SmartData Fabric® (SDF)
Virtual Graph Database and Interactive Graph Visualization

February 2024

Graph database and visualization are ideal

GRAPH DATABASE AND VISUALIZATION ARE IDEAL because it represents ALL data as a highly normalized data model, regardless of source, and it can present complex data in an easy-to-understand and work-with form, i.e.:

- Key data, e.g., sales revenue and costs
- Key entities, e.g., customers, employees, patients, products, sales persons, organizations and locations
- Key relationships between entities, e.g., bought, sold, lives at, works for and part of
- Highlight important entities and their role in the business, e.g., highest revenue customers, highest profit customers, products these customer buy (may not be the best-selling products) and sales persons for these customers (may not be the best sales persons)
- Focus on groups of interest for deeper analysis
- See data lineage and governance - where data comes from and who is responsible for it
- See how entities and data are connected within and across systems
- Perform link analysis
- Perform Social Network Analysis (SNA)

WhamTech takes a bottom-up approach

- Conventional graph database applications take a top-down approach
  - Start with graph visualization
  - Relegate data access to (a) ETL into a centralized data warehouse, or (b) copy into a data lake and the ELT, and then for (a) and (b), export into a graph database, or (c) directly ETL into a graph database - each attempts to retroactively address data quality, standardization and de-duplication
  - Graph visualization interacts with a physical graph database
- WhamTech SmartData Fabric® tackles data challenges from the bottom-up
  - Start with data sources - leave data where it resides
  - Enable data access, discovery, profiling, identification, classification, security, quality and standardization, and impose corporate data governance through index-based federated adapters and non-index-based federation servers
  - Map and capture data links/relationships at the entity and attribute-level using Link Indexes™
  - Use Link Indexes™ with content and master data indexes for virtual graph database, link analysis and SNA
  - Enable almost any application, including interactive graph visualization, to access data in sources through standard drivers and SQL
  - Perform graph queries combined with other queries directly through SQL or GQL-to-SQL translation, on index-based federated adapters in SmartData Fabric® accessing data sources
  - Graph visualization interacts with a virtual graph database

Problems with conventional graph database solutions

1. Limited data access
   - Federated access with conventional adapters used for incremental ETL, ELT or just copy, and eventually into a graph database
   - Copying data to a data lake and ELT or data warehouse using ETL, and eventually ETL into a graph database

   => Ultimately, all data is extracted and retroactively processed for data quality, standardization, de-duplication and links, and stored in a centralized database
2. Low data quality of federated data access – impacts query success and entity resolution for nodal representations, and subsequent links - GIGO
3. Difficult to scale - large amounts of data in a centralized database – less of an issue with Cloud
4. Near real-time updates usually not possible = latency
5. As data is extracted from sources, many inherent links and associations among data are lost
   - Example: All record data in one table indirectly associated with all record data in another table through PK-FK relationships
   - Chronological sequence of records in a data source
6. Probabilities of links and confidence levels of data that may have been in sources, may also be lost

Comparison between conventional physical and SmartData Fabric® virtual graph database

Data Store	Initial Build Performance	Features	Application flexibility	Query Performance	Storage	Update Performance	Query processing	Examples
Physical graph database	Low, as takes effort and time to prepare data for triple store	Unless data quality, standardization and MDM are addressed before or during ETL, can have inaccurate or incomplete data Can have a lot of data excluded – only what is in triple store	High – limited to data in graph database and available views – can be difficult to filter based on additional attributes	High, tends to use memory	Minimal, depending on additional attributed stored – triples contain minimal attributes	Low, as many locations to find and update – normally, a batch refresh required	SPARQL, GraphQL (GQL) and others	Neo4J, Palantir and Titan
SmartData Fabric® virtual graph database	Uses Link Indexes™ pre-built, as content indexes, and master data and indexes, are built	Combination of Link Indexes™, and content and master data indexes accessing data in sources	Extremely high, as also uses standard drivers and SQL on adapters to access data	High, as links are pre-built	Extremely efficient, as data stays in sources and minimal in indexes – only additional Link Indexes™ storage, which can be used for other purposes, e.g., MDM and query acceleration	High, in near real-time, as source content indexes, Link Indexes™, and master data and indexes, are updated	SQL and future GQL-to-SQL translation Also, may support native GQL in the future	Unique to SmartData Fabric®, but can be used in conjunction with most graph visualization software – SmartData Fabric® responsible for all virtual triple generation

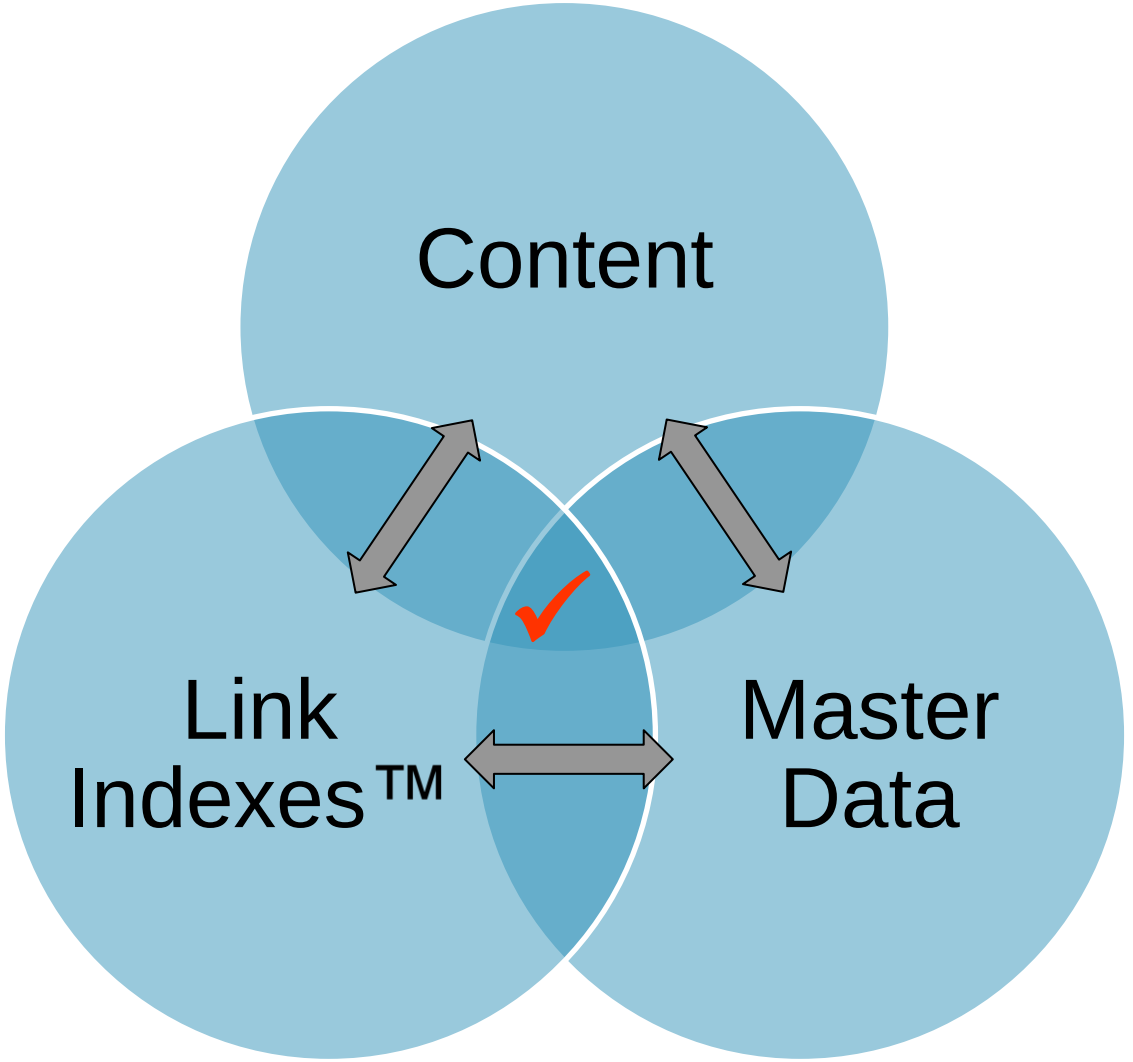
Virtual graph database and link analysis

- Three modes in SmartData Fabric®:
  - Virtual Graph Database – virtual triple store based on a standard semantic view aka Standard Data View (SDV)
    - Currently, works with standard SQL
    - Eventual SPARQL, GQL and/or OWL query support, with translation to SQL or native GQL
  - Graph Visualization (examples available in multiple presentations and demos)
    - SQL query generates a virtual triple store used directly by graph visualization software
    - Almost no computation needed
  - Link Analysis and Visualization (subject of other presentations)
    - SQL query generates a result set containing entities, relationships and multiple attributes accessed through graph visualization software
    - Link Analysis performed across SmartData Fabric involving multiple distributed indexes, including Link Indexes™

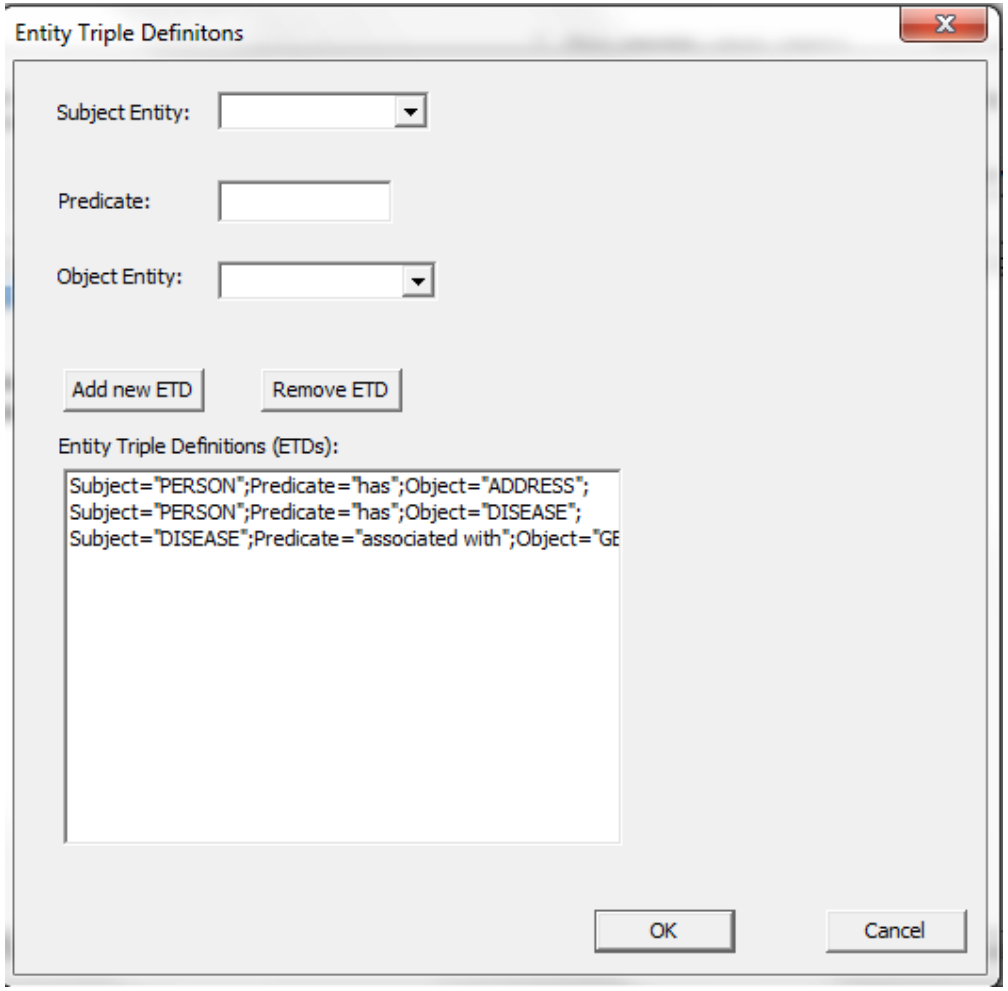
Virtual graph database is combination of three types of indexes

Content indexes basis for other indexes

All indexes resolve to “record numbers” – internal to SDF, but correlated with external/data source references, and can be combined



Virtual graph database



Entities are defined in the standard data view that data source indexes are mapped to, and can use these entity triple definitions to:

- Export triples to a triple store, maybe with other attributes, for graph database work
- Execute SPARQL, GQL and/or OWL, with translation to SQL, directly on indexes, using a combination of content, link and master data indexes

Both the above are ongoing projects

Future options may be to import RDF data models into SmartData Fabric® instead of manually defining them, e.g., HL7 FHIR RDF (https://www.hl7.org/fhir/rdf.html)

Cambridge Intelligence “KeyLines”

- Highly renowned software and very active in the graph visualization community
- Highly interactive thin client link/network visualization software
- Server-generated visualization
- Any device HTML display and interactivity
- Interaction allows generation of SQL queries with user-defined constraints across SmartData Fabric®
- Interaction refines and refreshes visualization in near real-time
- https://cambridge-intelligence.com/keylines/

Demos with virtual graph database and interactive graph visualization

- WhamTech Bitcoin - Blockchain Anti-Money Laundering Graph Visualization
- WhamTech EU General Data Protection Regulation (GDPR)
- WhamTech Financial Services
- WhamTech Virtual Master Patient Index (VMPI) for Huntington Medical Foundation
- WhamTech Virtual Master Person Index (VMPI)

The End