Data Preparation
A new perspective for BI & Analytics

November 2020

For two and a half decades or more, organisations have followed one of the two data warehousing principles of Inmon (CIF, GIF and DW2.0) and Kimball (Star Schema Fact models) for preparation of data. Both methods have intrinsic benefits but leave organisations with some challenges in accessing warehoused data. More recently Amazon S3 has provided data lake capability which similarly requires deep IT knowledge and somewhat prescriptive of data understanding to be of downstream value.

By taking the best of both data warehousing practices a new paradigm is possible however, one fundamental mindshift is required to bring the value of data closer to the surface for business self service, analytics and reporting. This mindshift is the breaking of the human paradigm of clustering data in a format that represents the data source (Transactions remain transactions, readings remain readings and functional records remain functional records) as this is the area that restricts data from being used in a more abstracted way.

To break this mould, we need to consider the incoming data as just that….. DATA rather than clustered structured content. When we do that, we are able to store the most granular elements of the data as individual pieces of a specific type and these types are finite in number irrespective of the business or source system.

This decomposition allows one to represent the incoming data clustered with data of like type from other sources meaning that the underlying data structures used to store the data remain stable and, with an element of business training, available for a less IT skilled user to consume.

CIntelligence (Patent Pending AU2016900704) utilises a combined practise of: -

1) *A 'single organisational context' to overcome nomenclature differences across different areas of the organisation (normalisation and standardisation of context);*
2) *A 'transactional de-composition' philosophy forcing the deconstruction of individual elements from source systems and thereby disassociating them for their original structural format constraints from their source of origination (information decomposition);*
3) *Recording elements of source data (as opposed to the whole record) within a common organisation (One-off build) context as one or more of a finite type of data;*
4) *Retention of time-based sensitivity for event driven recording; and*
5) *Multi-Dimensional representation of the same inputs allowing across source domain analysis and reporting.*

Using the above methodology, kudos must be provided to the originators for their contribution to the CIntelligence design as elements of each are evident. *From Inmon – a Single Organisational Context and from Kimball a Fact-Based Star/Snowflake schema model.*

As with these fore-running concepts, the CIntelligence product and methodology has been designed, implemented and made commercially available to enable greater consumption of enterprise data. Some elements of big data practice have also enabled this evolution in data warehousing, specifically MPP (Massive Parallel Processing), big data engines such as Hadoop and multi threaded/columnar data access and these, now common practices, allow the total capability of the new product(s).

To explain the above product in one sentence, one could label it as: -

'A functionally agnostic fine-grained Operational Data Store of factual detail (past, present and future) that represents any data source's specific elements in a single business context, irrespective of business type, source system or desired downstream use.'

Through adoption of the methodology a number of benefits are immediately realised by organisations namely:

- Stabilisation of the underlying data storage model means less discovery time for data engineers and scientists improving time to value for common functions to minutes as opposed to days or even weeks;
- Having a single structure filled with multi-faceted data allows an educated business user from any area of the business to self-serve in the reporting and retrospective modelling through any of the common Business Intelligence / Reporting Tool sets even if they have no understanding of the original source systems data structures and content;
- With a single structure representing the entire organisation and data no longer being constrained to source system structures, stored data never becomes redundant, so, in the event that source systems are changed, data from the original source system is stored side by side with that from the old system meaning zero redundancy over time;
- Having all data represented across multiple common type definitions, results in a single time perspective irrespective of granularity and the ability to remove a specific time dimension association of each record; and
- With all data conforming to a single context, business alignment across multiple business domains is achieved, this enables cross functional analysis and reporting to be performed with no data discovery time improving business turnaround.

How the integration is achieved is also derived from modern data interchange mechanisms, however, the maturity of an organisation in the data exchange domain will drive differences in implementation to achieve the same goals.

Simply put:

- a message structure is defined for source data expected and captured/loaded into the mapping engine;
- the input structure is then mapped using the organisations customised CIntelligence GUI mapping tool to its requisite destination(s) in the repository; and
- the CIntelligence custom services in the operating system that initiate the processing of that data on receipt.

The data interchange is a nuance change from existing ETL/ELT in that the separation of function is more explicit where data is 'Pushed' (P) at the source for delivery to the mapping 'Prepare' (P) engine which in turn delivers 'Load' (L) ready data elements. Implicitly, this PPL method isolation means that the 'L' component is only ever developed once, the prepare 'P' is an administrator level GUI-controlled function requiring little or no IT contribution and the Push is the only system specific development required for new data entering the repository. This process isolation aids organisations in delivering new data to the repository structure in time that is an order of magnitude faster that what business is used to. Experience has proved that improvements of up to 85% in turnaround time from source identification to active use of the data in the solution. We now talk in as little as hours to realise new content in the repository where history has typically been in days, weeks or even months.

Since the product is not dependent on a specific technology, it can be implemented with consideration of the existing investment in any current Business Intelligence / Reporting Tool sets from Teradata, Informatica, Cognos, Microsoft PowerBI, SAP Business Objects / Business Warehouse and even Amazon's S3 data lake.