



Modernizing the AdTech Stack

WITH NEXT-GEN DATA SOLUTIONS

For Media and Entertainment Use Cases

HYDROLIX | E-BOOK

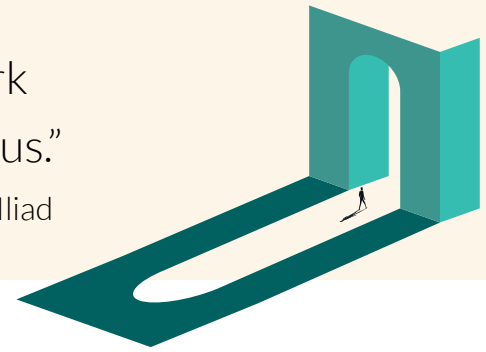


CONTENTS

Why the AdTech Data Stack Needs to Modernize	2
The Value of Real-Time Analytics and Monitoring	4
Advertising Campaigns	4
DSP Bidding.....	4
SSP Revenue Delivery	5
Publisher Inventory.....	5
Company Revenue.....	5
OTT Performance.....	5
Removing Problematic Content	6
Why Analytics Teams Need Unified, Accessible Data.....	7
Many Data Sources Without a Unified View	7
Data Retention Periods Are Too Short.....	7
Poor Query Performance.....	8
Challenges Adding New Parameters and Metrics	8
Engineering Teams Need To Rethink	
Legacy Data Systems.....	9
Privacy-First Models Will Lead to System Redesign.....	9
Machine Learning Will Impact the Entire Stack—	
and Increase the Volume of Data	10
Moving On From Batch: Modern Platforms Embrace Streaming.....	11
Object Storage Reduces Costs—But Needs to Be Performant	11
Choosing the Right Data Platform	11
Real-Time Data Ingest	12
Real-Time Data Transformation	12
Optimized for Analytical Queries.....	12
Optimized for Time	13
Support for Late-Arriving and Out-of-Order Data	13
Support for Ingesting Multiple Data Sources into One Table.....	13
Flexible Schema.....	14
Summary Tables.....	14
Long-Term, Cost-Effective “Hot” Storage.....	14
Ability to Independently Scale Up or Down	15
Separate Query Pools to Prevent Resource Contention	15
Powering AdTech at Scale with Hydrolix	16
Endnotes.....	17

“The roaring seas and many a dark range of mountains lie between us.”

– Homer, The Iliad



The AdTech market size was more than 845 billion dollars in 2023, with projections of more than three trillion by 2030.¹ That’s a tremendous amount of opportunity, but there are also roaring seas and many a dark range of mountains that will determine whether AdTech enterprises are successful or end up underwater.

The industry already faces major challenges, which include complex (and opaque) real-time bidding, made-for-advertising (MFA) sites, unsold inventory, low margins, and bot clicks. But that’s just the tip of the iceberg. There’s also the transition to cookieless, customer privacy concerns, a push towards sustainability, and the need for greater transparency across platforms.

In order to succeed, AdTech platforms, ad exchanges, DSPs, SSPs, even publishers—in short, all the players working to get the right ads in front of the right eyes—must harness and analyze billions of data points each day, often in real time. The volume of incoming data is greater than ever before, and AdTech data stacks have high storage and compute costs, leading to low margins and revenue loss. AdTech enterprises typically try to lower costs through more cost-effective tiered storage, but that data becomes more difficult to access and analyze, which means it turns into dark data and loses its value.

While the specific use cases vary for each part of the ad management chain, the underlying data needs are the same: teams must be able to analyze billions of data points, both for real time and historical use cases.

OVERVIEW

In this report, you’ll learn about the following:

- Why the AdTech data stack needs to modernize
- The value of real-time analytics and monitoring in AdTech
- Why analytics teams need unified, accessible data
- How engineering teams need to rethink legacy data systems
- Next-gen data solution features for AdTech use cases
- How Hydrolix can power AdTech platforms

Why the AdTech Data Stack Needs to Modernize

Serving the right ad to the right person in real time is a complex process, and many media experts don't think the process is working well. There is a major confidence crisis in programmatic advertising, and too little transparency in the process for enterprises to ensure that their ad efforts are effective.

Combined with macroeconomic challenges in the AdTech industry and many users and platforms saying no to cookies, AdTech platforms have little room for error. They must provide transparency and analytics both internally and to their customers, which means keeping a vast amount of data readily available for investigation and insights. However, they are also dealing with data infrastructure that comes with high storage and compute costs. In other words, the transparency and analytics that they must offer in order to be successful comes with prohibitively high costs. Neither the high costs nor the lack of data are sustainable, so AdTech enterprises must evolve from legacy data stacks to modernized solutions that provide cost-effective data at scale.

Let's start by looking at the current confidence crisis. 94% of experts currently use supply path optimization (SPO), which uses real-time bidding (RTB) and header bidding, to maximize the value of ads.²

AdTech enterprises must evolve from legacy data stacks to modernized solutions that provide cost-effective data at scale.

However, less than half (49%) of media experts think that SPO makes their ad spend more efficient.³ This is an astounding lack of confidence in the process. There has been a move towards quality path optimization (QPO), but the movement is in its infancy. A programmatic advertising cynic (and there are many) might be tempted to think it's the same old problem with a new acronym.

Programmatic advertising accounts for over 90% of US digital display advertising⁴ and global spending on programmatic advertising in 2024 is projected to be nearly 600 billion dollars.⁵ Advertisers want to maximize the value of their spend, and publishers want to maximize the amount of revenue they get from their ads, but too often, they must worry about ad fraud and the risk of damage to their brand. According to IAS research, 46% of industry experts list increased brand risk as a primary challenge of programmatic advertising, and 44% cite increased ad fraud as a challenge. Lack of transparency, hidden AdTech costs, and low ad views are also high on the list of challenges.⁶ Meanwhile, publishers only get about 50 cents on an advertising dollar, though in reality, the actual amount is often much less.⁷

When customers on both sides of the process (publishers and advertisers) are faced with poor results, it threatens the sustainability and long-term financial health of many of the players in the middle—DSPs, SSPs, and exchanges. The biggest publishers will have the power to return to direct deals and preferred partnerships, and with fewer cookies for collecting data across platforms, publishers will be able to collect first-party data while DSPs and SSPs, traditionally reliant on third-party data, will have new challenges to collecting valuable audience data.

These factors have in part led to the consolidation of DSPs and SSPs, and also caused many SSPs to close. As Ari Paparo, founder of Marketecture, told Digiday, “Fundamentally, the SSP business is not very attractive... It’s not growing, and it’s very competitive as publishers really treat you like a commodity.”⁸ Yahoo’s SSP, which closed in 2023, is one such example of a major closure.

Consolidation and closures can help reduce some of the inefficiencies in the ad management chain, but there is still intense pressure for both the largest and smallest players in the AdTech space. An SSP focused on building strong relationships with publishers can be laid low by MFA sites and ad fraud. A major publisher can shut down ads from a DSP when inappropriate ads or malvertising show up on their site. In order for DSPs and SSPs to thrive in an industry that traditionally has low margins, they must mitigate issues quickly, provide transparency to customers, and offer features that show demonstrable results. That means quality impressions and clicks for advertisers and increased ad revenue for publishers—with as much assurance as possible that their brands won’t be damaged.

To provide this transparency, AdTech platforms need cost-effective data solutions that provide both real-time and historical analytics while consolidating instead of siloing data.



The Value of Real-Time Analytics and Monitoring

Many AdTech teams rely on batched data and must wait days or even weeks to get reports. How are teams supposed to effectively monitor and evaluate the performance of their campaigns and supply paths when it takes so long to get actionable data? The irony is tremendous—real-time bidding and supply path optimization must happen with millisecond latency, and yet teams often do not have access to that data in a timely manner.

The process of improving transparency, trust, and performance begins with real-time analytics. However, in-memory databases and other solutions that offer millisecond latency for supply path optimization are extremely expensive at scale. From a budget perspective, they are not realistic solutions for analyzing and retaining billions of data points. But waiting days or weeks for a report is not an acceptable option, either.

The solution is delivering real-time analytics within a minute of data ingest. In the case of Hydrolix, data is typically ready for querying in 15-30 seconds, even when the system is ingesting millions of rows per second in real time.

AdTech enterprises must evolve from legacy data stacks to modernized solutions that provide cost-effective data at scale.

There is a prevailing myth that streaming data is more expensive than batch data, but with the right solution, that simply isn't the case. The issue is not that streaming is more expensive, but rather that many legacy data solutions don't have the right architecture for streaming, leaving engineering and data teams with the added work of developing complex pipelines for ingesting, transforming, and analyzing data. Too often, that leaves teams waiting for batch jobs that come in at the end of the day or week.

However, there are a wide range of jobs across the AdTech chain that benefit tremendously from real-time analytics and monitoring. Teams must be able to dissect performance and health across the following:

Advertising Campaigns

This includes assessing specific inventory, creative, and pricing that's generating the best performance. At the same time, campaigns must be adjusted in real-time based on unforeseen risks such as video rendering issues, ads running on fraudulent inventory, or broken CTA buttons.

DSP Bidding

Teams must be able to detect when DSPs are timing out, sending malformed bid responses, bidding below price floors, or serving broken creative, all of which negatively impact performance for advertisers as well as the overall margin for DSPs—and these issues must be mitigated immediately. DSPs rely on header bid

wrappers, which are small JavaScript files served by content delivery networks (CDNs) for RTB. In order to maximize RTB performance on a global scale, DSPs must monitor the CDNs they use and be prepared to switch out CDNs regionally if there are performance issues. Header bid wrappers and other tags must work with a wide range of devices, from mobile phones to laptops to connected TVs (CTVs), and then ads must display properly on end-user devices. There can be issues with SDKs, MRAID tags, and more, and troubleshooting these issues means monitoring a wide variety of device metrics.

SSP Revenue Delivery

SSPs must ensure they are serving the highest-quality demand across regions with the most inventory. They must also simultaneously keep inventory priced effectively and serve bid requests to DSPs where there is both demand and a high probability of ad placement.

Publisher Inventory

Publishers have many options for selling inventory, including private deals, open exchange, and programmatic guaranteed, and they must be able to assess which options generate the most revenue and how they can best optimize their inventory. This is especially important with changes in cookie usage already leading to a significant drop in ad revenue, including a 30% revenue reduction for cookieless users on Chrome.⁹

Company Revenue

Determining overall revenue is especially challenging after consolidations, including complex mergers that combine DSPs, SSPs, and publishers. Extracting gross and net revenue, margins, and other business health metrics is incredibly challenging when teams are navigating multiple disparate systems across each part of the company, leading to distrust and sometimes a single questionable Excel spreadsheet that becomes the source of truth. While company revenue doesn't necessarily rely on real-time data, if the reports rely on stale data sources that could be days or even a week old, revenue numbers will be neither accurate nor consistent.

OTT Performance

To troubleshoot and ensure optimal over-the-top (OTT) ad performance, which is a high-value source of revenue, teams need to collect a wide range of device information ranging from screen size to operating system to codecs. By correlating ad performance across geos, device types, and other dimensions, engineering and site reliability teams can find and fix issues with display ads and they can also find and mitigate fraudulent activity. Hydrolix includes an integration with ScientiaMobile, a real-time device detection solution that can identify more than 100,000 device profiles, allowing AdTech solutions to optimize the viewing experience and collect useful customer data.

Removing Problematic Content

DSP and exchange teams must also triage issues with problematic content in real time. This includes inappropriate content, malvertising, or simply ads that don't fit a publisher's brand, such as content reflecting extreme political views. These issues can be difficult to triage because there is often minimal context around the problematic ad. To fix the issue, publishers, analysts, and demand teams must analyze log information based on a small amount of context, such as when the ad was seen, a screenshot, and geo information.

After the issue is resolved, teams must replicate the issue, determine what happened, and perform forensic analysis on relevant ad data to fully report the issue to the publisher and ensure that the problem doesn't happen again.

Data must be unified, easy to access regardless of whether it's a minute or a year old, and most importantly, it must also be cost-effective.

Real-time analysis is critical for all of these jobs, and lack of access to that data can lead to lost revenue, brand damage, and strained partner relationships. And to provide the data needed for all of these jobs, AdTech platforms need to build on streaming data lakes that offer daily ingest at terabyte scale and beyond along with sub second query analysis—even on datasets that involve billions of data points. The data must be unified, easy to access regardless of whether it's a minute or a year old, and most importantly, it must also be cost-effective.



Why Analytics Teams Need Unified, Accessible Data

There are many jobs to be done to keep the ad management chain running smoothly, and those jobs are dependent on unified, accessible data. Unfortunately, for many AdTech platforms, there are numerous barriers to access, including:

- Many data sources in many different places with no cohesive view across systems
- Short retention periods limiting access to historical data
- Poor query performance
- Difficulty adding new metrics and parameters
- Analytics teams have limited ability to determine which data will be exposed

Let's take a closer look at each of these issues.

Many Data Sources Without a Unified View

AdTech platforms must manage many data sources, and this is especially the case for enterprises that have undergone consolidation. Mergers are challenging enough to manage even without consolidating and unifying data. Each acquired company represents a data silo that must be integrated into larger systems, and teams are struggling to pull data not just from different sources but also different tools. This includes both the solutions for storing and querying data as well as the tools used to visualize that data.

Fixing the issue will take time and a significant culture shift as teams from previously separate enterprises work through the growing pains of learning how to integrate and work together. But true integration is only possible if the data itself is consolidated, ideally into a platform that provides long-term, high-performance data access while remaining cost-effective. The solution should also be compatible with a wide range of visualization tools such as Tableau and Looker Studio. It should also be possible to group data sources together without needing to perform complex, costly JOIN operations. That means storing multiple data sources that should logically be grouped together in one table instead of storing them in separate tables that are difficult to compare.

Data Retention Periods Are Too Short

Real-time analytics is just part of the equation. In addition to understanding how services are performing now, you also need to be able to compare their performance over time, whether that's a week ago or a year ago. However, many teams working with data at terabyte scale are forced into short retention periods for their data because costs are too high otherwise. Sometimes data is quickly moved into tiered "cold" storage, where it's difficult to access and analyze. And sometimes, it's discarded altogether. For very high-volume use cases, it's not unusual to see data being discarded within seven days.

This is problematic when digital marketing campaigns should generally last six to twelve months.¹⁰

Another digital advertising software platform that partners with DSPs states that six months is required for actionable results. “After that amount of time, patterns can emerge, and meaningful insights can be drawn from your results. And pulling the plug too early will mean you don’t have enough reliable data to optimize with.”¹¹

Platform users should be able to drill down at any point in time to get a better understanding of how these campaigns are performing. For full transparency, that means getting data about where their ads were placed, how ads performed on each site, including metrics like win rate, impressions, clicks, and purchases for different publishers—regardless of whether that data was collected a minute ago or a year ago.

AdTech teams must also monitor long-term performance across their inventory, partners, and services in order to thrive. This includes identifying and taking actionable steps on low-margin inventory (for DSPs), fine-tuning ad blocking (for publishers), and addressing billing discrepancies (for exchanges).

Long-term monitoring and analytics are only possible with platforms that retain your data and keep it readily accessible in a cost-effective manner.

Poor Query Performance

If you have to step away for a coffee break while you’re waiting for the results of a query, that’s already subpar query performance. If you’re waiting hours or even days for query results and also have to deal with the possibility that the query might time out instead of returning actionable results, that’s even worse. These problems are magnified when you have to run multiple queries to get the specific results you need.

Degraded query performance becomes increasingly common as the volume of data increases. And many platforms simply aren’t designed to handle data at terabyte scale. To make that data storage cost-effective, it’s moved into tiered “cold” storage that must first be rehydrated before it can be queried. Even queries of “hot” data can be slow at terabyte scale if you’re using a platform that isn’t optimized for scale.

Modern OLAP (online analytical processing) databases that are columnar and designed for big data use cases can help solve this issue. Hydrolix is designed for sub-second query performance even on trillion-row datasets.

Challenges Adding New Parameters and Metrics

It should not be so challenging to add a field to a report, and yet this is a common problem in AdTech. What if you add new data sources that have different fields? Or simply realize the business value of an attribute, and want to build new metrics and reports around it? Typically, this involves a request to engineering that is slow and costly to complete, if it gets completed at all. To make matters more challenging, engineering

teams typically determine which data fields are exposed and available to analytics teams, leading to blind spots and incomplete data.

The issue is often related to database architecture. For many data solutions, particularly solutions that are row-based instead of column-based, adding new columns is a costly process that also adds expense to future queries. Data solutions that are either schemaless or have flexible schema make these kinds of operations much easier. The right data solution will add minimal overhead when it comes to adding new columns and fields. Ideally, a data solution should be built for high dimensionality, which might mean hundreds or even thousands of columns.

Engineering Teams Need To Rethink Legacy Data Systems

Rising data volumes are already putting tremendous pressure on engineering teams to rethink how data is ingested, stored, and analyzed. Larger macroeconomic trends affecting the AdTech and tech industry are leading to greater scrutiny of cloud costs, increased adoption of FinOps best practices, and more focus on maximizing the value of data.

Other forces at work are also increasing the urgency to modernize, including:

- Privacy-first probabilistic attribution models that don't allow fingerprinting
- Machine learning training and observability
- An evolution from batch to stream processing
- S3-compatible storage, which is reducing the cost to retain data at scale

Privacy-First Models Will Lead to System Redesign

With users going cookieless and continued concerns about user privacy, engineers will need to shift from fingerprinting users to privacy-first probabilistic attribution models. An example of the shift is reflected in Apple's [Required Reason API](#), which restricts access to device signals that can be used to identify users unless app developers have a specific, allowed reason for using those signals. Simply serving targeted ads will not be a sufficient reason.

However, there are other ways to collect data about users that can help make ads more targeted. Hydrolix includes an integration with ScientiaMobile, which allows engineers to collect and analyze hundreds of device-related fields—regardless of whether that device is mobile, a connected TV, or another device.

Probabilistic attribution models typically use machine learning to analyze user behavior and determine the likelihood of a conversion or revenue. In lieu of fingerprinting, they drive ad placement and can significantly impact on the success of ad campaigns.

Machine learning models benefit from having large datasets, so it's important to retain valuable user data which can be used for training probabilistic attribution models. Which leads to the next point.

Machine Learning Will Impact the Entire Stack—and Increase the Volume of Data

Spending on AI-enabled media reached \$370 billion by the end of 2022 and is projected to reach \$1.3 trillion by 2032, with more than 90% of all ad spend AI-enabled.¹² Every touchpoint and user interaction is an event which should be captured in a log line. And every log line represents a potential data point for a machine learning model to more accurately generalize how users will behave and whether they'll have a higher or lower probability of converting or generating revenue.

Spending on AI-enabled media is projected to reach \$1.3 trillion by 2032, with more than 90% of all ad spend AI-enabled.

AI is already being used for many use cases, including finding new ways to segment customer data, building new real-time bidding models, creating anonymized identity graphs to monetize customer data while addressing privacy concerns, and performing predictive analysis. And with the growing importance of first party data, one way to increase audience interaction is through chatbots that use natural language models. AI models can also help detect ad fraud and malicious behavior—but it's important to keep in mind that bad actors seeking to profit off of programmatic advertising will also utilize AI as well.

However, in order for AI models to make generalizations and be effective, they need many training runs across large datasets. Otherwise, models can be underfitted or overfitted (either of these issues can occur when data sample sizes are too small), leading to issues with accuracy. AI use cases need tremendous amounts of data, so long-term, cost-effective “hot” data is required for readily available analysis and training.

Not all of these use cases will use log data for training runs, and the amount of data needed to create a well-fitted model will vary, but taken altogether, AI use cases both need and generate a lot of data and will put tremendous strain on data stacks that are not designed for data at scale.

Moving On From Batch: Modern Platforms Embrace Streaming

Many data solutions don't have built-in real-time streaming, leaving engineering teams to either batch data or build complex pipelines that support streaming. This creates major challenges for analytical and monitoring use cases that need to act on data in near real time. To make the process even more challenging, many solutions simply ingest raw data, pushing normalization, enrichment, and other data transformation further down the pipeline. Not only does this increase complexity in the data pipeline, but it can also lead to lost context and more compute costs as data has to be processed and moved multiple times.

When it comes to ingesting log data, it should ideally be ingested once, normalized and enriched as needed before it's stored, and then be immutable. This combination of real-time streaming and transformation is built into Hydrolix, allowing you to store your data exactly once and then establishing it as a source of facts for any teams that need it—also known as a “store once, use many” model.

Object Storage Reduces Costs—But Needs to Be Performant

With data volumes rising, the cost of storage must go down. Many enterprises are turning towards cost-effective object storage, which uses commodity hardware and is horizontally scalable. However, object storage is distributed and relies on HTTP requests, unlike local storage. To enhance its performance, you need a solution that uses massive parallelism and builds on the strengths of object storage. Hydrolix is one such solution.

Choosing the Right Data Platform

AdTech solutions typically generate billions of events per day, so it's necessary to ingest and analyze data at terabyte scale. At this level of scale, many data platforms have issues with performance (such as query timeouts), high costs, or both. For that reason, it's important to choose a data platform such as a streaming data lake that is specifically designed to efficiently process huge datasets (think 100+ billion rows or more).

Here are some of the key features that teams should consider when looking to build an AdTech platform on a next-generation data solution. For a deeper technical dive into features that these next-gen platforms should include, see our [Powering Big Data with Next-Gen Data Solutions whitepaper](#).

Real-Time Data Ingest

Many of the use cases described in previous sections, from real-time dashboards to triaging issues in the ad management chain, require real-time data ingest. While real-time bidding requires millisecond latency, this level of performance has prohibitive costs at scale—and is not necessary for analysis use cases, where incoming data should be ready for querying in a matter of seconds.

Generally, AdTech solutions should use a data solution designed for streaming ingest use cases. Many data platforms are designed with batch ingest in mind and do not have true streaming capacities. They may rely on micro batching or external integrations for streaming ingest.

Real-Time Data Transformation

Many data platforms are designed for raw, unprocessed data, and you typically need to use an ELT (extract—load—transform) process to normalize the data later. However, there are a number of issues with this approach. Data often needs additional context at ingest time (such as information about its source, when it was ingested, and so on), and that context cannot be added later during an ELT process.

Data is also often left in its unprocessed form, making it difficult for humans to reason about, and it's much less efficient (and more computationally expensive) to query data in its raw form. Structured data is easier to analyze and understand, plus you'll achieve greater query efficiency.

ELT processes often involve complex pipelines that can be difficult to build and maintain. They are also more expensive computationally because data needs to be written to storage multiple times (at ingest and then again during the transformation process).

With Hydrolix, you can transform data in real time when it's ingested, allowing you to enrich and standardize your data before it goes into storage. [Learn how Hydrolix transforms data at scale.](#)

Optimized for Analytical Queries

AdTech solutions typically use OLAP (online analytical processing) databases as opposed to OLTP (online transaction processing) databases. OLAP databases are columnar, making it efficient to query only the columns you need. Next-gen data solutions must maximize the benefits of optimizations like partition pruning, predicate pushdown, and indexing to achieve sub-second query latency even on massive data sets.

Hydrolix is a columnar database with additional optimizations, including sparse indexing of columns, that makes retrieving query data extremely efficient. Read [Maximizing Query Performance for 100+ Billion Row Data Sets](#) to learn more.

Optimized for Time

Most AdTech analysis use cases, including A/B tests, ad campaigns, and issue mitigation, have a key commonality: they are all filtered by time. AdTech is event-based, and each event should have a timestamp that provides critical context.

For this reason, a cloud database that's optimized for time-based data is often a good fit for AdTech solutions. In the case of Hydrolix, data is partitioned by minimum and maximum timestamps, making it extremely efficient to query with time filters.

Support for Late-Arriving and Out-of-Order Data

When you're ingesting billions of events per day, some of those events will arrive late and out of order. This is especially true when these events come from thousands of different types of user devices and global networks with varying levels of reliability. Handling late-arriving data is a major challenge for many data management platforms (DMPs)—to the degree that some platforms simply do not provide support for it.

Late-arriving data can make data aggregations and reporting inaccurate and also impact query performance. An aggregation that doesn't account for the 10% of data that arrives late may suggest that all systems are performing reliably—when the reality is that the late-arriving data may be the source of the majority of performance issues and errors.

Hydrolix creates small partitions at ingest time and then merges these partitions for increased compression and storage efficiency over time. The merge service automatically sorts late-arriving and out-of-order data.

Support for Ingesting Multiple Data Sources into One Table

AdTech analysts and data scientists often have to work with data from many different sources. Too often, this data is fragmented and challenging to correlate, or you need complex and computationally expensive JOIN statements to do so. One way to avoid this problem—and bypass the need for complex JOIN statements—is to ingest multiple sources of data in one table. For example, you might want to ingest CDN data into one table, and all data on SSP performance into another table. In addition to making it much easier to compare data, it's also easier to generate reports and dashboards.

With Hydrolix, you can ingest data from multiple sources into one table, transform that data on the fly to add context as needed, and then easily compare that data—no joins needed. Each source can have different dimensionality (different fields), and you can also easily add new dimensions to reports.

Flexible Schema

AdTech teams often need to add new columns and dimensions to reports, and with many data platforms, this can be extremely challenging. To solve this problem, a data platform must have a dynamic, flexible schema which can be adapted along with a team's data needs. This is much easier to do with columnar data formats like Hydrolix, where new columns can be added to a table without impacting pre-existing data. Combined with the ability to add multiple sources to one table, it becomes much easier to correlate and report on your data.

Summary Tables

AdTech use cases often need to aggregate data or provide quick, high-level insights. These aggregations and insights are often used for reports and dashboards. It can be computationally expensive to regularly query the underlying data tables, especially when they consist of many billions of rows. Instead, that data can be aggregated into a summary table. Alternatively, some solutions choose to precompute and cache data into a separate table (known as a materialized view) for specific use cases such as more efficient querying.

Summary tables are a key feature of Hydrolix, and they also account for late-arriving, out-of-order data, unlike the majority of data solutions. You can also use table transforms to ingest and transform data for your specific use case. For example, that could include tables that ingest from all sources, summary tables, and customized tables that store data that's accessed on a regular basis.

Long-Term, Cost-Effective “Hot” Storage

Many AdTech use cases require long-term query analysis. Some examples include comparing campaigns year-over-year, using data for BI and AI use cases, and retaining customer data for future insights. However, ingesting billions of events each day means that you'll eventually be storing trillions of events. Unfortunately, this leads to prohibitively high storage costs with many data platforms.

To reduce costs, many AdTech solutions either move data to “cold” storage that isn't readily available for query access—or throw the data away altogether. This leads to missed opportunities for insights and problems with “dark” data, which is data that's collected but never analyzed.

Hydrolix is designed to maximize the performance of cost-effective object storage, reducing long-term storage costs and giving you “hot” storage performance (low latency queries) at “cold” storage prices. Hydrolix also uses high-density compression to dramatically reduce the size of your data footprint, so that users typically get 4x more data at 4x lower costs than other platforms.

Ability to Independently Scale Up or Down

AdTech solutions must deal with peak events such as major sales, Black Friday, holiday shopping, and more. During these events, data ingest and analysis often need to happen at much greater orders of magnitude, which means you need a data platform that can quickly scale to meet increased demand and then scale back down after the event for reduced costs.

Unfortunately, many data platforms are built using tightly-coupled architectures, which means that each part of the system can't be scaled independently. Scaling with these solutions is often challenging, and all parts of the system need to be scaled together, which leads to overprovisioning and wasted resources.

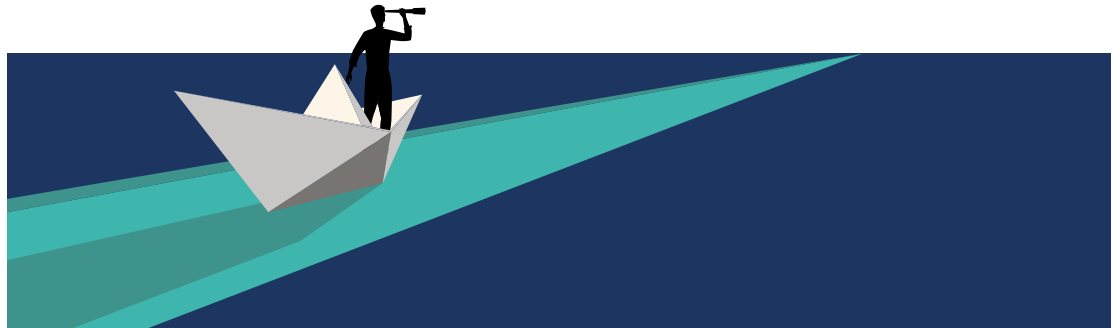
In contrast, each part of Hydrolix is independent and stateless, making it easy to scale up or down. Query and ingest can be scaled independently, while storage is pay-as-you-go and horizontally scalable.

Separate Query Pools to Prevent Resource Contention

When many stakeholders need to access data from the same source, there can be significant issues with resource contention. This is commonly referred to as the “noisy neighbor” problem where one team or stakeholder uses a disproportionate amount of compute resources, affecting the query performance of other teams.

This can be especially challenging for AI and BI use cases, which are resource-intensive, and other use cases like dashboards and ongoing cron jobs will constantly use query compute as well.

With Hydrolix, you can prevent resource contention by using [precision query scaling](#), which allows you to create separate query pools for different teams and stakeholders.



Powering AdTech at Scale with Hydrolix

The right data lake can help you navigate both the dark mountains and stormy seas of AdTech. Hydrolix is a streaming data lake that allows you to process, retain, and query event data at petabyte scale at a fraction of the cost of other data platforms. You can use Hydrolix to build B2B AdTech products, develop in-house solutions, and manage event data at scale, whether that's customer data, internal data, or both.

Hydrolix incorporates the core principles of modern cloud-native infrastructure and stateless, decoupled architecture to provide the following benefits:

- Data ingest, storage, and analysis using a columnar datastore with OLAP principles for petabyte-scale use cases.
- Hot storage performance at cold storage prices for campaign tracking, AI model training, and other long-term data use cases.
- Sub-second latency on ad hoc queries of large data sets for uncovering performance issues, finding and mitigating problematic inventory, forensic analysis, and more.
- Optimized for time-based use cases and designed to automatically sort out-of-order data.
- High-density compression algorithms that reduce your data footprint by 20-50x, reducing storage costs dramatically.
- Real-time data transformation at ingest time so you can structure and enrich events, normalize data, create summary tables, and build custom tables.
- All data lives in your own [virtual private cloud \(VPC\)](#), allowing you to control your data while reducing costs. Ensure the integrity of your data by managing it yourself.

When you build your AdTech solution with Hydrolix, you can give your customers long-term, affordable visibility, build transparency and trust, build powerful AI models to predict customer behavior and improve ad campaigns, and give marketers and publishers the tools they need to increase revenue. To build a next-generation AdTech platform, you need a next-generation data solution like Hydrolix.

Visit us at hydrolix.io to learn more about how Hydrolix is powering big data use cases.
Or email us at info@hydrolix.io to arrange a demo.

Endnotes

- 1 *AdTech Market Size, Share & Trends Analysis Report, 2030. (n.d.).* Grand View Research. Retrieved July 17, 2024, from <https://www.grandviewresearch.com/industry-analysis/adtech-market-report>
- 2 Integral Ad Science. (2022, November 16). *Constructing Your Quality Path: How Media Quality Strengthens Supply Path Optimization Strategies*. Integral Ad Science. Retrieved June 6, 2024, from <https://integralads.com/insider/constructing-your-quality-path-research/>
- 3 Integral Ad Science. (2022, November 16). *Constructing Your Quality Path: How Media Quality Strengthens Supply Path Optimization Strategies*. Integral Ad Science. Retrieved June 6, 2024, from <https://integralads.com/insider/constructing-your-quality-path-research/>
- 4 Lebow, S. (2024, February 14). *Guide to Programmatic Advertising: Market Size, Ad Types, & Buying Process*. eMarketer. Retrieved June 6, 2024, from <https://www.insiderintelligence.com/insights/programmatic-digital-display-ad-spending/>
- 5 *Global programmatic ad spend 2028. (2024, February 12).* Statista. Retrieved June 6, 2024, from <https://www.statista.com/statistics/275806/programmatic-spending-worldwide/>
- 6 Integral Ad Science. (2022, November 16). *Constructing Your Quality Path: How Media Quality Strengthens Supply Path Optimization Strategies*. Integral Ad Science. Retrieved June 6, 2024, from <https://integralads.com/insider/constructing-your-quality-path-research/>
- 7 Fou, A. (2023, April 14). *Publishers' Cut is ZERO Percent in Programmatic, How?* LinkedIn. Retrieved June 6, 2024, from <https://www.linkedin.com/pulse/publishers-cut-zero-percent-programmatic-how-dr-augustine-fou/>
- 8 Shields, R. (2023, February 13). *'Fundamentally, the SSP business is not very attractive': The fall out of ad tech's latest round of closures*. Digiday. Retrieved June 6, 2024, from <https://digiday.com/media/fundamentally-the-ssp-business-is-not-very-attractive-the-fall-out-of-ad-techs-latest-round-of-closures/>
- 9 Salmon, K. (2024, April 16). *Google's Privacy Sandbox causes 30% ad revenue drop for cookieless Chrome users*. CMOTech. Retrieved June 6, 2024, from <https://cmotech.uk/story/google-s-privacy-sandbox-causes-30-ad-revenue-drop-for-cookieless-chrome-users>
- 10 (n.d.). AdEdge Digital Marketing. Retrieved June 6, 2024, from <https://adedgemarketing.com/how-long-should-a-digital-marketing-campaign-last/>
- 11 *How long should an ad campaign run? - Digital Campaign Planning with Choozle. (2022, July 12).* Choozle. Retrieved June 6, 2024, from <https://choozle.com/blog/how-long-should-a-campaign-run/>
- 12 Mandese, J. (2022, April 6). *GroupM Benchmarks AI-Enabled Ad Spending At \$370B, Projects \$1.3T By 2032*. Media Post. Retrieved June 6, 2024, from <https://www.mediapost.com/publications/article/372705/groupm-benchmarks-ai-enabled-ad-spending-at-370b.html>