SEVENS TREE
GLOBAL

Sevens Tree Global — Product + Technology

1) Product portfolio and AI fit

1. Prop Trading (2026)
   - Funded accounts ($50k/$100k), evaluation, scaling.
   - AI use: trader scoring, risk telemetry, anomaly flags, coaching copilots, execution quality monitoring.
2. Retail Brokerage (2026)
   - Multi-asset DMA, fractional equities, FX, futures, crypto (jurisdiction-bound).
   - AI use: trade-assist copilots, best-execution routing hints, fraud/AML triage, CX automation.
3. Automated AI Portfolios (2026)
   - Quant strategies, thematic baskets, volatility-aware allocation.
   - AI use: regime detection, signal blending, RL for risk sizing, continuous model governance.
4. B2B/White-Label
   - Risk engine, surveillance, data feeds, OMS/EMS adapters, analytics APIs.

2) Reference architecture (high level)

- Presentation: Web, iOS, Android.
- Gateway: GraphQL + REST, OAuth2/OIDC, rate limiting, ABAC.
- Core: OMS/EMS, Risk Engine, Pricing, Accounts/Ledger, Compliance, Data Platform, Model Service Mesh, Workflow Orchestrator.
- Data: Event bus, Feature store, Vector DB, Lake/Lakehouse, Tick store, Secrets/HSM.
- ML: Offline training, online inference, monitoring, governance, registry.
- Infra: Kubernetes, autoscaling, GPU pools, IaC, service mesh, observability.

3) Systems required by product line

3.1 OMS/EMS and market connectivity

- Protocols: FIX 4.4/5.0, OUCH/ITCH where applicable, crypto exchange REST/WebSocket.
- Routing: Smart Order Router, internal crossing, LP selection, throttle control.
- Latency: colocation for listed markets, regionally sharded for FX/crypto.
- Tools: quickfix engines, bespoke Python/Go routers, kernel-bypass NICs if truly needed.

3.2 Risk and capital allocation

- Real-time: per-account drawdown, margin, kill-switch, circuit breakers.
- Portfolio: ES/ETL, factor exposure, stress ladders, liquidity haircut.
- Intraday models: drift detectors, change-point detection, adversarial trader patterns.
- Allocation: Bayesian or RL-driven scaling subject to concentration and liquidity caps.
- Interfaces: gRPC for low latency, Kafka topics for audit.

3.3 Pricing, data, and reference services

- Tick plant: normalized L1/L2, corporate actions, calendars, entitlements.
- Derived data: implied vols, greeks, synthetic crosses, TWAP/VWAP.
- Time sync: PTP/NTP hardening.
- Storage: columnar tick store + object lake; retention tiers.

3.4 Compliance and surveillance

- Trade surveillance: layering/spoofing/banging-the-close, abusive messaging patterns.
- AML/KYC: sanctions, PEP, adverse media, device fingerprinting.
- Reporting: IIROC/CSA obligations, CAT/EMIR equivalents where applicable.
- Explainability: model cards, feature attributions, decision logs.

3.5 Client platform

- Account opening: eKYC, liveness, document extraction, risk questionnaire, suitability.
- Portfolio UI: P/L, risk, factors, fills, slippage, statements.
- Copilots: task-scoped agents for onboarding, funding, and education. Guarded prompts.
- Accessibility and localization baseline.

4) AI stack (training → inference → governance)

4.1 Data and features

- Lakehouse: transactional tables for reproducibility.
- Feature store: offline/online parity, point-in-time joins, drift stats.
- Vector DB: RAG for policy, playbooks, runbooks, market microstructure notes.
- Labeling: weak supervision for surveillance; human-in-the-loop queues.

4.2 Models

- Time series: regime switchers, HMMs, temporal CNN/Transformers, volatility forecasting.
- Execution: microprice, queue position, short-term alpha, order placement policy via RL.
- NLP: entity/relation extraction for news, embeddings for retrieval, sentiment calibrated to realized vol.
- Tabular: gradient boosting for fraud/AML and suitability.
- LLMs: instruction-tuned finance assistants with tools for KYC, tickets, KB search. No direct trade autonomy.

4.3 Inference and safety

- Online inference: low-latency gRPC microservices, GPU/CPU mix, canary release.
- Safety layers: policy sandbox, role-based tools, allowed-actions lists, output attestation, prompt injection shields.
- Human oversight: kill-switches, four-eyes on model config changes.

4.4 Monitoring and governance

- Metrics: AUROC/PR for surveillance, hit-ratios for signals, slippage vs benchmarks, regret for RL.
- Drift: input, prediction, and data quality monitors.
- Explainability: SHAP or permutation importance for tabular; probative tests for sequence models.
- Registry: signed model artifacts, lineage, approvals, rollback.

5) Concrete tool options

- Orchestration: Airflow, Dagster, or Prefect.
- Event bus: Kafka or Redpanda.
- Lake/Lakehouse: S3/GCS + Iceberg/Delta.
- Feature store: Feast or Tecton.
- Vector DB: pgvector, Milvus, or Pinecone (jurisdiction matters).
- Model serving: BentoML, Ray Serve, KServe.
- LLM serving: vLLM/TensorRT-LLM for open models; managed endpoints if required.
- Training: PyTorch + Lightning/Accelerate; XGBoost/CatBoost for tabular; RLlib for RL.
- Observability: OpenTelemetry, Prometheus, Tempo/Jaeger, Loki.
- SecDevOps: Terraform, Crossplane, Vault + HSM, SOPS, OPA/Gatekeeper, Kyverno.

- CI/CD: GitHub Actions or GitLab CI with supply-chain signing (Sigstore/Cosign).
- Identity: Keycloak/Auth0, SCIM, device posture checks.
- KYC/AML vendors: document + liveness + sanctions APIs; keep swappable behind a facade.
- Payments: compliant PSPs, open-banking rails where available.
- Testing: property-based tests, market simulators, chaos engineering for OMS/EMS.

6) Security model

- Data: field-level encryption, tokenization for PII, envelope keys in HSM.
- Access: least-privilege, JIT elevation, strong approvals, hardware keys for prod.
- Network: mTLS everywhere, private link to vendors, egress allow-lists.
- App: threat modeling, code scanning, SAST/DAST, SBOMs, reproducible builds.
- Incident: SOAR playbooks, tabletop drills, immutable logs, breach notification timers.

7) Performance SLOs

- OMS/EMS: ≤1 ms median enrichment, ≤250 µs in-DC hops, ≥99.95% availability.
- Risk checks: ≤10 ms from event to limit decision.
- Pricing: ≤50 ms tick propagation p95.
- Inference: ≤20 ms p95 for surveillance; ≤5 ms for execution hints where colocated.
- Client APIs: 200–400 ms p95; real-time streams loss <0.1%.

8) Data contracts and lineage

- Contracts per topic/table with schema versioning and SLAs.
- Lineage from raw → features → model → decision, signed manifests.
- Retention: hot, warm, cold with legal hold paths.

9) Build vs buy (fast path)

- Build: risk engine, allocation logic, signal research infra, SOR, analytics, governance glue.
- Buy: KYC/AML, sanctions, market data feeds, clearing/custody, some OMS components if time-to-market dominates.
- Hybrid: surveillance (rules + ML), payments, client comms.

10) Compliance runway (Canada first)

- Policies: suitability, conflicts, best-execution, record-keeping, complaints.
- Controls: trade/comm surveillance, marketing review, books and records retention.

- Automation: scheduled regulatory reports, audit trails, attestations, model approvals.
- Privacy: data residency options, DPIAs, deletion workflows.

11) Delivery plan and milestones

- Q4 2025: MVP risk engine + OMS simulator, feature store live, data contracts v1.
- Q1 2026: Prop trading GA, surveillance v1, trader copilot v1, SOC2 Type I readiness.
- Q2 2026: Brokerage onboarding, SOR to live venues/LPs, CAT/IIROC reporting automations.
- Q3 2026: Automated portfolios v1, RL risk sizing shadow mode, SOC2 Type II start.
- 2027: EU/Asia licensing tracks, latency domains expansion, B2B risk engine.

12) Procurement checklist (minimum viable stack)

- Cloud accounts with org policies, GPU quota, KMS/HSM.
- Managed Kubernetes, private networking, service mesh.
- Kafka cluster, Lakehouse storage, Feature store, Vector DB.
- CI/CD with signing, artifact registry, IaC repos.
- Secrets + vaulting, PAM, SIEM/SOAR.
- Market data vendors, exchange/LP connections, KYC/AML vendor.
- Monitoring stack with traces, metrics, logs, model telemetry.
- DR site runbooks and RTO/RPO tests.

13) KPIs that matter

- Prop: pass-rate, risk breaches per 1k sessions, payout ratio, alpha decay half-life.
- Brokerage: cost-to-serve per account, execution quality vs NBBO/benchmarks, churn, fraud rate.
- AI: model lift vs rules, drift alerts per week, time-to-rollback, inference cost per 1k calls.
- Ops: MTTR, change failure rate, on-call pages per week.

14) Appendices (ready to expand)

- Data schemas: trades, orders, positions, ticks, KYC, funding, surveillance alerts.
- Model cards: purpose, training data, metrics, risks, fallback, retirement plan.
- Runbooks: SOR outage, LP failover, model revert, exchange halt, data vendor drop.

Technical Glossary (Points 1–14)

1) Product Portfolio and AI Fit

- Prop Trading: Proprietary trading where a firm funds traders and shares profits.
- Funded accounts ($50k/$100k): Predefined trading capital allocations.
- Evaluation: Testing period to qualify for real capital.
- Scaling: Increasing trader capital based on performance.
- Risk telemetry: Real-time monitoring of account risk data.
- Anomaly flags: Alerts for unusual/risky trades.
- Copilots: AI assistants helping traders.
- Retail Brokerage: Platform for individuals to trade securities.
- Multi-asset DMA: Direct Market Access across asset classes.
- Fractional equities: Buying parts of shares.
- Jurisdiction-bound: Regulated by local authority.
- AML (Anti-Money Laundering): Laws to prevent illegal money flows.
- CX automation: Customer experience handled by AI/chatbots.
- Quant strategies: Algorithmic, data-driven investment methods.
- Thematic baskets: Securities grouped by themes (e.g., green energy).
- Volatility-aware allocation: Adjusting portfolio weight based on risk.
- Regime detection: Identifying market conditions.
- RL (Reinforcement Learning): AI optimizing by trial and error.
- Model governance: Compliance and oversight of AI systems.
- B2B/White-Label: Backend solutions resold under another brand.
- Risk engine: Core system enforcing capital/risk limits.
- Surveillance: Monitoring trades for fraud/manipulation.
- OMS/EMS: Order Management System / Execution Management System.
- APIs: Software connectors for external apps.

2) Reference Architecture

- GraphQL/REST: API styles (GraphQL = flexible queries, REST = standard endpoints).
- OAuth2/OIDC: Authentication/identity protocols.
- Rate limiting: Restricting request frequency.
- ABAC: Attribute-Based Access Control.
- Ledger: Record of balances and transactions.
- Model Service Mesh: AI deployment infrastructure.
- Workflow Orchestrator: Automation of business processes.

- Event bus: Messaging backbone (e.g., Kafka).
- Feature store: Repository for ML variables.
- Vector DB: Embedding database for search/AI.
- Lakehouse: Data lake + warehouse hybrid.
- Tick store: Tick-by-tick market data storage.
- Secrets/HSM: Secure key storage (Hardware Security Module).
- Inference: AI prediction at runtime.
- IaC (Infrastructure as Code): Automated infrastructure setup.
- Observability: Monitoring logs, metrics, traces.

3) Systems by Product Line

Market connectivity

- FIX 4.4/5.0: Standard trading protocol.
- OUCH/ITCH: Nasdaq trading protocols.
- WebSocket: Real-time data connection.
- Smart Order Router (SOR): Finds best venue for trades.
- LP (Liquidity Provider): Firms providing market liquidity.
- Colocation: Hosting servers near exchanges.
- Kernel-bypass NICs: Low-latency network cards.

Risk/capital allocation

- Drawdown: Drop from peak balance.
- Circuit breakers: Trade halts under stress.
- ES/ETL: Expected Shortfall / Expected Tail Loss.
- Factor exposure: Portfolio’s sensitivity to factors (e.g., interest rates).
- Bayesian scaling: Probability-based capital adjustments.
- gRPC: High-performance communication protocol.
- Kafka topics: Streams of messages.

Pricing/data

- L1/L2: Level 1 = top of book, Level 2 = full depth.
- Greeks: Options risk metrics (delta, gamma, etc.).
- TWAP/VWAP: Execution benchmarks (time/volume-weighted).

- PTP/NTP: Precision/Network Time Protocol.

Compliance/surveillance

- Layering/spoofing: Market manipulation tactics.
- PEP: Politically Exposed Person.
- CAT/EMIR: US/EU regulatory frameworks.
- Model cards: Standardized AI model documentation.

Client platform

- eKYC: Electronic Know Your Customer.
- Liveness: Detecting a real person in ID verification.
- Slippage: Price difference between expected vs executed trade.

4) AI Stack

Data/features

- Lakehouse transactional tables: Versioned AI training data.
- RAG (Retrieval-Augmented Generation): AI retrieves before answering.
- Weak supervision: Imperfect but scalable labeling.

Models

- HMMs (Hidden Markov Models): Probabilistic sequence models.
- Microprice: Order book fair-value measure.
- Gradient boosting: Tree-based ML method.
- Instruction-tuned LLMs: Finance-trained large language models.

Inference/safety

- Canary release: Gradual rollout of new model.
- Prompt injection shields: Protecting LLMs from malicious prompts.
- Four-eyes principle: Dual approvals on sensitive changes.

Monitoring/governance

- AUROC/PR: Model performance metrics.
- SHAP: ML explainability method.

5) Concrete Tool Options

- Airflow, Dagster, Prefect: Workflow orchestration.
- Kafka/Redpanda: Event streaming.
- Feast/Tecton: Feature stores.
- Milvus/Pinecone/pgvector: Vector databases.
- BentoML, KServe: Model serving tools.
- vLLM, TensorRT-LLM: LLM deployment optimizers.
- PyTorch, XGBoost, CatBoost, RLlib: ML frameworks.
- Prometheus, Jaeger, OpenTelemetry: Observability tools.
- Terraform, Crossplane: IaC tools.
- OPA, Kyverno: Kubernetes policy engines.
- Sigstore/Cosign: Supply chain signing.
- Keycloak/Auth0: Identity providers.
- Chaos engineering: Testing system resilience.

6) Security Model

- PII: Personally Identifiable Information.
- Envelope keys: Encryption keys wrapped in master key.
- mTLS: Mutual TLS encryption.
- SAST/DAST: Static/Dynamic application security tests.
- SBOMs: Software Bill of Materials.
- SOAR: Security Orchestration, Automation, Response.

7) Performance SLOs

- ≤1 ms median enrichment: Speed benchmark per order.
- p95: 95th percentile latency measure.
- MTTR: Mean Time to Recovery.

8) Data Contracts and Lineage

- Schema versioning: Tracking changes in data format.
- Lineage: Provenance of data from raw → model → decision.
- Retention tiers: Hot/warm/cold storage strategies.

9) Build vs Buy

- Build: In-house core (risk, allocation, research infra).
- Buy: Outsourced functions (KYC/AML, data feeds, custody).

- Hybrid: Blend (e.g., ML surveillance + rules).

10) Compliance Runway

- Suitability: Matching products to client profile.
- Best-execution: Ensuring clients get market-best prices.
- Books/records retention: Mandatory archival.
- Attestations: Compliance sign-offs.
- Data residency: Data stored in required region.
- DPIAs: Data Protection Impact Assessments.

11) Delivery Plan

- MVP: Minimum Viable Product.
- General Availability (GA): Full release.
- SOC2 Type I/II: Security certification (design vs operation).
- CAT/IIROC reporting: Canadian compliance filings.
- Shadow mode: Running ML model in background, not live.

12) Procurement Checklist

- GPU quota: AI compute allocation.
- KMS/HSM: Encryption key management.
- PAM: Privileged Access Management.
- SIEM: Security Information and Event Management.
- DR site: Disaster Recovery site.
- RTO/RPO: Recovery time / data loss tolerances.

13) KPIs

Prop

- Pass-rate: Trader qualification ratio.
- Alpha decay half-life: Duration of strategy edge.

Brokerage

- NBBO: National Best Bid and Offer.
- Churn: Customer attrition.

AI

- Drift alerts: Model/data mismatch notifications.
- Time-to-rollback: Speed of undoing bad model.

Ops

- Change failure rate: % of changes causing incidents.
- On-call pages: Alerts received by engineers.

14) Appendices

- Data schemas: Structure of trade/order datasets.
- Model cards: Documentation of model details/risks.
- Runbooks: Emergency playbooks for failures.