SMARTDATA FABRIC® SOFTWARE PRODUCT DATA SHEET DETAIL

REVISION 1.1

SmartData Fabric® enables solutions that solve data problems, and fills the three main gaps in the current data management and analytics markets.

Data problems statement

- Data is everywhere in multiple locations/countries, sources, source types, formats and platforms.
- Cannot/will not copy/move ALL data to a single location, e.g., multi-jurisdictional/data sovereignty and third-party ownership.
- Some data is not readily useable and needs discovery, identification, classification, security, cleansing, transformation, standardization and normalization.
- Data needs to be connected and integrated.
- Many data sources do not have the query capabilities or load capacity for external queries by conventional data virtualization/federated data access/query engines.
- Large investment in existing systems and solutions.
- Need for rigorous enterprise-wide access control, and data governance, security and privacy.

The three main gaps in the current data management and analytics markets

1. Data issues are not addressed BEFORE the first query is made on data sources or data lakes, including data and master data management. Most data and master data management occurs AFTER data is extracted from sources on query result-sets/data extracts or landed in data lakes.
2. Not all customers want to, or cannot, copy data to a centralized location, data lake or large data warehouse – even conventional data virtualization is typically two-tier and tends to land data in a data lake, repository or cache.
3. No middle tiers that allow independent edge processing to address data source and data issues, including data and master data management, and create and maintain derived values and pre-aggregated, pre-calculated and pre-joined views. These views can accelerate reporting, BI and analytics queries, and entity views such as customer 360.

Conventional data solutions are a choice between a rock and a hard place

ROCK = CONVENTIONAL DATA VIRTUALIZATION/ FEDERATION

HARD PLACE = DATA WAREHOUSE

Each conventional data solution has its own pros and cons.

Conventional data virtualization/federation has major cons of data source:

CDV1. Data quality.
CDV2. Query capabilities.
CDV3. Query load.

A data warehouse has major cons of:

DW1. All data needs to be copied to a central location.
DW2. ETL and source schema transforms to a one-size-fits-all target schema.

Data lakes are in-between a rock and hard place, whereby they address some IT issues by centralizing data, but have major cons of:

DL1. All data still needs to copied to a single data lake or distributed data lakes.
DL2. Not helping with data management => still requires ETL to a data warehouse/materialized views and then data marts, ELT or ***an additional data management layer or fabric/mesh on top***.

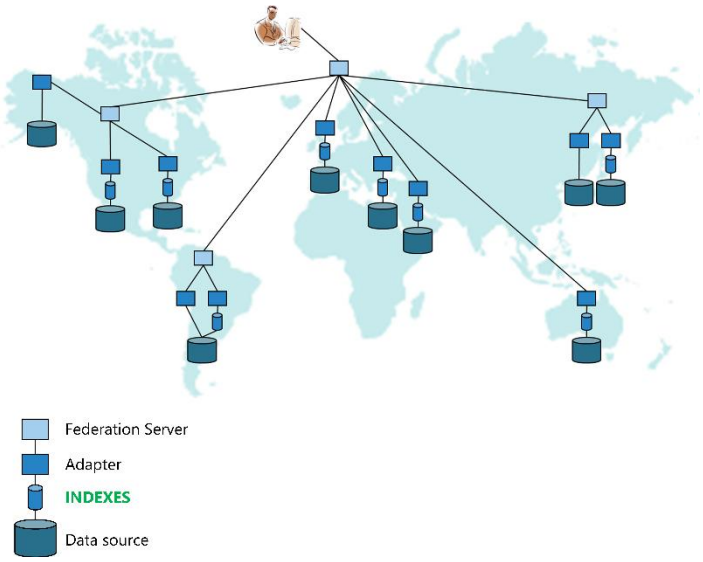
SmartData Fabric overcomes the cons and supports the pros of the above three approaches.

Note: The following text includes **blue-highlighted terms** that are specific to SmartData Fabric®.

SmartData Fabric® core architecture and components

- SmartData Fabric® consists of a multiple federated data source adapters and federation servers that access multiple adapters and other federation servers as if they are adapters.
- The highly flexible architecture enables a true data fabric/mesh across multiple data sources and almost any data source with structured, semi-structured data.
- There are **three types of federated data source adapters: indexed, conventional or non-indexed, and hybrid**, which is a combination of indexed and conventional.

- Each data source has at least one adapter and some data sources may have multiple adapters for scale and performance.



- There are significant benefits to using **indexed adapters** that can overcome data source and data issues, by addressing query processing capabilities, performance and load, and data quality and standards, and add features such as Master Data Management (MDM), data monitoring and event processing.
- Indexed adapters and the independently configurable federated architecture leverages, complements and is agnostic to existing systems, data sources, data virtualization software and query engines – avoids vendor lock-in.

SmartData Fabric® business benefits

- Lower Total Cost of Ownership (TCO) as lower implementation and maintenance costs.
- Faster and simpler implementation compared to the individual data source query optimization or ETL effort required of conventional data solutions.
- Return on Investment (ROI) much higher than alternatives because of additional capabilities, flexibility for current and future needs, and lower TCO.
- Incremental/phased implementation - start by addressing urgent needs first.
- True data fabric/mesh that enables and supports multiple use cases.
- Works with existing and future systems.

SmartData Fabric® comparison with conventional data solutions and data lakes

SmartData Fabric® combines the best of conventional data solutions and overcomes the worst of these data solutions.

Feature	Conventional data virtualization/query engines, e.g., Dremio, Presto and Trino	Data warehouse, e.g., Oracle, Snowflake and Teradata	Data lake, e.g., Hadoop and variants, incl. MapR	SmartData Fabric® true data fabric/mesh
Leave data in sources in original schema/format – ownership, compliance and cost	✓	✗	✗	✓
Clean, transformed and standardized data, views, and data and business objects	✗	✓	✗	✓
Uniform high-performance SQL queries on independent indexes and views	✗	✓	✓ or ✗	✓
Avoid high query loads on source systems	✗	✓	✓	✓
Advanced access control, data governance and security, regardless of source	✗	✓	✗	✓
Avoid additional ETL to a data warehouse/mart/views or ELT in data lake	✗	✓	✗	✓
Easy to add/remove data sources and schema-on-read flexibility	✓	✗	✓	✓
Avoids data latency	✓	✗	✓ or ✗	✓
Integrated and automated Master Data Management (MDM)	✗	✓	✗	✓
Pre-process and query unstructured data/text as part of SQL	✗	✓ or ✗	✗	✓
Actively monitor data sources, process events and support interoperability	✓ or ✗	✗	✗	✓

This remainder of this product data sheet will expand on the above-listed features:

Leave data in sources in original schema/format – ownership, compliance and cost

- SmartData Fabric® has several – a few unique -ways of dealing with raw data in and from sources:
  - Leave data where it is and **read, transform and index (RTI)** query-only data, and indexes point to the original data in sources. WhamTech call these **Virtual Keys (VKs)**. RTI cleans, transforms, standardizes and secures raw source data that is used for index values.
  - RTI data to indexes as with VKs, but store the clean, transformed, standardized and optionally secure data in the indexes instead of pointers to the raw data in sources. WhamTech calls these **Non-Virtual Keys (NVKs)**. The advantage over VKs is high performance.
  - Not all data needs to be indexed for queries. In fact, most data do not, so, WhamTech has **Virtual Columns (VCs)** that are data left on the source that is not needed to be indexed, but can be part of an eventual result-set. When a query result-set is created, VCs are used to "fetch" VC data from sources.
  - Conventional or non-indexed adpaters that submit queries directly to data sources, but then are subject to the query capabilities, performance and load of the source system, and quality and standards of source data.
  - Regardless of how data is isolated, read and combined to form result-sets, source data is cleaned, transformed, standardized and secured.

Clean, transformed and standardized data, views, and data and business objects

- As described in the above feature description, **WhamTech addresses data quality and standards BFEORE the first query is made on indexed adapters** – not conventional or non-indexed adapters.
- **Queries are made on standard data views, and data and business objects**, that are, many times, associated with APIs that are enabled and maintained in the adapters.
- In the case of indexed adapters, standard data views, and data and business objects are mapped to clean, transformed and standardized data in indexes and indexed views (see next feature section), or metadata in the case of conventional or non-indexed adapters.

Uniform high-performance SQL queries on independent indexes and views

- For indexed adapters, SmartData Fabric® does not need to modify and optimize a high-level query from an application to fit each individual data source that may be limited in query capabilities, performance and load capacity. Instead, **SmartData Fabric® uses uniform PostgreSQL query processing across all data sources, regardless of data source query support, or data quality and standards**.
- SmartData Fabric® is able to take advantage of **additional index types and views for query processing and can maintain indexes and indexed views in near-real-time, for pre-aggregation, pre-calculation and pre-joins**. This improves scalability and performance.

Avoid high query loads on source systems

- **Indexed adapters absorb all the query load with NVKs and most query load (>90%) with VKs**.
- With indexed adapters, **indexed views also greatly reduce query load because of the near real-time updates of pre-aggregated, pre-calculated and pre-joined indexed views**.

Advanced access control, data governance and security, regardless of source

- Developed for a large government contractor, **SmartData Fabric® enables advanced AD/LDAP-based IAM, SSO, RBAC, TLS, RLS and CLS**.
- Applies to both indexed and conventional or non-indexed adapters.

Avoid additional ETL to a data warehouse/mart/views or ELT in data lake

- **RTI does not transform data source schemas unlike ETL to a one-size-fits-all data warehouse schema**, and is similar to ELT in the sense that it works with data source schemas as they are, but not in a centralized or distributed data lake. Much of the work associated with ETL is because of schema transforms rather than data transforms.

Easy to add/remove data sources and schema-on-read flexibility

- Data sources can be added or removed one at a time without impacting the rest of the data fabric/mesh. This also allows **a phased approach to implementation**.
- Schema-on-read through multiple simultaneous views allows great flexibility in how data is consumed by applications. **Different applications or even different queries by the same application can use different views**.

Avoids data latency

- **As data sources are accessed in real-time and indexes updated in near real-time, the latest data and associated views are available for queries**.
- Most data lakes and data warehouses contain aged data unless they are updated in near real-time.
- **Archived or snapshot result-sets can be accommodated through data extracts**. Integrated and automated Master Data Management (MDM).
- **SmartData Fabric® seamlessly and automatically integrates master data**, which bridges the gap between data physically in sources or data lake and virtual logical data warehouse/mart-type views.
- Key entity data, such as a "person" with personal attributes, are deduplicated, best content captured and normalized/connected to associated data.
- ETL from source data to a data warehouse is normally where such a process occurs, but **with SmartData Fabric®, master data can be used to virtually represent data warehouse/mart-type views of data, e.g., a single person 360° view, product, organization, etc.**
- **MDM can also be used in virtual or physical graph database to logically represent entites as single nodes AND also retain the actual physical connections between entity data and associated attributes, and other entites**. This is particularly valuable for solutions such as entity (person, product, organization) 360° views, fraud, cybersecurity, intelligence, GDPR and ontological/semantic representations for analytics.

Pre-process and query unstructured data/text as part of SQL

- Indexed adapters can preprocess unstructured data using text analytics using Spark NLP, for example, and build extracted entity indexes and unstructured search indexes.
- Unstructured search and structured queries can be combined to provide additional query options.

Actively monitor data sources, process events and support interoperability

- **Because indexes and indexed views are maintained in near real-time, they can be monitored for predefined significant changes and trigger events**. Rest APIs can work in conjunction with MQ orchestration, e.g., RabbitMQ and BPM software, to drive updates and notifications, e.g., KPI updates for operational dashboards.
- **Other options include being able to write back to data sources**, which can be simple or complex, depending on the update.

SmartData Fabric® core software

- Data source adapters are based on open-source PostgreSQL[1] and Trino[2] data and query processing technologies.
- Federation servers are based either on PostgreSQL or Trino, each bringing their own unique capabilities. PostgreSQL-based federation brings extensive SQL support, horizontal scaling, etc., whereas Trino brings scalable, high-performance, in-memory, cross-data source join capabilities.
- Indexed adapters are based on PostgreSQL.
- Both PostgreSQL and Trino support conventional adapters that can work in combination with index-based data source adapters, aka hybrid adapters.
- Seamless and automatic MDM is supported by PostgreSQL federation servers. MDM for Trino federation servers is under development.
- PostgreSQL runs on Windows and Linux. Trino runs on Linux only.

- All adapters and federation servers run on physical and virtual machines, open-source Docker containers, and on on-premises, data centers and multiple cloud platforms.
- Most configuration client tools currently run on Windows only, even though adapters and federation servers run on both Windows and Linux. OS-agnostic browser-based configuration client tools are under development.
- Automation of many currently manual configuration steps is under development with more automation planned.

For more details and supporting material for the above features, please contact Gavin Robertson, CTO and SVP, WhamTech, Inc., gavin.robertson@whamtech.com, +1 (972) 991-5700 x706 (o)

[1] https://www.postgresql.org/.
[2] https://trino.io/.