Database Engineering Open Data Lakes and Out-of-the-Box Solutions What Will Be Cool in 2022

A closer look at three trends that will help enterprises get the most from their data.

- By Dipti Borkar
- December 10, 2021

Over the last year we've seen the rate of cloud adoption grow exponentially thanks in part to the cloud data warehouse, the cloud data lake, and the newer data lakehouse. In 2022 we'll see the next phase of growth in the cloud -- architecting for performance, the open data lake to augment the cloud data warehouse, and out-of-the-box cloud solutions to drive innovation.

For Further Reading:

Executive Q&A: A Closer Look at Open Data Lake Analytics

Executive Q&A: Data Lakehouses

I Have a Data Warehouse, Do I Need a Data Lake Too?

Trend #1: Database engineering is cool again!

The debate about the data warehouse versus the data lake seemed to headline the past year, with a pronounced movement to the data lake. Now, in 2022, it's time to make database engineering cool again -- on the data lake. That means the database benchmarking wars will be back in action.

As the performance of disaggregated query engines on data lakes meets or exceeds the performance of tightly coupled data warehouses, workloads will migrate to the data lake for lower cost and greater flexibility. Just look at Databricks' recent benchmark announcement on setting the official world record in 100TB TPC-DS performance benchmarking. Three key reasons they broke the record: open data formats (which allows for ecosystem development of data formats and standardization), a C++-based MPP architecture product for data lakes, and improved performance of both large and small queries.

In addition to performance, the flexibility and lower cost of disaggregated query engines will drive more workload migration from data warehouses to loosely coupled data lakes. As more innovation continues in the data lake space, community-driven open source query engines such as Presto will become more widely adopted for SQL on the data lake, especially as the next generation of Presto will be faster (rewritten in C++) and more scalable (MPP system). That, coupled with the fact that the data lake (AWS S3, etc.) is ubiquitous and extremely cheap, means we'll see a pronounced movement to a data lake architecture in 2022.

The database engineers who can build a data lake stack with data warehousing capabilities (transactions, security) but without the compromises (lock-in, cost) will win.

Trend #2: The rise of the open data lake for warehouse workloads

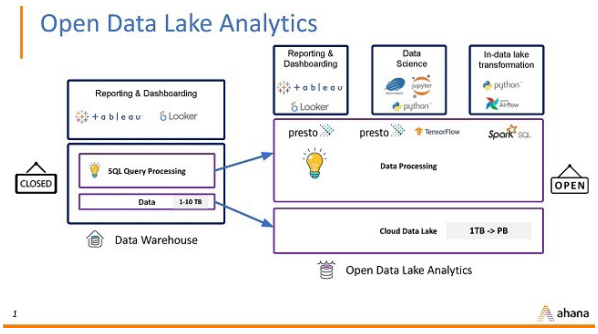
Some data warehouses are locking people into propri-

etary formats. As users start feeling the burden of higher costs as the size of their cloud data warehouse grows, they'll start looking for cheaper and open options that don't lock them into a proprietary format or technology. In 2022, it'll be all about the open data lake analytics stack, the stack that allows for open formats, open source, open cloud -- and no lock-in.

It's not news that more data is being generated than ever before, but enterprises now face a critical decision -- store their data in a proprietary database that can support their analytical workloads (but is complex, expensive, and requires lock-in) or dump their data into a data lake that has no analytics capabilities but is open, simple, and less costly. I talk to customers every day who want a data lake approach but want to run their warehouse workloads on that data lake.

In 2022, I believe we'll see a massive shift from the cloud data warehouse (CDW) to the cloud data lake because now you can run your warehouse workloads on the data lake.

The open data lake analytics stack includes a query engine that can talk to many different data formats, a catalog for metadata management, a transaction manager to ensure ACID support, and cloud storage. The data application, computational resources, and storage are decoupled, allowing for much more flexibility in addition to better performance and scale, especially when it comes to the query engine.



As users start feeling the burden of higher costs as their cloud data warehouses (CDWs) grow, they will start looking at the open data lake analytics stack to address costs, as well as give them the flexibility and openness that CDWs don't provide.

The data lake analytics stack will be the new data management architecture for analytics and AI, simplifying data infrastructure and accelerating business innovation. Not only will this give companies more flexibility and cost savings, but it will also unlock even more workloads that before couldn't be run on the traditional CDW, enabling data platform teams to support a wider range of data applications.

We're just in the beginning. Next year will bring even more innovation to the data lake in such areas as compliance, data governance, and security products/tools from cloud vendors.

Trend #3: A post-pandemic, data-driven strategic shift to out-of-the-box solutions

The pandemic has brought about massive change across every single industry. What sticks out to me is how fast the successful "pandemic" companies were able to pivot from their traditional business model. Look at Uber as an example. During the pandemic, Uber Rides all but disappeared. So what did Uber do? They pivoted to put a massive focus on their meal business -- Uber Eats -- to keep business going. They were able to do that quickly because of their flexible data infrastructure, which helped them quickly make product changes and go to the market to deliver exceptional results.

Uber is just one example. What this pandemic exposed for today's digital-native company is that you have to be agile and quick to react if you want to succeed. Being cloud-native is just the first step. The pandemic spurred even faster cloud adoption and data-driven strategies.

In 2022, we'll see even more companies focus on building an infrastructure that allows them to analyze business/product data and pivot quickly to accommodate new go-to-market programs. Most companies can't be Uber -- that requires significant resources -- but they can be smart about where their IT teams focus. That means less time focused on managing complex, distributed systems and more time focused on delivering innovation and business-driven decisions. As a result, we'll see more out-of-the-box cloud solution providers that reduce the complexities of the cloud so companies can focus on what they do best -- delivering value to their customers.