Microsoft PolyBase

Big Data complexity

Today, organizations face unprecedented complexities in data warehouse management due to the explosion in volume, variety, and velocity of data—both inside and outside the enterprise. Of the anticipated tenfold growth in data over the next five years, more than 85 percent is expected to come from nonrelational data, including sources such as social media, clickstreams, cameras, sensors, scanners, RFID, eGov, weather, traffic, and more. The latest trend for organizations is to use Hadoop as a data lake where they throw all of their data, structured and unstructured, and then figure out how to perform analytics on top of that data later. As part of a company's modern data warehouse, IT and business units need to find a way for Hadoop to coexist with their existing data warehouse investments.

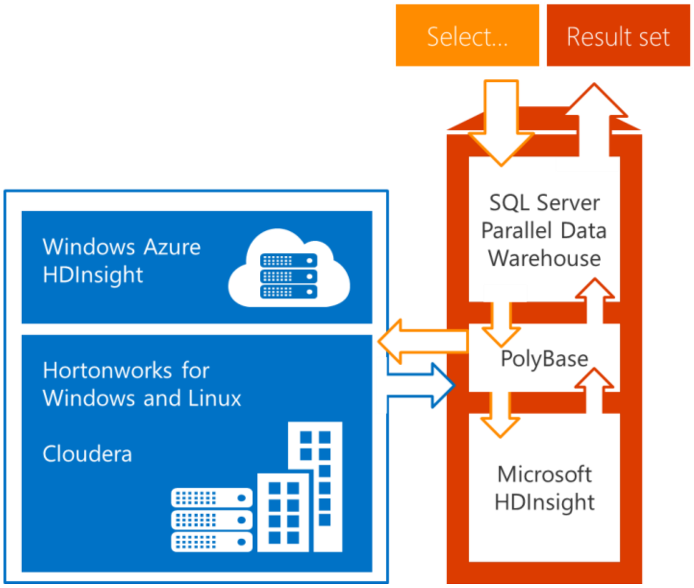
To get the most out of the volume and variety of your data lake information, take advantage of the **Microsoft Analytics Platform System (APS)**—a turnkey appliance for the modern data warehouse. APS is the industry's only solution that combines a best-in-class relational database management system, in-memory technologies, Azure cloud integration, and enterprise-ready Hadoop into a single appliance. APS includes PolyBase, a fundamental breakthrough in data processing that enables seamless querying of data that is stored in Hadoop and the data in your SQL Server Parallel Data Warehouse by using Transact-SQL (T-SQL)—without the requirement of additional manual processes, skills, or training in Hadoop.

Integrated query across SQL Server Parallel Data Warehouse and Hadoop

Normally, IT is burdened with the task of prepopulating the data warehouse with Hadoop data, or users need to undergo extensive training in MapReduce to analyze both relational and nonrelational data in a single view. PolyBase simplifies data analysis by giving users a way to query Hadoop data using standard T-SQL—without the use of MapReduce.

PolyBase is available only within the Microsoft Analytics Platform System, a no-compromise, all-in one modern data warehouse that features SQL Server Parallel Data Warehouse alongside HDInsight, the 100-percent Apache Hadoop distribution from Microsoft based on the Hortonworks Data Platform (HDP) for Windows, for seamless operation.

PolyBase can query Hadoop clusters in-place, eliminating the need to extract, transform, and load (ETL) data into the relational data warehouse. PolyBase works with HDInsight in the appliance along with Microsoft Azure HDInisght for hybrid cloud solutions. In addition, PolyBase also supports Hadoop distributions from Hortonworks and Cloudera within your organization.



Solution advantages

PolyBase makes it faster and easier to extend to new data sources and easily integrate data types for analysis and insight.

Simplifying the use of Hadoop data

PolyBase is one of the most exciting technologies to emerge in recent years because it unifies data in SQL Server PDW with Hadoop data at the query level.

- **Running high-performance queries:** PolyBase accepts a standard T-SQL query that joins tables from SQL Server PDW with optimized access to external table sources in a Hadoop cluster. It then seamlessly returns the results to the user.
- **Archiving data warehouse data to Hadoop:** PolyBase makes it easy to off-load less frequently accessed SQL Server PDW data to your Hadoop data lake architecture. You can then use PolyBase to query the archived data without having to load it back into SQL Server PDW.
- **Exporting relational data to Hadoop:** PolyBase can take the results of complex data warehouse queries and easily copy them to your Hadoop cluster for further analysis with other tools.
- **Importing Hadoop data to PDW:** PolyBase enables easier import of data between Hadoop with PDW using the simple CREATE TABLE AS SELECT T-SQL statement. There is no need to learn other tools like SQOOP.

PolyBase processes Hadoop data in-place, avoiding the need for costly ETL processes.

Big Data insights for anyone

APS is the only data warehouse solution that combines in-memory technologies and Hadoop with native end-to-end Microsoft BI integration through PolyBase, allowing users to create new insights themselves with tools they already know. This gives your Hadoop data ubiquitous connectivity across the entire SQL Server ecosystem.

Enhanced performance for Hadoop Queries

As a leading-edge innovation, PolyBase enhances the performance of querying Hadoop with APS.

- **Push-down computation:** PolyBase extends the SQL Server Parallel Data Warehouse engine with push-down computation, which parallelizes and distributes complex Hadoop query execution plans for better performance.
- **HDInsight optimized for APS:** Microsoft and Hortonworks have teamed together to optimize HDInsight integration on APS. PolyBase takes advantage of the high-speed connections between PDW and HDInsight engines for greater performance.

Supported Hadoop distributions

PolyBase can access data from other Hadoop vendors through its support for the following Hadoop distributions:

- HDInsight in the appliance (based on Hortonworks 1.3)
- Linux distribution 2.0 from Hortonworks
- Windows Server distribution 2.0 from Hortonworks
- Linux distribution 4.3 from Cloudera
- HDInsight in Windows Azure

More information

For more information about the Microsoft Big Data solution, visit the following:

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

Microsoft

The PolyBase technology and HDInsight are powered by Java.

Java is a registered trademark of Oracle and/or its affiliates.