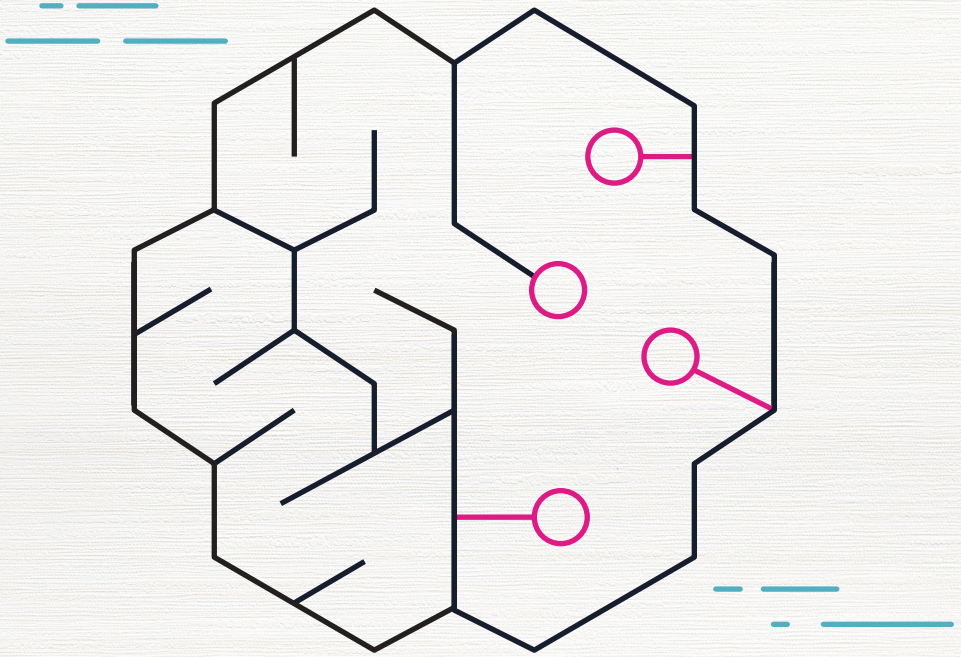
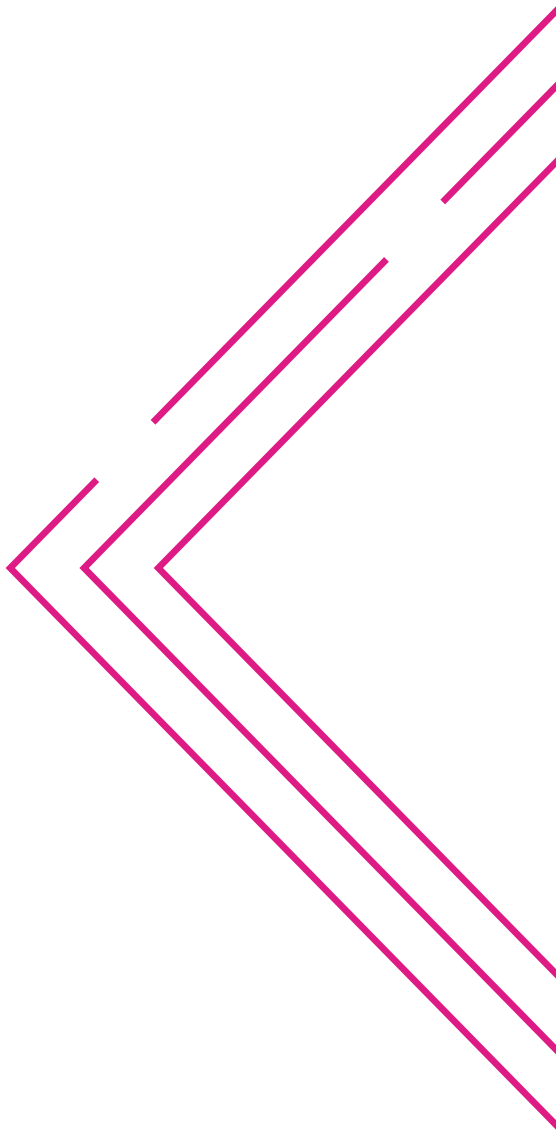
aws startups

Managing the data process for machine learning

A reference guide focusing on managing the data for your machine learning practice

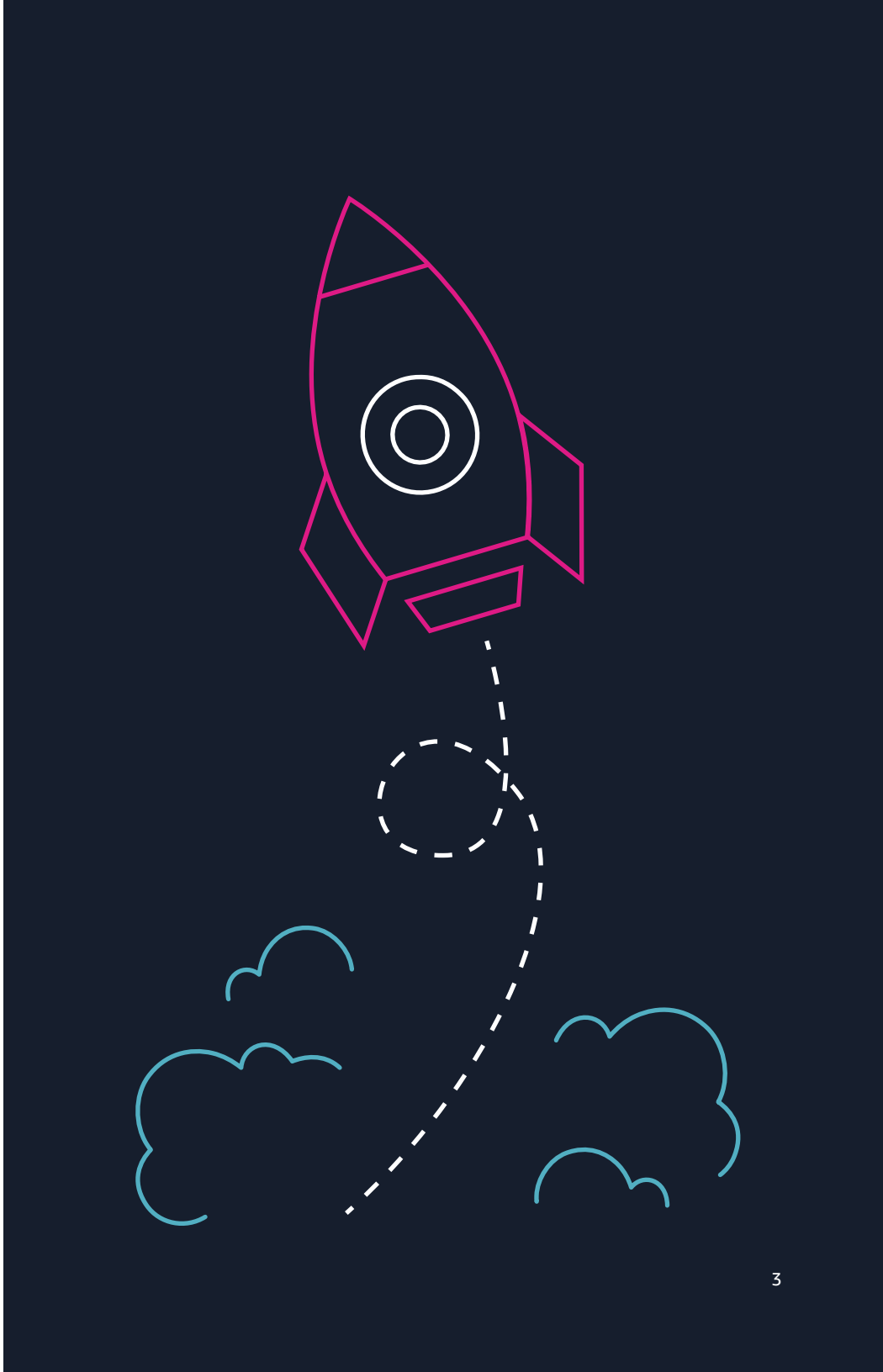


Contents



Introduction

In this e-book, we provide insights and practical guides for the core part of machining learning practice - data processing. We start out by providing a deeper understanding of machine learning data such as data acquisition, data quality and data sizing, followed by the relevant data management and processing services AWS offers along with the successful customers' case-studies in leveraging these services for managing their machine learning data. We hope the e-book lays the foundation for the machine learning data process and paves the way for you to continue the success in your machine learning practices in the subsequent stages.



Understanding the Machine Learning Data

Modern machine learning techniques rely heavily on data to build models. Understanding the various aspects of data for machine learning is critical to the success of your machine learning adoption. In this chapter, we look at a few of these important aspects related to machine learning data to have a better understanding when you start your machine learning projects.

Data Sources for Machine Learning

Your machine learning data sets can usually come from external data sources or internal data sources. External data sources can be further categorized as free open datasets versus private datasets offered by data providers.

External Public Machine Learning Data Sources

These are datasets produced or collected by organizations such as international bodies or agencies, national or local governments, research institutes etc. via historical operational data, research projects or machine learning challenges etc. Here are a few examples of popular public machine learning data sets:

- AWS Data Exchange
- Registry of Open Data on AWS
- Data | World Bank
- Data and maps – European Environment Agency (EEA)
- Public Data Sets: Amazon Web Services
- Google Public Data Explorer
- Competitions – Kaggle
- UCI Machine Learning Repository

These data sets are usually open to the public to access, use, and contribute to. For a list of these public data sources fit for machine learning, you can find them here and here.

External Private Machine Learning Data Sources

These are vendors who provide more proprietary and potentially value-added datasets for specific industries or research domains. These vendors usually offer the datasets through an exploration platform (e.g., Explorium), a marketplace (e.g., Datarade), or a machine learning SaaS service (e.g., Calligo). AWS also offers a service – AWS Data Exchange, which makes it easy to find, subscribe to, and use third-party data in the cloud.

Internal Machine Learning Data Sources

These are datasets within organizations acquired via customer data, business operational data, or their intellectual properties. These datasets can be stored in various internal systems e.g., databases, data warehouses, document stores etc. or in external vendors' SaaS platforms such as CRM or ERP systems. These are typically the main machine learning data sources companies try to leverage, but they are usually spread in silos, and are challenging to manage. We will discuss how you can leverage AWS services to consolidate and manage these datasets in more details in later chapters.

Data Quality of Machine Learning Datasets

"Garbage In, Garbage Out" has been a saying around since the early days of computing. This saying is even more relevant than ever in the age of artificial intelligence. Machine learning, as an important branch of artificial intelligence, uses large amounts of training data to build the machine learning models. By its very nature, a machine learning model is very sensitive to data quality. According to Dr. Andrew Ng's Data Centric AI reference, 80% of the AI developer's *time is spent on data preparation and time spent optimizing data is likely to improve model performance more than algorithmic efficiency*. Thus, as the first step in your machine learning project, you should check for *consistency, accuracy, compatibility, completeness, timeliness, and duplicate or corrupted records* of your machine learning data.

Measurement and Assessment of Data Quality

In the data science world, there are many methodologies used to measure and assess data quality. For example, data scientists follow these commonly used data quality measurement processes such as **Benchmarks**, **Consensus**, **Cronbach's alpha test** and **Review**[1] and use statistical notions such as Precision, Bias and Accuracy to assess data quality[2]. In a more practical sense, understanding what may impact your data quality (e.g., Data Measurement and Collection Errors, Noise, Outliers and Missing data)[2], and defining good metrics to measure them (e.g., Ratio of Data to Errors, Number of Empty Values)[3] will help you improve your data quality.

80%

of the AI developer's time is spent on data preparation and time spent optimizing data is likely to improve model performance more than algorithmic efficiency.

Improving Machine Learning Data and Model Quality

The best way to improve your machine learning model starts with good quality data sets for training, validation and testing. Thus, pre-processing your data sets (e.g., cleansing, transformation and imputation) is an important step. Amazon SageMaker provides several features such as SageMaker Data Wrangler in SageMaker Studio and SageMaker Processing to help visualizing and preparing the data to improve your machine learning data. SageMaker Clarify also helps you detect bias in your Machine Learning (ML) models to improve the model quality. We will describe these features in more details in the later chapters. There are also other considerations going beyond data cleansing and transformation [4]. One important thing to consider is to continue monitoring your ML data and models (e.g., to prevent model drifting) and retrain your model with refreshed data sets. Research towards automated data quality management for machine learning[5] has led to the development of Amazon SageMaker Data Quality Monitoring and Model Quality Monitoring features to continuously improve your machine learning models.

Data Sampling Size for Training Machine Learning Models

We know machine learning relies heavily on data. One commonly asked question is *"how much data do I need"*? Machine learning is complex enough with different learning paradigms (e.g., supervised and unsupervised learning), application domains (e.g., text analysis, natural language processing and image processing etc.) to solve different problems (e.g., classification, forecasting, anomaly detection, clustering etc.). There are also other factors to consider such as machine learning algorithms and frameworks used for training the models. Besides these factors, your specific requirements on model accuracy and training time and cost will also determine the data size you may need. So, **there is no one single answer to this question. It depends. However, we provide some general rule of thumb guide below, and you can adjust through experimentation to suite your needs**.

There are many researches on this topic about sampling data size or machine learning training data sets[6,7,8]. For simplicity as a general guideline, a simple rule of thumb is that the size of dataset should be at-least about 10x it's dimension and should be independent of the model used[9,10,11]. For other commonly seen machine learning domains or problem types, the general guide of sample data size is listed in the table to the right.

Use them as a general guide for your initial machine learning project planning. You should adjust the data set size based on the actual specific problems to solve, framework and algorithms used and other requirements to consider mentioned above.

Now that you understand the value and different aspects of machine learning data, in the following chapters we will discuss how AWS machine learning related services can help you collect, process and manage machine learning data to embark on your machine learning projects.

	Learning paradigm or domain	Problem types	Data size rule of thumb	Note	References
1	Supervised learning	Linear regression, binary or multi-class classifications	10:1 Ratio of data samples: dimension	e.g., XGBoost	9, 10, 11
2	Supervised learning	Time series forecasting	50-100 Observations	e.g., using ARIMA model	12, 13
3	Computer vision	Image classifications	1000 Images per class	e.g., CNN or DL on ImageNet datasets	14, 15, 16
4	Textual analysis	Document classification	>100 Documents per category*	e.g., document classification	17
5	Natural language processing (nlp)	Sentiment analysis	200-300 Words**	e.g., sentiment analysis using Yelp Academic Dataset	18

*,** NOTE: these numbers are mostly based on experimental data as a reference.

Managing Your Data with AWS

Now that you have gained a better understanding of the value and importance of data to Machine Learning, you will now learn about how to effectively use your data for ML. In this section, we will explore the common steps in the Machine Learning process, along with the AWS services that can help you easily and efficiently achieve the step.

Collecting Your Data

As mentioned earlier, machine learning today relies on data to build insightful and accurate models — and the first step is to identify what data is needed for your ML model, and evaluate the various means available for collecting that data to train your model. AWS provides you with a number of ways to ingest data in bulk from static resources, or from new, dynamically generated sources, such as websites, mobile apps, and internet-connected devices.

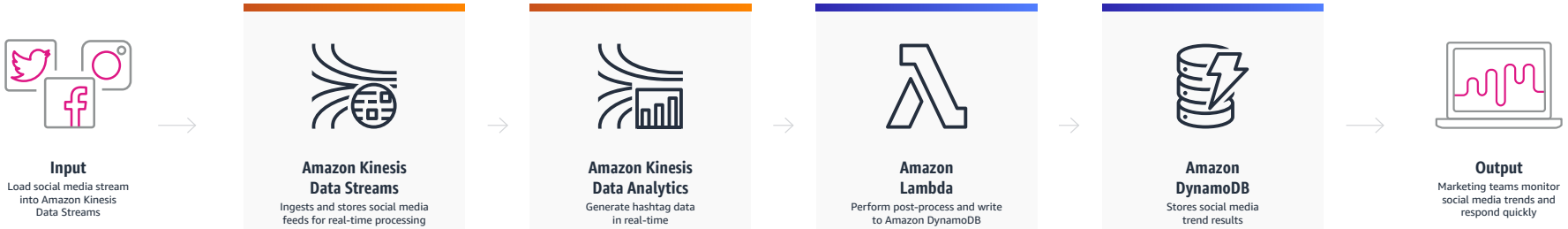
Amazon Kinesis Suite: *Streaming data* is data that is generated continuously by multiple data sources, which typically send in the data records simultaneously, and in small sizes (order of Kilobytes). Streaming data includes a wide variety of data such as application log files, e-commerce purchases, information from social networks, etc. Kinesis helps you collect, process, and store this streaming data:

- Amazon Kinesis Data Streams is AWS' scalable real-time data streaming service. KDS continuously captures GBs of data per second from thousands of sources and makes the data collected available in milliseconds to enable real-time analytics use cases.

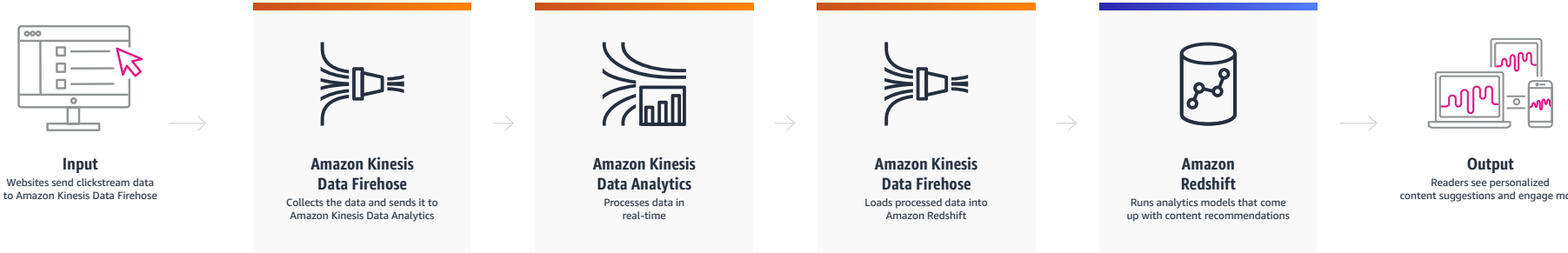
- Amazon Kinesis Data Firehose lets you easily load streaming data into data lakes, data stores, and analytics services. Kinesis Data Firehose is fully managed — it scales automatically to match the throughput of your data. It can also batch, compress, transform, and encrypt your data streams before loading.

- Amazon Kinesis Data Analytics enables you to easily and quickly build queries and sophisticated streaming applications in three simple steps: setup your streaming data sources, write your queries or streaming applications, and setup your destination for processed data.

Example: Analysis of streaming social media data



Example: Clickstream analytics



Cloud Data Migration: When moving data into the cloud, you need to understand where you are moving it for different use cases, the types of data you are moving, and the network resources available, among other considerations. AWS provides a portfolio of data transfer services to provide the right solution for any data migration project, which include hybrid cloud storage, online data transfer, and offline data transfer needs. See our Cloud Data Migration page for more details.

Third Party Tools: AWS wants to support all our customers and their different use cases. We understand that not all of your data may be native to the AWS cloud — as such, AWS also allows you to integrate with third-party data platforms. For example Snowflake, a popular data warehouse platform, supports integration with Amazon Simple Storage Service (Amazon S3), our massively scalable object-storage solution. This is done by giving Snowflake access to S3 via S3's access management policies, which we'll touch on in a later section.

Case Studies: Let's take a look at some Startups that have successfully developed ingest pipelines for collecting data to train their Machine Learning models. First is Depop, an alternative shopping experience driven by its marketplace for unique fashion[19]. Depop relies on Amazon Kinesis Data Firehose and Amazon Managed Streaming for Kafka to stream in its vast inventory of 25 million items and transactions. Leveraging managed services for ingesting data allowed Depop to focus on developing customer services, rather than managing infrastructure. Next is axialHealthcare, a company which ingests prescription and claims data for analysis and refers cases for intervention by teams of licensed healthcare practitioners[20]. axialHealthcare's cloud-based contact center uses Amazon Kinesis Data Streams to monitor agent status events and Amazon S3 to store call recordings as objects.

depop | axial Healthcare

Performing Analysis with Your Data

Understanding your data is vital to developing your ML models. Exploratory Data Analysis[21] refers to the process of performing initial investigations on data in order to discover patterns, spot anomalies, test hypothesis, etc. Additionally, by analyzing features in the context of our models, you gain intuition about which features are more important than others. Some features will improve model performance, while other features show no improvement or even reduce model performance. AWS provides services that are ready out of the box to explore the data that you've collected and stored in the cloud.

A key aspect to understanding your data is to identify patterns. These patterns are often not evident when you are only looking at data in tables. The correct visualization tool can help you quickly gain a deeper understanding of your data. Before creating any chart or graph, you must decide what you want to show. For example, charts can convey information such as key performance indicators (KPI), relationships, comparisons, distributions, or compositions.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run. With Athena, you can query your data in its raw form directly without additional transformations.

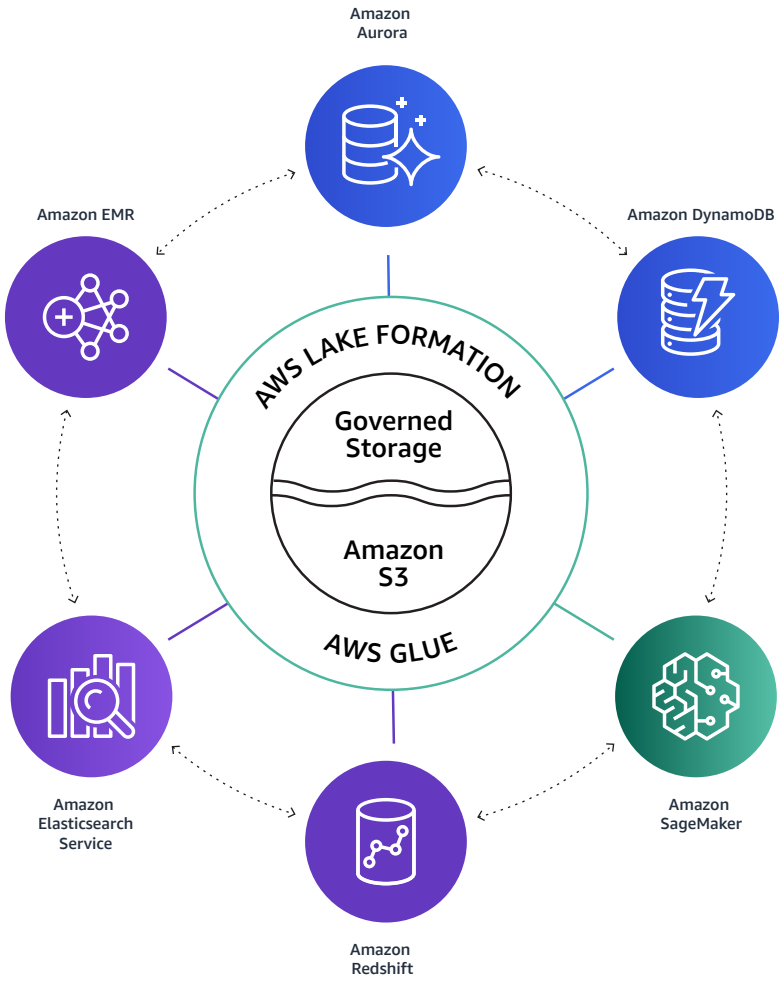
Amazon QuickSight is a scalable, machine learning-powered business intelligence service that lets you easily create and publish interactive dashboards. QuickSight also offers ML Insights, which leverages AWS's proven machine learning and natural language capabilities to help you uncover hidden insights and trends in your data, identify key drivers, and forecast business metrics. You can also create predictive dashboards by connecting QuickSight to your ML models built in Amazon SageMaker.

Amazon SageMaker Studio provides a single, web-based visual interface where you can perform all ML development steps. Studio features one-click Jupyter notebook that can be spun up quickly. The underlying compute resources are fully elastic, so you can easily spin up or down the appropriate compute power that you need. You can start analyzing data sets directly in your notebook environment with tools such as pandas, a popular Python open source data analysis and manipulation tool. For data visualization, you can use open source libraries such as Matplotlib and Seaborn.

Managing Your Data — Data Lake

In the past, accumulated data from operations resided in various data silos, thus making it very difficult to do analytics. Data silos present multiple challenges - the data needed for a given workload may be split across multiple silos and inaccessible; the silo where the data resides might not meet the cost requirements for a given workload; the silos may require different management, security, and authorization approaches, increasing operational cost and risk. With a data lake, data (structured and unstructured) across the enterprise can be centrally stored and cataloged in a highly scalable, available, secure, and flexible data store that can handle extremely large data sets for use cases such as machine learning and analytics. The Lake House architecture on AWS enables organizations to build data lakes and the surrounding purpose-built analytics stores and services, enabling data movement (from the inside out, from the outside in, and around the side) between the data lake and the surrounding purpose-built analytics stores and services to derive insights from their data[22,23].

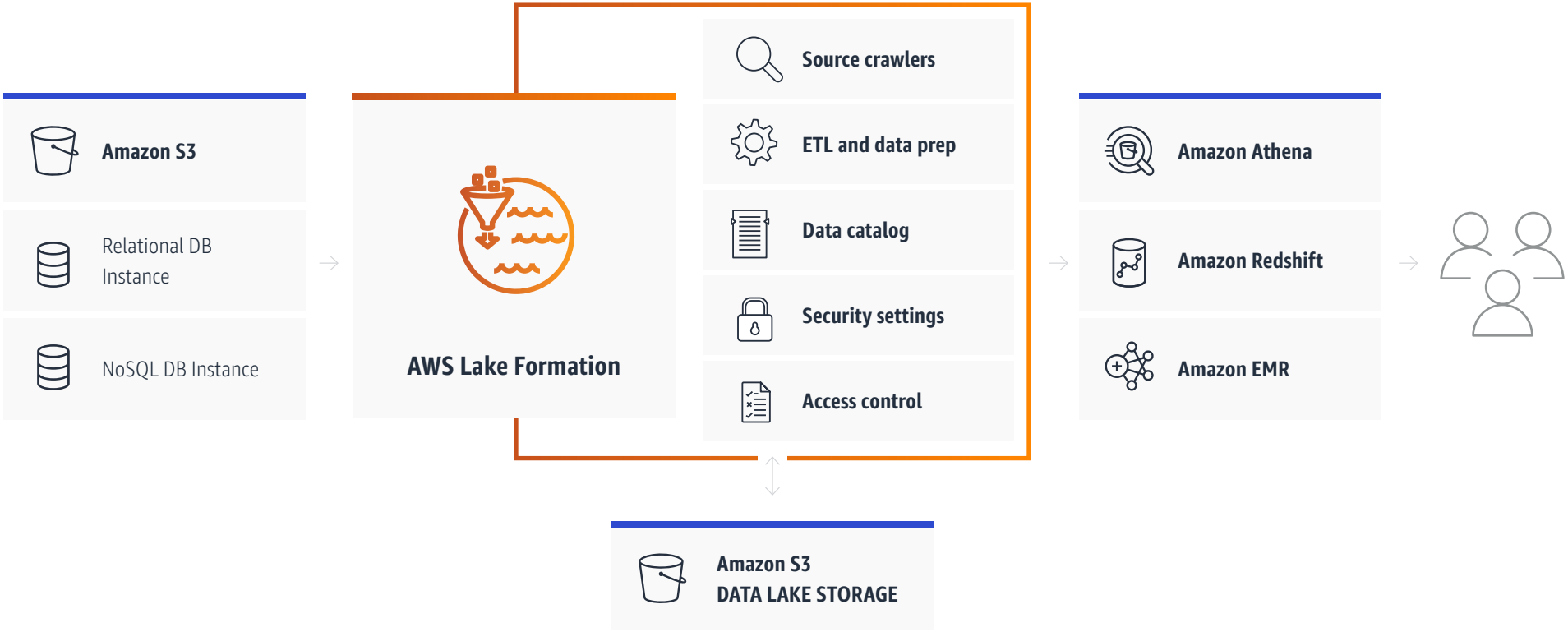
Central to the Lake House architecture is the data lake, and it starts with Amazon S3. Amazon S3 is the largest and most performant object storage service for structured and unstructured data and the storage service of choice for building data lakes[24]. You can build a scalable and cost-effective data lake of any size in a secure environment where data is protected by 99.999999999% (11 9s) of durability. Amazon S3 offers a wide range of cost effective S3 Storage Classes, which support different data access levels at corresponding rates. You can use S3 Storage Class Analysis to discover data that should move to a lower-cost storage class based on access patterns, and configure an S3 Lifecycle policy to execute the transfer. You can also store data with changing or unknown access patterns in S3 Intelligent-Tiering, which tiers objects based on changing access patterns and automatically delivers cost savings. Customers have been building data lakes on Amazon S3 longer than any other cloud provider and today, more data lakes run on AWS than anywhere else, and using Amazon Athena, a serverless interactive query service, to analyze all their data.



In addition to the data lake, customers use a combination of AWS purpose-built database and analytics services like Amazon EMR, Amazon Elasticsearch Service, and Amazon Redshift to ensure they are using the right tool for the job to get high performance and scale at the lowest possible cost. Customers use AWS Glue, a serverless data integration service, to move data between these systems. With AWS Lake Formation, customers can manage the security and governance of all their data regardless of whether it sits in the data lake or in purpose-built analytics stores. More specifically, AWS Lake Formation helps customers build data lakes with a single security and governance control across all the data, data sensitivity controls using fine-grained access controls, and centralized audit controls for the entire data lake.

One of the purpose-built data stores outlined in the Lake House Architecture is for machine learning use cases. The data journey begins with raw data ingested into the data lake and goes through the data transformation/ feature engineering process to create features for model training and inference. Once created, these features need to be stored, discovered and shared among data scientists for their own respective model training/ inference where different models can share common features across teams (as shown in the following figure). More importantly, to improve the models, data scientists will add new features over time and don't want to re-create or re-compute the existing features because doing so would be time consuming and adds latency to model predictions.

How it works



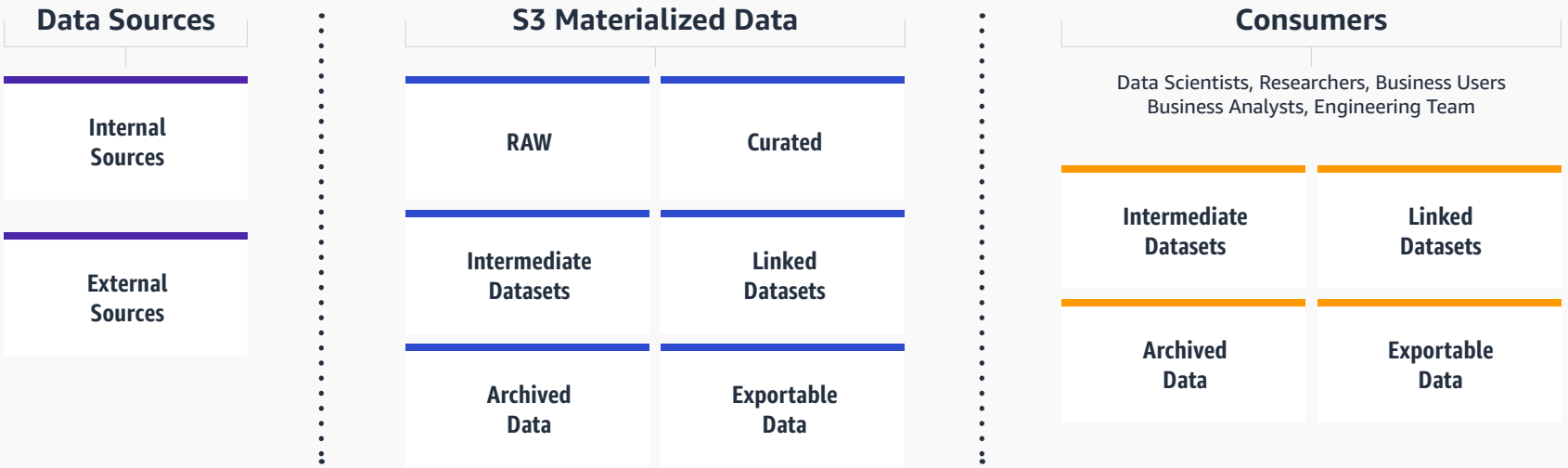
Amazon SageMaker Feature Store can help data scientists share common data features with other data scientists across teams for model training and inference. SageMaker Feature Store supports both offline feature store for training and online feature store for online inferencing[25]. SageMaker Feature Store serves features in large batches for model training and also serve features with low millisecond latency reads for real-time inference use cases[26]. Because Feature Store offers a central repository for model features, you have a consistent view of your features. In other words, the exact same features are available for training and inference, so your features never get out of sync between training and inference. You can visually search for and discover features in SageMaker Studio. All members of your team can share features in the repository, promoting reuse and eliminating rework. Feature Store also offers a unified set of features definitions across teams, making it easier for teams to work together.

The diagrams below conceptually illustrate model training using dedicated standalone feature engineering versus a shared feature store.

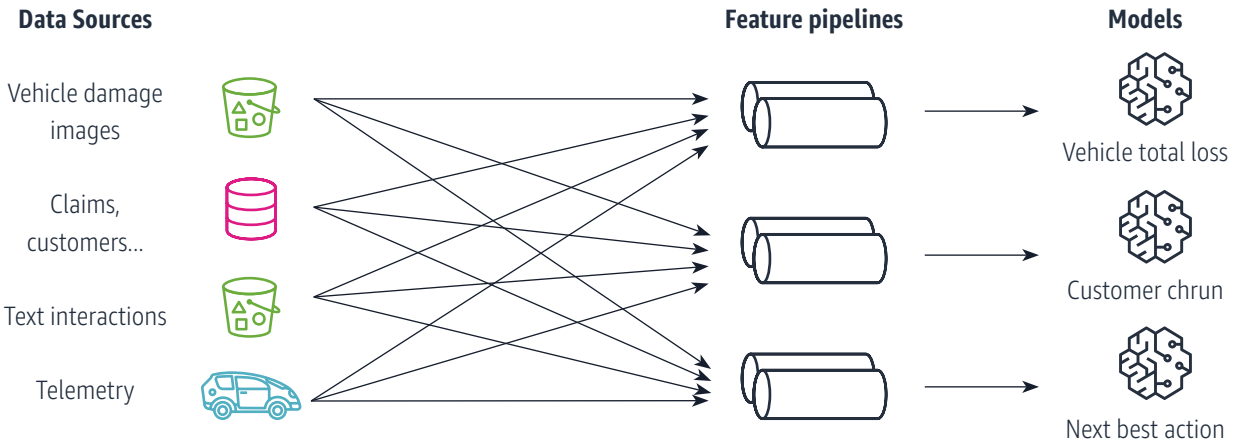
Case Studies: HappyFresh, an online grocery platform, developed their Data Lake on S[27]. The startup stores its clickstream data on Amazon S3 and uses AWS Glue to extract the data and process it for analysis of customer shopping patterns. Managing data with S3 allowed HappyFresh to focus on personalization services, and ensure customers are not losing valuable time searching for products.

happyfresh

Data Journey

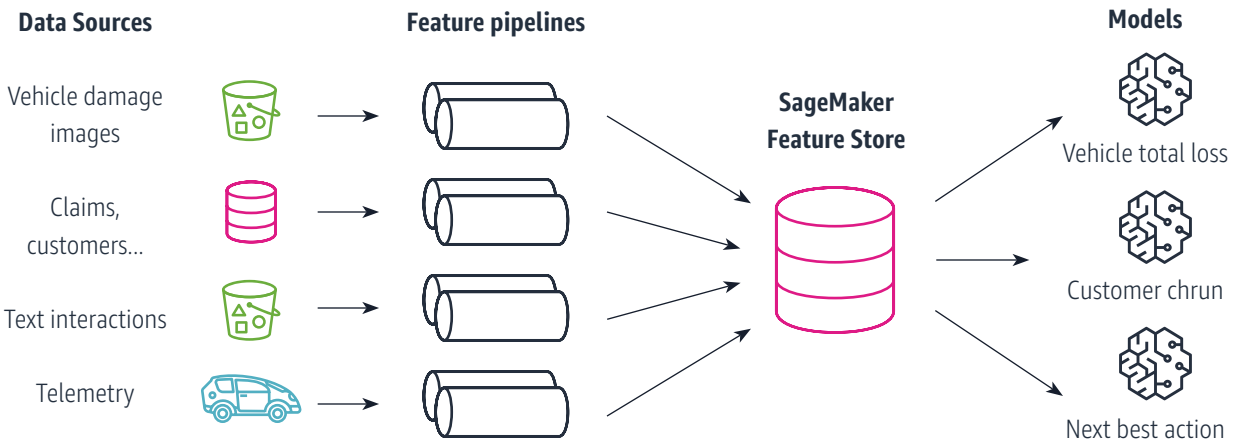


Standalone feature engineering for each new model



- Feature duplication
- Slow time to market
- Inaccurate predictions

Build features once, reuse them across teams and models using Feature Store



- Feature groups discoverable via search
- Reproducible feature transformations
- Extract accurate training datasets
- Low latency lookups for interference
- Consistent features for training and interference

Processing Your Data

Once you've collected data, you'll need to consider how you will process it to get it into a form that your models can effectively use. Examples of processing steps include converting data to the input format expected by the ML algorithm, rescaling and normalizing columns, cleaning and tokenizing text, and many more.

ML models are only as good as the data that is used to train them. After you collect data, the integration, preparation, processing, etc. of that data is critical. Training data that is optimized for learning and generalization is key to successful and accurate models. Data preparation should start with a small, statistically valid sample, and iteratively be improved with different data preparation strategies, while continuously maintaining data integrity[28]. AWS provides several services that you can use to annotate your data, extract, transfer, and load (ETL) data at scale.

AWS Glue: Once you have your data, you'll need to prepare and combine it for analytics and machine learning, in what's commonly referred as "Data Integration". AWS Glue provides both visual and code-based interfaces to make data integration easier. You can configure Glue to run periodically on a schedule, or through a trigger — for example, you may want to run Glue when new data lands in your S3 bucket.

AWS Batch: processing power. AWS provides GPU instance families, such as G4 and P4, which allow customers to run scalable GPU workloads. With /AWS Batch, you can run multiple batch computing jobs on AWS efficiently and at scale. It dynamically provisions the optimal number and type of compute resources based on the batch jobs you submit. Moreover, AWS Batch ensures that submitted jobs are scheduled and placed onto the appropriate instance, hence managing the lifecycle of the jobs. With the addition of customer-provided AMIs, AWS Batch users can take advantage of this elasticity and convenience for jobs that require GPU.[29]

Amazon EMR: Many organizations use Spark for data processing[30] and in these situations, Spark clusters are typically run in Amazon EMR, a managed service for Hadoop-ecosystem clusters that reduced the need to do your own setup, tuning, and maintenance. EMR lets you run custom jobs with your desired compute, memory and storage parameters. It provides automated cluster setup and autoscaling, and supports Spot instances for cost-savings. With EMR, you can quickly and cost-effectively perform operations on vast amounts of data, including importing, exporting, and joining your data sets.

Case Studies: Guru is a startup that provides knowledge management software[31]. Guru uses Amazon OpenSearch service to manage the storage and scaling of its Elasticsearch cluster. The company was also able to use Amazon EMR to develop an experimentation framework for improving the search result relevance of its search engine. Next, AiCure is an AI and advanced data analytics company that monitors patient behavior and enables remote patient engagement in clinical trials[32]. AiCure leverages AWS Step Functions and AWS Batch to continuously improve AI models and inferencing at scale to make data actionable.

Amazon SageMaker

Data schemas are a great first step for organizing and working with your data. But keep in mind that schemas evolve, code gets old, and queries get slow. Data Centric AI focuses on refining your data to provide high quality data for your downstream consumers, including ML engineering and data science teams — and this is an iterative approach. Data quality can halt a data processing pipeline in its tracks. If these issues are not caught early, they can lead to misleading reports, biased AI/ML models, and other unintended data products.

SageMaker Ground Truth: Accurately labeled data is vital to a supervised model's success. If there are incorrect labels, you ML model will learn from 'bad' examples, which leads to inaccurate predictions. SageMaker Ground Truth helps you efficiently and accurately label your data. It uses a combination of automated and human data labeling. You can also build custom workflows to define the user interface (UI) for data labeling jobs. To help you get started, Amazon SageMaker provides custom templates for image, text, and audio data labeling jobs — blog.

SageMaker Data Wrangler is designed specifically for machine learning, data analysis, feature engineering, feature-importance analysis, and bias detection. Data Wrangler provides over 300 built-in data transformations for feature engineering and bias mitigation.

SageMaker Processing Jobs can run any Python script or custom Docker image on the fully managed, pay-as-you-go AWS infrastructure using familiar open source tools such as scikit-learn or Apache Spark. This service can parallelize your custom scripts or docker images over many SageMaker instances in a cluster. With SageMaker Processing, you simply need to provide your script, and specify your instance type and cluster size.

SageMaker Clarify: In addition to detecting bias with SageMaker Data Wrangler, SageMaker Clarify helps select the best columns (aka "features") for model training, detects bias in our models after training, explains model predictions, and detects statistical drift of model prediction inputs and outputs.

Prepare →

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler NEW
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

SageMaker Feature Store NEW
Store, update, retrieve, and share features

SageMaker Clarify NEW
Detect bias and understand model prediction

Build →

SageMaker Studio Notebooks
Jupyter notebooks with elastic compute and sharing

Built-in and Bring-your-own Algorithms
Dozens of optimized algorithms or bring your own

Local Mode
Test and prototype on your local machine

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker Jumpstate NEW
Pre-built solutions for common use cases

Train & tune →

One-click training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare every step

Automatic Model Tuning
Hyperparameter optimization

Distributed Training Libraries NEW
Training for large datasets and models

SageMaker Debugger NEW
Debug and profile training runs

Managed Spot Training
Reduce training costs by 90%

Deploy & manage →

One-click deployment
Full managed, ultra low latency, high throughput

Kubernetes & Kubeflow Integration
Simplify Kubernetes-based machine learning

Multi-Model Endpoints
Reduce costs by hosting multiple models per instance

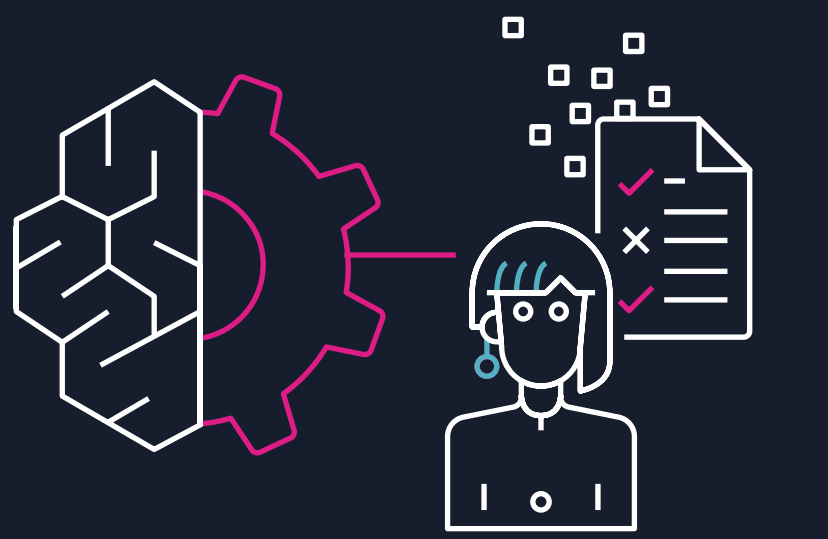
SageMaker Model Monitor
Maintain accuracy of deployed models

SageMaker Edge Manager NEW
Manage and monitor models on edge devices

SageMaker Pipelines NEW
Workflow orchestration and automation

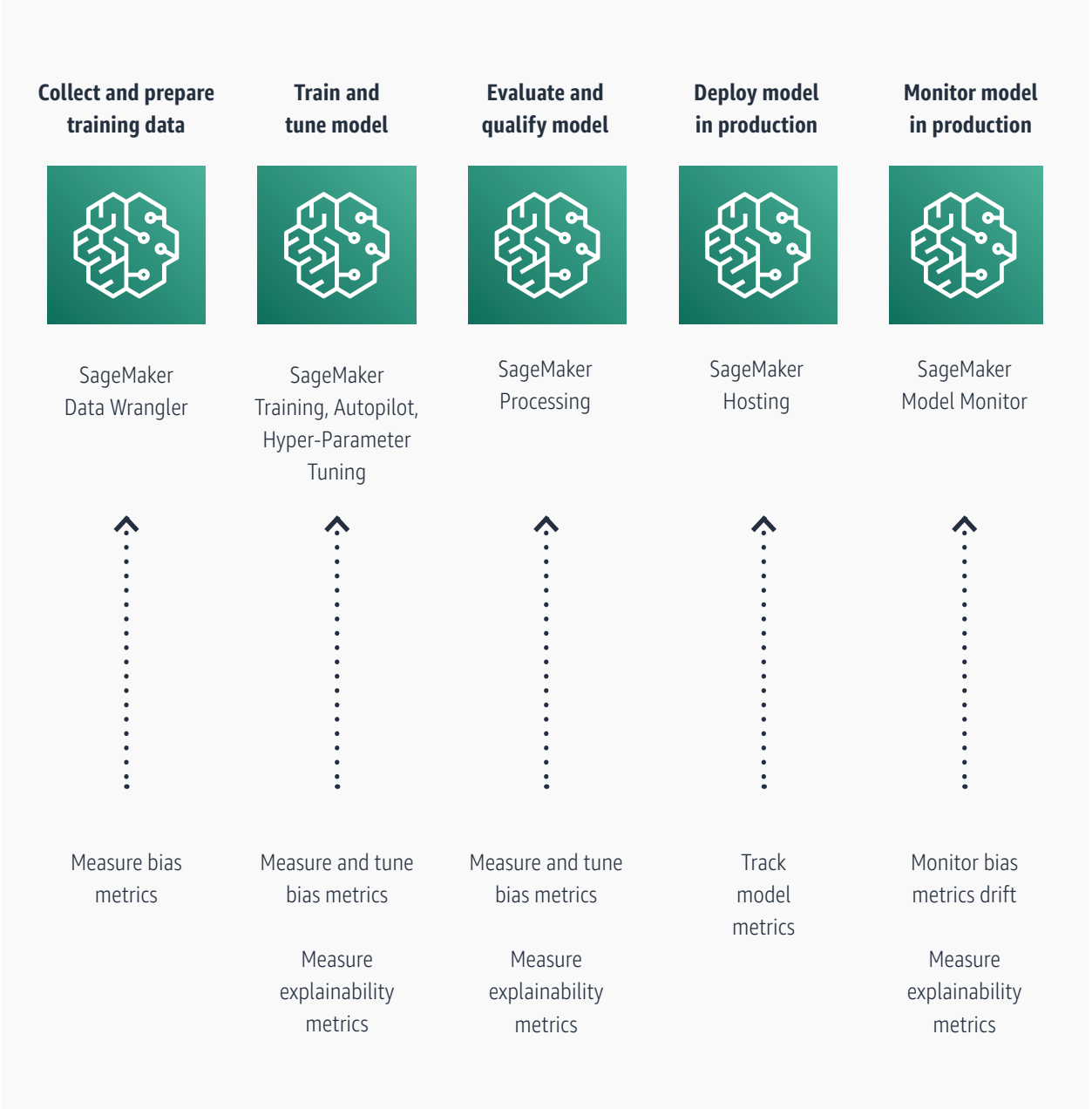
Putting Data in Practice

After learning how to manage your data to get the most value when doing Machine Learning, we will now explore AWS tooling that are helpful when training and tuning your models.



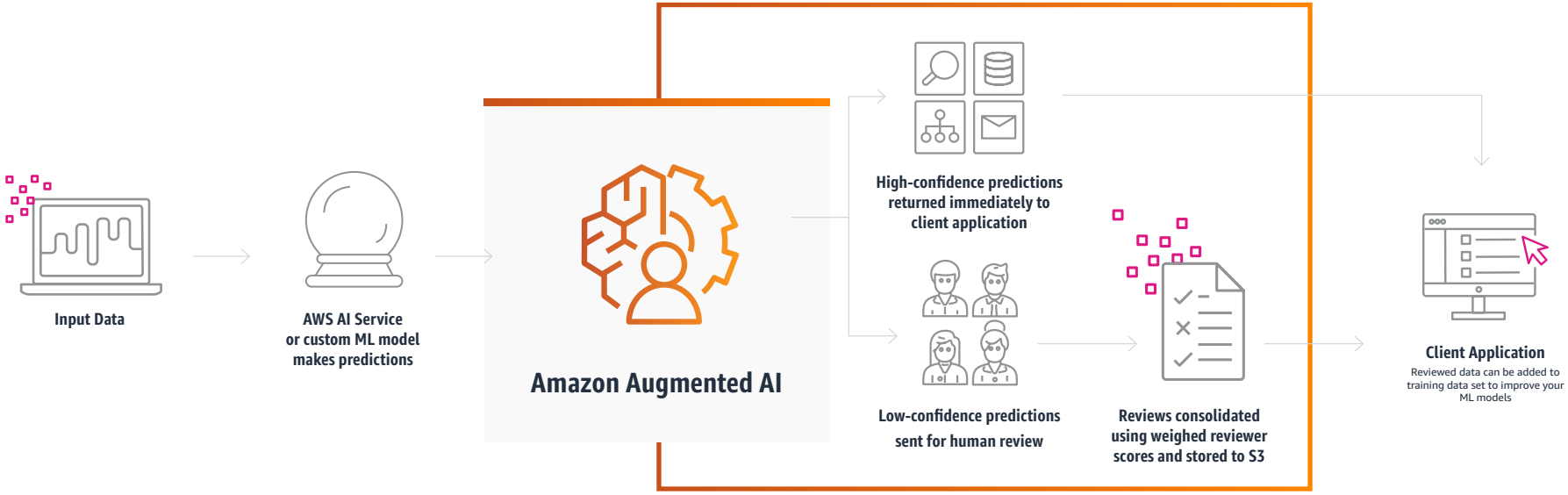
Amazon Augmented AI (Amazon A2I) makes it easy to build the workflows required for human review of ML predictions. Amazon A2I brings human review to all developers, removing the undifferentiated heavy lifting associated with building human review systems or managing large numbers of human reviewers.

Many machine learning applications require humans to review low confidence predictions to ensure the results are correct. For example, extracting information from scanned mortgage application forms can require human review in some cases due to low-quality scans or poor handwriting. But building human review systems can be time consuming and expensive because it involves implementing complex processes or "workflows", writing custom software to manage review tasks and results, and in many cases, managing large groups of reviewers.



Amazon A2I makes it easy to build and manage human reviews for machine learning applications. Amazon A2I provides built-in human review workflows for common machine learning use cases, such as content moderation and text extraction from documents, which allows predictions from Amazon Rekognition and Amazon Textract to be reviewed easily. You can also create your own workflows for ML models built on Amazon SageMaker or any other tools. Using Amazon A2I, you can allow human reviewers to step in when a model is unable to make a high confidence prediction or to audit its predictions on an on-going basis.

- Amazon A2I can also be incorporated into data set labels. This blog details how to use Amazon A2I to review and augment low confidence data.



Amazon SageMaker Studio: As mentioned earlier, SageMaker Studio provides a single web-based visual interface where you can perform all the machine learning development steps, improving your data science team's productivity. All the machine learning development activities including notebooks, experiment management, automatic model creation, debugging, etc. can be performed in Studio. In this guide, we covered various AWS services for the different stages of managing your data. As seen in the table below, Studio natively integrates SageMaker's features in data management and machine learning. By using Studio, your team can consolidate parts of your workloads into a single view for easier management and faster collaboration.

- When file workloads/performance is important to your team, Amazon FSx for Lustre can be an input data source for Amazon SageMaker. When FSx for Lustre is used as an input data source, Amazon SageMaker ML training jobs are accelerated by eliminating the initial S3 download step. SageMaker jobs can get started as soon as the FSx for Lustre file system linked with the S3 bucket is created without needing to download the full machine learning training data set from S3. Data is lazy loaded as needed from Amazon S3 for processing jobs. Another benefit is reduced TCO by avoiding the repeated download of common objects (saving S3 request costs) for iterative jobs on the same dataset.

Projects	**automates the model building and deployment pipelines using continuous integrations and continuous delivery (CI/CD)**
Data Wrangler	**aggregate and prepare data for machine learning**
Feature Store	**fully managed, purpose-built repository to store, update, retrieve, and share machine learning (ML) features**
Pipelines	**create, automate, and manage end-to-end ML workflows at scale**
Experiment and Trials	**organize, track, compare, and evaluate your machine learning experiments**
Model Registry	**track lineage & metadata of models**
Endpoints	**track your prediction endpoints**

Conclusion

In today's machine learning landscape, data is critical to build successful models. In this guide, we covered potential data sources for you to use on machine learning models, as well as the methods to measure your data quality and volume. We then discussed how AWS services and features can help you throughout the different steps needed to manage your data and prepare it for machine learning.

As you get more comfortable with the data management process and start looking towards using the data for machine learning, please visit AWS' Machine Learning page to explore machine learning services, hear about our customer's success stories, and browse common use cases. At AWS, our mission is clear: we aim to put machine learning in the hands of every developer — we're thrilled to help you understand and leverage your data.

Please reach out to your AWS Account Manager and Solutions Architect so they can help accelerate your data and ML journeys.

Free offers and services you need to build, deploy, and run machine learning applications in the cloud.

Get Started >>

References

[1] How to Define and Measure Your Training Data Quality.
[2] Assessing the Quality of Data.
[3] How to Measure Data Quality – 7 Metrics to Assess the Quality of Your Data .
[4] Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations.
[5] Towards Automated Data Quality Management for Machine Learning.
[6, 7] Predicting Sample Size Required for Classification Performance.
[8] How large a training set is needed?
[9] Sample size planning for classification models.
[10] Data set size versus data dimension, is there a rule of thumb?
[11] How much training data do you need?
[12] What should be the minimum number of observations for a time series model?
[13] Intervention analysis with applications to economic and environmental problems.
[14] What is the minimum sample size required to train a Deep Learning model - CNN?
[15] How Do You Know You Have Enough Training Data?
[16] How many images do you need to train a neural network?
[17] Text Analysis 101: Document Classification.
[18] How Much Text Do We Really Need for Sentiment Analysis?
[19] Depop Case Study.
[20] AxialHealthcare Case Study.
[21] What is Exploratory Data Analysis?
[22] What is a Lake House approach?
[23] Derive Insights from AWS Lake House.
[24] Data Lake Storage on AWS.
[25] Build a Secure Enterprise Machine Learning Platform on AWS.
[26] Create, Store, and Share Features with Amazon SageMaker Feature Store.
[27] HappyRefresh Case Study.
[28] Well-Architected Machine Learning Lens - Data Preparation.
[29] Deep Learning on AWS Batch.
[30] Data processing options for AI/ML.
[31] GURU Case Study.
[32] AiCure Case Study.


aws
startups