REPORT REPRINT

451 Perspective: The key components of an abstracted data architecture

JUNE 09 2020

By Matt Aslett

Abstracted data architecture has evolved as enterprises have exploited the separation of compute and storage by enabling the analysis of data in cloud-native storage layers. This report highlights the key components required to make that architecture a reality.

451 Research®
Now a Part of
S&P Global Market Intelligence

Introduction

The separation of compute and storage is an ongoing trend in enterprise data and analytics. Back in 2016 we noted that, as a growing volume of data was originated and stored in the cloud, there was growing interest in the ability to analyze data in cloud storage services without first extracting it into a data warehouse. In 2018 we described this as the evolution of an abstracted data architecture, noting that this abstracted data architecture was the logical continuation of the separation of compute and storage, as well as the use of hybrid cloud/on-premises infrastructure and multi-region distributed data management. In this report, we take a closer look at the components required to make this abstracted data architecture a reality, including the evolution of distributed query engines, OLAP cube systems, semantic modeling, federated query and data virtualization.

451 TAKE

There are multiple components that could be part of an abstracted data architecture, depending on the extent to which it is being adopted by the individual enterprises. While enterprise knowledge graph, stand-alone data virtualization products or a data warehouse designed to analyze data in cloud storage should be considered for an enterprise-wide initiative, they would likely be considered overkill for a small-scale trial or departmental project. The latter could likely be fulfilled by a distributed query engine, especially if equipped with query federation functionality to enable querying of data outside the cloud, while the analytics acceleration and semantic modeling functionality provided by OLAP-on-cloud and cloud data-lake analytics offerings are likely to prove attractive as adoption expands along with the number of queries and concurrent users and use cases

The separation of compute and storage

A prerequisite for an abstracted data architecture is that it be designed to take advantage of the separation of compute and storage. As we previously described, this is an approach that is designed to maximize the use of low-cost storage services by spinning up separate compute engine services to analyze the data without moving it all into a separate data warehouse.

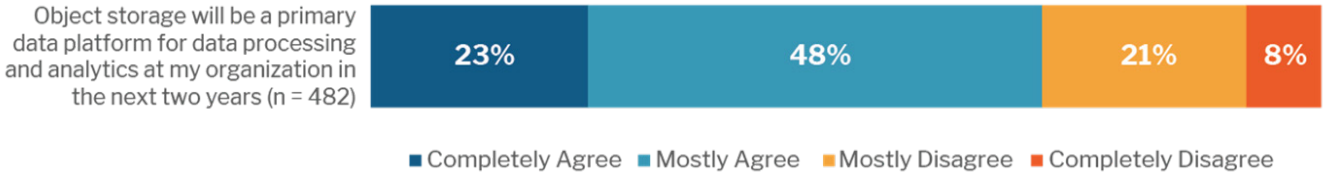
The approach of separating compute and storage was pioneered by Google with its Dremel research project, which provided an ad hoc SQL query system to analyze data stored in Google's Colossus file system (the successor to Google File System). This Dremel project was later commercialized as Google BigQuery, and inspired the emergence of distributed query engines that enabled enterprises to bring their existing SQL analysis skills and tools to data stored in Hadoop (initially), as well as cloud storage services and on-premises object storage.

Although this separation of compute and storage is closely associated with the adoption of cloud storage services, it is important to recognize that it is not limited to the cloud, and is also supported in on-premises environments through the deployment of cloud-native architecture.

451 Research first highlighted this architectural shift toward separation of compute and storage in 2016, and since then has seen the trend accelerate, to the extent that 71% of respondents to our Voice of the Enterprise: Data & Analytics 2H 2019 agreed either completely or mostly that object storage will be a primary data platform for data processing and analytics at their organization in the next two years.

Object Storage as a Primary Platform for Data Processing and Analytics

Source: 451 Research's Voice of the Enterprise: Data & Analytics 2H 2019



Analytics compute engine

The second prerequisite for an abstracted data architecture is the ability to spin up compute engine services with which to analyze the data in the separate storage layer. That analysis could be performed via a Hadoop/Spark service spun up to run against raw data in cloud storage, such as Amazon EMR, Azure Databricks or Google Cloud Dataproc.

Alternatively, and particularly for SQL-based analytics, it could be via a data-warehousing interface providing access to curated data sets in the cloud storage layer, such as Google BigQuery, Snowflake, Actian Avalanche, Yellowbrick, Cloudera Data Warehouse, Teradata Vantage or Amazon Redshift Spectrum.

A third choice would be an interactive query service designed to translate SQL queries to run against raw data on cloud storage services. Potential options include the Presto open source distributed query engine (including Starburst's Enterprise Presto distribution, the Amazon Athena Presto-based interactive query service and Qubole Managed Presto), as well as Apache Impala (part of Cloudera Data Platform), Apache Arrow and Microsoft Azure Data Lake Analytics.

There are multiple options for querying data in cloud storage, including Hadoop/Spark services, data warehouse services and SQL query engines. The options do not end there, however. Depending on the performance requirements, enterprises may also look at the potential to accelerate interactive analytics through the creation of OLAP cubes or virtual data sets.

Analytics acceleration and semantic modeling

The performance considerations include not just the speed requirements for any individual query, but also those that come from large-scale deployments with multiple concurrent users and queries. Key players in this space, including AtScale, Kyligence and Kyvos Insights, emerged from what was once known as the OLAP-on-Hadoop category, but has evolved into OLAP-on-cloud as increasing volumes of data are stored in cloud storage rather than the Hadoop Distributed File System. Additionally, Dremio argues that its approach of creating virtual datasets achieves the same goal while avoiding the need to create OLAP cubes.

In addition to accelerating known queries by preprocessing the required data, the likes of AtScale, Kyligence, Kyvos Insights and Dremio also provide functionality for creating a semantic model. This provides advantages for large-scale deployments that cannot be matched by the use of SQL query engines, by providing a layer for creating standard business definitions that can be applied across all data, even if it is accessed using multiple different business intelligence and virtualization tools for multiple use cases (potentially acting as a searchable catalog of available data assets).

Although an increasing volume of data is stored in cloud and distributed object storage, not all data is in cloud storage. Data from 451 Research's Voice of the Enterprise: Data & Analytics, Data Platforms 1H20 survey highlights that the majority of existing database and data-warehousing workloads are currently running on on-premises, noncloud infrastructure.

As hybrid IT architecture becomes the norm, the latest survey also highlighted that one-quarter of respondents expect to retain existing database workloads on-premises – either unchanged or modernized using cloud-native frameworks. The ability to access that data and federate queries across multiple data locations without having to combine all related data in a single environment is another important aspect of an abstracted data architecture.

Federated query and data virtualization

The ability to federate queries across multiple data sources is available, to a degree, with all the options that deliver the abstracted data architecture. For example, Google BigQuery enables federated querying of data in other Google Cloud databases, as well as Apache ORC and Parquet files in cloud storage, while the U-SQL language enables Azure Data Lake Analytics users to query data in relational data stores as well as Azure Data Lake.

Similarly, Teradata supports query federation via Vantage, while AWS also provides support for federated querying in both Amazon Redshift and Amazon Athena. The latter takes advantage of Presto's federated query functionality, which is similarly supported by Qubole's Managed Presto, as well as Starburst's Enterprise Presto distribution.

Query federation is also available at the analytics acceleration layer. For example, AtScale has added query federation to its offering by federating its adaptive cache functionality, allowing users to create OLAP cubes to query and combine data from multiple separate databases, Hadoop or cloud storage environments without moving the data from one environment to the other.

Similarly, while Dremio's Data Lake Engine is primarily positioned to enable enterprises to accelerate analytics on data stored in data lake storage and a variety of Hadoop cloud services, it also enables data from relational databases to be joined with data in data lake storage, as well as the ability to push queries to relational databases.

When it comes to joining data from multiple sources, another option comes in the form of stand-alone data virtualization products and services delivered by the likes of Denodo, TIBCO, Red Hat and Informatica. These possess the functionality required for strategic adoption across an enterprise that goes beyond query federation, with the addition of security, modeling capabilities and cataloging, for example.

Other potential components/alternative approaches

The above list of components that could be used as part of an abstracted data architecture is extensive, but not necessarily exhaustive. There are many ways to crack a nut, and many ways to make an abstracted data architecture a reality.

For example, Alluxio's data orchestration platform doesn't neatly fit into any of the categories above, but is used by enterprises to accelerate analytics in conjunction with Presto or Apache Spark, and to orchestrate the separate compute and storage layers of an abstracted data architecture in both single and multicloud deployments.

Molecula is another potential competitor with its Cloud Data Access platform based on the Pilosa distributed bitmap index, which stores a mathematical representation of the source data to enable access and analytic acceleration without pre-aggregating, federating, copying, caching or moving source data.

Additionally, if an enterprise is specifically looking to deliver a single semantic model to all data in an abstracted data architecture, then it might also consider the various products and services designed to enable enterprise knowledge graphs that abstract underlying data stores and provide a graph-based map of all available data across the organization, as well as the relationships between that data.