Data Flow for Analytics and Machine Learning

Extracting insights and actionable information from data requires a broad array of technologies that can work with data efficiently, scalably, and cost-effectively. AWS offers a comprehensive set of services to enable data science at enterprise scale and handle every step of the data processing chain—from extraction and streaming, to tagging and enrichment, through to Data Lakes, Business Intelligence, Machine Learning (ML) and Intelligent Devices. These services are powerful, flexible, and yet simple to use, enabling organizations to turn their raw data into a core corporate asset that is leveraged across all aspects of the business.

The adoption of data science in enterprises is creating significant shareholder value. Business decisions are more data-driven, and new products and product lines are created using actionable insights unveiled by Business Intelligence, Real-time Analytics, Machine Learning and Deep Learning. Discover the complete array of tools for pipelining, storing, enriching and serving data, as well as scalable computing environments for Data Lakes, Data Warehousing, Machine Learning and Intelligent Devices, with AWS.

Sources

Structured

Data that are highly normalized with common schema and stored in relational databases, powering transactional line-of-business applications. These data are easily accessible through SQL or data extraction tools.

ERP · CRM · LOB APPLICATIONS

Semistructured

Data that contain identifiers without conforming to a predefined schema, often stored in NoSQL databases, such as JSON and XML. These data are easily accessible but require some preparation in order to make them ready for data science.

MOBILE · SOCIAL · SENSORS · POS TERMINALS

Unstructured

Data that do not conform to a data model and are typically stored as individual files. The most common uses are text documents (such as email), pictures, audio and video.

PHONE CALLS · IMAGES · VIDEOS · EMAILS

Pipelining

Batch Load

Extracts data from various data sources at periodic intervals and moves them to the Data Lake. This process normally involves database queries and includes some transformation (also known as ETL or ELT).

Glue

Streaming

Ingests data that are generated continuously from multiple sources, such as log files, telemetry, mobile applications, IoT sensors and social networks. Data can be processed over a rolling time window and pipelined into a Data Lake.

Kinesis · Lambda · IoT

Real-time Analytics

Enables actionable insights for time-critical business processes that rely on streaming data analysis, including Machine Learning algorithms such as anomaly detection.

Neural Networks

Pretrained Deep Learning models that process unstructured data and prepare them for analysis. This includes images, video and speech recognition.

Rekognition · Rekognition Video · Translate · Transcribe · Comprehend

Data Lake

Cloud-scale centralized and scalable architecture that enables enterprise data science. Data Lakes are managed by Data Operations teams and include processes that categorize, standardize, aggregate, annotate and enrich data in order to facilitate the work of data scientists. Enrichment often includes the injection of third party sources, such as geographical, demographic and meteorological data.

S3 · Glue

Preprocessed data

Raw data extracted or streamed with minimal transformation. Data Lakes provide long-term storage of raw data for future use cases or for compliance and audit purposes.

CATEGORIZE · STANDARDIZE · AGGREGATE · ANNOTATE · ENRICH

Enriched data

The enrichment process provides an authoritative, single source of truth that can be consumed by data scientists, business analysts and ML/DL developers.

Third party sources

GEOGRAPHICAL · DEMOGRAPHIC · METEOROLOGICAL

Machine Learning

ML is the application of statistical and probabilistic techniques to derive analytical models that can make inferences and predictions. Data scientists build and train ML models using algorithms that are selected depending on the characteristics of the data and the use case; periodic retraining is necessary in order to keep the models updated with the latest available data.

Model Development

Sagemaker · EMR

Build

Connect to training data from Jupyter notebooks, perform data discovery and select the most appropriate framework, algorithms or ensemble.

Train

Spin up, manage and tear down the infrastructure needed to easily scale and train models, all the way to high-Terabyte or Petabyte scale. Includes automatic model tuning.

Deploy

Spin up elastic, highly available environments for models trained.

Run highly scalable, managed Spark clusters with EMR and take advantage of Spark MLlib or use SageMaker directly from EMR to train models at scale or using DL libraries.

Business Intelligence

Data Warehouses

Data warehouses are repositories of normalized data that are consolidated and aggregated from multiple sources and optimized for analytical use. Data warehouses provide the foundational technology for Business Intelligence.

Redshift · QuickSight

Data Lake Access

Data stored in the Data Lake can also be made directly searchable and queryable.

Athena · Redshift Spectrum · QuickSight

Once created, an ML model becomes production-grade software that can be hosted either in the cloud or at the edge.

Managed Endpoint

When hosted in the cloud, an ML model is made available through an API or REST endpoint that runs on elastic and highly available infrastructure. Data scientists often deploy multiple models and perform A/B testing to evaluate different options.

SageMaker

Intelligent Devices

When deployed to edge devices, ML models must rely on a transport and management layer to ensure that models are transferred reliably and updated successfully in the device fleet, which can have millions of edge points.

Greengrass

IoT Fleet Manager