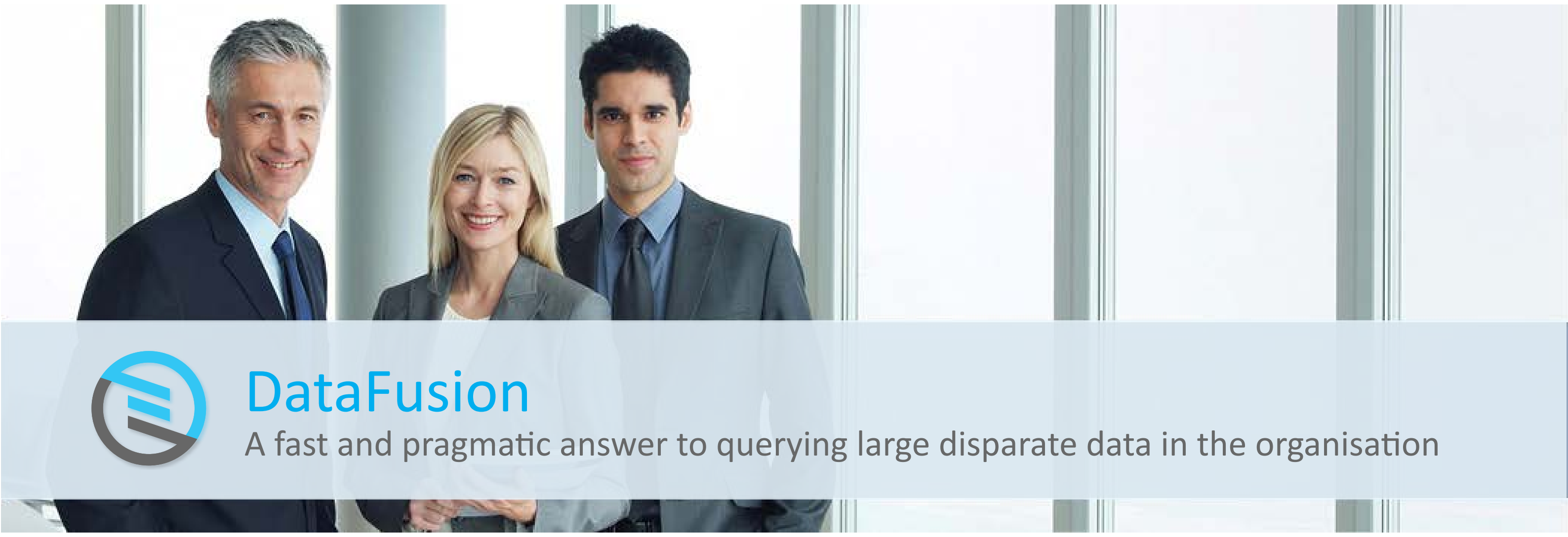
StreamFinancial



DataFusion

A fast and pragmatic answer to querying large disparate data in the organisation

The Challenge

Many organisations have the requirement to query, in near or real time, highly secure data that is stored across many different systems. These systems leverage different technologies, jurisdictions, teams and geographies. The queries require agile delivery with the user able to specify the complex cross system joins at short notice.

Financial institutions are an egregious example of how difficult it is to make use of the high quality data already existing in operational platforms. The same is also true of many large corporates, asset managers and governments.

> *"One of the most significant lessons learned from the global financial crisis that began in 2007 was that many banks lacked the ability to aggregate risk exposures and identify concentrations quickly and accurately at the bank group level, across business lines and between legal entities." BCBS239 Risk Aggregation & Reporting*

The Data Warehouse

The traditional approach – data is centralised in a physical data warehouse in order to apply more control and governance, eliminate inconsistencies and achieve the necessary efficiencies of scale.

Smaller implementations can work but many failed programmes have shown this approach to have significant disadvantages.

- The rate of change of the central data warehouse will not support the natural rate of change of the feeding systems
- Data is a translated copy of the golden source with no clear ownership and doubtful accuracy
- Very difficult and slow to change in response to new requirements or additional data sources
- Expensive hardware, software and implementation

To overcome these issues, some large organisations have had to resort to adding another data warehouse (and data copy) in response to each new and complex requirement!

The Data Lake

On realising the limitations of Data Warehouses, organisations have responded by building Data Lakes. These feed the raw data of all the source systems into a large file system based repository with a soft schema.

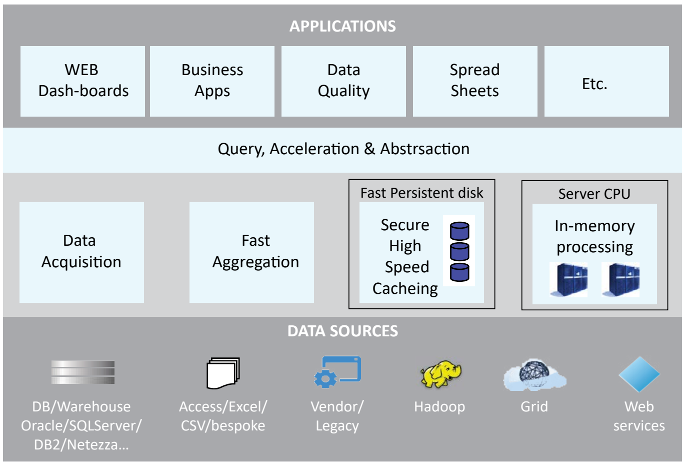
This brings some theoretical advantages; data from new sources can be added with little difficulty and the software is better able to query this data.

However, there are some serious challenges with this approach which has some disadvantages compared to the traditional Data Warehouse.

- No common schema The writer of the query needs to understand the schemas of all the source systems and all of the change histories – an impossible task for a large organisation
- Data Security The security model for such a large data set, with copies of data from many owners and jurisdictions, is exceedingly difficult to manage
- Performance Many of the technologies chosen struggle to provide the concurrency required for heavy usage

The Solution - Data Virtualisation

What is really needed is an Internet of things approach but with a usable and well understood schema.



APPLICATIONS

WEB Dash-boards | Business Apps | Data Quality | Spread Sheets | Etc.

Query, Acceleration & Abstrsaction

Data Acquisition | Fast Aggregation

Fast Persistent disk: Secure High Speed Caching

Server CPU: In-memory processing

DATA SOURCES

DB/Warehouse Oracle/SQLServer/ DB2/Netezza... | Access/Excel/ CSV/bespoke | Vendor/ Legacy | Hadoop | Grid | Web services

DataFusion

An agile and low cost route to the virtual data warehouse

- Users can query against their golden source data using a well understood schema with exceptional query and concurrency performance.
- Golden source data and access controls remain with the owners of the operational platforms
- Operational platforms can be protected from big queries
- Operational platforms can remain diverse in technology, team and location
- Operates with billions or even trillions of rows of fast moving data

"If there is any megatrend in the Gartner Data Warehouse Magic Quadrant (MQ) report, it's the emergence of the logical data warehouse. Essentially, this concept refers to the federation of various physical DW assets into one logically unified virtual data warehouse." - Gartner Data Warehouse Magic Quadrant Report 2013

This is not a new idea – but a successful implementation requires a number of key elements.

- Single and flexible logical schema to query against
  - Schema translation and anonymisation federated to source system teams
- Ultra-high performance for responsive querying across distributed and heterogeneous platforms
  - Federated querying at full granularity
  - High performance caching to protect source systems and responsiveness
- Both federated and role based security models
- Agile deployment
- High availability

Without these attributes, users (in an attempt to perform their duties) have to make uncontrolled data copies – this is potentially worse than the controlled copy in a data warehouse.

DataFusion

Provides an accurate and timely enterprise wide view, of high volumes of disparate data held in multiple highly specialised systems, geographically spread in many jurisdictions.

- Volume- very large scale
- Variety- disparate systems, locations, jurisdictions
- Variability- changes in shape and format
- Velocity- accessed quickly for decision-making
- Veracity- accurate & complete with full lineage

Flexibility

DataFusion provides a highly flexible way to query very large data sets held in disparate heterogeneous data sources. For example, DataFusion can select a few billion rows from hundreds of billions, join these to attributes from another data source and then group them as required; all in a few seconds.

Further attributes from additional data sources can be joined with ease - even from CSV files, Access data bases or spread sheets.

Unlike other technologies, DataFusion employs both a soft schema (usually restricted to NoSQL solutions) and the widely used SQL query language used by databases.

Open

DataFusion has a plug-in framework which allows it to integrate with and be called by any technology.

This open design allows for easy leveraging of legacy infrastructure, analytics, vendor platforms and other best-of-breed technologies, such as:

- SQL Server
- CSV
- Access
- Excel
- MySQL
- Active Directory
- "R" Analytics
- WEB services
- Vendor platforms
- Python
- Bespoke Market Data
- Oracle
- Matlab
- Hadoop
- ODBC
- Qlik

Lineage

DataFusion's flexibility and open architecture offers an easy route to providing fully granular lineage information for all data sources and making this available in the final schema.

High Performance

- DataFusion can perform highly distributed high performance in-memory queries, including joins, whilst maintaining the data security and flexibility of highly compressed on-disk storage.
- DataFusion takes advantage of its own optimised vector processing engine to both query and stream data highly efficiently on the network.
- DataFusion allows the processing to be taken to the data across multiple technologies. This is a key feature in improving the performance of real-world queries.

Agile and Reliable

Agile deployment is only achieved when the architecture has been designed to support it. In practice, this means that all components need to support side-by-side deployment with instant roll-back. For reliability, all services are run in highly available groups.

Security

DataFusion supports both federated (identity based) and role based security models. This is achieved by integrating Kerberos, Active Directory and other security infrastructure. All Data can be encrypted at rest.

For more information: info@stream-financial.com