Data Architecture Challenge Project

Module 1: Real-Time Market Data Ingestion

Choice: Time-Series Database + Stream Processor

1. **For High-Frequency Trading Feeds (5M+ records/sec) / For Bloomberg/Reuters API Data**
    - **Time-based Database**: **TimescaleDB** or **InfluxDB** (Time-Series Storage)
        - Optimized for time-stamped data with compression.
        - Supports sub-millisecond alerts via continuous aggregates.
2. **For Social Media Sentiment (JSON)**
    - **MongoDB** (Document Store)
        - Flexible schema for semi-structured JSON.
3. **Historical Data (5-Year Retention)**
    - **S3 / Data Lake** (Cold Storage)
        - Cost-effective for infrequently queried data.

Module 2: Client Portfolio Management

Choice: ACID-Compliant RDBMS + Document Store

1. **For Transaction Logs (ACID Required)**
    - **PostgreSQL** or **Oracle Database**
        - Strong consistency for GDPR-compliant audits.
        - Postgre SQL's JSONB handles semi-structured reports.
    - For analysis, use **Data Warehouse** to support, such as **SnowFlake**.
2. **For PDF/Excel Reports / For SEC Filings (Compliance)**
    - **Document DB: MongoDB GridFS** or **AWS S3**
        - Efficient binary storage with metadata tagging.
3. **For Audit Trails**
    - Use **OLAP** with Log system, such as **ELK**.

Module 3: Research & Analytics

Choice: ES + PostgreSQL + Data Lake

1. **For Equity Research Reports (Unstructured Text)**
    - **Elasticsearch with indexing on documents**
        - Blazing-fast full-text search with NLP capabilities (e.g., keyword highlighting, synonym matching).
        - Integrates with **Python/R** via Elasticsearch DSL.
2. **For Analyst Ratings (Relational Tables)**
    - **PostgreSQL**

        - Structured data with JOINs for cross-referencing ratings with market data.
        - Optional: Snowflake excels if analytics require massive scaling.
3. **For Alternative Data (Satellite Imagery, Shipping Logs)**
    - **Data Lake** (S3/MongoDB)
        - MongoDB for Schema flexibility for heterogeneous formats (e.g. JSON logs, image metadata).
        - S3 for Cost-effective storage for large binaries raw files (e.g. satellite images) with SQL querying.

Module 4: Regulatory Reporting Solution

Choice: Batch processing + Spark

1. **For Batch Processing (MiFID II/FATCA)**
    - **Snowflake**
        - Petabyte-scale columnar storage optimizes batch aggregation for CSV uploads.
        - Analysis: Handles FATCA's massive compliance logs efficiently.
    - **S3 with Hadoop**
        - Handles for batch processing for overnight regulatory submissions.
2. **Sub-10s queries on streaming data**
    - **Spark**
        - Handles for engine with In-memory processing
        - Real-time transaction for large data analysis and ETL pipeline

---

Module 5: Fraud Detection Solution

Choice: Graph Database + Kafka + Flink

1. **For Dark Web scraping feeds**
    - **Neo4j** (Graph Database)
        - Detects complex fraud networks via relationship analysis.
2. **For employee access patterns**
    - Real-time processing with different patterns.
3. **For Real-Time Anomaly Detection**
    - **Apache Kafka + Flink**
        - Processes AML alerts in real-time with ML models.
        - Use **Kinesis** as a substitute for Kafka.
        - Flink is for streaming processing on data from Kafka or Kenisis.
4. **For time-series data storage**
    - Time-series DB.

> Noted by Wu Wenhan.