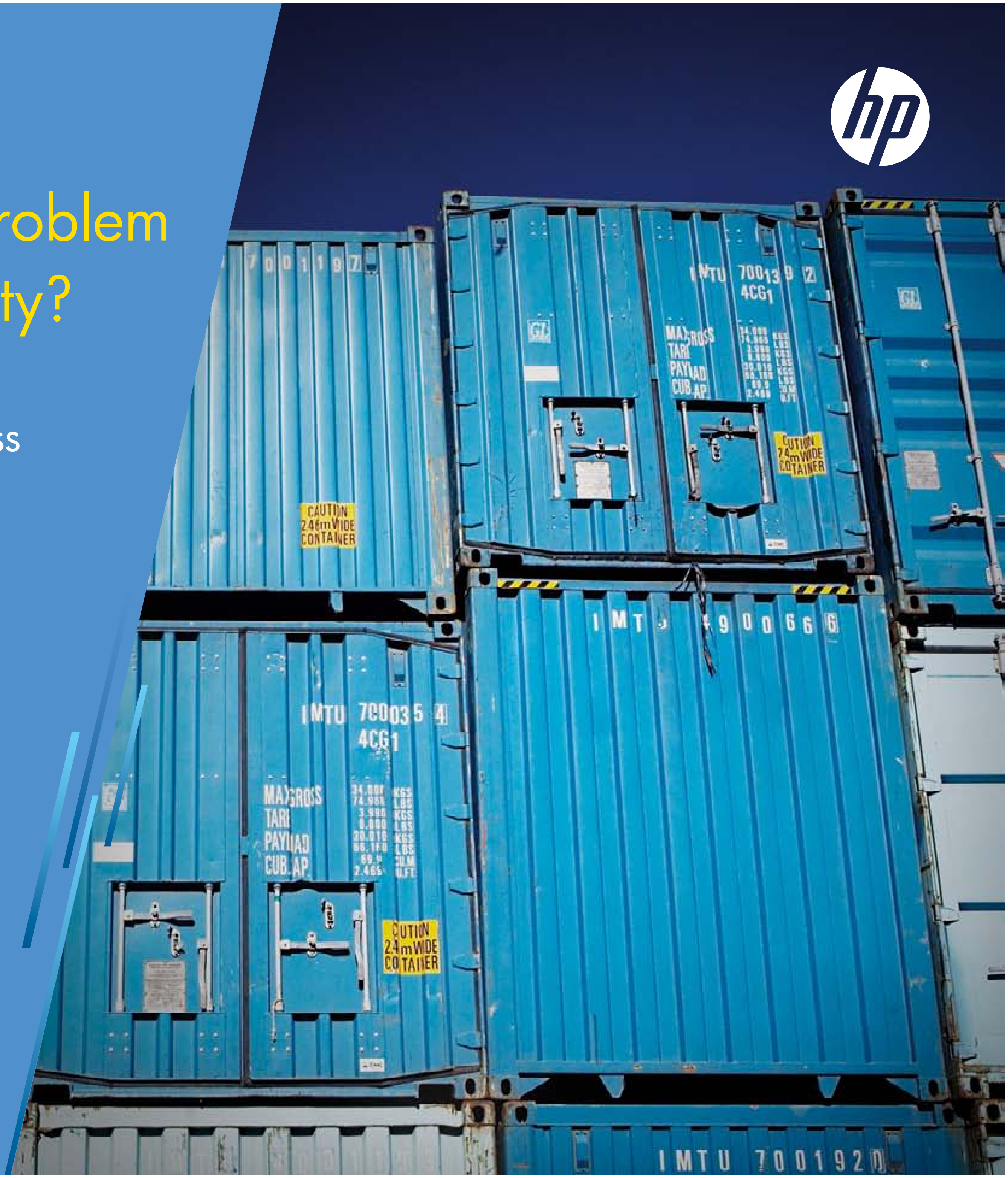
Big data—a big problem or a big opportunity?

Large-scale storage and cross analytics of big data

Business white paper

Enter →



Executive summary

Big data—a growing surge

- **4 trillion** SMSes, which is 4 PB per month worldwide
- **30 billion** pieces of content shared on Facebook every month
- **48 hours** of video uploaded onto YouTube every minute
- **40 percent** projected growth in global data generated per year vs. five percent growth in global IT spending

”Big data” is a term that describes enormous amounts of structured and unstructured data, which can possibly run into 10s–100s of terabytes (TB), petabytes (PB), or exabytes (EB). While structured data refers to records that can be stored in a database, unstructured data refers to all other data types. In addition to emails, documents, and transactions, businesses now need to archive and analyze multimedia material—including image, video, and audio formats—to track and manage product brands in the social media space. Moreover, keeping pace with the rapidly growing storage ecosystem are changing compliance demands, rising storage costs, new risk management challenges, and advanced business intelligence (BI) requirements.

More data will be created in the next 12–18 months than has been created during the entire human history. As data is gathered from more sources, BI solutions must be able to store and extract meaningful information from it. The solutions must be able to store, manage, access, analyze, and cross-analyze heterogeneous data, quickly and cost-effectively. Currently, organizations use a different data warehouse product for each data type, such as transactional, text, and email—creating multiple data silos. This approach has made it difficult to see the big picture and spot defects, trends, and opportunities.

Given the fast-changing dynamics of today’s market environment, you need to unify storage and analysis of all types of big data. The EDMT Solution—an integrated solution for data sourced from emails, documents, multimedia, and transactions—is the pragmatic answer to this need. This pioneering solution is brought to you by the combined efforts of BMMsoft, HP, Intel®, and Sybase. The solution’s powerful analysis and cross-analysis capabilities deliver sophisticated tools for detecting fraud, strengthening security, and predicting business trends.

Big data—its diversity and implications

Numbers that sum up the value of big data to the enterprise

- **$300** billion USD—potential annual value to U.S. healthcare
- **€250** billion EUR—potential annual value to Europe's public sector administration
- **$600** billion USD—potential annual consumer surplus from using personal location data globally

In today's economic climate, you need to predict, rather than react to critical events that can affect your businesses. For this, it is vital to integrate and analyze data from different sources. But traditionally, the different types of big data have been managed by multiple, fragmented systems that are disconnected from real-time business operations. This makes storing, indexing, and analyzing of the various data groups across isolated systems slow, complex, and inefficient.

BI processes built on such legacy frameworks can't support the dynamic and expanded business ecosystems that are developing even as new opportunities emerge from market upheavals. Cost is also a concern. With the explosion of electronically stored information, every organization must look for new ways to cut storage costs. Adding to the challenges are ever-demanding government regulation, compliance mandates, and frequent litigation.

What can help is a comprehensive, centralized database, capable of cost-effectively storing a variety of data—doing so intelligently by building relationships between the different types of data as the data is being loaded. This enables effective analysis of timely and relevant data to provide insights for supporting decisions such as launching new products in the market, detecting credit card fraud, or offering variable pricing options for energy usage. And this exchange of information throughout the business ecosystem can reveal new ways to use information to draw on new business prospects.

CASE STUDY

Big data—its diversity and implications

CASE STUDY—MEDIA INDUSTRY AND ENTERPRISE-READY BIG DATA

Customer: ID_Media (IDM) enables political organizations to use the Internet as a channel to communicate to voters through blogs, discussion groups, and email communications. The company often runs campaigns to inform voters of potential legislation and garner feedback.

Challenge: Due to IDM's limited capability to filter responses and categorize emails, their email campaigns often did not ask for responses.

Solution: With the EDMT Solution's advanced analytic capabilities, the IDM team was able to:

- Provide not only the context of email messages, but also cross-correlate the responses from registered voters with the specific demographics
- Access the specific language of respondents to tailor future messaging according to the concerns of those both for and against a certain legislation or candidate
- Determine which recipients want to be removed from the mailing list by categorizing recipients for removal from the mailing list via contextual search of responses rather than utilizing an unsubscribe button

Result: IDM was able to more quickly respond to their customers about the voters' opinions to proposed legislation and provide detailed information about the support or opposition—according to both location and other demographics characteristics of the voter. Going forward, IDM can continue to accumulate the "voice of the voter" for future campaigns.

N

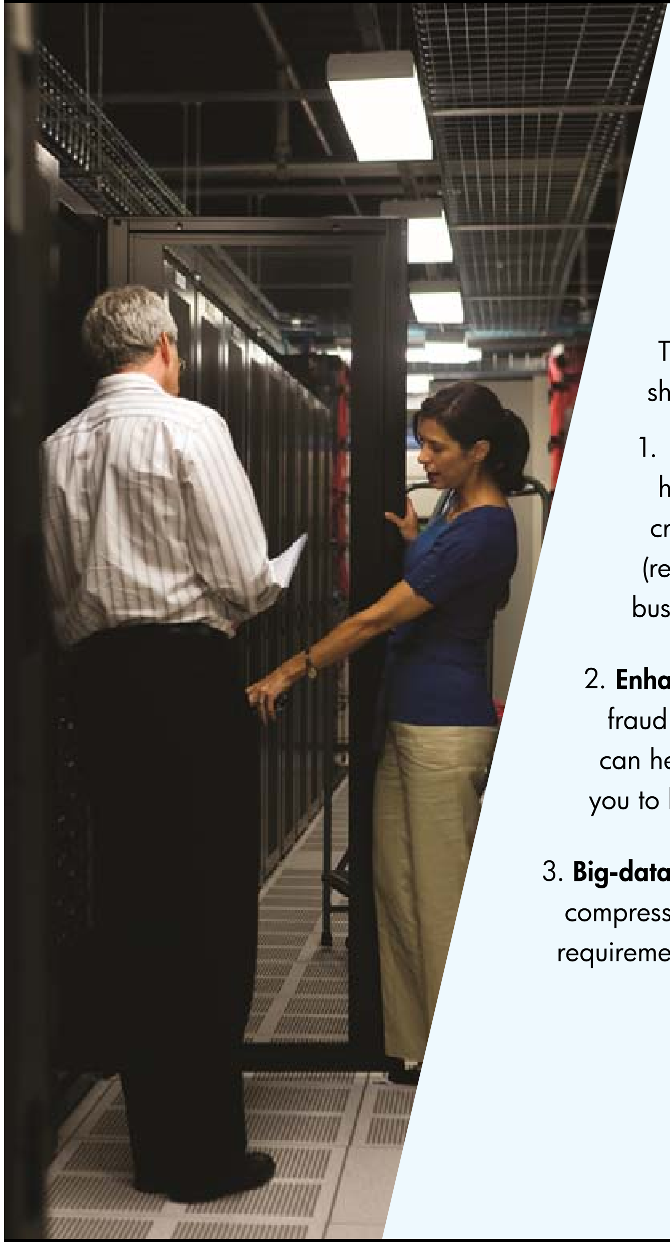
- S
- (
- S data globally

In today's our businesses.
For this, it i ypes of big
data have b ess operations.
This makes s complex,
and inefficier

BI processes b tems that are
developing eve explosion of
electronically sto to the challenges
are ever-demand

What can help is ta—doing so
intelligently by buil nables effective
analysis of timely a cts in the
market, detecting cre of information
throughout the busine .

Top 3 capabilities to look for in a big data solution



What you can get from big data depends on your specific industry and market. Government and intelligence agencies need instant analysis of real-time and historical big data from all available sources. Telecom and hosting service providers need to increase profit margins through data analysis. Financial services organizations need to meet strict regulatory requirements, fight fraud, and respond to litigation and audit demands. And healthcare organizations need to meet the regulatory requirements of the health insurance portability and accountability act of 1996 (HIPAA) and be prepared to defend against fraud.

To help each industry sector proactively meet its challenges and capitalize on opportunities, a big data solution should have inherent capabilities to enable these three crucial outcomes:

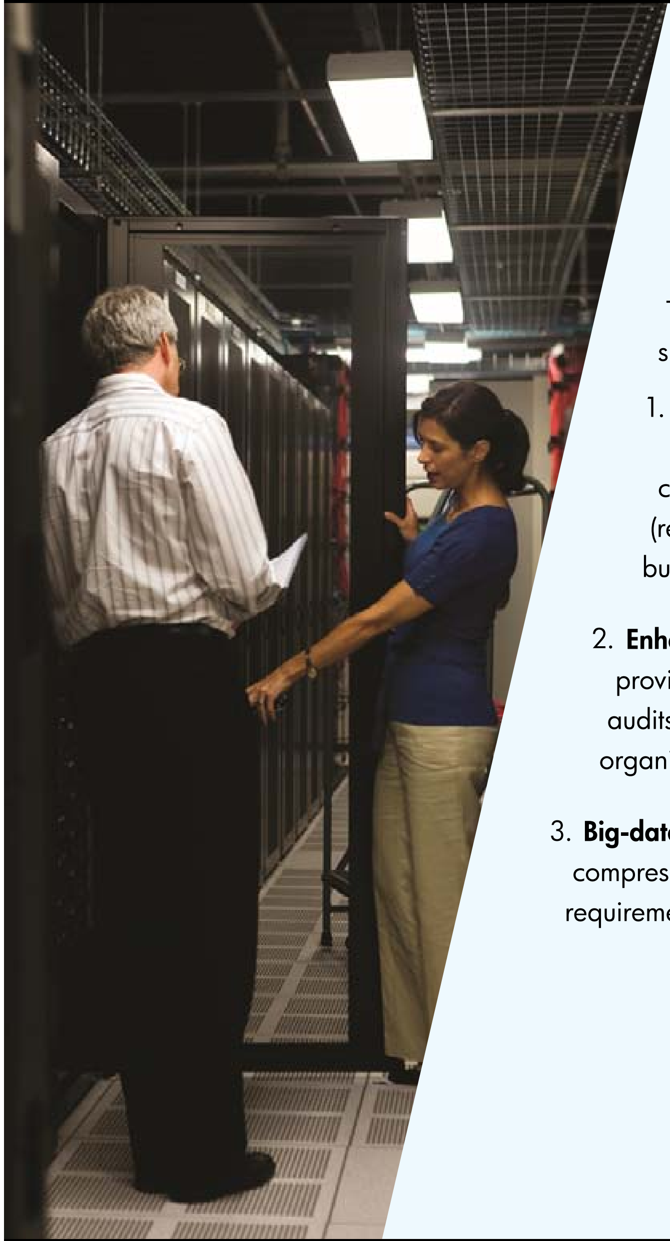
1. **Improved business intelligence:** To begin with, the solution should enable real-time, low-latency capture and high-speed ingesting and indexing of data from all sources. Next, the solution should enable you to store, cross-access, cross-analyze, and cross-reference all structured and unstructured data, as well as new (real-time) and historical data. Finally, the solution should be able to connect the intelligence scattered across the business ecosystem, reconcile data silos, and allow the knowledge gained to flow across the business.

2. **Enhanced security:** The solution should enable better audit; governance, risk management, and compliance (GRC); fraud and threat detection; and e-discovery by providing complex risk analysis capabilities and real-time alerts. This can help you be ready for internal or external audits at all times. The database should also be tamper proof, enabling you to better manage risks, defend your organization against fraud, and meet compliance requirements.

3. **Big-data-friendly storage:** The solution should help you decrease the rising storage costs for big data by efficiently compressing enormous volumes of data on cost-effective disk storage. The solution should also help you meet GRC requirements with fewer resources, as well as support active disaster recovery and archiving of data cost efficiently.

CASE STUDY

Top 3 capabilities to look for in a big data solution



Wh
age
hosti
strict r
need t

To help
should h

1. **Improv**
high-sp
cross-ac
(real-time
business e

2. **Enhanced s**
providing c
audits at all t
organization

3. **Big-data-frienc**
compressing en
requirements wit

CASE STUDY—AEROSPACE INDUSTRY AND ENTERPRISE-READY BIG DATA

Customer: Serbia and Montenegro Air Traffic Services Agency (SMATSA) is an air traffic services agency responsible for improving the safety of air navigation and supporting increase in air traffic.

Challenge: SMATSA was tasked with real-time storing, classifying, retrieving, and analyzing of their fast growing data volumes from disparate sources, such as emails, flight information, passenger information, data exchange services, telemetry and database records.

Solution: With the EDMT Solution's advanced storage and analytic capabilities, SMATSA was able to:

- Enable early detection of potential security problems
- Reduce the search time for new and historical data to less than two seconds
- Capture emails, CDR, telemetry and other data in real time (with latency less than two seconds)

Result: SMATSA is able to improve transportation security, meet compliance requirements with ease, and save 91 percent of costs. In addition, the agency now promotes collaboration with government, FAA, and law enforcement agencies; financial institutions; and public health authorities to increase security of air transportation.

and intelligence
Telecom and
ons need to meet
.S organizations

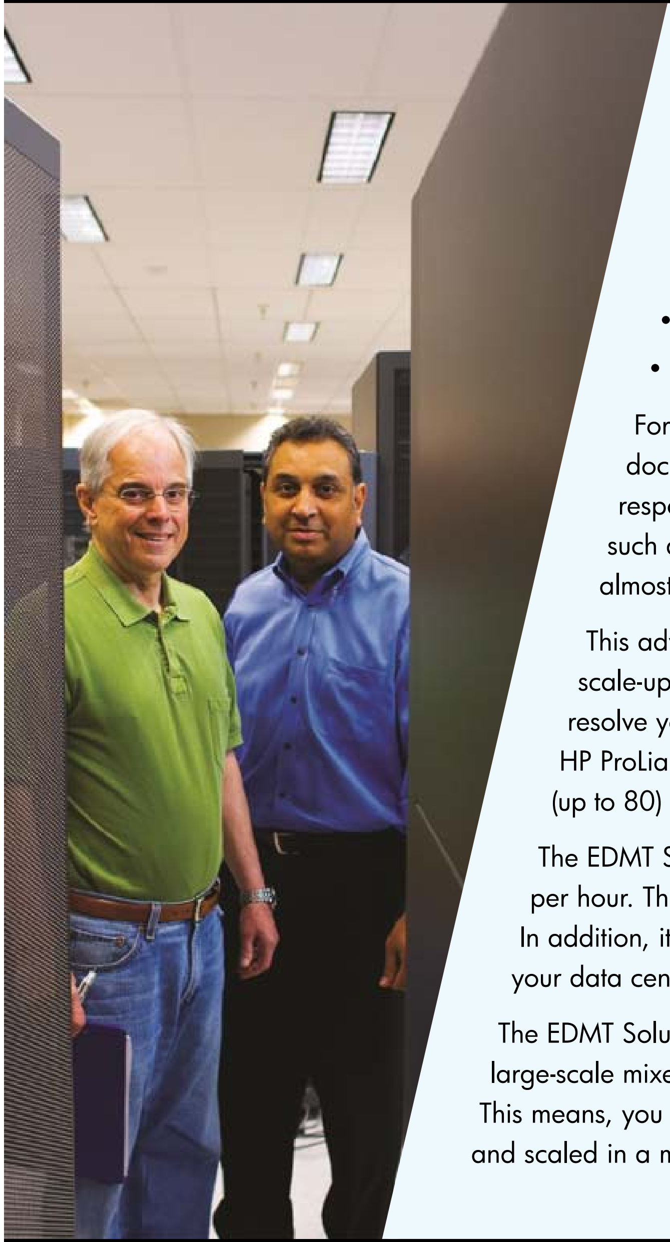
g data solution

capture and
to store,
as new
cross the
ess.

discovery by
al or external
efend your

efficiently
neet GRC
efficiently.

The EDMT Solution—when industry leaders come together



In response to the pent-up demand for a fast, practical, and out-of-the-box big data solution, trusted technology players BMMsoft, HP, Intel, and Sybase have come together to present the EDMT Solution. This is a visionary solution for large-scale storage and cross analysis of structured and unstructured data; others handle SQL or text, but not both. The solution can be deployed on a small, medium, or large scale, and it brings with it enormous potential to:

- Detect market opportunities to reduce competitive threats
- Discover the causes of problems in the past, present, and future
- Help fulfill regulatory obligations

For starters, the EDMT Solution can replace standalone products addressing data warehousing, email archiving, and document management—with the real-time analytic archiving functionality necessary for e-discovery and litigation response. Going beyond, it offers data mining of mixed data for revenue-generating and revenue-protecting activities, such as fraud detection, security threat assessment, social media analysis, CRM, and product analysis. It can do this at almost no extra cost and in real time.

This advanced solution leverages the power of HP Converged Infrastructure—where industry-leading software, powerful scale-up Intel or Linux servers, and highly reliable storage work together seamlessly to enhance your IT infrastructure and resolve your big data problems. The high-speed storage and analytics engine from BMMsoft is certified on the powerful HP ProLiant DL980 server running Linux, which comes with a large RAM size (up to 4 TB) and a large number of cores (up to 80) to be able to handle the unpredictable workload of big data capturing, parsing, ingesting, indexing, and searching.

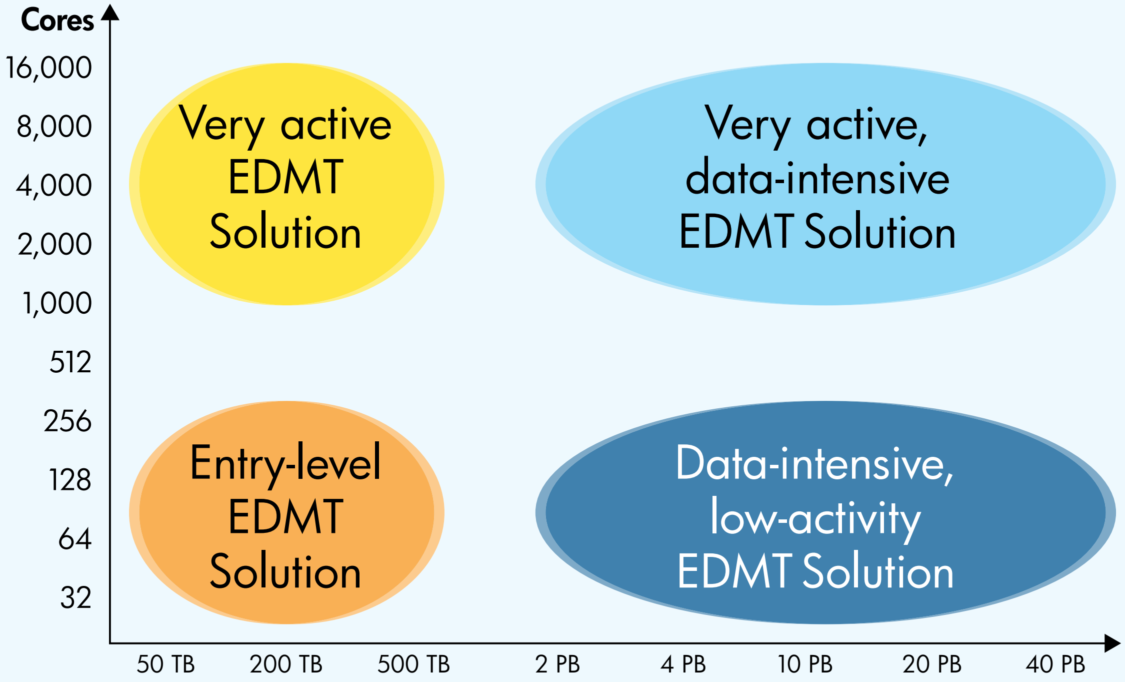
The EDMT Solution on the ProLiant DL980 server is the world's fastest real-time big data loader with a capacity of over 14 TB per hour. The server offers 200 percent boost in server availability and the resilient system fabric needed to increase uptime. In addition, it can help reduce operating costs by $48,300 USD per 100 users over three years, as well as extend the life of your data center with HP Thermal Logic, which enables energy savings.

The EDMT Solution also combines HP MSA 2000 G3 Storage and Sybase IQ. The result is a high-speed database for large-scale mixed SQL and text analytics, with the efficiency to store up to 10,000 emails and one million SMSes for $1 USD. This means, you get to store source files at the cost of tape storage. Moreover, storage and server capacity can be configured and scaled in a modular fashion to meet your organization's unique needs suitably, now and in the future.

The EDMT Solution—when industry leaders come together

The EDMT Solution—add servers and storage as needed

Contrast this with multiple, disparate point products (often as a result of multiple acquisitions) that require custom integration—that too at great expense and time (months), and the never-ending burden of expensive maintenance of nonstandard products.


Cores
16,000
8,000
4,000
2,000
1,000
512
256
128
64
32
Very active EDMT Solution
Very active, data-intensive EDMT Solution
Entry-level EDMT Solution
Data-intensive, low-activity EDMT Solution
50 TB
200 TB
500 TB
2 PB
4 PB
10 PB
20 PB
40 PB

CASE STUDY

The EDMT Solution—when industry leaders come together

The

Co that require
custo of expensive
maint

CASE STUDY—TELECOM INDUSTRY AND ENTERPRISE-READY BIG DATA

Customer: T365 is a telecom company with 80 percent of the world's inter-carrier data transfer billing and cross-charging business. It provides billing arbitration or cross-charging between data carriers for SMS, MMS, email, and other data transfers that use services of multiple carriers.

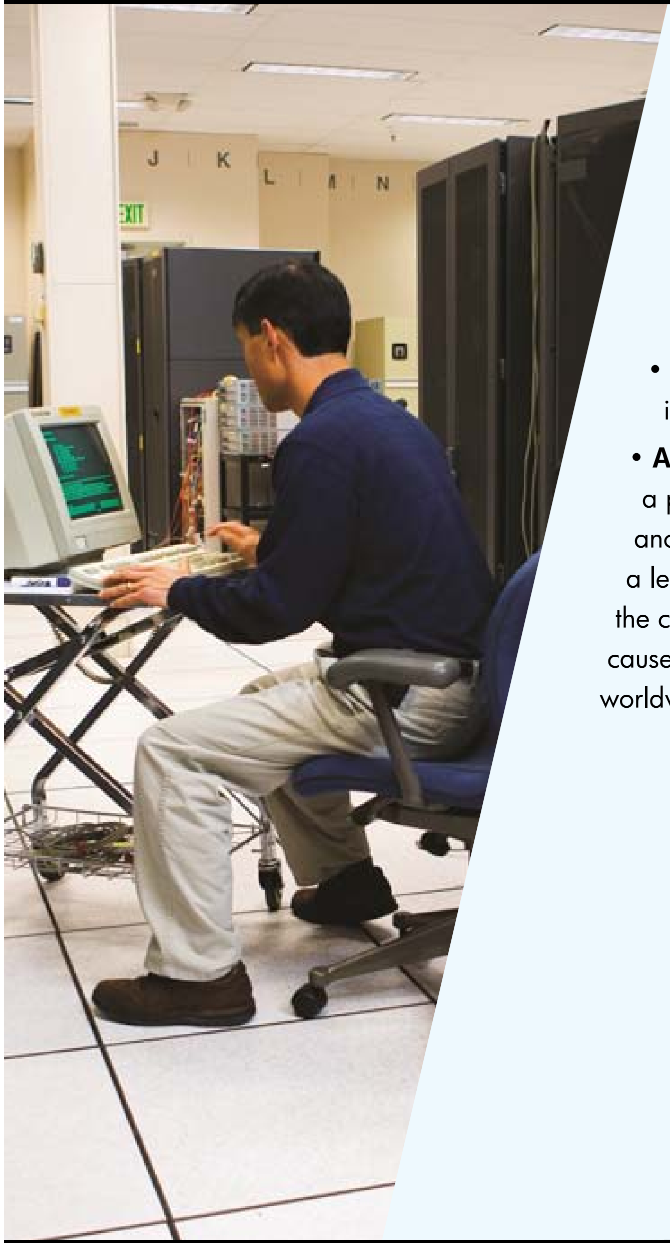
Challenge: 80 percent of the world's telecom traffic presents enormous data volumes to be processed, loaded, indexed, and calculated. This made it difficult for T365 to add additional data types, expand business scope beyond just billing, and support future growth.

Solution: The EDMT Solution provided the ability to load SMS/MMS data at very high speeds, as well as instantly search and cross-correlate billing information. This allowed T365 to:

- Load 520 million SMS and MMS data per hour (500 GB/h)—which is more than the combined SMS/MMS traffic in the US
- Perform monthly billing for US and Europe within minutes
- Reduce the search time for new and historical data to less than two seconds
- Add new types of messages without changing the application or data model
- Handle the payload for new applications such as mobile banking, compliance, fraud detection, and litigation
- Connect with enterprise system such as banks and telecoms seamlessly

Result: T365 is able to keep up with SMS, MMS, and packet data growth; meet billing requirements; and be ready for future needs. In addition, the company can now collaborate with financial and government authorities to increase security.

Applying the EDMT Solution to solve real business problems



Following are some of the queries that were executed during the "Petabyte Data Warehouse" performance sizing demonstration—highlighting the strengths of the EDMT Solution, as well as their relevance in solving real-world problems.[1]

Mining unstructured data

- **Query performed:** The "popular stocks" query identifies which securities are most discussed among traders. All correspondences, such as email, blog entries, or instant messages are searched to determine which securities were the most frequently mentioned on a particular day.
- **Strength demonstrated:** Ability to load large amounts of new unstructured data as well as query simultaneously in real time.
- **Application in real-world solutions:** A similar query could identify "insider trading" by finding securities for which a prevailing recommendation to "buy," "hold," or "sell" was discussed prior to the company's financial report. In another business environment, this query could be mining millions of emails and blog entries looking for the source of a leak or for the origin of a rumor affecting the company's reputation in the marketplace. In the context of litigation, the complete record of all electronic communications from the parties involved could be searched to help bolster the cause of action or uncover the evidence to support a claim. And in a "homeland security" context, online postings worldwide could be loaded and searched in real time to identify terrorist plans or money-laundering activities.

[1] Source: "Petabyte data warehouse—performance sizing report," InfoSizing, February 29, 2008.

Applying the EDMT Solution to solve real business problems

Features that make the EDMT Solution unique and easily scalable

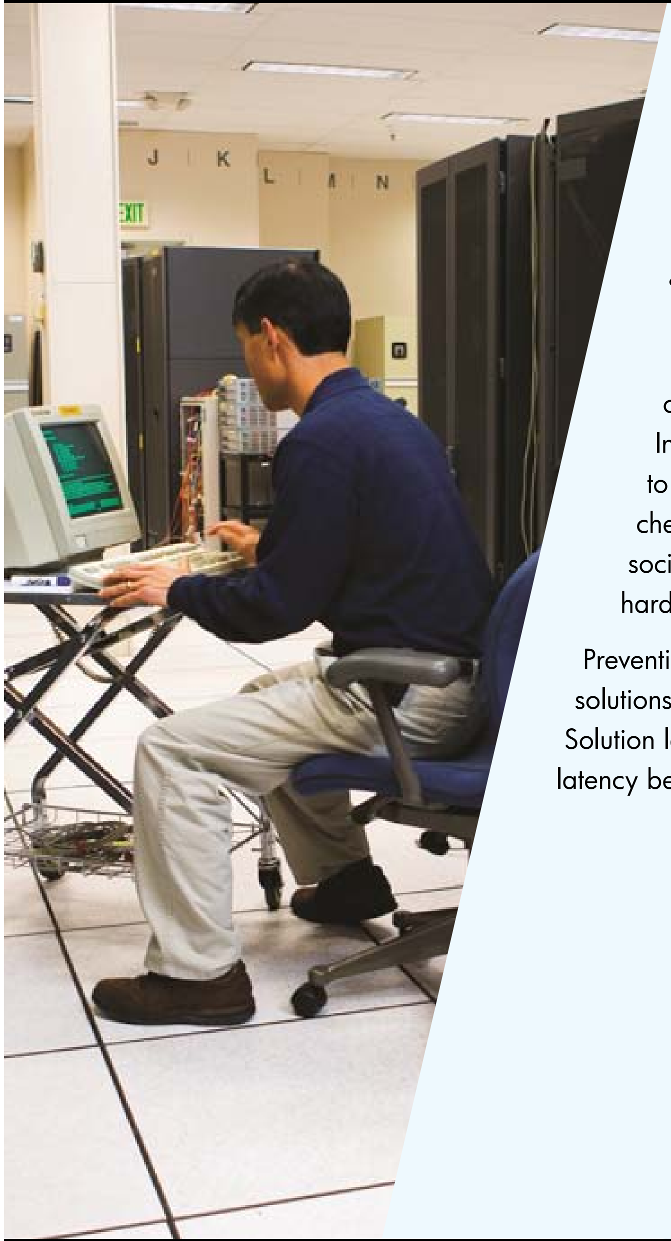
- Provides real-time alerting and notification of events referenced in any type of data
- Facilitates real-time data capture and response with high ingest speed and capacity—10 billion transactions and over 500 million messages, files, and emails per hour[2]
- Stores different types of data in a single SQL database, making unified search and analysis possible
- Presents a complete enterprise-ready solution, offering full ACID and full SQL to enable enterprise applications to search data precisely
- Reduces cost by 90 percent through extreme data compression—1 PB of data compressed to 250 TB on disk[3]
- Enables use of lesser resources—apart from reducing storage size and electricity costs, it cuts 5000 tons of CO_2 emissions per year[4]

Mining transactional data

- **Query performed:** The "portfolio growth" query determines the hypothetical growth of a portfolio over the past year. The query targets ten securities and assumes the following trading strategy: When the 20-day moving average of a security crosses over the five-month moving average, a tenth of the portfolio is invested; and when the 20-day moving average crosses below the five-month moving average, the portfolio is sold.
- **Strength demonstrated:** Ability to process and answer multiple streams of queries concurrently (up to 50 streams in the test) within a few seconds, despite the query's high degree of complexity and the high volumes of data being searched.
- **Application in real-world solutions:** A similar query could be useful when billions of dollars are being spent on complex government programs with inadequate tracking and auditing. The query could help detect fraudulent investment schemes by verifying that paid returns are the result of gains derived from actual purchases and sales within existing portfolios. In another example, the transactional data for orders between car manufacturers and their part suppliers could be loaded and analyzed to identify healthy areas of the industry that can be best rescued by an influx of capital, or areas where failures are chronic, decay is unavoidable, and new funding would be wasted.

[2, 3, 4] Source: "Petabyte data warehouse—performance sizing report," InfoSizing, February 29, 2008.

Applying the EDMT Solution to solve real business problems



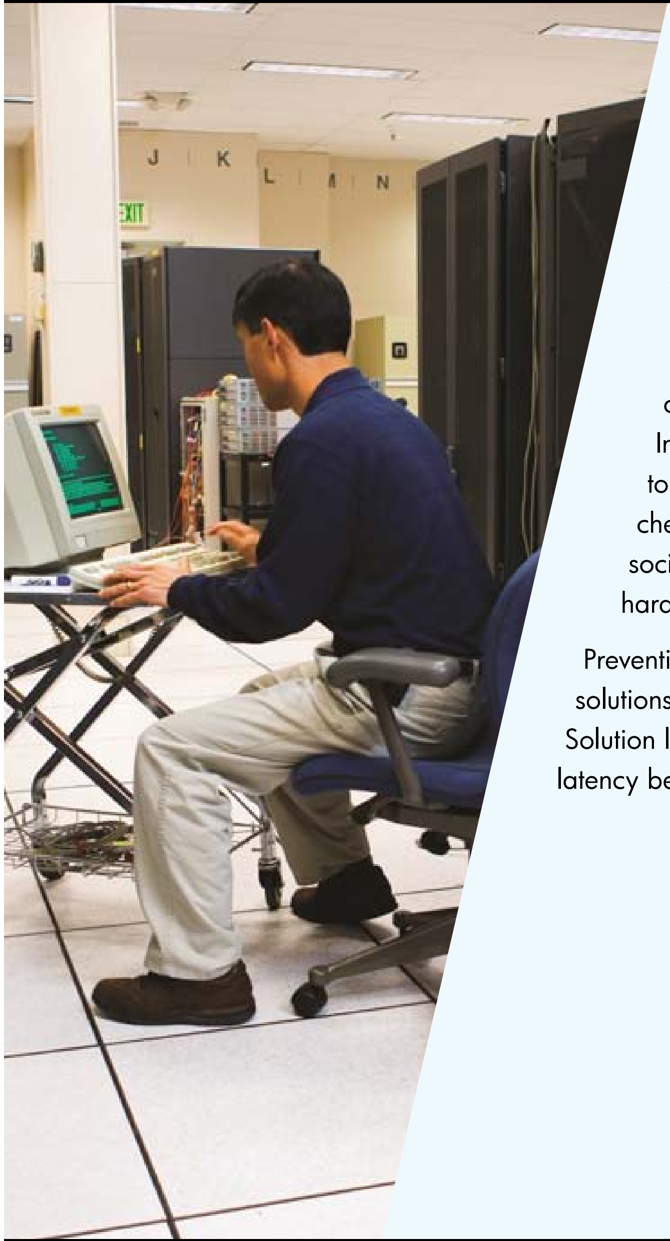
Mining mixed data

- **Query performed:** The "secret" query, by mining all transactional and unstructured data, determines if there is any correlation between quotes (bids) tendered and emails urging secrecy about the same stock.
- **Strength demonstrated:** Ability to combine structurally different data and correlate billions of emails, attachments, blog entries, and instant messages with trillions of conventional business records.
- **Application in real-world solutions:** A similar query could draw a correlation between the level of increase in the sale of a new product and the number of online postings containing a positive mention of that product. Conversely, the ability to include new online postings and new sales data in the query in real time could give a consumer product vendor an early warning about a manufacturing defect with the potential for a product recall. In another example, cross analysis of data could make real-time use of blogs, emails, and other online "chatter" to detect early rumors and threats, and correlate them with transactional data such as stock trades, purchases of chemicals, or shipments of weapons and ammunition. By effectively integrating all modern communications and social networking into the BI framework, cross correlation of mixed data could provide fact-based early warning of hard-to-catch events representing potential threats of terrorist attacks, criminals operations, or financial meltdowns.

Preventing problems instead of repairing the resulting damages is common sense. This requires a new breed of BI solutions that can detect events, problems, or opportunities and react, prevent, or exploit them—in real time. The EDMT Solution leads the real-time BI space with results that were demonstrated by measuring between 0.5 and 3 seconds of latency between "event" and "reaction."

CASE STUDY

Applying the EDMT Solution to solve real business problems



Mir

• Q
the

• Stre
attac

• Appli
in the
Conver
consume
In anothe
to detect
chemicals,
social netw
hard-to-catc

Preventing prob
solutions that co
Solution leads th
latency between

CASE STUDY—EDUCATION AND HEALTHCARE INDUSTRY USING BIG DATA

Customer: Belgrade University School of Medicine (BUSM) is the largest institution for training physicians in Southern Europe.

Need: To improve organization and teaching process, BUSM needed to rapidly generate and update documents for accreditation.

Challenge: BUSM was overwhelmed with constant growth of documents, email communication, and database transactions.

Solution: The EDMT Solution offered real-time storing, categorizing, retrieving, and analysis capabilities within a unified repository, which allowed BUSM to:

- Facilitate real-time, recorded collaboration between professors, students, departments, and patients
- Offer full insight into email communication and related documents for any particular teaching subject or department
- Promote fast, complex analysis of data, using combined numeric and full-text-search analysis
- Generate and update documents rapidly for accreditation

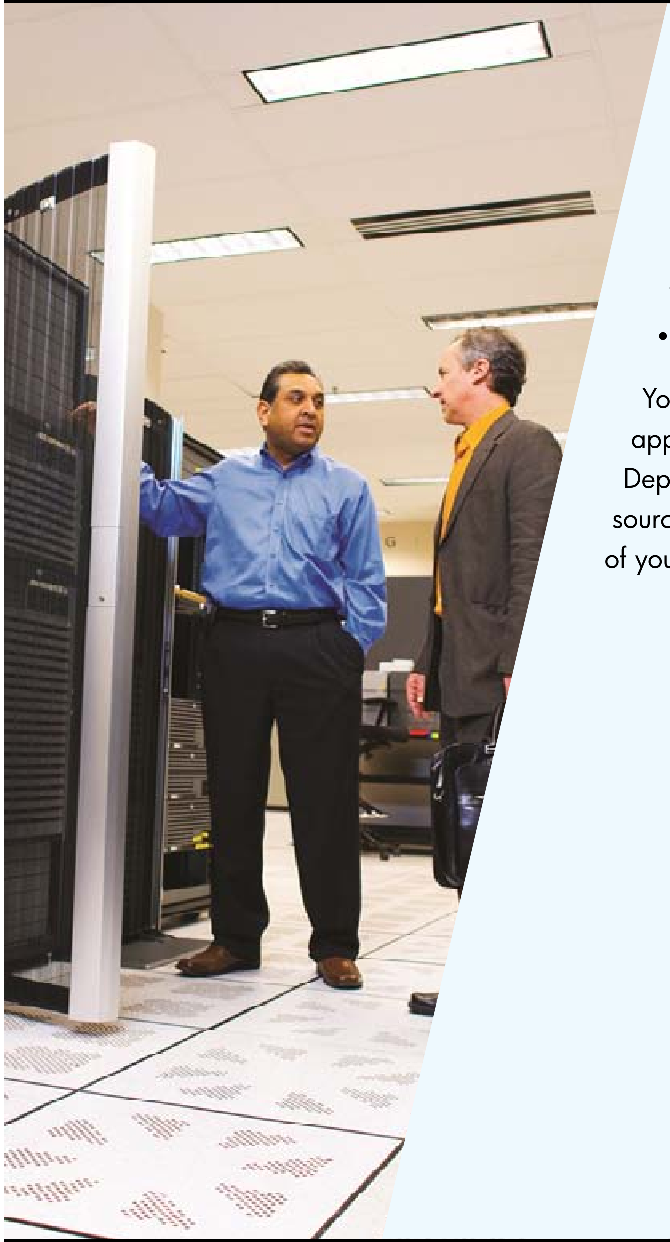
Result: BUSM was able to reduce search time by 90 percent by categorizing documents, emails, and database transactions stored in a unified repository. The institution was also able to cut purchase and maintenance costs by 76 percent.

etermines if
same stock.
emails,

of increase
that product.
could give a
product recall.
nline “chatter”
purchases of
cations and
arly warning of
l meltdowns.

eed of BI
ne. The EDMT
3 seconds of

Open the door to a "big opportunity"



Discover the benefits of adopting a dynamic and cost-effective approach to resolving your big data challenges. And learn more about deploying a massively scalable storage platform, comprising components from four industry leaders you trust. Visit:

- The EDMT Solution page at www.bmmsoft.com
- The HP and Sybase partner page at www.sybase.com/partner/hp; www.hp.com/solutions/sybase
- The HP ProLiant DL980 Server page at www.hp.com/go/dl980
- The Intel® Xeon® page at www.intel.com/itcenter

You can start with a half-day, one-day, or one-week technical training that can help you explore recommended approaches to your specific e-discovery, investigation, or compliance projects. Subsequently, you can use EDMT Deployment Services to deploy a standard solution on premise within a week, and connect it to your standard data sources. Alternatively, our consultants can customize your EDMT Solution and deploy it on premise, host it at a location of your choice, put it on the cloud, or offer it as software as a service (SaaS).

Get connected

www.hp.com/go/getconnected

Get the insider view on tech trends, alerts, and HP solutions for better business outcomes

Has this document helped you? Tell us how. Click here

4AA3-9959ENW, Created April 2012