IDC PERSPECTIVE

Choosing the Right Database Technology in the Age of Digital Transformation

Carl W. Olofson

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: Choosing the Right Database Technology

This IDC Perspective considers the dizzying range of options facing most users looking to build or implement new data applications addressing new workloads, modernize their existing workloads, or introduce new data-driven capabilities into the enterprise. It examines the data landscape, showing where the various types of workloads and technologies fit, and provides a list of example vendors and products that address those various needs.

Key Takeaways

- There are distinct workloads emerging as the IT world evolves and digital transformation progresses.
- Each workload demands specific qualities in a data management system, and those qualities are further refined by the specific requirements of the user.
- While many data management products can address most of the workloads under consideration, most are excellent at only one or a few.
- Identifying the right data management technology for each workload can be a challenge.

Recommended Actions

- Clearly define the workloads to be addressed, and identify their key requirements.
- Look for data management products that are aimed at those workloads, and highlight those requirements, not just the ones that provide such functionality in a "grab bag of stuff."
- Narrow down the list to a few key providers, taking into account the history, reputation, and business viability of each provider.
- Consider each provider's list of partners to determine how well they round out the overall solution.
- Research the availability of talent that can support the proposed technology solution.

Source: IDC, 2019

June 2019, IDC #US45179219

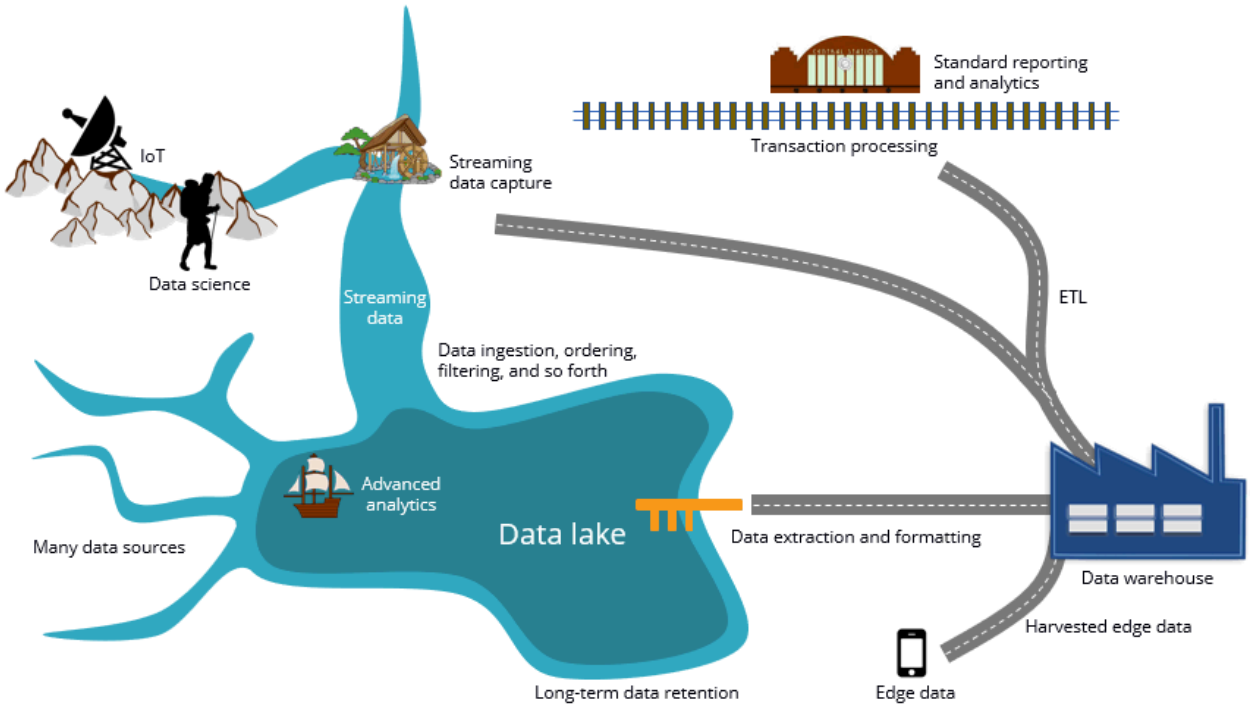
SITUATION OVERVIEW

The data management landscape can be very confusing. So many vendors, with very different technologies, tend to make competing claims in a wide variety of use cases. This makes it hard to determine which technologies address which problem areas the best. While a wide range of technologies can be used to do a variety of jobs at a simple level, specialized technologies are usually best when dealing with tasks that require specific functionality operating with a high degree of precision on large volumes of data, and that feature scalability, data structure flexibility, integrity, or some combination of these.

Figure 2 illustrates the data management landscape in question, where we see IoT data being collected, streaming data captured or simply helping fill a data lake that has many sources, extracting and formatting some of that data for a structured data warehouse, and transporting other data to the warehouse via ETL. The data lake performs data ingestion, ordering, filtering, and other data quality tasks; supports some advanced analytics; and is also used for long-term data retention. Data science explores the landscape, collecting data and building analytic models. Well-defined data follows a straight path through transaction processing, which feeds standard reporting and analytics. Edge data is managed on its own in an operational context, but elements are later harvested for analytics.

FIGURE 2

The Data Management Landscape



Source: IDC, 2019

Figure 3 lists the various types of data usage patterns, along with an explanation of their uses, the technologies involved, and examples of vendors' software products and services appropriate to each type. Note that these are only examples, not exhaustive lists, and there are quite a few other products that are appropriate to each type. Appearance of a vendor or product name in this list does not represent either endorsement or recommendation by IDC. The absence of a vendor or product from this list does not imply that such a vendor or product is any less qualified for the type and use indicated than those that are listed.

FIGURE 3

Data Management Types and Technologies

Type	Uses	Technologies	Examples
Edge data IoT	Manage and share operational data and support Agile development.	NoSQL: key-value stores and document databases	MongoDB, Redis, Azure Cosmos DB, Couchbase, InfluxData, IBM Cloudant
Data scie	Explore data, find patterns and connections, and build analytic models.	Curated data in Hadoop (HDFS) or object storage with Spark, Presto, and so forth	Databricks, Cloudera Data Science Workbench, Qubole, Starburst, Anaconda, MapR
Streaming data capture	Respond to streaming internal or external events "in the moment."	IMDB both relational and NoSQL; scalable SQL	MemSQL, VoltDB, eXtremeDB, SAP HANA, SQLStream, RethinkDB
Advanced analytics	Perform deep or specialized analytics (e.g., complex queries, pattern recognition).	Highly scalable RDBMS, graph DBMS, DM with time series support	Neo4j, TigerGraph, Kx, Riak TS, Kinetica, OmniSci
Data lake	Collect data, then define, filter, dedup, sort, and format it. Also, collect long-term data.	Hadoop and affiliated open source	Cloudera, Hortonworks, Google Cloud Dataproc, Amazon EMR, Azure HDInsight
Data warehouse	Manage highly structured and well-understood data for strategic business analysis.	RDBMS optimized for size and complex SQL analytics	Teradata, Amazon RedShift, IBM Db2, Oracle Database, Snowflake, Yellowbrick, Microsoft Azure SQL Data Warehouse
Transaction processing and analytics	Manage data of record with recoverability and analytic capability.	Databases supporting schematic data and atomicity, consistency, isolation, and durability (ACID) transactions	Oracle Database, Microsoft Azure SQL Database, Amazon Aurora, Splice Machine, IBM Db2, SAP HANA, MariaDB, EDB Postgres, Percona, InterSystems Caché, MarkLogic

Source: IDC, 2019

The technologies referenced in Figure 3 are as follows:

- Data lakes, including Hadoop and its affiliated Apache projects as well as other technologies, are used to collect and organize data and perform certain classes of analytics, often using Spark or Presto.
- NoSQL stores include document databases, key-value stores, and graph databases:
  - Key-value stores are usually used for application or session state management, where data sharing across applications is not necessary.
  - Document databases are used similarly as key-value stores, but where data sharing is required, and also for easing search and query functions. Those that support atomicity,

consistency, isolation, and durability (ACID) transactions may be used to manage data of record, such as financial transactions.

- Graph databases support what data modelers call "recursive relationships," which are relationships between objects of the same type. They are used to discover relationship patterns and deep indirect relationships between entities.

- Relational database management systems (RDBMSs) store data according to relational set theory as relations (tables) consisting in sets of tuples (rows), each having attributes (columns). These RDMSs can be classified as transactional, analytic-transactional, analytical, and deep analytical:
  - Transactional RDBMSs perform operational roles, managing data using transactions having ACID properties.
  - Analytic-transactional RDBMSs also perform operational roles using ACID transactions, but they support complex analytic queries as well.
  - Analytical databases are specialized for performing complex analytic queries.
  - Deep analytical databases support complex analytic queries on very large databases and often also feature extreme scalability.

The vendors and products mentioned in Figure 3 are as follows:

- **Amazon Web Services (AWS)** offers a data lake platform based on Hadoop called AWS EMR, a set of relational database services for transaction processing called Aurora, and a data warehouse platform called RedShift. AWS also offers a document data management system called DynamoDB and another called DocumentDB that emulates MongoDB.
- **Anaconda** is an open source data science platform that is also available in a commercial version called Anaconda Enterprise.
- **Cloudera** offers a commercial distribution of Hadoop and associated open source offerings in various packaged configurations, one of which is the Data Science Workbench.
- **Databricks** offers a Spark-based analytic platform as a managed cloud service.
- **EnterpriseDB** offers a commercial package for enterprise relational data management based on the open source PostgreSQL called EDB Postgres.
- **eXtremeDB** is a product from McObject designed for high-speed data ingest and analytics and is often used for streaming data management.
- **Google** offers a data lake product called Dataproc on the Google Cloud Platform (GCP). The GCP also features a highly scalable RDBMS called Google Cloud Spanner.
- **Hortonworks**, like Cloudera, offers various platform configurations of Hadoop and related technologies. Cloudera and Hortonworks are in the process of merging.
- **IBM** offers the RDBMS Db2 for both analytic-transactional and data warehouse workloads. The company also offers the document data management system Cloudant as a managed cloud service.
- **InfluxData** is maker of InfluxDB, which is a database management system (DBMS) designed to handle events with native time series support.
- **InterSystems** is maker of Caché, which is a multimodel DBMS with an object-oriented services layer and rich scripting language that supports ACID transactions and complex objects.
- **Kinetica** is a maker of a high-speed RDBMS powered by GPUs that can be used for advanced analytics and also ACID transaction processing.
- **Kx** provides a high-speed time series DBMS.

- **MapR** offers a commercial package based on Hadoop and associated technology and also provides an indexed file system as an alternative to HDFS, with database layers that support random retrieval and update and complex SQL queries.
- **MariaDB** is a provider of commercial RDBMS product packages based on MySQL and PostgreSQL.
- **MarkLogic** is a maker of a multimodel DBMS based on a document data management system that can be used with or without a schema and supports ACID transactions, complex queries, and complex data objects.
- **MemSQL** is a provider of an all in-memory RDBMS that can be used to capture and process streaming data.
- **Microsoft** offers a document-based multimodel data management system called Azure Cosmos DB, a data lake platform based on Hadoop called Azure HDInsight, a transactional RDBMS called Azure SQL Database, and a data warehouse specialization of that product called Microsoft Azure SQL Data Warehouse.
- **MongoDB** offers a commercial package based on its open source document data management system and provides multiple document and ACID transaction support.
- **Neo4j** offers a commercial package based on an open source graph database that boasts deep and complex graph analysis.
- **OmniSci** provides a commercial packaged version of its open source RDBMS, which is powered by GPUs for extreme speed analytic processing.
- **Oracle** offers a supported package of the open source RDBMS MySQL, as well as its flagship product, Oracle Database. Oracle Database is available as a software-only product or as a package within its bespoke hardware platform, Oracle Exadata. It is also available as a managed cloud service in the Oracle Cloud, including self-managing, self-tuning, and self-repairing versions called Oracle Autonomous Database. Oracle Database can be used as a analytic-transactional database system and as a data warehouse.
- **Percona** offers commercial packages based on both MySQL and PostgreSQL.
- **Qubole** offers a complex analytic environment as a managed cloud service.
- **Redis Labs** offers the open source key-value store Redis as a managed cloud service.
- **RethinkDB** provides a high-speed JSON document data system offered as a managed cloud service.
- **Riak** provides a managed service for the open source key-value store Riak and also an optimized version for time series called Riak TS.
- **SAP** provides an in-memory compressed columnar RDBMS called SAP HANA, optimized for vector processing, aimed at both fast analytics and ACID transaction processing.
- **Snowflake** provides a data warehouse platform offered as a managed cloud service.
- **Splice Machine** is a highly flexible and scalable RDBMS that runs on a Hadoop foundation, offering both analytic and ACID transaction support.
- **SQLStream** is a SQL-based streaming data event capture for real-time streaming data analytics. SQLStream is a subsidiary of Guavus Inc.
- **Starburst** offers a managed cloud service for analyzing large data collections using Presto.
- **Teradata** offers a shared-nothing clustered RDBMS designed for the most demanding enterprise data warehouse workloads, delivered as a database platform called Vantage, both in various on-premises configurations and as a cloud service.

- VoltDB offers a commercial package based on its own open source all in-memory RDBMS that supports ACID queries and also high-speed data ingest.
- Yellowbrick offers a data warehouse platform designed for high-speed ingest and query execution.

ADVICE FOR THE TECHNOLOGY BUYER

As may be seen, the variety of available technologies for data management is truly mind-boggling. The most important thing is to identify the problems to be solved, find the best technologies to address those problems, and not to be obsessed with the desire to manage all data on the same platform.

The following specific actions are recommended:

- Identify the jobs to be done and the types of data involved. Does the data require a schema? Is Agile development important? If so, a NoSQL data management system, such as a document system or key-value store, may be appropriate. If data query and analytics need to be done with the live data, a schematic DBMS may be necessary.
- Identify other requirement parameters, such as query speed, throughput, scalability, geographic distribution, availability, recoverability, and ACID transaction support. These will help narrow the list.
- Look within your organization as well as the available labor pool to make sure that there is sufficient talent available to manage the technology that you are considering.
- Consider your existing vendor relationship. If it is solid, you may want to give that vendor extra consideration.
- Look at the partners that work with the vendor under consideration. Make sure they have experience in helping build the kind of solution that you are looking for.
- Make sure that whichever vendor you choose has a product road map that fits with your future plans. This includes things such as adding further functionality and migration to the cloud.
- Make a short list, and do some sort of proof of concept before settling on one vendor. Make sure your team is comfortable with the choice.

LEARN MORE

Related Research

- *Data Management Lessons Learned at Google Cloud Next '19: Data Management Systems Perspective* (IDC #US45037819, May 2019)
- *MongoDB Acquires Realm, Greatly Expanding Its Mobile Footprint* (IDC #IcUS45035819, April 2019)
- *A Unified Approach to Driving Value Using Data for Digital Transformation: Lessons Learned from IBM Think* (IDC #US44887919, March 2019)
- *Significant Data Management Announcements in 4Q18* (IDC #US44514518, December 2018)
- *IDC TechBrief: Enterprise Data Platforms* (IDC #US44493518, December 2018)

Synopsis

This IDC Perspective considers the dizzying range of options facing most users looking to build or implement new data applications addressing new workloads, modernize their existing workloads, or introduce new data-driven capabilities into the enterprise. It examines the data landscape, showing where the various types of workloads and technologies fit, and provides a list of example vendors and products that address those various needs.

"The data management landscape can be very confusing. So many vendors, with very different technologies, tend to make competing claims in a wide variety of use cases," says Carl Olofson, research vice president for Data Management Software research at IDC. "While a wide range of technologies can be used to do a variety of jobs at a simple level, specialized technologies are usually best when dealing with tasks that require specific functionality operating with a high degree of precision on large volumes of data and that feature scalability, data structure flexibility, integrity, or some combination of these."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com