Introducing Microsoft Technologies for Data Storage, Movement and Transformation

Supporting Business Intelligence

Summary: The intent of this article is to introduce Microsoft technologies for data storage, movement, and transformation. These technologies support business intelligence by enabling your organization to store and access heterogeneous, quality data for analytics and reporting, in the cloud or on premise.

For an overview of the Microsoft business intelligence technologies, see Introducing Microsoft BI Reporting and Analysis Tools.

Writer: Carla Sabotta

Technical Reviewer: Douglas Laudenschlager, Sreedhar Pelluru

Copyright

Contents

Introduction

The intent of this article is to introduce Microsoft technologies for data storage, movement, and transformation. These technologies support business intelligence by enabling your organization to store and access heterogeneous, quality data for analytics and reporting, in the cloud or on premise.

For a brief overview of the Microsoft analytics and reporting tools, see [Introducing Microsoft BI Reporting and Analysis Tools](#).

Relational Data Storage

SQL Server Database Engine

The Database Engine is the core SQL Server service for storing, processing, and securing data. The Database Engine provides controlled access and rapid transaction processing to meet the requirements of the most demanding applications within your enterprise.

You use the Database Engine to create relational databases for online transaction processing or online analytical processing data.

You can also set up a hybrid solution by storing SQL Server database files as Windows Azure Blobs. This allows you to create a database in SQL Server running on-premises or in a virtual machine in Windows Azure with a dedicated storage location for your data in Windows Azure Blob Storage. This feature provides several benefits for data virtualization, data movement, security and availability, and low costs and maintenance for high-availability and elastic scaling.

By running SQL Server in Azure Virtual Machines, you can create a server in the cloud. With Azure as a hosting environment, you can reduce the total cost of ownership of deployment, management, and maintenance of your enterprise database applications. Administrators and developers can still use the same development and administration tools that are available on-premises. In addition, you can provision or deprovision virtual machines to achieve elasticity on demand and connect them to your on-premises infrastructure by using [Azure Virtual Network](#).

More Information

[SQL Server Database Engine](#)

[What's New in Database Engine](#)

[SQL Server Data Files in Windows Azure](#)

[In-Memory OLTP – Common Workload Patterns and Migration Considerations](#)

[SQL Server 2014 and Windows Azure Blob Storage Service: Better Together](#)

[SQL Server in Azure Virtual Machines](#)

[Provisioning a SQL Server Virtual Machine on Azure](#)

[Virtual Machines Pricing](#)

Azure SQL Database

SQL Database is a relational database service in the cloud based on the Microsoft SQL Server engine, which supports existing SQL Server tools, libraries and APIs.

You can scale up or scale out to thousands of databases, and protect your data with auditing, restore, and geo-replication. SQL Database provides you with the ability to restore your data to any point-in-time from any transaction up to 35 days. You can also replicate your data to an Azure region you choose and implement a geographic disaster recovery policy that meets the needs of your business.

More Information
[SQL Database Overview](#)

[Create your first Azure SQL Database](#)

[Overview: management tools for SQL Database](#)

[Connecting to SQL Database: Links, Best Practices and Design Guidelines](#)

Azure SQL Data Warehouse

SQL Data Warehouse is a distributed database in the Azure Cloud that can process up to petabyte volumes of relational and non-relational data. Query compute is scaled independently. You can dynamically deploy, grow, shrink, and pause compute.

This distributed database is based on the SQL Server Database Engine, with stored procedures, user-defined functions, table portioning, indexes, and collations. And with Polybase, you can also query non-relational data in Azure blob storage and Hadoop's File System (HDFS).

Massive parallel processing (MPP) architecture and the columnstore index technology enable breakthrough performance for SQL Data Warehouse.

More Information
[What is SQL Data Warehouse?](#)

[Get started: provision a SQL Data Warehouse](#)

[Get Started: Connect to SQL Data Warehouse](#)

Non-Relational Data Storage

DocumentDB

Microsoft Azure DocumentDB is a fully-managed, true schema-free NoSQL document database service. DocumentDB delivers consistently fast reads and writes, schema flexibility, and the ability to easily scale a database up and down on demand.

DocumentDB natively supports JSON documents. By default, it automatically indexes all the documents in the database and does not expect or require any schema or creation of secondary indices. DocumentDB enables complex ad hoc queries using a SQL language, supports well defined consistency levels, and offers JavaScript language integrated, multi-document transaction processing using the familiar programming model of stored procedures, triggers, and UDFs.

Application scenarios may include user data for interactive web and mobile applications as well as storage, retrieval, and processing of application JSON data. A database can store any number of JSON documents, as DocumentDB is well suited for applications that run at scale on the internet.

More Information
Introduction to Microsoft Azure DocumentDB

DocumentDB Learning Path

DocumentDB Pricing

Azure Storage

Azure Storage delivers the scalable, durable, and highly available storage that large-scale applications need, and provides a storage foundation for Azure Virtual Machines.

The scalability enables you to store and process a wide range of data sizes, from hundreds of terabytes for big data scenarios to small amounts of data for a small business web site. Wherever your needs fall, you pay only for the data you're storing. And, the storage is elastic. You can design applications for a large global audience, and scale those applications as needed - both in terms of the amount of data stored and the number of requests made against it.

An Azure storage account gives you access to Azure storage services. With a standard account, you have access to the following.

- **Blob storage**. Storage for large amounts of unstructured data.
- **Table storage**. A NoSQL database, which is an alternative to traditional relational databases.
- **Queue storage**. A messaging solution for asynchronous communication between application components, whether they're running in the cloud, on a computer or on a mobile device.

- **File Storage**. Cloud-based file shares that enable you to migrate legacy applications to Azure quickly.

The following are additional features of Azure Storage.

- Automatically load-balances your data based on traffic. As demand increases, appropriate resources are automatically allocated.
- Accessible anywhere in the world, from applications running in the cloud, on a computer or on a mobile device.
- Supports a variety of operating systems and programming languages. Data sources are also exposed via simple REST APIs.

To ensure durability and high availability, your Azure storage account is always replicated.
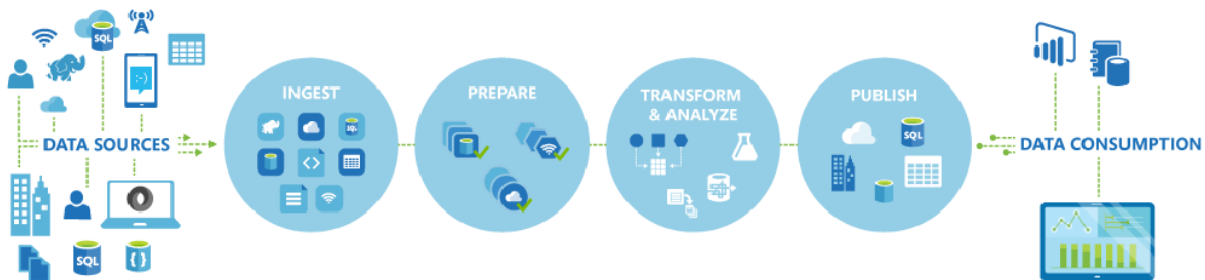
More Information
[Introduction to Microsoft Azure Storage](#)

[Get started with Azure Storage in five minutes](#)

[Azure Storage Pricing](#)

Data Movement and Transformation

Azure Data Factory

Data Factory is a cloud-based data integration service that orchestrates and automates the movement and transformation of data. Just like a manufacturing factory that runs equipment to take raw materials and transform them into finished goods, Data Factory orchestrates existing services that collect raw data and transform it into ready-to-use information.

Data Factory works across on-premises and cloud data sources and SaaS to ingest, prepare, transform, analyze, and publish your data. Use Data Factory to compose services into managed data flow pipelines to transform your data using services like [Azure HDInsight (Hadoop)](#) and [Azure Batch](#) for your big data computing needs, and with [Azure Machine Learning](#) to operationalize your analytics solutions. Go

beyond just a tabular monitoring view, and use the rich visualizations of Data Factory to quickly display the lineage and dependencies between your data pipelines. Monitor all of your data flow pipelines from a single unified view to easily pinpoint issues and setup monitoring alerts.

The Data Factory service provides reliable and complete view of your storage, processing, and data movement services. It helps you quickly assess end-to-end data pipeline health, pinpoint issues, and take corrective action if needed. You can also visually track data lineage and the relationships between your data across any of your sources, and see a full historical accounting of job execution, system health, and dependencies from a single monitoring dashboard.

Using the Azure Preview Portal, you can view your data factory as a diagram, view activities in a pipeline, view input and output datasets, and more.

Why and When to Use Azure Data Factory
See Customer Case Studies to learn directly from our customers how and why they are using Data Factory.

More Information
Introduction to Azure Data Factory Service

Build your first pipeline using Azure Data Factory

Data Movement Activities

Data Transformation Activities


Microsoft Power Query for Excel
Microsoft Power Query for Excel enables you to combine and refine data, as well as discover data across a wide variety of sources including relational, structured and semi-structured, OData, Web, Hadoop, Azure Marketplace, and more. Power Query provides data transformation capabilities such as merging data from multiple sources, cleansing data, and removing and splitting columns.

More Information
Introduction to Microsoft Power Query for Excel

Shape data (Power Query)

9

Enterprise Information Management with On-Premise Data
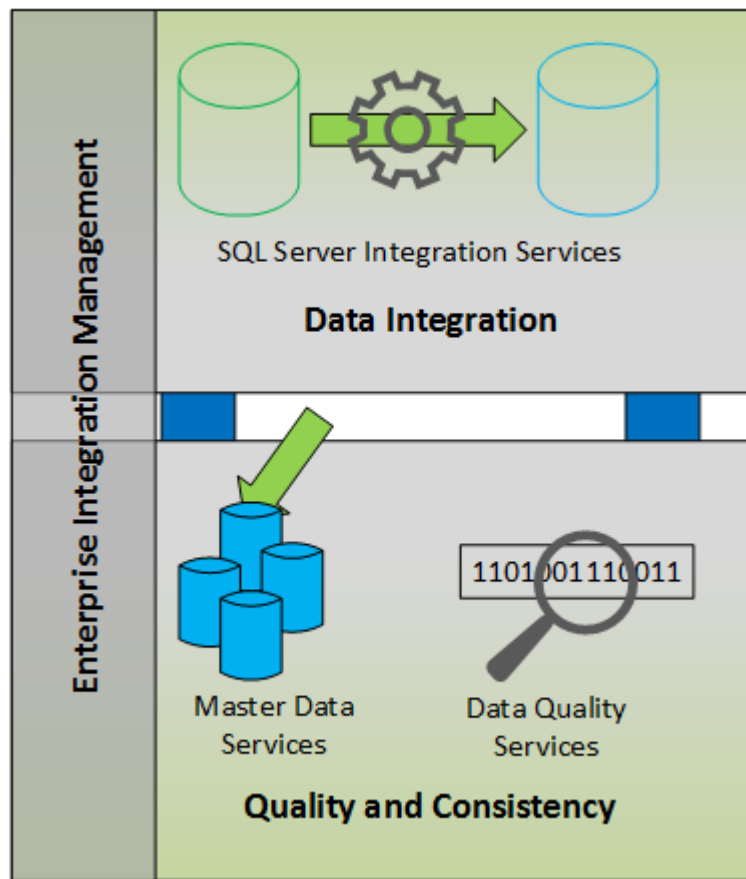
Managing information in an enterprise typically involves integrating data from heterogeneous sources across and outside the enterprise and beyond, cleansing the data to make the data more suitable for its intended use such as data analysis, and then storing a master copy of the data in a centralized location.

SQL Server 2014 and 2016 provide all the components needed for an effective Enterprise Information Management (EIM) solution.

- SQL Server Integration Services
- SQL Server Data Quality Services
- SQL Server Master Data Services

Data Quality Services functionality is built into a component of SQL Server Integration Services (SSIS) and into features of Master Data Services (MDS).

SQL Server Integration Services

SQL Server Integration Services (SSIS) is a platform for building enterprise-level data integration and data transformations solutions. You use SSIS packages to solve complex business problems by copying or

downloading files, sending e-mail messages in response to events, updating data warehouses, cleaning and mining data, and managing SQL Server objects and data.

With SSIS, you can extract and transform data from a wide variety of sources such as XML data files, flat files, and relational data sources, and then load the data into one or more destinations.

Adding a Data Streaming Destination to an SSIS package enables you to consume the output of the package as a tabular result set. You can publish the SSIS package as a SQL view in a SQL Server database and then query the view to return the tabular result set. You can also expose the output of the SSIS package as an OData feed by using the Power BI for Office 365 Admin Center. A data steward can consume the feed from SSIS package by using the Power Query.

In SSIS 2016, there is support for AlwaysOn Availability Groups that is a high-availability and disaster-recovery solution. To provide high-availability for the SSISDB database to which you can deploy your SSIS projects, you can add the SSISDB database to an AlwaysOn Availability Group. When a failover occurs, one of the secondary nodes automatically becomes the new primary node.

More Information
SQL Server Integration Services

SSIS Tutorial: Creating a Simple ETL Package

Enterprise Information Management using SSIS, MDS, and DQS 2014 Together [Tutorial]

Scale-Out SQL Server Integration Services 2014 Environment: A Sample Pattern

Data Streaming Destination

Power BI for Office 365 Admin Center Help

SQL Server Master Data Services

Master Data Services (MDS) enables you to discover and define non-transactional data, with the goal of compiling a maintainable master data set. Business workflows can be applied through the use of business rules. The result is reliable, centralized data that can be analyzed, resulting in better business decisions.

You can define multiple versions of your data. An uncommitted version can be unlocked for updates and then locked while you validate your data. Once the data is validated, you can commit the version. Committed versions of the data form an auditable record of changes.

Transaction logging records changes to data records and attribute values. The date, time, and user who took the action, along with other details, are recorded.

MDS security ensure that users have access to the specific master data necessary to do their jobs, and to prevent them from accessing data that should not be available to them. Security is based on local or Active Directory domain users and groups, and allows you to use a granular level of detail when determining the data a user can access.

With the MDS Add-in for Excel, you can load filtered data from MDS into Excel, where you can work with it just as you would any other data. When you are done, you can publish the data back to MDS, where it is centrally stored. With subscription views, you can make the data available as SQL Server views that can be used by any tool that can use standard SQL Server views.
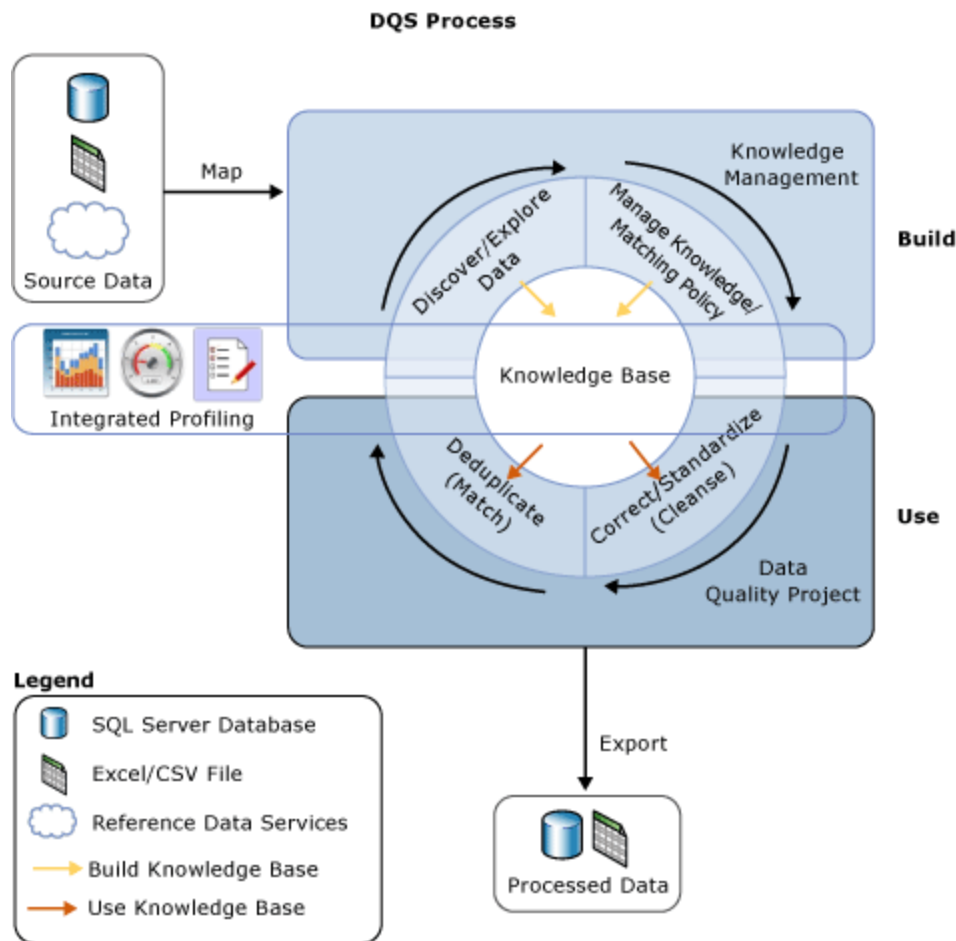
More Information

[Master Data Services](#)

[Master Data Services Blog](#)

SQL Server Data Quality Services

SQL Server Data Quality Services (DQS) enables you to make data more reliable, accessible, and reusable. DQS can improve the completeness, accuracy, conformity, and consistency of your data, and so improve the information value of the data and make the data more suitable for its intended use such as data analysis, reporting, data mining, and warehousing.

Incorrect data can result from user entry errors, corruption in transmission or storage, mismatched data dictionary definitions, and other data quality and process issues. Aggregating data from different sources that use different data standards can result in inconsistent data, as can applying an arbitrary rule or overwriting historical data.

DQS enables you to perform data cleansing by using cloud-based reference data services provided by reference data providers. DQS also provides you with profiling that is integrated into its data-quality tasks, enabling you to analyze the integrity of your data.

12

DQS Process

More Information
Introduction to Data Quality Services.

Moving Data from On Premise to Cloud

The following articles provide guidelines and strategies for moving data from on-premise sources to
Azure.

- Load data into SQL Data Warehouse (See the **Initial Loading into SQL Data Warehouse from SQL
  Server** section.)
- Move data from an on-premise SQL server to SQL Azure with Azure Data Factory
- Azure SQL Data Warehouse loading patterns and strategies

Integrating Technologies

Microsoft BI technologies integrate with a variety of Microsoft technologies for data movement,
processing, and storage.

13

Azure Data Factory enables you to build workflows (pipelines) to process data from on-premises, cloud-based and internet data sources. The data you produce can be consumed by BI tools such as Power Query in Excel and other applications.

From Power BI Designer and Power Query, you can connect to Azure Blob storage, Azure Table storage, and Azure SQL Database. You can also connect to SQL Database from Power BI Preview.

BI technologies that integrate with SQL Data Warehouse include Power BI, Azure Machine Learning and Azure Stream Analytics.

From SQL Server Integration Services (SSIS), you can connect to SQL Server, Azure SQL Database, and Azure Storage as well as Azure Data Factory and Data Quality Services to extract, cleanse, transform, and load your data on premises or in the cloud. You can also connect to Excel, Oracle, Teradata, HDInsight and HDFS, and OData sources.

More Information: Integrating with BI technologies
Connect to Microsoft Azure Table Storage (Power Query)

Connect to Microsoft Azure Blob Storage (Power Query)

Connect to a Microsoft Azure SQL Database (Power Query)

Connect to a SQL Server database (Power Query)

Data Sources in Power BI Designer

Integrate Services with SQL Data Warehouse

Conclusion

Microsoft offers a variety enterprise-ready technologies for data storage, movement, and transformation, on-premise and in the cloud. These technologies support business intelligence by enabling your organization to access heterogeneous, quality data for analytics and reporting.

For more information:

- Introducing Microsoft BI Reporting and Analysis Tools
- Using Power BI in a Hybrid Environment

Did this paper help you? Please give us your feedback.

Send feedback.