DATA-CORE SYSTEMS

The End of ETL

A White Paper on PolyBase

Data-Core Systems Inc.

www.datacoreautomation.com

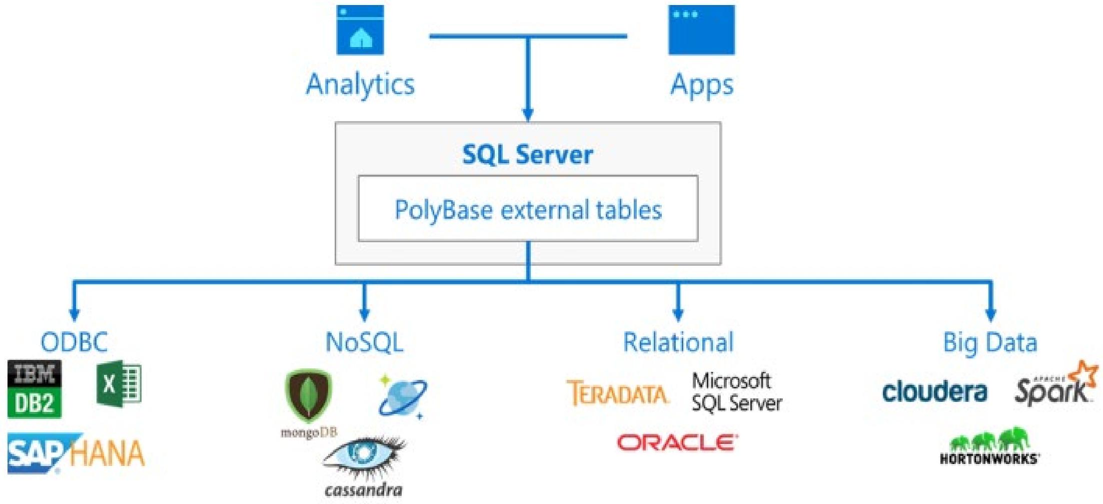
Introduction

PolyBase is a fantastic piece of technology that allows users to near seamlessly tie relational and non-relational data together. It has been around since 2016, but Microsoft has greatly enhanced the capabilities of PolyBase in 2019 with the ability to connect to more data sources. Instead of only flat files, PolyBase can now bring in data residing on other relational and non-relational data sources such as other SQL Servers, Oracle, Teradata or MongoDB.

PolyBase is also a key feature in Azure SQL Data Warehouse to bring in data from flat files sitting on an HDFS Cluster. It treats these sources as external tables which can be queried through T-SQL just like any local table stored in SQL database.

Enhanced PolyBase — Query over any type of data

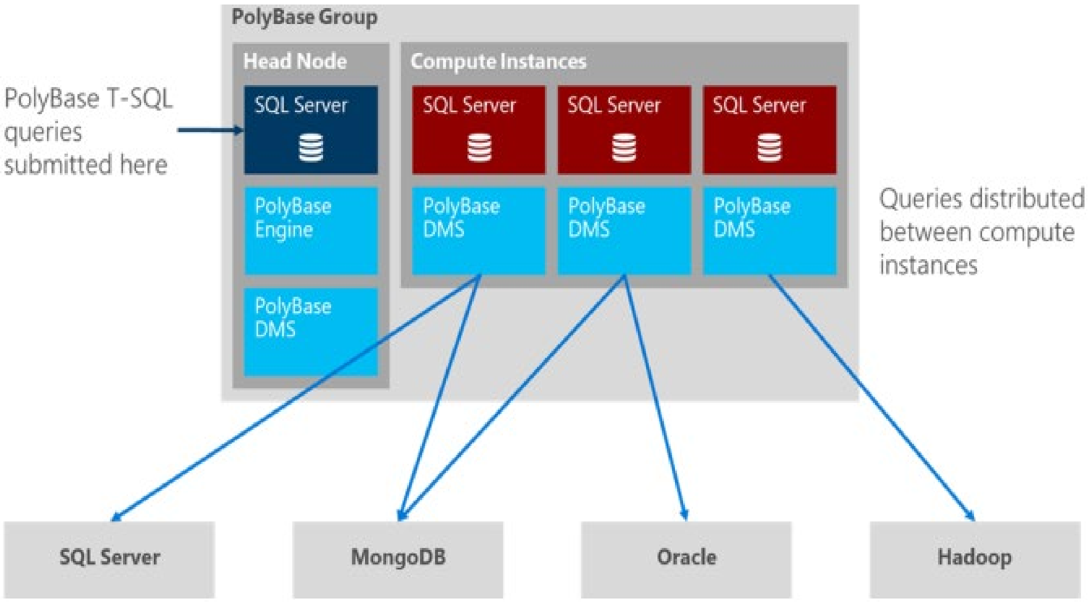
First added to the SQL Server database engine in SQL Server 2016, PolyBase allowed customers to query big data stored in HDFS-compatible Hadoop distributions and file systems such as HortonWorks, Cloudera, and Azure Blob Storage from Transact-SQL by defining an external table to represent HDFS data in SQL Server. Users can write Transact-SQL queries that reference the external table as if it were a normal SQL Server table; when the query is executed, data from the external table is retrieved and displayed to the user. SQL Server 2019 extends capabilities of PolyBase with new connectors and can create external tables that link to a variety of data stores, including SQL Server, Oracle, Teradata, MongoDB, or any data source with an ODBC driver.



Once external tables are created in the SQL Server, it can use Active Directory to control access to data sources, granting access to external tables to Active Directory users and groups. PolyBase already optimizes performance by using push-down computation—operations including projections, predicates, aggregates, limit, and homogeneous joins are all pushed to the source system, and the results of these operations are returned to SQL Server—improving performance by reducing network traffic. In SQL Server 2019 Big Data Clusters, the SQL Server engine has gained the ability to read HDFS files natively, and by using the SQL Server instances on the HDFS data nodes to filter and aggregate data locally.

Data Ingestion Technique

The performance and capacity of PolyBase increases proportionally with scale-out of SQL Server instances; many SQL Server instances can be added to a PolyBase group, under the control of a group head node. The head node distributes the workload across the PolyBase group's compute instances; this enables parallel ingestion and processing of external data. It's like sipping water from a glass with multiple straws as opposed to a single straw. This results in faster consumption.



One of the new features in SQL Server 2019 is the SQL Server Big Data Clusters, which includes an improved version of PolyBase. It treats data sources as external tables, allowing querying through T-SQL, similar as any local table stored in the SQL database. Fetch in data residing on other relational or non-relational data sources such as other SQL Servers, Oracle, Hadoop, Azure Data Lake, Azure Blob Storages or MongoDB.

With all the features available, it allows data scientists to continue using big data ecosystem tools while utilizing easy, real-time access to the high-value data in the SQL Server because it is all part of one integrated, complete system.

SQL Server 2019 big data clusters are a compelling new way to utilize the SQL Server to bring high-value relational data and high-volume big data together on a unified, scalable data platform. Enterprises can leverage the power of PolyBase to virtualize their data stores, create data lakes, and create scalable data marts in a unified, secure environment without needing to implement slow, costly ETL pipelines. This makes data-driven applications and analysis more responsive and productive.

Hybrid transaction/analytical processing (HTAP) is an emerging application architecture that "breaks the wall" between transaction processing and analytics. It enables more informed and real time decision making.

Recent advances in research, hardware, OLTP and OLAP capabilities, in-memory technologies, scalable transactional management, and products enable transactional processing and analytics, or HTAP, to operate on the same database.

Industry landscape and trends

Data integration has been accomplished in the past by using tools like SSIS, PowerShell, BCP and others to copy data from one data source to another. Data virtualization allows the data to be queried at its original source. This means that any query against such data becomes a real-time query while at the same time, it avoids the redundancy of storing data in two places.

Recognizing that different storage technologies are more appropriate for different types of data, a modern enterprise is likely to have data stored in a mixture of relational and non-relational data stores often from several different vendors. A challenge for developers, data scientists, and business analysts is that to extract business value from this data, they typically need to combine data from disparate sources. They typically do this by bringing all the relevant data from the source systems together on a single platform. Once the developers have cleaned and modified the data, they can load the data into their database or warehouse keeping it a much thinner approach.

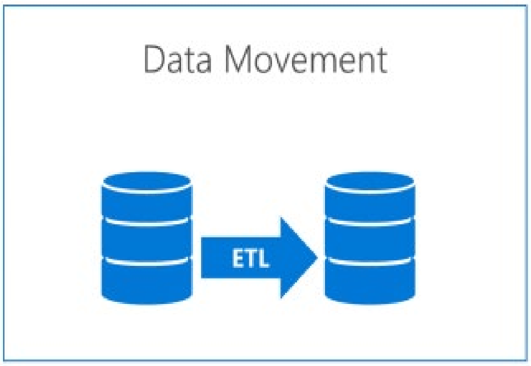
Traditional ETL

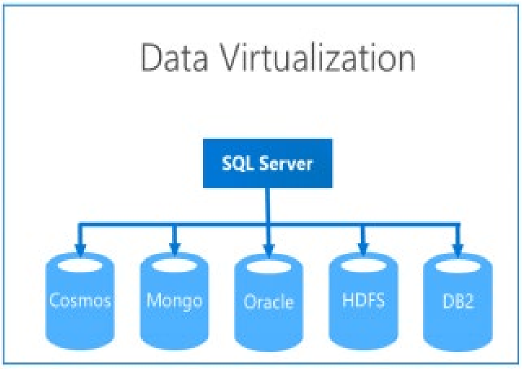
In traditional business intelligence systems, copies of data are created and loaded into a reporting platform with extract-transform-load (ETL) processes; reporting and analysis is carried out on the copies. Whilst enabling enterprises to extract business value from their data, ETL processes have several common issues:

- Expensive to develop, maintain, and support
- Slow ETL processes introduce an inherent delay
- Must be secured
- Require huge storage

Data Virtualization

An alternative to ETL is data virtualization. Data virtualization integrates data from disparate sources, locations and formats, without replicating or moving the data, to create a single "virtual" data layer that delivers unified data services to support multiple applications and users. The virtual data layer, sometimes referred to as a data hub or data lake, allows users to query data from many sources through a consistent interface. Users' access to sensitive data sets can be controlled from a single location, and the delays inherent to ETL need not apply; data sets can be up to date.





Benefits

- Directly query Hadoop or any data source from SQL Server Management Studio using T-SQL queries and the same old constructs developers have been using.
- No knowledge of Hadoop or MapReduce required
- No additional software is needed in the users Hadoop or Azure environment
- Optimal performance using distributed query processing
- Provides seamless integration with SQL BI tools and Analytical tools like Power BI.
- Organizations can reduce time to insight

Conclusion

SQL Server PolyBase continues to evolve in SQL Server 2019; new features and new technology make SQL Server an even more compelling tool for the enterprise. Developers, analysts, data scientists, and database administrators can be even more productive than ever before with new features across the database engine. With a greater range of supported platforms and usage scenarios, including SQL Server containers on Kubernetes and Service Fabric, organizations have even more flexibility when designing their data ecosystems.

Data is at the core of what we do.

Our world is being re-imagined through Analytics, Artificial Intelligence and Automation. Data-Core Systems is a digital transformation solution provider helping businesses reshape their future. We are a proven partner with a passion for client satisfaction, combining technology innovation, business process expertise and a global, collaborative workforce.

DATA-CORE SYSTEMS

Data-Core Systems Inc.

1500 John F. Kennedy
Blvd. Suite 624
Philadelphia, PA 19102

Tel: 215 243 1990

Toll Free: 877 327 4838

Fax: 215 243 1978