HYPHA DISTRIBUTED COMPUTE

COBOL Legacy Schema Inference & Extraction Benchmark

April 2026

Executive Summary

Hypha is a hardware-agnostic distributed compute platform that schedules operations across heterogeneous commodity hardware without containers, VMs, or specialist infrastructure. This paper documents a verified benchmark demonstrating Hypha's capability to infer the schema of unknown fixed-width COBOL data, validate structural integrity, and extract 100,000,000 records to structured CSV format — without access to source code, copybooks, or documentation.

The complete pipeline was built and executed on commodity hardware located in Danville, Virginia. All results are verified and reproducible.

The Problem: Legacy COBOL Data Is Locked

Financial institutions, government agencies, and large enterprises operate critical workloads on COBOL systems that collectively process an estimated $3 trillion in daily transactions. The data produced by these systems is predominantly stored in fixed-width flat files — a format that encodes information by byte position rather than delimiters.

The challenge facing organizations today:

- Original COBOL programmers are retired or unavailable
- Documentation and copybooks (field definitions) are lost or incomplete
- Existing modernization tools require source code access or proprietary mainframe infrastructure
- Manual reverse engineering by COBOL specialists costs tens to hundreds of thousands of dollars per engagement
- Data cannot be queried, analyzed, or migrated without first understanding its structure

The result is that organizations are unable to access, analyze, or migrate data that is critical to their operations. The data exists. The structure does not.

What Hypha Did

Starting with a raw fixed-width data file containing no documentation, no source code, and no field definitions, Hypha executed a three-phase pipeline:

Phase 1: Distributed Statistical Profiling

Hypha sharded the input data into 100 equal segments of 1,000,000 records each and dispatched parallel profiling operations across available compute agents. Each operation analyzed every byte position across its assigned shard, computing character class rates (numeric, alpha, space, other) and collecting sample values.

A reducer operation aggregated results across all shards, producing a validated position profile for the complete 100,000,000-record dataset.

Phase 2: Schema Inference

From the aggregated position profile, Hypha derived field boundaries by identifying contiguous regions of consistent character class. No human interpretation was required. The inferred schema:

Field	Byte Range	Type	Inferred Meaning
1	0 – 11	Numeric	Account number (12 digits)
2	12 – 13	Alpha	Account type code (CK, CD, IR, MM, SV)
3	14 – 58	Numeric	Amounts / numeric data (45 digits)
4	59 – 60	Alpha	Status flags (AN, CN, HN)

Field type codes observed: CK (Checking), CD (Certificate of Deposit), IR (IRA), MM (Money Market), SV (Savings). Status flags observed: AN (Active/Normal), CN (Closed/Normal), HN (Hold/Normal). These semantic labels were inferred from sample values; they were not provided as input.

Phase 3: Extraction and Validation

Using the inferred schema, Hypha extracted all 100,000,000 records to structured CSV format. The extracted data was immediately query-able using standard tools (DuckDB demonstrated in verification).

Verified Results

Metric	Result
Total records processed	100,000,000
Shards	100
Shard failures	0
Bad width records	0
Record width	61 bytes (uniform, zero exceptions)
Schema inference	4 fields, statistically certain boundaries
CSV extraction — bad rows	0
Extraction time	~69.9 seconds
Query verification	DuckDB — structured output confirmed
Hardware	Commodity Linux workstation + mini-PC

Zero failures across 100 shards. Zero bad records across 100,000,000 rows. Schema inferred without source code. Data query-able within the same session.

Why This Matters

The conventional approach to COBOL data extraction requires one or more of the following:

- Access to the original COBOL source code and copybooks
- A running mainframe environment to execute the original programs
- Engagement of COBOL specialists at significant cost and timeline
- Proprietary migration tools tied to specific vendor ecosystems

Hypha requires none of these. Given a raw fixed-width file, Hypha can:

- Infer the record structure statistically across the full dataset
- Validate structural integrity at scale (zero-exception record width verification)
- Extract structured, queryable data without human schema definition
- Execute the complete pipeline on commodity hardware in minutes

This capability is immediately applicable to any organization holding fixed-width legacy

data it cannot read. The total addressable market includes thousands of financial institutions, insurers, and government agencies operating COBOL workloads today.

Architecture Overview

Hypha is a hardware-agnostic distributed compute platform. Key architectural properties relevant to this benchmark:

- Operations (ops) are the unit of work — not containers, not VMs
- Agents pull work from a central controller and execute ops independently
- Self-healing epoch fencing and lease TTL recover failed or stalled work without human intervention
- No shared filesystem required — agents operate on locally available data
- Fully deterministic results — identical input produces identical output

The benchmark hardware consisted of commodity Linux systems. No mainframe, no specialized hardware, no cloud infrastructure. The same pipeline runs on any hardware Hypha supports, which includes x86, ARM, and mixed fleets.

What Is Not Claimed

This paper makes no claims beyond what was directly observed and measured. Specifically:

- The account type and status flag labels are inferred from sample values, not confirmed against source documentation
- The numeric field at bytes 14–58 contains 45 digits of encoded data; interpretation of subfields (balance, date, transaction amount) would require additional domain analysis
- Benchmark data was synthetic; results on production data may vary
- No claim is made regarding the cost of mainframe alternatives beyond general industry context

Conclusion

Hypha demonstrated a complete, verified pipeline for extracting structured data from unknown fixed-width COBOL files: distributed statistical profiling, automated schema

inference, integrity validation, and CSV extraction — 100,000,000 records, zero failures, ~69.9 seconds, commodity hardware.

The capability is real. The numbers are reproducible. The problem it solves costs enterprises millions of dollars annually.

HYPHA | Patent Pending

© 2026 TKT, Inc. All Rights Reserved.

HYPHA technology is the subject of pending patent applications. The contents of this white paper, including its text, illustrations, diagrams, and technical descriptions, are protected by applicable copyright laws. No portion may be reproduced, distributed, or transmitted without prior written permission from TKT, Inc.