MYRROR

Briefing Note

December, 2024

Technology Innovation and the Impact on the ETL Process Mining Approach

Overview

The applicability of Extra-Transform-Load (ETL) process discovery technologies is being questioned as new technologies are adapted into legacy systems and the demands for real-time system and data visibility becomes progressively more pronounced.

Technology Changes and ETL

Whilst ETL has been the foundation of System Visibility and Data Processing for at least the past two decades, ETL is becoming progressively less relevant as technology changes impose increasing pressures on ETL's batch processing methodology.

The first dynamic is two inter-linked factors - the exponential growth in data volume, coupled with the exponential growth in frequency of data processing. For brevity shall call this, the **Exponential Dynamic (ED)**.

The ED's explosive growth is resulting in data and event log volumes that overwhelm ETL's periodic batch processing and event-log sampling capabilities.

The second dynamic is the **AI Revolution (AI-R).** There is little doubt, AI is having a once-in-a-generation transformative impact on system use and process automation. This generational change is destined to extend its reach and accelerate its impact as AI models become more progressively more efficient and sophisticated.

All AI models are dependent on high quality, continuously streaming data to operate. The demand for constantly supplied, high quality data will only increase as AI extends its data-driven, decision-making functionality across all areas of system activity - and this will inevitably be the case.

The AI-R represents a seismic shift in both the way in which new systems will be designed and implemented and the way in which the enormous investment in legacy implementations will be leveraged in the future.

ETL's batch processing of periodic samples of historical system activity is simply non-compatible with the demands of the AI-R

The third dynamic is what I shall call the **Real-Time Requirement (RTR).** RTR is the need for immediate visibility over system and data change activity.

RTR is particularly relevant when considering system intrusion threats and the containment of malicious activity. ETL's batch processing of "snapshots" of historical system events is incompatible with the RTR demands of continuously processing systems.

The currently being witnessed simultaneous convergence of the:

- Exponential Dynamic
- AI Revolution
- Real-Time Requirement

is resulting in a fundamental shift in the methodologies to be applied to the management of data processing and the securing of activity visibility across widely distributed, heterogenous systems.

There is little doubt that this shift will be marked in the historical record of the emergence and development of Information Technology systems as the period when system and data management migrated from “Batch Processing First” to “Streaming is Mandatory”

This migration will additionally show the period when ETL’s two-decade periodic sampling and batch processing methodological dominance was replaced with the new methodology of the continuous capture of all system end-point behaviour.

“Streaming is Mandatory” does not mean the “End of Batch”

The migration to “Streaming is Mandatory” does not mean the end of periodically batch processing samples of historical system activity. What is means, is “Batch” becomes optional to “Streaming”

Batch processing will not be eliminated by Streaming is Mandatory, rather it will become a specific-use capability rather than the default approach.

Batch will still be used for specialised activities such as:

- cold-data processing
- historical data analytics
- cost-optimised report generation
- testing and training of development models

Streaming is Mandatory will replace batch processing where immediate, visibility is required over data processing and system visibility activity. This will include such activities as

- real-time system analytics
- “always-on” payment processing
- intrusion detection and malicious activity containment

AI-R, RTR and System Semantics

The increasing sophistication of intrusion software renders the need for real-time system visibility and malicious activity containment a mandatory requirement.

This imposes new demands on system and data processing visibility. It is no longer sufficient to hope to identify what has happened on the system by the periodic sampling of a selection of historical event logs, transforming these into a common format, conducting an analysis of the sampled logs and generating an output in a proprietary format.

What is now required is the real-time capture of the complete universe of system activity, rendering this activity into a formal system representation and extracting an understanding of the system’s semantics by preserving the context in which the activity took place across both different data types and different executing platforms.

System visibility capture using ETL tooling is not designed for this task. ETL extracts a sampling event logs and uses SQL to transforms these into a tabular format. The resulting table provides core system activity information - User, Vendor, Event Type, Case ID, Timestamp and other information originally aggregated into the event log(s) to assist with system management, not to generate a representation of the executing semantics.

Both the AI-R and RTR require the capture of richer activity information than can be provided by a sampling of event log activity. This becomes particularly relevant when dealing with

- concurrent activities executing across heterogeneous legacy environments
- continuous access to real-time data changes versus the periodic export of historical activity
- real-time AI event driven decision making rather than polling for system changes

A View of The Future

The shift to AI driven decision making coupled with User demand for immediate visibility over both system and data change activity and the semantics relating to these changes is rendering conventional ETL methodologies anachronistic - some may argue these changes are calling into question the viability the conventional ETL business case.

Whether this is the case or not is not yet clear. What is clear:

- across all verticals, entire implementations are migrating to streaming architectures and real-time processing
- AI requires access to continuously streaming real-time data
- schema and logic change that previously required manual re-coding is being replaced with machine generated and deployed system updates
- AI is able to generate data processing pipelines without the need for conventional process modelling tools

What is also clear is that the generational shift from conventional “Batch Processing First” to “Streaming is Mandatory” is accelerating and accelerating quickly.

This will result is a fundamental reassessment of the economics and viability of the business models of many of the ETL vendors. This change, in turn, will result in the rapid acceleration of new opportunities for real-time data management and process execution visibility across continuously executing, event driven architectures.