**IT@Intel White Paper**

**Intel IT**

IT Best Practices

BI Data Warehouse Strategy

March 2013

# Using a Multiple Data Warehouse Strategy to Improve BI Analytics

Intel IT's strategy for multiple business intelligence (BI) data warehouses enables our business groups to solve more high-value business problems, achieve greater operational efficiencies, and improve their competitive performance in global markets.

**Chandhu Yalla**
BI Engineering Manager/Architecture Owner
Intel IT

**Ajay Chandramouly**
Big Data Industry, Engagement Manager
Intel IT

**Charles Eden**
Senior Technical Integrator/Program Manager
Intel IT

## Executive Overview

**Intel IT is implementing a strategy for multiple business intelligence (BI) data warehouses to provide significantly more powerful analytics capabilities to business groups across Intel. By providing an array of BI platforms, we are helping Intel mine a broader range of data faster, deeper, and more cost-effectively. This expanded architecture enables our business groups to solve more high-value business problems, achieve greater operational efficiencies, and improve their competitive performance in global markets.**

For several years, Intel's BI needs were addressed with a "one size fits all" approach delivered by a centralized enterprise data warehouse (EDW). That solution is no longer adequate. With big data, data sets vary widely and are predominantly unstructured, complex, and in volumes that cannot be managed with traditional relational database methods.[1] To address these limitations, we have introduced this BI strategy to enable business groups to realize greater value from diverse and optimized data sets.

Intel's multiple BI data warehouses provide a dynamic range of BI analytic capabilities, including:

- EDW, for analysis of enterprise-wide structured data
- Apache Hadoop*, for analysis of raw, unstructured data
- Extreme data warehouse (XDW), for analysis of structured and semi-structured data
- In-memory, for real-time analysis of streaming volume data sets
- Custom, independent data warehouses, for analysis of structured, normalized data

Our multiple data warehouse BI strategy has enabled us to move from an expensive, one-size-fits-all approach to a more cost-effective, multi-tiered data warehouse architecture that better matches the business requirements and types of data available to our business groups. This strategy enabled us to avoid using EDW for security BI and Design (HSD) use cases, resulting in a cost avoidance of nearly USD 250,000 in the first year.

[1] "Insight Everywhere: The Growing Importance of Big Data and Real-time Analytics." January 2012

## Contents



## IT@INTEL

The IT@Intel program connects IT professionals around the world with their peers inside our organization – sharing lessons learned, methods and strategies. Our goal is simple: Share Intel IT best practices that create business value and make IT a competitive advantage. Visit us today at www.intel.com/IT or contact your local Intel representative if you'd like to learn more.

# BUSINESS CHALLENGE

**Performing cost-effective business intelligence (BI) analytics in an era of big data is an ongoing challenge for Intel and other organizations striving to improve their business decision making. Intel IT's overarching BI goal remains constant: provide the right data to the right people at the right time. Our BI strategy and its implementation are evolving to accommodate a wide variety of business use cases where BI can help solve high-value business problems with actionable insights in near real time.**

Our strategy is shifting as Intel and its business groups continue to collect an enormous, diverse, and rapidly expanding volume of external and internal data containing potentially valuable insights buried within it. A major portion of this data is large and unstructured, creating up to 90 percent of enterprise data.[2]

The ability to mine and analyze data in various forms from many sources gives us deeper and richer insights into business patterns and trends. It helps drive operational efficiencies and competitive advantage in manufacturing, product groups, security, marketing, and IT. At Intel IT, we strive to provide the optimum pairing of business-group BI requirements with the technologies that can perform the tasks most efficiently.

## Merging Diverse Data

Over the past decade, Intel's decentralized enterprise resource planning (ERP) system was aligned to the various lines of business (LOB). While these separate data warehouses satisfied many business requirements, as a group they were inconsistent in collecting and storing data. For example, the formats for product and customer names varied across different databases, making it difficult and costly to share data between groups. As a result, the data was shared only rarely.

To provide a more comprehensive and accurate view of the company, we implemented a centralized enterprise data warehouse (EDW) to deliver broader and more sophisticated BI reporting and data analysis for guiding business decision making. We have allocated the bulk of our BI investments to front-end tools and data management technology to help improve the integrity of our data.

Using a centralized EDW has brought new levels of data standardization to the company, strengthening our BI capabilities. For example, having consistent product-name conventions and other identifiers have helped us achieve "a single version of the truth" and enable us to merge data from diverse sources to gain business insights from the converged data.

This centralized architecture has proved effective at Intel, with the EDW a stockpile for aggregating all enterprise analytics data, regardless of use case. But the EDW's key limitation—the inability to deal with raw, unstructured, and semi-structured data—has become more evident in recent years.

## Big Data's Impact

Across the global IT industry, big data is prompting a reevaluation of BI data warehouse architecture.[3] The issue is how to effectively manage data sets whose volume, variety, and velocity are beyond the ability of traditional database tools to analyze the data.

A new generation of big data tools is capable of collecting, processing, and analyzing unstructured and semi-structured data in a timely manner, which means businesses can derive meaning from previously unexplored data sets. With this ability, they can achieve deeper and richer insights than were previously possible in the traditional EDW.

At Intel, our own EDW continues to provide valuable business reporting and ad hoc querying with its structured data, but a deluge of big data at Intel is escalating both storage

[2] "Mining Big Data in the Enterprise for Better Business Intelligence," Intel IT white paper, July 2012.

[3] "Data Warehousing Architecture Takes Logical Turn in Big Data Era," Search Business Intelligence, November 2012.

and processing costs. Here, as elsewhere, the introduction of expensive big data platforms has sometimes proved tempting to our business groups, even though the specific BI requirements may not justify the cost.

To offer the optimum solution that best matches BI requirements to platforms, Intel IT's strategy is to provide a range of BI data warehouses that can accommodate a variety of needs across Intel's business groups. We have extended our BI strategy to include solutions from traditional relational databases, Not Only SQL (NoSQL) database systems, data warehouse appliances, data marts, and big data tools—all co-existing with the EDW as a key part of our multi-tiered BI data warehouse strategy.[4]

## SOLUTION

**Expanding our BI strategy to include multiple BI data warehouses provides Intel business groups with a broader range of BI solutions to support diverse business needs. The key benefits include accelerating the decision making process with insights mined from much larger data sets, while promoting agility at a much lower cost.**

[4] "Enabling Big Data Solutions with Centralized Data Management," January 2013.

Our strategy takes into account that knowing what big data can mean to Intel—and managing it as a corporate asset—is more important than consolidating it in a more traditional, single data warehouse. Expanding our data warehouse architecture uses the value of the EDW for shared enterprise data, yet also extends BI benefits to cases where the unstructured data is evolving, requires special handling, or is focused on a limited audience.

Our EDW remains an important part of our BI strategy. This central data warehouse is optimum when business groups seek cross-functional, integrated views of the enterprise data. At the same time, it provides a single interface between users and the data, making it easier to get to the necessary information and to develop BI solutions requiring consolidation.

In addition to the EDW, our multi-tiered data warehouse strategy enables us to optimize the BI solution when the data and results are for a single business group. We accomplish this by selecting the type of warehouse that best fits the group's business requirements; for example, in cases where process and data changes frequently. Employing various BI platforms—when compared to a single EDW solution—can reduce capital costs while offering more rapid development and local control. Our strategy has enabled us to avoid using the EDW for LOB-specific security BI and Design (HSD) use cases, resulting in a cost avoidance of nearly USD 250,000 in the first year.

However, BI data warehouses capable of tackling big data solutions are not the optimal solution in every BI use case. For example, depending on the use case, it is often more expedient to keep data in a data warehouse close to the current transaction system and data users, minimizing latency problems and the potential failure points that come with each additional data movement.

### Guiding Principles and Cooperation across Business Groups

Intel IT employs a consistent, data-driven process that encourages cross-organizational collaboration for business, data, reusability, architectural, and support requirements. As shown in Figure 1, we work with our business groups to identify the optimum BI data warehouse for their specific requirements. Maintaining this multi-tiered data warehouse architecture means a single, standard process for BI activity is no longer a restriction.

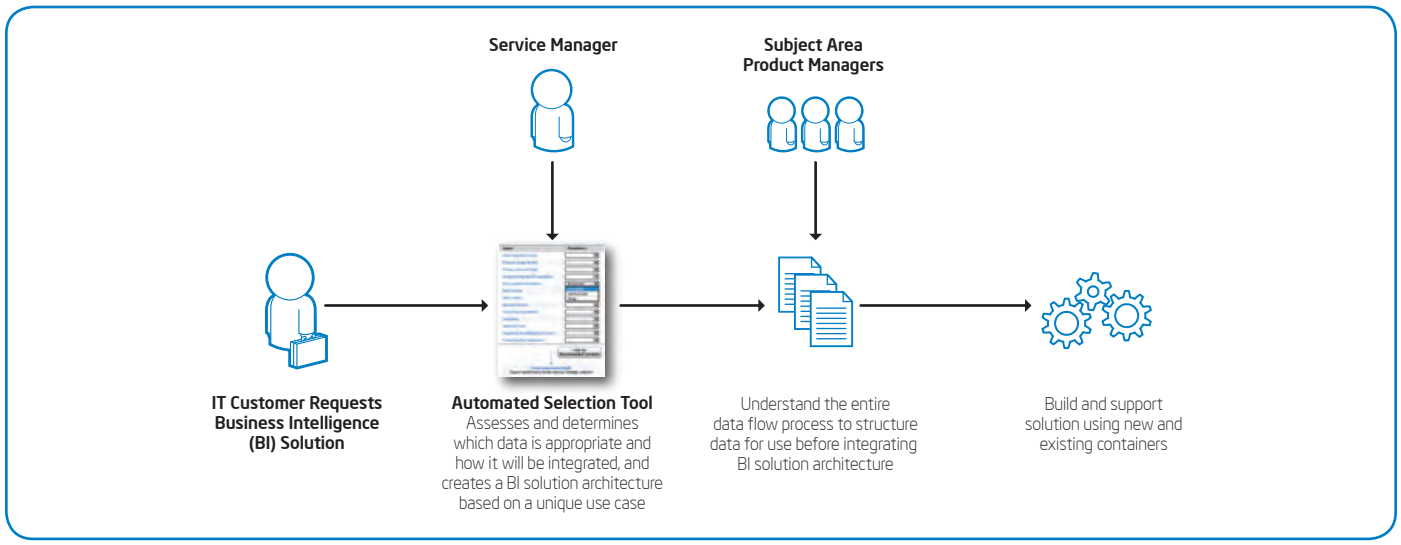


Figure 1. We use a proven engagement process with business groups, to determine the best business intelligence data warehouse for a particular use case.

Using multiple data warehouses requires keen management and oversight from Intel IT working with the business groups for the following reasons:

- Selecting a BI data warehouse without complete analysis can result in sub-optimal performance.
- Governance and agility varies according to the specific BI data warehouse.
- Decisions about the use of a particular BI data warehouse may not serve larger cross-organizational needs.
- BI solutions often involve multiple groups making decisions.

Put simply, there is a downstream effect for every decision made regarding selection of an appropriate BI data warehouse. Having clear policies in place for defining and managing all types of data is a critical first step.

## The Process for Optimizing BI Data Warehouse Selection

Our IT Business Intelligence Management team standardizes and centralizes the collection, storage, processing, and distribution of the information Intel needs to run its business.[5] This team strives to achieve the best BI procedural decisions for Intel's business groups, employing a consistent and data-driven process. This enables us to address the business process, value, architecture, data warehouse opportunities, and size of the investment necessary to support each particular BI use case.

When we engage with an IT customer, our first step is to fully understand the business requirements. We ask a series of questions to determine the specific BI data warehouse that is the optimal match for these requirements (Figure 2).

For example, it is not always necessary to move data. If all relevant data for a BI solution resides in the operational data store, or business warehouse, it should be used there and not moved to the EDW. If all the data required for visibility, aggregation, reporting, or a dashboard already is in a single location, it should be processed there.

A series of basic considerations can quickly narrow the data warehouse options.

- Is the information time-critical?
- Can the solution be delivered from the source?
- Can the data be sourced from the group's operational database solution?
- Does the data need to be integrated with other groups' data?
- Does security of the data require encryption or is it privacy-related?
- Does the solution require big data capabilities?

If integration with other data sets is required, the value of integrating that data with other enterprise subject areas in the EDW is considered.

To support our data-driven approach, we use an automated selection tool to help determine the BI data warehouse best suited to the requirements of the business use case. On the next page, Figure 3 shows the automated selection tool containing the type of information we collect to determine the optimal BI data warehouse for specific use cases.

[5] "Enabling Big Data Solutions with Centralized Data Management," January 2013.

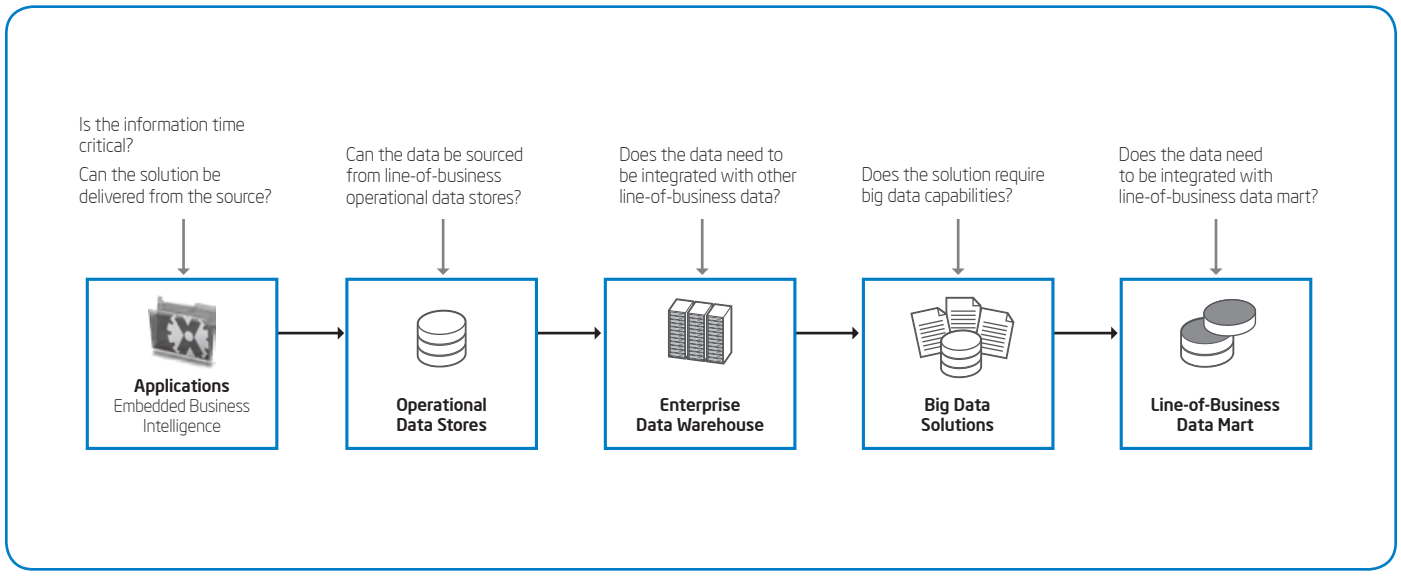


Figure 2. We ask a series of questions to determine which business intelligence data warehouse is the optimal match for a specific business use case.

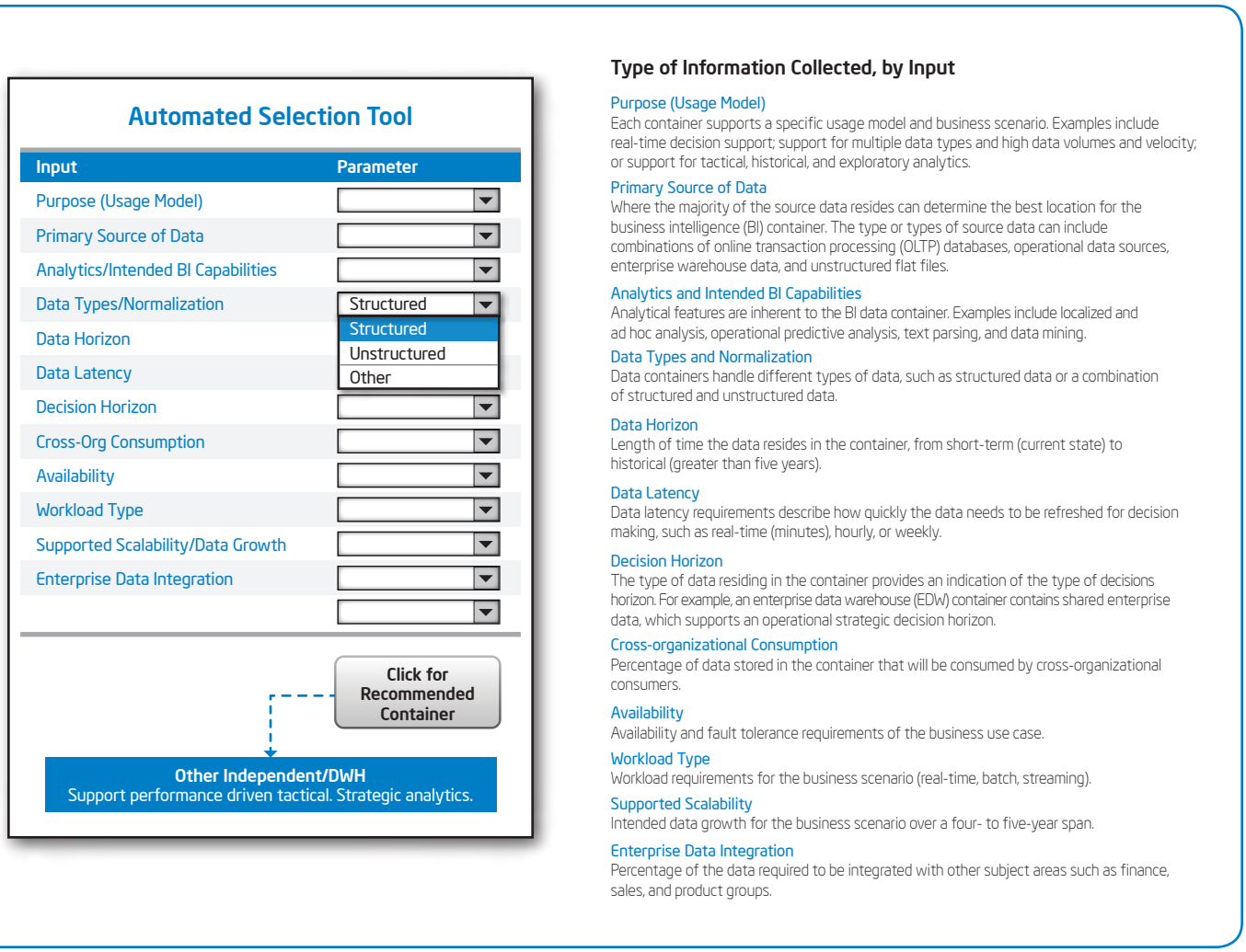


Figure 3. Our automated selection tool accelerates the decision making process for determining the optimal business intelligence data warehouse for a proposed use case. Changing the attributes can change the specific data warehouse best suited for the use case.

## Matching BI Data Warehouse Attributes to the Business Use Case

The selection of available BI data warehouses is based on providing business value while managing costs. To enable Intel to make the most efficient and cost-effective use of its data resources, our selection of BI data warehouses supports various levels of agility, performance, costs, consumption, and data types and states. Our strategy includes data platforms ranging from relational databases and BI data warehouse solutions to big data platforms.

These options include the following:

- EDW, for analysis of enterprise-wide structured data
- Apache Hadoop*, for analysis of raw, unstructured data
- Extreme data warehouse (XDW), for analysis of structured and multi-structured data
- In-memory, for real-time analysis of streaming volume data sets
- Custom, independent data warehouses, for analysis of structured, normalized data

The following sections and Table 1 (on the next page) summarize the attributes and characteristics for each BI data warehouse option.

### EDW: FOR ENTERPRISE-WIDE STRUCTURED DATA

We use the EDW to accommodate highly shared enterprise data and whenever a BI solution requires cross-application views or business data integration. EDW offers superior performance and serves as a data hub for downstream data mart and reporting solutions.

This BI data warehouse is primarily batch driven, with support for near real-time analysis.

### APACHE HADOOP*: FOR RAW, UNSTRUCTURED, SENSOR-TYPE DATA

We use the Intel® Distribution for Apache Hadoop software (Intel® Distribution) for strategic predictive analytics, text data mining, and behavioral analysis. Increasingly, it is the preferred BI data warehouse for large volumes of variable, multi-dimensional, structured, or unstructured data, or for analysis of multi-structured data where join conditions are unknown and the goal is to detect patterns.

The Intel Distribution has significant advantages, including support of security integration, high availability, and multi-tenancy conditions. It offers streamlined setup, management, security, and troubleshooting for Apache Hadoop clusters, as well as integration with existing management and analysis tools.

### XDW: FOR HIGH-VOLUME, STRUCTURED DATA

Intended for predictive analytics and operational reporting, we use the XDW for BI analysis requiring large volumes of data, where the goal is high performance at a lower cost than the high-end compute power of the EDW.

This BI platform is well suited for a LOB solution where governance and control are less than those found in an EDW solution.

This BI data warehouse supports integrated, normalized, granular, and historical master and transactional data, as well as both streaming and batch-driven solutions where data latency is dependent on the solution design. It can function as an alternative data warehouse for subject areas requiring special handling of classified or sensitive data not shared outside of a specific business group.

### IN-MEMORY: FOR REAL-TIME ANALYSIS OF STREAMING DATA SETS

Although we have not yet deployed an in-memory BI solution, we plan to use this solution for business use cases that require quick results to meet specific requirements of high business value. As an independent data mart with vendor-supplied applications that utilize an in-memory database, we recommend this type of solution to our business groups when extreme query performance is required and sub-second update latency is desired.

High expense may be a limitation of this data warehouse, due mostly to the condition that all data must fit into main memory—5x data compression is common—with a maximum of tens of terabytes. However, its unsurpassed velocity may also yield the highest business value due to its ability to enable faster and better-informed business decisions.

### INDEPENDENT DATA WAREHOUSES: FOR CUSTOM BI ANALYTICS

Intel business groups with custom requirements can rely on generic, independent BI data warehouses established for specific purposes, such as for ERP business warehouse operational data.

These solutions can produce ad hoc predictive analytics, formatted reports, dashboards, and write-back. Prospective business-analysis scenarios include HR sensitive data, credit reporting, and factory data, with the BI warehouse either self-managed or with central IT support.

Table 1. Comparative attributes of the various business intelligence data warehouse options at Intel

| | Enterprise Data Warehouse (EDW) | Big Data Warehouse Using Apache Hadoop* | Extreme Data Warehouse (XDW) | In-Memory DW in progress | Independent DWs |
|---|---|---|---|---|---|
| Positioning and Intended Use | Highly shared enterprise data requiring cross-organizational integration | LOB data | High-volume, LOB data; some shared data | Enterprise shared data based on combined OLTP/DW workloads | LOB data (business warehouses and operational stores) |
| Relative Performance | Best for structured data | Best for unstructured data | Better | Excellent | Good |
| Agility Factor | Slower (highly governed) | Highly agile (low governance) | Agile (medium governance) | Depends on OLTP source | Depends on the platform |
| Data Type/Normalization | Structure, denormalized | Raw/structured, unstructured data | Structured, denormalized | Structured hybrid or column | Structured, normalized |
| Real-time versus Batch Analytics | Batch/near real-time | Batch | Mix streaming/batch | Real-time/batch | Near real-time/batch |
| Historical Data Horizon | Long-term horizon; >5 years | Short- to very long-term horizon; <3 years to >5 years | Mid-term horizon; 3 to 5 years | Very short-term horizon; 6 months to 1 year | Mid-term horizon; 3 to 5 years |
| Supported Scalability | < 300 TB | >1 PB | >500 TB | <20 TB | < 2 to 5 TB |
| Summarized Data | Raw and summarized | Mostly raw | Raw and summarized | Very little to high | Very little to high |
| BI Capabilities Usage | ▪ Data mining<br>▪ Ad hoc<br>▪ Formatted reports<br>▪ Dashboards<br>▪ OLAP/MOLAP | ▪ Text parsing<br>▪ Data mining<br>▪ Temporary data<br>▪ Web<br>▪ Sensor ad hoc<br>▪ Sandbox<br>▪ Predictive Analytics | ▪ Data mining<br>▪ Ad hoc<br>▪ Predictive analytics<br>▪ Formatted reports<br>▪ Dashboards<br>▪ ROLAP/OLAP | ▪ Predictive analytics<br>▪ Formatted reports<br>▪ Dashboards<br>▪ HOLAP/OLAP | ▪ Ad hoc<br>▪ Formatted reports<br>▪ Dashboards<br>▪ Write-back<br>▪ MOLAP/OLAP |

**BI**-business intelligence; **DW**-data warehouse; **EDW**-enterprise data warehouse; **HOLAP**-hybrid online analytical processing; **LOB**-line of business; **MOLAP**-multidimensional online analytical processing; **OLAP**-online analytical processing; **OLTP**-online transaction processing; **PB**-petabyte; **ROLAP**-relational online analytical processing; **TB**-terabyte; **XDW**-extreme data warehouse

## CONCLUSION

**Providing multiple BI data warehouses greatly expands the ability of business groups across Intel to mine the enormous amounts of raw and unstructured data. Matching the use case with the most appropriate BI platform is enabling us to achieve substantial cost savings.**

The previous approach of relying on a single, centralized data warehouse became both costly and limited for our expanding BI needs, so revising our BI strategy to accommodate multiple data warehouses can significantly enrich the decision making process across the company and enhance business performance.

In the several business use cases where this strategy has been employed, BI solutions have been generated avoiding the use of the more-costly EDW platform. We will continue to analyze the financial benefits, basing them on the overall benefits derived from these use cases.

Since introducing the multiple data warehouse strategy, we have been able to document savings. For example, by applying the strategy, we have been able to avoid using the EDW for LOB-specific security BI and Design (HSD) use cases, producing a cost avoidance of nearly USD 250,000 in the first year.

With our BI options significantly expanded, it is important that business groups work closely with Intel IT so their BI use cases are executed accurately, quickly, and with the least amount of cost.

## RELATED TOPICS

**Visit www.intel.com/it to find white papers on related topics:**

- "Enabling Big Data Solutions with Centralized Data Management"
- "Insight Everywhere: The Growing Importance of Big Data and Real-time Analytics"
- "Integrating Apache Hadoop* into Intel's Big Data Environment"
- "Mining Big Data in the Enterprise for Better Business Intelligence"

## CONTRIBUTORS

John Martin, Intel IT
Moty Fania, Intel IT
Craig Chvatal, Intel IT
Alan Gonsalves, Intel IT

## ACRONYMS

| | |
|---|---|
| BI | business intelligence |
| DW | data warehouse |
| EDW | enterprise data warehouse |
| HOLAP | hybrid online analytical processing |
| LOB | line of business |
| MOLAP | multidimensional online analytical processing |
| NoSQL | Not Only SQL |
| OLAP | online analytical processing |
| OLTP | online transaction processing |
| PB | petabyte |
| ROLAP | relational online analytical processing |
| TB | terabyte |
| XDW | extreme data warehouse |

**For more information on Intel IT best practices, visit www.intel.com/it.**