Microsoft

The Microsoft Modern Data Warehouse

Contents

Executive summary

Data warehousing technology began as a framework to better manage, understand, and capitalize on data generated by the business. The traditional data warehouse pulled all data into a central, schema-driven "repository of truth" for analytics and reporting, and it worked extremely well for many years. However, the world of data is rapidly evolving in ways that are transforming the industry and motivating enterprises to consider new approaches to business intelligence (BI).

The traditional data warehouse is under pressure from the growing weight of explosive volumes of data, the expansive variety of data types, and the real-time processing velocity of how data is being used to grow and operate the business. These changes are so seismic that Gartner reports, "Data warehousing has reached the most significant tipping point since its inception. The biggest, possibly most elaborate data management system in IT is changing."[1]

The modern enterprise needs a logical architecture that can smoothly scale to meet these volume demands with real-time processing power and the ability to manage any data type to rapidly connect the business to valuable insights. This means that the traditional data warehouse needs to evolve into a modern data warehouse.

The traditional data warehouse

The traditional data warehouse was designed specifically to be a central repository for all data in a company. Disparate data from transactional systems, ERP, CRM, and LOB applications could be cleansed—that is, extracted, transformed, and loaded (ETL)—into the warehouse within an overall relational schema. The predictable data structure and quality optimized processing and reporting. However, preparing queries was largely IT-supported and based on scheduled batch processing.

Web 2.0 significantly grew business-related data generated from e-commerce, web logs, search marketing, and other sources. These sources remained business-generated and business-owned. Enterprises expanded ETL operations to compensate for the new data sources, ultimately also expanding the schema model.

Yet even with these growing complexities, the core business value of the traditional data warehouse was the ability to perform historical analysis and reporting from a trusted and complete source of data (Figure 1).

[1] Gartner, *The State of Data Warehousing in 2012*, http://www.gartner.com/id=1922714, February 2012.

Figure 1: Framework for the traditional data warehouse

Key trends breaking the traditional data warehouse

Together, four key trends in the business environment are putting the traditional data warehouse under pressure. Largely because of these trends, IT professionals are evaluating ways to evolve their traditional data warehouses to meet the changing needs of the business. The trends—increasing data volumes, real-time data, new sources and types of data, and cloud-born data—are discussed below (Figure 2). Also discussed is logical information architecture as a new approach to data warehousing in response to these trends.

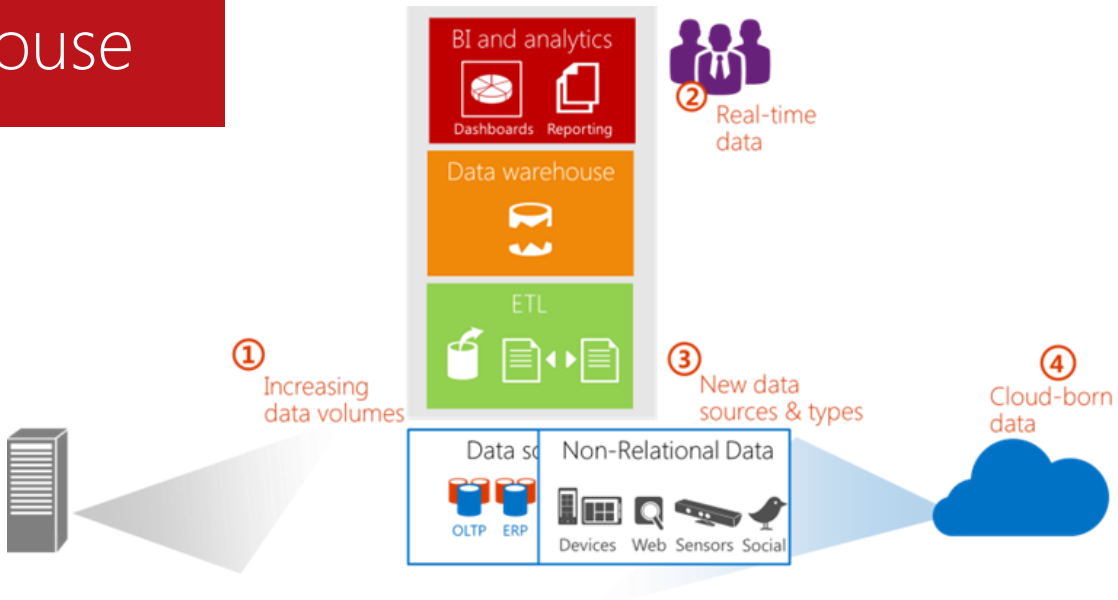

Figure 2: Four key trends breaking the traditional data warehouse

Increasing data volumes

The traditional data warehouse was built on symmetric multi-processing (SMP) technology. With SMP, adding more capacity involved procuring larger, more powerful hardware and then forklifting the prior data warehouse into it. This was necessary because as the warehouse approached capacity, its architecture experienced performance issues at a scale where there was no way to add incremental processor power or enable synchronization of the cache between processors.

However, data **volume** is expanding tenfold every five years. Much of this new data is driven by devices from the more than 1.2 billion people who are connected to the Internet worldwide, with an average of 4.3 connected devices per person. Devices including smartphone also provide support for remote monitoring sensors, RFID, location-based data, transactions and more.

For the modern business, the prospect of bigger, more powerful hardware and ever-larger forklift migrations is not a viable return-on-investment scenario. Enterprises are looking for an alternative to volume growth that does not break the budget.

Case Study: Hy-Vee Supermarkets

Hy-Vee operates a growing chain of employee-owned supermarkets in eight states in the midwestern United States. To boost its competitiveness, the company sought to increase its data warehouse performance so it could deliver store-level purchasing data more quickly to its business analysts and managers.[2]

"However, the process was not at all consistent. It simply took too long to load the files, and query times were too slow. We need to get that data to our employees for analysis first thing in the morning. If they don't have it on time, they don't have the most updated data for analyzing promotions." – Tom Settle, Assistant Vice President, Data Warehousing

Real-time data

The traditional data warehouse was designed to store and analyze historical information on the assumption that data would be captured now and analyzed later. System architectures focused on scaling relational data up with larger hardware and processing to an operations schedule based on sanitized data.

Yet the velocity of how data is captured, processed, and used is increasing. Companies are using real-time data to change, build, or optimize their businesses as well as to sell, transact, and engage in dynamic, event-driven processes like market trading. The traditional data warehouse simply was not architected to support near real-time

[2] Microsoft Case Studies, *Hy-Vee Boosts Performance, Speeds Data Delivery, and Increases Competitiveness*, http://www.microsoft.com/casestudies/Microsoft-SQL-Server-2008-R2-Enterprise/Hy-Vee/Hy-Vee-Boosts-Performance-Speeds-Data-Delivery-and-Increases-Competitiveness/710000000776, May 2012.

transactions or event processing, resulting in decreased performance and slower time-to-value.

Case Study: Direct Edge Stock Exchange

Among stock exchanges, low latency—the speed at which a stock trade can be processed—is supreme. Direct Edge wanted to reduce the already low latency of its system, while supporting vastly larger trading volumes. With a 40-terabyte warehouse growing 2 terabytes per month with targets for hundreds of terabytes generated from over 100 million trades per day, Direct Edge had to offer its customers the fastest, most reliable service it could.[3]

"That's because the amount of profit that your customers make depends on the speed at which a transaction is cleared. Latency also determines how many transactions an exchange can handle in a day. The higher the transaction volume, the greater the profits for the exchange." – Steve Bonanno, Chief Technology Officer

New sources and types of data

The traditional data warehouse was built on a strategy of well-structured, sanitized and trusted repository. Yet, today more than 85 percent of data volume comes from a variety of new data types proliferating from mobile and social channels, scanners, sensors, RFID tags, devices, feeds, and other sources outside the business. These data types do not easily fit the business schema model and may not be cost effective to ETL into the relational data warehouse.

Yet, these new types of data have the potential to enhance business operations. For example, a shipping company might use fuel and weight sensors with GPS, traffic, and weather feeds to optimize shipping routes or fleet usage.

Companies are responding to growing non-relational data by implementing separate Apache Hadoop data environments, which requires companies to adopt a new ecosystem with new languages, steep learning curves and a separate infrastructure.

Cloud-born data

An increasing share of the new data is "cloud-born," such as clickstreams; videos, social feeds, GPS, and market, weather, and traffic information. In addition, the prominent trend of moving core business applications like messaging, CRM, and ERP to cloud-based platforms is also growing the amount of cloud-born relational business data. Simply stated, cloud-born data is changing business and IT strategies about where data should be accessed, analyzed, used, and stored.

Business and IT leaders are seeking a new approach to their business intelligence and data warehouse strategies that focuses on the logic of information. This approach builds on existing best practices to add

[3] Microsoft Case Studies, *Stock Exchange Chooses Windows over Linux; Reduces Latency by 83 Percent*, http://www.microsoft.com/casestudies/Windows-Server-2008-R2-Enterprise/Direct-Edge/Stock-Exchange-Chooses-Windows-over-Linux-Reduces-Latency-by-83-Percent/4000008758, November 2010.

Logical information architecture

semantic data abstraction based on distributed processing and address the areas of data storage, virtual (any data) management, distributed processes, active system self-monitoring, service level tracking, and management based in metadata. Gartner[4] calls this next evolution in approach the **logical data warehouse**.

Change can either be a challenge or an opportunity. If an enterprise is experiencing any of the following scenarios, it may be ready to evolve to a modern data warehouse:

- The data warehouse is unable to keep up with explosive volumes.
- The data warehouse is falling behind the velocity of real-time performance requirements.
- The data warehouse is slower than desired in adopting a variety of new data sources, slowing time–to-value
- The platform costs more, while performance lags.

Evolve to a modern data warehouse

The modern data warehouse lives up to the promise of business intelligence from all data for business that is growing explosively, changing data types and sources and processing in real-time, with a more robust ability to deliver the right data at the right time.

A modern data warehouse delivers a comprehensive logical data and analytics platform with a complete suite of fully supported, solutions and technologies that can meet the needs of even the most sophisticated and demanding modern enterprise—on-premises, in the cloud, or within any hybrid scenario (Figure 3).

[4] *What is a Logical Data Warehouse* http://www.compositesw.com/solutions/logical-data-warehouse/ *Guest Post: Father of Logical Data Warehouse* http://blogs.gartner.com/merv-adrian/2011/11/03/mark-beyer-father-of-the-logical-data-warehouse-guest-post/

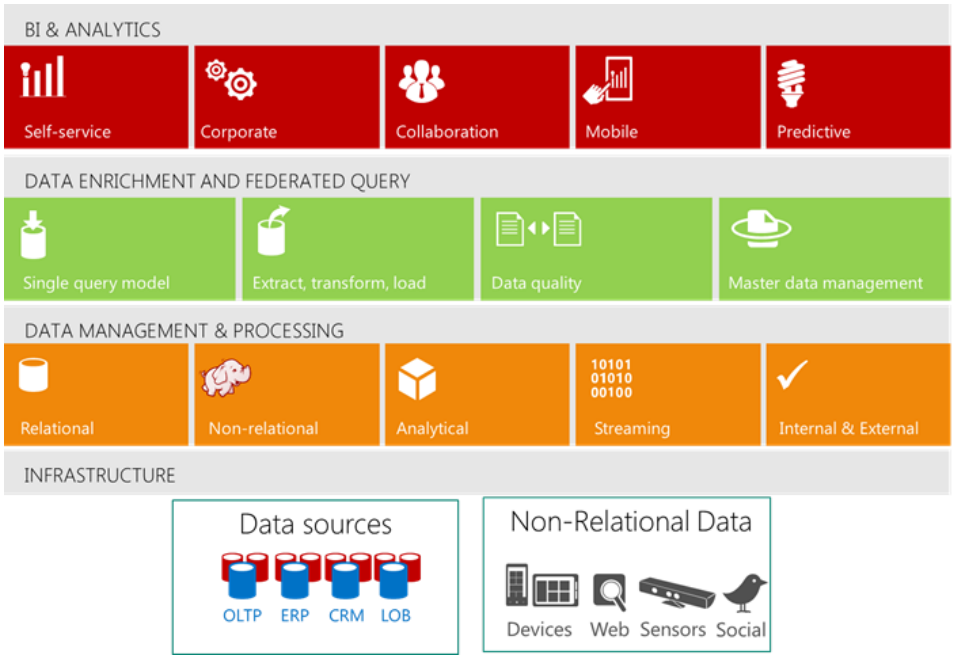

Figure 3: layers of a modern data warehouse framework

Data management and processing

The modern data warehouse starts with the ability to handle both relational and non-relational data sources like Hadoop as the foundation for business decisions. It can handle data in real-time using complex event processing technologies. It can easily augment data internal data with data from outside the organization. Finally, it provides an analytic engine for predictive analysis and interactive exploration of aggregated data from different perspectives.

Data enrichment and federated query

Next, the modern data warehouse has the ability to enrich your data with Extract, Transform and Load (ETL) capabilities as well as supporting credible and consistent data through data quality and master data management services. It also provides a single query mechanism across these different types of data through a federated query service.

Business intelligence and analytics

The modern data warehouse needs to support the breadth of tools that organizations can use to get actionable results from the data. This includes self-service tools that make it easy for business users to analyze data with tools they are familiar with already. Corporations need tools to take self-service solutions and operationalize them for broader use in their organization. Business users need a way to create and share analytics in a team environment across a variety of devices. Finally, the

platform needs to support predictive analytic models for assisting in real-time decision-making.

The Microsoft Modern Data Warehouse

The Microsoft Modern Data Warehouse can meet the needs of today's enterprise to connect agile and responsive BI to business decision makers. Highlights include the ability to handle:

- All volumes
- Real-time performance
- Any data

All volumes

Based on SMP technologies, traditional data warehouses processed queries sequentially. When more data was needed, a larger, more powerful machine was installed and the previous warehouse was forklifted into the new hardware, representing a hard limit to data size without a material new investment. In addition, SMP could be problematic on very large databases, with issues surrounding scalability of processors, cache synchronization between processors, and system performance when running concurrent loads. In total, this meant very expensive platforms with hard data size limits that slowed in performance approaching scale.

While Microsoft SQL Server is the most ubiquitous SMP database technology in the industry the Microsoft Modern Data Warehouse is designed to scale to the most demanding enterprise requirements—from 1 terabyte up to 6 petabytes with performance at linear scale.

Case Study: Hy-Vee Supermarkets

Hy-Vee boosts query performance by 100 times, and gets critical business data to analysts faster.[5]

"Using the previous system, analysts were working with data that was two weeks old, so it was difficult for them to react to trends. Now, they can view yesterday's sales data each morning. So if we're in the middle of a promotion for a certain product, analysts can come into the office in the morning and analyze how that

[5] Microsoft Case Studies, *Hy-Vee Boosts Performance, Speeds Data Delivery, and Increases Competitiveness*, http://www.microsoft.com/casestudies/Microsoft-SQL-Server-2008-R2-Enterprise/Hy-Vee/Hy-Vee-Boosts-Performance-Speeds-Data-Delivery-and-Increases-Competitiveness/710000000776, May 2012.

item has been selling, and they can order more products if they need to." – Tom Settle, Assistant Vice President, Data Warehousing

Scale out relational data

Microsoft has redesigned SQL Server into a multiple parallel processing (MPP) architecture with parallel data warehouse (PDW) distributed processing technology to handle the rigors of the modern data realities. MPP architecture enables extremely powerful distributed computing and scale. This type of technology powers supercomputers to achieve raw computing horsepower. As more scale is needed, resources can be added for a near linear scale-out to the largest data warehousing projects (Figure 5).

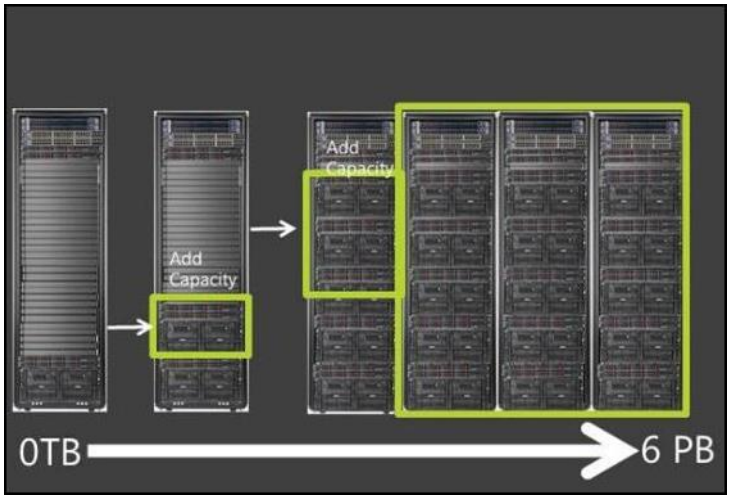

Figure 5: Scaling out relational data with the MPP architecture

MPP data architecture uses a "shared-nothing" architecture, where there are multiple physical nodes, each running its own instance of SQL Server with dedicated CPU, memory, and storage. This results in performance many times faster than traditional architectures. Customers like Hy-Vee who have upgraded their SQL Server are able to easily scale out their SQL Server data warehouse from 11 terabytes to several times that size without the need to forklift by adding incremental resources. [7].

An MPP engine enables near linear scale to support very large databases—up to the multi-petabyte capacity—with no forklift of prior warehouse data required to upgrade or grow. Capacity is added as data grows, incrementally and on a continual basis, simply by adding incremental hardware.

MPP addresses the issues related to SMP scalability of processors and synchronization of the cache between processors with its shared-

nothing architecture. As T-SQL queries go through the system, they are broken up to run simultaneously over multiple physical nodes, which can deliver the highest performance at scale through parallel execution (Figure 6). MPP architecture also enables high concurrency on complex queries at scale, which can be optimized for mixed workloads and near real-time data analysis.

The MPP architecture for the Microsoft Modern Data Warehouse is integrated within SQL Server 2012 Parallel Data Warehouse.

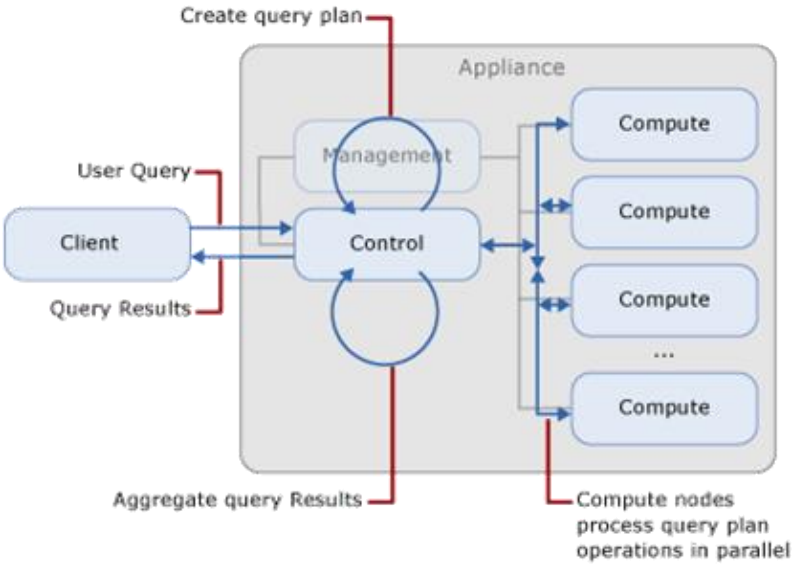

Figure 6: SQL Server 2012 PDW parallel query process

Scale out non-relational data

The traditional data warehouse added non-relational data by installing and maintaining a side-by-side, separate Hadoop ecosystem. Apache Hadoop is an open source software library framework that allows for distributed processing of large data sets across clusters of commodity computers. Hadoop offerings have driven the Big Data industry conversation because of their ability to manage large amounts of non-relational data from clusters of cost effective hardware. Hadoop uses the Hadoop Data File System (HDFS), which can support non-relational data using the MapReduce programming language. Adding scale to an existing Hadoop cluster is a matter of adding incremental Hadoop clusters (Figure 7).

Hortonworks Data Platform for Windows is available as a stand-alone software offering an Hadoop environment with cost effective hardware. HDInsight integrates Hadoop within a parallel data warehouse processing (PDW) appliance to take advantage of MPP distributed processing. Scaling to the cloud is also easily enabled with the HDInsight Hadoop service on Windows Azure. The Microsoft Modern

Data Warehouse empowers the business to scale out non-relational data with deployment agility unmatched in the industry.

Real-time performance

The traditional data warehouse was less concerned with query performance than data integrity because most analytics dealt with historical data. However, the modern enterprise works in real time and needs a data platform that can keep pace with demand without losing performance to deliver timely insights, and stream data for near real-time processing applications.

In-Memory Columnstore performance

The traditional data warehouse, which grew out of the concept of data records or rows, used a row-store based data storage design. However, rowstores are not optimal for many star schema based queries. Columnstore technology on fact tables within a star schema improves query performance for large tables by reducing the amount of data that needs to be processed through I/O.

In-Memory Columnstore changes the primary storage engine to an updateable and indexed in-memory columnar format, which groups, stores, and indexes data in compressed column segments (Figure 8).

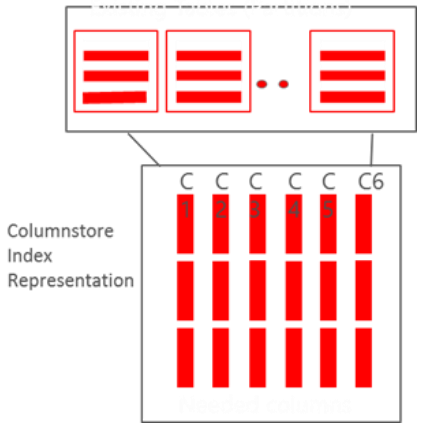

Figure 8: In-Memory Columnstore in the Microsoft Modern Data Warehouse

In-Memory Columnstore improves query performance over traditional data warehouses because only the columns needed for the query must be read. Therefore, less data is read from disk to memory and later moved from memory to processor cache. Columns are heavily compressed, reducing the number of bytes to be read or moved.

In addition, In-Memory Columnstore maximizes the use of the CPU by taking advantage of memory in processing the query, accessing data held in-memory. In-Memory Columnstore also accelerates processing speed by using the secondary columnar index to selectively query and

access columnar compressed data, further reducing the footprint and I/O to the physical media per node.

Combined, these techniques result in massive compression (up to 10 times), as well as massive performance gains (up to 100 times). In-Memory Columnstore can improve query performance even based on existing hardware investments. Customers like the Bank of Nagoya are able to leverage In-Memory Columnstore to dramatically boost query performance of key bank systems that distribute live data to the local branches to improve customer service.

In-Memory Columnstore technology is integrated into SQL Server Parallel Data Warehouse 2012 AU1 to improve the in-memory performance of every compute node in the network.

Case Study: Bank of Nagoya

Bank of Nagoya gained a 600-fold improvement in query performance by using SQL Server, which allows branches to instantly access data when talking to customers.[6]

"By using In-Memory Columnstore, we were able to extract 100 million records in 2 or 3 seconds versus the 30 minutes required previously." – Atsuo Nakajima, Assistant Director, Systems Development Group

Streaming insights with complex event processing

The traditional data warehouse provided a strategy for storing and analyzing historical data for trends and reporting. However, the modern enterprise moves in real time and needs data that can work in real time—not simply provide historical perspectives, but play an active role in optimizing operations.

Complex event processing applications enable the use of real-time streaming data created by technologies such as RFID, sensors, and other streams that support event-driven transactions. Examples of CEP applications include manufacturing process optimization, financial trading applications, web analytics, and operational analytics.

Microsoft StreamInsight is a powerful platform to develop and deploy CEP applications with low latency and sub-zero processing of large event streams (Figure 9).

[6] Microsoft Case Studies, *Bank of Nagoya Dramatically Accelerates Database Queries and Increases Availability*, http://www.microsoft.com/casestudies/Case_Study_Detail.aspx?CaseStudyID=710000000344, April 2012.

Figure 9: Using Microsoft StreamInsight to process large event streams

StreamInsight uses a high-throughput stream processing architecture and the Microsoft .NET Framework-based development platform to enable companies to quickly implement robust and highly efficient event processing applications.

With the Microsoft Modern Data Warehouse, companies can take advantage of game-changing performance 100 times faster than the traditional data warehouse and the ability to support real-time processing.

Any data

The traditional data warehouse managed historical relational data, such as ERP, CRM, and LOB outputs, with the key objective of establishing a central repository as a source of truth for the business. With Web 2.0 came a flood of new business data—including e-commerce, search marketing, collaboration, and mobile—so IT established costly ETL and data enrichment operations to bring this information into the data warehouse. This new business data expanded the relational schema model, which resulted in additional complexity.

What is Big Data?

"Big Data" is a term for the collection of data sets so large and complex that they cannot easily be managed by traditional data warehouse technologies. Big Data is the world of data that exists outside of the traditional data warehouse and enterprise. It is generated by devices; blogs and social feeds; mobile applications; clickstreams; ATM, RFID, and sensors; feeds for eGov, weather, traffic, and market sites; and so much more. Big Data is unstructured, unsanitized, and non-relational. Big Data is not generated or owned by the business.

Big Data is valuable to the business because it brings an enterprise into context with the world in which it operates, competes, and sells. Big Data offers the opportunity for the enterprise to engage with outside data in near real time to enhance, optimize, and move the business forward.

According to Gartner, "Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to

enable enhanced decision making, insight discovery, and process optimization."[7]

Common scenarios for Big Data

The popularity of Big Data is based predominantly on the tidal wave of new scenarios, data sources, and opportunities to integrate non-relational data from outside of an enterprise into its business analytics (Figure 10).

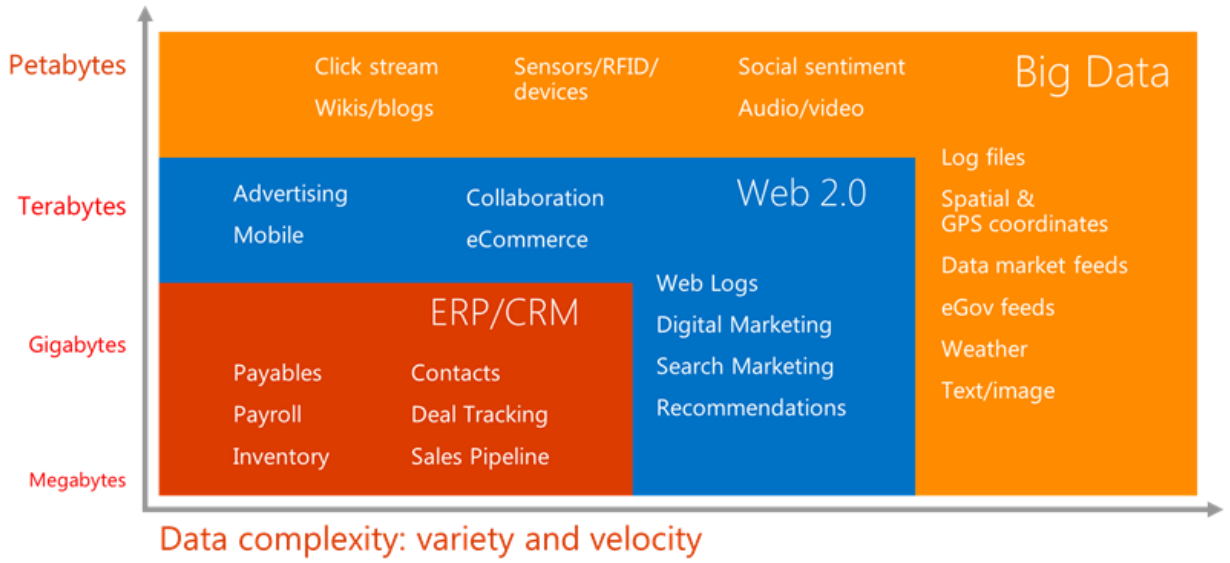

Figure 10: Complexity of Big Data in the modern business environment

Big Data can drive value in a wide range of emerging scenarios where new data sources or uses are changing how business is done. Example scenarios include IT infrastructure optimization, manufacturing process optimization, legal discovery, social network analysis, traffic flow optimization, web app optimization, integration of location-based information, churn analysis, natural resource exploration, weather forecasting, healthcare, fraud detection, life science research, advertising analysis, and smart meter monitoring.

The Microsoft Modern Data Warehouse can unlock the big value of Big Data.

What is Hadoop?

Apache Hadoop is an open-source solution framework that supports data-intensive distributed applications on large clusters of commodity hardware. The key benefit of Hadoop is the ability to process any non-relational data.

[7] Beyer, Mark A. and Douglas Laney (for Gartner), *The Importance of "Big Data": A Definition*, http://www.gartner.com/id=2057415, June 21, 2012.

There are several market solutions that customize or package Hadoop Data Platforms, such as Hortonworks, RMap, Cloudera, IBM, and others. The Hadoop framework is composed of a number of components, including:

- Data storage based on HDFS, Hbase, NFS, and CloudStore.
- Query processing based on MapReduce framework.
- Data access using Hive (SQL-like), Pig (data flow), Avro (JSON), Mahout (machine learning), and Sqoop (data connector).
- Data management based on Oozie (workflow), EMR (managed services), Chukwa or Flume (data management), and Zookeeper (system management).

The MapReduce framework is a programming model for taking data on a Hadoop file system and processing it as sets of key-value pairs. Applications written for Hadoop primarily use mapper and reducer interfaces, including tools like Apache Hive that provide a data warehouse infrastructure on top of the files, along with a SQL-like query language called HiveQL.

Case Study: Direct Edge Stock Exchange

Direct Edge, one of the largest equities exchanges in the world, wanted a better, faster BI solution for creating financial analysis reports. The company implemented a data warehouse and BI solution based on Microsoft SQL Server 2008 R2 Parallel Data Warehouse and Apache Hadoop. The solution provides more visibility into data and can deliver reports in seconds rather than hours, helping to drive better business growth.[8]

"Due to PDW's smooth integration with Hadoop, Direct Edge can use unstructured data for Big Data analysis, unlocking new analytic scenarios. Our analysts have a much deeper understanding of trading data. For example, they can better understand monthly fluctuations in trading fee revenue." – Richard Horchron, Chief Technology Officer

Seamless integration of Hadoop non-relational data

The traditional data warehouse did not anticipate or integrate Hadoop. Adopting Hadoop meant setting up and maintaining a separate, side-by-side Hadoop data warehouse next to the relational data warehouse. This significantly increased the learning curve and costs associated with development and maintenance, while slowing time-to-value.

[8] Microsoft Case Studies, *Stock Exchange Gains Deeper Understanding of Data and Drives New Business Growth*, http://www.microsoft.com/casestudies/Microsoft-Excel-2010/Direct-Edge/Stock-Exchange-Gains-Deeper-Understanding-of-Data-and-Drives-New-Business-Growth/710000002540, May 2013.

The Microsoft Modern Data Warehouse integrates Hadoop to provide the ability to seamlessly manage relational and non-relational data from a shared query model, infrastructure and ecosytem.

(Figure 11).

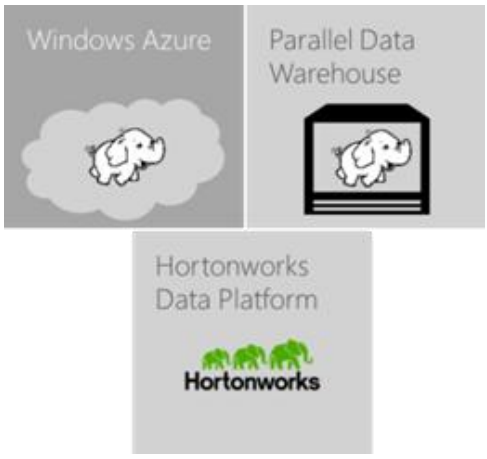

Figure 11: Creating an agile Hadoop cluster with Windows Azure and Hortonworks HDP

Hortonworks HDP for Windows

Hortonworks Data Platform (HDP) for Windows is a 100% Apache Hadoop software solution architected for the enterprise that implements a stand-alone Hadoop environment using cost effective hardware clusters. HDP for Windows enables the power of Hadoop with the simplicity and management of Microsoft. HDP for Windows enables seamless integration with the Microsoft BI tool ecosystem and is the only Hadoop distribution available for Windows Server.

HDInsight within the PDW appliance

HDInsight is HDP for Windows based software offering within a SQL Server PDW (parallel data warehouse) appliance. HDInsight installs a dedicated Hadoop region directly over the fabric layer of the appliance alongside the distributed PDW query engine sharingmetered resources for CPU, memory, and storage. The HDInsight region is a logical layer with boundaries for workload, security, metering, and servicing. HDInsight embeds Hadoop non-relational data processing directly into the parallelized distributed processing network. This enables seamless processing and scale within an integrated ecosystem and footprint.

The HDInsight region addresses customer requirements for relational and non-relational data strategies within a logical framework. HDInsight supports a number of key scenarios, including using it as a staging area for relational processing, enabling trickle loading, or using Hadoop as cold data storage.

HDInsight includes the following components:

Storage integration	M/R + abstractions	Job submission	Data
HDFS Web HDFS	MapReduce Hive (SQL over Hadoop) Pig (data flow)	WebHCat Oozie (workflow)	Hive OBDC Sqoop (connectors)

Windows Azure HDInsight

HDInsight is a 100 percent compatible Apache Hadoop solution to the cloud. The HDInsight service is a modern, enterprise-ready, cloud-based solution available only on Windows Azure.

The HDInsight service enables the business to seamlessly process Hadoop data in the cloud. Deep integration with Microsoft BI tools such as Power Pivot, Power View, Power Map or other Power BI cloud services enables business users and decision makers to easily analyze Hadoop data for insights.

Using a rich library of Powershell scripts, IT can deploy and provision a new Hadoop cluster deployment in only minutes, with easy upgrading to larger clusters without losing data. Using Secure Node, HDInsight offers enterprise-class security, scalability and manageability.

Integrated Relational and Non-Relational Data

For the traditional data warehouse to integrate their Hadoop solutions with their existing data warehouse, IT would typically need to pre-populate the warehouse with Hadoop data through an extensive data mapping and data movement project. A common alternative to these expensive ETL (extract, transform, load) operations has been to require extensive user training on MapReduce in order to query their Hadoop data.

Microsoft introduced PolyBase to address the unification of the traditional data warehouse and the new Hadoop offerings with the ability query relational and non-relational data in Hadoop with a single, T-SQL-based query model that can support both relational and non-relational data in parallel—resulting in dramatic performance improvements over traditional data warehouses and Apache Hive solutions.

PolyBase was pioneered and created in Jim Gray Systems Labs by David DeWitt, Professor Emeritus of Computer Sciences (University of Wisconsin, Madison). Dr. DeWitt is known for revolutionary research in parallel databases, benchmarking, and object-oriented and XML databases. PolyBase supports multiple third party Hadoop distribution

including Hortonworks Data Platform, Hortonworks Linux, and Cloudera Linux CHD. PolyBase also supports integration with BI in Excel Services, Power BI for Office, and SQL Server Reporting and Analysis Services which gives you the ability to query any third party Hadoop through the familiar Microsoft BI tools. **Integrated query model with PolyBase**

PolyBase is an integrated query processor available within SQL Server Parallel Data Warehouse 2012 AU1. PolyBase makes it possible to import and export data between HDFS and relational sources using a single (T-SQL) query and processing model without the need to learn MapReduce or HiveQL. The PolyBase Data Movement Service (DMS) works with the HDInsight HDFS bridge to parallelize and distribute the query processing of complex non-relational queries, improving performance and enabling the processing of Hadoop data in-situ (or "in place"), without the need for expensive ETL processes (Figure 12).

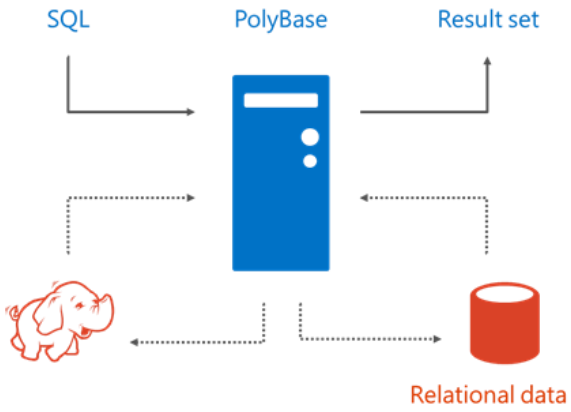

Figure 12: Integrated query model, powered by PolyBase

PolyBase enables the business to lower training and development costs, eliminate the cost of supporting an additional ecosystem and improves overall time to value for new data.

The traditional data warehouse was an on-premises operation, and the larger it grew, the more IT infrastructure and resources were required to support it. The Microsoft Modern Data Warehouse provides companies with several deployment scenarios and strategies to fit their unique business needs and plans. Rather than requiring an enterprise to buy an expensive offering or forcing the business to go to the cloud, Microsoft has a breadth of offerings that span delivery vehicles and can be mixed-and-matched as the business and data warehouse evolve.

Deployment options and hybrid solutions

Box software

As the business environment evolves, there are many reasons why companies might want to combine powerful software and custom-built hardware for their data warehouse installations, or use this software to extend existing investments in infrastructure. Companies may have negotiated hardware and software licensing deals, or perhaps they have unique security requirements. Regardless of the reason, software can provide these companies with the highest levels of flexibility in hardware size, configuration and tuning.

The software foundation for the modern data warehouse is the ubiquitous, industry-leading SQL Server, the most widely deployed database, delivering the required 9s of availability and reliability with its AlwaysOn functionality and failover technologies. Already an industry standard, SQL Server features ongoing investments and improvements for technologies such as In-Memory Columnstore (Figure 13).

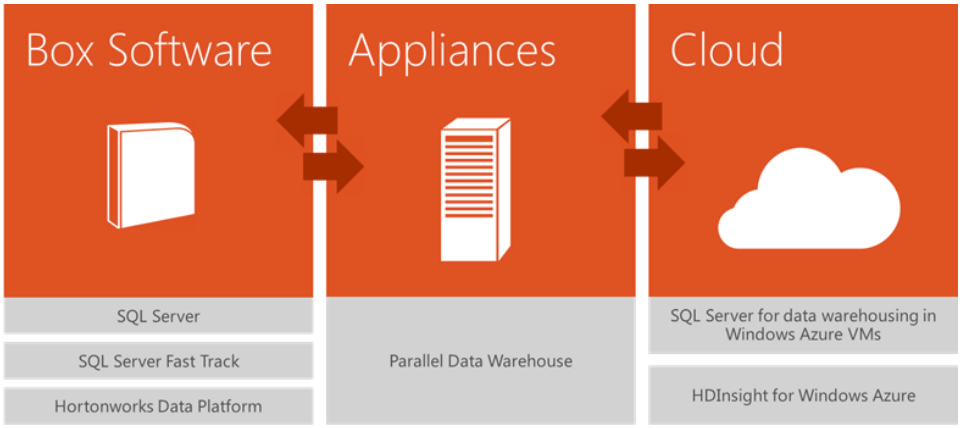

Figure 13: Maximizing data warehouse investments with box software

Adding to this foundation is Microsoft SQL Server Fast Track, a reference architecture and set of prescriptive guides delivered through Fast Track partners that simplify the process of building a data warehouse and integrating with over 10 hardware partners. Fast Track gives customers step-by-step instructions of how to build hardware servers with the right mix of CPU, I/O, and storage. It also provides guidance for tuning the software for optimal performance. Fast Track partners such as HP, IBM, Dell, Cisco, Hitachi XIO, Violin, NEC, Huawei, Fujitsu NetApp, and EMC bring additional expertise and best practices to on-premises deployments.

Hortonworks Data Platform (HDP) for Windows is the result of a partnership between Hortonworks and Microsoft to bring the benefits of

Apache Hadoop to Windows. Hortonworks Data Platform for Windows delivers an enterprise-ready Hadoop solution that deeply integrates Microsoft tools and applications to empower the business to access Hadoop data using familiar tools. Hortonworks Data Platform for Windows offers unprecedented choices for deploying Hadoop only within the Enterprise.

Prebuilt appliance

An appliance is prebuilt hardware with preinstalled software, configured and tuned for use. The value of an appliance is the ability to quickly add incremental plug-and-play resources tuned for optimal performance.

SQL Server Parallel Data Warehouse is built on MPP architecture configured with In-Memory Columnstore, PolyBase, and HDInsight. Setup is highly streamlined, and companies can simply plug in the appliance without building specialized infrastructure from disparate hardware or seeking experts to install and tune the software. This saves time and money on research and deployment and minimizes the need to hire expensive technical consultants.

SQL Server 2012 Parallel Data Warehouse provides highly scalable hardware architecture, allowing companies to start with a small data warehouse of 1 terabyte that linearly scales out to as many as 6 petabytes of data storage. The appliance is designed to work with 2 to 64 nodes for maximum scalability. Each node runs its own instance of Microsoft SQL Server 2012 with dedicated CPU, memory, networking, and storage configurable for any combination of Hadoop or relational processing. This means that companies can add capacity to the initial rack and, if necessary, simply add more racks to the appliance.

Parallel Data Warehouse also offers:

- Distributed architecture that integrates both MPP and SMP data warehouses.
- HDInsight dedicated Hadoop region within parallel data warehouse appliance.
- PolyBase integrated relational/non-relational query processing model.
- Integration with Microsoft BI tools, including Analysis Services, Integration Services, and Microsoft SharePoint with optional Power BI for Office 365.
- Comprehensive toolset for ETL, BI, MDM, and streaming data with StreamInsight.
- Interoperability with non-Microsoft BI and ETL tools, such as SAP Business Objects, SAS, Informatica, and Microstrategy.

Microsoft has partnered and extensively co-engineered appliance solutions with Dell, HP, and Quanta. The following table provides information about baseline rack designs.

	Dell Parallel Data Warehouse appliance	HP Enterprise Data Warehouse appliance	Quanta Parallel Data Warehouse appliance
			
Servers	PowerEdge R620	ProLiant Gen8 DL360	STRATOS S810-x52L
Computer nodes	Up to 9 per rack (3 minimum)	Up to 8 per rack (2 minimum)	Up to 8 per rack (2 minimum)
Racks	¼ to 6	¼ to 7	¼ to 7
Raw disk capacity (uncompressed)	0TB – 1.2PB	0TB – 1.2PB	0TB – 1.2PB

Cloud-based deployment

In the past, deploying a data warehouse has been a costly, strictly on-premises endeavor. The IT organization would purchase and install state-of-the-art hardware servers, optimally balanced and tuned for CPU, I/O, and storage. IT also would install the software and tune it for performance. Even before loading data, the company could spend months and hundreds of thousands of dollars on infrastructure with ongoing maintenance, support, and replacement.

A cloud deployment implements the same BI strategy but replaces on-premises infrastructure with a Windows Azure cloud infrastructure maintained by Microsoft. Customers save the cost of maintaining on-premises infrastructure in exchange for a low monthly service fee, lowering TCO. More importantly, high-value IT resources spend more time building the business and less time supporting infrastructure.

The Microsoft Modern Data Warehouse offers the option to deploy directly to the cloud with the elastic scalability of Windows Azure. SQL Server Enterprise for data warehousing can be installed and hosted in the cloud on Windows Azure Virtual Machines. This image takes advantage of best practices from the Fast Track reference architecture to tune SQL Server for data warehousing in Windows Azure. Users can provision a highly tuned data warehouse image within minutes without knowing Azure storage configurations or needing expertise on how to optimize SQL Server for data warehousing workloads. This is an ideal solution for customers who want to deploy a data warehouse quickly without having to manage a hardware infrastructure.

Customers also can benefit from deploying non-relational Hadoop data in the cloud using the HDInsight Service on Windows Azure. The HDInsight Service provides an Hadoop solution that can seamless process data of all types through Microsoft's modern data platform, which provides the simplicity, ease of management and enterprise-ready Hadoop service in the cloud.

Users can deploy and provision an HDInsight Hadoop cluster in minutes instead of hours or days with the full elastic scalability of Windows Azure.

Finally, with Windows Azure, customers can keep cloud-born data in the cloud and reduce capital and operating expenses by using cloud and hybrid scenarios like simple smart cloud backup, disaster recovery, and extension of on-premises applications.

Conclusion

The opportunities of Big Data are as big as the challenges. The most sophisticated traditional data warehouse is changing to meet the requirements of the modern data enterprise. Volume increases are expected to continue. Business velocity will continue to change business operations and customer interactions. Data will become even more diverse and more available than ever before. Big Data can mean big impact to the business. To tap into the immense new opportunities of Big Data, the modern enterprise needs a modern data platform. The Microsoft Modern Data Warehouse delivers this platform, solutions, features, functionality and benefits that empower the modern enterprise in three essential areas:

- All volumes (with nearly unlimited elasticity).
- Real-time performance (at scale).
- Any data (seamless integration across relational and non-relational).

Get started today

Sign up for a free architectural design session for data warehousing with your Microsoft rep.

Try HDInsight at www.microsoft.com/bigdata.

Try SQL Server for data warehousing in Windows Azure Virtual Machines at www.windowsazure.com.

Try SQL Server 2014 at http://www.microsoft.com/en-us/sqlserver/sql-server-2014.aspx.

For more information

SQL Server website: http://www.microsoft.com/sqlserver/

SQL Server TechCenter: http://technet.microsoft.com/en-us/sqlserver/

SQL Server DevCenter: http://msdn.microsoft.com/en-us/sqlserver/

SQL Server data warehousing: http://www.microsoft.com/sqlserver/en/us/solutions-technologies/data-warehousing.aspx

Fast Track data warehousing: http://www.microsoft.com/sqlserver/en/us/solutions-technologies/data-warehousing/fast-track.aspx

SQL Server Parallel Data Warehouse: http://www.microsoft.com/sqlserver/en/us/solutions-technologies/data-warehousing/pdw.aspx

Microsoft Big Data solution: http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/big-data-solution.aspx

Join the conversation

www.microsoft.com/sqlserver