**Secure Data Handling in Big Data Systems: A Practical Framework for Compliance**

**Introduction**

In today’s data-driven world, big data fuels innovation—but also raises serious concerns around data security and privacy compliance. With regulations like GDPR (EU) and HIPAA (US), businesses are now legally responsible for how they collect, store, and process sensitive information.

This report outlines a simple, practical framework to securely handle big data while ensuring compliance with global data protection laws.

* **Understanding Compliance Requirements**

1. **GDPR (General Data Protection Regulation)**

Applies to any company handling personal data of EU citizens.

Key requirements:

* User consent before data collection
* Right to access or delete data
* Data breach notification within 72 hours

1. **HIPAA (Health Insurance Portability and Accountability Act)**

Governs health-related data in the US

Requires:

* Strict protection of PHI (Protected Health Information)
* Access control, audit trails, and data encryption
* **Framework: Secure & Compliant Big Data System**

We break the framework into 6 key layers, each addressing specific risks and legal expectations:

**1. Data Ingestion Layer**

* Use secure protocols (HTTPS, SFTP) for incoming data.
* Mask or encrypt sensitive data (like names, phone numbers) at the point of entry.
* Collect only necessary information (data minimization).

**2. Data Storage Layer**

* Store data in encrypted form (AES-256 recommended).
* Use role-based access control (RBAC) to ensure only authorized users can view or edit sensitive data.
* Maintain versioned backups and log every access.

**3. Data Processing Layer**

* Run computations in isolated containers (e.g., Docker, Kubernetes).
* Use pseudonymization or anonymization techniques to protect personal identities.
* Avoid unnecessary duplication of raw data.

**4. Access Management**

* Enforce multi-factor authentication (MFA) for sensitive dashboards and tools.
* Build custom access policies (e.g., analysts get summary data, admins get full access).
* Track and store user access logs for auditing.

**5. Monitoring & Auditing**

* Monitor access in real-time using SIEM tools (e.g., Splunk, AWS CloudTrail).
* Run automated scans for vulnerabilities or unusual data access.
* Maintain a data audit trail for at least 6–12 months.

**6. Compliance Automation**

**Use automated tools to:**

* Handle user data requests (right to delete, access)
* Generate reports for data protection officers
* Monitor compliance checkpoints across systems
* **Real-World Use Case (Example)**
* Imagine a healthcare startup using Spark on AWS to process patient data:
* Ingestion is encrypted and tracked
* PHI is masked before analysis
* Access is restricted via AWS IAM roles
* Audit logs are stored in CloudWatch
* Users can request deletion of their data (GDPR Article 17)