



May 24-25, 2018
Bangalore, India

STORAGE DEVELOPER
CONFERENCE

Amalgamation of Cognitive Computing inside Object Storage for Security Compliance

Smita Raut
IBM

Acknowledgement:
Sandeep Patil, IBM

Agenda

- ❑ What is GDPR
 - ❑ GDPR for unstructured data
 - ❑ Challenge in GDPR
- ❑ Cognitive services to identify personal information
- ❑ Overview of Swift Object storage
 - ❑ Significance of object metadata
 - ❑ What is a Swift middleware
- ❑ Approaches for GDPR compliance solution
- ❑ Conclusion

What is GDPR

European Union (EU) General Data Protection Regulation (GDPR) compliance involves **personal data** and its **protection** (article 4, section 1) by any organization that conducts business with personal data of data subjects, in or from the 28 EU member states.

GDPR Compliance Requirements

As derived from- <https://gdpr-info.eu/>

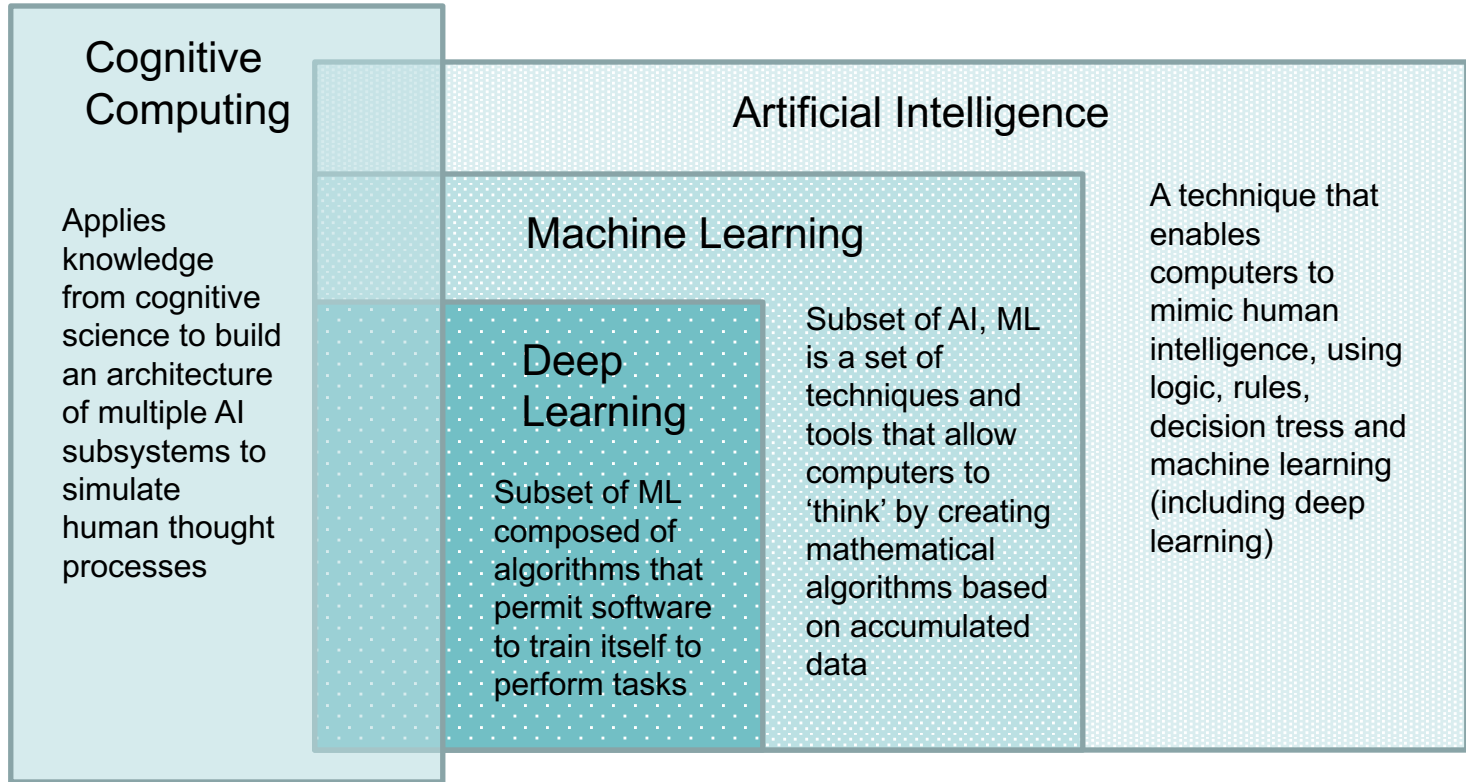
- ❑ **GDPR (Article 4, Section 1)**
 - ❑ GDPR compliance centers around **Personal Data**
 - ❑ Personal data resides in the form of either structured data (like databases) or unstructured data (like files, text, documents, etc.).
 - ❑ Personal data subject to GDPR is commonly stored in an unstructured data format. **Object** is an emerging form of unstructured data and hence this is directly applicable to **Object Storage**
- ❑ **GDPR Requirement Article 32 (Secure personal data)**
 - ❑ Securing personal data of EU residents is one of the key requirements of GDPR. One way to accomplish this is by using **data encryption**
- ❑ **GDPR article 17 (Right to erasure and to be forgotten)**
 - ❑ GDPR requires businesses to address right to erasure of data categorized as personal.
- ❑ **GDPR Requirement Article 15 (Right of Access)**
 - ❑ Need to control and audit access to data categorized as personal data through mechanisms such as secure authentication, authorization, and audit logging.

Challenge in GDPR for Unstructured Data

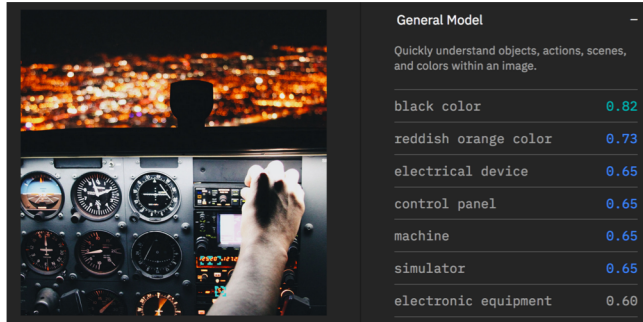
- ❑ Unstructured data can be text, documents, images etc.
- ❑ Personal information needs to be identified from this unstructured data so that GDPR regulations can be applied on it
- ❑ Autonomous identification of personal information is a challenge

Overcoming The Challenge...

Introduction to Cognitive Computing



Cognitive Services Examples

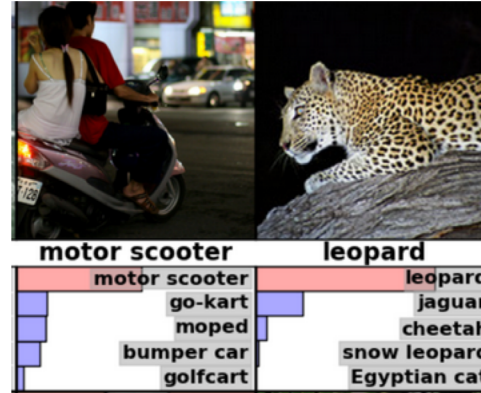


IBM Watson's
Visual
Recognition
&
Natural
Language
Understanding

Sentiment Emotion Keywords **Entities**

Extract people, companies, organizations, cities, geographic features, and

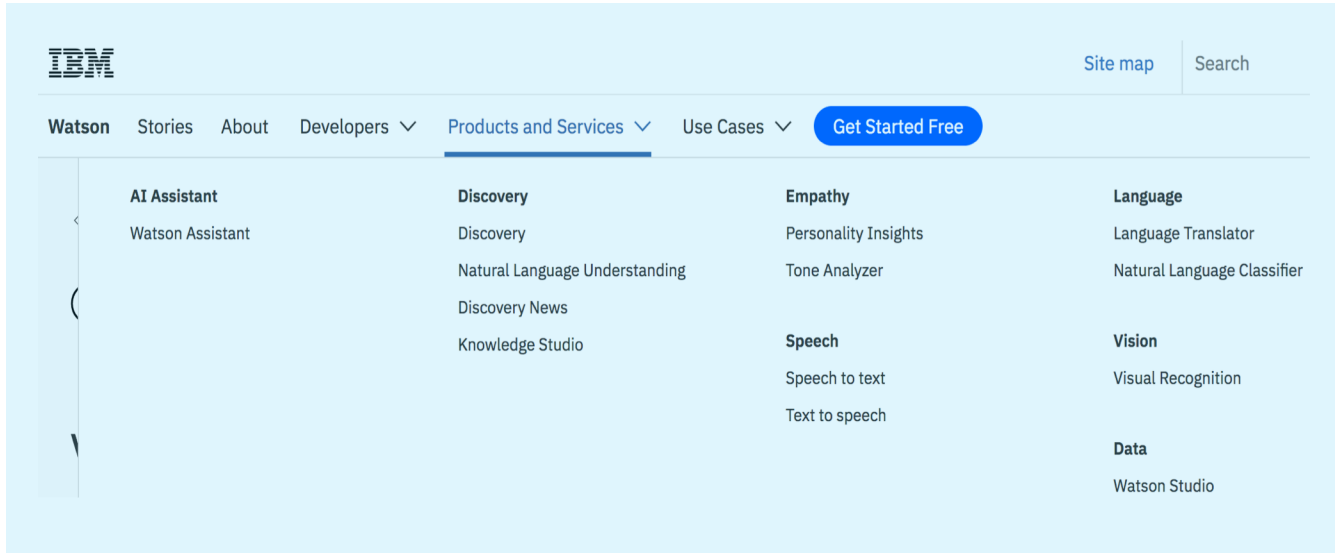
Name	Type	Score
Anza-Borrego Desert	GeographicFeature	0.84
Myrtle Botts	Person	0.83
Colorado Desert	GeographicFeature	0.58
Albert S. Evans	Person	0.51



TensorFlow's
Image Recognition
&
Text Classification

	sentence	sentiment
0	Next to "Star Wars" and "The Wizard of Oz," th...	10 1
1	I can't help but laugh at the people who prais...	1 0
2	Based on a true story, this series is a gem wi...	10 1
3	Van Dien must cringe with embarrassment at the...	1 0
4	This film had such promise!! What a great idea...	4 0

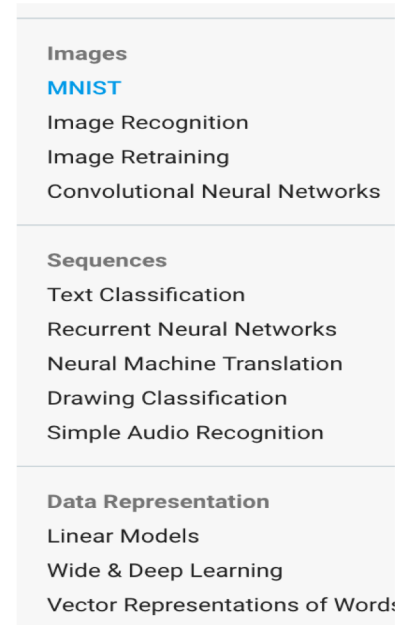
Cognitive Services Examples (cont...)



The screenshot shows the IBM Watson website. The top navigation bar includes the IBM logo, links for 'Watson', 'Stories', 'About', 'Developers', 'Products and Services' (which is expanded), 'Use Cases', and a 'Get Started Free' button. The 'Products and Services' dropdown menu is organized into four columns: AI Assistant (with 'Watson Assistant'), Discovery (with 'Discovery', 'Natural Language Understanding', 'Discovery News', and 'Knowledge Studio'), Empathy (with 'Personality Insights' and 'Tone Analyzer'), and Language (with 'Language Translator' and 'Natural Language Classifier'). Below these, there are sections for 'Speech' (with 'Speech to text' and 'Text to speech'), 'Vision' (with 'Visual Recognition'), and 'Data' (with 'Watson Studio').



The TensorFlow logo is displayed in white on an orange background. Below the logo are two orange buttons: 'GET STARTED' and 'PROGRAM'.



A list of TensorFlow examples categorized by type:

- Images**
 - [MNIST](#)
 - Image Recognition
 - Image Retraining
 - Convolutional Neural Networks
- Sequences**
 - Text Classification
 - Recurrent Neural Networks
 - Neural Machine Translation
 - Drawing Classification
 - Simple Audio Recognition
- Data Representation**
 - Linear Models
 - Wide & Deep Learning
 - Vector Representations of Words

Microsoft Azure

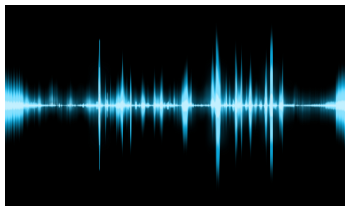
Vision APIs

Speech APIs

Language APIs

Search APIs

Identifying Personal Data Using Cognitive Computing



Speech To Text

Natural
Language
Processing

Text URL

My name is Smita. I live in Pune and work with IBM.
I am a software engineer and presenting at SNIA SDC.

Sentiment Emotion Keywords **Entities**

Extract people, companies, organizations, cities, geographic features, and

Name	Type	Score
Smita	Location	<div><div></div></div> 0.89
software engineer	JobTitle	<div><div></div></div> 0.63
Pune	Location	<div><div></div></div> 0.57
IBM	Company	<div><div></div></div> 0.46

So What Is The Problem?

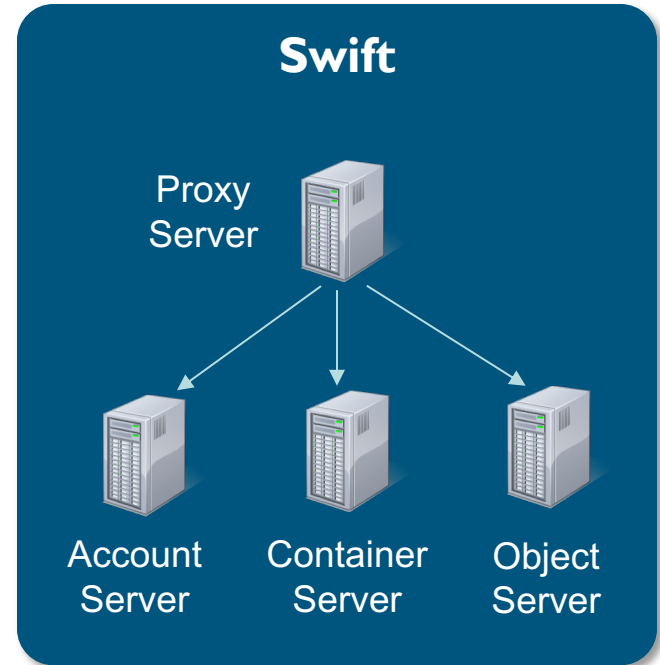
Problem Statement

- ❑ Tools and techniques to identify personal data exist, but they happen on a server while the data is on storage. We need to have identification embedded where object is.
- ❑ We need a smart object store who self identify personal data and apply GDPR rules ensuring compliance

Solution!

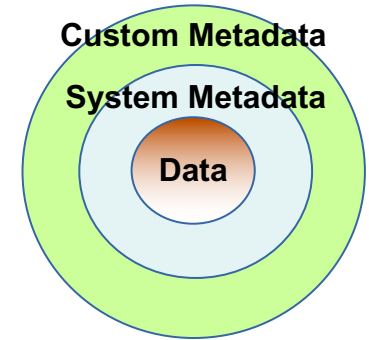
OpenStack Swift Object Storage

- ❑ OpenStack Swift cluster consist of multiple type of processes, of which object, container and account servers are backend and Proxy server acts as interface for the cluster.
- ❑ Depending on the request type i.e. Object or Container or Account, Proxy server chooses the responsible backend server which will serve depending on distributed circular hash, called as Ring.



Object Metadata

- ❑ Data is stored as individual objects with unique identifier
- ❑ Typically, Objects consist of an object identifier (OID), data and metadata
- ❑ Object data is unstructured – images, text, audio, video
- ❑ Metadata consists on system metadata and user defined custom metadata that can be extensive



Data + Metadata = Object



Object type = image

System Metadata

- Filename: taj1234.jpg
- Created: 01 Aug 2016
- Last Modified: 03 Aug 2016

Custom Metadata

- Subject: Taj Mahal
- Place taken: India
- Category: Travel
- Allow Sharing: yes

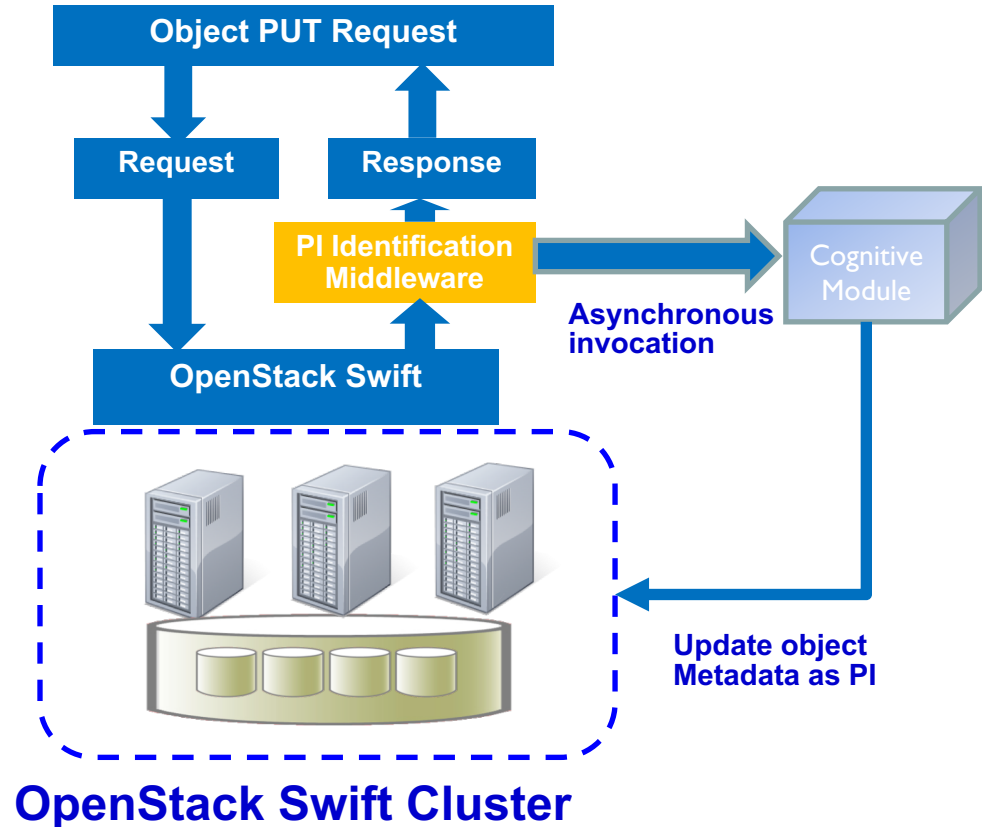
OpenStack Swift Middleware Framework

- ❑ OpenStack Swift is based on **WSGI** specifications and Middleware is a WSGI feature which extends functionality of any WSGI application.
- ❑ Middlewares are heavily used in swift, for purposes such as logging, tempurl, tempauth, quotas etc.
- ❑ If there is a middleware in an application pipeline, every request and response is passed through the middleware when a request is served.
- ❑ One can write custom middleware for OpenStack Swift cluster

Different ways to address the challenge

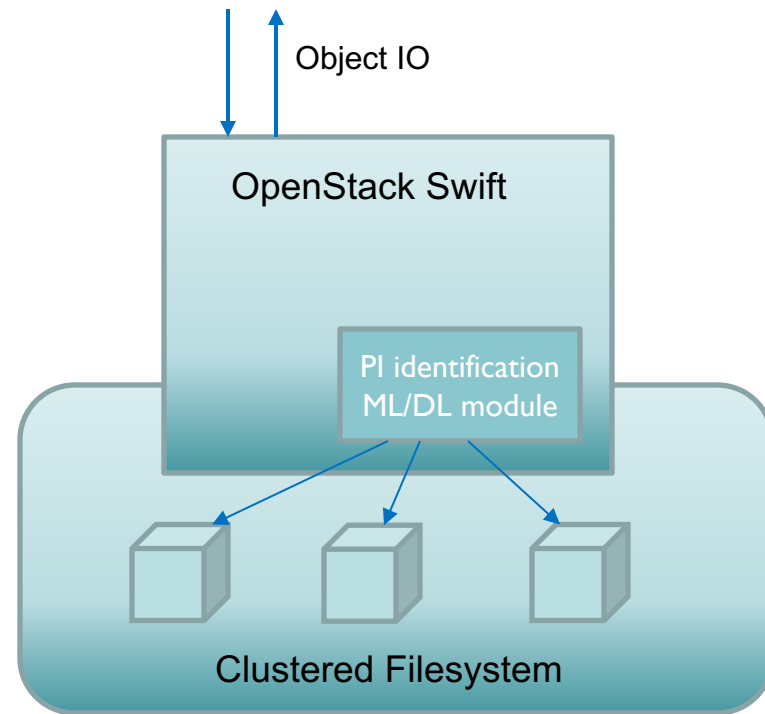
Intercepting Object PUTs

- ❑ Personal data identification middleware intercepting object PUT
- ❑ Asynchronously invokes cognitive module based on ML/DL
- ❑ Cognitive module analyzes object data and tags it as PI
- ❑ GDPR rules like encryption, geo-fencing etc. are applied on PI tagged objects



Scanning the Object Storage

- ❑ A PI identification module based on ML/DL algorithms is deployed in a swift cluster as a service/daemon
- ❑ The module scans the object storage, identifies personal data, tags the objects accordingly in background and applies GDPR rules on it
- ❑ If swift uses a clustered filesystem as its backend storage, then the PI identification module can leverage multiple nodes for parallel processing
- ❑ Object IO continues unaffected while the service is doing its work



Doing it for your data in Public Cloud

❑ Storlets

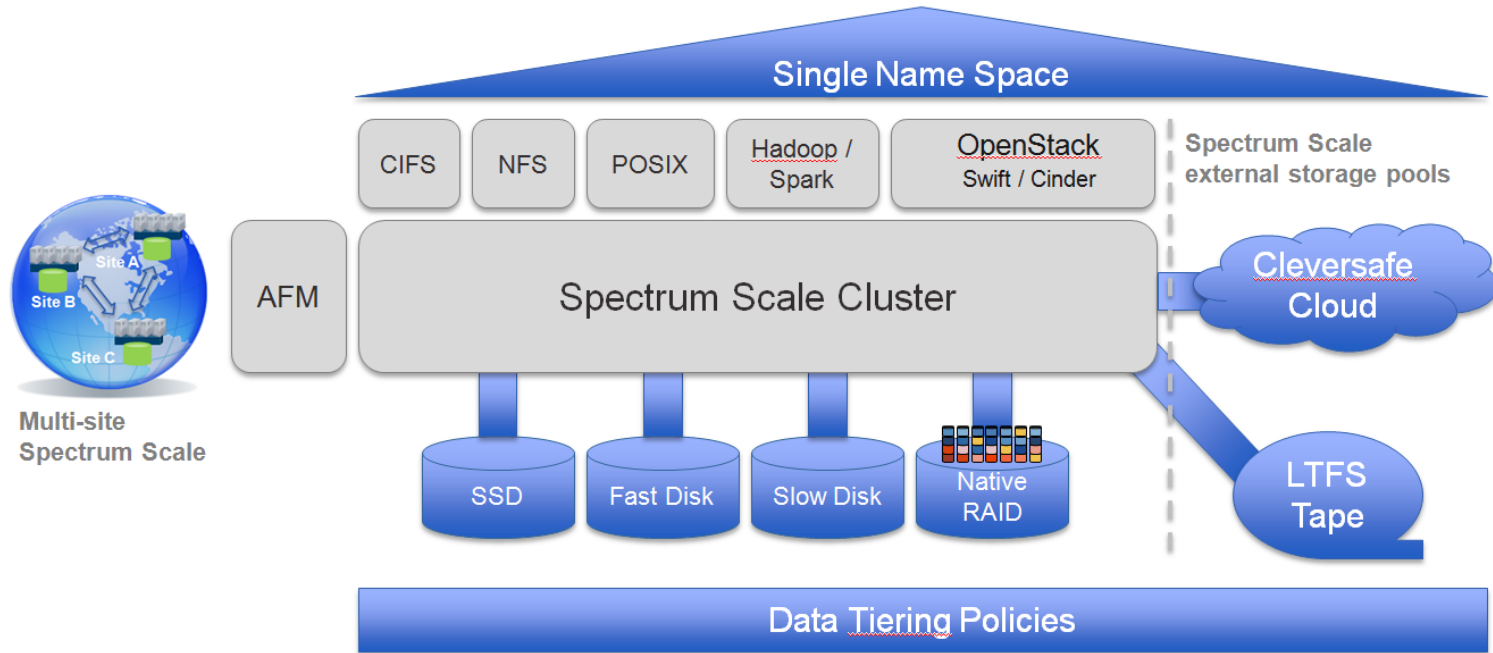
- ❑ The **Storlet Engine** extends OpenStack Swift by enabling the push down of **filtering**, **transforming** and **analysis** tasks to Swift instead of bringing the data to computation
- ❑ The executed code - called Storlet - is user defined, and is isolated from the Swift system using **Docker Linux containers**
- ❑ Storlets can be invoked on data objects during PUT, GET and COPY



Conclusion

- ❑ Identification of PII took place at the storage level
- ❑ Appropriate measure were taken to follow GDPR laws such as encryption, geo-fencing, etc.

IBM Spectrum Scale – the Data Management Solution



Thank You!