amazon web services

EBOOK:

Building a Data Lake on AWS

Contents

What is a Data Lake?

Today's organizations are tasked with managing multiple data types, coming from a wide variety of sources. Faced with massive volumes and heterogeneous types of data, organizations are finding that in order to deliver insights in a timely manner, they need a data storage and analytics solution that offers more agility and flexibility than traditional data management systems. Data Lake is a new and increasingly popular way to store and analyze data that addresses many of these challenges. Data Lake allows an organization to store all of their data, structured and unstructured, in one, centralized repository. Since data can be stored as-is, there is no need to convert it to a predefined schema and you no longer need to know what questions you want to ask of your data beforehand.

A Data Lake should support the following capabilities:

- Collecting and storing any type of data, at any scale and at low costs
- Securing and protecting all of data stored in the central repository
- Searching and finding the relevant data in the central repository
- Quickly and easily performing new types of data analysis on datasets
- Querying the data by defining the data's structure at the time of use (schema on read)

Furthermore, a Data Lake isn't meant to be replace your existing Data Warehouses, but rather complement them. If you're already using a Data Warehouse, or are looking to implement one, a Data Lake can be used as a source for both structured and unstructured data, which can be easily converted into a well-defined schema before ingesting it into your Data Warehouse. A Data Lake can also be used for ad hoc analytics with unstructured or unknown datasets, so you can quickly explore and discover new insights without the need to convert them into a well-defined schema.

> "Data Lake allows an organization to store all of their data, structured and unstructured, in one, centralized repository."

Benefits of a Data Lake on AWS

There are a variety of benefits to hosting your Data Lake on AWS, including:

Cost-Effective Data Storage
Amazon S3 provides cost-effective and durable storage, allowing you to store nearly unlimited amounts of data of any type, from any source. Because storing data in Amazon S3 doesn't require upfront transformations, you have the flexibility to apply schemas for data analysis on demand. This enables you to more easily answer new questions as they come up and improve the time-to-value.

Easy Data Collection and Ingestion
There's a variety of ways to ingest data into your Data Lake, including services such as Amazon Kinesis, which enables you to ingest data in real-time; AWS Import/Export Snowball, a secure appliance AWS sends you for ingesting data in batches; AWS Storage Gateway, which enables you to connect on-premises software appliances with your AWS Cloud-based storage; or through AWS Direct Connect, which gives you dedicated network connectivity between your data center and AWS.

Security and Compliance
When you host your Data Lake on AWS, you gain access to a highly secure cloud infrastructure and a deep suite of security offerings designed to keep your data secure. As an AWS customer, you will benefit from a data center and network architecture built to meet the requirements of the most security-sensitive organizations. AWS also actively manages dozens of compliance programs in its infrastructure, helping organizations to easily meet compliance standards such as PCI DSS, HIPAA, and FedRAMP.

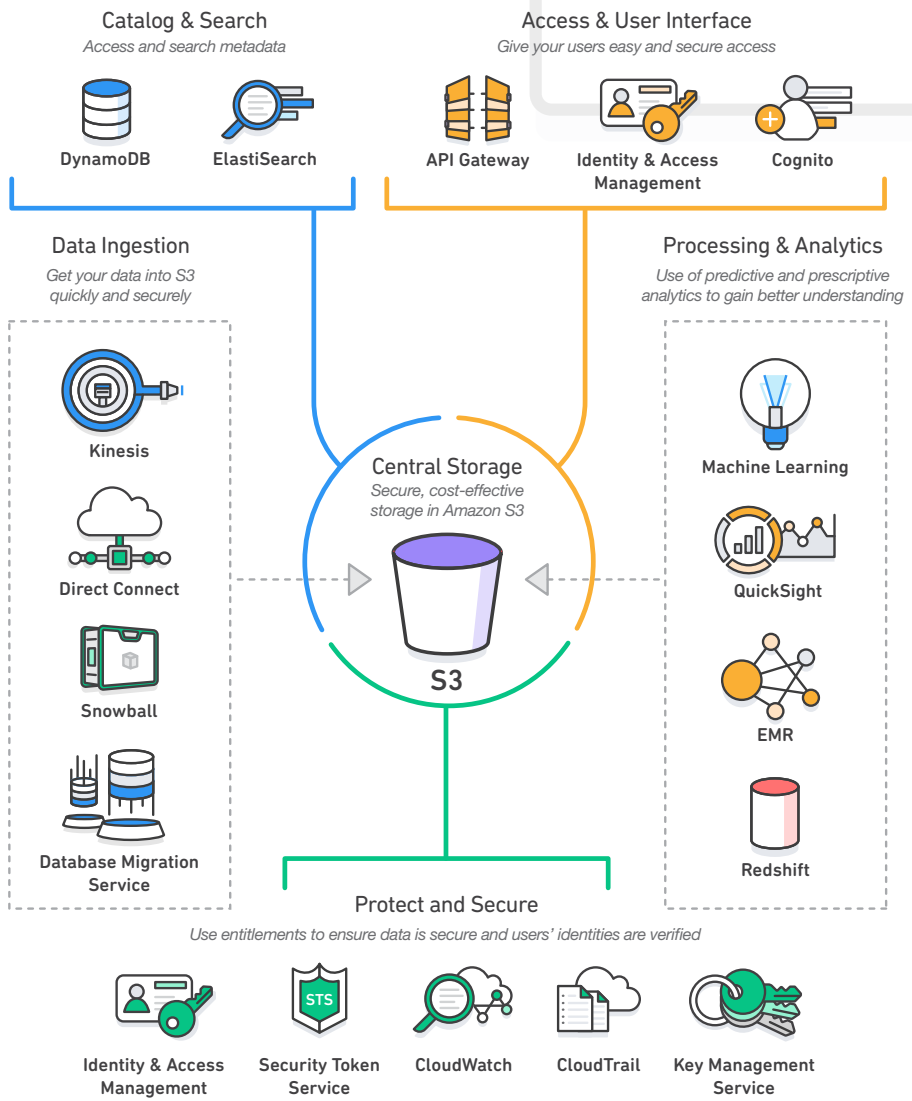
Most Complete Platform for Big Data
AWS gives you fast access to flexible and low cost IT resources, so you can rapidly scale virtually any big data application including data warehousing, clickstream analytics, fraud detection, recommendation engines, event-driven ETL, serverless computing, and internet-of-things processing. With AWS, you don't need to make large, upfront investments in time and money to build and maintain infrastructure. Instead, you can provision exactly the right type and size of resources you need to power big data analytics applications.

Building a Data Lake on AWS

A Data Lake solution on AWS, at its core, leverages Amazon Simple Storage Service (Amazon S3) for secure, cost-effective, durable, and scalable storage. You can quickly and easily collect data into Amazon S3, from a wide variety of sources by using services like AWS Import/Export Snowball or Amazon Kinesis Firehose delivery streams. Amazon S3 also offers an extensive set of features to help you provide strong security for your Data Lake, including access controls & policies, data transfer over SSL, encryption at rest, logging and monitoring, and more.

For the management of the data, you can leverage services such as Amazon DynamoDB and Amazon ElasticSearch to catalog and index the data in Amazon S3. Using AWS Lambda functions that are directly triggered by Amazon S3 in response to events such as new data being uploaded, you easily can keep your catalog up-to date. With Amazon API Gateway, you can create an API that acts as a “front door” for applications to access data quickly and securely by authorizing access via AWS Identity and Access Management (IAM) and Amazon Cognito.

For analyzing and accessing the data stored in Amazon S3, AWS provides fast access to flexible and low cost services, like Amazon Elastic MapReduce (Amazon EMR), Amazon Redshift, and Amazon Machine Learning, so you can rapidly scale any analytical solution. Example solutions include data warehousing, clickstream analytics, fraud detection, recommendation engines, event-driven ETL, and internet-of-things processing. By leveraging AWS, you can easily provision exactly the resources and scale you need to power any Big Data applications, meet demand, and improve innovation.



Featured Data Lake Partners

47Lining

47Lining is an AWS Advanced Consulting Partner with Big Data Competency designation. Big Data solutions developed by 47Lining are built from underlying AWS building blocks like Amazon Redshift, Amazon Kinesis, Amazon S3, Amazon DynamoDB, Amazon Machine Learning and Amazon Elastic MapReduce (EMR). With 47Lining, organizations don't need to start building a Data Lake from scratch.

NorthBay

NorthBay Solutions LLC is an AWS Advanced Consulting Partner with Big Data and Mobile Competencies. With over 5 years of experience and a team of over 200 employees, NorthBay has the expertise and manpower to create a Data Lake that meets the unique needs of any organization. NorthBay has furthered their expertise by aligning with AWS since the start, and were among the first organizations to architect a Data Lake on AWS.

Cloudwick

Cloudwick is the largest enterprise Big Data-as-a-Service provider and currently manages over 50,000 Big Data clusters on AWS. As an Advanced Consulting Partner with the APN Big Data Competency, Cloudwick has the expertise to make moving workloads and architecting your Data Lake simple by leveraging their proven 3-step methodology for performing Big Data migrations to AWS.

AWS Case Study: MLB Advanced Media

Major League Baseball Fields Big Data, and Excitement, with AWS (click here to read full study)

MLB Advanced Media (MLBAM) wanted a new way to capture and analyze every play using data-collection and analysis tools. It needed a platform that could quickly ingest data from ballparks across North America, provide enough compute power for real-time analytics, produce results in seconds, and then be shut down during the off season. It turned to AWS to power its revolutionary Player Tracking System, which is transforming the sport by revealing new, richly detailed information about the nuances and athleticism of the game—information that's generating new levels of excitement among fans, broadcasters, and teams.

It was a legend-making play for fans of baseball—a sport built on legends going back 150 years. In the third inning of the winner-takes-all Game Seven of the 2014 World Series, the San Francisco Giants and Kansas City Royals were tied at two. The Royals' Eric Hosmer hit the ball hard, driving it toward centerfield. If the ball cleared the infield, the hit could have sparked a rally.

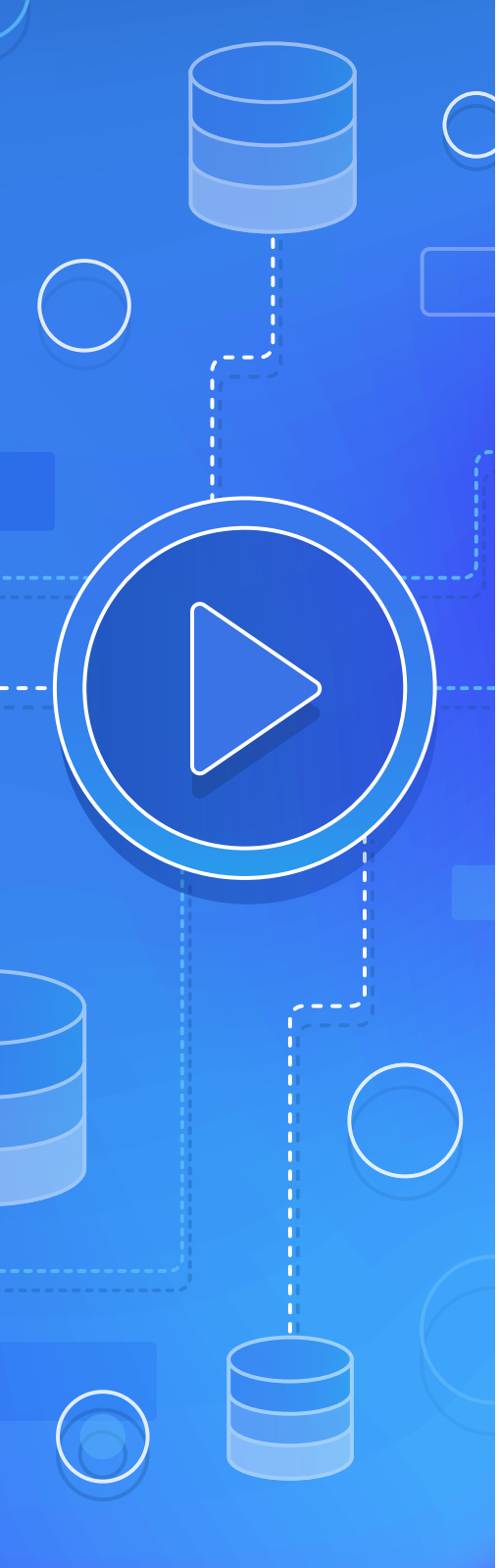
But Giants second baseman Joe Panik made an amazing dive to snatch the ball, resulting in two outs—including Hosmer, who was thrown out at first base after a diving attempt to beat Panik's throw. A possible Royals rally fizzled, and the Giants went on to win the game—and the World Series—by a single run.

Panik's play fueled plenty of talk on social media and in bars and broadcast booths. But more details about the play emerged from a system hosted in the cloud—a new big-data solution called the Player Tracking System, which MLB Advanced Media (MLBAM) created using Amazon Web Services (AWS).

The solution, which revealed Hosmer could have made it safely to first base by running through the base instead of diving, captures and analyzes the subtle complexities of every play in games. Launched into full production at all 30 MLB ballparks for Opening Day of the 2015 season, the Player Tracking System is generating new excitement with data delivered within seconds after the action occurs, including information sent to broadcast companies under the brand name "Statcast."

> "We believe the Player Tracking System powered by AWS will deliver new and more exciting information to apps and devices, and that will appeal to a younger generation of fans, who are used to video games and who have a lot of expectations about the viewing experience. It delivers a new level of excitement to baseball."

Dirk Van Dall
Vice President of Multimedia
Technology Development, MLBAM



Getting Started

For more information about Data Lakes on AWS, visit:

> Big Data on AWS

> About Data Lake on AWS

> Building a Data Lake on AWS (Video)

About AWS

For 10 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform. AWS offers over 70 fully featured services for compute, storage, databases, analytics, mobile, Internet of Things (IoT) and enterprise applications from 35 Availability Zones (AZs) across 13 geographic regions in the U.S., Australia, Brazil, China, Germany, Ireland, Japan, Korea, and Singapore. AWS services are trusted by more than a million active customers around the world – including the fastest growing startups, largest enterprises, and leading government agencies – to power their infrastructure, make them more agile, and lower costs.

To learn more about AWS, visit aws.amazon.com.