IBM dashDB for analytics MPP service

A fully managed data warehouse in the public cloud offering high speed and massive scalability

Highlights

- Scale out your data warehouse in the cloud for larger data sets
- Reach new performance heights with the MPP cluster architecture
- Run analytics and business intelligence tools faster with MPP

Contents

IBM® dashDB™ for analytics MPP is a high-performance, massively scalable cloud data warehouse service fully managed by IBM. dashDB MPP is designed to enable simple and speedy information management, analytics and business intelligence operations in the cloud. It offers the same ease of use as other dashDB for analytics configurations and adds the ability to handle much larger data sets.

The massively parallel processing (MPP) architecture found in dashDB for analytics uses a networked cluster of servers working in parallel to speed up query fulfillment. In the dashDB for analytics MPP cluster, multiple servers work on the same query simultaneously with processing at each server parallelized across all the CPUs. Furthermore, the dashDB for analytics MPP cluster provides more storage capacity for each data set. The resulting performance boost can save you valuable time and resources as you extend the reach of your data warehouse in the cloud.

Because dashDB for analytics MPP is fully managed by IBM, users are free to store, manipulate and analyze their data without the added complexity of network cluster maintenance and database management.

dashDB for analytics: A data warehouse in the public cloud fully managed by IBM

dashDB for analytics is the fully managed cloud data warehouse for builders from IBM. These builders are the developers, database administrators, business analysts, data scientists and others who are bringing new solutions, architectures and applications to market every day. IBM manages the setup, configuration, tuning and disaster recovery operations for the dashDB for analytics service. This means you can immediately begin creating your newest project without spending time and resources building out data warehousing infrastructure.

IBM dashDB for analytics is designed for performance and scale, utilizing technologies, including IBM BLU Acceleration®, embedded Netezza® in-database analytics and SoftLayer® bare metal infrastructure to help provide a high-speed, flexible environment for data management and analytics. dashDB for analytics is available through the IBM Bluemix® platform, which can help make it easier to spin up a dashDB for analytics service as you need it. The platform also helps users seamlessly connect to the many other cloud services available through Bluemix.

dashDB for analytics is designed with the greater business intelligence ecosystem in mind. It's compatible with advanced analytics tooling, including R predictive analytics, with RStudio fully integrated and IBM Watson Analytics™. It connects directly with other IBM cloud data services like IBM Cloudant® and IBM Bluemix Lift. Plus, it works with a wide range of third-party business intelligence (BI) toolsets, including Looker, Aginity Workbench, Tableau, MicroStrategy and many more.

dashDB for analytics is ideal for:

- Augmenting your existing on-premises data warehouse to create a hybrid environment
- Analyzing JSON data from mobile applications
- Running predictive analytics on your data stored in the cloud
- Creating a full enterprise data warehouse on the cloud

dashDB for analytics lets you begin building almost immediately without requiring hardware or software setup. And if needed, you have 24x7 support from IBM specialists.

Why MPP for cloud data warehousing?

dashDB for analytics MPP provides all the benefits of the standard dashDB for analytics service but with even more speed and scalability to handle much larger data sets. This offering of dashDB for analytics is augmented by an MPP architecture, a performance-driven environment for large-scale data warehousing in the cloud.

Customer voice: RSG Media

"The tremendous growth of data is redefining today's competitive advantage. With IBM's dashDB and Cloudant, we can leverage a modern and complete cloud-based data analytics portfolio, which allows us to accelerate our delivery of products and services for analytically savvy media companies. With less time and money spent on IT pains, we can direct our focus on our strategic imperative to provide innovate ways to maximize revenues for media companies' content, advertising and promotional inventories."

Mukesh Sehgal, President and CEO, RSG Media

MPP allows the data warehouse to leverage multiple servers and processors in a network cluster so you can process data simultaneously. In a standard architecture, parallelization occurs only at the processor level. With an MPP architecture, a query is broken up into pieces so multiple servers with their own local storage and compute capacity can work on separate pieces of the data. This team effort can reduce I/O requirements and significantly accelerate the querying process.

With MPP, performance improvements are increased with each new server added to the network cluster. For example, if a query takes one hour in a standard architecture using a single server, it would take approximately 15 minutes with an MPP cluster comprised of four servers. Adding additional servers can further reduce the time needed to process a query. With a total of five servers, the query would need only 12 minutes; with six servers, 10 minutes; and so on. Therefore, with dashDB for analytics MPP, scaling out can be as simple as adding additional servers to your cluster.

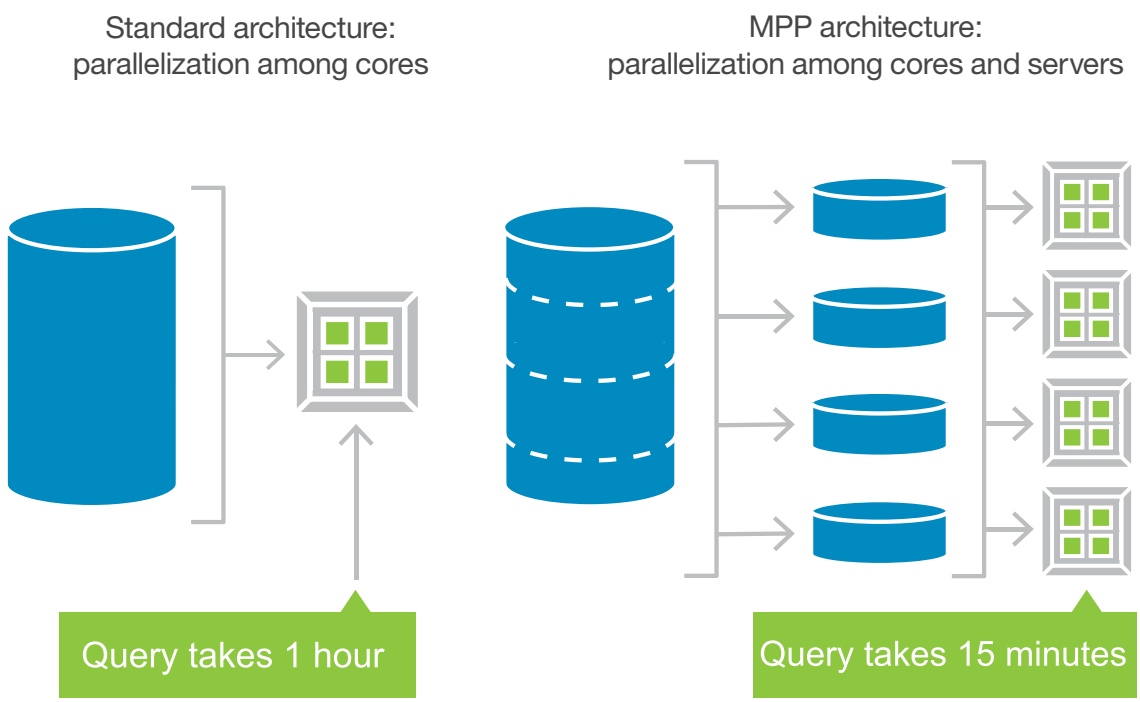

Figure 1: MPP helps accelerate queries in IBM dashDB for analytics.

Massively high performance in an IBM cluster

The dashDB for analytics MPP service uses the massively parallel architecture described above to achieve greater performance and scalability. These benefits are further compounded by leveraging the breakthrough BLU Acceleration dynamic in-memory column store technology from IBM, which dashDB for analytics MPP extends to the network cluster.

Each individual server working on a query in MPP can leverage BLU Acceleration to minimize I/O and achieve an order of magnitude in speedup compared to conventional row-store databases. The ultra-high speed of BLU Acceleration is made possible by a number of key technologies, including the following:

- **Dynamic in-memory processing:** Even when a data set does not fit entirely in memory, dashDB for analytics still processes data quickly using a series of patented algorithms that enable in-memory acceleration. While every workload is different, dashDB for analytics requires the RAM size to be only five percent of the original preload source data size to run at in-memory optimized speeds.
- **Actionable compression:** dashDB for analytics performs a broad range of operations, including joins and predicate evaluations directly on compressed data, which can improve memory and cache bandwidth, and save CPU costs.
- **Parallel vector processing:** dashDB for analytics is CPU optimized and designed for the latest generation of microprocessors. Both multicore parallelism and single instruction, multiple data (SIMD) vector instructions help dashDB for analytics maximize hardware performance.
- **Data skipping:** BLU Acceleration enables dashDB for analytics to intelligently avoid scanning entire ranges of column data that don't qualify for analysis, which can save time and resources.

Whenever dashDB for analytics MPP is performing distributed joins or aggregation processing, data is exchanged between servers entirely within the BLU Acceleration runtime in native columnar format. This exchange is achieved by using a highly parallel infrastructure optimized for columnar data exchange.

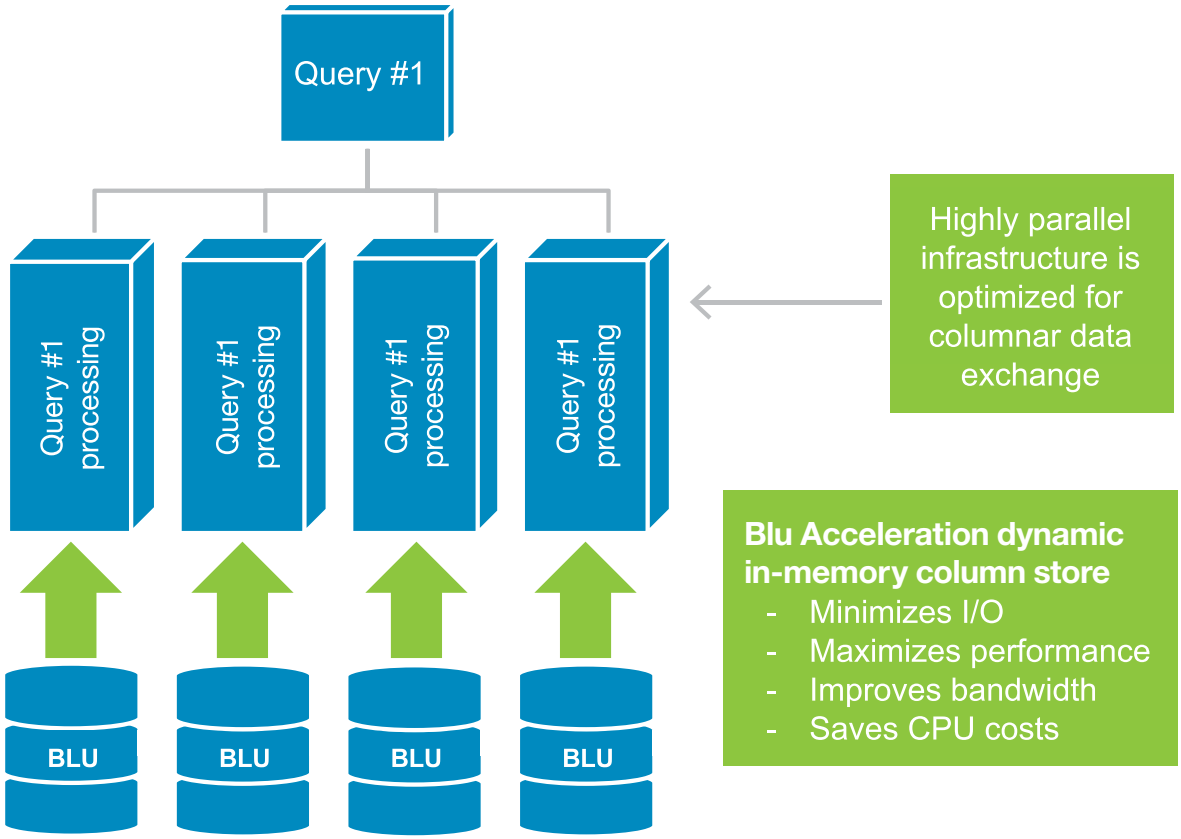

Figure 2: Using BLU Acceleration in an MPP scale out

In the MPP architecture, the use of a common table dictionary helps data remain in an optimized form when being exchanged over the network during query processing. This method can significantly reduce network traffic and increase the overall effective network bandwidth.

MPP can be extremely effective

An IBM data warehouse benchmark has shown that a deep analytic workload achieved a performance speedup of 10 times on a dashDB for analytics MPP instance. This was achieved using a three-server configuration compared to a dashDB for analytics Enterprise single-server instance using a 4 TB server. The benchmark measured the throughput performance of 60 concurrent query streams generated by an IBM Cognos® application in combination with queries from a public benchmark.[1]

This stunning speedup was achieved by the ability of dashDB for analytics MPP to leverage multiple servers in parallel, using the massively parallel network cluster architecture described above. Further, dashDB for analytics MPP is optimized for Intel Xeon E5 v3 and later processors, which allows efficient scaling up to use a greater number of cores, helping deliver faster response times and an improved memory bandwidth. dashDB for analytics MPP also upgrades the I/O subsystem, which can improve input/output operations per second (IOPS).

Snapshot of dashDB for analytics MPP performance benefits

- Higher memory-per-core ratio (10.6 GB per core)
- Improved I/O subsystem with higher IOPS
- MPP parallelism for heavy group-by and join queries
- New enhanced workload manager (WLM) configuration

The advantages of a multilingual data warehouse

There are almost as many dialects of SQL as there are database products on the market, and dashDB for analytics is designed to speak more database dialects than any other data warehouse.

Whether you've coded an application to Oracle, IBM DB2®, PostgreSQL, Netezza or are starting a brand new project for the cloud, the flexible language support of dashDB for analytics covers virtually all the major SQL extensions. In addition to SQL language variants, dashDB for analytics provides support for a wide range of application interfaces including:

- ADO
- Embedded SQL
- JDBC
- .NET (C#)
- Node.js
- ODBC
- OLE DB
- Oracle Call Interface (OCI)
- Perl
- PHP
- PL/SQL for Oracle and Netezza variants
- Python
- Ruby
- Scala
- SQL*Plus scripts
- Visual Basic

Primary use cases

The dashDB for analytics MPP service is highly flexible and can be implemented for a large variety of business use cases. Here are five broad scenarios where dashDB for analytics can help you gain more value from your data:

1. **Stand-alone cloud data warehouse**
   The scalability and performance available using dashDB for analytics mean you can use it as a stand-alone, fully managed cloud data warehousing service, regardless of your size. You can also use it to help build a data mart, a development environment or an enterprise data warehouse.
2. **Development and QA systems**
   If you have a powerful data warehouse on premises that you're using for critical workloads, you may not want your developers testing new code there. With dashDB for analytics, your developers can experiment, build new application code and test it on the cloud without disrupting on-premises operations. Because dashDB for analytics is compatible with Oracle, DB2, Netezza and PostgreSQL, you can have confidence that code developed and tested on dashDB for analytics should run well on premises, too.
3. **Augmenting the existing data warehouse through a hybrid strategy**
   With dashDB for analytics, you can build a hybrid data warehouse that extends on-premises data warehouse environments to the cloud. Since you pay only for the capacity you need, the platform is elastic and can cost-effectively grow with your business needs.
4. **Analysis of NoSQL data**
   You can synchronize JSON documents within Cloudant to structured data within dashDB for analytics, helping provide a way to bring BI and analytics to your unstructured data.
5. **Data science data store**
   The dashDB for analytics service maintains a robust set of predictive analytics algorithms for data scientists and analysts, and includes built-in R runtime and RStudio. These algorithims make dashDB for analytics an optimal data warehouse to support data analysis and statistical software development.

Getting started with the dashDB MPP service

Users new to dashDB for analytics can create a new entry-level dashDB service quickly through the IBM Bluemix platform. Simply log in to the Bluemix platform using a Bluemix ID and navigate to *dashDB for analytics* in the service catalog. To upgrade and add the storage, performance and scale advantages of dashDB for analytics MPP, contact your IBM Cloud Data Services sales representative, or send an email to dashDB_Info@wwpdl.vnet.ibm.com.

A dashDB MPP use case example: Hybrid data warehouse

One dashDB for analytics MPP client maintains customer data for sports and entertainment venues in a large number of small on-premises SQL server data marts. The company leverages this data for analytics to help improve its operations. Now, the client is incrementally moving its data marts to the cloud, but it needs to scale beyond the confines of a single server and do it quickly.

The dashDB for analytics MPP service is helping the company seamlessly migrate its data using IBM Bluemix Lift. Bluemix Lift allows the company to accelerate its analytics and reporting without having to manage the system itself. In the long term, the company plans to consolidate its data on dashDB and move away from its legacy on-premises environment entirely. The client intends to use IBM Bluemix Data Connect to shape and prepare the data for new analytic techniques, such as Watson Analytics.

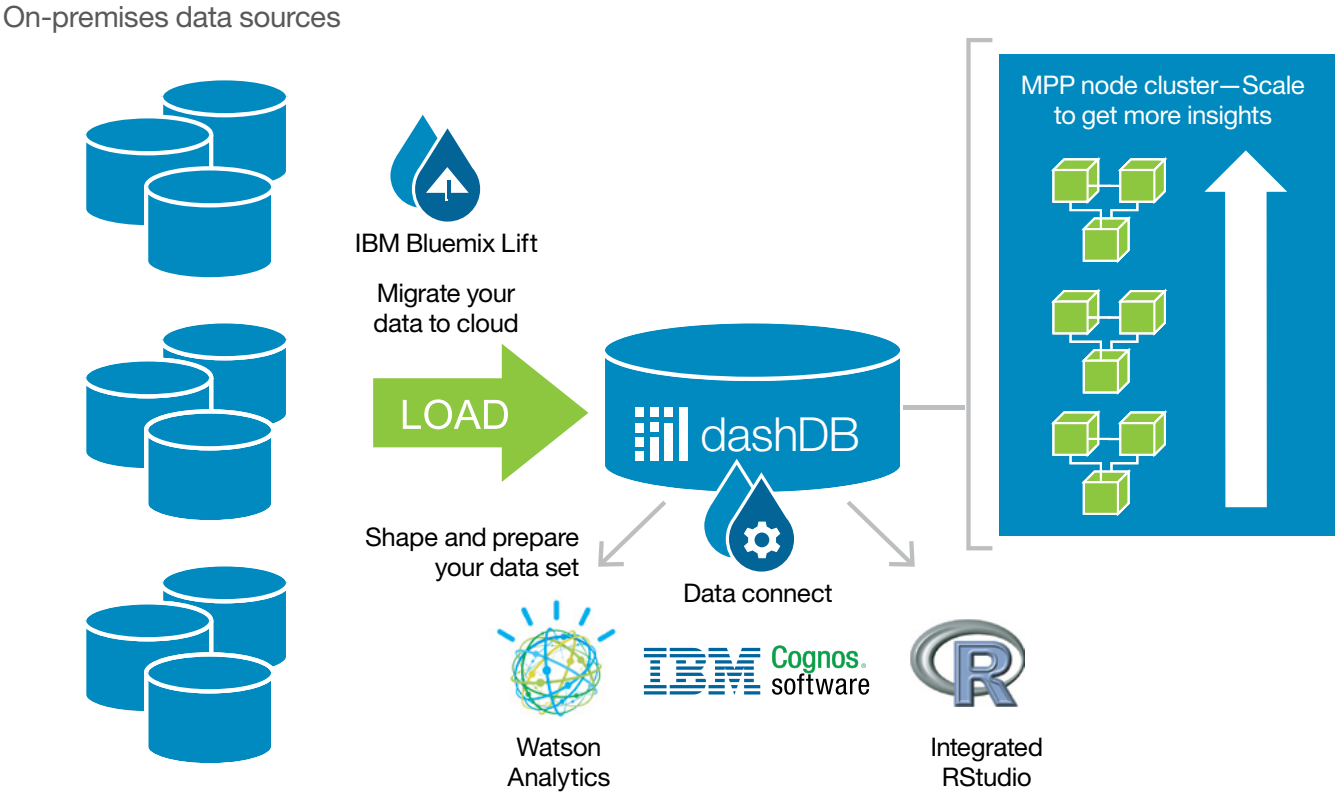

Figure 3: Moving on-premises data marts to the cloud with dashDB MPP

Creating tables and selecting distribution keys in dashDB for analytics MPP

dashDB MPP uses a hashing function to distribute table data across database servers. To achieve optimal data distribution and performance, a distribution key should be specified for any tables that do not have an explicit primary key. Otherwise, you can employ a default distribution key provided by the dashDB for analytics MPP service.

There are two primary approaches for selecting the optimal distribution key in dashDB for analytics MPP:

1. You can colocate the rows of your fact table with the rows of your biggest frequently joined dimension table, optimizing the performance of the joins.
2. You can look for an identifying column that contains a large number of unique values to achieve an even data distribution across the MPP cluster. This approach helps optimize performance and ensure storage is fully utilized.

About IBM Cloud Data Services

IBM Cloud Data Services provides developers with a comprehensive set of rich, integrated data services covering content, data and analytics. Cloud Data Services offerings can speed up time to market, improve uptime and deliver higher value to developers of web and mobile applications.

For more information

To learn more about dashDB for analytics, please contact your IBM representative or IBM Business Partner, or visit dashDB.com.

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit **ibm.com**/financing.

IBM Corporation
IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
March 2017

CDW12346-USEN-01