INSIGHTS

Techstrong Research

Modern Data Processing Requires Serverless Computing

SPONSORED BY aws

DATA IS THE NEW OIL and organizations are sitting on a treasure trove of it but like most natural resources the challenge is refining that data in a way that makes it simpler for organizations to consume. Organizations of all sizes rely on their data to surface actionable insights to empower them to discern patterns, trends, and correlations that support informed decision-making and strategic planning. Efficient and timely data processing helps organizations streamline operations, foster innovation, adapt to changing market conditions and customer needs, gain a competitive edge, manage risks, and ensure compliance with regulations.

But building data processing and analytics pipelines that unlock the value of data for the entire organization is not an easy or straightforward journey. Data systems are often sprawling and complex, with diverse data sets spread out across data lakes, data warehouses, databases, SaaS applications, and on-premises systems. Historically, harnessing and processing this data has required IT teams to provision and manage massive amounts of IT infrastructure. However, with the rise of serverless computing it's now possible to process data on demand using IT infrastructure that automatically scales up and down as needed.

Serverless computing accomplishes that goal by eliminating the need to manage underlying infrastructure, including tasks like server provisioning, scaling, and infrastructure maintenance, enabling faster development and deployment of data processing operations. Serverless improves IT agility making it simpler to respond to

changing data processing requirements as needed in a way that reduces total cost of ownership. A serverless strategy makes it feasible to adroitly process massive amounts of data.

In contrast, legacy approaches to managing IT infrastructure will simply not be able to keep pace with the need to make better informed decisions based on data that now often needs to be processed in near real-time.

THE DATA VOLUME CHALLENGE

As the volume of data grows from increasingly diverse sources, there is a need for organizations to move quickly to process this data to ensure they are able to make faster, well-informed business decisions. Per Statista, data volume is projected at 147 zettabytes in 2024, a 22.5% increase from the previous year. This growth can be attributed to the expansion of connected devices, the widespread use of cloud solutions, and technological advancements.

To process data at that level of scale, organizations need to elastically provision resources to manage the information they receive from multiple sources. Unfortunately, most organizations today are struggling to get a handle on where all their data resides, how to connect to that data, and act on that data effectively. Rather than having to over provision IT infrastructure to process and analyze what in many cases is already petabytes of data, a serverless computing platform automatically scales resources up or down in response to workload changes, making it suitable for unpredictable and large surges in data volume.

For example, AWS Lambda can automatically scale up 1,000 concurrent executions every 10 seconds in response to sudden, large and often unpredictable increases in data volume. Similarly, Amazon Elastic Container Service (ECS) when configured as a serverless computing platform with AWS Fargate makes it possible to automatically scale containers to run thousands of tasks to process data using up to 6 vCPUs configured with as much as 120GB of memory.

Data processing is a multi-step process with different stages across the pipeline including data collection, preprocessing, transfer, analysis, insights and storage. Since individual tasks across the pipeline are often done by purpose-built services there arises a need to coordinate all these tasks in order to achieve business outcomes. Serverless integration services like Amazon EventBridge, AWS Step Functions, Amazon SQS and Amazon SNS play a key role in ingesting data that is available in various formats and from multiple sources, and connecting and managing the interactions of all the services within a data processing application.

DATA PROCESSING COSTS

Legacy approaches to processing data at scale can be especially expensive as data volumes grow, and the need for scalability becomes paramount. When processing data at scale, organizations often have to provision and maintain servers even during idle periods. Additionally, many organizations require real-time or near-real-time data delivery, and experience unpredictable spikes in data volume and availability which can be both challenging and cost prohibitive to manage. An efficient architecture is key for processing, and becomes more essential with growing data volume and velocity.

Serverless allows organizations to pay only for the computing resources they actually use, on a per-execution basis. This can result in cost savings compared to traditional server-based approaches, where companies have to provision and maintain servers even during idle periods. Additionally, the elimination of operational burden allows data engineers and developers to focus on building data processing solutions, which further accelerates operations and reduces total cost of ownership. According to Deloitte, customers can reduce their TCO by 48% when using serverless solutions on AWS.

DATA PROCESSING TYPES

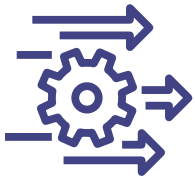
A serverless computing platform is used for several types of data processing applications including streaming data, large-scale parallel processing, and batch processing. Listed below is an example of a solution that works well for each of these application types.

STREAMING DATA

Streaming data includes a wide variety of data such as log files generated by mobile or web applications, e-commerce purchases, in-game player activity, click streams from social networks, financial trades, geospatial services, and telemetry from connected devices. Streaming data can be processed in real-time or near real-time to surface actionable insights. Because streaming data is generated in a nearly continuous, incremental manner it's not uncommon for organizations to process 2-3 terabytes every day. The challenge is this type of data needs to be processed quickly, in a specific order on a record-by-record basis or over sliding time windows.

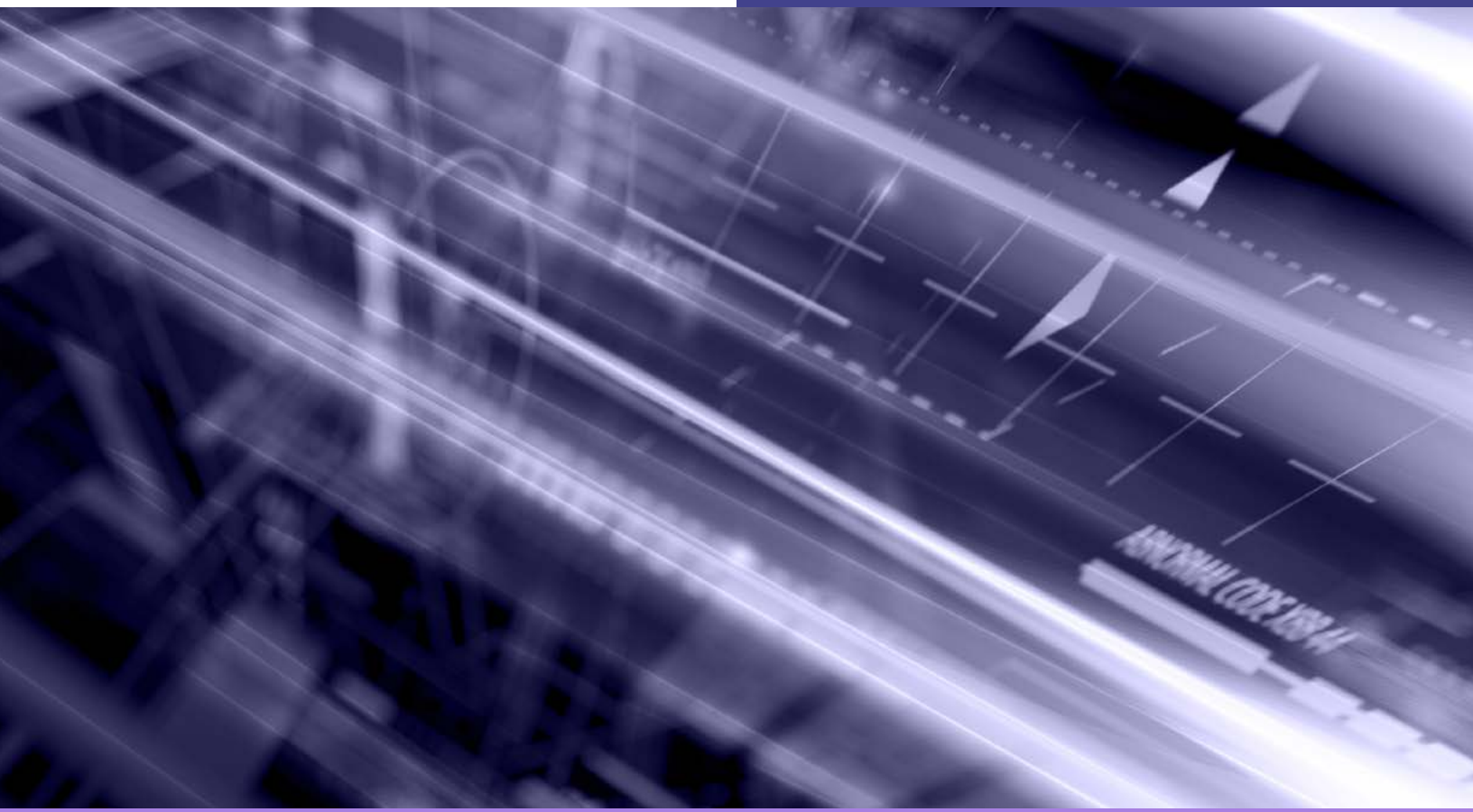
AWS Lambda is ideal for processing streaming data applications. It integrates natively with Amazon Kinesis and Kafka (through Amazon Managed Streaming for Apache Kafka (MSK), self-managed Kafka clusters, and Confluent Cloud) as a consumer to process ingested data. You can invoke Lambda functions from messages in Kafka topics or Kinesis data streams to integrate into downstream serverless workflows. AWS Lambda provides automatic scaling of resources to match data volume, reduced time to market for launching new services, reduced costs by only charging for compute time when functions are invoked, and flexible integrations with a suite of other serverless offerings and event triggers. With downstream event-driven architectures (EDA), you can easily create, modify, or remove various producers and/or consumers without worrying about planning for scale, and using any programming language.

LARGE-SCALE PARALLEL DATA PROCESSING



This requires running a series of business logic on every single piece of data in a dataset in parallel. Examples of running a business process for multiple entries in a dataset include creating a thumbnail for thousands of images in parallel, processing hundreds of thousands of invoices, running predictive analysis, or model simulation with hundreds of thousands of input data. With ever-growing data and the need to get insights faster, it's critical to master technologies that process large volumes of data in a distributed fashion. Many of the technologies that offer distributed processing require additional tooling and skills that platform teams and developers would need to purchase and learn which can hinder developer productivity.

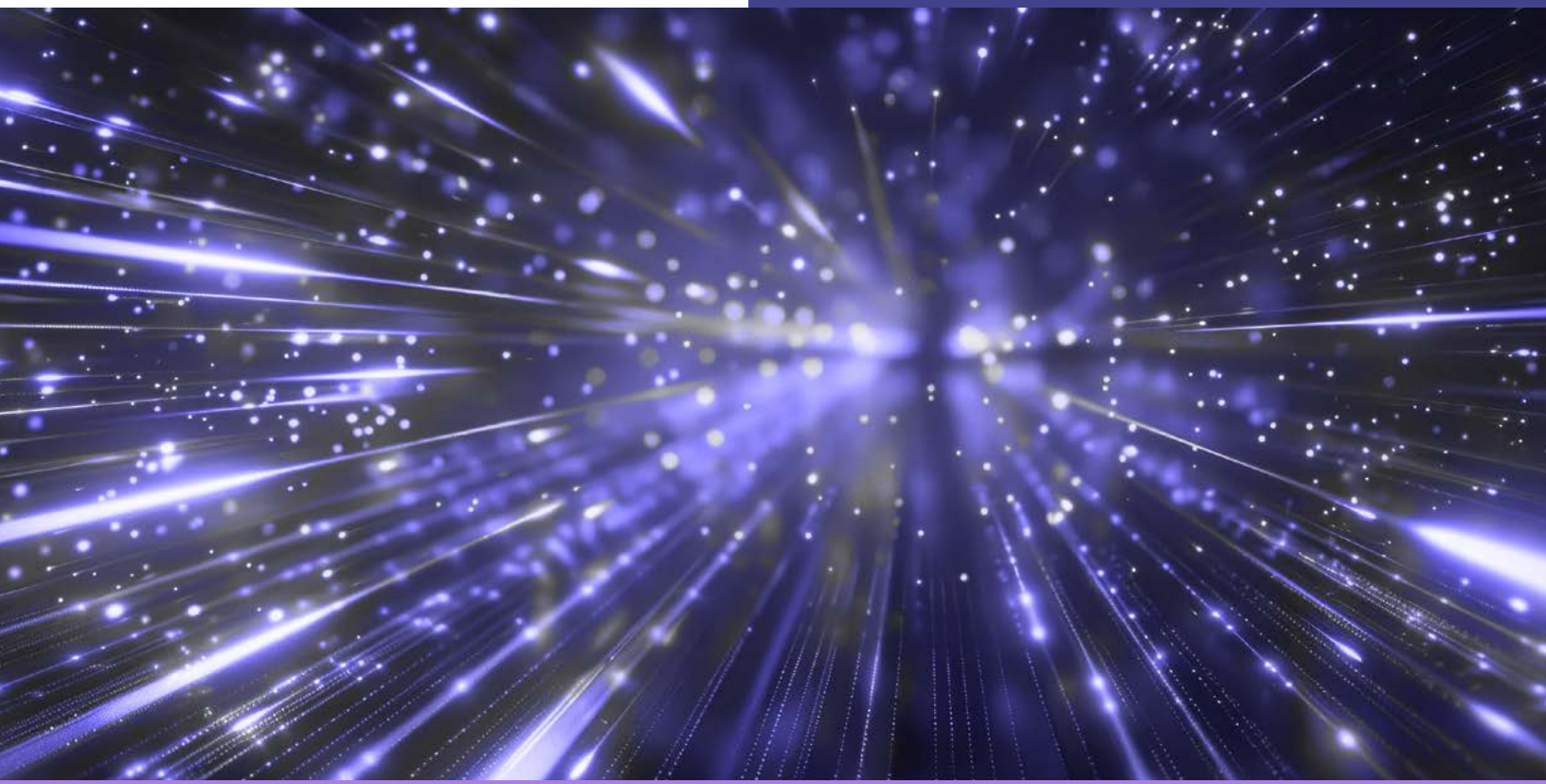
AWS Step Functions in combination with AWS Lambda is well-suited to process large-scale data in parallel. Distributed Map is a feature of AWS Step Functions which can iterate over data and batches the data enabling you to process it in parallel. It integrates with Amazon S3 natively so you can operate on millions of S3 objects in parallel, and AWS Lambda runs business logic with unparalleled concurrency and speed while offloading infrastructure management to AWS. With a pay-per-use pricing model, you can reduce your infrastructure cost of running serverful workflow orchestration and compute services. AWS Step Functions integrates with over 11K AWS APIs, third party APIs, and on-premises systems making it easier for developers to use AWS services they are familiar with to integrate with the data processing workflow.



BATCH PROCESSING

Many organizations still rely on traditional batch processing to enable file intake processes, queue-based processing, and transactional jobs, in addition to heavy data processing jobs. These workloads are typically processed on a recurring time interval, such as hourly, nightly, or monthly, so there is a need to invoke large quantities of compute for a relatively short duration of time, and on completion, scale down those resources. Of course, the faster the available compute is at the beginning of processing; the more rapidly such jobs can be completed. Data engineering teams that focus on batch processing are often deeply skilled in the nuances of data storage, processing logic, and transformations, but may not be deeply proficient in the nuances of operating large fleets of cloud infrastructure.

Using **Amazon ECS** with **AWS Fargate** and **AWS Batch** is a very capable serverless batch computing solution for background, asynchronous tasks or for data processing. Once data is present in a data repository, a job scheduler is employed, either through AWS Batch, or in combination of AWS Batch with Airflow or AWS Step Functions, to manage the creation, retry logic, and general organization of the process. As needed, Amazon ECS with AWS Fargate scales out either a long running-service or asynchronous jobs that will shut down once completed. AWS Lambda can also be used for shorter-lived jobs to complement Fargate tasks. Serverless compute resources automatically and quickly scale out to thousands of parallel compute units, and upon completion scale down, and offer a pay-per-use pricing model. Serverless compute integrates with a variety of data orchestration and data storage options so you can utilize storage as inputs and outputs for any batch processing workload with reduced operational overhead.



SUMMARY

Serverless data processing solutions are built using different serverless services in a variety of common architecture patterns depending on the type of workload, use case, and business need. Organizations often use one or both of AWS's primary serverless compute services which also integrate with other serverless offerings from AWS that play a big role across storage and analytics, including serverless data storage, serverless databases, and big data analytics solutions.

Using a serverless strategy, IT leaders are better able than ever to streamline data processing pipelines, increase agility, and optimize costs, without letting the management of IT infrastructure get in the way. With automatic resource scaling, serverless compute efficiently manages unpredictable surges in data volume without over-provisioning resources, thereby reducing unnecessary costs. And, with reduced operational overhead, data engineering teams can focus on building the core application, reducing time-to-market. This, in addition to serverless's pay-for-value billing model helps reduce total cost of ownership.

Via AWS serverless compute services organizations can leverage AWS best practices and expertise to improve performance, scalability, availability, and security, further reducing operational overhead, and improving reliability and security. AWS offers a broad range of serverless services and flexible integrations, so you can pick the right solution for the right workload.

Learn more at https://aws.amazon.com/serverless/.