Optimizing Query Performance in Cloud Data Warehousing

A Comparative Analysis of Microsoft Azure Synapse, Amazon Redshift, and Oracle Autonomous Data Warehouse

By Cruz Bello

Research Context & Problem

Context

- Cloud data warehouses now essential for modern analytics
- SMEs face challenges selecting optimal platforms
- Performance varies significantly across platforms

Research Gap

- Conflicting benchmark results in literature
- No neutral comparison under identical conditions
- Limited guidance for SME adoption decisions

Research Objectives

1. Benchmark three leading platforms using TPC-DS and TPC-H at 10GB scale
2. Evaluate platform-specific optimization strategies
3. Develop evidence-based decision framework for SMEs

Research Context & Problem

Context

- Cloud data warehouses now essential for modern analytics
- SMEs face challenges selecting optimal platforms
- Performance varies significantly across platforms

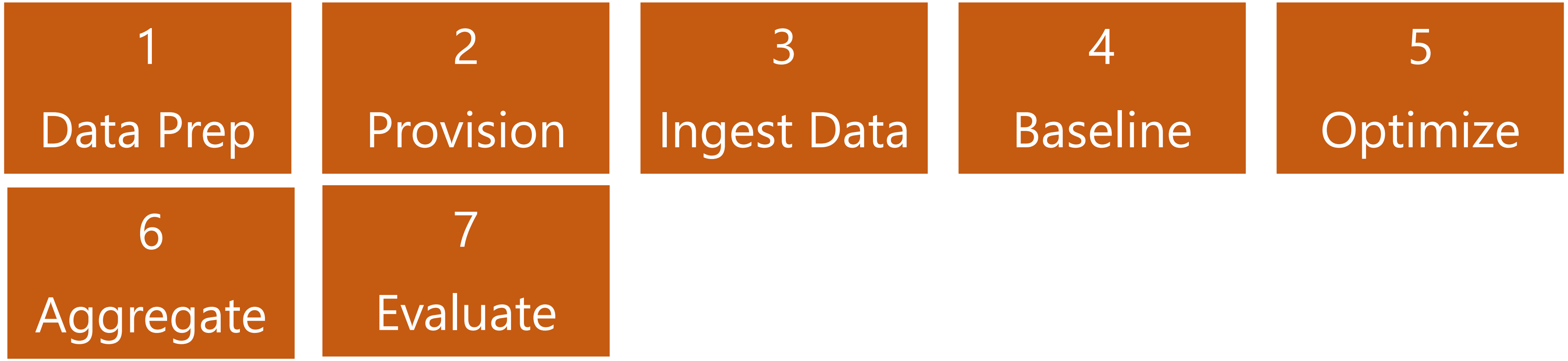
Research Gap

- Conflicting benchmark results in literature
- No neutral comparison under identical conditions
- Limited guidance for SME adoption decisions

Research Objectives

1. Benchmark three leading platforms using TPC-DS and TPC-H at 10GB scale
2. Evaluate platform-specific optimization strategies
3. Develop evidence-based decision framework for SMEs

Research Methodology
Seven-Phase Experimental Pipeline



1 Data Prep

2 Provision

3 Ingest Data

4 Baseline

5 Optimize

6 Aggregate

7 Evaluate

Benchmarks

- TPC-DS: 99 queries, 24 tables
- TPC-H: 22 queries, 8 tables
- Scale Factor: 10GB (SME-representative)

Statistical Analysis

- Kruskal-Wallis H-test (α=0.05) - Tests whether there are statistically significant differences in median query latency across the three platforms and complexity level
- Dunn's post-hoc comparisons - After Kruskal-Wallis detects differences, Dunn's test identifies which specific pairs of platforms differ from each other.
- Cliff's Delta effect sizes - Quantifies how large the performance difference is between platforms, not just whether it's statistically significant.
- n=20 iterations per query

Platform Configurations

Strategy: Minimum production-tier matching to reflect real-world SME adoption constraints

Platform	Configuration	Cost/Hour	Justification
Azure Synapse	DW200c	$1.60	Vendor minimum Gen2 production tier
Amazon Redshift	2 × RA3.large	$1.08	Minimum RA3 cluster (32GB per node)
Oracle ADW	4 ECPUs	$1.04	Minimum ADW ECPU configuration

Baseline Performance Results

Baseline Performance Results
TPC-DS Performance

Platform	p50 (s)	p99 (s)	QPH
Oracle ADW	0.014	0.225	148,072
Redshift	0.019	0.229	191,168
Synapse	4.080	27.342	2,414

Baseline Performance Results
TPC-H Performance

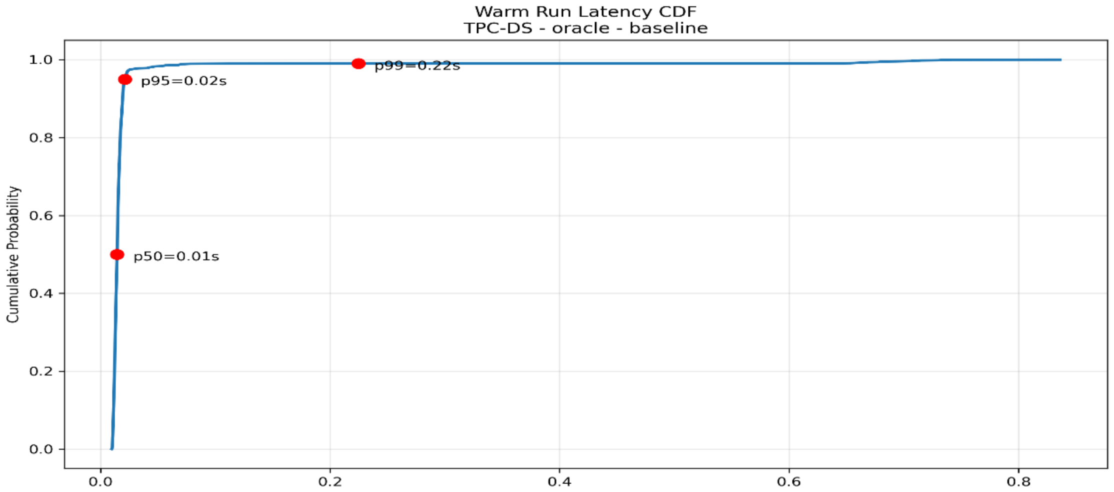
Platform	p50 (s)	p99 (s)	QPH
Oracle ADW	0.016	9.221	21,189
Redshift	0.018	3.271	71,338
Synapse	5.389	22.495	2,110

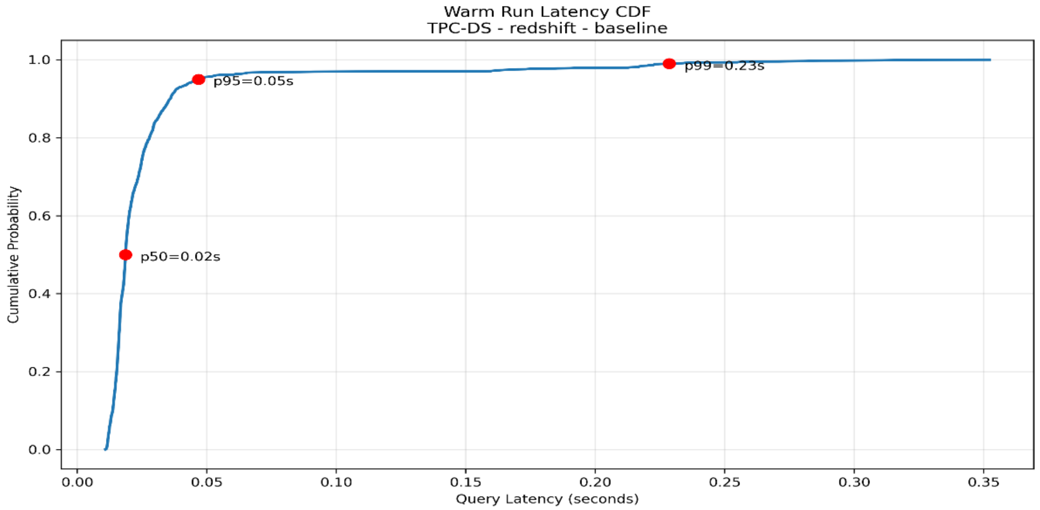
Oracle ADW
Lowest median latency, exceptional consistency

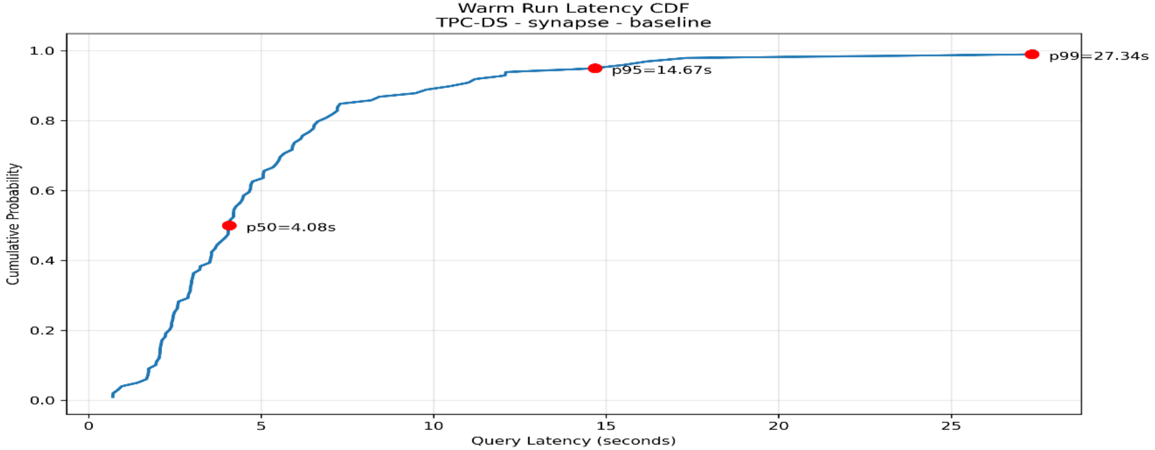
Redshift
Competitive performance, superior TPC-H throughput

Synapse
DW200c inadequate for production workloads

Baseline Performance latency Results







Optimization Intervention Results

1. Compression Tuning

-0.97%
Oracle ADW (TPC-DS)

-3.38%
Redshift (TPC-DS)

-49.23%
Synapse (TPC-H)

2. Distribution Key and Partition Optimization

-3.98%
Redshift (TPC-DS)
Top: query 20 (-30.90%)

+123.01%
Synapse (TPC-DS)
79 queries degraded

+1.32%
Oracle Partitioning (TPC-DS)
Mixed results (query 34 -98.12%,query10 +37.00%)

3. Materialized Views

-0.61%
Oracle (TPC-DS)

-9.60%
Redshift (TPC-H)

+1.16%
Synapse (TPC-DS)

Key Finding: Optimization effectiveness highly variable and platform-dependent. Resource constraints (Synapse DW200c) can negate traditional optimization strategies.

Concurrency Scaling Performance

TPC-DS Throughput Scaling (1 → 20 concurrent sessions)

12.2×	1.8×	1.4×
Oracle ADW	Redshift	Synapse
148K → 1.8M QPH	191K → 354K QPH	2.1K → 3.0K QPH
Latency: 0.014s → 0.020s	Latency: 0.019s → 0.201s	Latency: 5.6s → 63.6s

Oracle ADW

- Superlinear throughput scaling
- Minimal latency degradation (1.4×)
- Autonomous resource management
- Excellent p99 stability (0.108s → 0.153s)

Redshift & Synapse

- Redshift: p99 explosion (0.228s → 1.548s)
- Workload queue limitations evident
- Synapse: 11.3× median latency increase
- Resource constraints prevent scaling

Implication: Autonomous optimization (Oracle) provides superior multi-user performance. Manual tuning required for Redshift. Synapse requires higher tier provisioning.

Statistical Analysis

Statistical Analysis with Optimizations (TPC-DS):

Kruskal-Wallis H-test: Would still reject H_0 ($p \approx 0.000$). The gap between Synapse and others remains enormous.

Dunn's Post-Hoc Comparisons:

- **Synapse vs. Oracle ADW:** Extremely significant ($p \approx 0.000$)
- **Synapse vs. Redshift:** Extremely significant ($p \approx 0.000$)
- **Oracle ADW vs. Redshift:** Now more competitive. Oracle has better p50 and p99, but Redshift has higher QPH.

Cliff's Delta Effect Sizes:

- **Synapse vs. Oracle ADW:** $\delta \approx +1.0$ (Large)
- **Synapse vs. Redshift:** $\delta \approx +1.0$ (Large)
- **Oracle ADW vs. Redshift: δ** ≈ -0.15 (Negligible/Small) - Very similar performance profiles

Key Research Findings

1. Performance Disparities
Oracle ADW demonstrated lowest median latency (0.014s TPC-DS) and exceptional consistency. 61× throughput advantage over Synapse's entry-level configuration.

2. Optimization Effectiveness Varies
Compression: 3-49% improvements. Distribution keys: -30% to +123% variance. Materialized views: 0.6-68% per-query effects. Results highly dependent on platform maturity and resource provisioning.

3. Resource Provisioning Critical
Azure Synapse DW200c proved inadequate for production workloads. Traditional optimization strategies can degrade performance under resource constraints.

4. Autonomous vs Manual Tuning
Oracle's autonomous optimization provided consistent performance without manual intervention. Redshift and Synapse require DBA expertise for optimization.

5. Workload Sensitivity
Platform performance varies by query complexity. Oracle excels at simple queries (caching), Redshift competitive on moderate joins, all platforms struggle with complex nested subqueries at 10GB scale.

Platform-Specific Recommendations

Oracle ADW	**Amazon Redshift**	**Azure Synapse**
Best For:	Best For:	Best For:
•Operational simplicity priority	•Technical teams	•Azure ecosystem users
•Multi-user analytics	•Cost-conscious deployments	•Power BI integration
•Limited DBA expertise	•OLAP-heavy workloads	•Adequate provisioning (≥DW500c)
Considerations:	Considerations:	Considerations:
•Higher cost ($1.04/hr base)	•Requires DBA tuning	•DW200c inadequate
•Provision 8+ ECPUs for TB-scale	•Distribution key critical	•Requires tier planning
•Excellent consistency	•Best cost efficiency	•Schema design critical

Decision Framework

- Operational simplicity priority: Oracle ADW (**autonomous optimization, compression yields modest improvements; check heavy outliers**)
- Cost optimization with technical expertise: Amazon Redshift (**Distribution key produced the largest overall improvement in your runs, mean −3.98%**)
- Azure ecosystem lock-in: Synapse with ≥DW500c provisioning (**Tuning had mixed effects**)
- High concurrency requirements: Oracle ADW (**12.2× scaling**)

Research Contributions
•✓ Neutral tri-platform comparison under controlled conditions
•✓ Explicit cold/warm run separation methodology
•✓ Marginal attribution of optimization effects
•✓ Telemetry-driven root cause analysis
•✓ Evidence-based SME decision framework

Conclusions & Future Directions

Key Takeaways

Platform Choice: No universal winner; depends on workload and expertise

Provisioning: Entry-level tiers may prove false economies

Optimization: Highly context-dependent; test before deploying

Limitations
- 10GB scale factor (SME-focused but limits TB/PB generalizability)
- Single region deployment (US-East)
- Temporal validity (October 2025 snapshot)
- Resource imbalance across configurations

Future Research Directions
- Larger scale factors (SF100-SF1000) for enterprise validation
- Multi-region performance and cost analysis
- Real-world workload traces beyond synthetic benchmarks
- Cost-per-query economic modeling
- Synapse Spark pool evaluation for complex analytics

(Click the arrow when in Slide Show mode)