IBM Analytics
Technical White Paper

IBM dashDB: Enterprise MPP Service

Fully managed cloud data warehousing with 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 better and faster with MPP

Contents

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

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

Because dashDB 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: IBM’s cloud data warehouse service

dashDB is IBM’s fully managed cloud data warehouse service for builders—the developers, database administrators, business analysts, data scientists, and more 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 service, so you can get straight to creating your newest project without spending time and resources building out data warehousing infrastructure.

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

dashDB 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 DataWorks; and it works with a wide range of third-party BI toolsets including Looker, Aginity Workbench, Tableau, and many more.

Whether you're augmenting your existing data warehouse appliance to create a hybrid environment; analyzing JSON data from mobile applications; running predictive analytics on your data stored in the cloud; or creating a full enterprise data warehouse on the cloud, dashDB lets you get straight to building without any need for hardware or software setup, and with the aid of 24/7 support from IBM experts.

Why MPP for cloud data warehousing?

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

MPP works by allowing the data warehouse to leverage multiple servers and processors in a network cluster to 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 that multiple servers, each with their own local storage and compute capacity, are working on separate pieces of the data. This team effort reduces I/O requirements and drastically speeds up the querying process.

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

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 utilizing just four servers. Adding one more server, for a total of five, reduces the query time to 12 minutes; six servers reduces the query time further to 10 minutes; and so on. Therefore, with dashDB MPP, scaling out is as simple as adding additional servers to your cluster.

Massively high performance in IBM's cluster

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

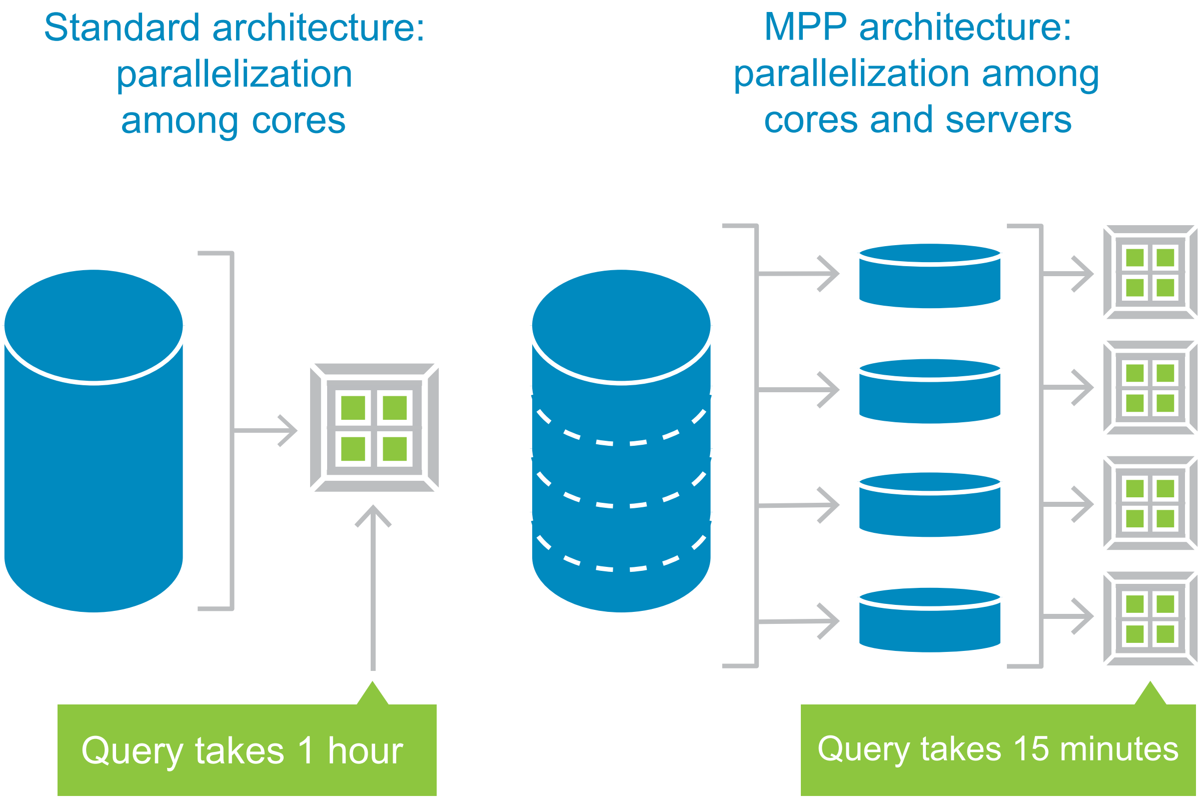
Standard architecture: parallelization among cores
MPP architecture: parallelization among cores and servers
Query takes 1 hour
Query takes 15 minutes

Figure 1: MPP for IBM dashDB

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

- **Dynamic in-memory processing:** Even when a dataset does not fit entirely in memory, dashDB still processes at lightning fast speeds using a series of patented algorithms that enable in-memory acceleration. While every workload is different, dashDB only requires RAM size to be 5 percent of the original pre-load source data size in order to run at in-memory optimized speeds.
- **Actionable compression:** dashDB performs a broad range of operations—including joins and predicate evaluations—directly on compressed data, therefore improving memory and cache bandwidth, and saving CPU costs.
- **Parallel vector processing:** dashDB is CPU optimized and designed for the latest generation of microprocessors. Both multi-core parallelism and SIMD vector instructions enable dashDB to maximize hardware performance.
- **Data skipping:** BLU enables dashDB to intelligently avoid scanning entire ranges of column data that don't qualify for analysis, preserving time and resources.

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

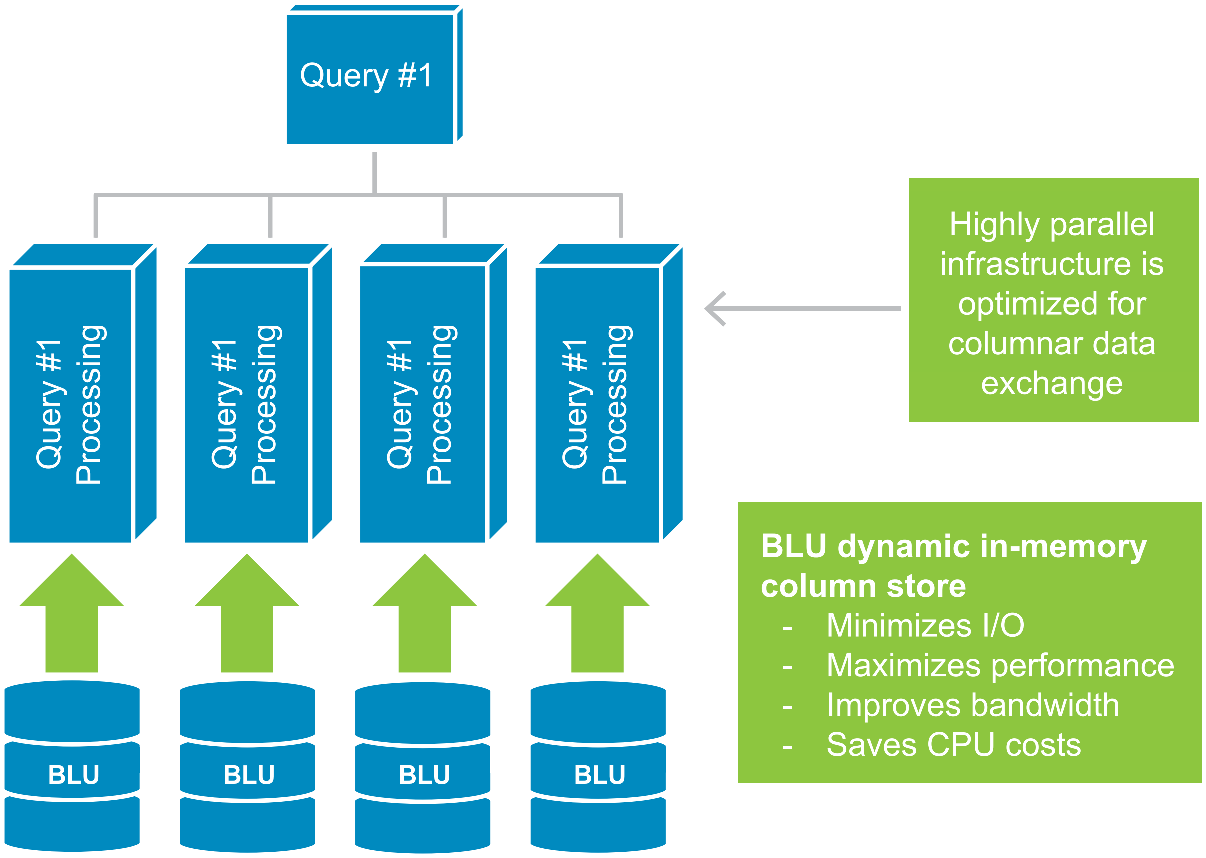

Figure 2: BLU Acceleration - MPP scale out

In the MPP architecture, use of a common table dictionary further enables data to remain in an optimized form when being exchanged over the network during query processing, significantly reducing network traffic and increasing the overall effective network bandwidth.

dashDB MPP: The proof is in the pudding

An IBM data warehouse benchmark has shown that a deep analytic workload achieved a performance speedup of 10 times on a dashDB MPP instance (three-server configuration) when compared to a dashDB Enterprise single-server instance (4TB server configuration). 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.

This stunning speedup was achieved by dashDB MPP's ability to leverage multiple servers in parallel, using the massively parallel network cluster architecture described above. Further, dashDB MPP uses the latest generation of Intel Xeon E5 v3 processors, which allows efficient scaling up to a greater number of cores, delivering faster response times and improved memory bandwidth. dashDB MPP also upgrades the I/O subsystem with better IOPS (input/output operations per second).

Snapshot of dashDB MPP Performance Benefits

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

The ultimate "polyglot" data warehouse

There are almost as many dialects of SQL as there are database products on the market, and dashDB speaks more database dialects than any other data warehouse.

Whether you've coded an application to Oracle, IBM DB2®, PostreSQL or Netezza—or are starting a brand new project for the cloud—dashDB's flexible language support covers all the major SQL extensions you need. In addition to SQL language variants, dashDB 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 MPP service is highly flexible and can be implemented for a large variety of business use cases. Here are five broad scenarios where dashDB can help you gain more value from your data:

1. **Standalone cloud data warehouse**
   dashDB's scalability and performance mean you can use it as a standalone, fully managed cloud data warehousing service—regardless of your size. You can also use it to help you 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, your developers can experiment, build new application code, and test it on the cloud without disrupting on-premises operations. Because dashDB is compatible with Oracle, DB2, Netezza and PostgreSQL, you can have confidence that code developed and tested on dashDB will run well on premises, too.
3. **Augmenting the existing data warehouse through a hybrid strategy**
   With dashDB, you can build your hybrid information management strategy and extend on-premises data warehouse environments to the cloud. Since you pay for capacity as you need it, the platform is elastic and can grow with your business needs.
4. **Analysis of NoSQL data**
   You can easily synchronize JSON documents within Cloudant to structured data within dashDB, providing a way to bring BI and analytics to your unstructured data.
5. **Data science data store**
   The dashDB service maintains a robust set of predictive analytics algorithms for data scientists and analysts, and includes R runtime and RStudio built in. This makes dashDB an optimal data warehouse to support data analysis and statistical software development.

Getting started with the dashDB Enterprise MPP service

Users new to dashDB can create a new entry-level dashDB service quickly and easily through the IBM Bluemix platform. Simply log in to the Bluemix platform using a Bluemix ID, navigate to "dashDB" in the service catalog, and complete the "Add Service" form. To upgrade and get the storage, performance and scale of dashDB Enterprise MPP, contact your IBM Cloud Data Services sales representative, or send an email to dashDB_Info@wwpdl.vnet.ibm.com.

A dashDB MPP scenario: Accelerating analytics

One dashDB MPP customer maintains customer data for sports and entertainment venues in a large number of small on-premises SQL Server data marts, and they leverage this data for analytics to help improve operations. The company is now incrementally moving their data marts to the cloud, but they need to scale beyond the confines of a single server in short order.

The dashDB MPP service is enabling them to seamlessly migrate their data, and to accelerate their analytics and reporting, without having to manage the system themselves. Long term, the customer plans to consolidate their data on dashDB and move away from their legacy on-premises environment entirely.

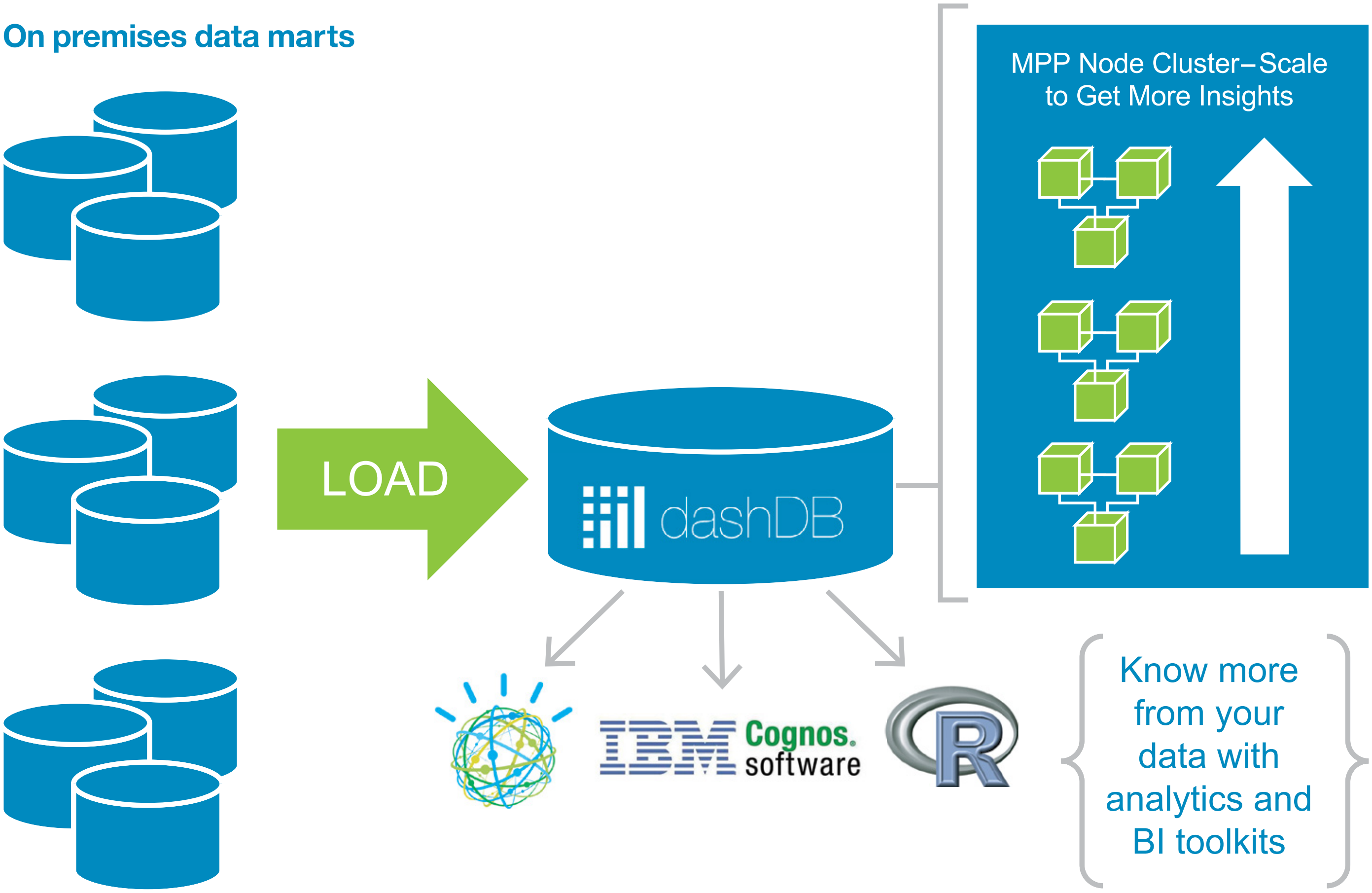
On premises data marts
LOAD
dashDB
MPP Node Cluster–Scale to Get More Insights
IBM Cognos. software
R
Know more from your data with analytics and BI toolkits

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

Creating tables and selecting distribution keys in dashDB MPP

dashDB MPP utilizes a hashing function to distribute table data across database servers. In order 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 MPP service.

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

1. You can collocate the rows of your fact table with the rows of your biggest frequently joined dimension table, optimizing 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 optimizes performance and ensures storage is fully utilized.

Additional information about choosing the right distribution keys can be found here.

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 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, please contact your IBM representative or IBM Business Partner, or visit www.dashdb.com

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. We’ll partner with credit-qualified clients to customize an IT financing solution to suit your business goals, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward.

For more information, visit **ibm.com**/financing

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
July 2015

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

[1] Approximate pre-load uncompressed data.