## **Data Lake Use Case for Financial Risk & Compliance Data Lake on Google Cloud**

### **1. Data Sources**

* **Core Banking Systems:** Transactions, deposits, loans
* **Trading Platforms:** Market orders, positions, P&L
* **CRM/KYC Databases:** Customer profiles, risk scores
* **External Data:** Credit bureaus, market data feeds, sanction lists (via APIs)

### **2. Data Ingestion**

* **Batch Ingestion:**
  + **Cloud Storage Transfer Service** or **Transfer Appliance** for large data dumps
  + **Cloud Composer (Apache Airflow)** to orchestrate batch ETL jobs
* **Streaming Ingestion:**
  + **Pub/Sub** to ingest real-time transaction or trade streams
  + **Dataflow (Apache Beam)** for real-time data transformation and enrichment

### **3. Data Storage (Data Lake Core)**

* **Cloud Storage (GCS):**
  + Raw zone (landing zone)
  + Processed zone (curated, cleaned data)
  + Enriched zone (aggregated, analytics-ready data)
* **BigQuery:**
  + Analytical data warehouse on top of the lake
  + Stores flattened, query-optimized datasets

### **4. Data Processing & Analytics**

* **Dataflow:** Real-time stream processing (e.g., detect anomalous transactions)
* **Dataproc:** Spark/Hadoop-based batch processing (e.g., large-scale risk simulations)
* **BigQuery ML:**
  + Build fraud detection and credit risk models
  + Use SQL-based ML for faster iteration

### **5. Data Governance & Security**

* **IAM + VPC-SC:** Identity-based access control and network perimeter
* **Cloud DLP:** Masking/tokenization of PII (GDPR/CCPA/GLBA compliance)
* **Cloud Audit Logs:** Full traceability of data usage
* **Dataplex:** Unified data governance and metadata catalog

### **6. Consumption Layer**

* **Looker** or **Data Studio:** Interactive dashboards for compliance, audit, and risk teams
* **BigQuery BI Engine:** In-memory acceleration for low-latency dashboards
* **Notebooks (Vertex AI Workbench):** Data scientists explore and train advanced models

In response to increasing regulatory demands and the growing need for real-time risk insights, this initiative proposes the implementation of a modern, cloud-native **Financial Data Lake** on **Google Cloud Platform (GCP)**. The objective is to consolidate siloed data sources—including transactional systems, trading platforms, CRM/KYC databases, and third-party feeds—into a unified, governed environment that supports advanced analytics, machine learning, and regulatory reporting.

The proposed solution leverages GCP services such as **Cloud Storage** for scalable data lake storage, **Pub/Sub** and **Dataflow** for real-time ingestion and processing, and **BigQuery** as the analytical engine for generating insights and compliance reports. **BigQuery ML** and **Vertex AI** will be used to develop predictive models for fraud detection, credit risk scoring, and early-warning systems. Data governance and security are enforced using **IAM**, **Cloud DLP**, **VPC Service Controls**, and **Dataplex**, ensuring full compliance with GDPR, GLBA, and other financial regulations.

This architecture enables significant improvements in **compliance efficiency**, **risk transparency**, and **data-driven decision-making**, while reducing time and cost associated with legacy data management systems. It sets the foundation for future innovation in AI-driven finance, ultimately strengthening the institution’s resilience and regulatory posture.