

# Kognitio Pablo

Delivering real-time OLAP performance with Kognitio Pablo  
A paper by Dr. Mark Whitehorn



**Kognitio provides solutions to business problems that require acquisition, rationalization and analysis of large and/or complex data**

# Contents

1. Introduction	3	3. Kognitio Pablo	7
2. OLAP	4	3.1 Pro - intuitive interface	8
2.1 Pro - intuitive interface	6	3.2 Pro - speed	8
2.2 Pro - speed	6	3.3 Pro - single version of the truth	8
2.3 Con - multiple versions of the truth	6	3.4 Pro - pre-aggregation now a major pro	9
2.4 Con - pre-aggregation time	7	4. Implications of Kognitio Pablo	9
2.5 Con - stale data	7	4.1 Real-time analysis	9
		4.2 Fine control	10
		4.3 Snapshots	10
		4.4 Investment	10
		5. Summary	11

## About the author

Dr. Mark Whitehorn is a well-recognized commentator on the computer world, publishing articles, white papers and books. He has been writing about computers since 1987 and his column in Personal Computer World is one of the longest-running database columns in the world. He specializes in database technology, data warehousing and OLAP and has written eight books – several co-authored with Bill Marklyn, the designer of Microsoft's Access. Their first book, *Inside Relational Databases*, is a best seller (for a database book at least) and is now available in several versions. Mark's most recent book, *Fast Track to MDX*, was written in collaboration with Robert Zare and Mosha Pasumansky of Microsoft. Rob and Mosha are both heavily involved with Microsoft's Analysis Services product and Mosha was one of the originators of the MDX language.

**This paper is designed to give an introduction to OLAP, its pros and cons as well as Kognitio's approach to real-time analysis with WX<sub>2</sub> and Pablo**

# Collecting data is no longer an issue; the real trick is turning it into a business asset and extract the valuable information within it

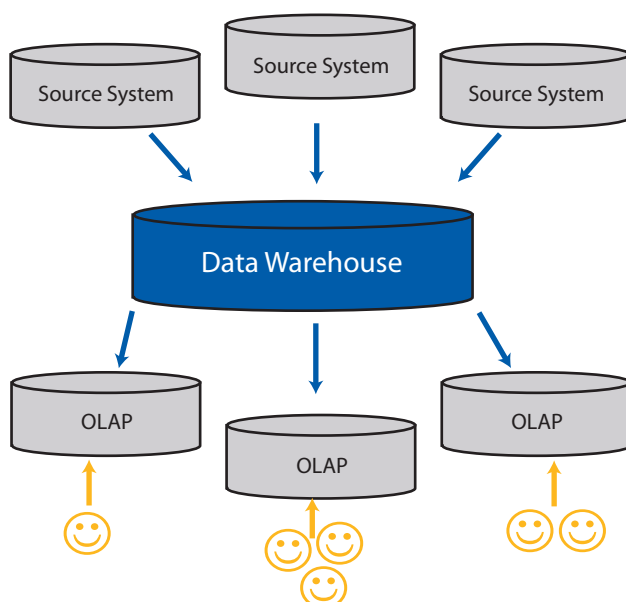
## 1. Introduction

Gone are the days when collecting data was a problem. Players in virtually every sector - telecommunications, healthcare, manufacturing, retail - collect quantities that only a few years ago would have been unimaginable. Collecting data is no longer an issue; the real trick is to turn it into a business asset, to extract the valuable information that's hidden within it.

Most of the data an enterprise collects is initially captured by the operational systems that run the business – the finance system, HR, the CRM system and so on. As a new employee starts work, or a new order is placed, the relevant data is created or amended. These operational systems are typically relational databases. You can think of relational databases as meticulous, completely trustworthy employees who sadly, with the best will in the world, are not the sharpest tools in the box. In other words, they will meticulously record what is going on, but you wouldn't ask them to give you an overview of what is happening. It isn't that relational databases are incapable of answering these complex questions, it's just that the reply takes far too long to arrive.

So, for analysis we typically pull a copy of the data from the operational systems and move it to a centralised repository called a data warehouse and from there into specific OLAP systems.

**Figure 1: A centralized data warehouse pulls data from sources and uses it to populate OLAP systems**



# Most OLAP systems are based around a multi-dimensional database

## 2. OLAP

OLAP (On-Line Analytical Processing) is, as the name suggests, not a specific technology – rather it is an umbrella term that embraces many different processes and techniques that allow us to extract information from data. Most OLAP systems are based around a multi-dimensional database. In physical terms a multi-dimensional database is described as a star schema. In terms of implementation it can either take the form of a fact table and a series of dimension tables held in a relational database engine (ROLAP), or of a set of data held in a purpose-built multi-dimensional database engine (MOLAP).

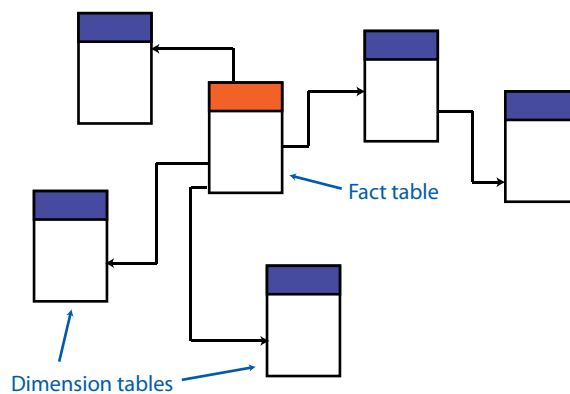


Figure 2: A star schema with a fact table and dimension tables

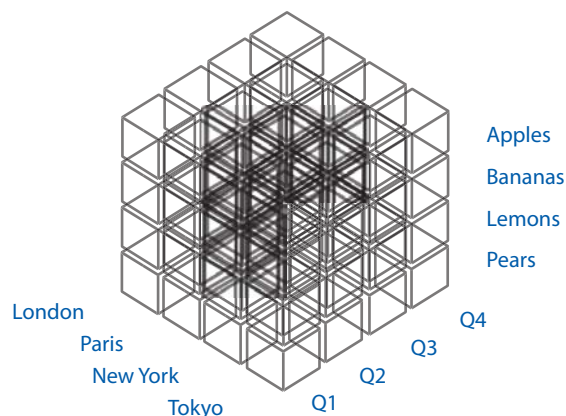


Figure 3: A 'cube' of multi-dimensional data: the cube isn't merely three-dimensional as this diagram shows – drawing further dimensions is tricky. As its name suggests, it has multiple dimensions.

# To make data analysis as fast as possible, OLAP data is typically pre-aggregated before it is queried

Whichever implementation is used, the result can be called a 'cube' of data, described as a multi-dimensional database; the descriptions below (about measures, aggregations etc) apply equally to either implementation.

Multi-dimensional databases make use of measures and dimensions. Measures are usually numeric; examples are cost price, sale price, number of items sold, profit. Dimensions are the factors by which the data is divided up during analysis, for instance time, customer or geographical region. Think of a bar chart like the one shown below: measures are the numerical values (e.g. Total Sales) typically found on the Y axis, the dimension (e.g. Time) appears on the X axes.

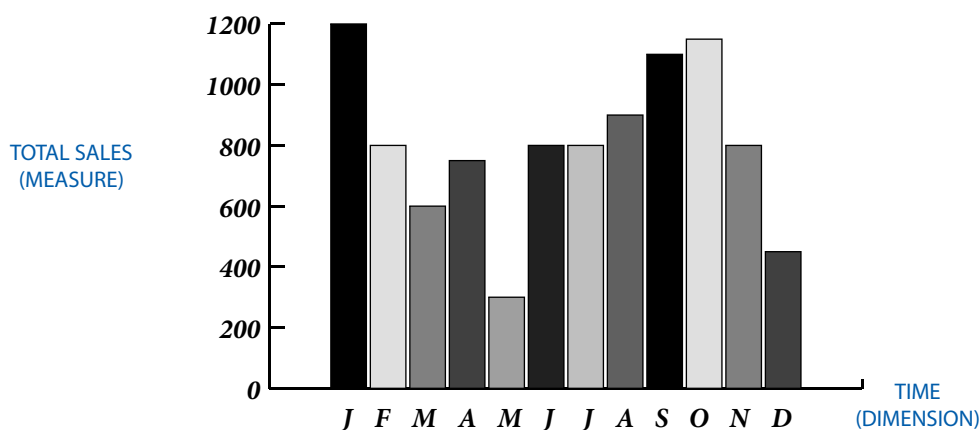


Figure 4: Total Sales is a measure and Time is a dimension, aggregated by month

Once the structure is defined, data is imported into it from the data warehouse and then pre-aggregated before it is ready for querying. Why does it have to be pre-aggregated? Well, dimensions are often hierarchical – take Time for example. Sales to customers may be recorded as having occurred on a particular day. However we often want to analyze sales by day, month, quarter and year. In order to allow this, and to make the analysis as rapid as possible, OLAP data is typically pre-aggregated before the users query it. So, when the user poses a question to the system, instead of having to calculate the answer, the system simply fetches the answer and sends it to the user. This is, in the main, why multi-dimensional databases are so much faster than relational ones.

So far, so good. In fact it's great; the entire process described above is relatively common practice and works very well, giving a huge advantage over trying to run complex analytical queries against the operational databases. However, like any system it has both pros and cons.

# MDX is the communications layer between the intuitive user interface and the underlying data structure

## 2.1 Pro - Intuitive interface

First among the pros is that intuitive interfaces are available which make it easy for users to formulate and run analytical queries against these OLAP systems. Various front-end analytical tools can be used for this, including Microsoft's Excel which will be familiar to many users, and third-party suppliers also offer bolt-on tools for Excel. A familiar interface flattens the learning curve for users and speeds the way to increased efficiency. Another pro is the adoption of the MDX (Multi-Dimensional eXpressions) query language: MDX is the communications layer between the intuitive user interface and the underlying data structure. Most users never want or need to learn MDX (or even to know it exists) but it is the standard communication layer between the user interfaces and the data.

## 2.2 Pro - Speed

The second major pro is speed - it's awesome. Response times of under five seconds are the norm, allowing 'train of thought' analysis. Answers to analytical questions are notorious for raising further questions: "Yes, I see that sales in the west shot up in June. Is that in all stores or in a localized area? Which products were popular? And was it green ones or pink ones?" When answers are presented so promptly, users can ask the further questions that are immediately formulated, gaining a far better understanding of the data and what it says about business activity. Research by IBM in 1984 demonstrated that speeding up response times significantly increases productivity<sup>1</sup>.

## 2.3 Con - Multiple versions of the truth

A con is the potential for multiple versions of the truth to arise. If multiple data sources are used and multiple OLAP cubes produced from them, then it is common for different definitions of, for instance, the dimension 'customer' to be produced. For example, one cube might contain a list of all customers; another a list of current customers (those who have bought something within the last two years). Even when these both cubes are supplied with accurate data about advertising spend, they will yield different values for 'advertising spend per customer'. In truth this is not an inherent problem with OLAP since it can be eliminated by careful control of definitions with the centralized BI (Business Intelligence) system. However, in practice it often is a major problem in many systems - the centralized control is either not there or is imperfectly applied. The net result is that many systems are plagued by 'multiple versions of the truth'.

<sup>1</sup> "Factors affecting programmer productivity during application development" A. J. Thadhani. IBM SYSTEMS JOURNAL, VOL 23, NO 1. P 19, 1984

"A comparative study of system response time on program developer productivity" G. N. Lambert. IBM SYSTEMS JOURNAL, VOL 23, NO 1. P 36, 1984

# Kognitio offers an alternative with WX<sub>2</sub> and Pablo

## 2.4 Con - pre-aggregation time

Another con is the time taken to pre-aggregate data. Once a large data set has been imported it can take hours before the aggregations are complete and the data ready for querying. Traditionally the importing and aggregating was performed during the quiet overnight period but in an increasingly 24/7 global business world 'overnight' may no longer exist.

## 2.5 Con – stale data

There is an inherent 'cost' (financial, temporal, processing etc) in building traditional OLAP cubes which means that, in practice, many companies tend to rebuild them less frequently than daily, perhaps weekly or monthly. This reduces costs but means that users often have to work with 'stale' data.

In summary, OLAP is hugely important to multi-dimensional analysis but with its undoubted advantages come inevitable disadvantages as outlined above. If we are looking for an alternative way to implement OLAP, in general terms the replacement has to be significantly better – if it isn't, there is no point in replacing the existing systems. The best result would be where all the pros are retained and one or more of the cons are converted into pros. In other words, an alternative solution is only useful if it increases the ratio of pros to cons.

## 3. Kognitio Pablo

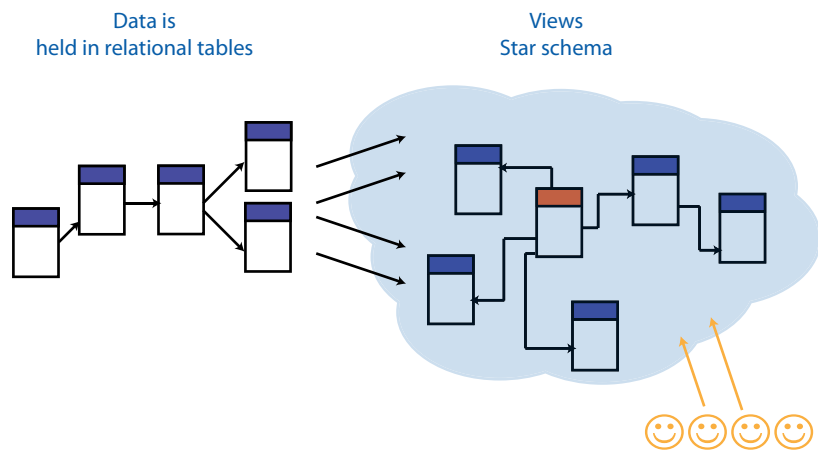
Kognitio offers an alternative that does just this with WX<sub>2</sub>, a fast in-memory analytical database and Pablo, its MDX connector. Essentially, the data is held not as a star schema but as relational data in a relational database engine (for reasons explained below). Kognitio WX<sub>2</sub> is conventional in the sense that it is relational but highly unconventional in that it can process analytical queries at mind-blowing speeds.

As well as the raw data in the relational tables, there is also a set of non-materialised views created on the data. (These views are essentially just queries that are stored within the database itself.) One of these views, when opened, shows the data organized into a classic fact table, the others correspond to the dimension tables.

Kognitio Pablo

© Kognitio Limited 2011

# Kognitio's MDX interface, Pablo, takes the OLAP model to a new level of efficiency, performance and agility



*Figure 5: The data is held in relational tables and the views turn that into a star schema*

In other words, the data is held as relational tables and the views provide a dimensional view of the same data. Creating such a set of views is perfectly possible in a conventional relational database engine, but it would be effectively useless because it would run insufferably slowly. However, when backed by a sufficiently fast relational engine, it becomes immediately apparent that this has the potential to significantly change the pro/con ratio for OLAP. If we examine those pros and cons again:

### 3.1 Pro - intuitive interface

The raft of good interface tools out there use MDX as the query mechanism. Kognitio's MDX interface, Pablo, takes the OLAP model to a new level of efficiency, performance and agility. Essentially this means that WX<sub>2</sub> can interact with all the usual suspects in terms of great UI tools for OLAP.

### 3.2 Pro - speed

Speed is as much a pro as it is for conventional OLAP: the speed of WX<sub>2</sub> and Pablo is broadly comparable to that of OLAP backed by a multi-dimensional database engine. It depends, of course, on the hardware, CPUs, memory and so on, but Kognitio states that on similar hardware similar performance can be expected.

### 3.3 Pro - a single version of the truth

Since each 'cube' is virtual and draws its data from the same central relational core (where each piece of data is held once and once only), the potential for manifesting versions of the truth is reduced. In practice this is not an absolute – it would still be perfectly possible to have multiple cubes using different definitions. However there is an important switch of emphasis; in the Kognitio approach it requires an active decision to make the definitions different.

# WX<sub>2</sub> offers real-time analysis. For many customers, this is a killer feature

## 3.4 Pre-aggregation becomes a major pro for Kognitio

However we've saved the best for last because the overwhelming advantage of this approach is that the pre-aggregation time is zero. Data is only ever aggregated on the fly by the views, which means that as soon as data is loaded into the database held in Kognitio's WX<sub>2</sub> relational database engine, it can be queried as a multi-dimensional set of data. However, there is more to this than just putting the cubes on-line as soon as the data is loaded, there are highly significant implications of working in this way.

## 4. Implications of Kognitio Pablo

### 4.1 Real-time analysis

We said above that Kognitio holds the data in relational tables. This is the conventional way to hold data in operational databases such as a finance or HR system. However, it is a very uncommon way to hold analytical data because such databases are typically far too slow when queried. Having solved the speed issue, Kognitio (and its customers) can capitalize on one of the major strengths of relational data – it is very easy and fast to update. And as soon as the data held in WX<sub>2</sub> is updated, because there is no pre-aggregation stage, the analytical queries run against it will return answers that reflect the new data. This is one of the most remarkable features of the WX<sub>2</sub> engine – it offers real-time analysis. For many customers this is a killer feature. Conventional OLAP systems simply cannot deliver real-time analysis. Instead the data is refreshed at regular intervals (typically overnight), pre-aggregated and then made available. So, even on first delivery, it can be 16 hours out of date and by the end of the working day will be 24 hours behind reality. Kognitio's WX<sub>2</sub> engine - with Pablo - offers to keep your analytics up-to-date by delivering real-time OLAP.

It is worth noting that in practice one problem with achieving real-time OLAP is that it is often difficult to assemble consistent data at the same time. As a trivial example, imagine that the order entry system runs with a lag of five minutes behind reality and the CRM system runs about four hours behind. If data is loaded when it becomes available, the potential exists for orders to appear in the analytical system for which there is no customer data. In practice this can be managed by determining the lag and artificially delaying the data load from the systems with the least lag and thereby providing a consistent data set. In other words, the ETL (Extract, Transform & Load) process can be adjusted to ensure consistent data.

# WX<sub>2</sub> stores data in relational tables and uses views to present that data to users

As discussed above, Kognitio stores the data in relational tables and uses views to present that data to the users. These views can be thought of as filters and formers that can extract and manipulate the data exactly as required. However it is important to note that the views themselves are simply descriptions of the desired result, they don't affect the underlying data. The same is true of software that you use to look at images in your digital photo archive. You can call up a picture and see it in sepia or black and white, crop it down and so on. Unless you save your work, these are just 'views' of your original picture.

The database views work in the same way and in practical terms we can have as many simultaneous views as we need. We can use them in a variety of ways, for example, to give us fine control over time and to provide specific snapshots of the data.

## 4.2 Fine control

In some cases, the system may be required to run in real-time but also be capable of presenting users with a consistent set of data over a particular period as well as showing them a constantly changing set. For example, the users may ask for one analytical set to remain static between 9am and 1pm. This is easily achieved with Kognitio's system: data can be loaded into the engine as and when it becomes available but the views that pull data from the engine can be defined to show only the data that is time/date-stamped prior to, for example, 9am today. At 1pm these views are replaced by an updated set which display data including the morning's newly loaded data. The process can be repeated at chosen intervals to suit prevailing business requirements.

## 4.3 Snapshots

We can also use views to show the state of the data at a particular point in time. There is no need to populate a structure with the appropriate data set or to re-aggregate, a snapshot can be achieved simply by a specific set of views.

## 4.4 Investment

Innovation that provides new features and also reduced complexity is rarely found in the bargain basement. If you really don't need any of the unique features that Kognitio's WX<sub>2</sub> engine and Pablo offers, then it makes sense to look elsewhere for your analytical solution.

# With Pablo, Kognitio has introduced a game-changing analytical environment

## 5. Summary

Companies now have access to a vast array of information sources and a need to collate and review them ever more quickly, 24 hours a day, 365 days a year. Traditional architectures struggle to combine the flexibility to analyze with the immediate need for the information.

With Pablo, Kognitio has introduced a game-changing environment that enables business users to analyze all of their information from their platform of choice without the need for costly and time-consuming transformations and builds. Fast, simple and available 24/7.

All of the information, when it is needed, where it is needed.

## About Kognitio

Kognitio is an innovative, technology-rich company, providing leading-edge solutions to business problems that require the acquisition, rationalization and analysis of large or complex data.

Kognitio's offering is centered around WX<sub>2</sub>, the fastest and most scalable analytical database on the market, which is available as a software license, a data warehouse appliance and via DaaS (Data Warehousing as a Service). With its industry-leading analytical database offering, WX<sub>2</sub>, Kognitio is able to rapidly turn a company's raw data into valuable business insight, empowering its customers to realize comprehensive answers to critical business questions.

Kognitio's DaaS model allows its customers to focus on running their businesses and increasing their bottom line. By also adopting Kognitio's outsourced approach, customers are able to reduce start-up time and costs, as well as avoid expensive product acquisition costs.

## About Kognitio WX<sub>2</sub>

Kognitio WX<sub>2</sub> is the most powerful and scalable analytical database in the industry. It enables organizations to query, in detail, vast amounts of granular data in seconds. The software-only solution uses high-speed, Massively Parallel Processing (MPP) technology to deliver an extremely fast data mart/warehouse platform to companies seeking to gain intelligence from their data.

Kognitio WX<sub>2</sub> runs on low-cost non-proprietary, industry-standard hardware, does not use indices or data partitions and can be scaled to handle hundreds of terabytes of data with performance that delivers answers in real time. This technology delivers the most comprehensive, cost-effective Business Intelligence database platform in the industry and offers delivery mechanisms from pure product through to fully managed services.

More than fifteen years of development have been focused on providing and refining the best tool for corporate Business Intelligence users to freely engage with ever-increasing volumes of data and/or disparate sets of data. Kognitio WX<sub>2</sub> enables business to work its data harder: to take more benefit out; in shorter times scales; with considerably less effort; and without the need for a complex large-scale IT installation. Tests have shown Kognitio WX<sub>2</sub> to run up to 60 times faster than typical databases and at a lower cost of ownership when compared to the lifecycle cost of other solutions.