Solving the Data Gravity Problem in RF Engineering

December 08, 2025

Executive Summary

The modern electromagnetic spectrum (EMS) has transformed from a static, regulated utility into a dynamic, contested, and exponentially expanding domain of operation. Across telecommunications, aerospace, and defense sectors, the critical asset is no longer merely the hardware used to transmit or receive signals, but the digitized data representing the spectrum itself. As 5G and 6G networks push into millimeter-wave frequencies, and as military adversaries deploy sophisticated cognitive electronic warfare (EW) systems, the volume and velocity of Radio Frequency (RF) data have precipitated a crisis in information management. This crisis is defined by "Data Gravity", the tendency of massive datasets to become immovable, effectively trapping intelligence within the localized infrastructure where it was captured.

Current RF engineering workflows are buckling under this weight. High-fidelity In-Phase and Quadrature (I/Q) recordings, essential for training Artificial Intelligence (AI) models and validating complex waveforms, now routinely exceed terabytes in size for mere minutes of capture. Traditional file systems, network architectures, and proprietary vendor software are mathematically incapable of scaling to meet this demand. Engineering teams are forced to rely on physical logistics, shipping hard drives via "sneakernet", creating unacceptable latency in decision-making loops and fostering dangerous data silos.

This white paper presents an exhaustive analysis of the data gravity challenge in RF engineering. It argues that the solution requires a fundamental architectural paradigm shift: moving away from hardware-centric, file-based workflows toward a software-defined, purpose-built Enterprise RF Data Lake. By decoupling massive binary storage from lightweight metadata indexing, and by enforcing open standards such as the Signal Metadata Format (SigMF), organizations can neutralize the paralyzing effects of data gravity. This architecture not only resolves immediate storage bottlenecks but also serves as the foundational "feature store" required to unlock the strategic potential of AI, ensuring compliance with Department of Defense mandates like the Modular Open Systems Approach (MOSA) and securing a competitive advantage in the era of spectrum dominance.

1. The Physics of Data Gravity in the Electromagnetic Spectrum

1.1 The Gravitational Pull of Massive Datasets

The concept of "Data Gravity," first articulated by Dave McCrory, provides the theoretical framework for understanding the current crisis in RF engineering. The core theorem posits that data possesses "mass." As a dataset grows in size, it exerts a gravitational pull on applications, services, and compute resources, attracting them toward the data's physical location. The larger the dataset, the more difficult and costly it becomes to move, and the more it dictates the architecture of the systems around it.[1]

In the context of typical enterprise IT, banking ledgers, customer relationship management (CRM) databases, or web server logs, data gravity is a manageable phenomenon. These datasets consist of text and structured rows that scale linearly and can be compressed efficiently. In the domain of Radio Frequency engineering, however, data gravity is exponential and physically imposing. RF data is not a log of an event; it is a high-fidelity digital reconstruction of physical reality, sampled at rates that can exceed billions of measurements per second.

This accumulation of mass creates a "black hole" effect in spectrum operations. Once a high-bandwidth recording is committed to storage, the gravitational pull is so strong that the data rarely leaves its resting place. The cost of bandwidth (egress fees), the latency of transfer (speed of light limitations over fiber), and the sheer operational risk of moving petabytes of information render traditional cloud-first or centralized processing strategies unfeasible.[3] For RF engineers, this means that valuable signal intelligence collected at the tactical edge, whether on a naval vessel, a satellite ground station, or a remote test range, remains trapped on local solid-state drives (SSDs). It becomes "dark data," inaccessible to the broader enterprise for trend analysis, cross-mission correlation, or machine learning training until physical media can be manually transported.[4]

1.2 The Velocity of Capture: Quantifying the Mass

To fully grasp the magnitude of the data gravity problem in RF, one must move beyond abstract concepts and examine the raw mathematics of signal digitization. Unlike video or audio files, which benefit from perceptual compression algorithms (like H.264 or MP3) that discard imperceptible data, RF analysis typically demands raw, uncompressed I/Q data. Preserving the phase and amplitude integrity of the signal is non-negotiable for demodulation, beamforming, and signal characterization tasks.

The data rate for a complex I/Q signal is derived from the fundamental Nyquist-Shannon sampling theorem, which dictates that the sampling rate must be at least twice the maximum frequency component of the signal of interest. In practice, to account for anti-aliasing filter rolloff, a multiplier of 1.25x the target bandwidth is often used.[5]

The formula for throughput is:

$$\text{Data Rate} = \text{Sample Rate (Samples/s)} \times \text{Bit Depth (Bytes/Sample)} \times \text{Channel Count}$$

Let us apply this to the current generation of hardware used in defense and commercial research to visualize the scale.

Case Study A: The Research Standard (Ettus USRP X310)

The Ettus USRP X310 is a workhorse Software Defined Radio (SDR) in academic and military labs. It features high-speed ADCs capable of wideband capture.[6]

- **Target Bandwidth:** 160 MHz (typically utilizing dual 10GbE interfaces).
- **Sample Rate:** 200 Million Samples Per Second (MSps).
- **Bit Depth:** Standard I/Q sampling uses 16-bit integers for both the In-phase (I) and Quadrature (Q) components. This equates to 4 bytes per sample (2 bytes I + 2 bytes Q).
- **Channels:** 2 concurrent channels (MIMO).

$$\text{Throughput} = 200,000,000 \times 4 \text{ bytes} \times 2 = 1.6 \text{ Gigabytes per second (GB/s)}$$

At a capture rate of 1.6 GB/s, the storage accumulation is rapid:

- **1 Minute:** 96 GB
- **1 Hour:** 5.76 TB
- **24 Hours:** 138.24 TB

To put this in perspective, a single day of recording from one mid-range SDR consumes more storage than the entire text content of the Library of Congress.

Case Study B: The High-End Spectrum Monitor (Per Vices Cyan)

Moving up the value chain to ultra-wideband recorders used for 5G/6G development and advanced Electronic Warfare, we examine platforms like the Per Vices Cyan.[8] This class of SDR is designed for massive instantaneous bandwidth monitoring.

- **Target Bandwidth:** 1 GHz per channel.
- **Sample Rate:** 1.25 GSps (approximated for 1GHz bandwidth).
- **Bit Depth:** 16-bit I/Q (4 bytes/sample).
- **Channels:** 4 Independent Channels.

$$\text{Throughput} = 1,250,000,000 \times 4 \times 4 = 20 \text{ GB/s}$$

- **1 Minute:** 1.2 TB
- **1 Hour:** 72 TB
- **24 Hours:** 1.72 Petabytes (PB)

In this scenario, a single system generates nearly 2 Petabytes of data every day. This is the physical reality of data gravity. Standard 10 Gbps enterprise networks are mathematically incapable of moving this data in real-time (10 Gbps is roughly 1.25 GB/s, an order of magnitude slower than the capture rate). Even advanced 100 Gbps links would be saturated by a single device, leaving no bandwidth for the rest of the organization.

1.3 The "Binary Blob" and Metadata Detachment

The secondary dimension of RF data gravity is the "opacity" of the data structures. In the absence of specialized management systems, engineering teams often store these recordings as raw binary files, effectively massive "binary blobs".[4]

A 5TB file containing raw I/Q samples is opaque to traditional information retrieval tools. Standard operating system search functions (Windows Search, grep) cannot index the contents of a binary file to find "signals centered at 2.4 GHz" or "transmissions using QPSK modulation." Without external metadata describing the center frequency, sample rate, timestamp, gain settings, and geospatial location, the binary data is indistinguishable from random noise.

In legacy workflows, this critical metadata is frequently detached from the binary payload. It resides in handwritten laboratory notebooks, disparate text files (README.txt), or proprietary sidecar files that are not indexed centrally.[10] When a 50TB dataset is stored on a Network Attached Storage (NAS) appliance without a searchable index, it becomes undiscoverable. An engineer seeking specific test data from six months ago has no mechanism to query the archive. They are forced to download massive files to their local workstation, open them in analysis software, and visually inspect them, a process that wastes hundreds of engineering hours and creates redundant data movement that further congests the network. This lack of discoverability cements the data gravity trap: because the data cannot be queried remotely, it must be moved; because it is too heavy to move, it remains unused.

2. The Legacy Infrastructure Gap

The mismatch between the physics of RF data and the capabilities of traditional IT infrastructure has created a widening gap in operational capability. As sensor resolutions increase and multi-channel MIMO (Multiple Input Multiple Output) systems become standard, legacy storage and transfer methodologies are failing.

2.1 The "Sneakernet" Reality: Bandwidth vs. Latency

Faced with petabyte-scale datasets and bandwidth-constrained networks, the defense and aerospace industries have reverted to the "Sneakernet", the manual, physical transport of storage media between locations. While often joked about, the Sneakernet is a legitimate data transport protocol with distinct characteristics: infinite bandwidth but horrific latency.[11]

Consider a test range in the Nevada desert recording 500 TB of data during a week-long exercise. Transferring this data to a research lab in Massachusetts over a standard 1 Gbps dedicated link would take:

$$500 \text{ TB} \times 1024 \times 8 \text{ bits}/1 \text{ Gbps} \approx 4,096,000 \text{ seconds} \approx 47 \text{ days}$$

Even with a 10 Gbps link, the transfer takes nearly 5 days, assuming 100% network efficiency and no other traffic, an unrealistic best-case scenario.
In contrast, loading a crate with High-Density HDDs or AWS Snowball devices and shipping them via overnight air courier takes 24 hours. Thus, the Sneakernet remains the default "high bandwidth" solution.[12] However, this reliance on physical transport introduces severe operational risks that undermine modern agile workflows:

1. **Latency of Intelligence:** Critical threat data captured on Day 1 of an exercise is not visible to analysts until Day 7 or later. In a near-peer conflict scenario, this delay in the OODA (Observe-Orient-Decide-Act) loop is a critical vulnerability.
2. **Chain of Custody and Security:** Physical drives are vulnerable to loss, theft, damage, and tampering during transit. Managing the classification lifecycle of hundreds of physical disks is a logistical nightmare for security officers.
3. **Data Silos and Fragmentation:** Once drives arrive, they are often copied to local workstations rather than a central repository. This leads to version control chaos, where multiple engineers work on divergent copies of the same dataset, and valuable data remains trapped on individual desks, invisible to the wider enterprise.[4]

2.2 The Limitations of Hardware-Centric Vendor Ecosystems

The RF test and measurement market is dominated by established hardware giants such as Keysight Technologies, Rohde & Schwarz (R&S), and National Instruments (NI). These companies produce world-class sensors and analyzers, but their software ecosystems have historically exacerbated the data gravity problem rather than solving it.

- **Keysight PathWave:** PathWave is the industry standard for signal analysis. It offers powerful tools for demodulating and visualizing waveforms. However, its architecture is fundamentally file-centric and seat-licensed.[14] It excels at analyzing a signal *currently loaded in memory*, but it is not designed to manage a petabyte-scale repository of historical data. Furthermore, legacy versions often rely on proprietary file formats (.dat, .xdat) that lock the data into the Keysight ecosystem, making it difficult to index or process using third-party AI tools.[10]
- **Rohde & Schwarz:** High-end wideband recorders like the R&S IQW series are engineering marvels capable of sustaining massive write speeds.[16] Yet, they often function as "storage islands." The workflow typically involves a single operator capturing data to internal RAID arrays. Exporting this data for enterprise-wide access is a secondary, manual process. The proprietary nature of formats like .iq.tar (though documented) creates friction for seamless integration into open data lakes.[17]
- **National Instruments (NI) SystemLink:** NI has moved closer to a data management solution with SystemLink, which offers TDM (Technical Data Management) capabilities.[19] However, SystemLink is optimized for parametric data (scalar test results, voltage readings, pass/fail logs) rather than the massive, streaming I/Q binary blobs characteristic of SIGINT and EW operations. Its architecture is often tied to the .tdms format, which, while efficient, still represents a specific vendor standard rather than a universal open protocol.[20]

2.3 The Economic Failure of General-Purpose Data Lakes

When faced with massive data volumes, IT departments often propose generic enterprise data lake or SIEM (Security Information and Event Management) solutions like Splunk or Elastic (ELK Stack). While these tools are excellent for text-based logs, they are economically and technically unsuited for RF data.[9]

- **Pricing Models:** Platforms like Splunk often charge based on "Daily Ingest Volume." Ingesting 100 TB of binary RF data per day would result in annual licensing fees in the tens of millions of dollars.[22]
- **Technical Capability:** These tools index text. They do not natively understand complex I/Q number formats (pairs of 16-bit integers or 32-bit floats). They cannot natively render a spectrogram or waterfall plot from binary data without massive, custom engineering. Storing binary blobs in an index designed for text strings causes massive database bloat and performance degradation.[24]

3. Architecting the Solution: The Purpose-Built RF Data Lake

To solve the data gravity problem, RF engineering must adopt the "Data Lake" architecture, but it must be adapted specifically for the physics of signal data. The proposed solution, exemplified by architectures like **SigDrive**, relies on a fundamental separation of concerns: **Decoupling Compute, Storage, and Metadata**.[4]

3.1 Decoupling Binary Storage from Metadata

The core architectural innovation required to overcome RF data gravity is the separation of the heavy "binary blob" (the I/Q samples) from the lightweight "metadata" (the description of the signal).

1. Object Storage for Binary Blobs:
Instead of using traditional file systems (NTFS/ext4) or block storage, the system should utilize Object Storage (S3-compatible protocols like MinIO, Ceph, or AWS S3).26 Object storage is designed for massive scalability and handles unstructured binary data efficiently.

- **Scalability:** Object stores handle petabytes to exabytes of data in a single namespace by simply adding storage nodes. The underlying erasure coding ensures data durability without the massive overhead of RAID rebuilds.
- **HTTP API Access:** Data is accessed via RESTful APIs (GET/PUT/HEAD). This allows web-based applications to request specific *byte ranges* of a file. An analyst can visualize the first 10MB of a 5TB file without downloading the entire object, effectively neutralizing the latency penalty of the file size.[4]
- **Cost Efficiency:** Object storage allows for the use of dense, commodity hardware, significantly lowering the cost per terabyte compared to high-performance SAN or NAS filers.[28]

2. Relational Database for Metadata:
While the binary file sits in object storage, its metadata is extracted and normalized into a high-performance relational database, such as PostgreSQL.4

- **Indexing:** Key parameters (Center Frequency, Sample Rate, Timestamp, Bandwidth, Gain) are indexed for millisecond-latency searches.
- **Geospatial Intelligence:** Leveraging extensions like **PostGIS** transforms the RF archive into a geospatial system. Users can execute queries such as "Find all signals captured within 10km of coordinates X,Y between 08:00 and 12:00." This spatial indexing is impossible in flat-file architectures.[29]

The Workflow Shift:
In this decoupled architecture, when a user searches for data, they query the Metadata Database, not the file system. A search over a petabyte archive scans an index size measured in megabytes, returning results instantly. This decouples the discovery of data from the mass of data. The user identifies the relevant recordings first, and only then incurs the bandwidth cost of retrieving the binary payload.4

3.2 Tiered Storage Architecture and Lifecycle Management

Given the massive volume of RF data, storing everything on high-performance flash storage is financially ruinous. A purpose-built RF Data Lake must implement automated **Storage Lifecycle Policies**.[4]

- **Hot Tier (NVMe/SSD):** This tier ingests new data and hosts datasets currently active in analysis or AI training pipelines. It offers the highest throughput (GB/s) to support hungry GPUs but comes at the highest cost per TB.
- **Warm Tier (HDD/Object Store):** Data that has been ingested but not accessed in the last 30 days is automatically migrated to high-density spinning disks. This tier offers moderate performance at a significantly lower cost.
- **Cold Tier (Deep Archive/Tape):** A significant portion of RF data is "Write Once, Read Never", retained solely for compliance, forensic auditing, or long-term trend analysis. This data should be migrated to cold storage solutions like AWS Glacier or LTO Tape Libraries. These mediums offer the lowest possible cost per terabyte (often pennies per GB/month).[32]

The Data Lake software must abstract this complexity from the user. A file might physically reside on an LTO tape, but it appears in the user's search results. When requested, the system triggers a retrieval job, moving the data from Cold to Hot storage transparently.[34]

3.3 "Air-Gap Ready" Microservices Design

The defense and intelligence sectors operate under strict security constraints. Systems must frequently run in **Air-Gapped** environments, disconnected from the public internet (SIPRNet, JWICS, discrete labs). This creates a unique "Compliance Gravity" that prevents the use of convenient public cloud SaaS tools.[4]

To address this, the RF Data Lake architecture must be built on **Microservices** and **Containerization** (e.g., Docker, Kubernetes).[4]

- **Containerization:** By bundling the application with all its dependencies (libraries, databases, web servers) into a single deliverable image (e.g., a .tar of Docker images), the software can be deployed into a secure facility without requiring internet access to download packages from npm or pip.
- **Offline Operation:** The system must function without "phoning home" for license validation or updates. All update mechanisms must support physical media transfer ("Sneakernet Updates").

- **Sneakernet Ingestion Nodes:** Recognizing that sneakernet is unavoidable for initial transport, the system should include dedicated "Ingestion Kiosks." These are high-performance workstations where physical drives are docked. The system automatically mounts the drive, hashes the files (SHA-256) for integrity, parses the metadata, and begins the background upload to the central Object Store, automating the chain of custody.[4]

4. Standardization: The Strategic Imperative of SigMF

Architecture alone cannot solve the interoperability crisis. A common language is required to describe the data, ensuring that a file recorded by a USRP can be analyzed by a Keysight tool and trained on by a PyTorch model. This is where **SigMF (Signal Metadata Format)** becomes the strategic linchpin.

4.1 SigMF vs. Proprietary Formats

Historically, the RF industry has been plagued by a "Tower of Babel." Every vendor utilized proprietary formats: Keysight (.dat), Rohde & Schwarz (.iq.tar), BlueMidas (.tmp), and countless ad-hoc binary formats created by defense contractors.[4] This forced engineers to write complex, fragile conversion scripts (e.g., MATLAB parsers) every time they moved data between tools, often resulting in the loss of critical metadata like geolocation or center frequency.

SigMF solves this by mandating a strict separation of data and metadata [36]:

1. **.sigmf-data**: The raw binary I/Q samples. This file is agnostic to the standard; it is just the samples.
2. **.sigmf-meta**: A human-readable, machine-parsable JSON file describing the data.

By adopting SigMF as the **Canonical Schema** [4], an Enterprise RF Data Lake creates a universal interoperability layer.

- **Vendor Agnosticism:** Any sensor data is converted (transcoded or simply re-wrapped) to SigMF upon ingestion. The Data Lake speaks one language internally.
- **Extensibility:** JSON is inherently extensible. If a new sensor type emerges (e.g., a quantum RF sensor), new metadata fields can be added to the JSON namespace without breaking the core schema or requiring a database migration.
- **AI Native:** Modern AI frameworks are built to ingest JSON. Parsing a SigMF file in Python is a native operation (json.load), whereas parsing a proprietary binary header requires specialized, often compiled, libraries that may not exist for the target training environment.[38]

4.2 SigMF vs. VITA 49: Storage vs. Transport

A critical distinction must be made between **SigMF** and **VITA 49 (VRT)** to avoid architectural confusion.

- **VITA 49** is a *transport* protocol. It is packet-based, designed for streaming data over Ethernet (UDP/TCP) in real-time. It interleaves metadata (context packets) with signal data. It is optimized for "Data in Motion".[36]
- **SigMF** is a *storage* format. It is file-based, designed for "Data at Rest."

The optimal architecture utilizes both in their respective domains. VITA 49 moves data from the sensor to the recorder. The recorder then writes the data to disk as SigMF (or converts VITA 49 packets to SigMF streams). Attempting to use VITA 49 as a long-term storage format is inefficient for Big Data analytics because the metadata is buried inside the binary stream. To find the center frequency of a VITA 49 file, one might have to read gigabytes of packets to find the relevant context packet. In SigMF, that information is available instantly in the JSON sidecar.[40]

4.3 Strategic Alignment with MOSA and JADC2

The **Modular Open Systems Approach (MOSA)** is not just a best practice; it is a US legal requirement (Title 10 U.S.C. 4401) for major defense acquisition programs.[41] It mandates the use of open standards to ensure modularity and prevent vendor lock-in.

Adopting SigMF and a Data Lake architecture provides immediate MOSA compliance for the "Data Layer." It ensures that the government (the data owner) retains full utility of the data regardless of which defense prime contractor built the sensor. This is a prerequisite for **Joint All-Domain Command and Control (JADC2)**, which envisions a unified network where data flows seamlessly between Air Force, Navy, and Army assets. A common, open data format is the "lingua franca" required to build this Data Fabric.[4]

5. Operationalizing AI: From Raw I/Q to Cognitive EW

5.1 The "Data Readiness" Bottleneck

The Department of Defense and the commercial sector are aggressively pursuing AI/ML capabilities. Programs like the US Army's **Project Linchpin** aim to create "Trusted AI" pipelines for Electronic Warfare.[43] The objective is **Cognitive EW**: systems that can automatically detect, classify, and counter unknown signals in real-time.

However, AI models are only as good as their training data. The primary bottleneck in military AI is **Data Readiness**. Archives are filled with petabytes of RF recordings, but they are "swamp data":

1. **Unlabeled:** We have the recording, but we don't know what signals are inside.
2. **Uncurated:** We don't know which files contain valuable threat data versus empty noise floor.
3. **Incompatible:** Different recordings have different sample rates, bit depths, and gains, breaking the input requirements of neural networks.

5.2 The RF Data Lake as a Feature Store

A purpose-built RF Data Lake functions as the critical **Feature Store** for AI pipelines.[4]

- **Normalization:** By converting all ingest to a canonical SigMF schema, the Data Lake presents a uniform, clean dataset to the Data Scientist. The "data wrangling" phase is automated.
- **Annotation:** The system must provide integrated visualization tools (Web-based Spectrograms) that allow analysts to draw **Regions of Interest (ROI)** boxes around signals. These annotations are saved directly into the SigMF metadata.[4]
- **Query-Based Dataset Generation:** A data scientist can execute a precise query: *"Export all clips labeled 'Drone Control Link' with SNR > 10dB captured in 2024, resampled to 10 MSps."* The system automatically subsets the massive binary files, extracting only the relevant chips and converting them to the target format. This reduces the volume of training data from petabytes of raw storage to manageable gigabytes of high-value features, solving the data gravity problem for the AI model training loop.[4]

This capability is essential for **Project Linchpin**, which explicitly calls for infrastructure to support "RF Data Management: acquisition, generation, storage, access and retrieval for model training".[43]

6. Security and Compliance in the Defense Sector

6.1 The "ATO" Barrier

In the defense market, software cannot be deployed without an **Authority to Operate (ATO)**. This certification process validates the security posture of the system. A purpose-built Data Lake must be designed from the ground up to pass these rigorous checks (e.g., RMF, NIST 800-53).

Key security features must include:

- **Role-Based Access Control (RBAC):** Granular control over who can see which datasets. A contractor working on Radar System A should not be able to access SIGINT data from Mission B.[4]

- **Immutable Audit Logs:** Every action, upload, download, search, view, must be logged. If a classified file is accessed, the system must definitively prove who accessed it and when. This is critical for insider threat detection.[4]
- **STIG Compliance:** The underlying containers and databases must be hardened according to the Security Technical Implementation Guides (STIGs) set by DISA.

6.2 Data Sovereignty and Federation

Data gravity often intersects with legal and policy gravity. **Data Sovereignty** laws may prevent data collected in one country from being moved to a cloud server in another. In coalition operations (e.g., Five Eyes), partners may want to share *insights* without sharing raw *data*.

The Data Lake architecture supports **Federated Search**.[35] Multiple instances of the Data Lake can be deployed across different security domains or physical locations. A central "Manager of Managers" can query the metadata indices of all nodes without moving the binary data. An analyst can see that a specific signal exists on a UK node and request access, initiating a controlled, policy-driven transfer of that specific file, rather than a bulk replication of the entire archive.

7. Commercial Applications: LEO and 6G

While the defense sector drives the most extreme requirements, the commercial spectrum market faces identical physics.

7.1 LEO Satellite Constellations

Low Earth Orbit (LEO) satellite operators (Starlink, OneWeb, Planet) manage thousands of satellites downlink data to ground stations. The telemetry and spectrum monitoring data from these downlinks is massive. Ground stations must record spectrum to debug interference issues and monitor link health. The "Store at the Ground Station, Process at the Core" model is broken by data gravity. LEO operators require the same distributed Data Lake architecture to manage ground station recordings, processing interference analysis at the edge and sending only alerts to the Network Operations Center (NOC).[46]

7.2 6G Research and Testbeds

As the telecom industry moves toward 6G, frequencies are shifting into the Terahertz range, and bandwidths are expanding to 10 GHz+. The data rates for 6G testbeds will dwarf current 5G requirements.[48] Capturing "Golden Reference" datasets for 6G waveform validation will require storage throughputs exceeding 100 GB/s. A scalable, sharded Object Storage architecture is the only viable storage medium for this class of data.

8. Conclusion and Strategic Recommendations

The "Data Gravity" problem in RF engineering is a defining challenge of the next decade. It is not merely an IT infrastructure issue; it is a strategic bottleneck that limits the speed of innovation, the efficacy of AI, and the operational tempo of defense forces. Legacy approaches, sneakernets, proprietary files, and siloed workstations, are mathematically incapable of keeping pace with the exponential growth of sensor data.

To secure electromagnetic dominance, organizations must implement a **Purpose-Built Enterprise RF Data Lake**.

Strategic Recommendations:

1. **Adopt a Data-Centric Architecture:** Move away from application-centric workflows. The Data Lake should be the "System of Record," not the individual analysis tool.
2. **Standardize on SigMF:** Mandate SigMF as the delivery requirement for all new sensors and recording campaigns to ensure future interoperability and MOSA compliance.
3. **Invest in "Data Readiness":** Recognize that AI requires curated data. Invest in the infrastructure to label, annotate, and manage data *before* investing in the AI models themselves.
4. **Embrace Tiered Storage:** Stop treating all data as equal. Automate the lifecycle of data from NVMe to Tape to balance the massive costs of petabyte-scale retention.

By mastering the physics of data gravity, RF engineering teams can transform their massive data holdings from a crushing liability into a gravitational well for innovation, attracting advanced analytics, accelerating AI deployment, and enabling a new era of spectrum intelligence.

Feature	**Legacy (File-Based / NAS)**	**Generic IT Data Lake (Splunk/Elastic)**	**Purpose-Built RF Data Lake (SigDrive)**
Primary Storage	Local HDD / Shared NAS	HDFS / Proprietary Index	S3-Compatible Object Storage
Data Gravity	**High** (Data trapped in silos)	**Medium** (Centralized but expensive)	**Managed** (Tiered, Edge-Federated)
Metadata	Proprietary Header / Sidecar	Ingested Text Logs	Relational DB + JSONB (SigMF)
Searchability	File Name / Folder Structure	Text / Log based	Parametric (Freq, Time) & Geospatial
Cost Model	CapEx (Hardware/Storage)	Ingest Volume (Prohibitive for RF)	Storage Capacity (Economical)
AI Readiness	**Low** (Manual conversion required)	**Low** (Not signal-native)	**High** (Canonical Schema / Feature Store)
Visualization	Desktop Apps (PathWave, Matlab)	Custom Dashboards (Limited)	Web-based Spectrogram / Waterfall
Security	High (Air-gapped islands)	Low (Cloud-centric)	High (Containerized / Offline Ready)

Table 1: Comparison of RF Data Management Architectures

References

1. What is Data Gravity? | Talend, https://www.talend.com/resources/what-is-data-gravity/
2. What is Data Gravity and How Does It Impact Your Cloud Strategy? - BETSOL, https://www.betsol.com/blog/what-is-data-gravity-and-how-it-impacts-your-hybrid-cloud-strategy/
3. Data Gravity vs. Data Velocity - Interconnections - The Equinix Blog, https://blog.equinix.com/blog/2024/01/11/data-gravity-vs-data-velocity/
4. Strategic Business Plan_ SigDrive Enterprise RF Data Lake (1).pdf
5. How-to Calculate RF Data Rate / Throughput In LabVIEW From IQ Rate or Bandwidth - NI, https://knowledge.ni.com/KnowledgeArticleDetails?id=kA03q000001DtxDCAS&l=en-US
6. X300/X310 - Ettus Knowledge Base, https://kb.ettus.com/X300/X310
7. About USRP Bandwidths and Sampling Rates - Ettus Knowledge Base, https://kb.ettus.com/About_USRP_Bandwidths_and_Sampling_Rates
8. Cyan - Per Vices, https://www.pervices.com/cyan/
9. What is binary large object (BLOB) storage? - Google Cloud, https://cloud.google.com/discover/what-is-binary-large-object-storage
10. Fundamental challenges of scientific data - TetraScience, https://www.tetrascience.com/blog/fundamental-challenges-of-scientific-data
11. Long live the Sneakernet: Computing's most resilient network | ZDNET, https://www.zdnet.com/article/long-live-the-sneakernet-computings-most-resilient-network-6001018604/
12. AWS Snowball | Secure Edge Computing and Offline Data Transfer | Amazon Web Services, https://aws.amazon.com/snowball/
13. Git is Bad at Binary File Management – But is it Worse than Duplicate 'Versioned' Files?, https://www.reddit.com/r/git/comments/ek4kv2/git_is_bad_at_binary_file_management_but_is_it/
14. How edge computing and 5G allow you to make real-time, data-based decisions, https://blog.shi.com/business-of-it/how-edge-computing-and-5g-allow-you-to-make-real-time-data-based-decisions/
15. PathWave Manufacturing Analytics - Keysight, https://www.keysight.com/us/en/assets/3120-1500/technical-overviews/PathWave-Manufacturing-Analytics.pdf
16. R&S®IQW Wideband I/Q Data Recorder - Rohde & Schwarz, https://www.rohde-schwarz.com/products/test-and-measurement/data-recorder/rs-iqw-wideband-i-q-data-recorder_63493-548624.html
17. R&S®IQW Wideband I/Q Data Recorder - Rohde & Schwarz, https://www.rohde-schwarz.com/us/products/test-and-measurement/data-recorder/rs-iqw-wideband-i-q-data-recorder_63493-548624.html
18. Rohde & Schwarz iq-tar File Format Specification, https://www.rohde-schwarz.com/us/manual/rohde-schwarz-iq-tar-file-format-spec

ification-manuals_78701-37313.html

19. What Is NI SystemLink Software? - National Instruments, https://www.ni.com/en/shop/electronic-test-instrumentation/application-software-for-electronic-test-and-instrumentation-category/systemlink.html
20. A Comprehensive Solution to Large-Scale Data Management - NI - National Instruments, https://www.ni.com/en/shop/electronic-test-instrumentation/application-software-for-electronic-test-and-instrumentation-category/systemlink/automate-data-analysis/a-comprehensive-solution-to-large-scale-data-management.html
21. What is Splunk? Key Benefits and Features of Splunk | Fortinet, https://www.fortinet.com/resources/cyberglossary/what-is-splunk
22. Grand Challenges in AI in Radiology - PMC - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC10364978/
23. Splunk SIEM: Key Features, Limitations and Alternatives - Exabeam, https://www.exabeam.com/explainers/splunk/splunk-siem-key-features-limitations-and-alternatives/
24. Is splunk the best option for storing data? - Reddit, https://www.reddit.com/r/Splunk/comments/osrcr6/is_splunk_the_best_option_for_storing_data/
25. How to Solve 4 Elasticsearch Performance Challenges at Scale | by Julie Mills - Medium, https://medium.com/rocksetcloud/how-to-solve-4-elasticsearch-performance-challenges-at-scale-b56429ea27cc
26. What is Binary Large Objects? - Dremio, https://www.dremio.com/wiki/binary-large-objects/
27. Why Object Storage? A Systems Engineer Explains - Western Digital Blog, https://blog.westerndigital.com/why-object-storage/
28. Cloud Storage Cost: Comparing AWS, Azure, and Google - N2W Software, https://n2ws.com/blog/cloud-storage-cost
29. Cell Radio Frequency (RF) Propagation Algorithm - Homeland Security, https://www.dhs.gov/sites/default/files/publications/WEA%20-%20TCS%20Final%20Report.pdf
30. PostGIS: A powerful geospatial extension for PostgreSQL - Red Hat Developer, https://developers.redhat.com/articles/2025/10/02/postgis-powerful-geospatial-extension-postgresql
31. Petabytes Explained: Understanding Data Storage in Cloud Computing | Lenovo IE, https://www.lenovo.com/ie/en/glossary/petabytes/
32. Tiered Storage Takes Center Stage - Oracle, https://www.oracle.com/assets/oracle-tiered-storage-takes-center-194075.pdf
33. THE ESCALATING CHALLENGE OF PRESERVING ENTERPRISE DATA - Fujifilm, https://asset.fujifilm.com/master/americas/files/2022-08/5082fc03fc44d80b691c87a1a96febd5/Furthur_Market_Research_WP_080322_FINAL.pdf
34. Introduction The selection of data storage technologies has never been more robust. Today's choices range from ultra - Fujifilm,

https://asset.fujifilm.com/master/americas/files/2020-03/939958a474b8cc74bfa805b594fc0163/Horison_Tiered_Storage_2019.pdf

35. Edge Functions Vs Origin Compute: Latency and Data Gravity, https://itwplexus.co.uk/edge-functions-vs-origin-compute-latency-and-data-gravity#:~:text=The%20phenomenon%20of%20data%20gravity%20occurs%20as%20data%20volumes%20increase,to%20increased%20latency%20and%20costs.
36. SigMF: The Signal Metadata Format - Proceedings of the GNU Radio Conference, https://pubs.gnuradio.org/index.php/grcon/article/download/52/38/
37. SigMF, https://sigmf.org/
38. uSDR - Machine Learning (ML)-Driven RF Signal Detection | Crowd Supply, https://www.crowdsupply.com/wavelet-lab/usdr/updates/machine-learning-ml-driven-rf-signal-detection
39. Overview of VITA 49 VITA Radio Transport, https://www.vita.com/page-1855484
40. Stream support · Issue #60 · sigmf/SigMF - GitHub, https://github.com/sigmf/SigMF/issues/60
41. Modular Open Systems Approach (MOSA) - Defense Standardization Program, https://www.dsp.dla.mil/Programs/MOSA/
42. JADC2 Begins With Intelligence | AFCEA International, https://www.afcea.org/signal-media/jadc2-begins-intelligence
43. PEO IEW&S Artificial Intelligence and Software At Pace (AIS@P) Industry Day, https://peoiews.army.mil/wp-content/uploads/2025/01/AIS@P-MATOC_Industry-Day-1.7.25.pdf
44. Developing an AI/ML Operations Pipeline: Project Linchpin - PEO IEW&S, https://peoiews.army.mil/wp-content/uploads/2023/09/Project-Linchpin-Approved-for-Release-1.pdf
45. Edge Functions Vs Origin Compute: Latency and Data Gravity - ITW Plexus, https://itwplexus.co.uk/edge-functions-vs-origin-compute-latency-and-data-gravity
46. LEO Satellite Market Size, Share, Industry Trend Report, 2025 To 2030 - MarketsandMarkets, https://www.marketsandmarkets.com/Market-Reports/leo-satellite-market-252330251.html
47. Satellite Ground Station Market worth $82.72 billion by 2030 - Exclusive Report by MarketsandMarkets™ - PR Newswire, https://www.prnewswire.com/news-releases/satellite-ground-station-market-worth-82-72-billion-by-2030---exclusive-report-by-marketsandmarkets-302630223.html
48. On Challenges of Sixth-Generation (6G) Wireless Networks: A Comprehensive Survey of Requirements, Applications, and Security Issues - arXiv, https://arxiv.org/html/2206.00868v2
49. A Comprehensive Exploration of 6G Wireless Communication Technologies - MDPI, https://www.mdpi.com/2073-431X/14/1/15
50. Five things you should do to create an accurate on premises vs cloud comparison model, https://aws.amazon.com/blogs/aws-cloud-financial-management/five-things-you-should-do-to-create-an-accurate-on-premises-vs-cloud-comparison-model/

51. RF Challenge: The Data-Driven Radio Frequency Signal Separation Challenge | MIT, https://people.lids.mit.edu/yp/homepage/data/2024_rfchallenge_full.pdf
52. AI in RF Threat Detection: Opportunities and Challenges - CIO Influence, https://cioinfluence.com/security/ai-in-rf-threat-detection-opportunities-and-challenges/