IBM Watson Query

Deliver universal governed data access
and query any data, anywhere

Deliver data fast with
governed virtualization

Achieve a complete view
of your data

Become data-driven with
a data fabric architecture

The rise of increased data volume and variety has pushed organizations to store data within a myriad of repositories across different vendors and locations. While the practice of storing data where it fits best is good in principle, the fragmented nature of such environments takes its toll. In these fragmented environments, up to **74% of potentially valuable data is not used**.[1]

Physically moving this data through extract, transform and load (ETL) processes to centralize the data in an attempt to bridge the gap between these silos can be complex, time consuming and expensive. This is particularly true when queries need to be adjusted to fit the varying types of data or their repositories. The movement of unstructured data in a data lake to a data warehouse for analysis alone can be a significant time-sink and budget killer.

IBM® Watson® Query is a universal distributed query engine that can query disparate data sources and stores without physically moving them. This technology allows users to see data from all repositories within a single view, query it without moving it or replicating it, and govern it to ensure compliance with legislation. This brief dives deeper into the technologies that permit converged data queries, along with the data and AI platform and data stores that can augment their success.

Deliver data fast with complete and unhindered access through governed virtualization

Watson Query integrates data sources across multiple locations and turns all this data into one logical data view that can be governed. This virtual data view makes getting value out of your data easy, reducing ETL requests by 25-65%.[2]

How does Watson Query achieve this?

Optimize data access
Unify data across any cloud, warehouse, data lake or database, empowering users to fully use all your data for innovation.

Reduce costs
Gain independent scaling of computation and storage with zero downtime while reducing resource consumption by eliminating data movement.

Control access
Manage all permissions in a virtual layer, empowering data consumers to get the data needed for the task at hand without compromising privacy or regulatory requirements.

Achieve a complete view of your data

Watson Query allows users to quickly preview all connected data sources in one view. It automatically finds and matches tables across systems so you can query them as a single object and allows you to easily join and analyze all virtualized objects with the power of SQL.

Through its multi-worker topology and parallel processing mesh, Watson Query's data virtualization technology creates a highly efficient system for data processing. Queries are optimized to take advantage of multiple workers running within the service to enable parallel data ingestion and query processing. You can extend this even further through the mesh capability, allowing you to scale higher numbers of data sources and a higher degree of parallel processing.

Constellation

Data sources

Service cluster

Analytics/ML application

Caching
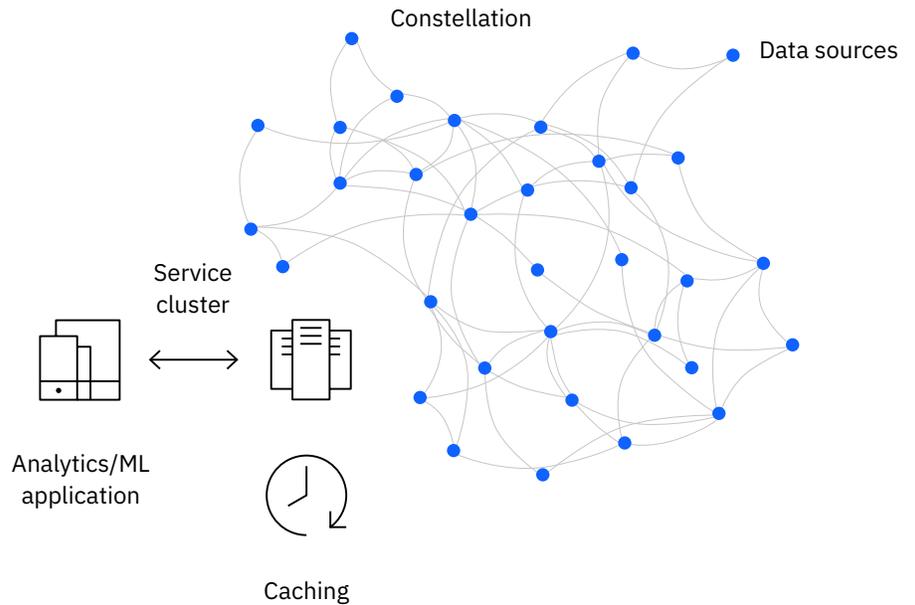
You can query multiple sources using a single query in Watson Query. The user query is broken down and optimized for each data source that it references. Those queries execute in parallel utilizing the processing mesh in a self-balanced way, then Watson Query combines and processes the results as they are received.
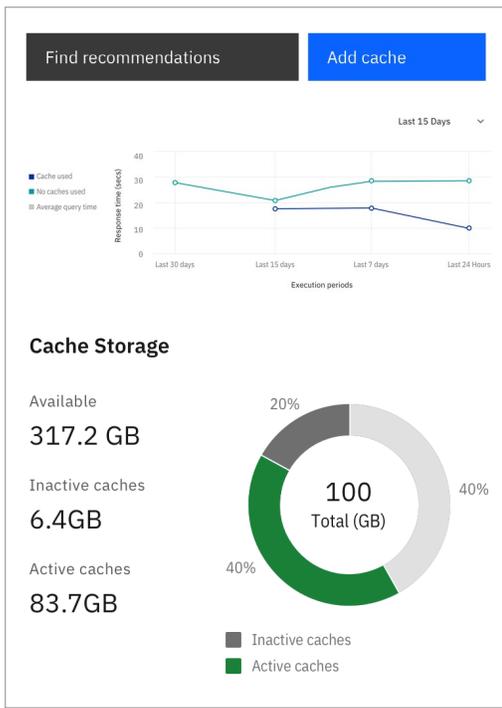
Application development made easy
As your data grows, so do your storage methods. This complicates the number of ways data is described, accessed and processed. Multiple dialects of SQL come into play, barring access to critical data. Watson Query supports multiple SQL dialects for incoming queries and translates that SQL into the native optimized query language for each connected data source. This allows you to write your queries the way you want, to access multiple data sources, vendors, and cloud providers to make application development easy and efficient, so one query can access multiple sources (from 2 to 2,000 and more).

Support for a wide range of data sources
Today, about 95% of businesses are operating in hybrid cloud environments, with about two-thirds using multiple cloud providers combined with on-premise data sources.[3] To be able to use data across these sources, Watson Query provides quick access across deployment methods, repositories and data types. Watson Query is able to query multiple sources, including on-premise data warehouses, SaaS databases and cloud object stores, with new sources constantly being added.

One notable source that traditional integration methods and technology struggle to connect applications to quickly are object stores. Instead of needing to move data from a data lake to a more costly data warehouse, Watson Query enables you to query data directly in the data lake without data duplication or movement.

Find a list of supported IBM and third-party data sources and data types in Watson Query here.

IBM Watson Query

| Find recommendations | Add cache |

Last 15 Days

Cache used
No caches used
Average query time

Response time (secs): 40, 30, 20, 10, 0

Execution periods: Last 30 days, Last 15 days, Last 7 days, Last 24 Hours

Cache Storage

Available
317.2 GB

Inactive caches
6.4GB

Active caches
83.7GB

20%
40%
40%

100
Total (GB)

Inactive caches
Active caches

Intelligent caching

Watson Query monitors your workload and recommends when caching would be most useful. Through our caching engine, it can generate recommendations, evaluate recommended and user-created caches against the workload and automatically maintain and use those caches to optimize query processing. The service contains an advanced, optimizer-integrated, data-caching solution that can cache tables and queries, offering pre-computed result sets so you can analyze faster. There's also flexibility in how you cache, and you can take advantage of easy-to-use cache management functions and multiple caching strategies to improve performance.

Watson Query helps you become data-driven with a data fabric architecture

Enable real time data integration
Watson Query executes distributed and virtualized queries across databases, data warehouses and data lakes. Watson Query is the tool of choice for quick, easy integrations to data sources thanks to its data virtualization capabilities. The amount of effort required for small and large ETL jobs is often the same; data virtualization helps improve efficiency when smaller, ad hoc requests occur. Learn more.

Provide privacy and governance at every query
Applying sweeping governance rules across data lakes, databases and warehouses is time consuming and delays users from getting access to the right data. Watson Query enforces governance policies when data is accessed across multiple sources, quickly providing data to your end applications through a single view without the need for any manual changes, data movement or replication. Learn more.

Conclusion
Simplify your data landscape with IBM Watson Query by connecting data across disparate sources and making it accessible through a single access point—regardless of data location, size, type or format. Achieve optimized and governed data access, reduced costs and controlled access of your data without movement, helping you to meet growing data demands at the scale and speed you need.

For more information
To learn more about IBM Watson Query, please contact your IBM representative or IBM Business Partner®, or visit ibm.com/products/watson-query.

Try IBM Watson Query today!

IBM Watson Query

1. Rethink Data: Put More of Your Business Data to Work–From Edge to Cloud, an IDC and Seagate Technology survey report, July 2020.

2. New Technology: The Projected Total Economic Impact Of IBM Cloud Pak for Data, a Forrester Study commissioned by IBM, December 2020.

3. What's Your Multicloud Strategy? Boston Consulting Group, 19 April 2021.