



The Microsoft Modern Data Warehouse



Contents

4	Executive summary
4	The traditional data warehouse
5	Key trends breaking the traditional data warehouse
6	Increasing data volumes
7	Real-time data
7	New sources and types of data
8	Deployment: cloud-and hybrid
8	Advanced Analytics & ML
9	Evolve to a modern data warehouse
10	The Microsoft Modern Data Warehouse
12	All volumes at high performance
16	Relational and non-relational
21	On-premises and cloud
24	Analytics and Machine Learning
26	Conclusion
26	For more information

Executive summary

Data has become the strategic asset used to transform businesses to uncover new insights. Traditionally, data has been gathered in an enterprise data warehouse where it serves as the central version of the truth. However, the world of data is rapidly evolving in ways that are transforming the industry and motivating enterprises to consider new approaches of gaining insights. Beyond the traditional sources from transactional systems, ERP, CRM, and LOB applications, new types of data sources are driving analytics that are transformative to the business. And it is coming from data generated by everything around us like social media apps, websites and connected devices. Collectively, IDC projects that this explosion of data will result in a 40 Zetabyte digital universe by 2020.

The challenge for IT organizations is their traditional enterprise data warehouse was never designed to incorporate this explosion of new types of data at this volume and velocity. To solve for this will require dramatic changes so much so 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 To drive the business forward, the modern enterprise needs to evolve their enterprise data warehouse so that it can take advantage of big data and do so in real time. Once all data has been incorporated, this lets business analysts and data scientists uncover new insights that impact the business. To do this, 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 are 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 operational 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, 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, new deployment models in the cloud and hybrid, and the need to do advanced analytics & machine learning—are discussed below. Each of these trends are driving organizations to examine how they evolve their existing data platform to become a modern data warehouse.

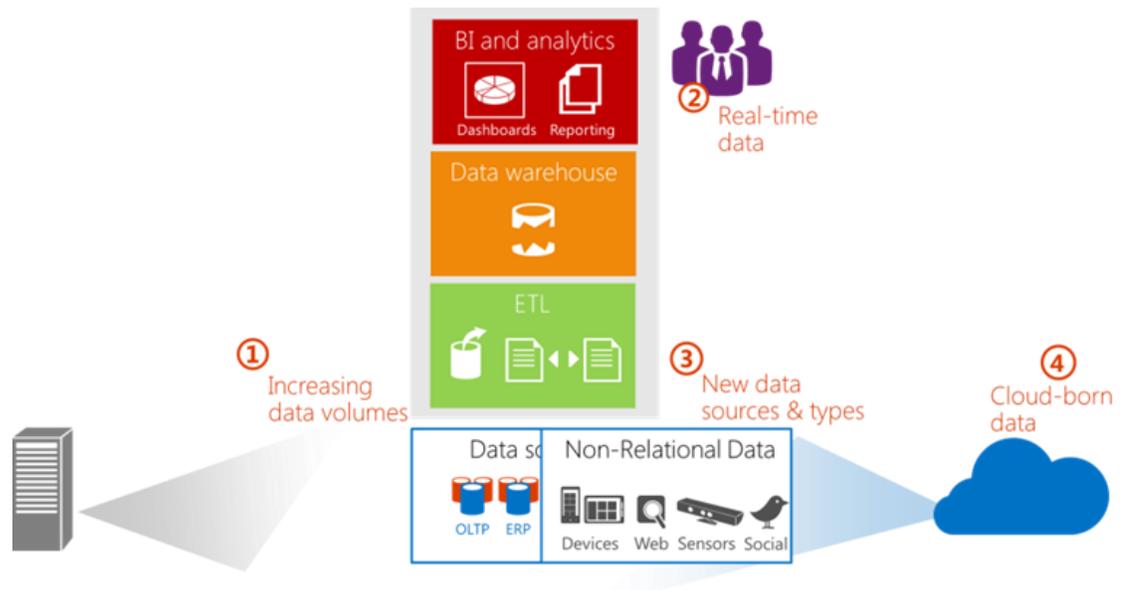


Figure 2: Four key trends breaking the traditional data warehouse

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. Internet of Things (IoT) Devices 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

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/71000000776>, May 2012.

"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 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.

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.

New Deployment models: cloud-and hybrid

Advanced Analytics and Machine Learning

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

The cloud has quickly become an integral part of many IT organizations with recent research from cloud solutions provider RightScale showing 93% of businesses using cloud technology.⁴ Forrester recently did a study where they found 47% of organizations increasing their cloud deployments for big data specifically.⁵ It makes sense because the cloud not only enables cost efficiencies, it gives you the scale to meet demands to process any amount of data now and in the future.

This trend means an increasing share of new data is simply “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, the cloud is changing business and IT strategies about where data should be accessed, analyzed, used, and stored.

Business and IT leaders are seeking new approaches to uncover insights and create new business opportunities. To do this, many organizations are implementing advanced and predictive analytics to figure out what is likely to happen from an increasingly varied set of data sources and types. However, the traditional data warehouse is not designed for these new types of analytics because the analytical style is inductive, or bottoms-up in nature. Instead of working through a requirements-based model of the traditional data warehouse where the schema and data collected is defined upfront, advanced analytics and data science uses the experimentation approach of exploring answers to ill-formed or nonexistent questions.⁶ This requires the examination of data before it is curated into a schema allowing the data to drive insight in itself. Gartner recommends both approaches and for established data warehouse teams to collaborate with this new breed of data scientists as part of a move towards the logical data warehouse.⁷

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:

⁴ 2015 State of the Cloud Report: See the Latest Cloud Computing Trends <http://www.rightscale.com/lp/2015-state-of-the-cloud-report?campaign=701700000012UP6>

⁵ The Forrester Wave™: Big Data Hadoop Cloud Solutions, Q2 2016 <https://www.forrester.com/report/The+Forrester+Wave+Big+Data+Hadoop+Cloud+Solutions+Q2+2016/-/E-RES126541#figure4>

⁶ Big Data Business Benefits Are Hampered by ‘Culture Clash’ <https://www.gartner.com/doc/2588415/big-data-business-benefits-hampered>

⁷ Big Data Business Benefits Are Hampered by ‘Culture Clash’ <https://www.gartner.com/doc/2588415/big-data-business-benefits-hampered>

- 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 not incorporating a variety of new data sources, slowing down the ability to do advanced analytics with the new breed of data scientists
- The platform costs more, while performance lags

Evolve to a modern data warehouse

The modern data warehouse lives up to the promise to get insights from a variety of techniques like business intelligence and advanced analytics from all types of data that are growing explosively, 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).

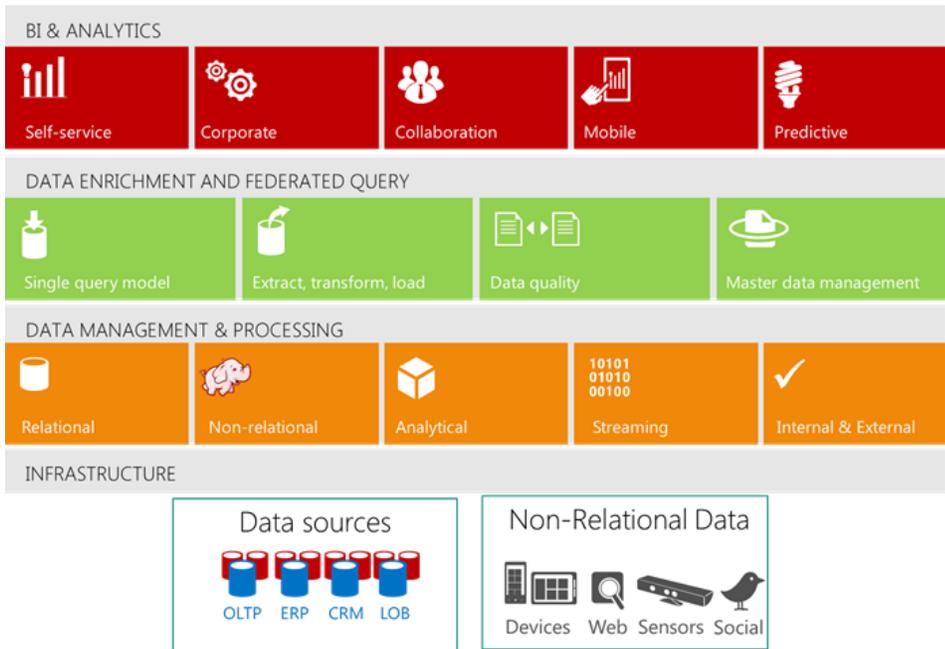


Figure 3: layers of a modern data warehouse framework

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 real-time streaming solutions. It can easily augment on-premises, internal data with data from outside the firewall. 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 or supports the ability to do data preparation through big data techniques as well as supporting credible and consistent data through data quality and master data management services. It also provides a way to easily query relational and non-relational data through a single federated query service.

Business intelligence and data science

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 analysts to analyze data with BI tools they are familiar with already. Business users need a way to create and share analytics in a team environment across a variety of devices. The platform also needs to support a new breed of data scientists who are running experimentation with the data, doing predictive analytic modeling, and assisting in real-time decision-making.

The Microsoft Modern Data Warehouse

The Microsoft modern data warehousing portfolio offers a complete set of solutions that span on-premises and cloud to enable choice and flexibility for your ever-evolving needs to encompass relational and non-relational data. By choosing the right deployment option for each of your workloads, you can meet your workloads where they are versus shoehorning them into a one-size-fits-all approach.

On-premises, you can deploy **SQL Server** software on your existing hardware for a data warehouse, or **SQL Server Fast Track** certified reference architectures can help speed up your time-to-value. For extremely high-volume and complex data warehousing requirements, Microsoft offers **Analytics Platform System (APS)**, a dedicated scale-out data warehousing appliance, pre-tuned and pre-configured on a choice of HPE, Dell, and Quanta hardware.

In the cloud, you can drive even more value and faster time to insight across **Cortana Intelligence Suite** which includes **Azure SQL Data Warehouse, Azure Data Lake, Azure HDInsight (Hadoop and Spark), Advanced Analytics, and Power BI** for approximately 10x the value of a traditional appliance plus analytics solutions.

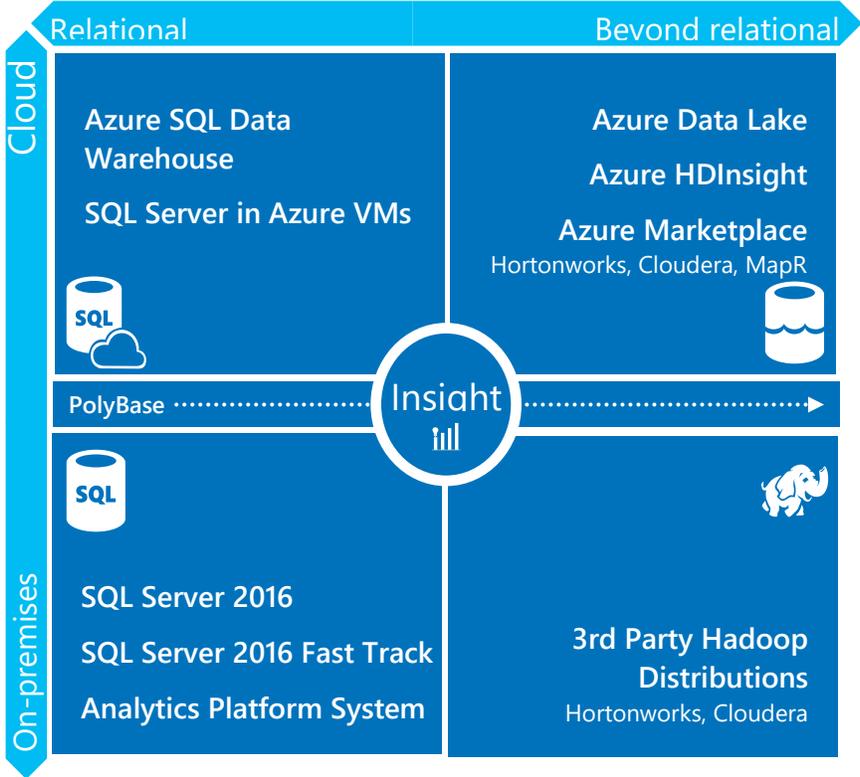


Figure 4: Microsoft's Modern Data Warehouse

With this end-to-end portfolio, you can:

- Easily manage relational and non-relational data at all volumes and high performance
- Enjoy a consistent experience across on-premises and cloud
- Gain insights from BI and advanced analytics across all your data wherever it resides

All volumes of data at high performance

Whether it be on-premises or cloud over relational or non-relational data, Microsoft modern data warehouse technologies have broken barriers to handle all volumes of data at high performance through in-memory columnstore, MPP technologies, and optimizations on the core query engine.

SQL Server 2016 as an example has ground-breaking performance optimizations and efficiencies, leading to new levels of performance and scale. Modern servers can support a large number of cores with sophisticated vector instructions, can hold terabytes of memory, and provide very high I/O bandwidth with local flash storage. Optimizing for the concurrency and parallelism inherent within such servers can provide dramatic speedups at scale, and often outperform large distributed databases.

For example, Microsoft recently collaborated with Intel to demonstrate stunning performance on a massive 100TB data warehouse using just a single server with four Intel Xeon E7 processors and SQL Server 2016. The system was able to load a complex schema derived from TPC-H at **1.6TB/hour**, and it took just **5.3 seconds** to run a complex query (the minimum cost supplier query) on the entire **100TB** database.

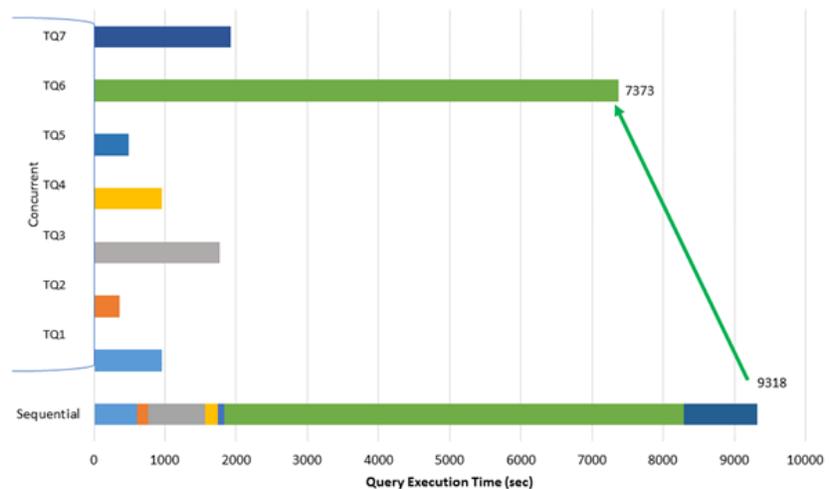


Figure 5: SQL Server benchmark

Industry standard database benchmarks further support SQL Server's performance leadership. It is the only database that holds the world record for the data warehouse benchmark TPC-H (non-clustered TPC-H **30TB, 10TB, 3TB, 1TB** scale factors). For customers, this means incredible value compared to other data warehouses with the highest performance and significant savings in total cost of ownership.

Case Study: Hy-Vee Supermarkets

Hy-Vee boosts query performance by 100 times, and gets critical business data to analysts faster.⁸

"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 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 on-premises and in the cloud

With the Analytics Platform System, Microsoft has also redesigned SQL Server into a multiple parallel processing (MPP) architecture with the analytics platform system 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 6: 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 were 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

⁸ 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/71000000776>, May 2012.

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.

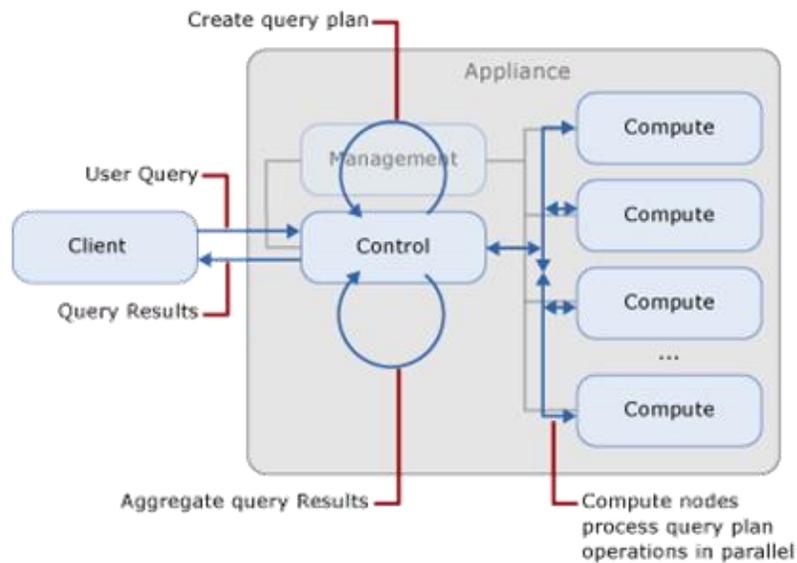


Figure 7: Analytics Platform System parallel query process

In the cloud, Azure SQL Data Warehouses leverages the same MPP architecture as the Analytics Platform System letting you combine the scaling power of this architecture with the elasticity of the cloud. A defining characteristic of cloud computing is elasticity – the ability to rapidly provision and release resources to match what a workload requires – so that a user pays no more and no less than what they need to for the task at hand. Such just-in-time provisioning can save customers enormous amounts of money when their workloads are intermittent and heavily spiked.

Azure SQL Data Warehouse is a fully managed DW as a Service that you can provision in minutes and scale up to 60 times larger in seconds. With a few clicks in the Azure Portal, you can launch a data warehouse, and start analyzing or querying data at the scale of hundreds of terabytes. Our architecture separates compute and storage so that you can independently scale them, and use just the right amount of each at

any given time. A very unique pause feature allows you to suspend compute in seconds and resume when needed while your data remains intact in Azure storage.

Case Study: PoundSand

PoundSand had to elastically scale their data warehouse to keep up with 100x increase in demand

"Getting featured in the iOS App Store was a big deal for a small company like ours as our users increased from 3,000 to 300,000 in 48 hours. To keep up with this 100x increase in workload, we simply added data warehouse compute capacity by moving a slider and our services just scaled in minutes—we didn't miss an insight," notes Paul Ohanian, CTO, PoundSand

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).

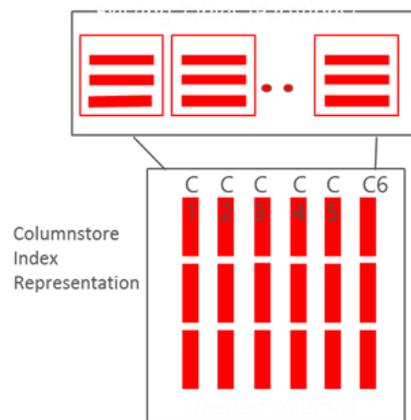


Figure 8: In-Memory Columnstore in the Microsoft Modern Data Warehouse

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 2016, Analytics Platform System, and Azure SQL Data Warehouse 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.⁹

"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

Relational and non-relational 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;

⁹ Microsoft Case Studies, *Bank of Nagoya Dramatically Accelerates Database Queries and Increases Availability*, http://www.microsoft.com/casestudies/Case_Study_Detail.aspx?CaseStudyID=71000000344, April 2012.

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.

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."¹⁰

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 9: 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.

¹⁰ Beyer, Mark A. and Douglas Laney (for Gartner), *The Importance of "Big Data": A Definition*, <http://www.gartner.com/id=2057415>, June 21, 2012.

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 Analytics Platform System 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.¹¹

"Due to APS'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

Microsoft's big data solutions is encompassed in the Azure Data Lake. Azure Data Lake includes all the capabilities required to make it easy for developers, data scientists, and analysts to store data of any size, shape and speed, and do all types of processing and analytics across platforms and languages. It removes the complexities of ingesting and storing all of your data while making it faster to get up and running with batch, streaming, and interactive analytics. Within this solution, there are three services: Azure HDInsight, Azure Data Lake Analytics, and Azure Data Lake Store.

Azure HDInsight-managed Apache Hadoop®, Spark, HBase, and Storm

The Azure Data Lake offers fully managed and supported 100% Apache Hadoop®, Spark, HBase, and Storm clusters. You can get up and running quickly on any of these workloads with a few clicks and within a few minutes without buying hardware or hiring specialized operations teams typically associated with big data infrastructure. You have the choice of running Linux or Windows with the Hortonworks Hadoop Data Platform, making it easy to move code and projects to the cloud. Finally, the rich ecosystem of Apache Hadoop-based applications provides security, governance, data preparation, and advanced analytics, letting you get the most out of your data faster. With Data Lake, Hadoop is made easy.

Azure Data Lake Analytics

The Data Lake analytics service is a new distributed analytics service built on Apache YARN that dynamically scales so you can focus on your business goals, not on distributed infrastructure. Instead of deploying, configuring and tuning hardware, you write queries to transform your data and extract valuable insights. The analytics service can handle jobs of any scale instantly by simply setting the dial for how much power you

¹¹ 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/71000002540>, May 2013.

need. You only pay for your job when it is running making it cost-effective. The analytics service supports Azure Active Directory letting you simply manage access and roles, integrated with your on-premises identity system. It also includes U-SQL, a language that unifies the benefits of SQL with the expressive power of user code. U-SQL's scalable distributed runtime enables you to efficiently analyze data in the store and across SQL Servers in Azure, Azure SQL Database and Azure SQL Data Warehouse.

Case Study: Plexure

Plexure uses Azure Data Lake to ingest and process massive amounts of big data from mobile, web, IoT, and retail transactions

"We ingest a massive amount of live data from mobile, web, IoT and retail transactions," says David Inggs, CTO at Plexure. Data Lake gives us the ability to easily and cost effectively store everything and analyse what we need to, when we need to. The simplicity of ramping up parallel processing on the U-SQL queries removes the technical complexities of fighting with the data and lets the teams focus on the business outcomes. We are now taking this a step further and exposing the powerful Data Lake tools directly to our clients in our software allowing them to more easily explore their data using these tools."

Azure Data Lake Store

The Data Lake store provides a single repository where you can capture data of any size type and speed simply without forcing changes to your application as the data scales. In the store, data can be shared for collaboration with enterprise-grade security. It is also designed for high-performance processing and analytics from HDFS applications and tools, including support for low latency workloads. For example, data can be ingested in real-time from sensors and devices for IoT solutions, or from online shopping websites into the store without the restriction of fixed limits on account or file size unlike current offerings in the market.

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. 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).

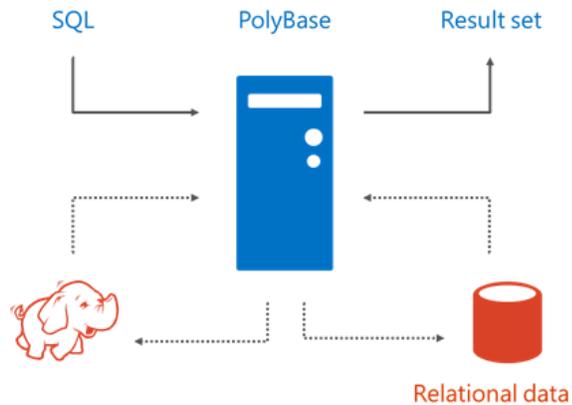


Figure 10: Integrated query model, powered by PolyBase

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. This capability is available as part of SQL Server 2016, Analytics Platform System, and Azure SQL Data Warehouse.

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.

Consistent experiences on-premises and cloud

On-premises

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.

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).

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, Dell, NEC, i, Fujitsu, NetApp, Pure Storage, Tegile, Lenovo and EMC bring additional expertise and best practices to on-premises deployments.

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.

Analytics Platform System is built on MPP architecture configured with In-Memory Columnstore, and PolyBase. 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

Cloud-based

consultants. APS 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 2016 with dedicated CPU, memory, networking, and storage. This means that companies can add capacity to the initial rack and, if necessary, simply add more racks to the appliance.

Microsoft has partnered and extensively co-engineered appliance solutions with Dell, HP, and Quanta.

A defining characteristic of cloud computing is elasticity – the ability to rapidly provision and release resources to match what a workload requires – so that a user pays no more and no less than what they need to for the task at hand. Such just-in-time provisioning can save customers enormous amounts of money when their workloads are intermittent and heavily spiked. And in the modern enterprise, there are few workloads that have a desperate need for such elastic capabilities as data warehousing and big data. Traditionally built on-premises with very expensive hardware and software, most enterprise Data Warehouse (DW) systems have very low utilization except during peak periods of data loading, transformation and report generation.

The Microsoft Modern Data Warehouse offers the most comprehensive options to deploy data warehousing and big data directly to the cloud with the elastic scalability of Azure.

Azure SQL Data Warehouse, is Microsoft's relational, MPP data warehouse that brings the true promise of cloud elasticity to data warehousing. It is a fully managed DW as a Service that you can provision in minutes and scale up to 60 times larger in seconds. With a few clicks in the Azure Portal, you can launch a data warehouse, and start analyzing or querying data at the scale of hundreds of terabytes. Our architecture separates compute and storage so that you can independently scale them, and use just the right amount of each at any given time. A very unique pause feature allows you to suspend compute in seconds and resume when needed while your data remains intact in Azure storage. And SQL Data Warehouse offers an **availability SLA of 99.9%** – the only public cloud data warehouse service that offers an availability SLA to customers.

Case Study: Integral Analytics

Integral Analytics left AWS Redshift to Azure SQL Data Warehouse because it can suspend compute and save them on their bottom line.

"When we learned about the pause and resume capabilities of SQL Data Warehouse and integrated services like Azure Machine Learning and Data Factory, we switched from AWS Redshift, migrating over 7TB of uncompressed data over a week for the simple reasons of saving money and enabling a more straight-forward implementation for advanced analytics. To meet our business intelligence requirements, we load data once or twice a month and then build reports for our customers. Not having the data warehouse service running all the time is key for our business and our bottom line," said Bill Sabo, managing director of information technology at Integral Analytics.

Customers also can benefit from deploying non-relational data in the cloud using the Azure Data Lake.

Azure Data Lake solves these two fundamental challenges of cloud-based big data solutions by providing Azure Data Lake Store, the industry's very first cloud Hadoop File System designed from the ground up for big data analytics, Azure Data Lake Analytics, a fully managed, clusterless big data analytics service that speeds the time required for a user to be productive on big data with minimal IT or software development skills, and Azure HDInsight which provides managed Hadoop and Spark clusters in the cloud.

Azure Data Lake Store is a single repository to build cloud-based data lakes to capture and access any type of data for high-performance processing and analytics and low latency workloads with enterprise-grade security. This lets you store data in a single place and use any type of analytics to process it such as Azure HDInsight (Hadoop and Spark), R Server, Hortonworks, Cloudera, and Azure SQL Data Warehouse. Unique to the Data Lake Store is that it can handle both the largest datasets as well as low latency data that needs to be ingested in real-time from sensors and devices for IoT solutions. While other cloud object stores have restrictions or fixed limits to account or file sizes, Data Lake Store can scale to individual petabyte-sized files, 200x larger than the file size limits of other cloud storage offerings. Because there are no fixed limits, developers will not have to create workarounds to their code when the data scales beyond the file limits.

Azure Data Lake Analytics is the first true clusterless big data analytics service that automatically scales out your queries on-demand letting you focus on the logic of your application and not the hardware or clusters running it. After writing your script/query, users simply set the dial to indicate how much power is needed. Data Lake Analytics will dynamically provision resources to run the code to process any amount of data. When the job completes, it winds down resources automatically making it cost-efficient because you only pay for your job while it is

running. Support for Azure Active Directory lets you manage access and roles simply and integrates with your on-premises identity system.

Case Study: Devicedesk

Devicedesk uses Azure Data Lake to get started on big data in the cloud without having to spend millions of dollars building their own big data clusters

"Giving organizations insights to how their IT investments are being utilized is a challenging problem that we solved with our solution, Insightcentr," says Anthony Stevens, CEO of Devicedesk an Australian start-up. "Azure Data Lake is instrumental because it helps Insightcentr ingest IoT-scale telemetry from PCs in real-time and gives detailed analytics to our customers without us spending millions of dollars building out big data clusters by scratch. We saw Azure Data Lake as the fastest and most scalable way we can get bring our customers these valuable insights to their business."

Azure HDInsight is a managed Hadoop and Spark cluster solution that provides an enterprise-ready Hadoop service in the cloud. Users can deploy and provision an HDInsight Hadoop and Spark cluster in minutes instead of hours or days with the full elastic scalability of Azure.

Hybrid

For enterprises that want to do data warehousing and big data both on-premises and cloud, Microsoft can accommodate with hybrid options. Unlike other purely on-premise or purely cloud implementations, Microsoft uniquely delivers a comprehensive offering that can span both on-premises and cloud. Hybrid Cloud gives you benefits of both the control & flexibility of on-premise and elasticity & redundancy of the cloud. It will also open up a wealth of cloud computing and advanced analytics that are accessible through Azure.

Advanced Analytics and Machine Learning

Microsoft provides a comprehensive set of Advanced Analytics and Machine Learning tools included as part of the Cortana Intelligence Suite in the cloud (Azure Machine Learning, Spark for Azure HDInsight), and R Server which can be deployed on-premises with SQL Server as well as in the cloud with HDInsight.

The integration of advanced analytics into a data warehouse is revolutionary. Today a majority of advanced analytic applications use a primitive approach of moving data from data warehouses into the application tier to derive intelligence. This approach incurs high latency because of data movement, doesn't scale as data volumes grow and burdens the application tier with the task of managing and maintaining analytical models. And deep analytics on real-time transactions are next to impossible without a lot of heavy lifting.

On-premises, SQL Server 2016 with R Services simplifies analytics in the way databases simplified enterprise data management, by moving analytics close to where the data is managed instead of the other way around. It introduces a new paradigm where all joins, aggregations and machine learning are performed securely within the database itself without moving the data out, thereby enabling analytics on real-time transactions with great speed and parallelism. As a result, analytical applications can now be far simpler and need only query the database for analytic results. Updating machine learning models, deploying new models, and monitoring their performance can now be done in the database without recompiling and redeploying applications. Furthermore, the database can serve as a central server for the enterprise's analytical models and multiple intelligent applications can leverage the same models. It is a profound simplification in how mission critical intelligent applications can be built and managed in the enterprise.

In the cloud, Cortana Intelligence Suite provides an end-to-end solution that combines big data stores (data warehousing and big data) with advanced analytics like machine learning, and business intelligence software in a single convenient monthly subscription. But more than a package, Cortana Intelligence Suite embodies the belief that the most impactful data-driven solutions will go beyond advanced analytics, and will include built-in intelligence that allow solutions to see, hear, speak, understand and interpret our needs using natural methods of communication like Vision, Speech, Text, Recommendations and Facial Detection.

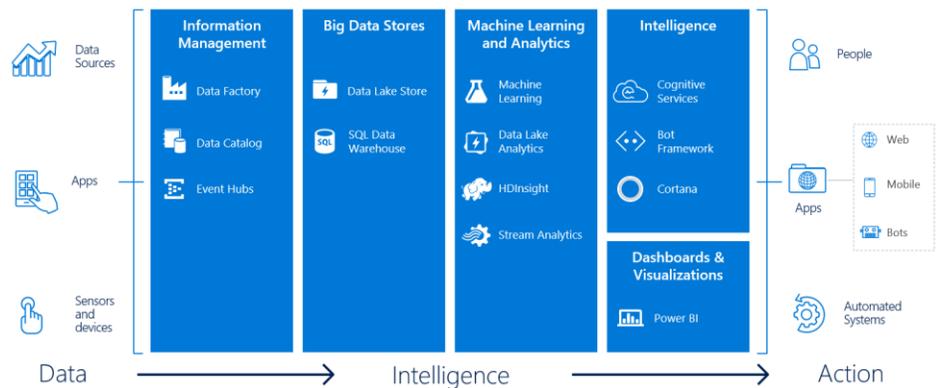


Figure 11: Cortana Intelligence Suite

Conclusion

The opportunities of Big Data and advanced analytics 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:

- Easily manage relational and non-relational data at all volumes and high performance
- Enjoy a consistent experience across on-premises and cloud
- Gain insights from BI and advanced analytics across all your data wherever it resides

For more information

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

Azure HDInsight www.azure.com/hdinsight

Azure Data Lake www.azure.com/datalake

Azure SQL Warehousing <https://azure.microsoft.com/en-us/services/sql-data-warehouse/>

Cortana Intelligence Suite <https://www.microsoft.com/en-us/cloud-platform/cortana-intelligence-suite>

SQL Server 2016 at <https://www.microsoft.com/en-us/cloud-platform/sql-server>

Modern Data Warehouse Solution Page:

http://www.microsoft.com/en-us/server-cloud/solutions/modern-data-warehouse/default.aspx#fbid=IsK_qrqtR35

Analytics Platform System:

<https://www.microsoft.com/en-us/cloud-platform/analytics-platform-system>

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

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