REPORT REPRINT

Coverage Initiation: Molecula seeks to simplify big-data infrastructure with Cloud Data Access Platform

MARCH 2 2020

By Paige Bartley

Data access and integration become more complex with escalating data volume, yet prevailing methods for integration frequently require creating duplicate data copies that further exacerbate the underlying problem. Molecula, born out of the Pilosa open source project, seeks to solve big-data virtualization with a novel approach that eliminates the need to pre-aggregate, federate, copy, cache or move source data.

Introduction

Many approaches to data integration and aggregation can, paradoxically, exacerbate data management problems by creating duplicate data copies in order to move or situate data into a staging area, such as a data warehouse, where it can be more easily controlled and accessed for insight initiatives such as business intelligence or analytics. Not only is this time-consuming and compute-intensive when large volumes of data are involved, but it creates inherent risk, particularly for regulated firms where each additional copy straying from the original source represents a new piece of data somewhere else in the IT stack that now must be adequately managed for controls such as permissions and privacy.

Molecula, based on the Pilosa open source project, is using a novel approach to data virtualization to effectively integrate data on-demand and in real time, without pre-aggregating, federating, copying, caching or moving source data. A bitmap indexing methodology stores a mathematical representation of the source data in question, without creating copies or moving the data itself, providing scale, performance and increased control.

451 TAKE

The IT environment that supports enterprise data insight initiatives can be thought of as a vertical stack, with databases and data persistence at the very bottom, and consumption methods such as visualization tools at the very top. It's frequently the middle layer – the plumbing that gets data from point A to point B – that spawns the most complexity. Various methods of shaping and staging data into meaningful datasets frequently require data copies and aggregation, sapping organizations of valuable time and multiplying data governance risk.

The ability to serve data into analytics engines and other environments, feeding directly off source data copies, is both a performance and governance concern. Molecula's approach is an attractive one, particularly for very large, complex enterprise IT environments, since it minimizes the tangle of products and data copies that are typically required in this environment to deliver the right data to the right place, at the right time.

Context

When we last wrote about Molecula's predecessor, Pilosa, it was still an open source project with no paid enterprise offering. That changed in May 2019 when Molecula Enterprise was launched: an offering based on the same foundation as Pilosa, but with additional IP and capabilities.

The company is led by CEO Higinio Maycotte, who started the Pilosa project following his earlier founding of Umbel, a data management platform that was originally designed primarily to serve the needs of the sports and entertainment market. From that heritage, Molecula Enterprise was designed to meet the requirements of especially large, complex and data-intensive environments and real-time use cases.

The relationship between the Pilosa open source project and Molecula remains a close one. The Pilosa community currently has approximately 2,000 members, and around 40% of Molecula Enterprise customers were originally users of open source Pilosa. This has provided a steady pipeline of business for the company.

Today, Molecula has a headcount of about 25 employees, with growth objectives to grow its staff to at least 65 by the end of 2020. The company is notable in its relatively short sales cycle and high success rate with POCs. Current enterprise sales cycle time sits at an average of approximately 100 days, and the company notes that it has yet to lose a deal in which it has participated in a POC.

The company is still early in its funding ambitions, with a $6m seed round currently under the its belt. However, a series A round is on the horizon, and the company is working to raise further capital. Business partners include Oracle, which has recognized Molecula as part of its startup accelerator program.

Technology

Molecula's flag in the mountain is its fresh take on data virtualization, which leverages the Pilosa project and the concept of virtual data sources (VDSs) to eliminate the need for data caching, data federation or data copies. Currently, the company has been granted nine patents, with 24 patents applied for total. Molecula Enterprise (now available as the 2.0 release) acts as a unified access layer for data, and is composed of four key components.

Pilosa open source project

Pilosa forms the core of Molecula Enterprise, being an open source, distributed bitmap index. Because Pilosa allows mathematical representations of source data to be made and stored in a highly compressed way, organizations can act on representations of source data rather than direct data copies.

Virtual data sources (VDSs)

VDSs provide the high-performance, masterless, distributed system for data representations to be accessed and leveraged by organizations. It takes only subseconds to get data from the source to a VDS, and once done, the process never needs to be repeated. Additional VDSs can easily be added when needed; 10 are included in the Molecula small enterprise license subscription package.

VDS Manager (VDSM)

VDS Manager is a management framework and wrapper for the Molecula ecosystem, used primarily to provide and manage the resources required to spawn and scale the VDSs. VDSM can also be used to manage access controls, plug-ins and the accompanying API. One VDSM comes with each package of Molecula.

Plug-ins

Molecula's plug-ins are what allow it to play nice with the tools on the ingestion and consumption ends of the data workflow, as well as provide additional capabilities for use cases such as compliance and model execution support. Open source and enterprise plug-in options allow for flexibility with organizations' existing IT assets and ecosystem.

Marketing

Molecula bases its messaging and positioning on three key themes: simplicity, acceleration and control. Simplicity is highlighted for making data easier to access, and in many cases, replacing multiple products and/or open source tooling. Acceleration puts emphasis on enabling shorter time to insight, since data that is easier to access – without duplicates or sprawl – is easier to derive value from.

In addition, because of the format in which data is represented in a VDS, queries can be extremely fast, essentially eliminating the latency that can often come with other approaches. And control, because the company slashes the typical large organization's footprint of data to be actively managed, helps organization minimize risk and achieve outcomes such as regulatory compliance.

In terms of end-user and target audiences, Molecula is primarily focused on the data engineer role, although it acknowledges that data scientists, analysts, CISOs and CDOs can be key influencers in the product procurement process. However, it is the data engineer that typically is tasked with the heavy lifting of creating the data pipelines that ingest, integrate and ultimately deliver quality data to other users downstream for purposes such as BI. Since these individuals are often tasked with stitching together a complex ecosystem of tools to achieve the data access step of the data workflow, they are the ones that most acutely feel the pain of cumbersome methods that generate data copies and have high latency.

Because the company has such a high success rate in POCs, a key part of the marketing strategy is simply getting to this stage. It's often Molecula Enterprise's ability to break through the latency 'floor' – via lossless compression and elimination of data copies – that tips the scale entirely in its favor, especially for prospects with very large and complex IT environments. While data control is an important aspect of what it does, it's still performance at scale that remains the foot in the door. Thus, marketing efforts focus on outreach to organizations that fit this large/complex profile, for example, large healthcare, technology and financial services organizations.

Partnerships also extend Molecula's marketing reach. As a small company targeting very large and complex organizations, credibility and name recognition can mean a lot in getting to the POC stage of the sales cycle. By partnering with enterprise-oriented technology firms such as Oracle (as part of the Oracle for Startups program), Molecula increases its visibility to a key target audience.

Competition

Given that Molecula itself leverages open source technology at its core, it should come as no surprise that a smattering of open source options dot the landscape for organizations leveraging a DIY approach to the data access part of the data pipeline. Some examples, which play various roles in data access, would include Apache Arrow, Apache Druid, Apache Kudu, Apache Parquet, Apache Zookeeper, Elastic and Presto. RedHat JBoss Data Virtualization is a popular data virtualization option as well.

Commercial competitors can best be broken largely into legacy or incumbent providers, focused on traditional data integration and virtualization, as well as newer providers that are taking varying approaches to data unification architecture. Some of the longtime providers include virtualization specialist Denodo, Informatica (with its PowerCenter offering) and Tibco. Newer approaches include AtScale, which specializes in query optimization and analytics performance for large data sets, as well as Dremio which aims to provide a self-service semantic layer for the data lake so that users can accelerate their analytics queries. Another virtualization specialist with similar positioning would be Gemini Data, which also espouses a zero-copy data approach.

Others that play in the loosely defined data fabric space can also theoretically overlap, in that they offer real-time and near real-time integration of data sources, helping support downstream outcomes such as self-service analytics and data science. Some examples would include K2View, which relies on architecture defined by so-called micro-databases that store data relevant to particular defined business entities, as well as Talend, which offers a diversity of data integration tooling, including lightweight options that facilitate rapid integration and business outcomes.

Because the core motive of Molecula is to drive faster and more efficient queries on large data sets, several other offerings in the market potentially overlap, particularly in the database space. In-memory databases focused on performance include Altibase, MemSQL, Pivotal (now part of VMware), SAP HANA and VoltDB, as well as some grid/cache vendors such as GigaSpaces, GridGain, Hazelcast, Redis Labs, ScaleOut Software and Software AG.

SWOT Analysis

STRENGTHS

Data copies don't just present a performance concern; they present a governance and compliance concern. Molecula's approach, which leverages mathematical representations of source data, massively shrinks the footprint of data that must be managed and speeds up query performance. This approach is particularly suited for very large, very complex data environments, as well as for organizations that face strict regulatory requirements for data privacy or protection.

WEAKNESSES

Molecula's differentiators are based on highly technical architectural qualities, which may be difficult to communicate in discussions with higher-level business roles who have the power to write checks. The target audience of data engineers acutely feels the pain of the business problem being addressed, but their power and voice varies. As data becomes the most important business asset, the company will need to hone its messaging with top-level decision-makers such as CEOs and CDOs.

OPPORTUNITIES

As regulations concerning data grow more complex and numerous, and real-time use cases become more common, organizations will find themselves painted into a corner with analytics and insight workflows that require sprawling estates of data duplicates. These organizations will be looking to not only meet compliance requirements, but also accelerate performance. For them, Molecula's promises of simplicity, acceleration and control will hold immense appeal.

THREATS

Relative to data integration incumbents and traditional data virtualization providers, Molecula lacks high-impact visibility. Its business relationships are of importance in building credibility, but if partners begin to perceive the company as a threat, it would be easy enough to cut ties. While Molecula considers its technology and IP its crown jewels and does not want them to fall into the wrong hands, it would not be unthinkable for a larger incumbent to purchase the firm just to eliminate competition.