Architecture Diagram Pack

Inside the Infrastructure Powering Next-Generation Market Research

RICCHE LTD

ricche.ai

Powered by the NVIDIA Accelerated Computing Stack

The Ricche Architecture Vision

Financial markets generate more data every day than most industries produce in a year. Tick-by-tick price movements, shifting order books, macroeconomic releases, satellite imagery, news sentiment -- the signals are everywhere, but the infrastructure to process them at scale has traditionally been locked behind the walls of the largest institutions.

Ricche is changing that. We are building purpose-built GPU-accelerated research infrastructure that brings institutional-grade computational power to quantitative market research. Every layer of our architecture is designed for one goal: turning raw market complexity into actionable research insights, faster than ever before.

> ***Our architecture will process market data through GPU-accelerated pipelines, targeting dramatically faster model training and sub-2ms inference latency.***

End-to-End Platform Architecture

Data flows left to right through five purpose-built layers. Each layer is independently scalable, fault-tolerant, and optimised for its specific workload -- from low-latency data ingestion to GPU-parallelised model training and real-time inference.

Data Ingestion Layer

Real-time streaming from global exchanges (NYSE, CME, LSE, Eurex, HKEX, SGX). Tick-level order books, OHLCV pricing, corporate actions, FX rates, and fixed income. Sub-second ingestion latency with exactly-once delivery guarantees. Automatic gap detection, reconnection, and data reconciliation.

GPU Compute & Storage

NVIDIA GPU clusters purpose-built for financial ML workloads. Distributed columnar storage (Parquet, Arrow) for petabyte-scale datasets. Smart tiering: hot data on NVMe, warm on SSD, cold on object storage. Auto-scaling compute allocation with workload-aware scheduling.

GPU Compute Infrastructure

At the heart of Ricche is a compute layer that would have been unimaginable five years ago. By building on the full NVIDIA accelerated computing stack, we compress research cycles from weeks to hours and enable experiments at a scale that was previously the exclusive domain of billion-dollar hedge funds.

NVIDIA CUDA **DGX SuperPOD** **NVIDIA GB200 NVL72** **NVIDIA H200** **BlueField-3 DPU** **TensorRT-LLM**

Triton Inference Server **Megatron-LM** **NeMo Framework** **NVIDIA NIM** **TensorRT** **RAPIDS cuDF** **cuML**

NVIDIA CUDA -- The Foundation

Every training run, every simulation, every feature computation will be CUDA-accelerated. Multi-GPU distributed training with mixed-precision (FP16/BF16) will deliver significant speedups over CPU-only infrastructure. Researchers will iterate on ideas in hours, not days. This is the engine that will make everything else possible.

RAPIDS -- Data at GPU Speed

Traditional data pipelines are CPU-bound bottlenecks. RAPIDS cuDF will eliminate them entirely. Feature engineering on financial time-series will run 10-50x faster than pandas. End-to-end GPU processing means data never leaves the GPU between preparation and training -- zero transfer overhead, maximum throughput.

TensorRT -- Production Speed

Research models are optimised for production inference through layer fusion, kernel auto-tuning, and INT8 quantisation. The result: sub-2ms inference latency for real-time signal generation. When markets move in milliseconds, every microsecond of latency reduction translates directly to better research outcomes.

NVIDIA NIM -- Production Deployment

NVIDIA NIM microservices will deliver optimised model inference with one-click deployment and dynamic scaling. Built-in model orchestration will handle multi-model serving with intelligent batching and GPU memory management. Zero-downtime version transitions with automatic rollback.

141 GB

HBM3e per H200

<2ms

Inference Latency

400G

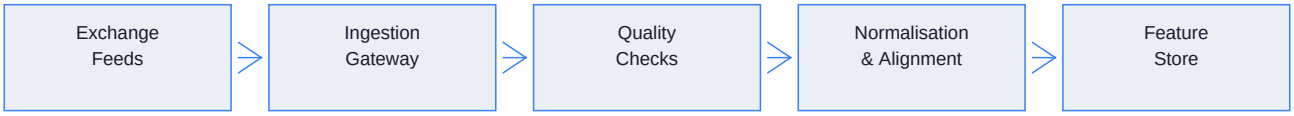
InfiniBand NDR

5

Asset Classes

Data Pipeline Architecture

Garbage in, garbage out -- in quantitative research, data quality is everything. Our multi-stage ingestion pipeline is engineered for institutional-grade reliability, because the best models in the world are worthless if they are trained on dirty data.



Global Market Coverage

- Equities: Developed and emerging markets across 30+ exchanges worldwide
- Futures & Commodities: Energy, metals, agriculture, financial futures (CME, ICE, LME)
- Options: Full listed chains with Greeks, implied volatility surfaces
- FX: Spot, forwards, crosses -- major and emerging market pairs
- Fixed Income: Government bonds, investment-grade credit, interest rate swaps
- Alternative Data: NLP-processed news, satellite imagery, social sentiment, web scraping

Real-Time Streaming

Event-driven pipelines with sub-second latency. Exactly-once delivery semantics ensure data integrity. Automatic reconnection and gap detection for exchange feeds. Designed to handle sustained throughput of millions of events per second.

GPU-Accelerated Batch Processing

Historical backfills and daily reconciliation powered by RAPIDS cuDF. Idempotent processing with full audit trails and lineage tracking. Parallel processing across GPU nodes for massive backfill operations. Data versioning will enable reproducible research at any point in time.

Research Workflow -- From Hypothesis to Discovery

Great research demands discipline. Every experiment at Ricche follows a structured workflow designed to eliminate noise, prevent overfitting, and ensure that only genuinely robust models progress to evaluation. This is not about running more experiments -- it is about running the right experiments, rigorously.

Stage 1: Research Hypothesis

Every project begins with a clearly defined hypothesis about market behaviour -- target instruments, timeframes, expected signal characteristics, and falsification criteria. Hypotheses are registered in our experiment tracking system before any code is written. This prevents the most dangerous trap in quantitative research: fitting a narrative to data after the fact.

Stage 2: Data Preparation

Relevant datasets are extracted from the feature store with strict temporal discipline -- no look-ahead bias, ever. GPU-accelerated feature engineering using RAPIDS produces training-ready datasets 50x faster than traditional tools. Proper train/validation/test splits with temporal walk-forward design ensure real-world applicability.

Stage 3: Model Development

Researchers develop models in CUDA-accelerated PyTorch environments on NVIDIA GPU clusters. Every hyperparameter, every metric, every model artifact is tracked automatically. Automated hyperparameter optimisation explores thousands of configurations. GPU speedups mean researchers test more ideas per week than traditional setups allow in a month.

Stage 4: Simulation Testing

Candidate models face rigorous stress-testing in GPU-parallelised simulation engines. Monte Carlo simulations evaluate behaviour across thousands of market scenarios including regime changes, liquidity crises, and black swan events. Agent-based simulations model market microstructure interactions. Only models that survive this gauntlet proceed.

Stage 5: Validation & Peer Review

Surviving models undergo out-of-sample validation, walk-forward testing, and formal peer review. Statistical significance tests (multiple hypothesis correction, bootstrap confidence intervals) guard against overfitting. Approved models enter controlled paper-trading evaluation with strict position-size constraints and continuous monitoring.

Current Technology Stack

We chose every tool in our stack for a reason. No resume-driven development, no hype cycles -- just the best technology for building world-class quantitative research infrastructure.

GPU & Accelerated Compute

NVIDIA CUDA **DGX SuperPOD** **NVIDIA GB200 NVL72** **NVIDIA H200** **BlueField-3 DPU** **TensorRT-LLM**

Triton Inference Server **NVIDIA NIM** **TensorRT** **RAPIDS cuDF** **cuML**

Machine Learning & Research

PyTorch JAX FlashAttention-3 vLLM Ray W&B

Data Engineering

KDB+/q Redpanda Apache Iceberg Polars DuckDB ClickHouse

Platform & Infrastructure

Kubernetes Terraform Temporal InfiniBand NDR 400G Rust CUDA C++ Python

Security & Observability

AES-256 Encryption at Rest TLS 1.3 Role-Based Access Control eBPF OpenTelemetry Prometheus Grafana Audit Trail

Infrastructure-as-code (Terraform) will ensure every deployment is reproducible. Temporal will orchestrate complex ML workflows with built-in retry and observability. Ray will distribute training and hyperparameter search across GPU clusters. Polars and DuckDB will handle analytical queries at blazing speed. ClickHouse will power sub-second OLAP queries on financial time-series. Apache Iceberg will provide the data lakehouse layer. Redpanda replaces Kafka with 10x lower latency and zero JVM overhead. Rust will power latency-critical paths.

Build the Future of Market Research With Us

We are looking for partners, collaborators, and visionary researchers who share our belief that the future of financial markets belongs to those with the best infrastructure.

info@ricche.ai | ricche.ai