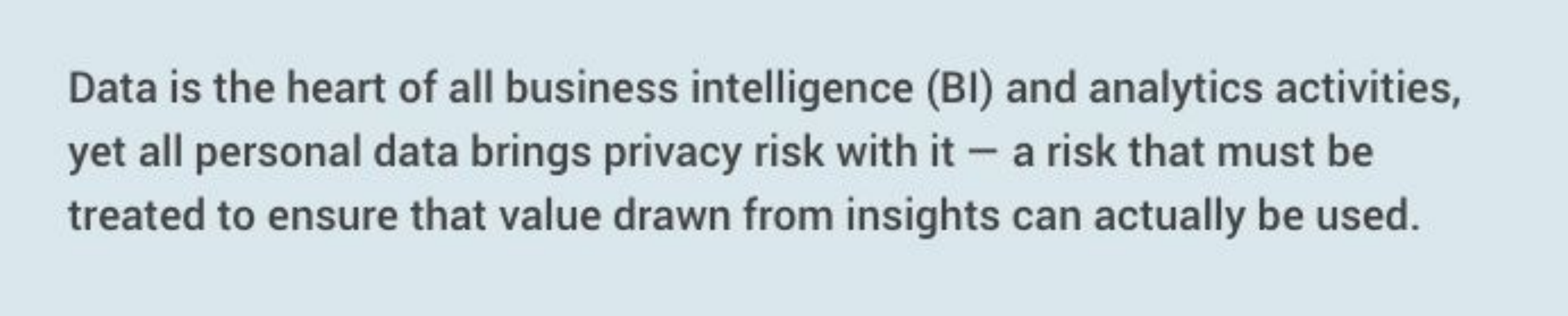
- Delivered as a software solution that automates the encryption process for any application on any database
- Dynamic access control
- Comprehensive compliance monitoring
- Requires that user defined functions (UDFs) are deployed

A Glimpse Into Privacy Preserving Analytics

Privacy Preserving Analytics

What is it?

- A computational method that allows for operations, processing and analysis of data without revealing the underlying data values or violating the data privacy contract.



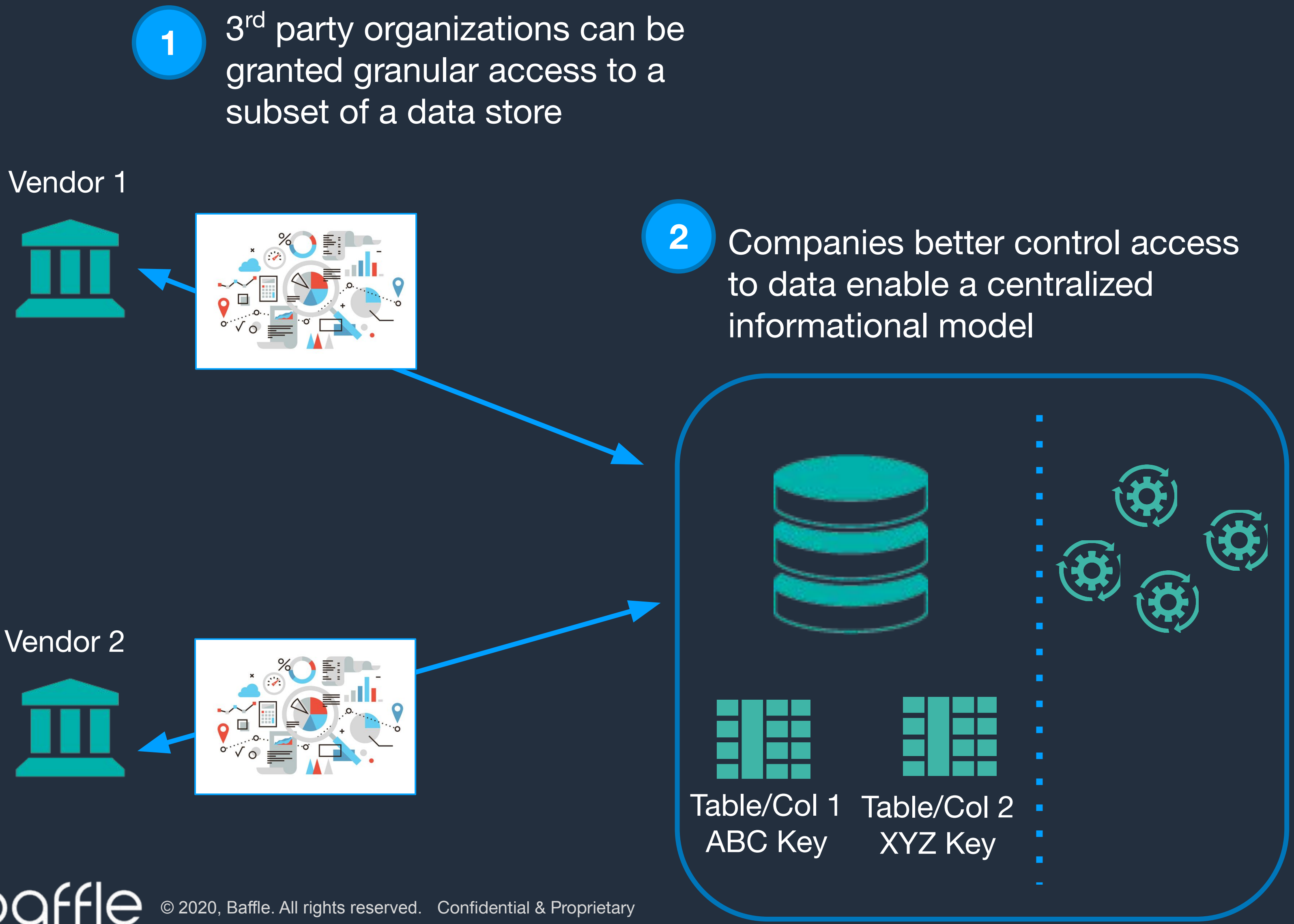
Data is the heart of all business intelligence (BI) and analytics activities, yet all personal data brings privacy risk with it — a risk that must be treated to ensure that value drawn from insights can actually be used.

Gartner Report on Privacy Preservation in Analytics

More info and resources: **https://baffle.io/privacy**

USE CASE

Data as a Service - 3rd Party Data Access Control

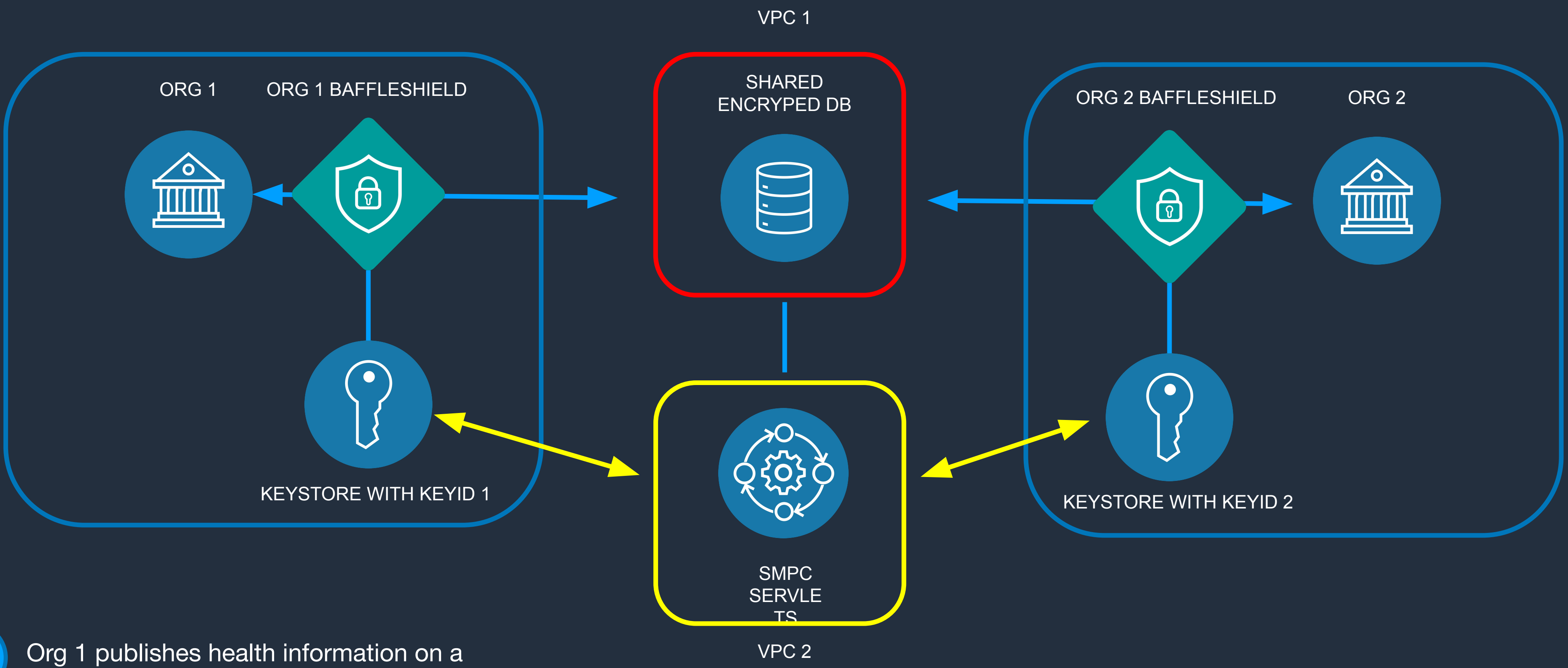

1
3rd party organizations can be granted granular access to a subset of a data store
Vendor 1
2
Companies better control access to data enable a centralized informational model
Vendor 2
Table/Col 1
ABC Key
Table/Col 2
XYZ Key

Key Benefits

- Organizations can control and minimize data sharing via a centralized data model
- Rather than spend time vetting 3rd parties via questionnaires and then giving the your data, allow them to securely integrate into your centralized data management structure
- Achieve the benefits of sourcing specific operations, without compromising your security posture

USE CASE

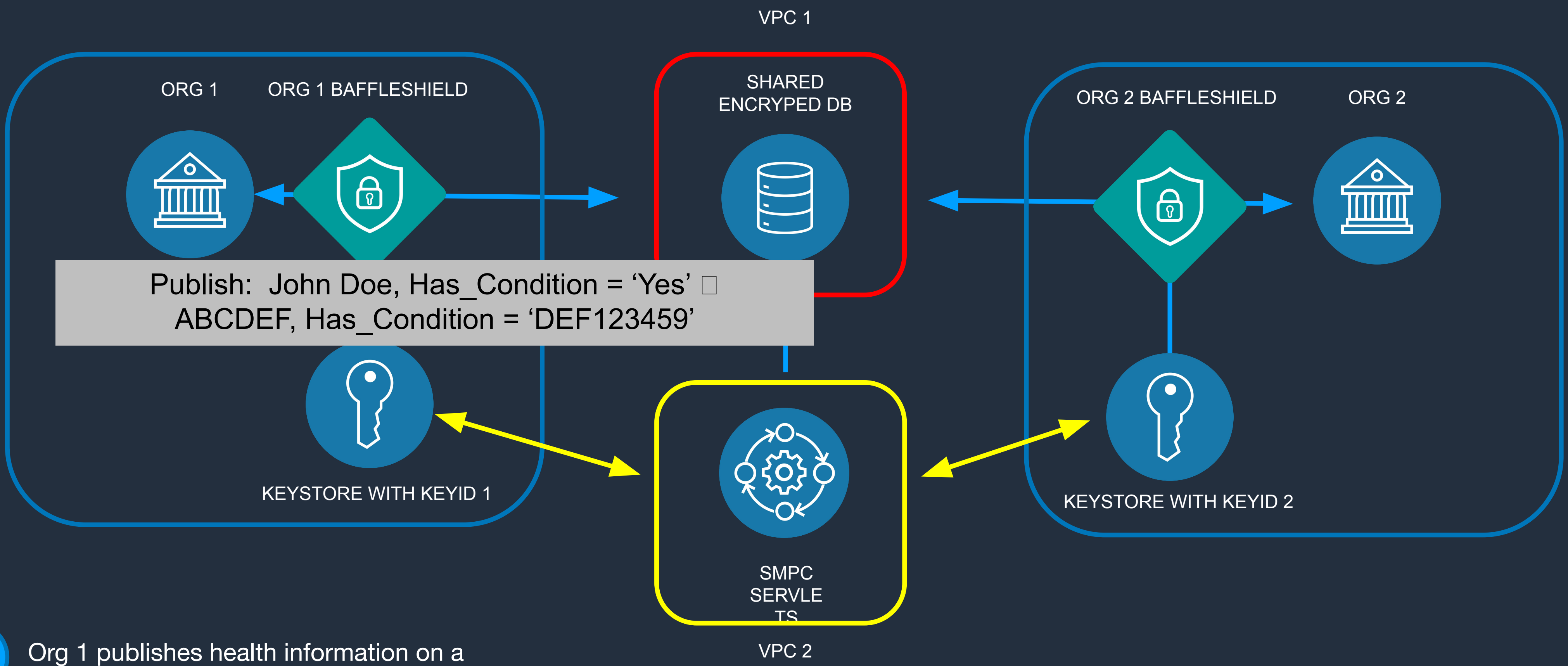
Healthcare Data Sharing



1 Org 1 publishes health information on a patient to a shared database encrypting the patient data with their own encryption key.

USE CASE

Healthcare Data Sharing

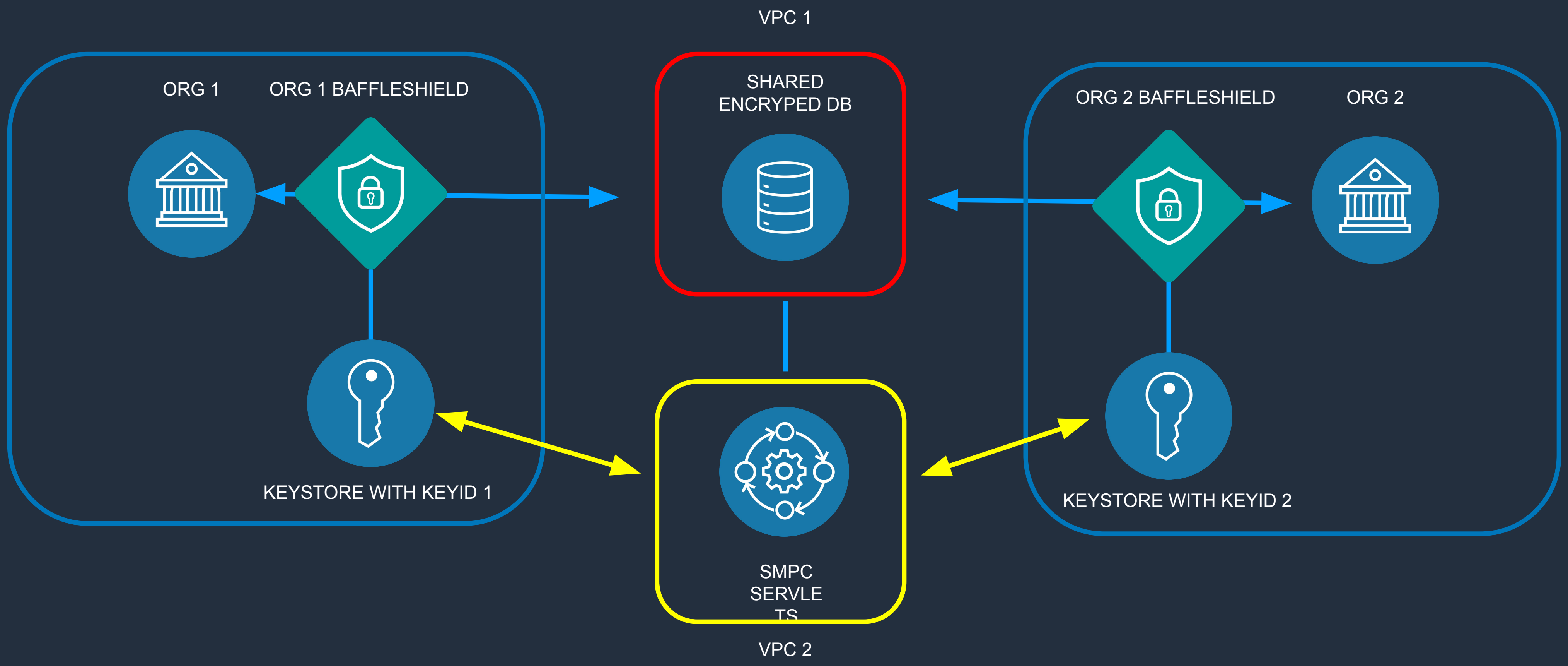


Publish: John Doe, Has_Condition = ‘Yes’ □
ABCDEF, Has_Condition = ‘DEF123459’

1 Org 1 publishes health information on a patient to a shared database encrypting the patient data with their own encryption key.

USE CASE

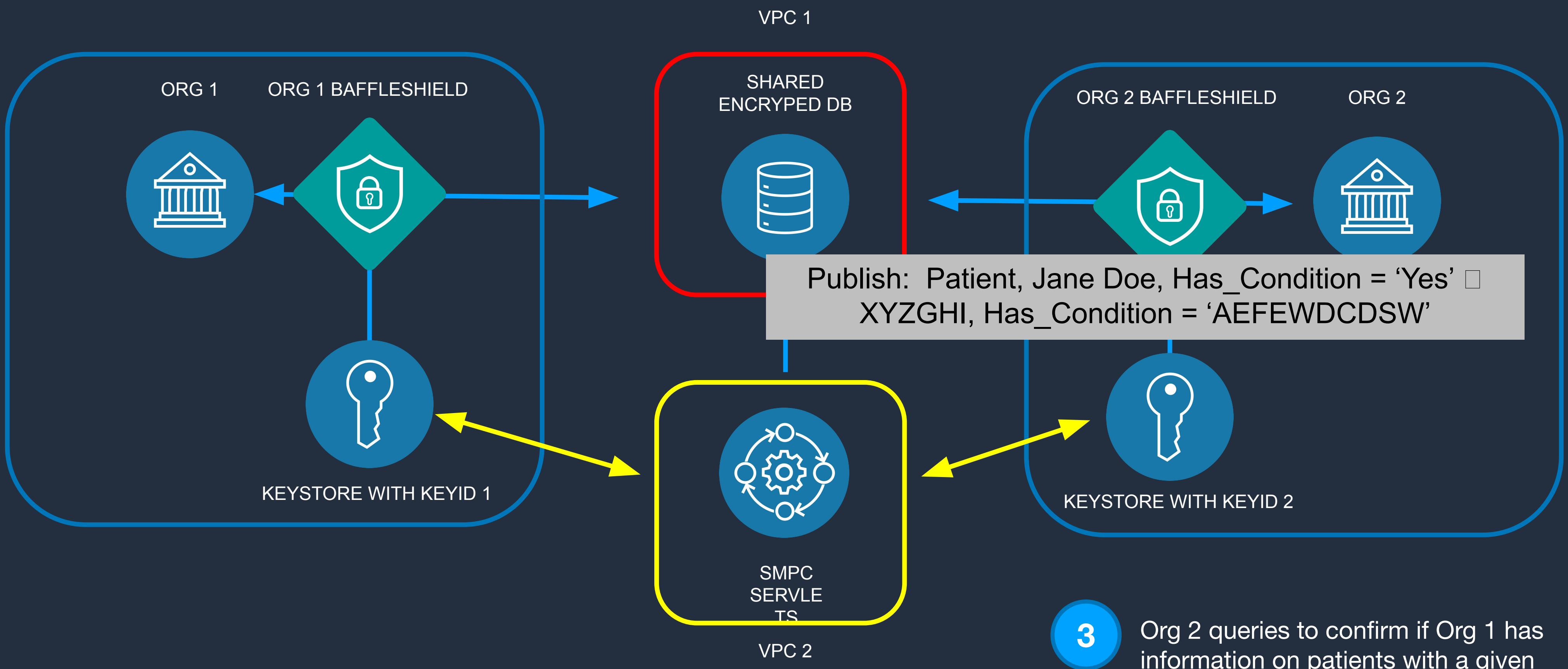
Healthcare Data Sharing



2 There are no encryption keys present in the shared database and no access to keys.

USE CASE

Healthcare Data Sharing

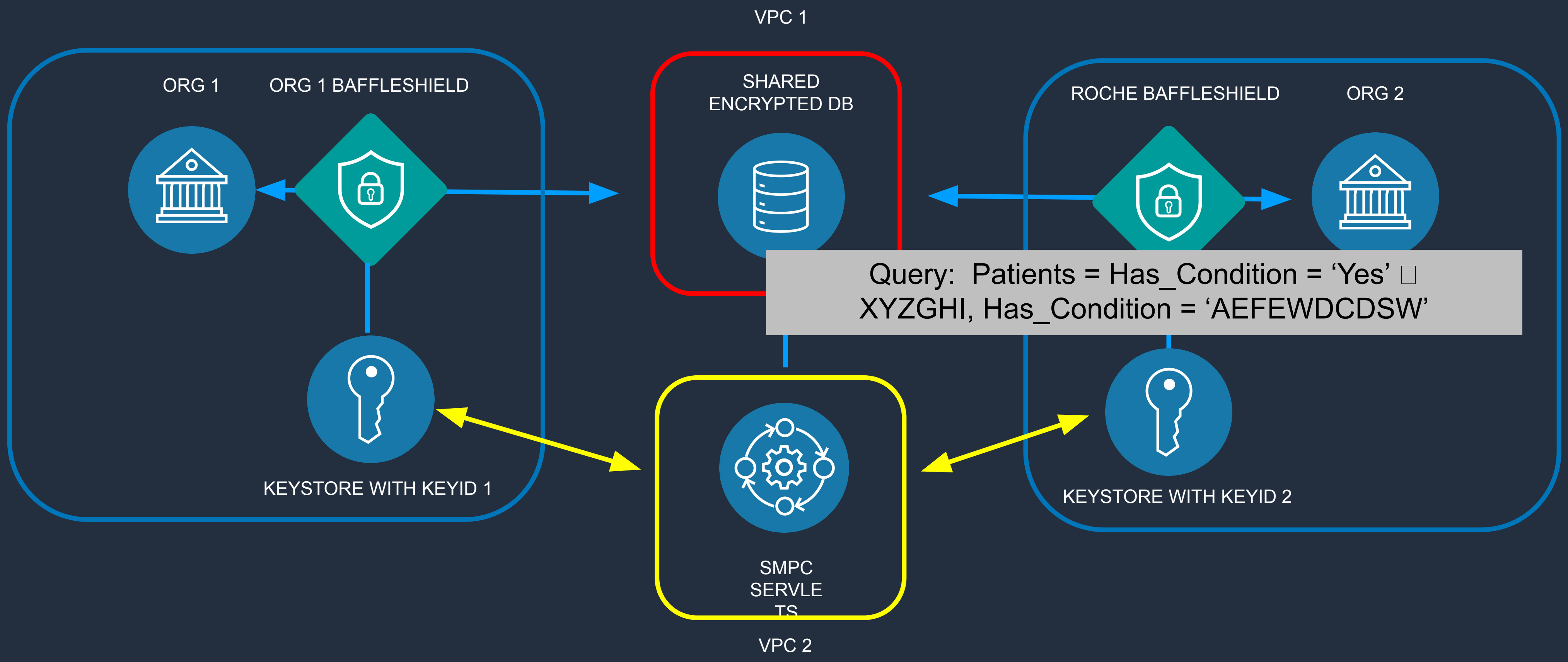


Publish: Patient, Jane Doe, Has_Condition = ‘Yes’ □ XYZGHI, Has_Condition = ‘AEFEWDCDSW’

3 Org 2 queries to confirm if Org 1 has information on patients with a given condition. The patient PHI is encrypted using Org 2’s encryption key.

USE CASE

Healthcare Data Sharing



Query: Patients = Has_Condition = ‘Yes’
XYZGHI, Has_Condition = ‘AEFEWDCDSW’

4 SMPC performs a comparison operation on using different keys without ever accessing the encrypted data values. The results are returned without decrypting the data.

Summary

- Leverage cloud data lakes to enable flexibility and accommodate data growth easily
- Implement data-centric protection methods to reduce the risk of data leakage
- Leverage de-identification capabilities to accelerate analytics and data monetization efforts that still comply with data privacy regulations
- Examine operational models that minimize impact to Devops and business data flows

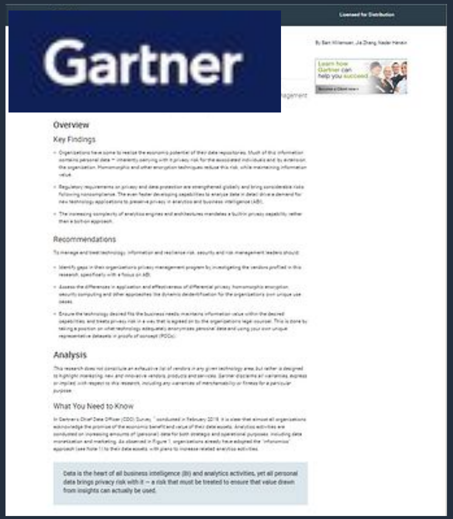
Data Privacy Resources

Simplifying Encryption White Paper

Introduction

High Level Architecture for Threat Intel Sharing

Gartner Report on Privacy Preserving Analytics



Licensed for Distribution

Gartner

Overview

Key Findings

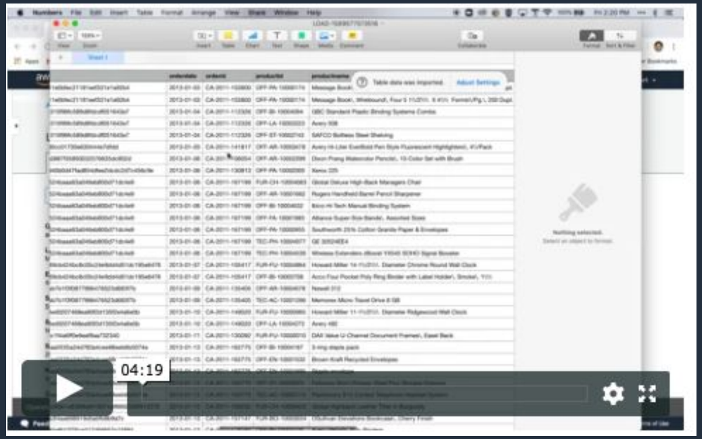
Recommendations

Analysis

What You Need to Know

Data is the heart of all business intelligence (BI) and analytics activities, yet all personal data brings privacy risk with it — a risk that must be treated to ensure that value drawn from insights can actually be used.

Video Talks and 1:1 Technical Consultation



Q & A

Thank You!

harold@baffle.io