IT@Intel White Paper

Intel IT
IT Best Practices
Enterprise Data Management
January 2013

Enabling Big Data Solutions with Centralized Data Management

Ajay Chandramouly
Big Data Industry Engagement Manager
Intel IT

Kelly Stinson
Data Management Program Manager
Intel IT

Executive Overview

In 2010 Intel IT created an IT Business Intelligence Data Management team to standardize and centralize the collection, storage, processing, and distribution of the information Intel needs to run its business. By architecting solutions that make enterprise data readily available for use across the company, we enable Intel to make more effective and extensive use of business intelligence (BI) tools to drive operational efficiency and competitive advantage.

By architecting solutions that make enterprise data readily available for use across the company, we enable Intel to make effective and extensive use of business intelligence tools to drive operational efficiency and competitive advantage.

Achieving business success in today's global economy requires the ability to make the right data available in the right form to the right people at the right time. In many organizations, this is difficult because each division collects and processes its own data differently. This makes it a challenge to know exactly what information is available, where to get it, and how to share it.

Having a dedicated data management team enables Intel to make its enterprise data available to all parties. Advantages of our approach include:

- A single source for data, providing easier access, integration, and cross-referencing of data.
- A customer engagement model that enables architecting data solutions that make the most intelligent use of enterprise data warehouses, big data platforms, and other solutions to address data volume, variety, variability, and velocity.
- Self-service solutions that give IT customers faster access to data and more autonomy to discover new possibilities and glean greater insights.

Through the centralization of data management, Intel IT is bringing greater agility, efficiency, speed, accountability, and consistency to how Intel handles enterprise data, improving how business groups access and use it for BI. We plan to continue improving our responsiveness to our IT customers and implementing new data platforms and self-service BI solutions that enable them to extract greater business value.

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

The external and internal data Intel collects and uses to run its business and compete in a variety of global markets continues to grow in volume, value, and importance. Cost-effectively extracting as much value as possible from this data efficiently and accurately for more than 95,000 employees presents a constant and considerable challenge.

To deliver new business intelligence (BI) to Intel's operations through rich analytics, Intel IT needs the ability to cross-reference seemingly unrelated data sets in new ways. At Intel, this was difficult prior to 2010 because even though master data was available and utilized to operate the company, each business group managed their own data and shared it among the other groups only when requested, which frequently led to costly duplication of efforts.

In addition, the generators of a particular set of data, such as our Sales and Marketing group, often didn't recognize the potential value of the data for another group, such as Finance. Equally troubling, business groups had no incentive to share data and thus provided little information on what data they possessed.

Another frequent issue involved data formatting. The transactional systems or other data sources used by a business group rarely delivered the data in formats that met the requirements for how the data was used by other business groups. Naming conventions for products, customers, and other data often varied, plus the use of non-standard and cryptic field names made it difficult to integrate and use data from multiple sources. The attempts that were made to pull data from a source and convert it for the needs of a downstream IT customer often resulted in many copies and forms of the same data. Because of the duplicate copies, the next IT customer that accessed the data frequently had trouble determining which data was definitive.

And with enterprise data now being measured in petabytes, the rising cost of data storage and processing is a significant factor. Similarly, with the increase in the amount of data and the advent of new technologies such as big data platforms, business groups often requested expensive technologies when their data didn't actually justify such a solution.

SOLUTION

To centralize the collection, storage, processing, and distribution of the information Intel needs to run its business, Intel IT formed an IT Business Intelligence Data Management team. Working closely with business groups, this team adopts the IT customer's viewpoint to make enterprise data available and usable at the highest quality level.

As part of the Intel IT Business Intelligence group, the IT Business Intelligence Data Management team's goal is to drive the best data management solution decisions for our customers through a consistent, data-driven process that facilitates cross-organization collaboration around business, data, reusability, architectural, and supportability requirements. The team functions as a data

broker between data source system owners and a broad spectrum of enterprise data customers. Responsibilities include creating and maintaining a trusted data management service that provides enterprise data governance and support. In addition, the team provides data management services for the latest advanced BI solutions our customers are using to find high-value line-of-business (LOB) opportunities. Examples of these solutions include advanced statistical techniques, data and text mining, and game theory.

A Single Source for Enterprise Data

Rarely does an IT customer want data from only a single subject area or business group. Most want to integrate and analyze data from several subject areas, such as sales orders, inventory, and location, for insights that help them make decisions. By providing a single source for this data, the IT Business Intelligence Data Management team helps save these customers from having to source the information themselves from various business groups and deal with disparate data types and naming conventions that make data difficult to interpret and use.

A big part of our single-source service is ensuring the use of the same master data through standardization of product names and other identifiers. Such standardization in field names saves considerable time compared to the complex mapping required when a group tries to use data named one thing in, for instance, sales orders and named something else in, say, inventory. To determine what master data is needed and establish naming conventions, the IT Business Intelligence Data Management team works with capability segment teams that are responsible for delivering web, application, and BI solutions in accordance to IT operational priorities. This collaboration results in an approved list of product names, customer names, and other identifiers that help ensure the ability to integrate data.

Having a standard "single version of the truth" eliminates the contradictions in numerical figures and other data from two different business groups, ultimately improving the accuracy of Intel's BI.

Right-Sizing the Solution to the Data

From a technical architecture perspective, getting the right data to the right customers at the right time makes all structured enterprise data available from the data warehouse. Data would be piped into the data warehouse, organized by subject area, and then created in the appropriate views to be used by business groups.

There were several problems with this approach:

- Increasingly, the enterprise is seeking to use unstructured, or big data, for BI. This requires a different platform than a traditional data warehouse.
- Many of our IT customers were already using data of variable quality and not using similar naming conventions to create subsequent versions. Transitioning these customers would mean shutting down their data warehouses and reengineering their data. This is time-consuming, disruptive and expensive, particularly for data the company is depending on for LOB operations.
- Keeping data already residing in a container close to the current transaction system and data users is often more expedient, minimizing latency problems and the potential failure points that come with each additional data movement.

Big Data, Big Opportunities

The term big data broadly describes collections of data sets that challenge traditional relational database approaches by either sheer volume or complexity. For example, even a few terabytes of unstructured data containing multiple data types from a variety of sources can be considered big data.

Big data tools, such as the open source solution Apache Hadoop*, enable the collection, processing, and analysis of large, heterogenic data sets in a timely manner. These tools enable organizations to derive meaning from previously unexplored sets of unstructured data, achieving deeper and richer insights compared to analysis of aggregated, partial snapshots of these data sets.

Increasingly, organizations are realizing that one size or solution does not fit all when it comes to handling structured and unstructured data. This means they may need to utilize an assortment of solutions ranging from traditional relational databases, NoSQL (generally interpreted as meaning "not only SQL") database management systems, data warehouse appliances, and big data tools.

Using and integrating these technologies to deliver business insights is a challenge for any modern organization. It requires new IT skills in Linux* and Java*, as well as rethinking the problem in terms of parallel computing constructs when working with big data.

Intel IT sees maximizing our ability to process big data and obtain actionable insights in near real-time as a way to solve high-value business problems, achieve operational efficiencies, and increase our competitive advantage. To quickly and effectively increase our ability to harness big data, we are developing in parallel both a big data methodology and center of excellence to acquire the necessary skills in solution design, engineering, development, administration, and data visualization.

Apache Hadoop* at a Glance

Hadoop is an open source framework for writing and running distributed applications that process large amounts of data. Instead of requiring one large supercomputer, Hadoop enables the coordination of local storage and computation across multiple servers that act as a cluster, which can typically include hundreds of servers. Each server works with a subset of the data. This enables Hadoop to run on large clusters of mainstream two-socket servers, such as those powered by the Intel® Xeon® processor E5-2600[Δ] product family.

By itself, Hadoop acts as a distributed computing operating system that provides a distributed file system across all nodes in the Hadoop cluster, as well as the ability to use other file systems. It also includes a distributed computational feature, MapReduce, which coordinates each of the servers in the cluster to operate on a part of the overall processing task in parallel. Numerous commercial and open source applications, toolkits, and data layers are available to operate on top of this core, providing job and task tracking, integration with other data processing solutions, scripting, and querying capabilities in SQL, as well as database functionality and content management for images, documents, and other types of data.

Hadoop is a top-level open source project of the Apache Software Foundation. Numerous commercial distributions are also available.

To learn more about Apache Hadoop, go to http://hadoop.apache.org

[Δ] Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to: Learn About Intel® Processor Numbers

To address these data management issues, we designed our program to:

- Support an assortment of data platforms ranging from relational databases and data warehouse solutions to big data platforms such as Apache Hadoop*.
- Maintain a data catalog of where the single version of the truth data resides.
- Establish a customer engagement process that enables us to address the business process, value, architecture, data container, location of the data container, and size of investment necessary to support each particular BI use case (see Table 1).

Our customer engagement process helps us consistently right-size the solution. For instance, our IT customers often request the use of the latest technology such as big data platforms like Hadoop. Reviewing their business needs and their data, we frequently find that a different approach more efficiently and effectively meets their needs. By proving such expertise in data platforms, we enable Intel to make better use of its data resources.

USING AGILE PROJECT METHODOLOGY

To provide greater flexibility and responsiveness to the needs of our IT customers, the IT Business Intelligence Data Management team is currently transitioning to an agile project methodology. The team members currently using this methodology produce iterative releases on projects typically every two to four weeks.

In some cases, we also use SCRUM methodology to enable incremental agile development. SCRUM is an iterative and incremental agile framework for completing complex projects, such as data management systems, that uses set periods of time called "sprints" and feedback loops to move a project along instead of traditional command-and-control styles of management.

Table 1. Use cases help Intel IT to determine the right solution for data management

Use Case	Business Intelligence (BI) Container
Real-time Online Transaction Processing (OLTP) reporting/embedded BI	OLTP applications
Real-time analysis of streaming volume data sets	In-memory data warehouse uses RAM instead of disk-based storage
Analysis of enterprise-wide structured data	Enterprise data warehouse
Analysis of structured/ multi-structured data	Low-cost data warehouse
Analysis of raw, unstructured, sensor-type data	Big data Hadoop*
Ad hoc analysis and reporting of enterprise operations-structured data	Transactional system proprietary (operational) data warehouse
Ad hoc analysis of business-area structured data	Traditional relational database

Focusing on IT Customer Needs

In addition to right-sizing solutions, our customer engagement process (see Figure 1) enables us to better tailor our services to the needs of our IT customers, providing a high level of service quality and accountability. A key element of this strategy is maintaining intact teams that effectively utilize the business and domain expertise they gain over time.

Our customer engagement process starts with the assignment of a service manager to each business group, such as Sales and Marketing, Supply Chain, or Design. We also assign subject area product managers who focus on normalizing data for use. Each of these product managers specializes in a subject area, such as location, inventory, finance, sales distribution, and item.[1] Since many of these service and product managers are from the business groups we serve, we are able to glean important knowledge about the data and data management practices of those groups.

[1] "Item" is an Intel-specific term that covers the manufacturing stages of a silicon product—from silicon wafer to finished product.

ROLE OF THE SERVICE MANAGER

The service manager assesses the business problems and programs within the business groups we serve to determine what data is necessary to support their needs, its time criticality, whether the data needs to be integrated with other data from other sources, what platform is required, whether an existing data container can be reused or a new one will be needed, and whether the resulting data will need to be integrated with a LOB datamart.

ROLE OF SUBJECT AREA PRODUCT MANAGER

Subject area product managers are responsible for making data available and usable on specific subject areas, as well as managing authorization to access and use it—an important security aspect. This means each subject area product manager must understand the entire pipeline through which the data flows. This includes how it's created, where it's created, the business process in which it's used, and all the hops it goes through. Each subject area product manager must also determine the best way to structure the data to be used by the various business groups. This requires a close relationship with the business groups and owners of the systems where the data is created.

MEETING THE NEEDS OF THE CUSTOMER

Service managers and subject area product managers collaborate to ensure that through proper data modeling the enterprise data needs of each IT customer are met. This means providing data in a form that takes into account how the customer uses the data and appropriately structuring it to be readily usable. For example, to enable insights on how Intel's inventory matches current sales projections, a subject area product manager would work with a service manager to structure inventory data to enable easy integration and correlation with other information, such as sales and marketing data.

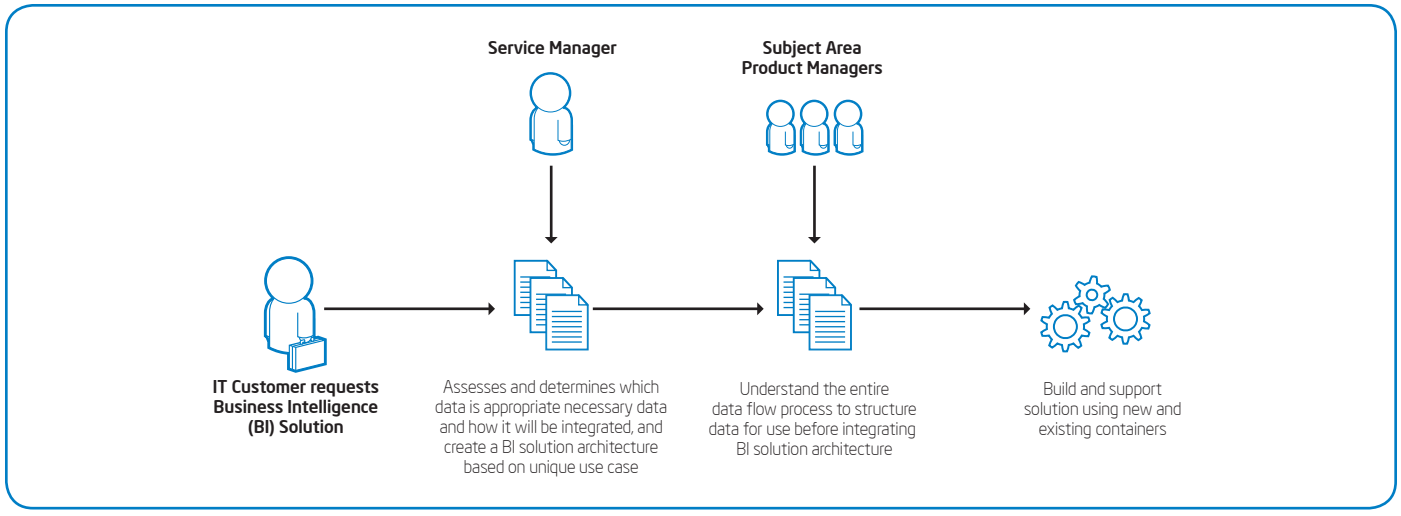

Figure 1. Intel IT uses a customer engagement process for matching customer needs with the right data management solution.

Self-Service Business Intelligence Solutions

One way to improve the velocity in which IT customers can obtain BI is to give them direct access to the information. Self-service solutions give the groups who use data the most within our IT customer base faster access to data tools and more autonomy to extract value and gain new, greater insights using visualization solutions and other techniques.

Properly designed, self-service BI solutions enable IT customers to work with enterprise information with minimal support from IT and without acquiring technical skillsets. They also allow IT customers to make their own modifications to respond more quickly to business changes.

Self-service solutions have advantages for the IT Business Intelligence Data Management team too. They enable us to focus more resources on addressing requests for solutions that involve more complex requirements, as well as investigating new data management services that could bring value to the organization.

USING SEMANTIC LAYERS

An important element of our self-service strategy is to take existing data subject areas and package them into semantic layers targeted at business groups. These semantic layers represent corporate data designed to allow our IT customers to access data autonomously using common business terms, such as "sales order," without having to know the underlying implementation details.

Business groups can use semantic layers to build reports, dashboards, or other BI solutions they require, including various forms of data visualization. Data visualization techniques are increasingly important, particularly with big data, for understanding the results of advanced analytics techniques. Semantic layers can also help with collecting data from multiple containers into a single location, eliminating the need for a solution developer to expend time and effort locating distributed data.

RESULTS

The creation of the IT Business Intelligence Data Management team and the centralization of data management has delivered greater agility, velocity, efficiency, accountability, and consistency for how Intel manages enterprise data and how business groups access and use it for BI. In this paper we provide just a few examples of recent results in these areas.

Agility

Our centralization and team approach to data management, as well as implementation of a Project Management Community of Practice specific to IT BI, has enabled easier adoption of agile methodology by the IT Business Intelligence Data Management team. The Project Management Community of Practice is a methodology for tying an organization's projects to business strategy to ensure projects support business objectives.

In the past year, we've overcome many of the challenges of moving to agile methodology—such as distributed project teams, limited agile experience, partial resource allocation, and resistance to change. Agile methodologies we have applied include: intact teams; small, frequent releases; and adding value at each step. In addition, we included new, aggregated metrics for data management results on our dashboard that let us measure our success in shifting to agile methodology.

Agility helps us to understand Intel's enterprise data and the ability to decide up front how we need to structure it for each usage based on our expertise. For example, when we get a new customer for a particular data set, the subject area product manager will often see how just a small adaptation, rather than a full-blown custom solution, will meet the customer's needs. We can then employ our agile methodology to create a solution in just a two-to-four-week sprint. This combination of data expertise and agility also enables us to do strategic builds ahead of time for key areas.

Velocity

One method of improving velocity is by streamlining the path data takes. Through improved data transformation techniques that eliminate a hop and move transformation one step closer to the source, we were able to let IT customers to see inventory changes in finished goods in near real-time. This enables them to more efficiently manage inventory. The solution also enabled us to make the transformed inventory data available and viewable to all customers at an enterprise level in near real-time.

Efficiency

One example of efficiency is a new channel-reporting solution that reduces data refresh by up to six hours in our revenue pipeline. The solution uses real-time data acquisition and other innovations to improve performance, including a file-less transfer process to move data. This technical release is the first of three planned project phases targeting the improvement of channel data availability and reduced time to access data to support key finance revenue decisions. By utilizing new technologies, we have built a reporting solution with fewer hops and reduced data latency to key downstream applications.

Accountability and Consistency

To improve accountability and consistency, as well as our ability to deliver a single version of the truth, we used a new tool from one of our vendors to develop data quality monitors for the revenue pipeline. Using this tool, we built approximately 50 data quality monitors and released them into production with just 20 hours of work by a single developer. These monitors enable us to clean up processes and missing records, discovering anomalies and issues much faster, and then fixing them immediately.

Based on our success with these first monitors, we built a monitor for our inventory subject area in 30 minutes and migrated through the path to production in just 7 hours, which is extremely fast. By enabling us to identify

duplicate records and confirming that our fix resolves these issues, this new monitor increases our confidence in the data quality. Data monitors are now being adopted for projects across the enterprise by our team. In the future, we plan to embed data quality monitors into all our data management areas, including sales order, inventory, customer, location, factory BI, and item.

CONCLUSION

By centralizing accountability for enterprise data and architecting solutions to make data more usable, Intel IT provides faster paths to higher quality data throughout the enterprise. This enables us to deliver improved BI results for our IT customers, helping them make better, timelier business decisions.

Over the next few years, Intel IT will continue improving our responsiveness to new data requests and providing faster turnarounds for our IT customers. At the same time, we will capitalize on the advantages of data centralization by increasing standardization and focusing on efficiencies that will enable us to drive down costs. In particular, we will continue to reduce the number of our current custom solutions and replace them with solutions designed for reusability.

We will also focus on new skill development. Open source big data solutions such as Hadoop, for example, necessitate new programming skills. At the same time we will look for new opportunities to take advantage of embedded BI solutions from our vendors, particularly ones that enable us to deliver self-service solutions that put our customers in charge.

We anticipate that the IT Business Intelligence Data Management team's efforts will enable Intel IT to find many new ways in the future to empower users with access to centralized enterprise data and the tools to extract better BI.

FOR MORE INFORMATION

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

- "Mining Big Data in the Enterprise for Better Business Intelligence"

ACRONYMS

BI	business intelligence
LOB	line-of-business
NoSQL	not only SQL
OLTP	Online Transaction Processing

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

Printed in USA ♻ Please Recycle 0113/ACHA/KC/PDF 328165-001US