Enterprise Data Warehouses and BI in the Age of Cloud Computing and Big Data

Six illustrated use cases

Microsoft

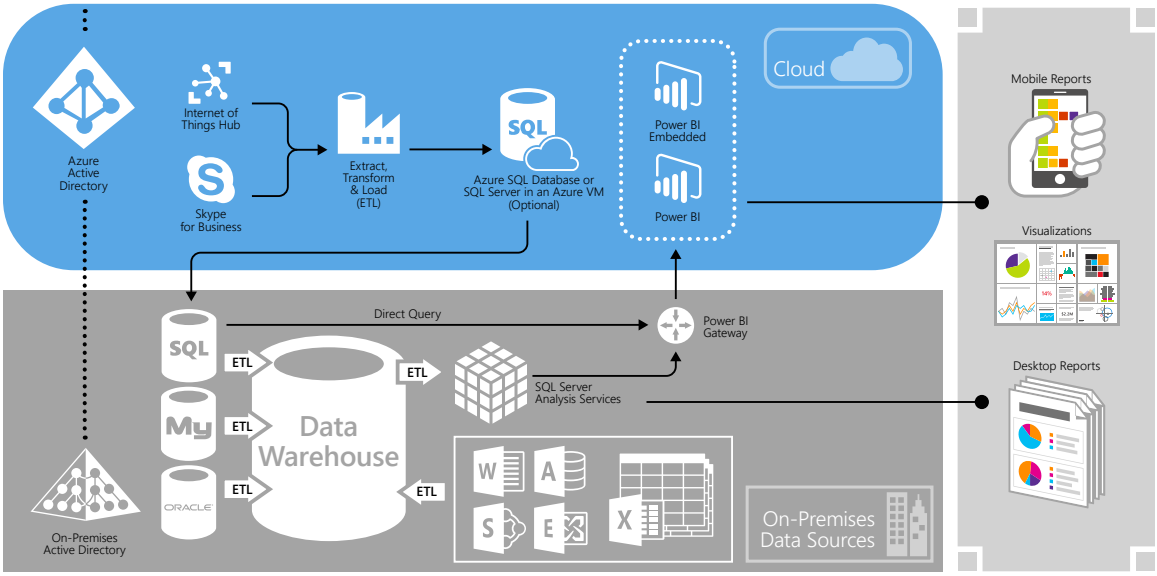
Table of Contents

Introduction

The massive growth of data and devices is having a profound impact on business. Every company in every industry around the world is being challenged to transform into a digital organization. If your organization is already engaged in this process of digital transformation, you understand its potential to improve every aspect of your business. Your ability to unlock value from data can help define your organization's success and uncover new ways to create and capture value and expand to new customer segments.

Never have the opportunities—and the challenges—been greater. Most companies have multiple vertical silos of information. Different apps gather overlapping, sometimes conflicting, data. Unstructured data, which can include Internet of Things (IoT) data, typically is collected without a predefined data model, but needs to be connected to the rest of your data.

Databases are the bedrock for provisioning this data, and data analytics has become the tool by which this data is converted into actionable information. Regardless of where the data is located or how it's stored, database professionals must meet organizational recovery time and recovery point objectives (RTO and RPO, respectively).

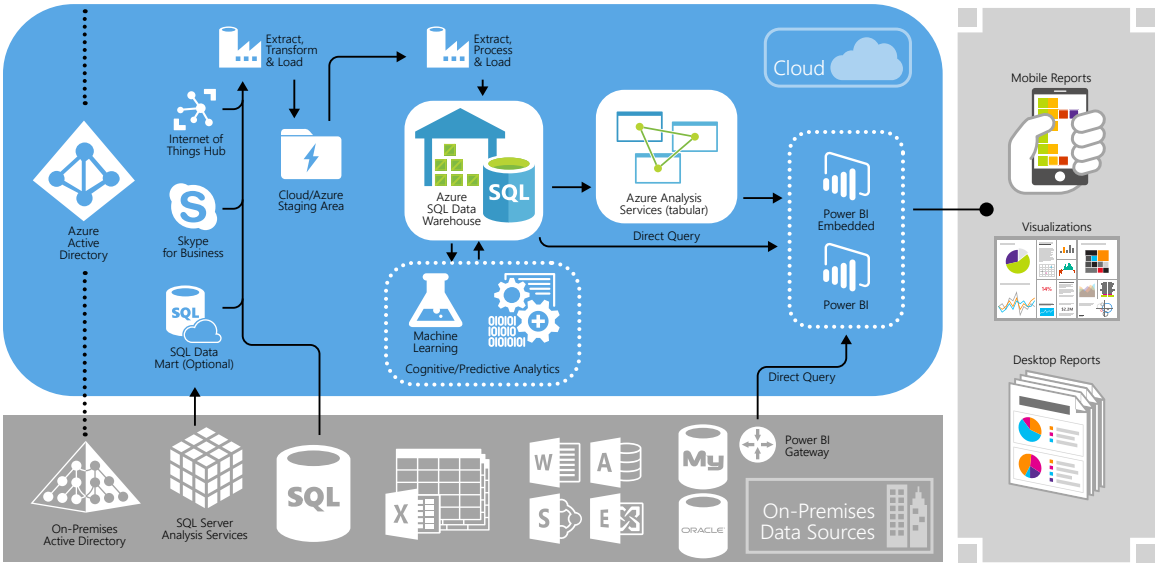


Traditional model. *Companies that rely on on-premises cloud analytics services typically scale up on-premises infrastructure and staff to meet performance metrics. As data volumes grow, however, this model stretches budgets and staff capabilities.*

Organizations that want to make data accessible for business value creation traditionally invest heavily to scale up infrastructure and add staff to help ensure that data is secure, persistent, and always available, and that the databases will perform as required, at all times.

Trying to bring together data sets while delivering timely, accurate information can overwhelm any IT team. Data keeps accumulating, requiring more and more investment in physical infrastructure and new ways to analyze the data. Costs mount based on the need for perpetual uptime and highly trained database and system professionals.

Moving to a Data Management Platform for Analytics



The new approach. *Increasingly, organizations are finding that cloud services, such as those in Microsoft Azure, can help them deploy a data warehouse solution faster. The Microsoft Azure SQL Data Warehouse solution is primed to take advantage of supportive cloud-based analytics services to deliver insight and reporting anywhere.*

Cloud-based data platform services in Microsoft Azure offer a new approach and a competitive edge. Organizations can bring together on-premises data sources with advanced cloud-based technologies to build in-the-cloud enterprise data warehouses, using familiar architectures and data flows. Because the data platform in Azure is built on the same on-premises SQL Server framework that you have used for years, you get consistency across your hybrid environment with common tools for development and management, and a common Transact-SQL (T-SQL) surface area. The benefits are clear:

- Protect sensitive data by keeping it on-premises when necessary.
- As needed, add incremental, supportive cloud technologies that are expensive to build and run on-premises, such as machine learning and streaming analytics.
- Enable a single destination for data processing and analysis.

Arguments against using cloud resources focus on its newness. Organizations grow cautious after promises of new and better information technologies don't work out. Cost savings are often overblown. Skills don't always transfer, which means hiring new people. Reporting tools must be replaced. Budgets must shift from capital to operational expenditures, which can disrupt organizational finance models.

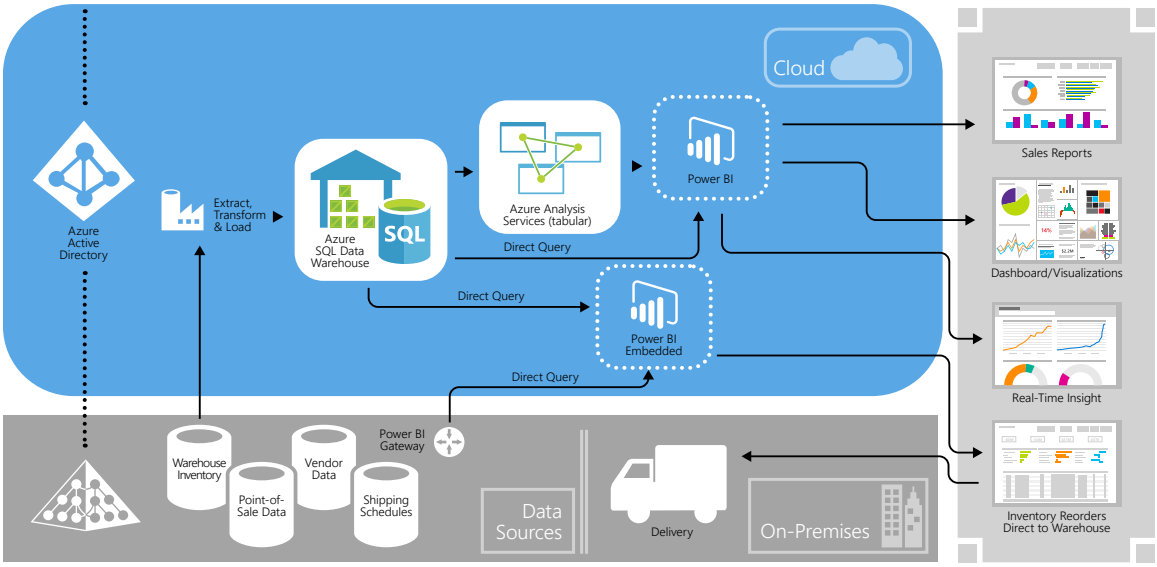
This guide is for the decision maker that asks: What would a datacenter without limits look like? How can we virtually expand our IT staff without adding to its budget? And if we're able to bring together data from disparate islands of information into one repository, are there tools to help with data integration and evaluation? Are there ways to find answers to questions we haven't asked yet?

This guide shows how enterprises are making cloud-based data warehouses and analytics work across key industries. Although the use cases are based on real customer challenges and solutions, company names have been changed.

Data, and access to timely information that describes what is happening within and outside of the organization, has become a critical resource. Data—used well—can make the difference in industry leadership. Data can help organizations:

- Improve visibility and make accurate predictions.
- Get the right products to the right place at the right time.
- Offer customers exactly what they want, when they want it.
- Fix problems proactively.
- Explore new business opportunities.

Retail use case: Inventory analysis



Retail businesses require both cached data models for high performance and direct query models for real-time information that keeps shelves stocked.

A major retailer, Wide World Importers, has long embraced the concept of just-in-time inventory control. On-premises point-of-sale (POS) devices capture transactions as they're scanned at checkout. Additional on-premises data stores contain information about vendors, shipping schedules, and existing warehouse inventory. As the POS data reports the decreasing inventory on the store shelf, the system sends automatic reorder instructions to vendors, who then arrange for deliveries to restock the store.

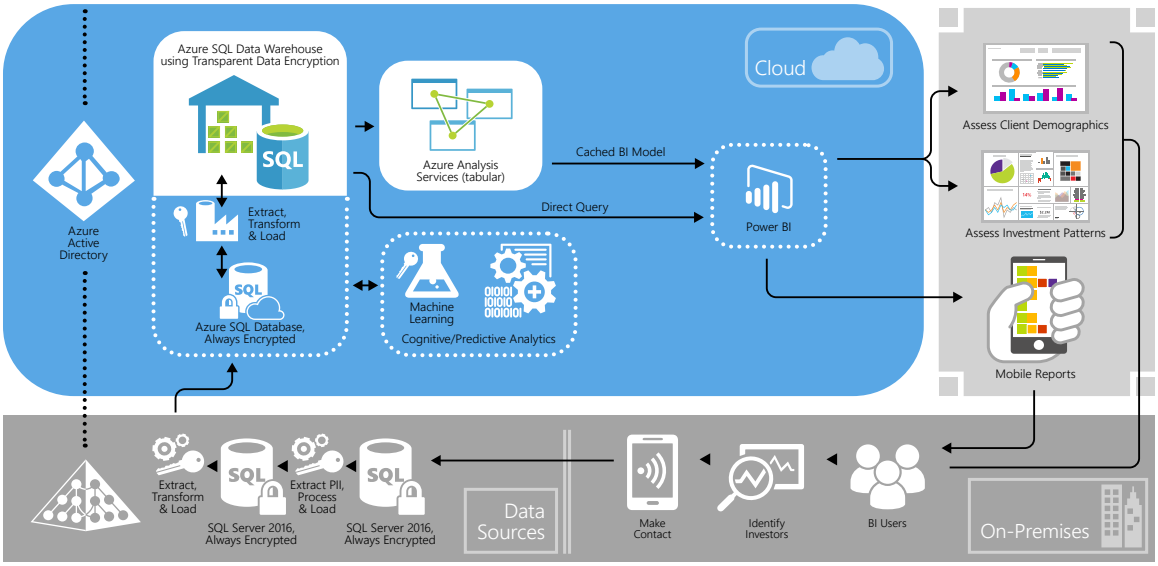
Although this may have worked well in the past, consumer purchasing patterns have changed based on increased use of the Internet and social media. To avoid empty shelves from unexpected bursts of purchase activity, Wide World Importers decided to enhance inventory control to better determine re-order points and deliveries.

Today, the company's solution streams POS and other on-premises data to an Azure SQL Data Warehouse. The data warehouse pushes the data through analysis tools in three separate streams:

- One data stream uses Microsoft Power BI to draw from models cached in Azure Analysis Services that define streamlined data sets, enabling near-instantaneous, highly visual activity and sales reports that work well on a decision maker's mobile device.
- Another data stream uses Embedded Power BI to directly query the POS and other on-premises data sources. Vendors monitor the resulting dashboards in an inventory control portal and can immediately reorder inventory.
- A third stream extracts data directly from the Azure SQL Data Warehouse and pushes it through Power BI to deliver real-time insight to top corporate decision makers.

The data warehouse and supportive services enable Wide World Importers to handle all data collection, processing, and transformation in one location, more economically. By using Azure Analysis Services to create the cached, in-memory models that showcase specific sales activity—including sales activity based on time, location, and ongoing product discounts—the big retailer can improve customer experience and reduce the likelihood of having empty shelves. Wide World Importers can also help control costs by pausing Azure Analysis Services servers when they are not needed.

Financial services use case: Client prospecting



Removing sensitive data on-premises enables a financial services company to use the cloud as a data collection point for deep analysis across industries and regions to identify high-value sales targets.

Contoso Consultants is a global financial planning company that manages billions of dollars in assets. Its large collections of data, which it stores both on premises and in the cloud, includes historical, transactional, and demographic data along with prospecting leads from various CRM apps that service the many offices scattered around the globe. The company is known worldwide for its high-touch sales cycle. This personalized approach helps build customer loyalty, but limits the amount of time financial planners can spend on identifying high-value prospects for special investment opportunities. Contoso is also not well equipped to identify underperforming offices. Contoso wanted to better use its data to support business development. It turned to Azure cloud services to help planners increase target and conversion rates—a key factor in meeting growth goals.

More data is better than too little, especially when you are modeling data to derive business insight. The challenge is integrating decades of data scattered across multiple data sources. Contoso opted to aggregate its data in an Azure SQL Data Warehouse protected by Always Encrypted[1] after first removing any personally identifiable information, which it must do to maintain compliance with the multitude of privacy and security regulations faced by any international organization.

After the anonymized data is loaded into the Azure SQL Data Warehouse, Contoso applies Azure Analysis Services, Azure Machine Learning,[2] and Power BI to coax the volumes of data into meaningful and actionable information, including predictive analytics. The company found that Azure Analysis Services helps identify behaviors among financial planners that could be adjusted for enhanced performance and increased revenues. The company's BI professionals were delighted that the hybrid cloud solution delivered a consistent user experience between on-premises SQL Server Analysis Services and cloud-based Azure Analysis Services. Since their skills transferred, they were able to come up to speed quickly, providing actionable insight to Contoso decision makers right away.

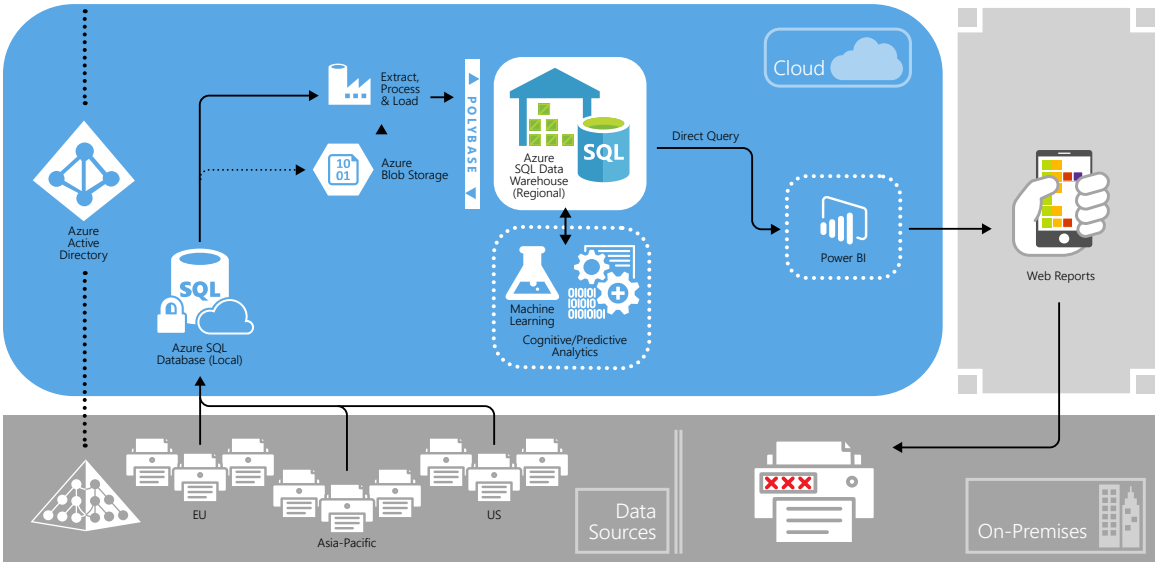
Azure Machine Learning

Machine Learning is a powerful tool that can work its way through terabytes of data, using algorithms developed and refined by BI professionals to return succinct and accurate information. It can help answer questions such as:

- How can I accurately identify the best high-value prospects for the portfolio of special investments?
- How can I better manage the prospect pipeline and get more customer conversions when we have more leads than the financial planners can handle?
- Which districts do I need to focus attention and resources on, and how can I help their financial planners?

[1] See Appendix B: Definitions for more information about Always Encrypted.
[2] See Appendix B: Definitions for more information about Machine Learning.

Manufacturing use case: Shop floor equipment monitoring



One manufacturer uses telemetry and other data to monitor the health of its printer units in customer settings, and to quickly alert customers if any printer ceases to function properly.

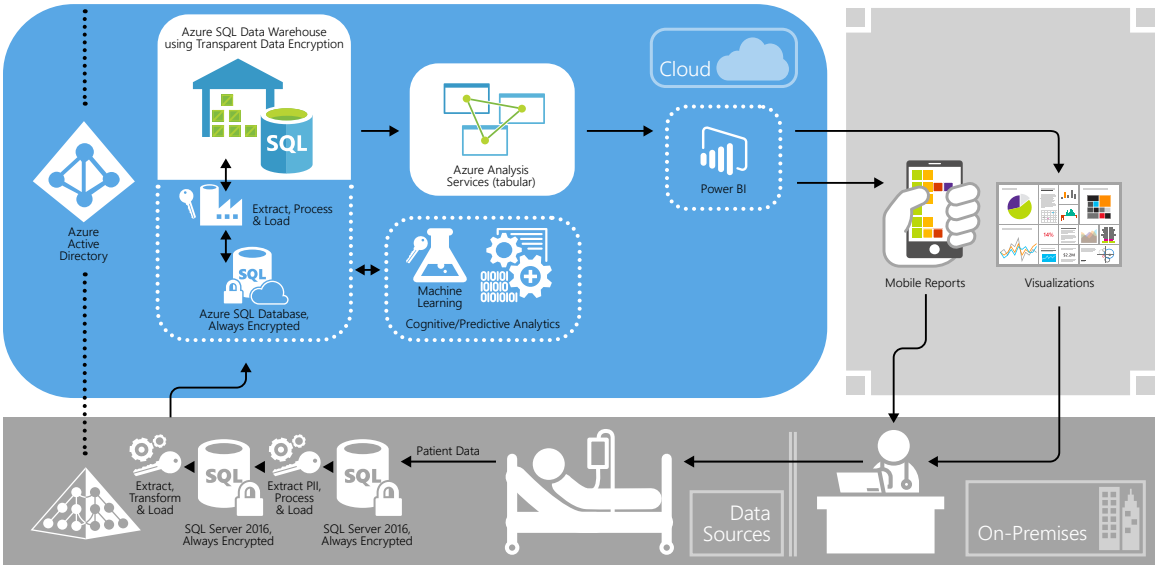
Fabrikam, Inc. helps customers manage both paper and digital communications with smart printers and copiers. Fabrikam collects telemetry data through agents installed on each printer or copier, field gateways, and Windows Communication Foundation (WCF) framework, which was introduced in the .NET 3.0 API, to collect data from devices and store it in compressed, encrypted form in a local Azure SQL Database.

From there, the Azure Data Factory scheme loads the data into Azure Blob storage to be used for other Azure services, while the aggregated results are stored in Azure SQL Data Warehouse. Fabrikam uses regional Azure SQL Data Warehouses as its analytical layer to take advantage of the flexibility and high performance features available with Azure Analysis Services and Azure Machine Learning Studio. It is here that the data is consumed by various apps to gain insight into unit usage and health.

Data scientists can interact with the data in Azure Blob storage and explore it using PolyBase as a language bridge between the Hadoop Distributed File System (HDFS), Azure Blob storage, and Azure SQL Data Warehouse to extract historical trends and other long-term insights. With PolyBase, data scientists can run queries against unstructured and structured data and extract a single result–without moving the data.

Power BI produces reports that help Fabrikam identify which part or parts of a smart copier or printer needs repair or replacement. The SQL Server Customer Advisory Team, the SQL Data Warehouse team, and the Azure Data Factory team have all been involved with this project from day one and have helped Fabrikam unblock issues and expedite the project.

Healthcare use case: Better patient outcomes



Anonymized data goes through multiple rounds of cognitive and predictive analysis to help improve patient outcomes.

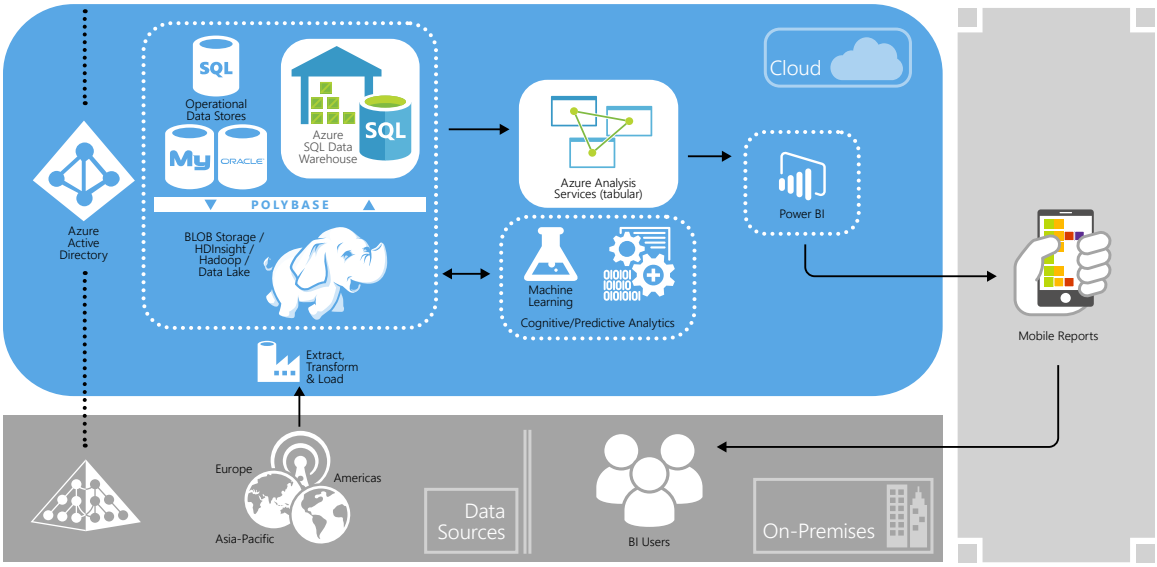
Healthcare organizations must protect patient privacy while striving to produce better patient outcomes. That's one reason Lamna Healthcare adopted Azure SQL Data Warehouse and Azure Machine Learning. In accordance with HIPAA requirements, patient data—including vital statistics and seizure history—is collected via wearables[3] and administered medications are recorded into an on-premises SQL Server 2016 database equipped with Always Encrypted. Personally identifiable information is stripped and then the system transfers anonymized data to a second SQL Server database configured with Always Encrypted, to start the predictive analysis process. Always Encrypted technology ensures that data is encrypted both at rest—while stored on disk—and in motion, as it flows from one repository to another.

The now-anonymous data is pushed to an Azure SQL Database that is also protected by Always Encrypted, where it goes through multiple rounds of cognitive and predictive analysis in the Azure Machine Learning Studio environment to improve understanding of patient conditions and predicted outcomes. Data is then pushed to an Azure SQL Data Warehouse protected by Transparent Data Encryption[4] (TDE), where data is finally ready to be sorted, organized, and presented as information on clinicians' mobile devices. The hospital's BI analysts and motivated care practitioners use reports and visualizations processed in Azure Analysis Services and rendered through Power BI to help improve diagnoses and enhance patient care—all while protecting patient privacy across a hybrid cloud environment.

[3] See Appendix B: Definitions for more information about Wearables.

[4] See Appendix B: Definitions for more information about Transparent Data Encryption.

Telecom use case: Developing new customer offerings



A telecom giant combines business data with data from industry trends and third-party social networking to explore areas for new product opportunities and markets.

Contoso Telecom, an international telecommunications company, generates a lot of data, especially from its cellular network. By harvesting more insight from existing data that has not been fully investigated, Contoso Telecom can explore new product opportunities and markets.

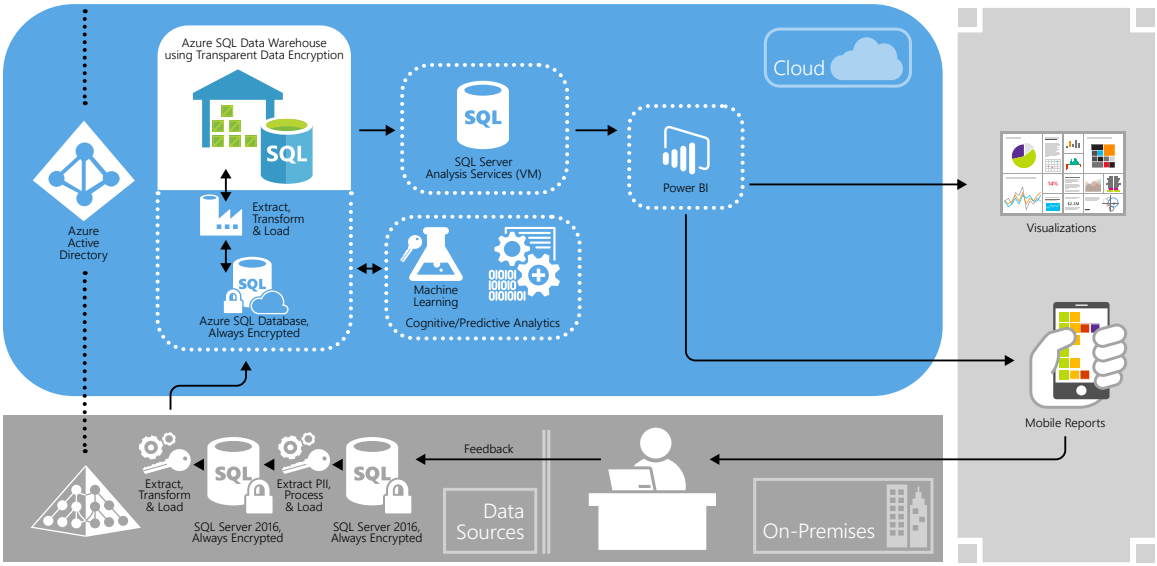
Contoso Telecom already uses data for insight into key metrics, including churn, average revenue per unit, subscriber counts, breakdowns between voice and data usage, volume of devices sold, the number and duration of SIM-only plans, frequency of porting in and porting out of existing phone numbers, use of roaming services, and prepaid versus postpaid plans.

The company then integrates the data using data models that incorporate social networking to derive insight into customer and technology trends. It builds customer personas to represent segments of the customer base, modelling what different personas would post on social sites, what they would spend at certain locations, preferences for plan pricing and spending habits, preference for pay-as-you-go versus data-only plans, likelihood to leave or return, use of international accounts, and preferences for free or flexible contracts.

Contoso Telecom promotes new plans or offers that appeal to the personas across all its channels to influence consumer behavior. Contoso also uses personas for new product development, scheduling network improvements, security and fraud investigations, and branch planning.

Azure SQL Data Warehouse is the telecom's central storage repository for network-generated tabular data, Hadoop serves as blob storage for semi-structured and unstructured data, and PolyBase acts as the language bridge for connecting these dissimilar data stores, enabling cross-product queries without moving the data. By adding Azure Machine Learning and Azure Analysis Services, Contoso Telecom can use data to predict future outcomes with reliability and efficiency. The combined, integrated data has helped it predict the results of multiple scenarios, including whether an increase in subscriber count will result in network overload and a greater number of outages, the frequency with which preventative maintenance needs to be performed, and which new cell phone plans will do well based on subscriber trends.

Government use case: Analyzing personnel data



Agencies can use Azure-based analysis tools on sensitive data by anonymizing it first on-premises.

Government agencies keep files on employees and others who participate in background checks as a condition of employment. Agencies collect sensitive data, including addresses, health and financial history, family connections, and other private details. Properly analyzed data can link an employee's activities to revenue and cost savings and help measure the impact on agency morale and productivity.

Two types of data can be measured: self-reported and ambient. Self-reported data requires voluntary data entry by the employee and might include timesheets, surveys, performance evaluations, and expense reports. Ambient data is information about behavior that can be automatically collected via RFID badges, emails, phone calls, or events added to a department calendar. For compliance reasons, agencies might choose the ambient approach, a hassle-free method of collecting neutral and objective data to monitor employee performance.

To extract accurate information, an agency needs to thoroughly track an employee's chain of action. To maximize the value of this data while protecting it from hackers and unauthorized viewers, agencies can use the security and analytical abilities of the SQL Server tools.

Data is captured and stored on-premises, close to the users. Data is protected in storage and in transit with the Always Encrypted security feature introduced with SQL Server 2016. While still on-premises, personally identifiable information is stripped from copies of the data as it's moved from one SQL Server database to another. The anonymized data is then uploaded to an Azure SQL Database, which is also protected by Always Encrypted. Here the data is analyzed for valuable insights in the Azure Machine Learning Studio environment. Portions of the raw data and new insights are loaded into an Azure SQL Data Warehouse, protected by TDE, where it can be used for additional analysis. Agencies can run analyses with SQL Server Analysis Services in a virtual machine in Azure, which meets all government compliance requirements. With Power BI, business information analysts can generate actionable reports. (At the time of publication, Azure Analysis Services is not available in the Azure Government cloud.)

Conclusion

The use cases in this paper show a variety of ways to use Azure data platform services, but they have aspects in common. Each organization described wants to enhance the consumer experience. Each organization wants to optimize the use of data it has collected. And each wants to control its bottom line. For these reasons, each organization has chosen to take advantage of the consistency between its on-premises SQL Server database and the Azure data platform, which includes instances of Azure SQL Data Warehouse and Azure Analysis Services, along with complementary tools. The consistency makes it easier to manage a hybrid cloud environment in which some critical components remain on-premises. It also enables organizations to add cloud components and services as needed, incrementally.

When moving some or all of your operations from on-premises to a cloud environment, it is critical to invest in tools that will provide both the ability to monitor your cloud operations and the ability to quickly grow or shrink to meet the organization's needs. Keeping data secure as it flows from initial collection to final report is enabled by the combination of Always Encrypted and TDE, neither of which requires changes at the app level. Organizations can control access to data using Active Directory, which stretches from ground to cloud.

Caching data in-memory in an Azure Analysis Services server provides nearly-instantaneous query results, with the level of "real time" limited only by the scheduling of the data refresh cycles. Azure Analysis Services shows all the relationships between tables and fields, complete with user-friendly names (for example, "authorized users," "cancelled sales," "commission charged"), which encourages self-service BI, even among users who are not trained data scientists.

Microsoft BI capabilities deliver an integrated and comprehensive cloud, on-premises, and hybrid BI platform that helps turn complex data into actionable insights. By co-locating data from various sources into an Azure SQL Data Warehouse and creating a single model using Azure Analysis Services, an enterprise can extract real-time answers with Power BI.

In summary, organizations using the Azure data platform are able to access the data they need in a timely fashion. They are able to meet RTO and RPO because the Azure cloud environment offers excellent uptime assurances, and they can benefit from built-in security, encryption, and the proven persistence of SQL Server 2016. They can also be confident about database performance, because Azure services can scale up and scale out on demand.

Appendix A: Further reading

About Azure SQL Data Warehouse

Azure SQL Data Warehouse is a massively parallel processing (MPP) relational database management system (RDBMS) data warehouse based on Microsoft SQL Server. It leverages much of the platform capability of Azure SQL Database, which is a market leader in cloud-deployed RDBMS. Since it is SQL Server-based, it has a very broad level of tool compatibility.

Unlike other cloud-based data warehouse solutions, Azure SQL Data Warehouse separates the compute nodes from the data storage. Compute is deployed using Azure virtual machines; data is stored in Azure Blob storage. This separation allows Azure SQL Data Warehouse to function like a truly elastic data warehouse, able to scale storage capacity and compute capacity within seconds. Azure SQL Data Warehouse can accommodate hundreds of terabytes of data. When all tables are architected as clustered columnstores, which is the default table type, capacity approaches 1 petabyte. For more details about the Azure SQL Data Warehouse, see Azure SQL Data Warehouse Capacity Limits.

Access to Azure SQL Database is offered as a service. Managed database services take care of scalability, backup, and high availability of the database. Microsoft Azure SQL Database includes built-in intelligence that learns app patterns and adapts to maximize performance, reliability, and data protection. Key capabilities include:

- Continuous learning of your unique app patterns, adaptive performance tuning, and automatic improvements to reliability and data protection.
- Scaling as needed, with virtually no app downtime.
- Management and monitoring of multitenant apps with isolation benefits of one-customer-per-database.
- Leverage open source tools such as cheetah, sql-cli, VS Code and Microsoft tools like Visual Studio and SQL Server Management Studio, Azure Management Portal, PowerShell, and REST APIs.
- Data protection with encryption, authentication, limiting user access to the appropriate subset of the data, continuous monitoring and auditing to help detect potential threats and provide a record of critical events in case of a breach.

Azure and the SQL Server big data ecosystem

Forrester has named Microsoft Azure a leader in big data and Hadoop Cloud Solutions noting the following features that contribute to the comprehensive, scalable, and integrated platform:

- **Traditional SQL Server tools:** Azure SQL Data Warehouse is fully integrated with SQL Server Analysis Services, Integration Services, and Reporting Services.
- **Cloud-based tools:** A number of new tools are available in Azure, including Data Factory, Stream Analytics, Machine Learning, and Power BI.
- **HDInsight:** The Azure SQL Data Warehouse MPP engine and architecture are fully integrated with Spark, R, Storm, HBase, and Hadoop.
- **Azure Data Lake:** This Hadoop Distributed File System (HDFS)—compatible distributed analytics service built on Apache Hadoop YARN dynamically scales, for petabytes of massive throughput.
- **PolyBase:** A new feature in SQL Server 2016, PolyBase uses familiar T-SQL language to enable you to run queries on external data and can also push query operations to Hadoop.

More use cases

Find out more about how other innovative companies are working with data warehouses and cloud-based analytics on the Microsoft website at azure.microsoft.com/en-us/case-studies.

Appendix B: Definitions

Always Encrypted: A technology employed by Microsoft and other database companies to encrypt database files at rest and in motion. The feature, designed to protect stored sensitive data such as credit card numbers or national identification numbers (for example, US Social Security numbers), works with Azure SQL Database or SQL Server databases. Always Encrypted allows clients to encrypt sensitive data inside client apps and never reveal the encryption keys to the database engine (Azure SQL Database or SQL Server). As a result, Always Encrypted provides a separation between those who own the data (and can view it) and those who manage the data (but should have no access). By ensuring that on-premises database administrators, cloud database operators, or other high-privileged but unauthorized users cannot access the encrypted data, Always Encrypted enables customers to confidently store sensitive data outside of their direct control. This allows organizations to encrypt data at rest and in use for storage on-premises and in Azure, to enable delegation of on-premises database administration to third parties, or to reduce security clearance requirements for their own database administration staff.

Always Encrypted makes encryption transparent to apps (requiring no changes to the app code). An Always Encrypted-enabled driver installed on the client computer achieves this by automatically encrypting and decrypting sensitive data in the client app. The driver encrypts the data in sensitive columns before passing the data to the database engine, and automatically rewrites queries so that the semantics to the app are preserved. Similarly, the driver transparently decrypts data, stored in encrypted database columns, contained in query results. Always Encrypted is available in SQL Server 2016 and Azure SQL Database. Prior to SQL Server 2016 SP1, Always Encrypted was limited to the Enterprise Edition.

Azure Analysis Services: A cloud-based analytics engine comparable to the engine in Microsoft SQL Server Analysis Services (SSAS); provides enterprise-grade data modeling in the cloud. Runs as a service (Platform as a Service, or PaaS) rather than having to mount a virtual machine with SSAS installed.

Azure Data Factory: A cloud-based data integration service that orchestrates and automates the movement and transformation of data. You can create data integration solutions using the Data Factory service that can ingest data from various data stores, transform and process the data, and publish resulting data to the data stores. Data Factory service allows you to create data pipelines that move and transform data, and then run the pipelines on a specified schedule (hourly, daily, weekly, and so on). It also provides rich visualizations to display the lineage and dependencies between your data pipelines, and monitors the pipelines from a single unified view to easily pinpoint issues and set up monitoring alerts. It is a platform somewhat like SQL Server Integration Services in the cloud to manage the data you have both on-premises and in the cloud. It provides access to on-premises data in SQL Server and cloud data in Azure storage (blob and tables) and Azure SQL Database. Access to on-premises data is provided through a data management gateway that connects to on-premises SQL Server databases.

Cloud: In cloud computing, the word "cloud" is used as a metaphor for the Internet. The phrase cloud computing means "a type of Internet-based computing" in which different services—such as servers, storage, and apps—are delivered to an organization's computers and devices through the Internet.

Machine Learning: A branch of artificial intelligence in which a computer generates rules underlying or based on raw data that has been fed into it. Azure Machine Learning Studio is a GUI-based integrated development environment for constructing and operationalizing Machine Learning workflow in Azure.

On-Premises: On-premises refers to software located within the physical confines of an enterprise as opposed to running remotely on hosted servers or in the cloud.

PolyBase: PolyBase is a technology that enables queries across relational data and non-relational data with a single result. You can use PolyBase in SQL Server 2016 to run queries on external data in Hadoop or Azure Blob storage. Queries are optimized to push computation to Hadoop. You can use PolyBase In Azure SQL Data Warehouse to query external data sitting in other data stores, both cloud and on-premises, including Azure Blob storage and Azure Data Lake Store.

Power BI: A free, self-service business intelligence cloud service that provides non-technical business users with tools for aggregating, analyzing, visualizing, and sharing data.

Power BI Embedded: An Azure service that enables ISVs and app developers to integrate Power BI reports into web or mobile apps to enable viewing outside the organization, without requiring a corporate login or an instance of Power BI installed on the viewing device.

Scalability: 1) The ability of a computer application or product (hardware or software) to continue to function well when it (or its context) is changed in size or volume in order to meet a user need; 2) The ability not only to function well in the rescaled situation, but to actually take full advantage of it.

Transparent Data Encryption (TDE): A technology that encrypts SQL Server, Azure SQL Database, and Azure SQL Data Warehouse data and log files, while they are at rest on the disk. TDE performs real-time I/O encryption and decryption of these files. The encryption uses a database encryption key (DEK), which is stored in the database boot record for availability during recovery. The DEK is a symmetric key secured by using a certificate stored in the master database of the server or an asymmetric key protected by an extensible key management (EKM) module. TDE provides the ability to comply with many laws, regulations, and guidelines established in various industries. This enables software developers to encrypt data by using AES and 3DES encryption algorithms without changing existing apps.

TDE does not provide encryption across communication channels. For more information about how to encrypt data across communication channels, see Enable Encrypted Connections to the Database Engine (SQL Server Configuration Manager).

Wearables: Relating to or noting a computer or advanced electronic device that is incorporated into an accessory worn on the body or an item of clothing: wearable gadgets embedded in fabric; a wearable heart-rate sensor.

Microsoft