Symbolic Generative AI for Fast, Accurate, and Efficient Data Integration & Query Migration.

Today, much of the time, risk and expense in digital transformation efforts is realized in data engineering and integration. Data engineers/architects merge ontologies manually to create a master data model or schema, with the intent that all elements within an organization will comply (e.g. governance). Unfortunately, the sizes of modern ontologies are far too large for humans to keep in their working memory and different sub-organizations often have different and sometimes competing requirements which leads to insufficient funding and motivation to provide quality data for others in the enterprise. This often leads to an endless stream of questionably effective meetings to reach consensus. Once an agreed upon master data model is established, systems undergo manual testing to ensure data integrity and system functionality. This testing can only cover a small fraction of potential errors. While many solutions exist to augment and semi-automate data integration, it remains a manual process that requires labor with sufficient skills to write queries and analyze results. This work is monotonous, error-prone, and often leads to high levels of attrition and labor churn. Even worse, when new data sources are added to the enterprise or upgraded, much of this work needs to be thrown away and then repeated. These issues lead to extended project completion time, high cost, inconsistent labor, and ultimately high risk.

Conexus AI provides an innovative solution using symbolic, generative artificial intelligence (AI). Instead of approaching data integration as a set of step-by-step, manually written, procedures, Conexus AI applies many patented algorithms to discover the mathematically provable unique, most efficient common data warehouse, free of human bias and error. These algorithms detect contradictions as they operate and do not require rework with system upgrade or addition, making them more scalable (consensus-free; fewer design meetings required), cost effective, and error-free. Symbolic AI is deterministic, transparent, and explainable, avoiding the many issues with the more common probabilistic generative AI solutions that are prone to hallucination, bias, and error.

Conexus AI is an on-shore, U.S. software and services company providing symbolic, generative AI solutions, primarily to Fortune 100 companies. Conexus AI is the only small business spin-off from the MIT Mathematics Department. Initially developed with funding from DoD and DARPA, the U.S. Department of Commerce (NIST) funded the development of Conexus CQL, a software program that implements Conexus AI's patented algorithms. Conexus AI has subsequently improved their software to meet the demands of commercial clients at scale, enabling previously impossible data integration efforts.

Conexus AI's symbolic generative algorithms have far-reaching applications, with new use cases being discovered by clients regularly, however, there are two principal offerings: 1) Efficient Generative Data Integration; and 2) Query Migration.

Efficient, Generative, Data Integration

Conexus AI CQL's chase algorithm generates precise data sets from disparate data sources, silos, and extra-enterprise partners that are optimal in data quality, in the sense that the data sets are mathematically guaranteed to preserve the semantics, including context, of the source data, while not generating any "extra" data that is not logically implied. Conexus AI avoids wasting resources attempting to manually extract ontologies and achieve consensus. Instead, relationships between data structures are defined as logical axioms. These relationships can be learned from log files, enterprise data exhaust, or defined by domain experts. Rather than attempting to bring domain experts together to achieve consensus, manual definitions are independently defined and merged by the Conexus chase algorithm, leading to an optimal integrated database as output. Logical contradictions in data are quickly identified and thus can be easily resolved at the enterprise level. We illustrate this methodology using an example involving spreadsheet integration from a large commercial client. While data integration is more commonly executed against large cloud-based data warehouses, the spreadsheet lends itself to a more intuitive interpretation.

Consider two structured data sources, A and B, both representing analysis and remediation of cyber event incident data but doing so differently. Established rules relate the two sources; most rules are simple, but some rules encode complex logic, such as privilege escalation and lateral movement. The Conexus' CQL constructs a consolidated schema and database with speed and autonomy. Deterministic fields are mapped and integrated, and redundancy eliminated in a way robust to source change. The chase algorithm is a form of generative AI that efficiently operates in-memory, infers logical relationships, and identifies any logical contradictions that may indicate a conflict, either due to logical or human error. The contradiction detector narrows the focus of attention to resolve conflicts quickly and mathematically prove that no new errors have been introduced through the data consolidation process.

Figure 1 illustrates this process, where blue fields represent those from source A, red fields represent those from source B, and the mixed red/blue tabular data represents the generated data model/warehouse, with gray schema elements being common to both sources. Yellow fields represent overlapping fields that may require human attention, for example, the same data occurring in both sources. Correlated, generated placeholders (?0 and ?1 shown in Figure 1 as red text in white cells) provides the ability to capture unknown relationships and differentiates the chase technique from SQL-based approaches. In addition, Conexus CQL emits its results to relational databases, graph databases, RDF and more, which provides enhanced indexing and search as well as allowing data to be exchanged among sources.

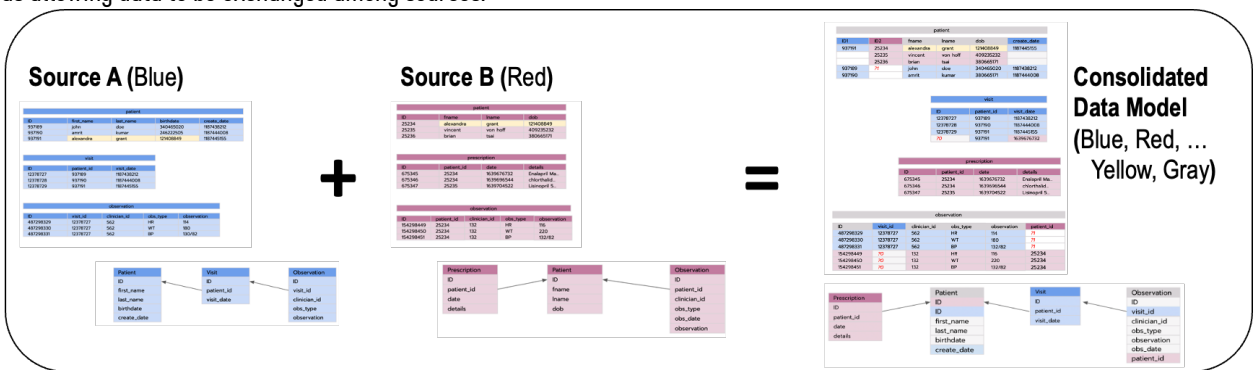


Figure 3. Generative Symbolic AI for Data Integration Example

The resulting integrated warehouse (or spreadsheet in the example) is the unique, most integrated warehouse, free of any contradictions between data sources. Due to the use of formal methods inherent in the Conexus symbolic AI, the accuracy of data integration is mathematically provable, avoiding the need for time-consuming and error-prone testing. This effectively accelerates transformation efforts, using less cost and labor than traditional methods, while also improving accuracy and data integrity. Finally, because the Symbolic AI operates using logical axiomatic data relations defined in parallel, there is no need for rework when systems are replaced, upgraded or added- the new system must be mapped to existing sources, and nothing else changes, which achieves tremendous time and cost savings, while again improving accuracy as shown in Figure 2.

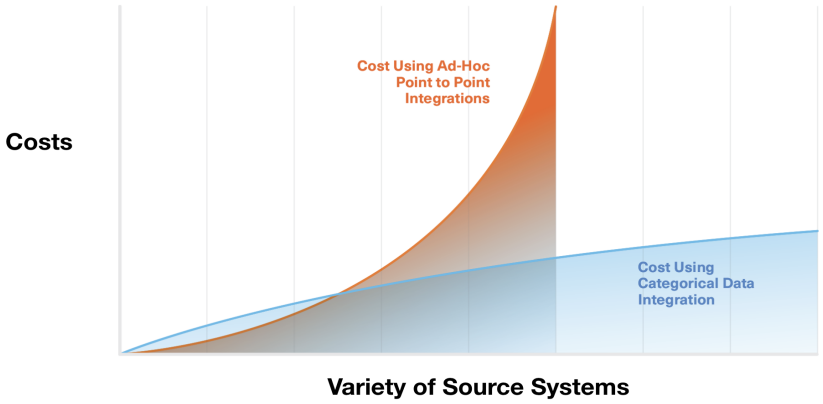
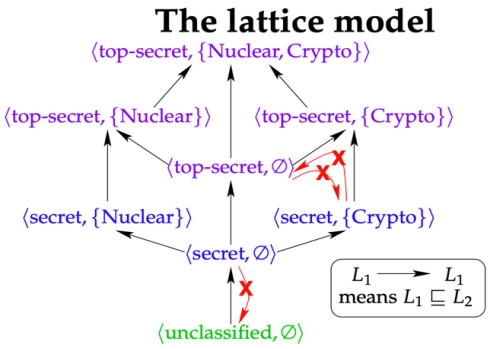


Figure 2. Over time cost comparison of Conexus AI (blue) and traditional integration (orange).



- **Information can only flow up the lattice**
  - System enforces "No read up, no write down"
  - Think of $\sqsubseteq$ as "can flow to" relation

Figure 3. Privacy & Security Model

Query Migration

Within any digital transformation effort, following data integration as outlined above or in any ETL process, all the existing queries and analytic dashboards were written against the legacy system and often need to be upgraded as well. The Conexus AI chase algorithm not only constructs the data warehouse, but provides an automated way to translate queries from the sources into the schema of the warehouse, generating new queries in SQL, Java, XML, or other common languages from existing queries in SQL, etc. This process can also be executed in an arbitrary ETL project by suppling the Conexus AI with either an SQL-based data dictionary or a set of examples sufficient for the Conexus AI to learn a data dictionary. In either scenario, using symbolic AI, Conexus CQL mathematically proves the queries are consistent and will always preserve semantic meaning and return the same results (as far as is mathematically

possible), eliminating the need for costly, time-consuming, and error prone testing, as well as the time and cost of translating the queries from old to new in the first place.

Another benefit of symbolic AI is the ability to encode privacy and security as a data model - that is, within the exact-transform-load (ETL) itself - as shown in Figure 3. By including privacy and security policies as ontologies directly in the data to be integrated or migrated, security tasks can be done at compile time, instead of run time. For example, migrating data from a database of chemical molecule equivalence with a classified database of chemical weapons precursors, the data dictionaries or CQL queries constructed by users, will be checked by the Conexus AI for respect of these policies and ensure that an equivalent molecule is not inadvertently disclosed in otherwise classified information, in the exact same way the Conexus AI checks that e.g. foreign keys will not dangle.

Client Success Stories

Uber - Conexus AI has helped their clients achieve previously intractable problems with speed and accuracy. For example, Uber's scale prevented integration of sensor data, application events and imported data, adversely affecting the performance of the platform. At the time, Uber's data included 15M trips per day, 75M active riders, 1T Kafka daily messages. Their data catalog included over 260K Hive tables, 15K Kafka topics, 6K MySQL and PostgreSQL tables, 3K other tables in addition to Schemaless, Cassandra, Gairos, Apollo, Pinot and more. Conexus AI CQL reconciled existing language and domain specific conventions and constraints and generatively created and propagated schemas and best practices throughout the company, resulting in more efficient and optimal data warehouses (https://www.uber.com/blog/dragon-schema-integration-at-uber-scale/).

Synchrony Bank - Conexus AI has helped their clients improve data security, accuracy, and integrity addressing several key issues. Synchrony's security policy limited a user's data visibility. Their legacy system included cross database queries in 3rd party tools such as Python, SAS, and Ab Initio, however SAS and Python did not create automatic logs and they only used a partial SQL solution for queries. This led to a culture of creating temporary tables which obscured inference between the data warehouse (DWH) and tables, which Synchrony refers to as "dark data." These issues resulted in hidden potential vulnerabilities and inconsistent data/analytics leading to flawed insights and financial decisions. For example, two bank clients might be placed into the same marketing campaign in error. Conexus AI CQL constructed an MRI-like miner that emits dynamic inferred relationships between data tables spanning many DWHs, identifying contradictions, measuring frequency and popularity of fields and connections, enabling them to repair damaged governance systems and improve insight accuracy, all while maintaining data security. Conexus AI deployed their solution into a sandboxed, air-gapped environment.

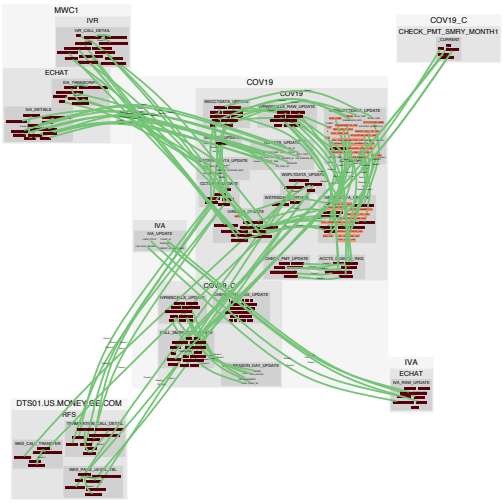


Figure 4. Discovering "dark" schema mapping

Empower Retirement is the 2nd largest retirement company in the United States, with 8.5 million plan participants and over 38,000 retirement plans under administration. Conexus AI assisted Empower to migrate pension data to a third party, discovering conflicting business rules between source and target, and repairing any conflicts. This project achieved 100% elimination of data errors downstream (a new standard in Data Integrity) by providing mathematical proof data is normalized and not missing values.

Conclusion

While there has been much recent attention focused on probabilistic generative AI which comes with significant risks in hallucination, bias, intellectual property, and uncertainty, Symbolic AI provides mathematically provable checks and balances. If we liken modern AI to a supped-up race car engine, Conexus' symbolic AI provides the brakes and steering. No race car driver will dare push their engine to its limits without steering and brakes. Similarly, we cannot reach the full potential of AI without the important checks, balances, and efficiency provided by symbolic AI in terms of contradiction detection, provable data quality, and speed of generative schema mapping.

Please contact Conexus AI to discuss solutions for your AI and data integration needs at ian@conexus.com or eric@conexus.com.