# Microsoft Analytics Platform System

## Solution Brief

# Contents

# Introduction

**For decades, the data warehouse has been at the center of the enterprise's decision support infrastructure, acting as the system of record for data analysis. Now the traditional data warehouse has reached a critical point, requiring major business-driven changes to the systems in place today.**

Key contributing factors include:

**Data growth.** Databases designed with traditional symmetric multiprocessing (SMP) architecture cannot scale to keep up with the amount of data that is expected to grow tenfold over the next five years without major investments in hardware, tuning, support, and maintenance.

**Non-relational data.** Organizations are using Apache Hadoop to store existing data and process new data types from sources like blogs, sensors, social media, and devices. This data can get isolated from users because it is not integrated with data in the traditional data warehouse.

**End-user expectations.** End users need results in near real time, and they expect their internal systems to match the speed of an Internet search engine.

The modern data warehouse needs to enable users to collect and analyze virtually all data, regardless of its size or type. It also needs to deliver performance, scale, and user accessibility to keep up with enterprise demand in this world of Big Data.

# Microsoft Analytics Platform System

Microsoft is ready to help organizations transition to a modern data warehouse with the Microsoft Analytics Platform System (APS), a no-compromise solution that unifies non-relational data from HDInsight, Microsoft's 100-percent Apache distribution of Hadoop based on the Hortonworks Data Platform, with relational data from Microsoft SQL Server Parallel Data Warehouse (PDW), a massively parallel processing (MPP) relational data warehouse, into a single integrated appliance. APS provides tier-one performance, low total cost of ownership (TCO), and accessibility to all users through some of the most widely-used business intelligence tools (BI) in the industry, such as Microsoft Excel. Organizations can take advantage of APS to achieve the following:

- Improved MPP performance with SQL Server PDW
- Seamless Hadoop integration with relational data using Transact-SQL (T-SQL) in PolyBase
- Integrated Hadoop performance through HDInsight (optional)

# Enterprise-ready Big Data

APS supports up to 6 petabytes of relational data using SQL Server PDW for large data volume needs. When it comes to working with a variety of data, data scientists often rely on the Apache Hadoop platform for Big Data solutions. APS now has the option of adding HDInsight directly into the appliance with failover redundancy for the head node, worker nodes, and the Hadoop Distributed File System (HDFS) storage.

By integrating HDInsight with APS, you can take immediate advantage of these enterprise-ready Big Data features:

- **High performance tuned within the appliance.** Microsoft and its hardware partners have optimized HDInsight for maximum performance and reliability. HDInsight takes advantage of the high-speed network connection within the appliance for maximum throughput.
- **End-user authentication with Active Directory.** With Active Directory, administrators can ensure that end users have the correct access permissions to the data within the HDFS.
- **Management and monitoring with Microsoft System Center.** Microsoft provides a free APS Management Pack that IT administrators can use to make sure HDInsight and SQL Server PDW are operating at optimal efficiency.
- **100-percent Apache Hadoop.** HDInsight uses the Hortonworks distribution of Apache Hadoop. Both Microsoft and Hortonworks contribute advances to the Apache Hadoop project and there is no proprietary technology as part of HDInsight.
- **Accessible insights for everyone with Microsoft BI tools.** With APS and HDInsight, you can make both your relational data and Hadoop data available to your entire organization using Microsoft BI tools such as SQL Server Reporting Services, Excel, and Office 365 with Power BI.

### Connecting islands of data with PolyBase

Hadoop solutions are often built in silos and can be complex. This can result in a steep learning curve for developers. Using Hadoop typically requires a significant training investment and high cost for integration.

PolyBase enables you to bring Hadoop solutions and the relational data warehouse together by taking advantage of the following capabilities:

- A single T-SQL query model for PDW and Hadoop with the rich features of T-SQL, including joins without extract, transform, and load (ETL) operations. SQL Server developers and database administrations can become instantly productive by using T-SQL, tying together siloes of Hadoop solutions with SQL Server PDW.

- Integration of the parallel processing of both the MPP database engine and Hadoop to enhance query execution performance that joins relational data with Hadoop data.

- Support for Microsoft Azure HDInsight and Azure Blob storage to enable new hybrid cloud scenarios.

- The ability to query leading non-Microsoft Hadoop distributions, such as Hortonworks and Cloudera, without getting locked into a vendor's proprietary Hadoop distribution.

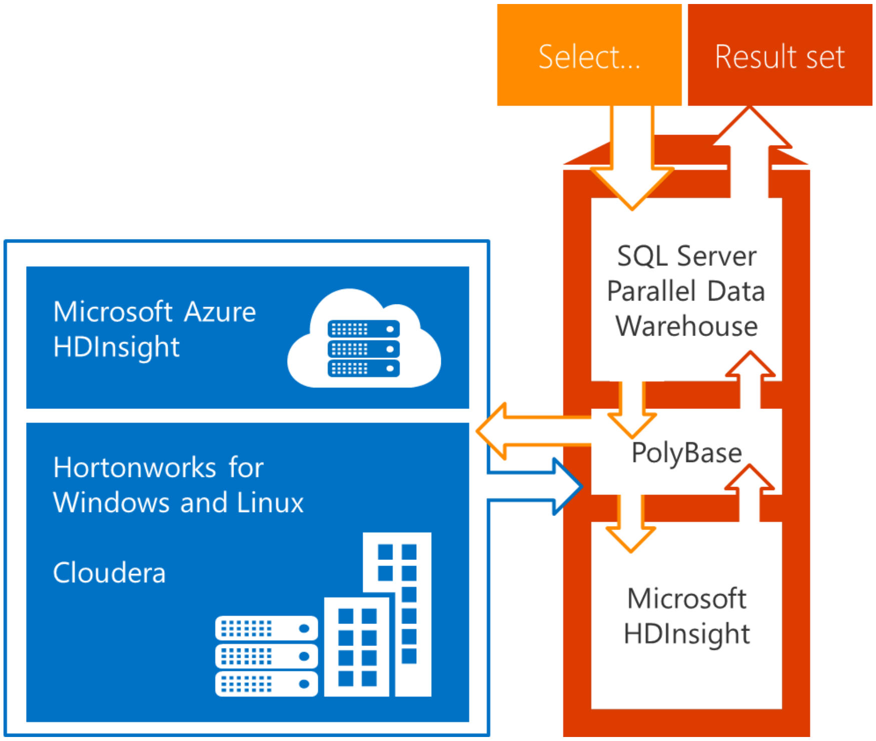

Select...
Result set
Microsoft Azure HDInsight
Hortonworks for Windows and Linux
Cloudera
SQL Server Parallel Data Warehouse
PolyBase
Microsoft HDInsight

*Figure 1 – Bringing together Hadoop solutions with the data warehouse*

With PolyBase, it's easier than ever to:

- Run high performance queries against Hadoop data.

- Archive data warehouse data to Hadoop.

- Export relational data to Hadoop solutions.

- Import Hadoop data into a data warehouse for broader access to users.

### Case Study: Shinsegae Corporation

Shinsegae Corporation, a major department store chain in Korea, needed better performance for customer data mining and basket purchase analysis. Shinsegae took advantage of the integration of PDW and Hadoop to combine 40 terabytes of data, and was pleased to see PolyBase performing nearly twice as fast as their best Hive/Hadoop environment.

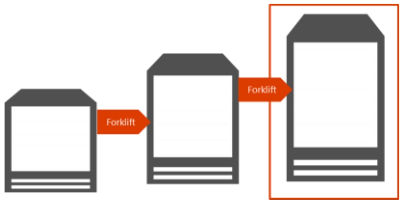
*"We are really satisfied with the performance of PolyBase to allow us to join relational and Hadoop data (weather data, board data, text data) faster and easier. PolyBase is a really powerful feature of PDW to deploy a Big Data system. PolyBase is one of the reasons we selected PDW as our Big Data platform." – HunDong, Kim, Big Data Team, Shinsegae Corporation*

# Next-generation performance at scale

APS was built to scale to multiple petabytes, handling data stored in both relational database management systems and Hadoop, to deliver the performance that meets today's near real-time and rapid insights requirements.

## Performance limitations and scale with traditional data

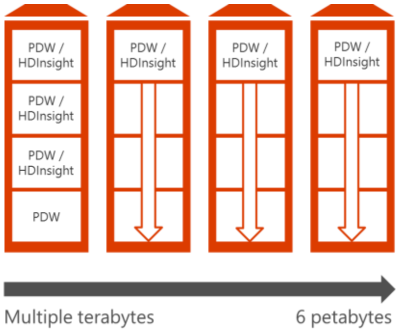
Today, your data warehouse is most likely built on the traditional scale-up SMP architecture, which uses tables organized in rowstore format with indexes.



**SMP architecture limitations.** An SMP scale-up solution runs queries on a single box that shares CPU, memory, disk, I/O operations, and more. To get more scale in a scale-up solution, you need to acquire a more powerful server every time. At some point, you reach diminishing returns after a certain scale.

**Rowstore limitations.** Rowstore stores data in traditional tables as rows. The values comprising one row are stored contiguously on a page. Rowstore tables work great for online transaction processing, but are sub-optimal for data warehouse queries.

Together, SMP technology and rowstore tables limit your data warehouse workload performance and capacity.



## Using APS to scale out your data to petabytes

With APS, you can start with a quarter rack and expand out to multiple racks to support data warehouse workloads up to 6 petabytes. With APS, you get the following capabilities:

- Multiple nodes with dedicated CPU, memory, and storage
- Ability to incrementally add hardware for near-linear scale to multiple petabytes
- Ability to handle query complexity and concurrency at scale

- No “forklift” of prior server to increase capacity
- Ability to scale out HDInsight and/or PDW

## APS delivers blazing-fast performance

The MPP architecture of SQL Server PDW takes advantage of distributing data across servers. In addition, each processor in the server can further provide parallel execution of the sub-queries on the server.

Up to 100x faster queries

Up to 15x more compression

By changing the primary storage engine to a new, updateable version of In-Memory Columnstore, data is grouped and stored one column at a time. The benefits to doing this are as follows:

- Reads only the columns needed for the query. This means SQL Server PDW reads less data from disk to memory and later moves data from memory to the processor cache for faster performance.
- High compression of column data reduces the number of bytes that must be read and moved for data warehouse queries.

The combination of MPP technology and In-Memory Columnstore gives you next-generation performance that is up to 100 times faster than traditional SMP data warehouse solutions and provides traditional indexes that have up to 15 times the compression of rowstore tables.

## Faster data loading performance

SQL Server PDW parallel data loading performance has improved by up to 60 percent over prior versions. One scale unit (equivalent to two compute nodes) in SQL Server PDW can import as much as 480 GB per hour. Additionally, Microsoft internal testing has validated data loading rates of 9.4 terabytes per hour on 40 compute nodes.

## Delivering greater concurrency with multiple workloads

The combination of MPP architecture, in-memory updateable columnstore tables, faster parallel data loading performance, PolyBase, and optimized HDInsight provides you greater concurrency with different types of modern data warehousing workloads that can fuel greater adoption of APS throughout your organization. Figure 2 shows examples of how APS handles multiple concurrent workloads within your organization.

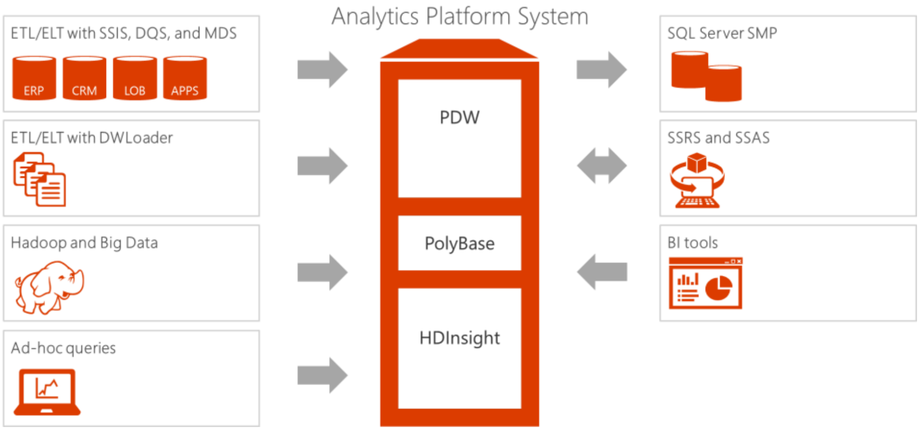


*Figure 2 – Examples of how concurrency and great performance with mixed workloads can fuel greater adoption of APS*

- Extract, transform, and load (ETL) and extract, load, and transfer (ELT) with SQL Server Integration Services (SSIS), Data Quality Services (DQS), and Master Data Services (MDS) provide intra-day uploads of clean and consistent data for SQL Server PDW.

- Using ETL and ELT with DWLoader provides SQL Server PDW with near real-time data from your operational data.

- PolyBase enables users to run faster queries against Hadoop data and import or export data between Hadoop clusters and SQL Server PDW.

- Faster query performance allows users to make concurrent ad-hoc queries against APS without impacting operational workloads.

- CREATE TABLE AS SELECT capabilities allow linked tables between SQL Server PDW and SQL Server SMP data warehouses for a spoke-and-hub data warehouse solution.

- The ability to quickly refresh online analytic processing (OLAP) cubes with direct query provides near real-time updates for business intelligence.

- Faster query performance through direct integration with Microsoft BI tools enables users to access APS with little impact on other users.

### Case Study: CROSSMARK

CROSSMARK needed faster and more detailed insight into terabytes of information about product supply and demand. They deployed a turnkey business intelligence solution from Microsoft and HP that is based on the Microsoft SQL Server Parallel Data Warehouse.

*"People can instantly create their own reports with SQL Server Power View and PowerPivot for Excel and… they can build those reports 50 percent to many times faster compared with the previous system." – David Reis, Director of Application Development*

## Engineered for optimal value

More than a converged system, PDW has reshaped the very hardware specifications required through software innovations to deliver optimal value.

Microsoft APS streamlines the hardware footprint and optimizes value for the modern data warehouse investment. Windows Server 2012 reduces cost through Storage Spaces by achieving levels of performance that are on par with storage area networks (SANs) through commodity storage drives. In-Memory Columnstore helps organizations reduce storage usage by up to 70 percent. As a result of these innovations, SQL Server PDW offers the lowest price per terabyte in the industry for a relational data warehouse appliance. By integrating Hadoop into the same rack as the relational data warehouse, organizations can save on consulting and configuration costs for Hadoop as well as hardware, energy, and general data center costs with an integrated appliance.

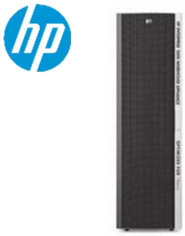
### Case Study: Royal Bank of Scotland

The Royal Bank of Scotland—the leading UK provider of corporate banking services—needed a powerful analytics platform to improve performance and customer services. The bank implemented a Microsoft SQL Server 2012 Parallel Data Warehouse appliance to increase productivity by 40 percent for faster response to business needs.

*"I knew that it would be easy for my team to transition from managing SQL Server databases to SQL Server 2012 PDW, and the solution cost about 85 percent less than products from other vendors." – Alan Grogan, Chief Analytics Officer*

# Hardware vendor choices

Microsoft has partnered and extensively co-engineered appliance solutions with Dell, HP, and Quanta. The following table provides information about baseline rack designs.

| | Microsoft APS by Dell | HP ConvergedSystem 300 for Microsoft APS | Microsoft APS by Quanta |
|---|---|---|---|
| | |  | |
| **Servers** | PowerEdge R620 | ProLiant Gen8 DL360 | STRATOS S810-x52L |
| **Compute nodes** | Up to 9 per rack (3 minimum) | Up to 8 per rack (2 minimum) | Up to 9 per rack (3 minimum) |
| **Racks** | ¼ to 6 | ¼ to 7 | ¼ to 6 |
| **Raw disk capacity (uncompressed)** | 0 terabytes to 1.2 petabytes | 0 terabytes to 1.2 petabytes | 0 terabytes to 1.2 petabytes |

# Conclusion

The challenges of Big Data are as monumental as the opportunities. Even as the traditional data warehouse strives to change to meet the requirements of the modern enterprise, data volumes continue to increase. Likewise, business velocity continues to accelerate, changing business operations and customer interactions along the way, and data continues to become even more diverse and available than ever before.

The Microsoft Analytics Platform System is a no-compromise modern data warehouse solution that seamlessly combines a best-in-class relational database management system, in-memory technologies, Hadoop, and cloud integration in a turnkey package built for Big Data analytics.

To learn more about APS, visit www.microsoft.com/aps

To learn more about the modern data warehouse solution, visit http://www.microsoft.com/en-us/server-cloud/solutions/modern-data-warehouse/default.aspx#fbid=IsK_qrqtR35