

Emerging Technologies Shaping the Future of Data Warehouses & Business Intelligence

Appliances and DW Architectures

Zukeran
technologies

John O'Brien
President and Executive Architect
Zukeran Technologies

Agenda

- What is an “appliance”?
- Data warehouse appliances
- Current BI appliances
- Data Warehouse Architectures

What is an “appliance”?

How do you make toast and waffles?

General Purpose

Appliance-oriented



Zukeran
technologies

Copyright © 2004-2006 Zukeran Technologies Corp., All Rights Reserved

3

An appliance is born...

- Horizontal versus Vertical Layers
- When does it flip?

Compatibility	Applications (Applications, Operational, BI, DSS)	Purpose
	Applications (Databases)	Purpose Specific Appliance
	Operating System (Windows, Unix, Linux, etc)	
	Computing Resources (CPU, memory)	
	Networking Resources	
	Storage Resources (Persistence)	

Characteristics of a good appliance

1. Personalization vs. Configuration
 - “How dark do you like your toast?”
2. Few or no settings, options or configuration
3. Clear specific purpose
 - Designed and architected for a specific purpose
4. Not general purpose
5. Easy of use
 - “Just plug it in and it works...”
6. Compatible with existing infrastructure

Proven appliances

- Google search appliances
- Symantec VPN appliance
- Decru Security appliance
- Barracuda Email appliance
 - block viruses, spam, phishing attacks and other mail-borne security threats.

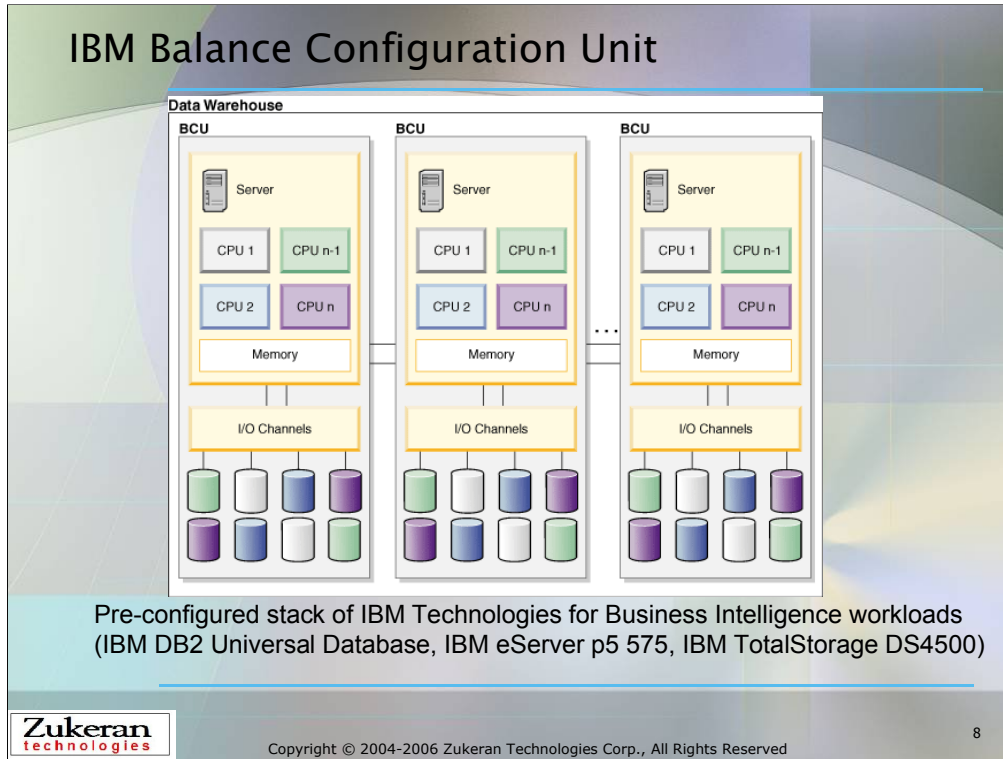


Pre-configured vs. Appliance

- Appliances are not a “pre-configuration” or assembly of tuned products and technologies
- An appliance is *built* for a specific purpose and can rely on pre-existing products or commodity parts
- Appliances have no user serviceable parts inside

Test: Does it matter what's inside the appliance?

- IBM Balanced Configuration Unit (BCU) is not an appliance but the benefits come from standardized preconfigured and pre-tested units.



IBM website:

"With the BCU, IBM has "one-upped" the data warehouse appliance makers..."

All the appliance advantages of price, performance and simplicity PLUS:

- Fully functional top-of-the-line DB2 data server enhanced for data warehousing, the IBM DB2 Data Warehouse Edition
- Built completely on industry-standard components and open systems
- Designed for modular scalability
- The IBM Data Warehousing Balanced Configuration Unit (BCU) reduces the complexity, cost and risk of designing, implementing, growing and maintaining a data warehouse and BI infrastructure.

A Balanced Configuration Unit (BCU) is composed of software and hardware that IBM has integrated and tested as a pre-configured building block for data warehousing systems. A single BCU contains a balanced amount of disk, processing power and memory to optimize cost-effectiveness and throughput. IT departments can use BCUs to reduce design time, shorten deployments and maintain strong price/performance ratio as they add building blocks to enlarge their BI systems.

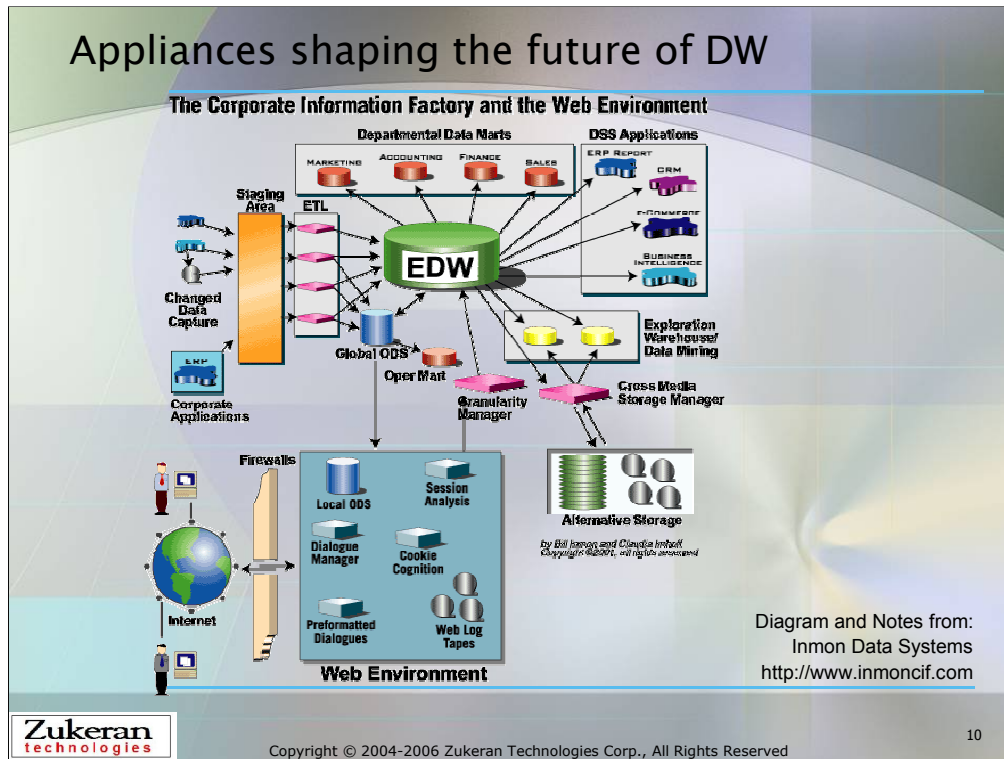
Future data warehouse appliances?

Will we see these specific appliances in the future?

Do they meet the criteria for success?

- ETL or ELT appliance ?
- BI appliance ?
- EII appliance ?
- OLAP appliance ?
- Audit appliance ?

Appliances shaping the future of DW



Operational Systems are the internal and external core systems that support the day-to-day business operations. They are accessed through application program interfaces (APIs) and are the source of data for the data warehouse and operational data store. (Encompasses all operational systems including ERP, relational and legacy.)

Data Acquisition is the set of processes that capture, integrate, transform, cleanse, reengineer and load source data into the data warehouse and operational data store. Data reengineering is the process of investigating, standardizing and providing clean consolidated data.

The Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data used to support the strategic decision-making process for the enterprise. It is the central point of data integration for business intelligence and is the source of data for the data marts, delivering a common view of enterprise data.

Primary Storage Management consists of the processes that manage data within and across the data warehouse and operational data store. It includes processes for backup and recovery, partitioning, summarization, aggregation, and archival and retrieval of data to and from alternative storage.

Alternative Storage is the set of devices used to cost-effectively store data warehouse and exploration warehouse data that is needed but not frequently accessed. These devices are less expensive than disks and still provide adequate performance when the data is needed.

Data Delivery is the set of processes that enable end users and their supporting IS group to build and manage views of the data warehouse within their data marts. It involves a three-step process consisting of filtering, formatting and delivering data from the data warehouse to the data marts.

The Data Mart is customized and/or summarized data derived from the data warehouse and tailored to support the specific analytical requirements of a business unit or function. It utilizes a common enterprise view of strategic data and provides business units more flexibility, control and responsibility. The data mart may or may not be on the same server or location as the data warehouse.

The Operational Data Store (ODS) is a subject-oriented, integrated, current, volatile collection of data used to support the tactical decision-making process for the enterprise. It is the central point of data integration for business management, delivering a common view of enterprise data.

Meta Data Management is the process for managing information needed to promote data legibility, use and administration. Contents are described in terms of data about data, activity and knowledge.

The Exploration Warehouse is a DSS architectural structure whose purpose is to provide a safe haven for exploratory and ad hoc processing. An exploration warehouse utilizes data compression to provide fast response times with the ability to access the entire database.

The Data Mining Warehouse is an environment created so analysts may test their hypotheses, assertions and assumptions developed in the exploration warehouse. Specialized data mining tools containing intelligent agents are used to perform these tasks.

Activities are the events captured by the enterprise legacy and/or ERP systems as well as external transactions such as Internet interactions.

Statistical Applications are set up to perform complex, difficult statistical analyses such as exception, means, average and pattern analyses. The data warehouse is the source of data for these analyses. These applications analyze massive amounts of detailed data and require a reasonably performing environment.

Analytic Applications are pre-designed, ready-to-install, decision support applications. They generally require some customization to fit the specific requirements of the enterprise. The source of data is the data warehouse. Examples of these applications are risk analysis, database marketing (CRM) analyses, vertical industry "data marts in a box," etc.

External Data is any data outside the normal data collected through an enterprise's internal applications. There can be any number of sources of external data such as demographic, credit, competitor and financial information. Generally, external data is purchased by the enterprise from a vendor of such information.

TDWI Exhibitors

NETEZZA
The Power to Question Everything™

Teradata
a division of  NCR


DATA AT THE SPEED OF BUSINESS

IBM®

Zukeran
technologies

Copyright © 2004-2006 Zukeran Technologies Corp., All Rights Reserved

11

BI Specific Appliance or Analytics Appliance



Copyright © 2004-2006 Zukeran Technologies Corp., All Rights Reserved

How do you make a BI specific appliance?

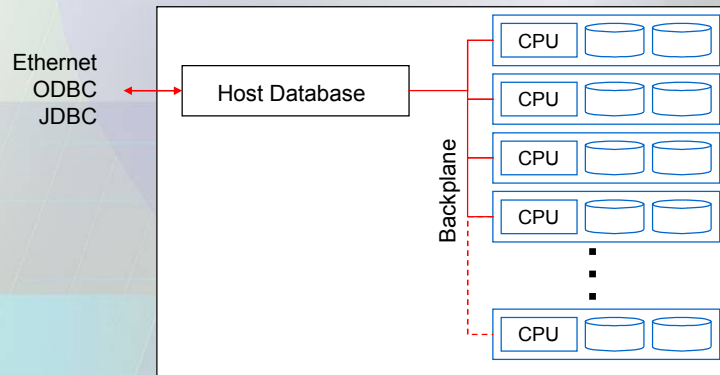
- Optimized for BI workloads
 - Store lots of data
 - Perform analytical queries
- Combination of Server + Database + Storage
- Ease of use
- Low maintenance and TCO
- Integrates with existing infrastructure
 - BI tools with ANSI SQL
 - Ability to load data easily
 - Compatible with standard operating procedures

BI appliances are here...

- ✓ High Performance & Price Performance
- ✓ On-Demand
- ✓ TCO and DW Budgets
- ✓ Scalability to petabytes
- ✓ Backup and Recovery
- ✓ Fast Loading and Unloading
- ✓ Licensing model
- ✓ Operations and Administration

BI Appliance: High Performance

- Massive Parallelism inherent to the architecture
- Improved I/O throughput rates with effective I/O



BI Appliance: TCO & DW Budgets

- Licensing costs
 - Appliance license replaces separate server and database licenses
- Reduced DBA expertise and administration
 - No tablespaces, extents to manage
 - No archive, redo logs to manage
 - No partitioning management
- Reduce need for storage expertise
 - SAN storage architects
 - LUNs, meta volumes

BI Appliance: Large Scale Databases

- 2TB to 100TB appliances are available today

How do you grow from 2TB to 100TB with appliances in the DW Architecture?

- Federated Architecture
- Data Redistribution

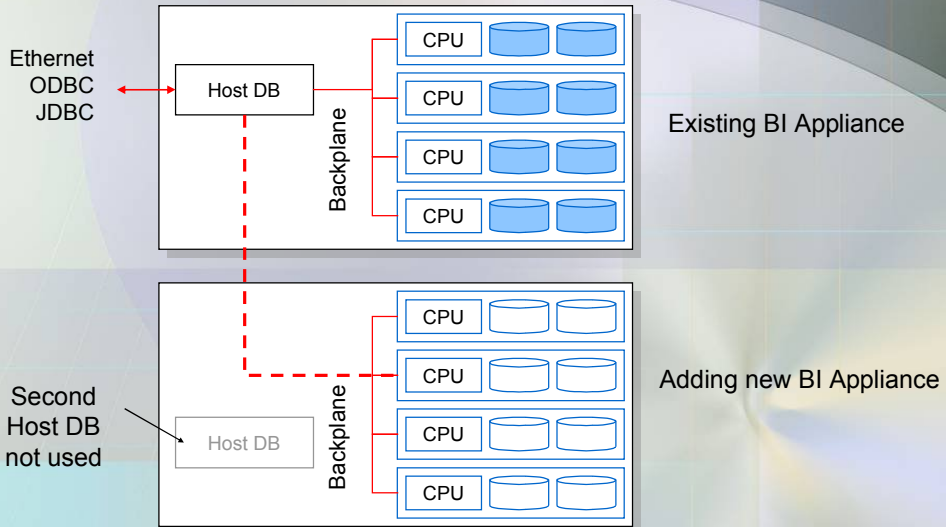
Scalability: Federated Architecture

- Logical areas of the data warehouse architecture move to new appliances
 - data marts, ODS, staging
- Subject areas or conformed dimensions stay together unless the capability to perform database joins across platforms exist
 - EII tool, remote tables
- Logical groups such as North America DW, Euro DW or corporate entities

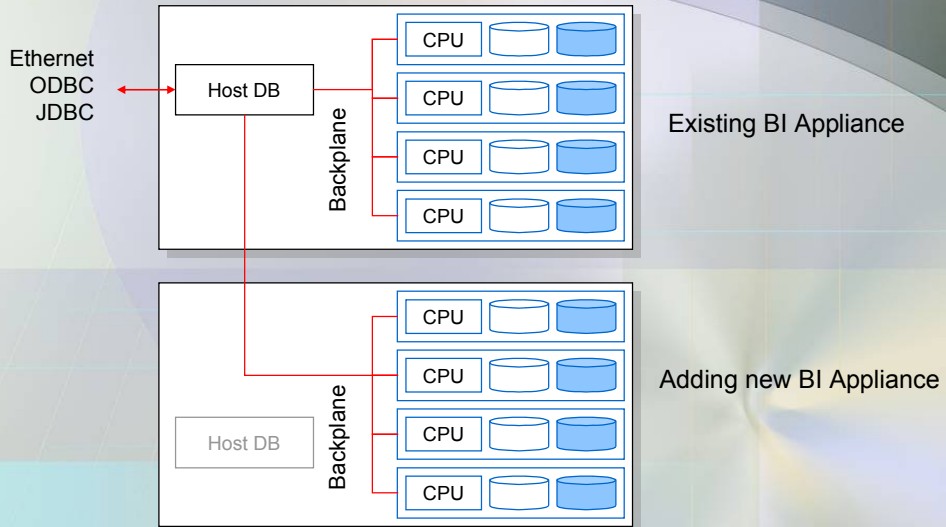
Scalability: Data Redistribution

- Adding additional appliances or upgrading to larger appliance may cause data to be redistributed
- Newer data can be loaded to the new appliance and allow the older appliance to become historical data.
- However, this is not an effective use of the appliance price-performance with relatively dormant data on high performance storage
- Appliance vendors should be able to evenly distribute data over more than one appliance and leverage a single host database architecture.

Scalability: Data Redistribution



Maintaining performance with scalability



BI Appliance: Backup and Recovery

- Built-in RAID 1+0 mirroring and striping
- The challenge is that large systems take massive amount of computing and network resources to backup changes in a high volume environment
- Recommended to compress and save load files since loads are faster than recoveries
- In a fully mirrored environment, chances of going to need to restore from a backup are low
- Veritas or other standard APIs are available backups

BI Appliance: Fast Load/Unload

- 500GB per hour load rates
- Near physical limitations disk I/O and Ethernet connection to appliance
- Utilizes high performance ODBC drivers and special loader utilities

BI Appliance: Licensing cost

- Appliance license cost
 - Annual maintenance cost as % of appliance cost
- Versus ---
- Server manufacturer annual maintenance cost
- Operating System license cost
 - Operating System vendor may be different from hardware
- Database license cost
 - Per CPU license
 - Per Concurrent user or Named user license
- Database options cost
 - Increases database maintenance cost as a %

BI Appliance: Operations

- How do they address:
 - User activity tracking
 - Query activity tracking
 - Explain Plans
 - Management Console
 - Data distribution profiles
 - Users, Roles, Privileges

What a fast database can do for you

- Before ETL tools
 - ETL was hand coded programs
 - ETL was code in database procedural languages
- ETL tools offered
 - Faster development with 3GL
 - Better performance than database code & SQL
- High Performance database appliance
 - Faster queries on large data sets
 - → High Performance SQL

ELT - A paradigm shift

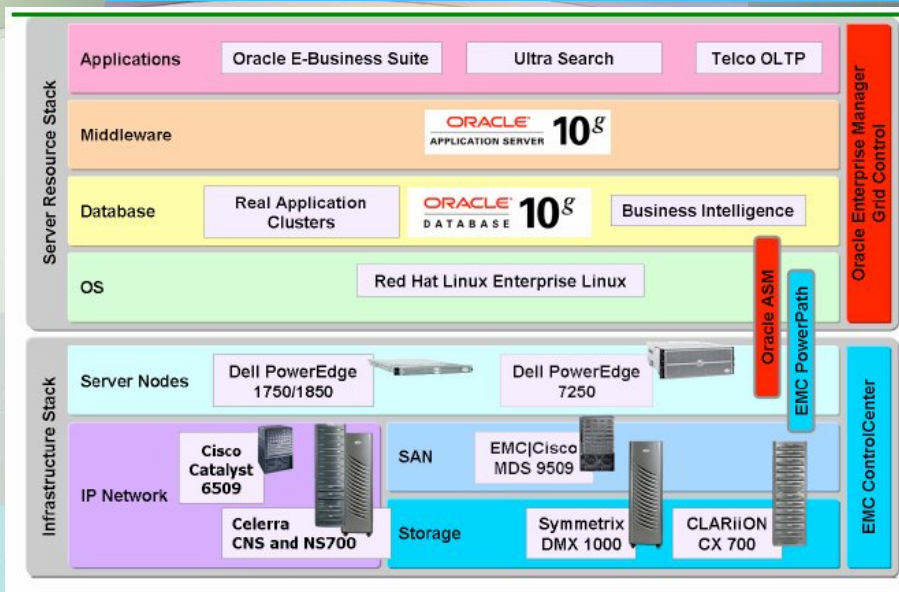
- Transformations typically are performed after Extraction and before Loading limiting the database workload which is thought to be query intense for data warehouses.
- High Performance BI appliances are being used to transform the data inside the high performance database and the results stored in the appliance or in other databases
 - Some companies point to ETL license savings as part of the ROI for appliances
 - Sunopsis is a tool being used in these cases

Sunopsis

ROLAP versus MOLAP

- Multidimensional database engines were created to make up for the performance deficiencies of relational OLTP databases at that time.
- MOLAP cubes are known for their high performance interactive capabilities, complex calculations and dimensional and hierarchical analysis but are challenged in:
 - Real-time environments
 - Large datasets
 - High degree of dimensional and hierarchical branches
- MOLAP cubes are also a duplication a data in another database format, the multidimensional database.
- ROLAP leverages on the underlying database for performance characteristics
- ROLAP metrics can be atomic level, pre-calculated, pre-rolled up depending on what works best

Oracle 10g RAC/GRID



Copyright © 2004-2006 Zukeran Technologies Corp., All Rights Reserved

Appliances don't do everything

- Data Architecture and Modeling
 - High performance databases shouldn't make up for poor data analysis and modeling efforts
- Good Requirements, Analysis and Reports
 - Make sure that the wrong answer doesn't just come back faster...

What the business wants from DW

- On-Demand
 - New products and projects
- Lower Cost
 - Overall TCO and long term maintainability
- Flexibility
 - Simpler infrastructure to build and go
- New capabilities
 - Scalability
 - Ability to store more data with less cost

Class Discussion