

The Ultimate in Price/Performance for Apache Spark™ ETL Pipelines

Spending Too Much on ETL?

More than 50% of cloud data platform spend is on ETL pipelines—not analytics. What if you could cut that in half?

- ETL represents the **largest line item** in data platform compute bills
- Most platforms treat lakehouse formats as “**just another file format**”
- Generic Spark runtimes **lack deep lakehouse integration**
- Data engineers **spend more time tuning infrastructure** than delivering insights

The Onehouse Advantage: Purpose-Built for Lakehouse Workloads

Onehouse Compute Runtime (OCR): Beyond Generic Spark

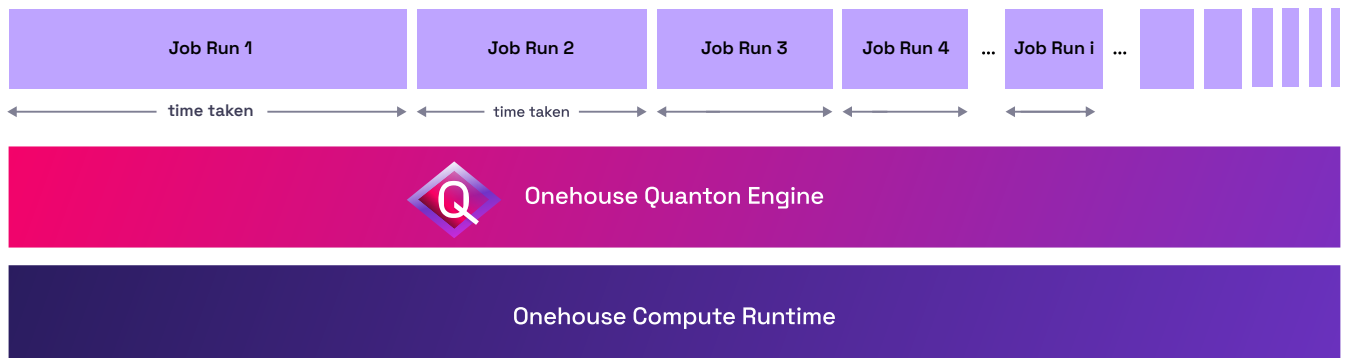
Runtime designed explicitly for lakehouse challenges

- **Up to 30x query performance** and 10x faster writes vs. open source
- **Adaptive Workload Optimizer:** Auto-tunes for fastest writes, reads, or balanced performance
- **Serverless Compute Manager:** Compute infrastructure optimized for the most challenging lakehouse workloads.
- **High-Performance Lakehouse I/O:** A radical rethinking of foundational lakehouse operations.

Quanton Query Engine: The ETL-Aware Execution Engine

Built for repetitive ETL patterns, not one-off queries

- **2-3x better price/performance** than leading market solutions
- **Minimal quantum processing:** Only processes differential work needed per ETL run
- Scales **proportional to work performed**, not table size



Guaranteed Results: Risk-Free Migration

Immediate Cost Savings (Guaranteed)

- **>50% reduction in Spark infrastructure costs** for existing Hudi users
- **>50% reduction in ETL spend** with zero code changes
- **Point existing** SQL transformations and dbt models to Onehouse
- **Simple lift-and-shift:** Use the same Spark/SQL interfaces you know

Zero-Friction Developer Experience

- **100% Apache Spark compatibility:** Migrate on/off platform easily
- Existing JARs, Python files, and spark-submit parameters **work unchanged**
- Built-in SQL Editor highlights **lakehouse-specific syntax**
- **Native** Airflow operators and dbt integrations

True Openness: Future-Proof Your Data Stack

Open Engines™: Best-of-Breed Compute Choice

Finally flip defaults to “open” for both data AND compute

- **Deploy open source engines** (Flink, Trino, Ray) with one click
- **>10x lower self-management costs** vs. DIY open source deployments
- **Free cluster included:** Scale up to 20 OCU's with unlimited queries
- **Data-first architecture:** Pick optimal engines for each workload

Universal Data Interoperability

- **All table formats:** Apache XTable™ (incubating) integration for compatibility across Apache Hudi™, Apache Iceberg™, Delta Lake
- **OneSync™ multi-catalog:** Sync across all major catalogs automatically
- **Never locked in:** Your data remains portable and engine-agnostic

Operational Excellence: Serverless Without Compromise

Managed Infrastructure in Your VPC

- **Serverless experience** with **BYOC security:** Best of both worlds
- Auto-scaling with deep lakehouse workload intelligence
- **Complete network control:** Data never leaves your compliance boundary
- **Leverage existing** cloud discounts, reserved instances, spot nodes

Expert-Level Observability & Support

- **E2E observability:** Engine layer down through storage
- **Cost transparency:** Per-job tracking and budget controls
- **Expert support:** Team that built planet-scale lakehouses at Uber
- **Built-in** Spark UI, comprehensive logs, performance bottleneck detection

Proven at Scale: Real Customer Impact

Production-Proven Results

“ With automated scaling and resources that adapt to our workloads, Onehouse helps us dedicate our teams to building out our core platform differentiators rather than keeping the data stack continuously optimized. ”

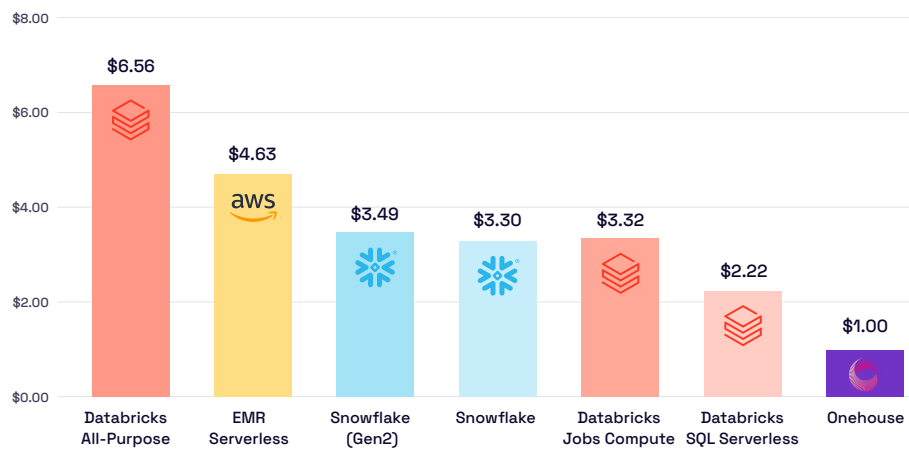
— Emil Emilov, Principal Software Engineer, Conductor

Use Cases Powered by Onehouse:

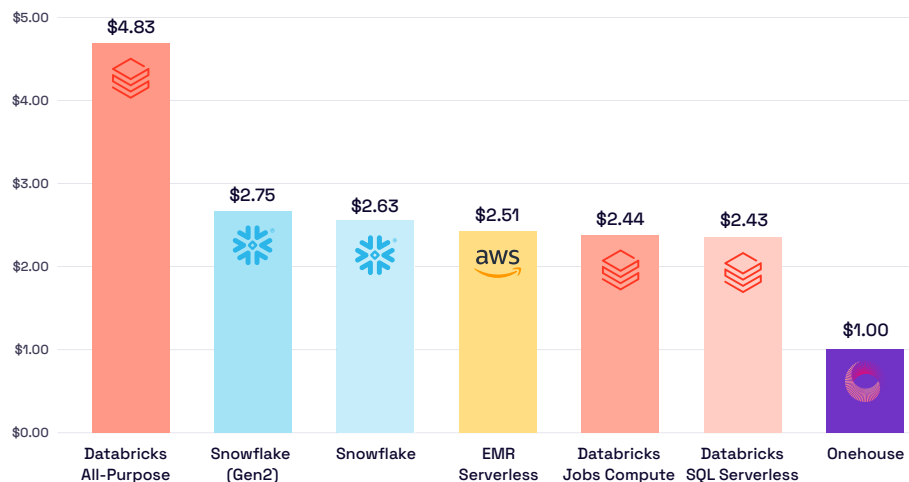
- Near-real-time clickstream and telecom analytics
- Blockchain ledger transactions
- IoT utility sensor monitoring
- Retail supply chains
- PB-scale marketing analytics platforms

Put to the test against the best

Ingest Price/Performance¹



ETL Price/Performance²



The Bottom Line

Onehouse is the only platform purpose-built for lakehouse ETL workloads, delivering guaranteed cost savings while keeping your data open and portable. Stop overpaying for generic compute—get the specialized lakehouse runtime your pipelines deserve.

¹ Benchmark run May 20, 2025, using Amazon EMR Serverless 7.5.

² Benchmark run December 2, 2025, using Amazon EMR Serverless 7.12.

Ready to Cut Your ETL Costs in Half?

Meet with us for your **FREE** Cost Analysis [▶](#)

