Future of Storage

Spring with IBM Storage
Zurich, April 3rd 2025

Dr. Robert Haas

Department Head, Hybrid Cloud Research,
IBM Research – Zurich

Senior Technical Staff Member

rha@zurich.ibm.com

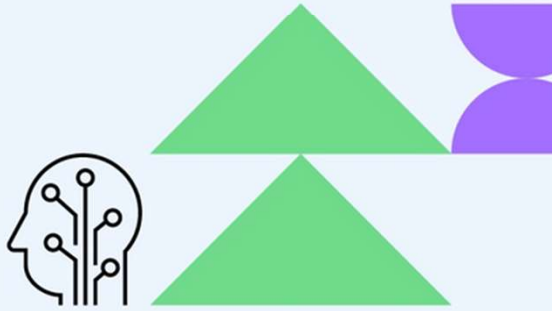
IBM **Research**

Spring with IBM Storage : Cyber Resiliency, data protection, hybrid cloud for VM's with AI and Openshift, and future technology from IBM Research.

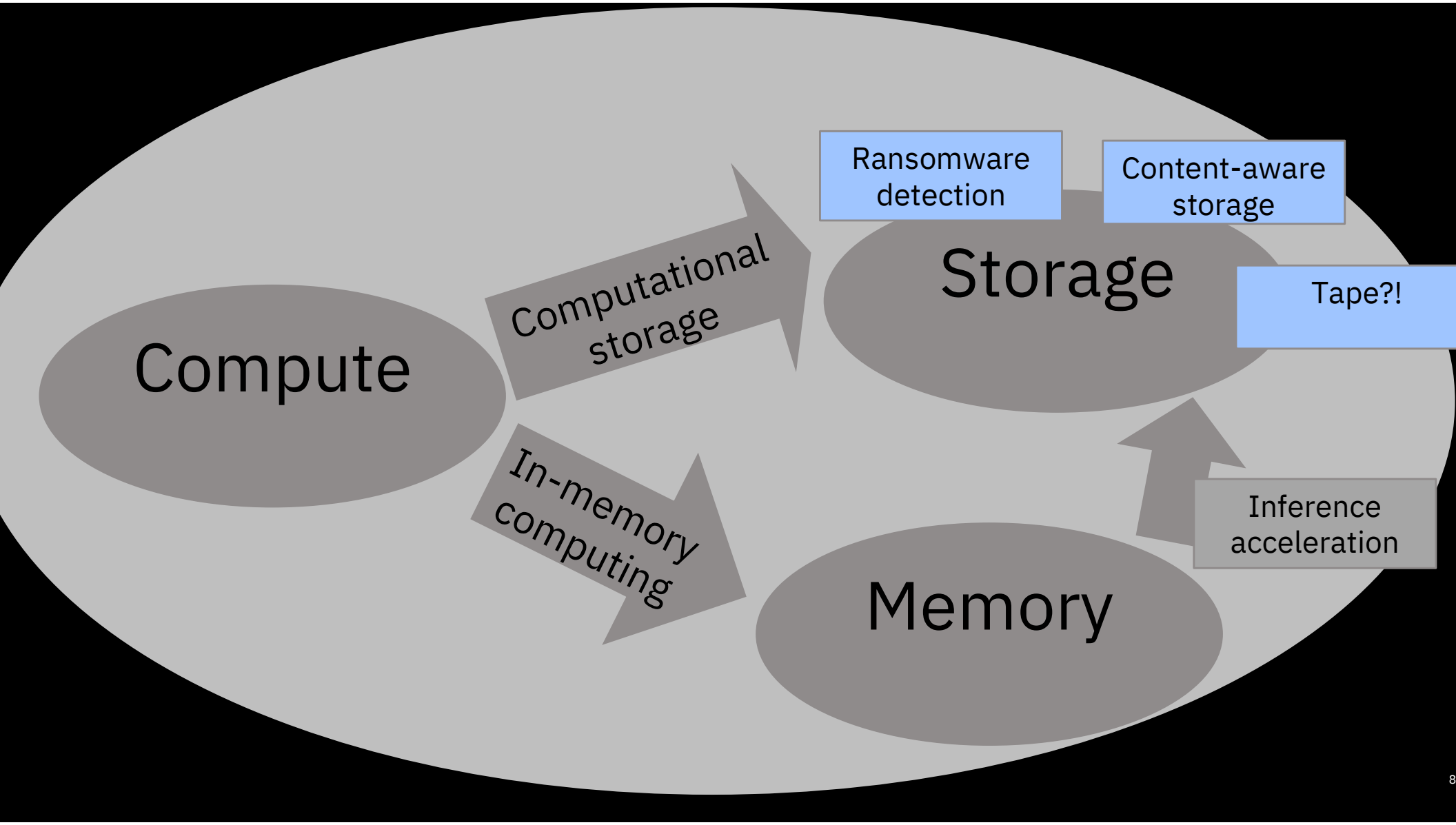
April 03, 2025
13:30 – 17:30

IBM Switzerland HQ Zurich
Vulkanstr. 106, ZH, 8048
Zurich

Register →



If you have any questions regarding the event, please send an e-mail to
jmas@ch.ibm.com


Compute
Computational storage
In-memory computing
Storage
Memory
Ransomware detection
Content-aware storage
Tape?!
Inference acceleration


Wana Decrypt0r 2.0
Ooops, your files have been encrypted!
English
What Happened to My Computer?
Your important files are encrypted.
Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.
Can I Recover My Files?
Sure. We guarantee that you can recover all your files safely and easily. But you have not so enough time.
You can decrypt some of your files for free. Try now by clicking <Decrypt>.
But if you want to decrypt all your files, you need to pay.
You only have 3 days to submit the payment. After that the price will be doubled.
Also, if you don't pay in 7 days, you won't be able to recover your files forever.
We will have free events for users who are so poor that they couldn't pay in 6 months.
How Do I Pay?
Payment is accepted in Bitcoin only. For more information, click <About bitcoin>.
Please check the current price of Bitcoin and buy some bitcoins. For more information, click <How to buy bitcoins>.
And send the correct amount to the address specified in this window.
After your payment, click <Check Payment>. Best time to check: 9:00am - 11:00am
Payment will be raised on
5/16/2017 00:47:55
Time Left
02:23:57:37
Your files will be lost on
5/20/2017 00:47:55
Time Left
06:23:57:37
About bitcoin
How to buy bitcoins?
Contact Us
bitcoin
ACCEPTED HERE
Send $300 worth of bitcoin to this address:
12t9YDPgwueZ9NyMgw519p7AA8isjr6SMw
Copy
Check Payment
Decrypt

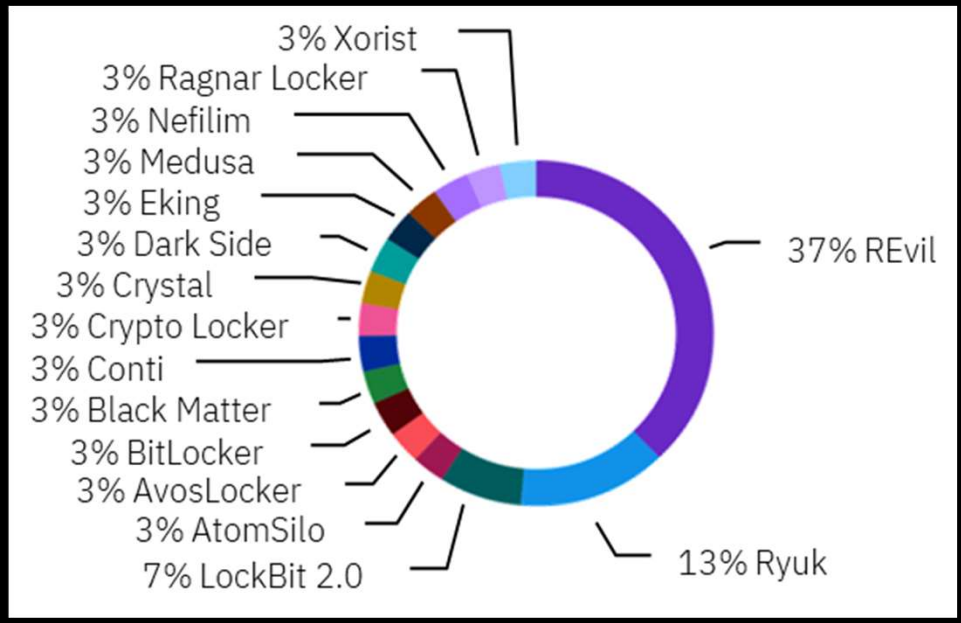
Ransomware cyber security threats

Growing number of ransomware families

Ransomware-as-a-Service (RaaS)

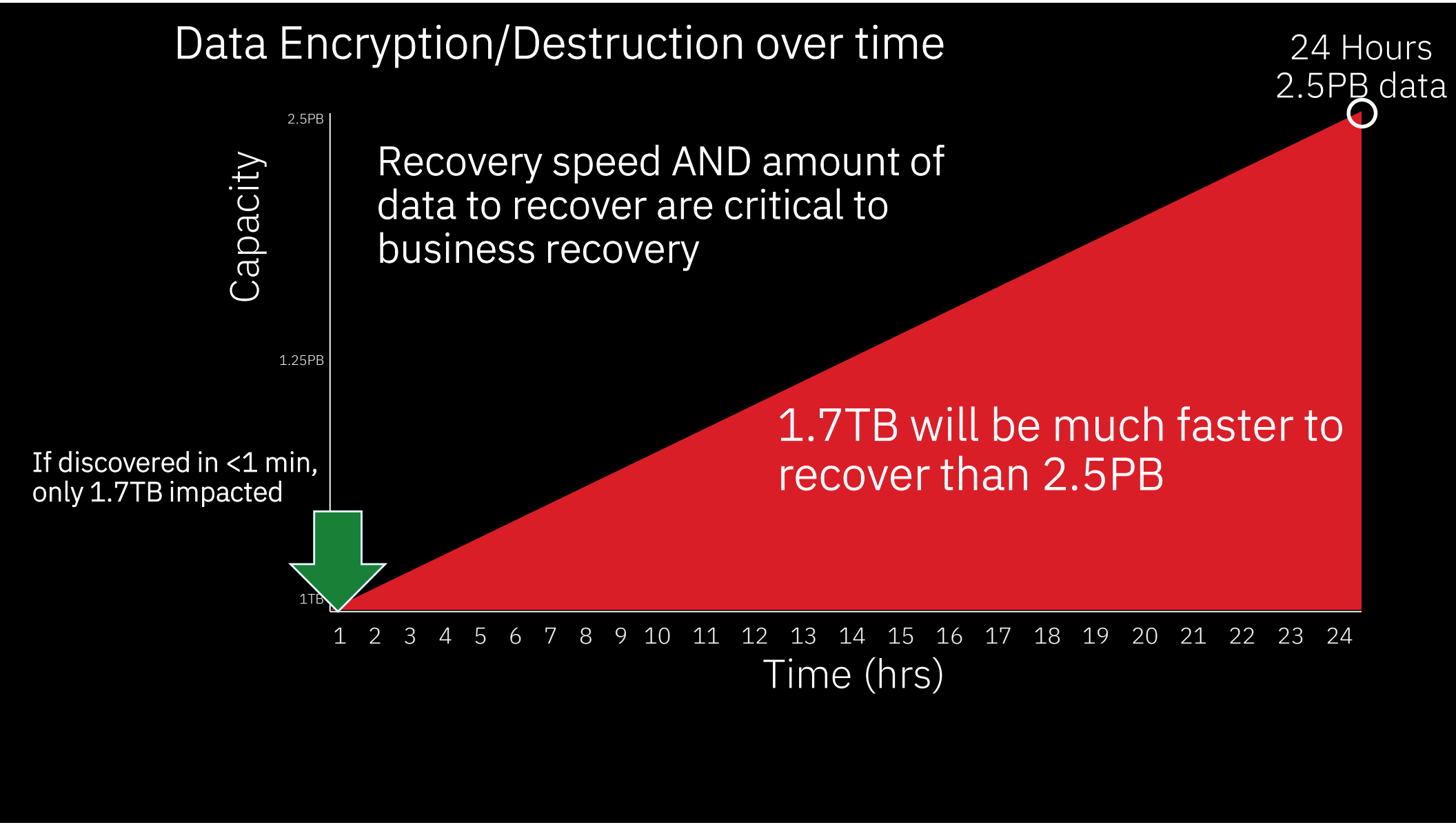
Obfuscation techniques

Attack surface increase



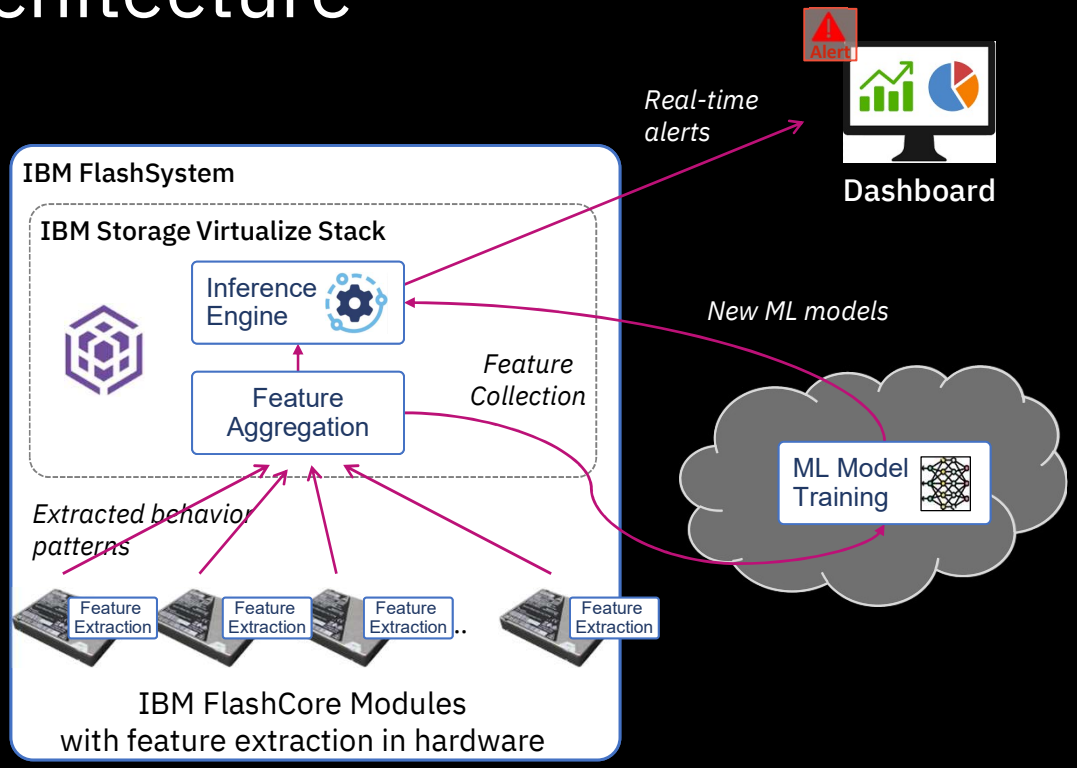
Source: IBM X-Force Threat Intelligence Index 2022

Data Encryption/Destruction over time



Ransomware detection architecture

- Feature collection on each IO operations in each FCM
- Aggregation of features from all FCMs
- Detect anomalous behavior, alerting and mitigation in real time
- Periodic retraining of ML models



Feature extraction in FCM

Base features → Windowing → Extracted features

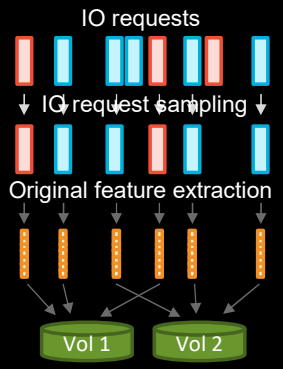
Extracted from IO operations

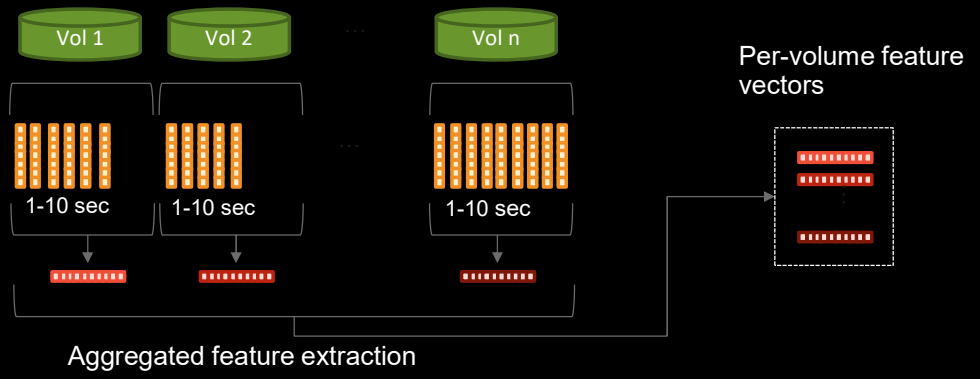
- Shannon Entropy of writes
- Read transfer size
- Write transfer size
- Read LBA
- Write LBA
- NVMe application tag (volume ID)

Aggregated features are extracted using a moving window over 1-10 seconds

Additional features extracted from windows for each volume:

- Mean/variance/Kurtosis of entropy of writes
- Mean/variance read and write transfer size
- Variance/Kurtosis of LBAs read and written
- Read and write IO rate





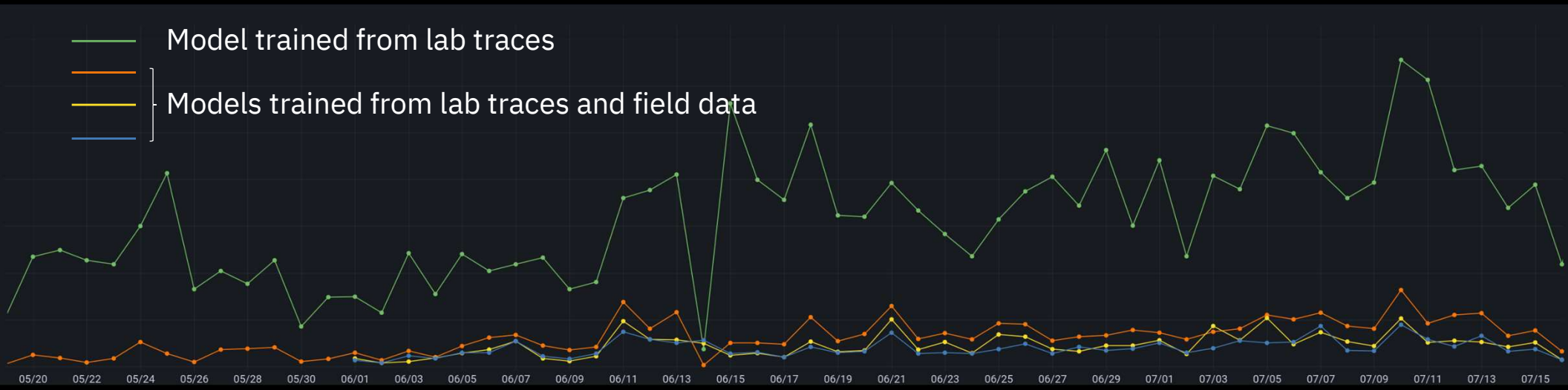
Measured ransomware detection time

Inference time for 1000 volumes every 2 sec in less than 10 ms

Detection in less than 1 min

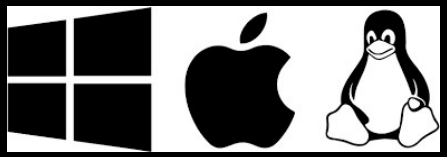
- Results measured while the inference engine is performing the feature vector classification for 1000 volumes in parallel and the evaluation classification results using majority voting.
- Evaluated the ransomware detection time in a KVM setup with a Windows 10 VM where the Conti ransomware was executed.

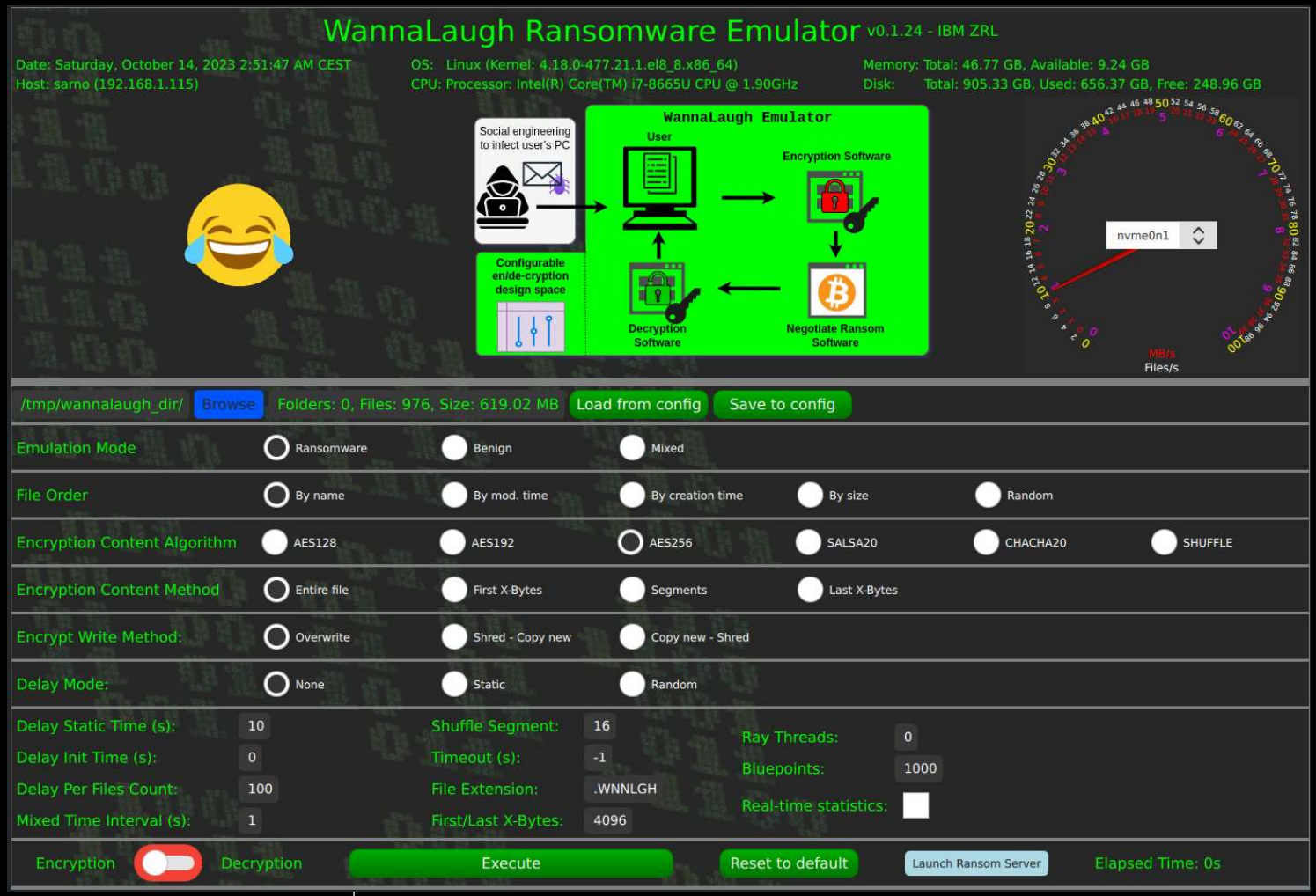
Improving classifier accuracy with field data



- Sets collected from real systems in the field can be used to retrain models
- Here, the FPR of the single-level classifier was reduced by 80%-90% with models trained that include field data

Introducing WannaLaugh





Outlook for block-level ransomware detection

Current features

- Ransomware detection on >1000 volumes
- Training with 50+ real ransomware and emulated ransomware strains in 200+ configurations
- Continuous ML model updates
- Filesystem-aware ML models
- 32k volumes and volume grouping

Outlook

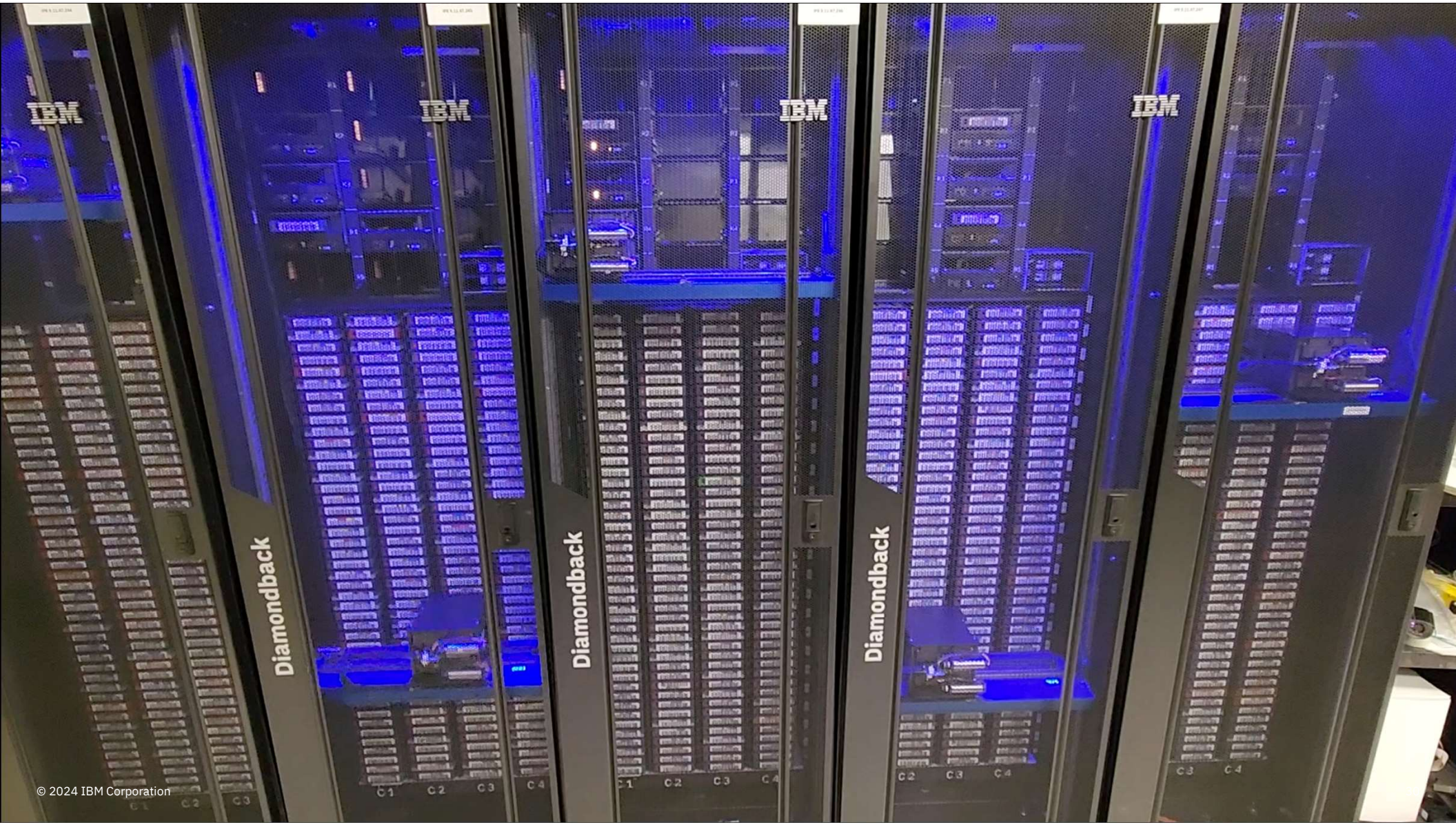
- Multi-variate time series processing
- ML models for wiperware and exfiltration
- Using host stats as additional features
- Operating System awareness, …

Shipping began in 1952

Construction began in 1173

INSIC 2024-2034 Tape Roadmap

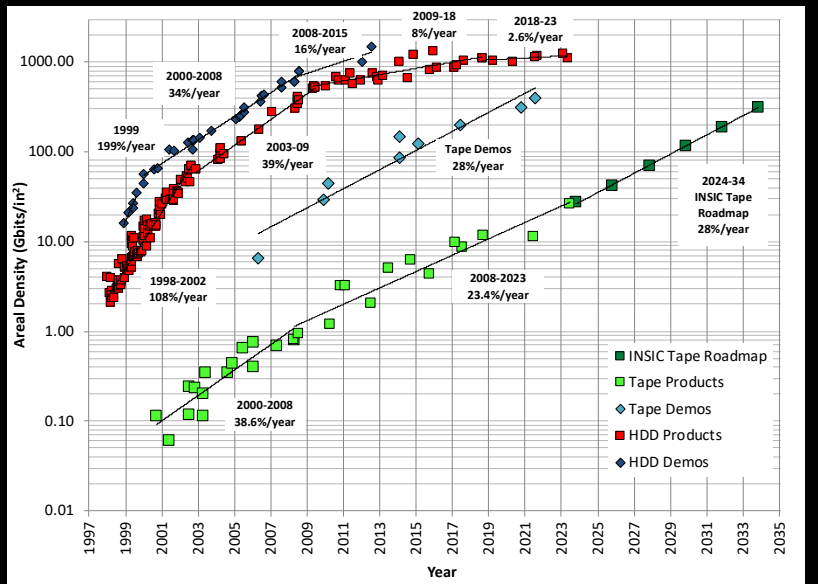
PARAMETER/Year	2024	2026	2028	2030	2032	2034	
1. Capacity (TB)	45	78	137	238	415	723	32% per year
2. Maximum data rate per channel (MB/sec)	12.5	8.3	10.9	14.5	19.1	12.6	
3. Maximum streaming drive data rate (MB/s)	400.0	529.0	699.6	925.2	1223.6	1618.2	15% per year
4. Minimum streaming drive data rate (MB/s)	90.7	200.0	220.5	243.1	268.0	591.0	1.2 m/s min
5. FC Speed Roadmap (GB/sec*)	128	256	256	512	512	1024	
6. Number of channels	32	64	64	64	64	128	
7. Tape speed (m/sec)	5.3	3.2	3.8	4.6	5.5	3.3	
8. Tape thickness	5.00	4.75	4.52	4.30	4.08	3.88	-2.5% per year
9. Data capacity reserve	3.0%	3.0%	3.0%	3.0%	3.0%	3.0%	
10. Tape length that is recordable (meters)**	1036	1092	1151	1213	1278	1347	40 meters reserve
11. Tape length total (meters)***	1076	1132	1191	1253	1318	1387	2.56% per year
12. Track density (TPI)	45,296	63,677	95,494	143,222	214,824	322,248	21.68% per year
Track pitch = 2.54×10^7/tpi (nm)	561	399	266	177	118	79	
13. Linear bit density (kfci)****	600	662	729	804	886	977	5.00% per year
fcmm = kfci/0.0254	23,622	26,043	28,713	31,656	34,901	38,478	
14. Areal density (Gb/in^2)	27.18	42.12	69.64	115.16	190.44	314.95	27.76% per year
15. Tape width in mm	12.65	12.65	12.65	12.65	12.65	12.65	
16. ECC and formatting overhead	20%	20%	20%	20%	20%	20%	0% per year
17. Servo track and layout overhead****	16.0%	10.4%	10.4%	10.4%	10.4%	10.4%	
18. Number of passes to write a tape	592	444	666	999	1498	1123	
19. Number of passes to end-of-life (media)	34560	37093	39812	42730	45862	49223	3.6% per year
20. Time to fill a tape in hours	31	41	54	71	94	124	14.78% per year
21. Number of data tracks	18,949	28,415	42,613	63,911	95,862	143,799	22.47% per year
22. Bit Aspect Ratio (BAR)	16	12	9	6	5	3	-14.26% per year
23. Uncorrectable Bit Error Rate (UBER)	1e-20	1e-20	1e-20	1e-20	1e-20	1e-20	



Tape Phenomenal Scaling

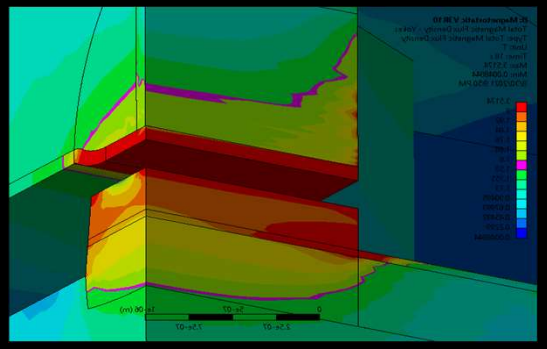
Product Year	IBM 726 1952	LTO9 2021	**TS1170 2023**	Demo 2017 Sputtered Tape	Demo 2020 SrFe Tape
Capacity	2.3 MB	18 TB	**50 TB**	330 TBytes	580 TBytes
Areal Density	1400 bit/in^2	11.9 Gbit/in^2	**26.1 Gbit/in^2**	201 Gbit/in^2	317 Gbit/in^2
Linear Density	100 bit/in	545 kbit/in	**555 kbit/in**	818 kbit/in	702 kbit/in
Track Density	14 tracks/in	21.9 ktracks/in	**47 ktracks/in**	246 ktracks/in	452 ktracks/in

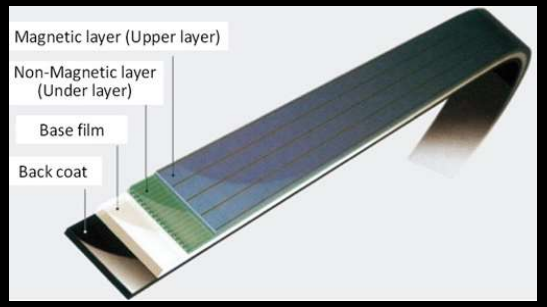
Areal Density **>18.6M** x



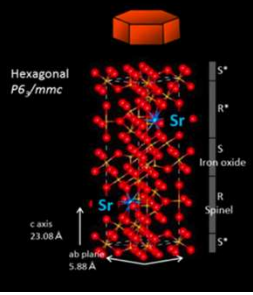
Tape Research Focus Areas

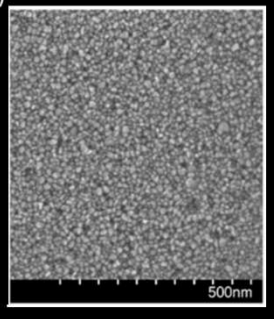
- Signal Processing and Error Correction Coding
  data and servo channel algorithms, iterative decoding
- Mechatronics and Control for Nano-Positioning
  track follow and reel-to-reel control
- Materials Science
  magnetic particles, advanced polymers for substrates, lubricants, wear coatings, thin films for transducer
- Tribology
  friction, lubrication & wear of the tape-head interface
- Mechanical Engineering
  tape path design, actuator design,
- Physics of Magnetism
  finite element and micromagnetic modeling
  read and write transducer design
- Tape System Reliability and Performance Modeling
- Tape System Software





Strontium Ferrite (SrFe)

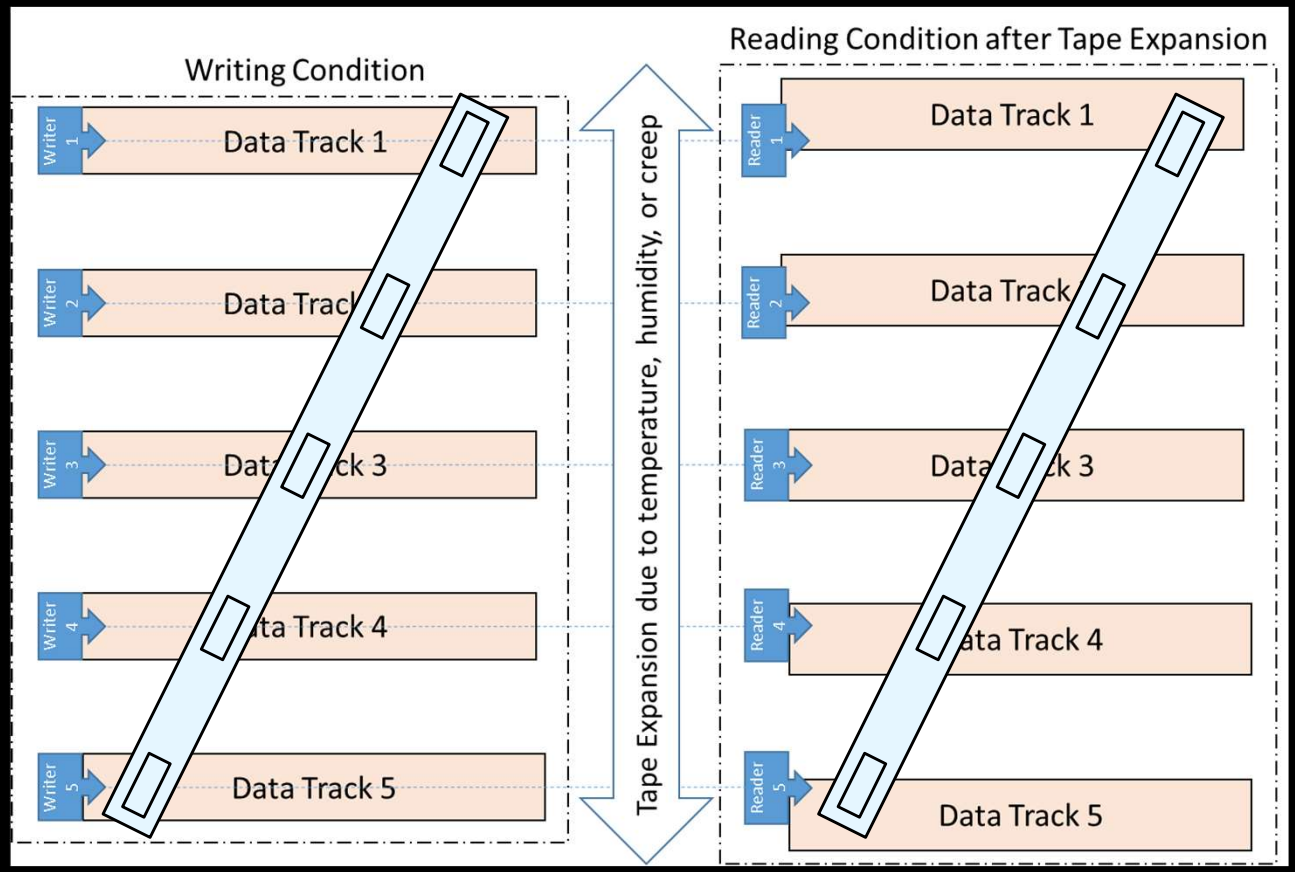




Tape Dimensional Stability (TDS) Compensation by Head Tilt

TDS Compensation by Head Tilt

- The drive continually monitors the condition of the tape
- As mismatches are observed the head tilt is changed
- By increasing or decreasing the angle of the head, the drive can follow expansion or contraction of the tape
- More flexibility to choose operating point



Special Issue on Past, Present, and Future of Storage

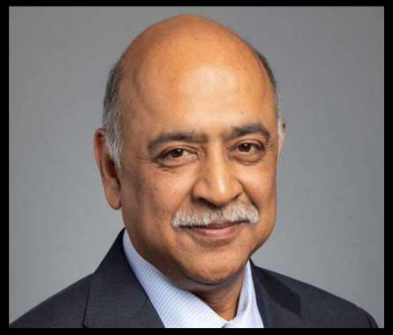
ACM Transactions on Storage , Feb 2025

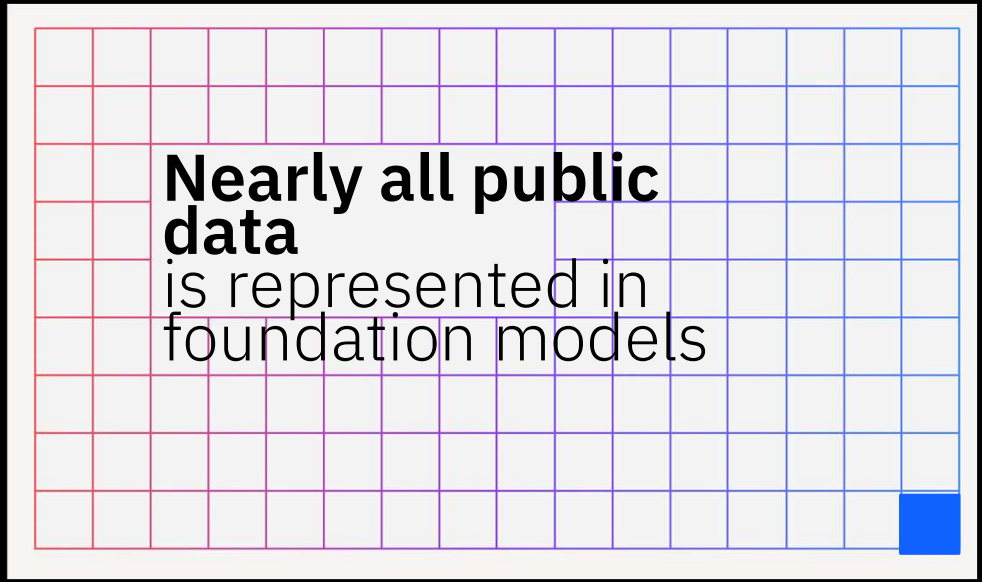
Organizations Need to Unlock Value From Data...
Wherever it Resides

Data is the fuel for an effective AI strategy

Only small fraction of enterprise data is used in Gen AI

“We fundamentally believe that core to the competitiveness of every company going forward will be their ability to use AI to unlock real-time value from their data wherever the data resides..”

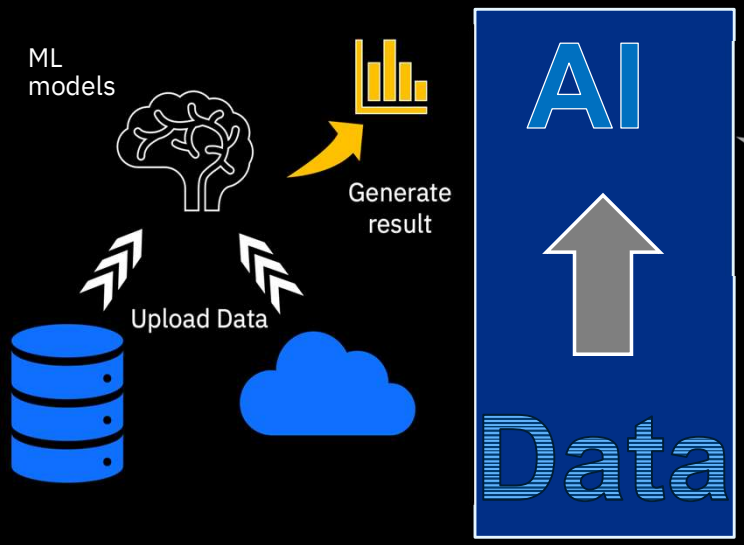




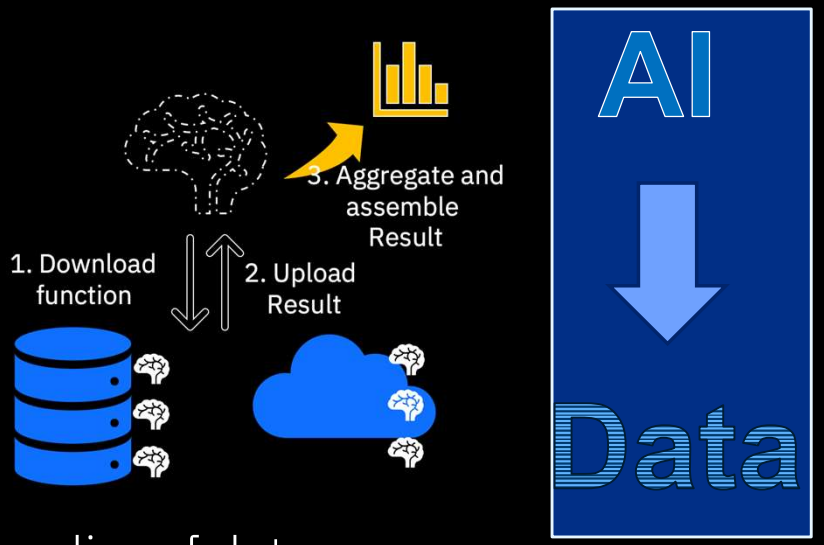
But only 1% of enterprise data!

Vision of Content-Aware Storage

Current Storage



Content-aware Storage



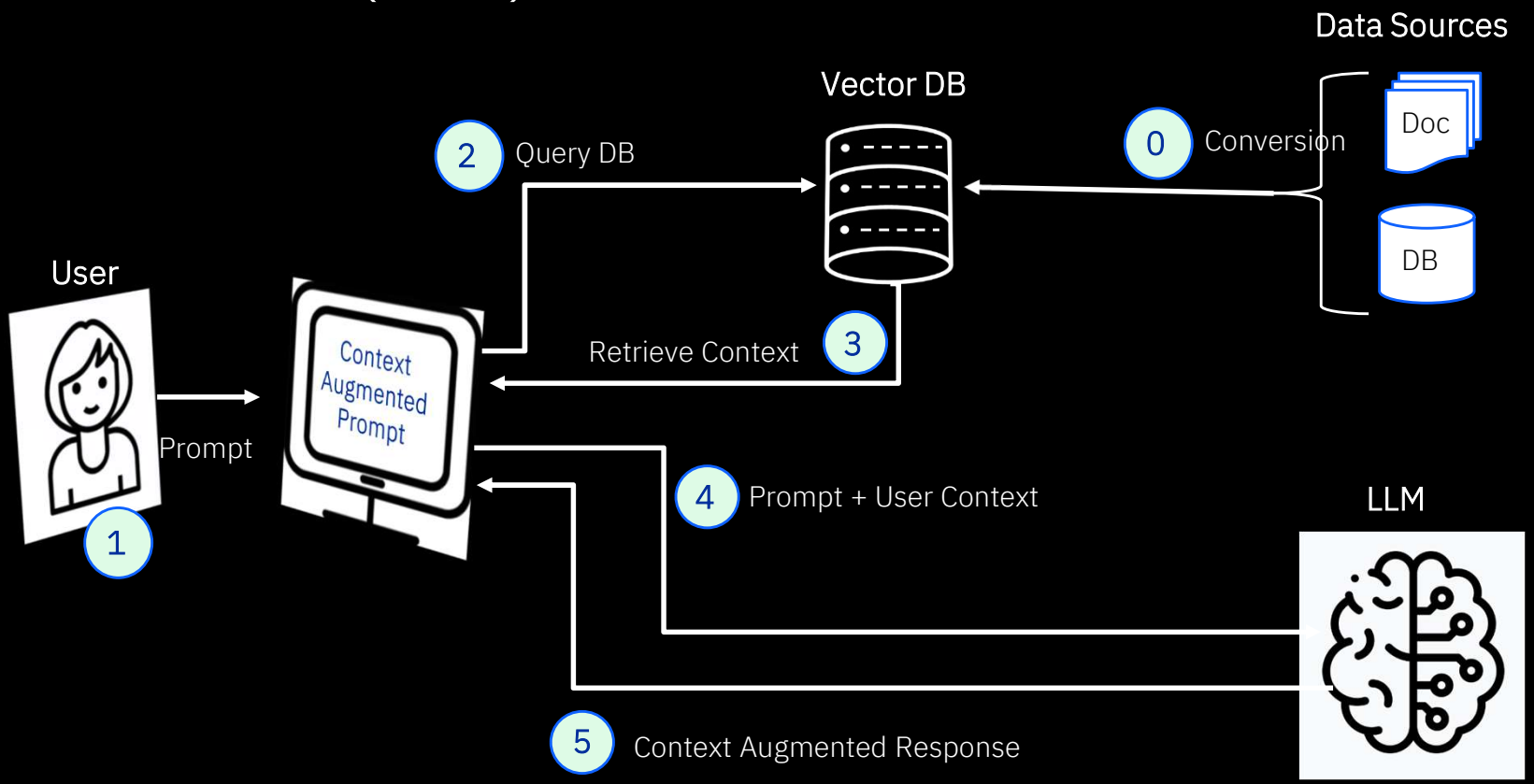
Passive supplier of data:

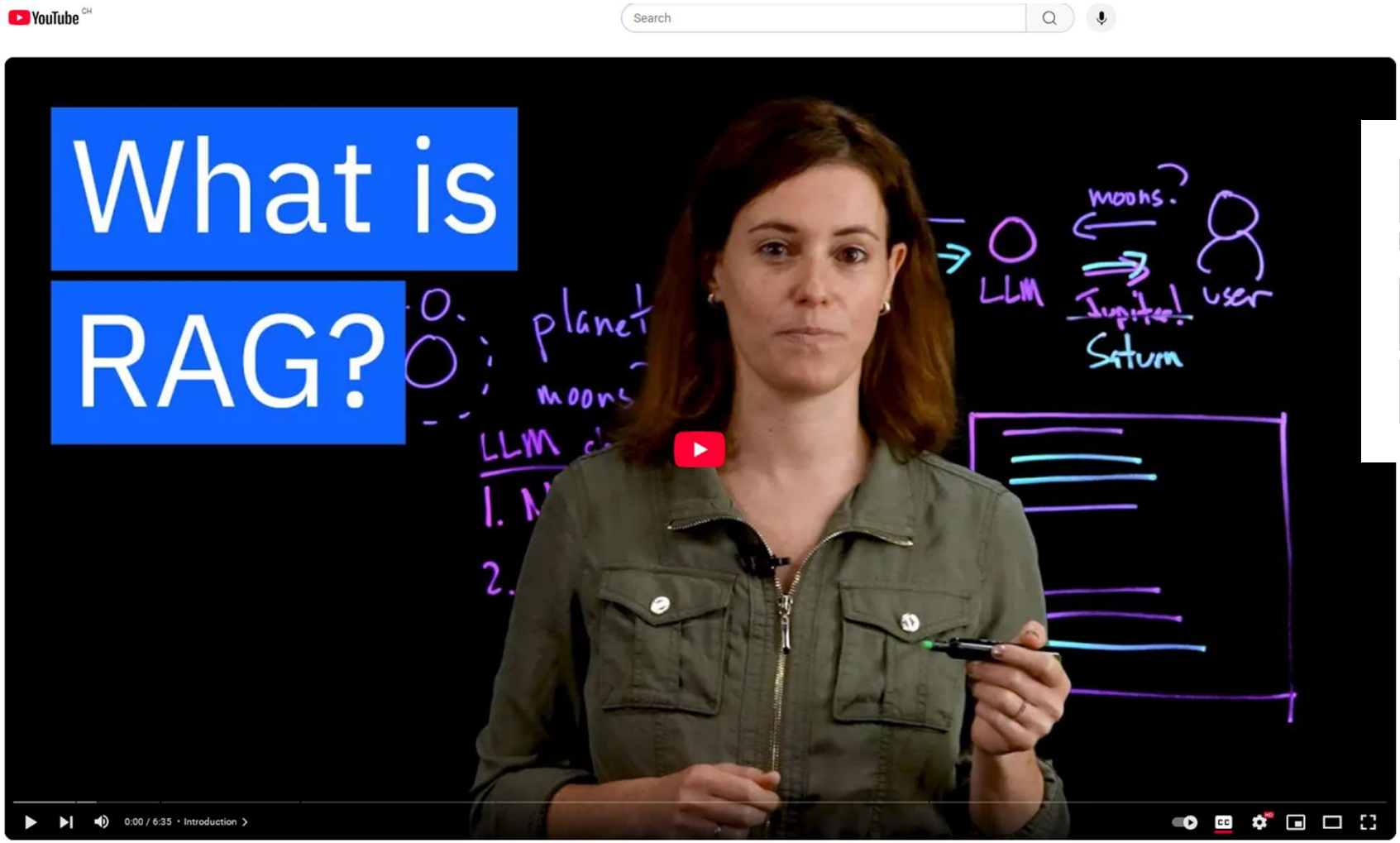
- No deep understanding of content
- No way to search or ask for content other than
  by explicit URL (e.g. S3 buckets)
- Bit-level compression

Active supplier of data:

- Responds to ad hoc natural-language queries aided by semantic memory
- Advanced content-aware compression and de-duplication
- Rapid adaptive re-inference based on ever-changing content

Content-Aware-Storage supporting Retrieval Augmented Generation (RAG)





What is Retrieval-Augmented Generation (RAG)?

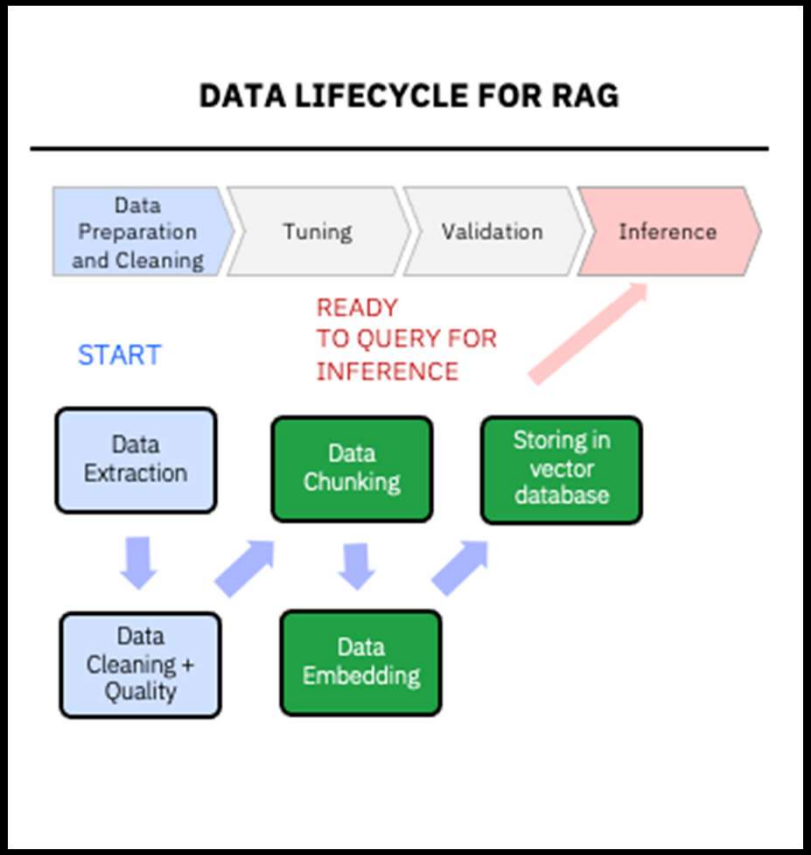
IBM Technology
1.12M subscribers
Subscribe

26K Share Save

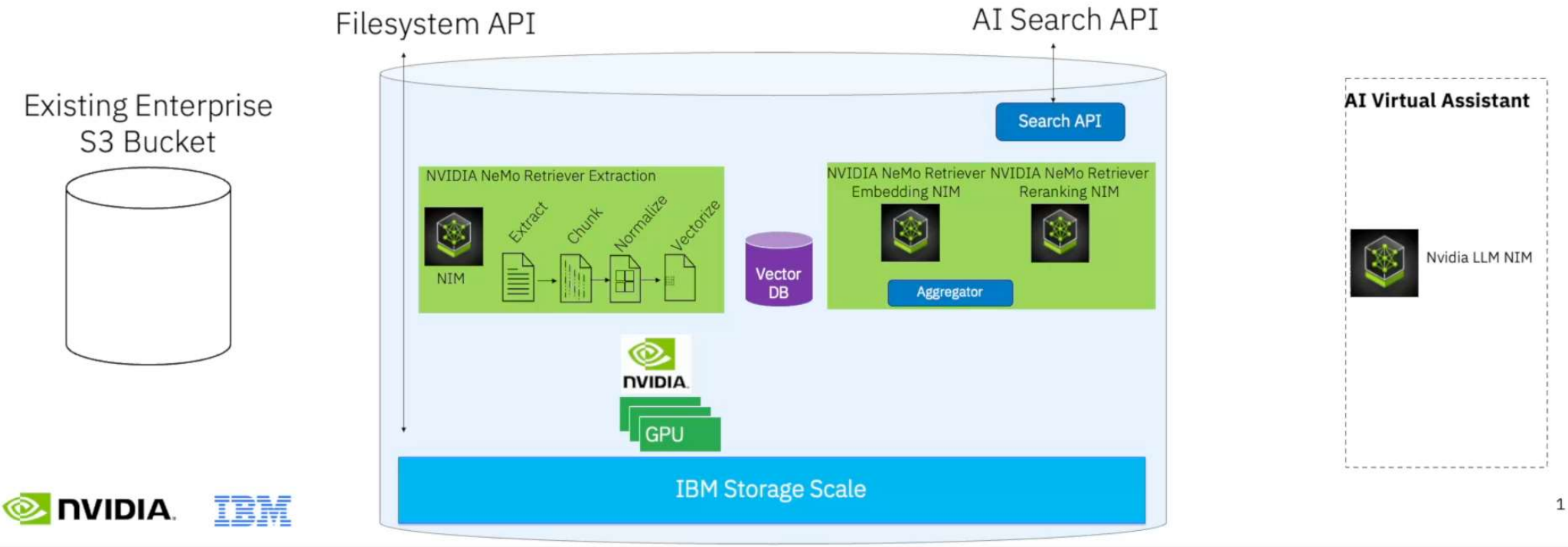
1M views 1 year ago
Ready to become a certified GenAI engineer? Register now and use code IBMTechYT20 for 20% off of your exam → https://ibm.biz/BdGhCF
Learn about the technology → https://ibm.biz/BdMsRT
...more

Current RAG implementations for enterprise today are complex, inefficient and costly

1. High cost: too many copies of data
2. Security exposure: too many replicas of data with Inconsistent ACL between raw data and embedding
3. Complex: too many technologies to integrate
4. Stale embedding: long latency between data change and updated vectors
5. Limited scalability

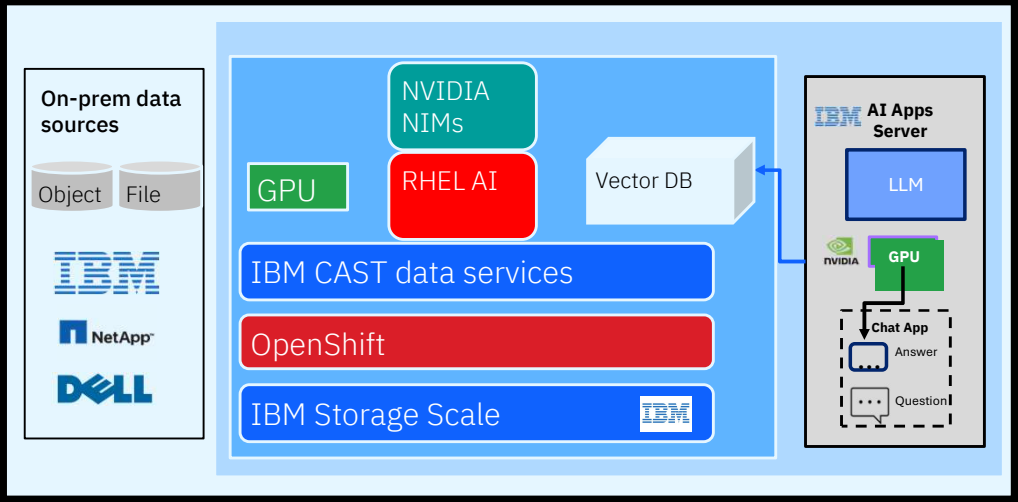


Demo : IBM Content Aware Storage Scale



IBM just announced Content-Aware Storage (CAST): a software-defined storage data service addressing GenAI challenges

Combine power of **IBM Granite** AI processing with **IBM's** AI storage software, & research innovations to a storage-based RAG solution



Simple:

Automated RAG solution

Enables Gen AI capabilities on unstructured data in any on-prem location

Efficient:

a) Cost/performance
b) Works with legacy data

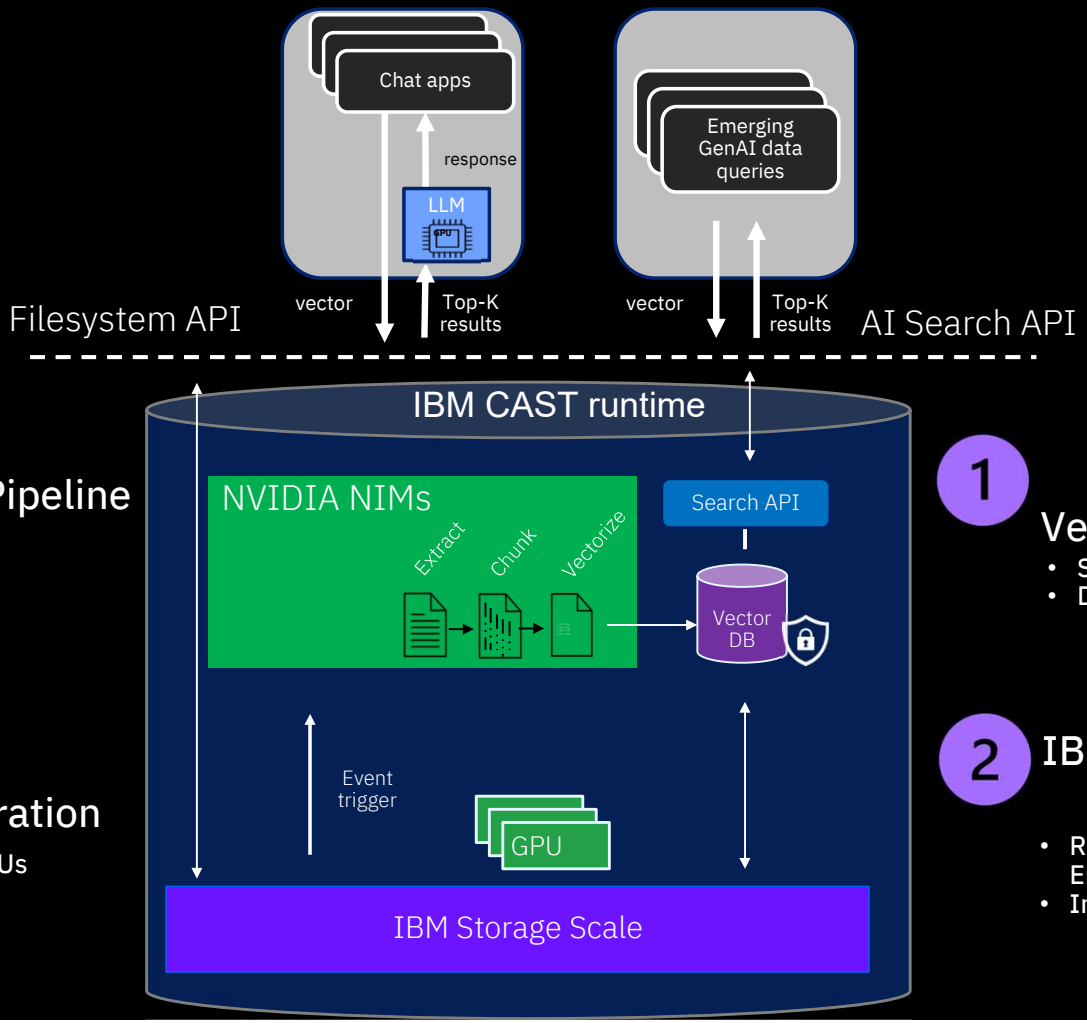
Only process incrementally changed data; High performance shared storage for data processing; GPU optimized storage for optimized NIMs performance

Secure:

Preserve data ACL; Data encryption for embedding

IBM CAST with NVIDIA NIMs Data pipeline

Integrating Nvidia Multimodal PDF Data Extraction and NIMs Blueprint with IBM AI storage



1 Vector Database

- Scalable to 1B vectors and beyond
- Datasource ACL preservation

2 IBM AI Storage and Gen AI runtime

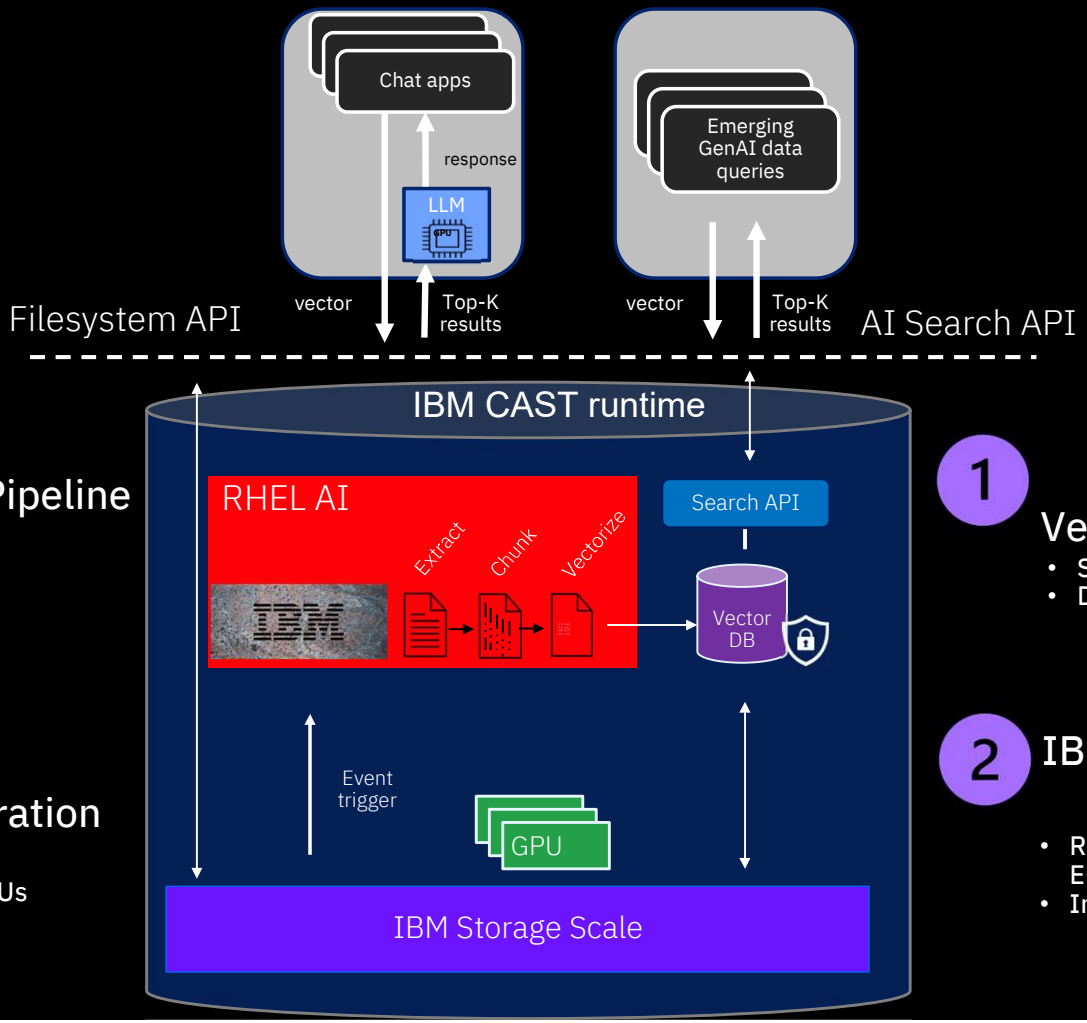
- RAG Virtualization for Existing Enterprise Data
- Incremental Data Processing

3 Data Processing Pipeline

4 Hardware Acceleration

- Nvidia L40S GPUs

IBM CAST with RHEL AI Data pipeline



3 Data Processing Pipeline

1 Vector Database

- Scalable to 1B vectors and beyond
- Datasource ACL preservation

4 Hardware Acceleration

- IBM Spyre
- Nvidia L40S GPUs

2 IBM AI Storage and Gen AI runtime

- RAG Virtualization for Existing Enterprise Data
- Incremental Data Processing

Swiss IT Magazine hat sich in den letzten Jahren klar als *das* IT-Magazin für die Schweiz profiliert.

Mit einer Auflage von monatlich 7000 Exemplaren erreicht Swiss IT Magazine CIOs und Informatik-Verantwortliche in den Schweizer Unternehmen ebenso wie IT-Professionals.

Online erreicht Swiss IT Magazine monatlich über 60'000 Besucher (Unique Clients) und erzielt 148'000 Page Impressions.

Dezember-Ausgabe (9.12.2024) von "Swiss IT Magazine" mit Schwerpunkt auf den Ausblick auf Technologie- und ICT-Trends 2025

“Die Rolle von Speichertechnologien in der neuen Welt von Cybersecurity und KI”,
von Robert Haas