Oracle Autonomous Data Warehouse dives into the Data Lake

Connecting with data across the cloud and Apache Iceberg

Published May 2023

Executive Summary

Trigger

Oracle Autonomous Data Warehouse (ADW) is adding a data lakehouse. With the spring 2023 release, Oracle is formally adding support for open source data lakehouse access with read-only support from Apache Iceberg and AWS Glue. It is also expanding access to cloud storage through several pathways; making it simpler to connect to and transform data from non-Oracle sources; and like another of its cloud database services, Oracle MySQL HeatWave, is adding support for the open source Delta Sharing protocol for sharing data. How will this impact ADW's reach?

Our Take

The core theme for the spring 2023 release is about turning Autonomous Data Warehouse into an analytics hub with upgraded support for connecting to data sitting in the data lake and other common data sources, and better developer tools. That encompasses expanding access to data sitting in cloud object storage, which is the de facto data lake. ADW already had an Amazon S3 API. Now it is expanding support for data sitting in cloud storage in several ways, starting with the ability to connect to other non-S3 cloud object stores that have an S3 API (which has become the industry's de facto standard). Additionally, Oracle has announced long-awaited support for reading data from Apache Iceberg, the leading multivendor-supported data lakehouse table format, which makes data sitting in object storage ACID-compliant. This is just the first step; we are looking forward to Oracle extending full write capabilities to iceberg as well.

Adding icing on the cake, Oracle is also dropping costs for data sitting in ADW premium storage by 75%, almost leveling the playing field with object storage costs. As for connecting to data from other sources, ADW is adding JDBC connectors that can work with other data warehouses, from cloud SaaS services to on-premises installations. And, in the spirit of Oracle's APEX low code/no code tool for developers, Oracle is extending the approach to data transformation and integration with a new Data Studio web-based development tool for ADW.

The lakehouse/Iceberg support puts Oracle on the right side of history, as it is joining a rapidly growing technology ecosystem that places Oracle on the same open data playing field as its rivals. And on that open data playing field, it can take the gloves off in competing with the TCO and SLA advantages of ADW.

Published May 2023

While we weren't looking

When Oracle introduced the Autonomous Database back in 2017, the headlines were all about whether, how, and how much it would impact the need for DBAs. And based on our discussions with a number of early adopters, the Autonomous Database has had significant impact there. It gained its early foothold with modest-sized businesses with limited DBA skillsets, aggressive pricing, and a simplified user experience that enabled customers to get up and running with just a few clicks, and a TCO that Oracle has benchmarked as undercutting popular cloud data warehousing services such as Amazon Redshift. ADW, a version of Autonomous Database optimized for analytic workloads, has subsequently begun making inroads with Oracle's core large enterprise installed base, and while it still accounts only for a tiny percentage of overall deployments, traffic has piled up. ADW is now processing an average of 9 billion queries hourly, which far surpasses that of Snowflake. Who knew?

ADW has also lived up to its promise for operational simplicity as reflected by the lower incidence of service calls and higher uptimes. Not to disrespect Oracle's other cloud database services, but ADW entertains *38x fewer service requests* when compared to Oracle Database Service "base" edition and Oracle Exadata Database Service. With a track record exceeding five years of service, ADW delivers Four 9's of availability (99.995% to be exact), which is among the tops in the industry. By contrast, providers such as Snowflake and AWS provide SLA commitments based on sliding scales. Snowflake's SLA commitment, which is based on query execution errors (just one source of issues in the user experience), specifies Three 9s for 1% query execution errors, and Four 9s for query error rates of 10%. As for Amazon Redshift, service credits are offered when uptimes are below Three 9s, based on overall monthly figures.

While ADW has not yet grabbed the spotlight like Snowflake or Databricks, its analytic footprint is much larger than you think.

Diving into the Lakehouse

With this release, Oracle is taking a number of steps to entrench the data lake (e.g., data sitting in cloud object storage) with ADW, and has taken its first dip into the data lakehouse.

The data lakehouse is about delivering the best of both worlds: the scale, flexibility, and openness of the data lake with the SLAs, repeatability, and mature governance of the data warehouse. This does *not* come with traditional approaches for treating data in cloud storage as external tables.

Published May 2023

The elevator pitch for data lakehouse is ACID compliance. It's about building confidence in the data sitting in the data lake, ensuring that it is current, consistent, and transactionally valid. For the lakehouse, ACID is *not* about turning the data lake into a transaction database.

There are side benefits to having ACID-compliant lakehouse tables, as they make it possible to enforce security and access control down to column and row-level, which was never possible with file formats sitting in object storage. It can also improve performance over traditional file scanning, although lakehouse tables will never match the performance of optimized native tables. But for the types of queries thrown at a lakehouse, in most cases performance should be *good enough*.

The technology to accomplish that is to overlay an ACID-compliant table structure on data stored in open source columnar file formats (Parquet is the most popular and commonly supported) that physically resides in economical, durable cloud object storage. Over the past year, we have found critical mass commercial ecosystems forming around open source data lakehouse table formats, led by Apache Iceberg and Delta Lake, and conclude that open source will drive this tier.

Until now, ADW supported the ability to copy data from Amazon S3-compatible object stores, but it required a fairly complex path for establishing credentials to access that data, and the data lacked ACID guarantees. It already supported native integration with Amazon Identity and Access Management, as it has also done with Azure and Google Cloud, making access to data sitting in cloud storage pretty straightforward.

With this release, Oracle adds formal data lakehouse table support along with tighter integration accessing and discovering data sitting in cloud object storage with a couple of key new features. First, it introduces read-only access to Apache Iceberg. Notably, this is a departure from the proprietary lakehouse table format adopted by Oracle MySQL HeatWave service (we hope that HeatWave will follow up with similar Iceberg support in short order). Iceberg support addresses the ACID requirement. And secondly, it makes AWS Glue a first-class citizen in ADW, extending the same ability to read metadata as it already has with that sitting in OCI's own data catalog.

Oracle ADW's read-only support is a good first step and puts it in league with counterparts like AWS Redshift and Google BigQuery. Ultimately, full lakehouse support would also extend to metadata integration along with full read *and* write support. That would make data sitting in the lakehouse a first-class citizen in ADW alongside data sitting in Oracle Database native tables. With full read/write support, Iceberg data could gain the same DML, security, and access control capabilities as data sitting in native tables. Among commercial providers, only a handful of cloud analytic services (e.g., Snowflake, Databricks, Confluent, and Starburst Data and a few others) and open source frameworks (e.g., Hive, Spark, and Flink) have gone that far.

In this release, Oracle adds another trick up its sleeve. It sliced the cost of storing data in ADW's native Exadata optimized storage by 75%. While exact comparisons with cloud object storage are tricky (they typically offer multiple tiered pricing structures), it significantly changes the equation for storing data in ADW native tables locally. For data sitting in cloud object storage, Oracle is saying, in effect, "If you can't beat 'em, join 'em." We expect that the practical effect won't necessarily be the downloading of reams of data from S3 et al. as there are reasons such as data egress costs, governance, requirements for data in other applications, and the question of why bother fixing something that ain't broke. But it provides an economic means of expanding ADW's native storage footprint, thereby making it economical to lengthen the lifecycle for retaining data locally in ADW.

Connecting to the rest of the analytics world

Oracle is promoting what it terms "multicloud" support with the upcoming release. A fuller discussion of Oracle's multicloud strategy is beyond the scope of this research note, although we'll provide this spoiler alert: look to OCI's interconnect with Azure as the indicator on where Oracle wants to go. For ADW, multicloud means having the ability to *access data* in other clouds, as opposed to *running analytics* in other clouds. In addition to existing JDBC connectors, the new release adds a host of native connectors to cloud data sources like Salesforce.com, Amazon Redshift, Foursquare, QuickBooks, Google Analytics, and so on.

This also ties in with a related announcement of a new Autonomous Database Data Studio, a suite of built-in tools, including Data Load, ELT-style Transforms, Data Analysis, Insights and Data Catalog. Its release is consistent with Oracle's renewed push for more love from developers. In the tradition of Oracle APEX, it is designed as a low code/no code visual environment; but it will also enable developers and analysts to drop in SQL coding as well. It complements Oracle's existing support for the popular dbt tool, which meets developers where they live.

There is yet another strategy that will make data connectivity more ubiquitous: support for data sharing. With the new release, Oracle is adding support for Delta Sharing, an open source protocol for cross-organization and cross-cloud data sharing; their support is bidirectional. For now, this is a modest start because Delta Sharing does not (yet) support monetization that would make disparate data sources available in a commercial marketplace. Our take is watch this space; OCI already has an applications marketplace, and we also expect that Delta Sharing will inevitably add a monetization component (in spite of discussions that we had with Databricks last year; things change).

The bottom line is that ADW can connect to the most popular data sources, although some (if they are deployed by specific cloud providers) may incur egress costs.

Published May 2023

Takeaways

The spring 2023 release of Oracle Autonomous Data Warehouse is all about extending the reach of the platform and evolving it into an analytics hub. Oracle has targeted data in cloud object storage, connectivity to other data sources, and data sharing.

As noted before, ADW has been able to pull data from cloud object storage, but support was limited to copying data and the processes for gaining access were hardly seamless. Oracle has addressed that in this release in a couple key ways. ADW now syncs, not only from OCI's own data catalog, but AWS's Glue as well. And it adds the ability to read from Apache Iceberg, which potentially could make data sitting in the data lake a first class citizen in ADW.

The latter advantage—making data lake data a first-class citizen, has huge potential importance for Oracle and ADW customers as it expands the reach of ADW to open data. Over the past year, a critical mass ecosystem of data and analytics vendors have coalesced around supporting data lakehouses, with Apache Iceberg being the leading multivendor platform, and Oracle is now the latest provider to get on the right side of history. The market is still at an early stage of awareness, so Oracle will be ready when its customers begin demanding lakehouse support. Oracle's support of Iceberg is at an early stage, as it is read-only and doesn't have the full DML support to perform writes; but then again, that's also where most (not all) of Oracle's data and analytics rivals are.

By supporting Iceberg, Oracle gains the same access to data as rivals such as Amazon Redshift, Google BigQuery, and Snowflake. As enterprises embrace the lakehouse, the query engine and the control plane, not the table format, will become the differentiators. It places Oracle on the same open data playing field as its rivals, where it can come out swinging in competing with the TCO and SLA advantages of ADW.

Author

Tony Baer, Principal, dbInsight

tony@dbinsight.io

Linked In https://www.linkedin.com/in/onstrategies/

About dbInsight

dbInsight LLC® provides an independent view on the database and analytics technology ecosystem. dbInsight publishes independent research, and from our research, distills insights to help data and analytics technology providers understand their competitive positioning and sharpen their message.

Published May 2023

Tony Baer, the founder and principal of dbInsight, is a recognized industry expert on data-driven transformation. *Onalytica* named him as a Top Cloud Influencer for 2022 for the fourth straight year. *Analytics Insight* named him one of the 2019 Top 100 Artificial Intelligence and Big Data Influencers. His combined expertise in both legacy database technologies and emerging cloud and analytics technologies shapes how technology providers go to market in an industry undergoing significant transformation.

Published May 2023