



Mastering Unstructured Data – Mitigating AI Risks in Organizational Files



Juha Sallinen / GDPR Tech

GDPR Tech

GDPR Tech, founded 12/2016

- Dedicated team with strong knowledge of practical data privacy work, focusing to the EU GDPR
- We focus to the Nordic and Baltic countries, but we do have global customers
- Training, consulting, projects, e-learning courses
- Data Management as a Service (DMaaS)
- Data Protection as a Service (DPaaS)

**PHOENIX
CONTACT**
INSPIRING INNOVATIONS

MPY

EvoLearn
Enabling growth & development

MOST
Digital

EVONIK
POWER TO CREATE

Holiday Club

WANNADO

FUJITSU



Unstructured data management challenges

Files, spreadsheets, documents, emails, listings from databases and apps. Data, which is not defined to a defined model or architecture



About 80% of enterprise data (various sources, IDC, Gartner)



People manage, search and protect (Fileserver, Teams, SharePoint, Box, Slack, WhatsApp, email, Zoom, CRM,...)



Data breach - City of Helsinki (capital of Finland) – how to avoid headlines in the first place?

Unstructured Data & Poor Classification

Sensitive data was stored on file servers where it shouldn't have been, and there was no clear overview of what was compromised.

Outdated Systems & Lack of Monitoring

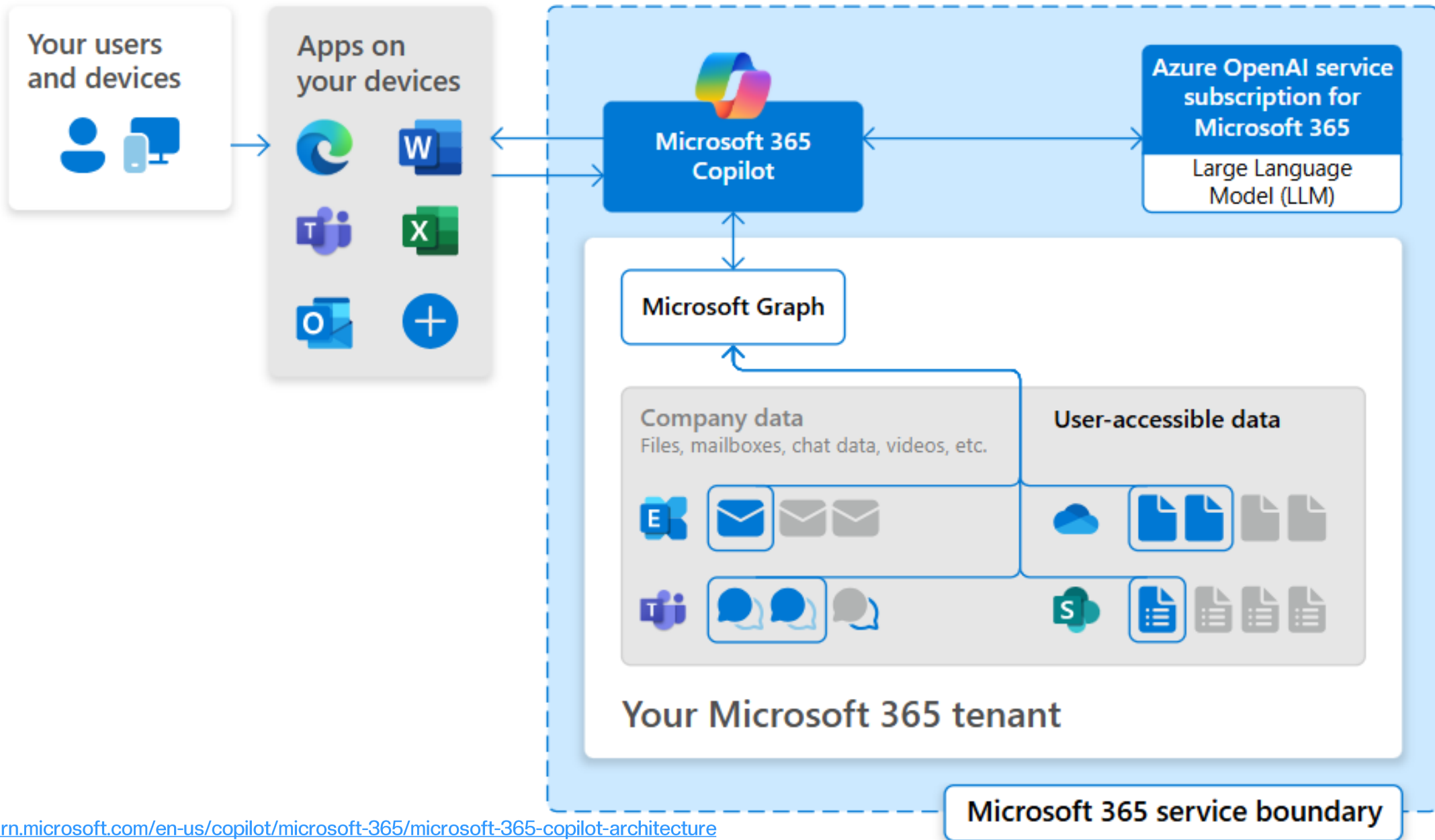
The breach occurred due to an unpatched remote access service with known vulnerabilities, and unauthorized access went undetected.

Lack of Awareness & Responsibility

Employees were unaware of proper data storage policies, and there was no clear accountability for securing sensitive information



Challenge: “AI uses the data that is accessible to the user”



As per the Microsoft guidance

Know your data – what we have

Protect the data – who has access

Data retention – how long to store



Story of an energy company

Year 2021

We run a data assessment for the company, against the file servers (data shares like H: drive or [\\server\projects](#))

CISO wanted to do this with us, but asked their legal department, if that is ok

Answer from the legal department

"There are NO files containing personal data because they have been instructed NOT to store any"

Worst access management ever

And on top of that, files containing SSN (social security numbers) openly accessible

Thousands of such documents, mostly open access to all users



"GDPR-HR"

"PII","GDPR-HETU","G...

"Luottamuk...

"Hetu-50"

"PII","G...

Tagname "Hetu-50"
Count of Absname 4271

Thousands of files containing 50 or more SSN

"IT"

"PII","GDPR-HR"

"PII","GDPR-...

"GDPR-HETU"

"Special-Category"

"PII","GDPR-HETU..."

"PII","GT-TH"

"PII","Luotta...

file \ Yhteinen \ Siirrot \ Frends \ SKS \ Testi

Overview User Activity Folder Activity Permissions Audit Logs

Effective Rights File System Access Control List Share level permissions Recommendations

User/UserGroup	Permission							Inherited From	Applies To
	FULL	RE...	WR...	MODI...	READ & EX...	LI...	SPECIAL		
Aki								Frends	This folder, Subfolders, files
Prod_Integration_Services...								Siirrot	This folder, Subfolders, files
Test_Integration_Services...								Siirrot	This folder, Subfolders, files
sql2014service								Siirrot	This folder, Subfolders, files
latuki@...								Siirrot	This folder, Subfolders, files
Yhteinen_denied@...								\	This folder, Subfolders, files
S-1-5-21-10...								\	This folder, Subfolders, files
Domain Users@H...								\	This folder, Subfolders, files
Ulkopuoliset@H...								\	This folder, Subfolders, files
Authenticated Users								\	This folder, Subfolders, files
Local System								\	This folder, Subfolders, files
BizTalkService								\	This folder, Subfolders, files
Domain Admins@...								\	This folder, Subfolders, files
Administrators								\	This folder, Subfolders, files

Story of a multinational accounting company

Year 2023

Pilot for the sites and data in Finland. Content from the file servers and Cloud was in the scope.

Later, in year 2024, we ran data cleaning, removing data based on defined rules

Only about 11 terabytes

On average, 1 terabyte is 1 million files.

But over 75 million files, which is not common



Customer data under product site in the cloud

https://xyz_group.sharepoint.com/sites/HRBusinessUnitsProducts/productName/Brochures/CustomerProjects/Customer_123/Product Name/2016_12_12_ProductName_employee_record_changes.xlsx
Large number of SSN stored in the wrong place



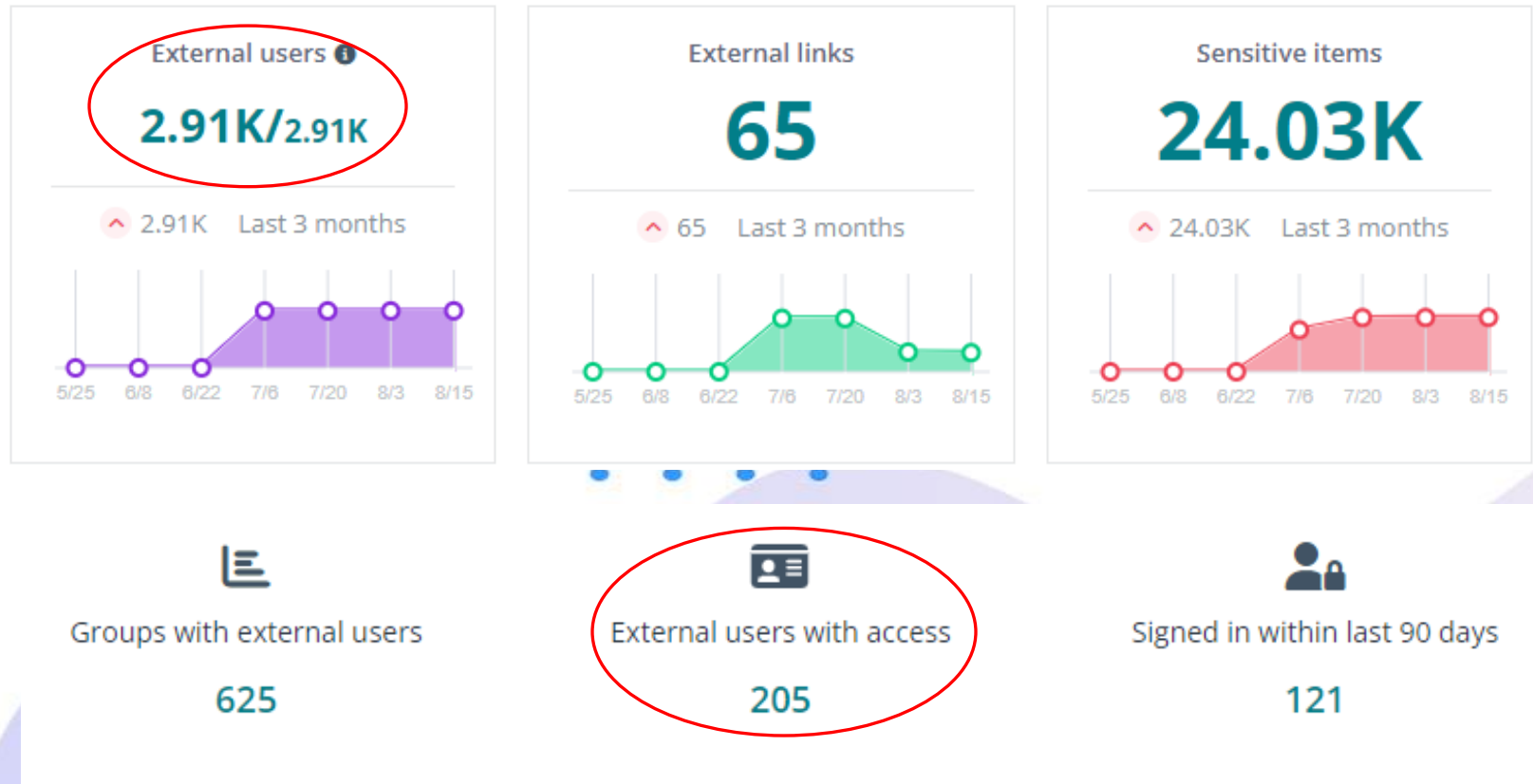
GDPR Tech

Story of a city and the education environment

City in Finland (top 10),
fileservers and cloud
data – 2023 - 2024

Data assessment of the
city data sources and later
education tenant

Education contained CV
and evaluation documents
with SSN and shared to
others



**“AI uses the data that is accessible to the user”
Who should define who has access?**

Who should define where to store data

Who should define how long to keep data?

Fix this!

Know your data – what you have and where it really should be (app, filer, cloud?). Make a playbook and share that. Test that users read guidelines.

Protect the data – who actually has access and who should (domain users, everyone and guest groups?). Find and remove deviations!

Manage data – what is the guidance how long to store and how do you implement that (forever or 10 years?). Monitor real life!

Don't be the next data breach organisation, with or without AI





In GDPR, the use of common sense is allowed

www.gdprtech.com

Data governance and privacy work from 2016

Is AI using my data for training?

Feature	Microsoft Copilot	Google Gemini
Base AI Model	Universal (GPT-4, proprietary models)	Universal (Google Gemini AI)
Trains on Org Data?	✗ No	✗ No
Uses Real-Time Org Data?	✓ Yes, via Microsoft Graph	✓ Yes, via Google Workspace
Data Security Policies	Uses Microsoft Purview, follows org security settings	Uses Google Workspace security settings
Admin Controls?	✓ Yes (can restrict access to sources)	✓ Yes (admins control data access)

Microsoft Graph is a secure connector that links all Microsoft 365 apps (like Outlook, Teams, OneDrive, and SharePoint), allowing them to share and retrieve data based on user permissions.

