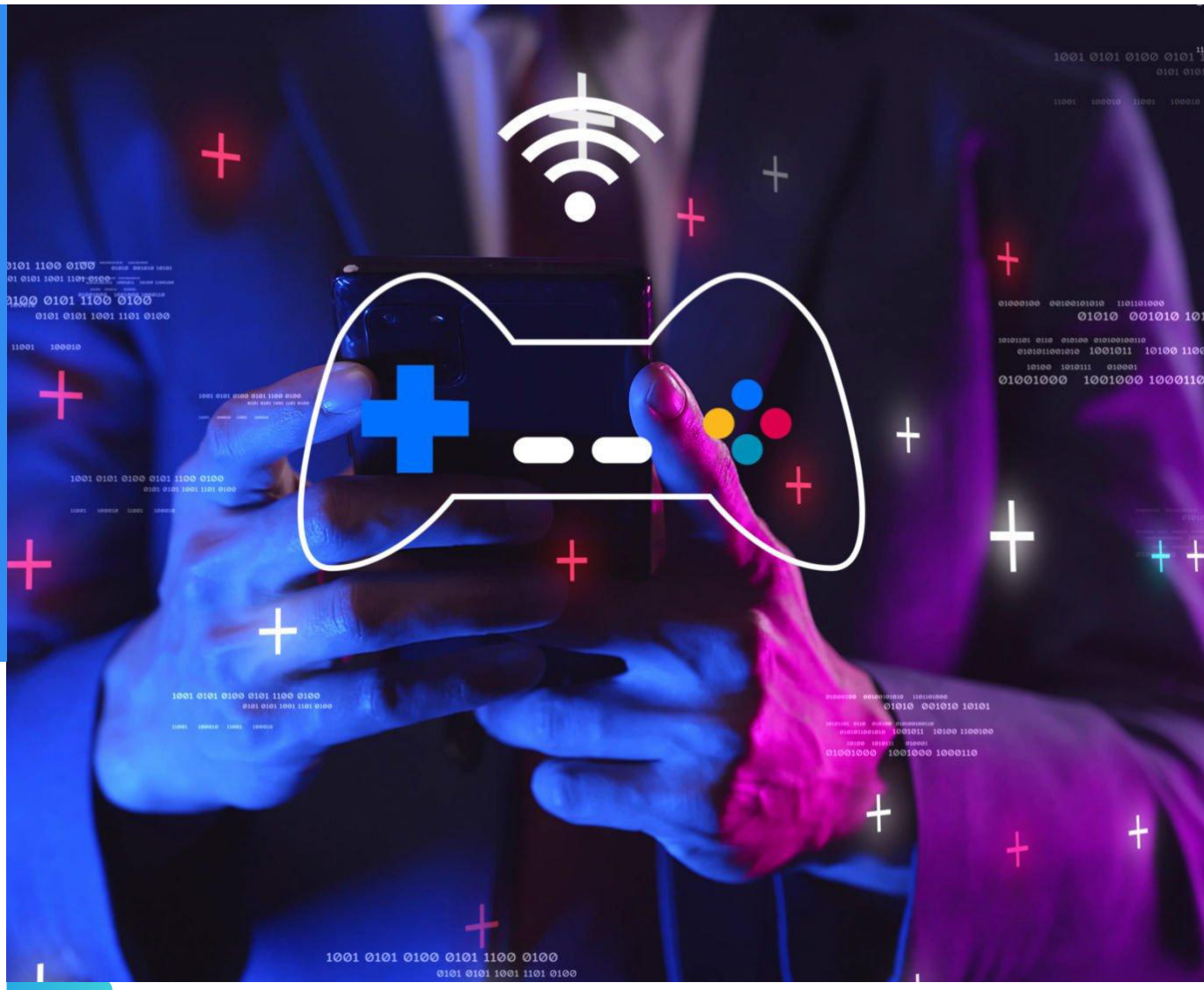
indium
Make Technology Work



Building a Data Lake Architecture for Gaming Applications

A Whitepaper

© Indium Software 2022

Data is the new oil fuelling all industrial segments, including gaming. Understanding user behavior, identifying trends and preferences, and analyzing games for problem-solving are some ways in which data is helping the gaming industry.

Large volumes of data help to develop machine learning algorithms for building descriptive and predictive models to improve the gaming experience.

Graphics and image recognition technology are other areas where data is transforming the gaming industry. Call of Duty (COD), for instance, is a game where big data is used to improve the gaming experience.

In addition to improving the gaming experience, data is also used for strategic planning, financials, and marketing communication.

With several titles produced every year across genres, retaining customer attention and ensuring their loyalty has become very important as well as challenging.

Knowing the pulse of the market, constantly improving the gaming experience, and increasing the difficulty level just enough to intrigue without losing customers is no child's play anymore.

Therefore, game developers have to leverage data to make timely, data-driven decisions.

As the importance of data increases in the gaming industry, so is the need for data lakes gaining popularity amongst game developers.

© Indium Software 2022

Benefits of Data Lake for Game Developers

Game developers need data storage solutions that allow ingestion, transformation, analysis, and visualization of structured and unstructured data in a secure, durable, and cost-effective manner.

A data lake combines data storage and governance to enable analytics, machine learning (ML), and visualization in a centralized and integrated data platform.

A cloud-based storage platforms such as Amazon Simple Storage Service (Amazon S3) can help game developers build a data lake that will allow:

- Ingestion of data from multiple data sources using ingestion methods such as batch, near real-time, and real-time
- Storing ingested data in a reliable, cost-effective, and centralized platform
- Enabling easy access to data assets stored in the data lake by building a comprehensive data catalog
- Ensuring data and business metadata security and governance
- Allowing monitoring, analysis, and optimization of cost and performance
- Providing visualization of data lineage
- Transforming raw data assets into optimized usable formats and querying them in place
- Enabling easy and secure sharing of processed datasets and results across teams

© Indium Software 2022

Data lakes provide real-time analytics that facilitates the tracking of crucial Key Performance Indicators (KPIs) to analyze performance, engagement, and revenue. This includes visibility into the player experience and Lifetime Value (LTV) during their lifecycle, empowering game developers with insights about player behaviors and experiences. This can help improve the game based on data to increase the player's delight and the game's success.

Building a Lake House Architecture

Data lakes are often built with data warehouses due to the latter's lower latency and improved SQL query performance when working with local data.

Data warehouse is especially common when daily aggregations need to be built with data extracted from business intelligence (BI) solutions.

From the time of logging to every keystroke is tracked in games, resulting in terabytes of data being generated and processed every day.

This data needs to be loaded onto different data stores, which could be a cache, a relational database, a NoSQL database, or log data.

Managing data can become very complex, making it difficult to get meaningful insights.

Integrating the data lake with the data warehouse can help to break siloes and allow historical data to be stored in data lakes.

This way, data warehouses with the latest data can be queried for faster responses, and the historical data in the data lake accessed only when needed.

© Indium Software 2022

Data Ingestion

Data ingestion is the first step and can be collected and processed from different sources such as in-gaming events, marketing data, and third-party services, including clicks, impressions, and installs.

Events could be streamed as they occur or loaded in batches. A combination is ideal for reducing the load on the network and keeping the costs low. That's why some degree of batching is often used with streaming to improve performance.

Both have their pros and cons. With real-time streaming, there is a lower risk of losing events due to client crashes or network outages, and it allows for real-time analytics capabilities.

But, since the data lake needs to aggregate data from various data sources available in different formats, batch processing may be an option to enable this and to process datasets of any size depending on the need.

Data Transformation

Since data is collected from different sources, they are available in different formats. For meaningful analytics, they need to be transformed into one format, but that is not an easy step.

Therefore, building a data pipeline is critical to allow the different data formats to evolve and match a standard format using ETL (extract, load, and transform).

Transformation engines can help with the large-scale data processing of batch and/or streaming data. The second option is to work with an unknown format.

In this case, a data classifier reads the data from the data store to identify known structures, and a custom classifier reads the data when a new format is spotted.

It is also essential to assess the volume of game data every month to enable the efficient transformation of data, the format and schema, the average file size, and the number of sources.

Further, identifying the stewards of the data, the type of transformations needed, understanding the consumers, ensuring its security, governance, auditing, and so on are also essential at this stage.

© Indium Software 2022

Data Cataloging

Modern data management and governance require data catalogs as a core component and technology for enabling data sharing and self-service analytics.

This enhances the data and analytics business value by facilitating the easy location, inventorying, and analyzing diverse data assets distributed widely.

It helps continuously improve the velocity and quality of data analysis, enabling self-service by business users to increase the productive use of data across functions. Therefore, it becomes essential to choose the correct data cataloging.

A data catalog lists the inventory of data assets across various data sources by collecting metadata and combining it with data discovery and management tools.

While helping with listing all data assets across all different data sources, it also improves the productivity of users to discover, understand, and consume data more and increases the chances of data lake adoption.

A centralized data catalog and deployment across various frameworks can improve the entire data management process.

© Indium Software 2022

Data Governance

Data needs to be available, trustworthy, usable, lineage, and secure. Ensuring this is data governance, which is covered in business policies and usually includes:

- Ownership of data and accountability
- Being properly structured in raw, curated, and processed formats with permissions and roles
- Monitoring to ensure there is no suspicious activity
- Enforcing rules and policies
- Automating data sharing, quality, alerts, and accelerating data access for authorized users with proper permissions.

Data Visualization

Visualization of data with gaming-specific dashboards can help to detect potential fraud or anomalies in player activity quickly and improve game security and experience. This helps with monitoring, visualizing, and auditing your data lake operations for continuous improvement.

Build an AWS Data Lake with Indium Software

Amazon Web Services offers a range of solutions for building an efficient and effective data lake. It provides multiple solutions right from building data lakes best suited to the game developer to data lifecycle management, workflow orchestration, monitoring, and cost optimization.

© Indium Software 2022

The AWS Well-Architected Framework helps assess the soundness of the decisions made and implement best practices when building the systems.

Amazon Simple Storage Service (Amazon S3) decouples computing and data processing, allowing a multitenant environment to be built. Other options, such as Amazon ElastiCache for Redis, Amazon Aurora, and Amazon DynamoDB, offer alternate storage solutions.

Indium Software is an AWS partner with a team of experts who can assess your needs and build a bespoke data lake architecture for gaming applications based on your data sources and expected outcomes.

To know more about how Indium can help your game development and marketing strategies with an efficient data management solution using AWS data lake architecture,

visit: https://www.indiumsoftware.com/aws-partner/

© Indium Software 2022

USA
Cupertino | Princeton
Toll-free: +1-888-207-5969

INDIA
Chennai | Bengaluru | Mumbai
Toll-free: 1800-123-1191

UK
London
Ph: +44 1420 300014

SINGAPORE
Singapore
Ph: +65 6812 7888

www.indiumsoftware.com

For Sales Inquiries
sales@indiumsoftware.com

For General Inquiries
info@indiumsoftware.com