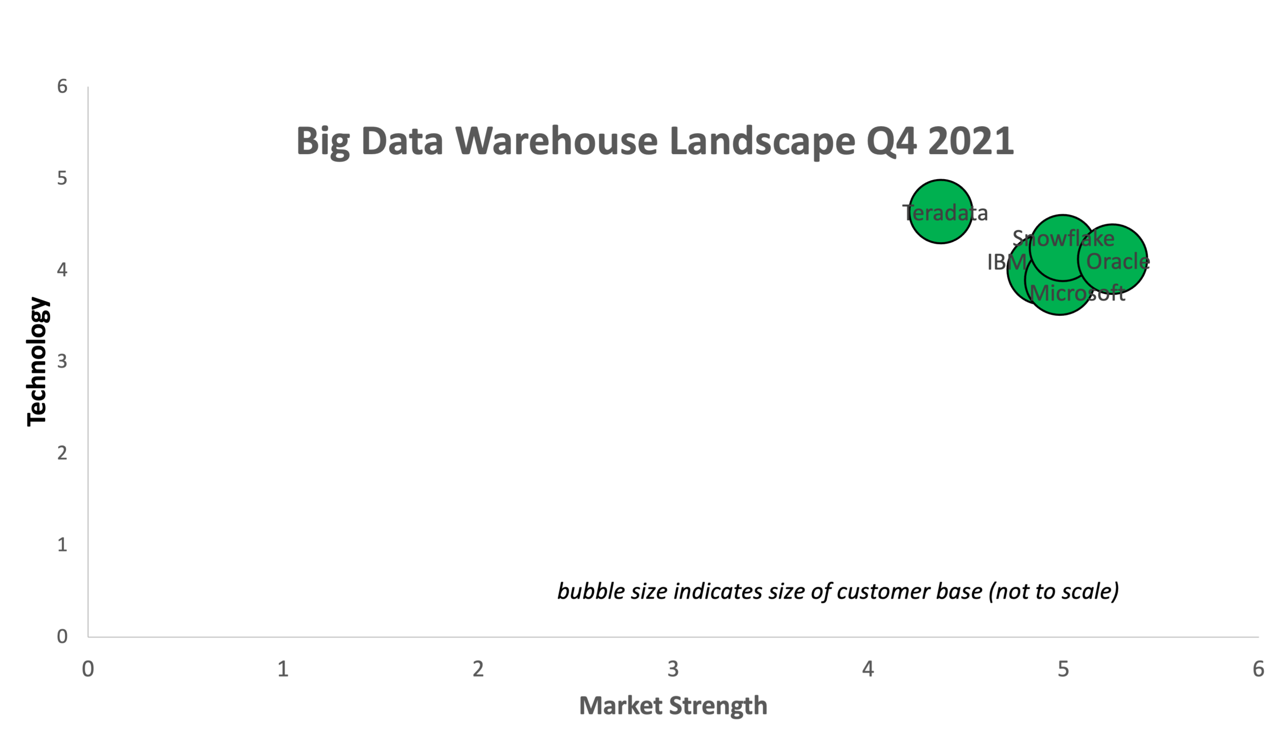
Big Data Warehouse Landscape Q4 2021

Every organisation needs to be able to track and monitor its business performance. The precise question depends on the industry, but most companies need to know what are their most (and least) profitable products, channels and customers, which of their locations (or website pages) generate the highest sales, and which assets are in need of maintenance. Gathering this information is hard, partly because of inconsistent of definitions and partly because of sheer data volumes. The subsidiaries of a multinational may classify their products differently based on the peculiarities of the local market, while mergers and acquisitions mean that it is hard to maintain consistently business rules, classification hierarchies and cost allocation rules. Some important data is stored outside the enterprise, e.g. by suppliers, government agencies, credit rating agencies and more. Data volumes have always been an issue for some industries, such as retail banking, Telco and retail, and more recently digital industries that have to track on-line sales and traffic. Data warehouses are databases where data is gathered together from multiple sources into a consistent form for the purposes of business analysis. The concept has been around since the 1980s, but the challenges of how to manage a data warehouse with regards to data quality, consistency and performance are constantly changing. As the volume of data that we store has grown then so data warehouses have grown greatly in size to reflect this. In 2003 the largest data warehouse in the world was 30 TB in size, yet just a decade later there were examples of petabyte sized data warehouses, a 30-fold increase in a decade. This is a trend that has continued relentlessly, with for example the data warehouse of taxi company Uber weighing in at over 100 petabytes by 2018.

Data warehouses were traditionally row-oriented in their design and mostly relational (SQL based). This is good for update and insert speed but less so for wide-ranging queries, so data warehouses increasingly adapted to use columnar databases, which have many advantages for analytic processing, albeit at the price of the speed of inserting and updating data. In order to deal with larger scale, multi parallel processing evolved, allowing data and query processing to be shared across multiple processing nodes and physical storage mechanisms. In recent years further different database constructs have emerged, with NoSQL (“not only SQL”) databases including graph databases, document databases and more. The rise of “big data’ file systems (Hadoop, Spark) adds a further level of complexity and size to data sources that a data warehouse design has to consider. Data warehouses today have to deal with traditional numeric data but also a wider range of data types and sources, such as text, images, video, time series data and sensor data. Architectures have adapted to spin off “data marts” from a corporate data warehouse, and more recently we see “data lakes” of big data sitting alongside, and potentially acting as feeds into, data warehouses.

For some time, the data warehouse market was mostly restricted to the large database vendors (Oracle, IBM, Microsoft) plus some specialist companies like Teradata. However, over the last few years things have become much livelier. Snowflake made great strides in the market by providing a native cloud-based data warehouse. Cloud data warehouses offer the promise of almost limitless processing power and storage, provided you have the budget to pay for this. The cloud vendors themselves have their own data warehouse solutions, such as Amazon Redshift and Google BigQuery, and newer competitors have emerged that also offer parallelised cloud-based data warehouses. Across industries there is greater use of digital data, whether that be cars or airplanes feeding performance data to maintenance centres, smart meters measuring electricity usage, physical equipment in factories sending status reports, medical equipment monitoring patient health, or streaming TV services trying to predict what new show you will like. All these new applications add to the volume of data that needs to be handled, and further challenges data warehouse architects. This demand has stimulated new and innovative products, and has led to industry growth of 8-12%, a rate virtually unaffected by the coronavirus pandemic.

The major vendors in the market are summarised in the diagram below.



The landscape diagram represents the market in three dimensions. The size of the bubble represents the customer base of the vendor, i.e. the number of corporations it has sold data warehouse software to, adjusted for deal size. The larger the bubble, the broader the customer base, though this is not to scale. The technology score is made up of a weighted set of scores derived from: customer satisfaction as measured by a survey of reference customers[1], analyst impression of the technology, maturity of the technology in terms of its time in the market and the breadth of the technology in terms of its coverage against our functionality model. Market strength is made up of a weighted set of scores derived from: data warehouse revenue, growth, financial strength, size of partner ecosystem, customer base (revenue adjusted) and geographic coverage. The Information Difference maintains vendor profiles that go into more detail. Customers are encouraged to carefully look at their own specific requirements rather than high-level assessments such as the Landscape diagram when assessing their needs.

A significant part of the "technology" dimension scoring is assigned to customer satisfaction, as determined by a survey of vendor customers. In this annual research cycle the vendors with the happiest customers were Teradata, followed by Magnitude. Our congratulations to them.

[1] In the absence of sufficient completed references, a neutral score was assigned to this factor.

Below is a list of the significant data warehouse vendors.

Vendor	Brief Description	Website
Actian	Actian's product is an analytic database on commodity hardware.	www.actian.com
Alibaba	The Oceanbase distributed cloud-based data warehouse is Alibaba's data warehouse offering.	www.alibabacloud.com/product/oceanbase
Amazon Redshift	Cloud-based data warehouse solution.	aws.amazon.com/redshift/
Cloudera	Enterprise cloud vendor incorporating Hortonworks, with a data warehouse offering.	www.cloudera.com
Databricks	Data "lakehouse" vendor.	databricks.com
Exasol	German data warehouse appliance vendor.	www.exasol.com
Greenplum	Appliance vendor aiming at high-end warehouses, now part of Pivotal, a subsidiary of EMC, itself acquired by Dell in 2015.	pivotal.io/big-data/pivotal-greenplum
HPCC	An open-source, massively parallel platform for big data processing, developed by LexisNexis Risk Solutions.	hpccsystems.com
IBM	DB2 is the data warehouse software offering from the industry giant, now available on cloud as well as on-premise.	www.ibm.com
InfoBright	Provides a columnar-database analytics platform.	www.infobright.com
jSonar	Boston-based NoSQL data warehouse vendor.	www.jsonar.com
Magnitude	Part of Magnitude Software, Kalido is an application to automate building and maintaining data warehouses.	magnitude.com
MarkLogic	Enterprise NoSQL database vendor.	www.marklogic.com
Microsoft	As well as its SQL Server relational database, Microsoft acquired Data Allegro and at the end of 2010 launched its Parallel Warehouse based on this technology.	www.microsoft.com
MonetDB	MonetDB is an open-source columnar database system for high-performance applications.	monetdb.cwi.nl
Neo4j	Open source graph database.	www.neo4j.org
Oracle	Database and applications giant with its own data warehouse offering.	www.oracle.com
ParStream	Columnar, in-memory, MPP database vendor aimed at analytic processing.	www.parstream.com
Pivotal	Owners of the Greenplum massively parallel data warehouse solution, now an open-source solution.	pivotal.io/big-data/pivotal-greenplum
Qubole	Markets the Qubole Data Service, which accelerates analytics workloads working on data stored in cloud databases.	www.qubole.com

Sand	Focuses on allowing customers to-effectively retain massive amounts of compressed data in a near-line repository for extended periods.	www.sand.com
SAP/Sybase	Sybase was a pioneer in column-oriented analytic database technology, acquired in mid-2010 by giant SAP. SAP also offers the in-memory database technology HANA.	www.sap.com
SAS Institute	Comprehensive data warehouse technology from the largest privately-owned software company in the world.	www.sas.com
Snowflake	Cloud-only data warehouse vendor.	www.snowflake.com
1010 Data	Provides column-oriented database and web-based data analysis platform.	www.1010data.com
Teradata	Teradata positions itself as a connected multi-cloud data platform company.	www.teradata.com
Vertica	Data warehouse appliance vendor Vertica was purchased by HP in 2011.	www.vertica.com
XtremeData	US vendor that provides highly scalable cloud database platform.	www.xtremedata.com
WhereScape	Not an appliance, but a framework for the development and support of data warehouses.	www.wherescape.com
Yellowbrick	Cloud data warehouse vendor.	www.yellowbrick.com