IBM Fluid Query 1.7

Unifying IBM PureData System for Analytics with the Logical Data Warehouse

Highlights

- Enables query access from IBM® PureData® System for Analytics to Hadoop platforms and Spark data
- Users can leverage a generic query connector to access any relational database using JDBC
- Provides query access between IBM PureData System for Analytics and Relational Database Management System (RDBMS) data in IBM Db2® products, IBM Informix®, Teradata, Oracle, SQL Server, Postgres, MySQL and other IBM PureData System for Analytics appliances
- Allows rapid transfer of database level data between Hadoop and IBM PureData System for Analytics
- Extends IBM PureData System for Analytics by allowing colder data to be offloaded to a lower cost-per-terabyte Hadoop environment
- Available at no additional charge with IBM Netezza® Platform Software (NPS) releases

Overview

How can data consumers, scientists and managers balance costs, while maximizing business insights? Data warehouse environments are rapidly changing to keep pace with demands for user self-service, increased agility, new data types, lower-cost solutions, the adoption of open source technologies, better business insight, and faster time to value.

These requirements have led IT organizations to consider adopting database appliances, Hadoop environments and data platforms on the cloud. The idea of a single, enterprise data warehouse holding all the data is no longer the prevailing architecture approach. Organizations need to use analytics across all data formats on a variety of platforms and repositories.

The architecture of a single enterprise data warehouse is evolving toward a Logical Data Warehouse that is used to describe the collection of data assets, which may reside in different forms, structures and platforms, yet all support the data requirements for analytics. The Logical Data Warehouse architecture approach gives organizations the ability to manage structured, semi-structured and unstructured data types, and different latency requirements—in the location where it makes the most sense—to drive a variety of analytics requirements. It also abstracts data access so applications need not change to gain insight and value from data across the Logical Data Warehouse. This may sound easy, but many organizations lack the tools to enable fluid data access to gain insights from all their data stores.

IBM Fluid Query 1.7 provides data access between Apache Spark or Hadoop and IBM PureData System for Analytics. Your current data warehouse, the IBM PureData System for Analytics, can be extended in several important ways over this bridge to enable additional Hadoop capabilities. The coexistence of IBM PureData appliances with the beneficial features of Hadoop is a best-of-breed approach where tasks are performed on the platform best suited for that workload. IBM PureData System for Analytics appliances can also directly query data in Db2 products, Oracle, Teradata, IBM PureData System for Operational Analytics and other IBM PureData appliances using native connectors, while providing access to any other structured data source that supports JDBC—such as Teradata, MapR or Microsoft SQL Server—using Fluid Query 1.7. This is an important step in data integration. Use the IBM PureData System for Analytics data warehouse for production-quality analytics where performance is critical to the success of your business, while simultaneously using Hadoop and Spark to discover the inherent value of full-volume data sources.

What is IBM Fluid Query?

IBM Fluid Query is a capability that unifies data access across the Logical Data Warehouse. Users and analytic applications need access to data in a variety of data repositories and platforms without concern for data location or access method, or the need to rewrite a query. IBM Fluid Query enables a data store to route a query, or part of a query, to the correct data store within the Logical Data Warehouse so the query can flow to the data instead of the data flowing to the query.

No matter where users connect within the Logical Data Warehouse, they can access all data through the same, standard API/SQL access. IBM Fluid Query powers the Logical Data Warehouse by giving users the ability to combine their data, even if it is spread across various sources, in a fast, agile manner to drive analytics and deeper insight, without having to understand how to connect multiple data stores, use different syntax or APIs, or change their application.

IBM Fluid Query allows access to data in Hadoop or Spark from IBM PureData System for Analytics appliances and enables the fast movement of data between Hadoop and IBM PureData System for Analytics appliances. Using query and data movement, IBM Fluid Query connects those appliances to common Hadoop systems like Cloudera and Hortonworks with the ability to access Spark through a Spark SQL connector. This provides an additional level of flexibility when accessing data residing on the Hadoop framework. With IBM Fluid Query, you can query against IBM PureData System for Analytics, Hadoop or both by combining results from IBM PureData System for Analytics database tables and Hadoop data sources to create powerful analytics combinations.

IBM Fluid Query 1.7 enables your existing IBM PureData System for Analytics applications to gain insight from even more data. You can now run your existing queries, reports and analytics against data on Hadoop or structured database data, in addition to the data in your appliance.

"The ability to answer complicated questions with data from disparate sources will allow our analysts to focus on answering business questions without having to worry about where the data lives or waiting on a project to perform the data integration for them."

— David Darden,
Business Intelligence Engineering Manager, Big Fish Games

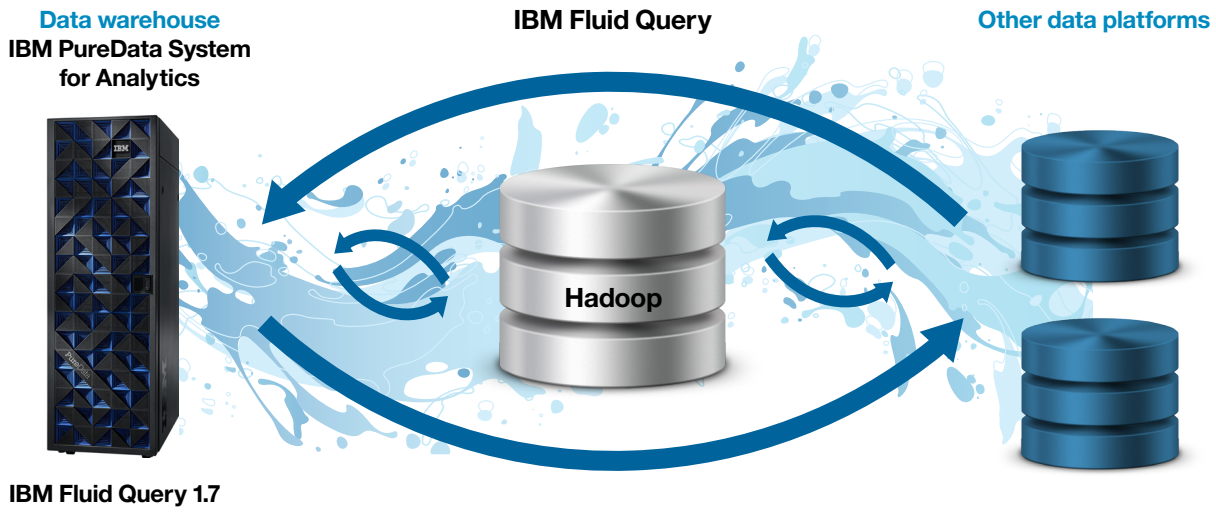

Figure 1: IBM Fluid Query is the capability that unifies data access across the Logical Data Warehouse.

Exploiting Hadoop functionality

IBM PureData System for Analytics delivers advanced analytics with speed and simplicity. A new data source evaluated first on Hadoop and where value is discovered, can then be loaded into the data warehouse. IBM PureData System for Analytics appliances also support workloads with high-performance standards enforced by service level agreements.

Hadoop distributed file environments excel at storing large data volumes and accessing both structured and unstructured data. Hadoop is well suited for data archiving, exploration and integration. It can easily handle data in different formats, since data can be stored without a defined schema. Hadoop can be used as a "Day 0" archive or as a staging area for storing and managing new data—making it an ideal platform for evaluating new data sources. Apache Spark exploits in-memory technology that works well for analyzing the exhaust from the Internet of Things logs to reach machine learning conclusions.

IBM Fluid Query bridges this gap. It is easy to install and gets you connected to leverage the strengths of Hadoop and IBM PureData System for Analytics environments. Only one percent of the queries in today's data analytic systems touch 100 percent of the data in your system, which makes that activity perfect for the lower cost and the performance Hadoop offers. At the other end of the query usage spectrum, 90 percent of current queries touch only 20 percent of the data, which matches well to the characteristics of the IBM PureData System for Analytics — reliability with better analytics performance.

"Installing and configuring IBM Fluid Query was a snap. And once we knew what features we wanted to deliver, our implementation was pretty straightforward..."

— Brian Weissler,
Director of Product Management, Aginity

What's new in version 1.7

Open access with a generic connector

- Users can leverage pre-defined templates to quickly access Informix, Oracle, SQL Server, Teradata, PostgreSQL and MAPR-DB via Hive.
- Support for Kerberos enables use with the generic connector.

Added flexibility for accessing Hadoop file formats on HDFS

- Select the specific file format and compression mode during import of data to HDFS from IBM PureData System for Analytics. Support for AVRO, Parquet, ORCfile and RCfile allow users to move, store and read compressed file formats using the IBM Fluid Query connector.
- Synchronize Hive and Big SQL during the import of data to Hadoop from IBM PureData System for Analytics automatically.
- Import a combination of multiple schemas and tables on Hadoop from IBM PureData System for Analytics.
- Support of Hadoop services for EOL and special characters can be extended, but is not currently supported by query frameworks on Hadoop.

Usability and currency

- Support for Big SQL 4.1
- Support for IBM Cloudera 5.5.1
- New Getting Started and Best Practices guides help simplify your implementation.
- New instructional videos help with your IBM Fluid Query implementation.
- Validate IBM Fluid Query environments on both IBM PureData System for Analytics and Hadoop with additional configuration checkpoints.

Data does not have to be forgotten, purged or unavailable. Instead, store the less active ("colder") data on Hadoop, and the more important, active data on the IBM PureData System for Analytics appliance. More specifically, consider storing all data on Hadoop and only the "hot" data on IBM PureData System for Analytics. Route ten percent of the queries that need 80 percent of the data to Hadoop, leaving the majority of the queries that only need access to the "hot" data on IBM PureData System for Analytics, an appliance known for its speed and simplicity.

IBM Big SQL also provides IBM Fluid Query capabilities within the Logical Data Warehouse. IBM Big SQL is a Structured Query Language (SQL) engine that provides seamless access to data across any system from Hadoop, using JDBC or ODBC, whether that data exists in Hadoop or a relational database. This feature provides developers with SQL skills access to data in Hadoop, and across the Logical Data Warehouse, without having to learn new languages or move massive amounts of data. The IBM Big SQL massively parallel processing (MPP) design takes full advantage of the Hadoop distributed file architecture and offers a higher degree of SQL compatibility than other vendors. By better adhering to SQL standards and the Hadoop physical storage design, IBM Big SQL facilitates data access within the Logical Data Warehouse.

IBM Big SQL supports query federation to many data sources including, but not limited to, IBM PureData System for Analytics, Db2 products, IBM PureData System for Operational Analytics, Teradata and Oracle. This allows users to send distributed requests to multiple data sources within a single SQL statement. For environments using IBM PureData System for Analytics as the data warehouse, the powerful combination of IBM Fluid Query and IBM Big SQL delivers the ultimate in flexibility in data access between data stores in the Logical Data Warehouse.

IBM Fluid Query use cases

1. **Use Hadoop as a "Day 0" archive for data discovery, analytics and exploration with the IBM Fluid Query connection enabling data movement to and from IBM PureData System for Analytics.**
2. **Access structured data from familiar sources like Db2 products, Oracle, Teradata, Microsoft and other IBM PureData System for Analytics appliances.**
3. **Run multi-temperature queries by combining "hot" data located on IBM PureData System for Analytics with "cooler" data on Hadoop.**
4. **Move data from IBM PureData System for Analytics to Hadoop for capacity relief, exploratory analytics, database backup or disaster recovery as an alternative to a "Day 0" Hadoop archive.**
5. **Move archive data to Hadoop using IBM Fluid Query for query while leveraging IBM Big SQL or Hive to locally query the data on Hadoop.**

IBM Fluid Query specifications

Supported Hadoop providers Apache Hadoop providers and Apache Spark 1.2.1, 1.3 and 1.4	
Cloudera 4.7, 5.3, 5.3.5, 5.5.1	**Hortonworks 2.1, 2.2, 2.3, 2.5**

Supported relational database management systems

IBM PureData System for Analytics N1001, N2002, N3001, Db2 10.1, Db2 10.5, Db2 10 for IBM z/OS®, Db2 Warehouse on Cloud, IBM PureData System for Operational Analytics, Informix, Oracle 11g, 11g, Release 12c, Teradata, Microsoft SQL Server, PostgreSQL and MySQL

Minimum system requirements

System	**Software**
IBM PureData System for Analytics N1001	NPS 7.0.2 and Netezza Analytics 2.5*
IBM PureData System for Analytics N1001	NPS 7.0.4 and Netezza Analytics 2.5.4*
IBM PureData System for Analytics N2002	NPS 7.1 and Netezza Analytics 3.0*
IBM PureData System for Analytics N3001	NPS 7.2 and Netezza Analytics 3.02*

** IBM Netezza Analytics 3.2.1 is required to work with the latest Hadoop distributions supporting Java 1.7 or later.*

Conclusion

IBM Fluid Query powers the Logical Data Warehouse, giving users the ability to combine numerous types of data from various sources in a fast, agile manner to drive analytics and deeper insight, without having to understand how to connect multiple data stores, use different syntaxes or APIs, or change their application.

Why IBM?

IBM offers a complete set of capabilities to enable today's Logical Data Warehouse. IBM is a strategic advisor that can help clients:

- *Innovate by effectively addressing a broad and ever-evolving set of data management needs*. Through our strategic acquisition strategy and organic development at IBM Research, we have amassed true best-in-class data warehousing solutions that address virtually every information need.
- *Boost the success of their data warehouse initiatives*. IBM combines proven innovations, a sharp focus on integration and world-renowned industry experts to help ensure the success of data warehouse initiatives. Plus, IBM is an industry leader in providing a comprehensive solution portfolio that targets key aspects of data warehousing—including data integration, governance and security.
- *Modernize your data warehouse to fuel real-time business*. IBM's commitment to innovation focuses on designing the right mix of data platforms and integration capabilities for our clients' changing business requirements.

About IBM PureData System for Analytics

IBM PureData System for Analytics, powered by Netezza technology, integrates database, server and storage into a single, easy-to-manage appliance that requires minimal setup and ongoing administration while producing faster and more consistent analytics performance. IBM PureData System for Analytics dramatically simplifies business analytics by consolidating all analytics activity in the appliance, right where the data resides, for industry-leading performance.

Visit: **ibm.com**/software/data/puredata/analytics to see how our family of expert integrated systems eliminates complexity at every step and helps you drive true business value for your organization.

IBM offers a variety of services for our IBM PureData System for Analytics clients. There are services offerings which accelerate and define your vision and strategy, and help you implement your solutions, maximize performance and ensure operational efficiency. Other services can help enhance your business effectiveness and alignment to increase enterprise skills and adoption. Learn more at **ibm.com**/software/data/services/dw.html

About data warehousing and analytics solutions from IBM

IBM provides the broadest and most comprehensive portfolio of data warehousing, information management, and business analytic software, hardware and solutions to help customers maximize the value of their information assets and discover new insights to make better, faster decisions to optimize their business outcomes.

For more information

To learn more about IBM Fluid Query please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/software/data/puredata/analytics

© Copyright IBM Corporation 2017

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
July 2017

IBM, the IBM logo, ibm.com, Db2, Informix and IBM PureData are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Netezza is a trademark of IBM International Group B.V., an IBM Company.

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 MERCHANT-ABILITY, 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.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

Please Recycle

WAS12388-USEN-04