

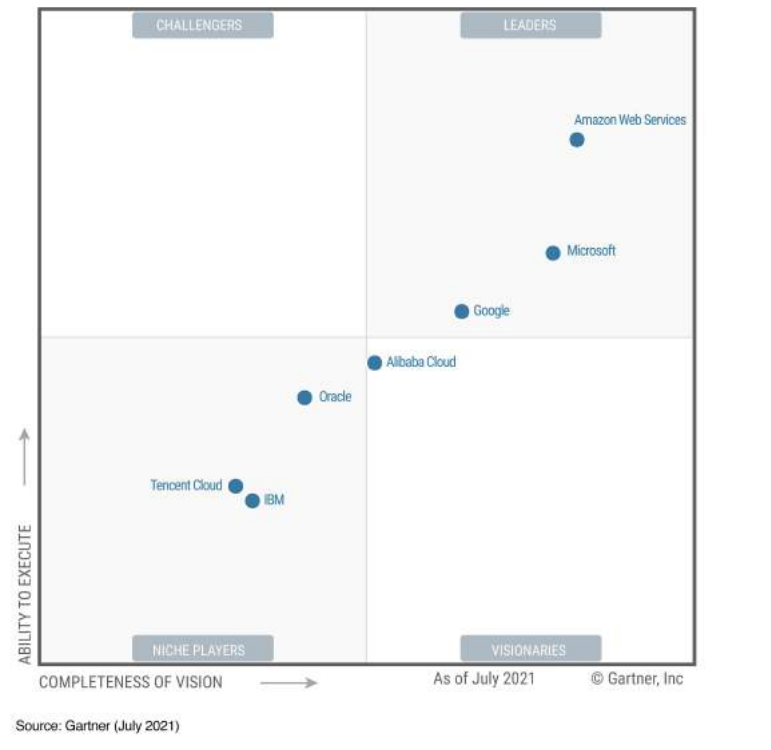
AWS.ML создание и запуск моделей ML в облаке AWS

Aleksandr Patrushev
Sr. AI/ML Solutions Architect

Agenda

- AWS cloud overview
- ML in VideoTech
- Architecture review
- ML pipelines and optimization
- Summary

Figure 1: Magic Quadrant for Cloud Infrastructure and Platform Services



Gartner recognizes AWS as a Leader for the 11th straight year

Magic Quadrant for Cloud Infrastructure and Platform Services

Gartner, Magic Quadrant for Cloud Infrastructure & Platform Services, Raj Bala, Bob Gill, Dennis Smith, Kevin Ji, David Wright, 27 July 2021. Gartner and Magic Quadrant are registered trademarks of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose. This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from AWS.

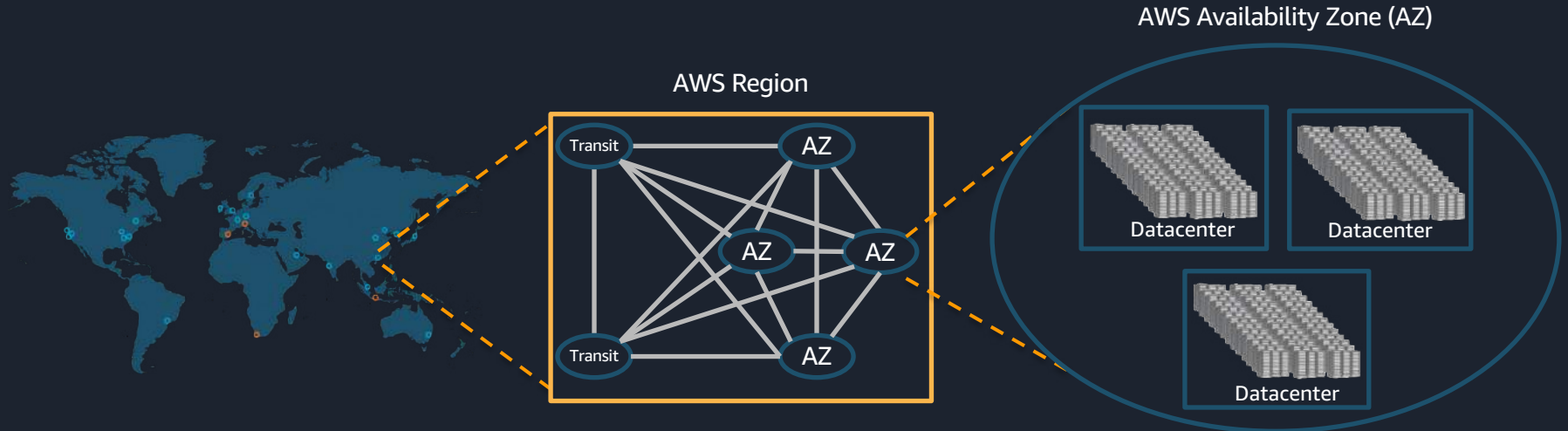
Regional expansion

- First 5 years: 4 regions
- Next 5 years: 7 regions
- 2016–2020: 14 regions
- Coming soon: 8 regions



AWS Region design

AWS Regions are comprised of multiple AZs for **high availability**, **high scalability**, and **high fault tolerance**. Applications and data are replicated in real time and consistent in the different AZs.



A Region is a physical location in the world where we have multiple **Availability Zones**.

Availability Zones consist of one or more discrete data centers, each with redundant power, networking, and connectivity, housed in separate facilities.

AWS network design

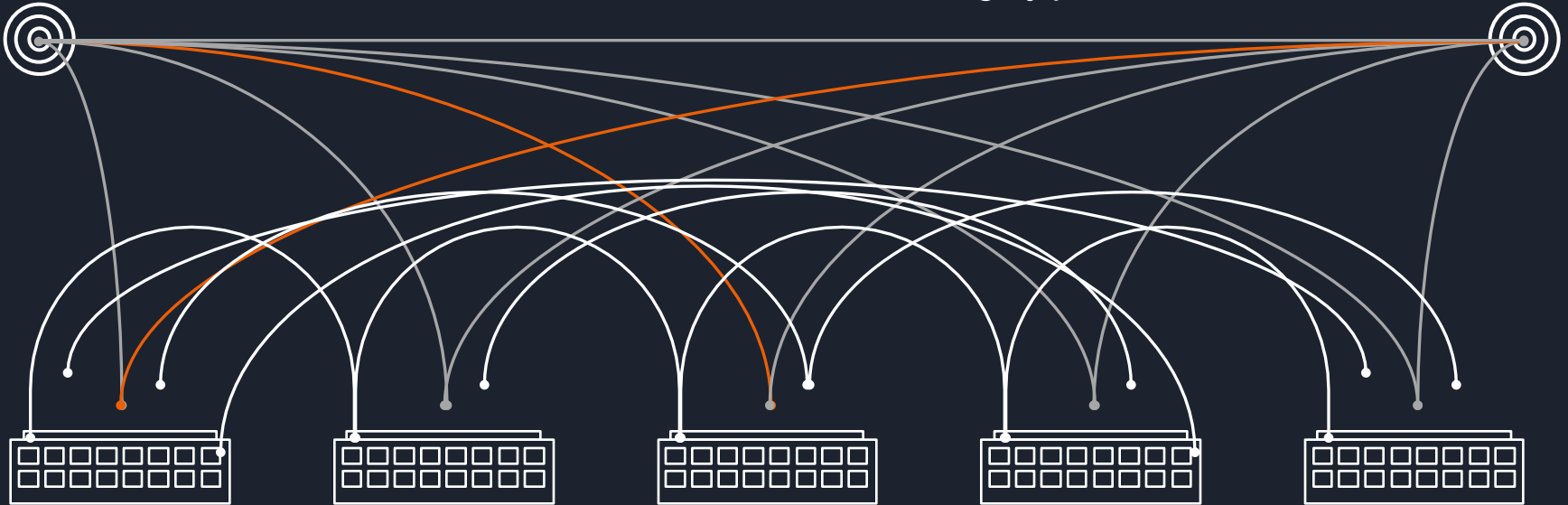
— Intra-AZ connections

— Inter-AZ connections

— Transit center connections

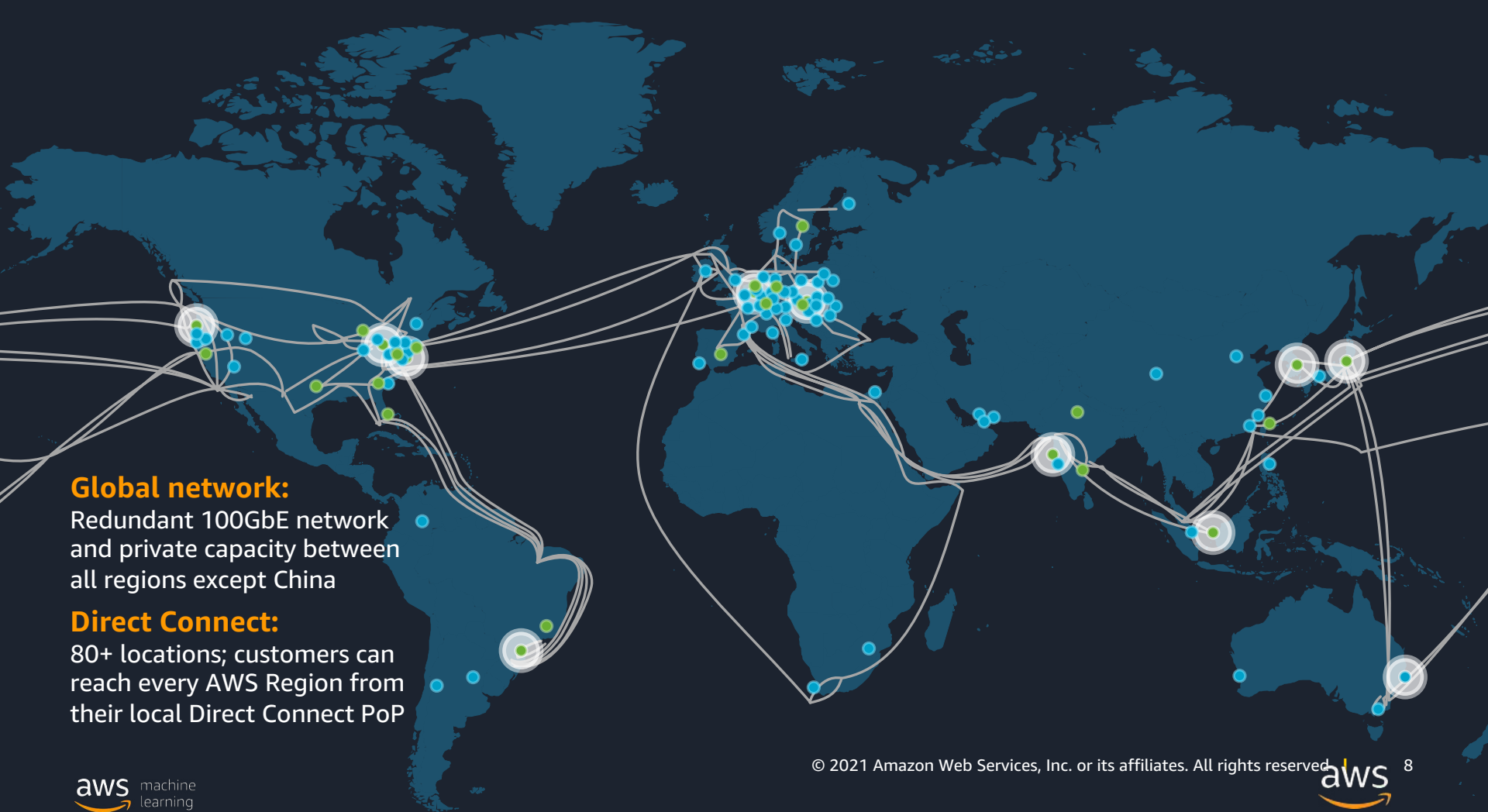
At least 2 redundant transit centers

Highly peered & connected



A world map with a dark blue background. Numerous small blue and green dots are scattered across the map, representing points of presence. Several of these dots are grouped together and enclosed in white circles with a grey border, indicating Regional Edge caches. The dots are most densely clustered in North America, Europe, and East Asia.

Amazon CloudFront
uses a global network of
230+ points of presence
(218 Edge locations and 12
Regional Edge caches) in 77
cities across 37 countries



Global network:

Redundant 100GbE network and private capacity between all regions except China

Direct Connect:

80+ locations; customers can reach every AWS Region from their local Direct Connect PoP

AWS custom hardware



Routers



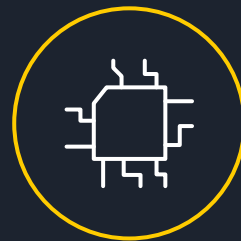
Load
balancers



Custom compute
servers and
semiconductors



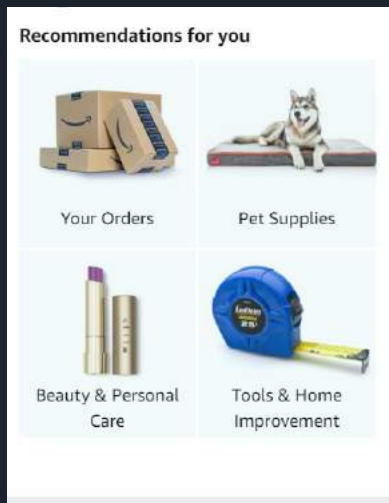
Custom
software



Silicon

200+ fully featured services

Amazon's machine learning innovation at scale



4,000 products per minute sold on Amazon.com



1.6M packages every day



Billions of Alexa interactions each week











First Prime Air Delivery on **Dec. 7, 2016**

Что значит «создание и запуск моделей ML»

1. Новые данные / Новый алгоритм
2. Существующие данные / Новый алгоритм
3. Новые данные / Существующий алгоритм
4. Существующие данные / Существующий алгоритм

Media use cases enabled by Machine Learning

Search	Search & Discoverability		Automate the creation of rich metadata index (object, activity, faces), extracted from audio visual content, integrated into asset management systems
Quality Control & Compliance	Subtitling & Localization		Automate the creation of subtitles, captions, transcription and translations of audio content
	Content Moderation		Detect potentially inappropriate content to avoid issues in global markets, and to increase brand safety for advertisers
	Compliance & Redaction		Detect and pixelate faces captured incidentally to preserve the privacy of non-persons of interest in news feeds and security footage.
Monetization	Contextual Advertising		Make advertising relevant to the context of the content. Avoid brand damage through unintended associations.
	Dynamic Ad Insertion		Optimize Ads based on consumer viewing patterns. Tune the creative in-flight based on real data and consumer feedback.
	Highlight Generation		Identify actors/ players, detect specific activities (goal, foul, car crash...) and automatically create a clipping around the detected scenes.
	Content Recommendation		Analysis disparate data and enable the consumer to make personalized content choices, predict consumer behavior

AWS Media Services and products



Amazon Interactive
Video Service



AWS Elemental
MediaConnect



AWS Elemental
MediaConvert



AWS Elemental
MediaLive



AWS Elemental
MediaPackage



AWS Elemental
MediaStore



AWS Elemental
MediaTailor



AWS Thinkbox
Deadline



Render Farm
Deployment Kit



Amazon
Nimble Studio



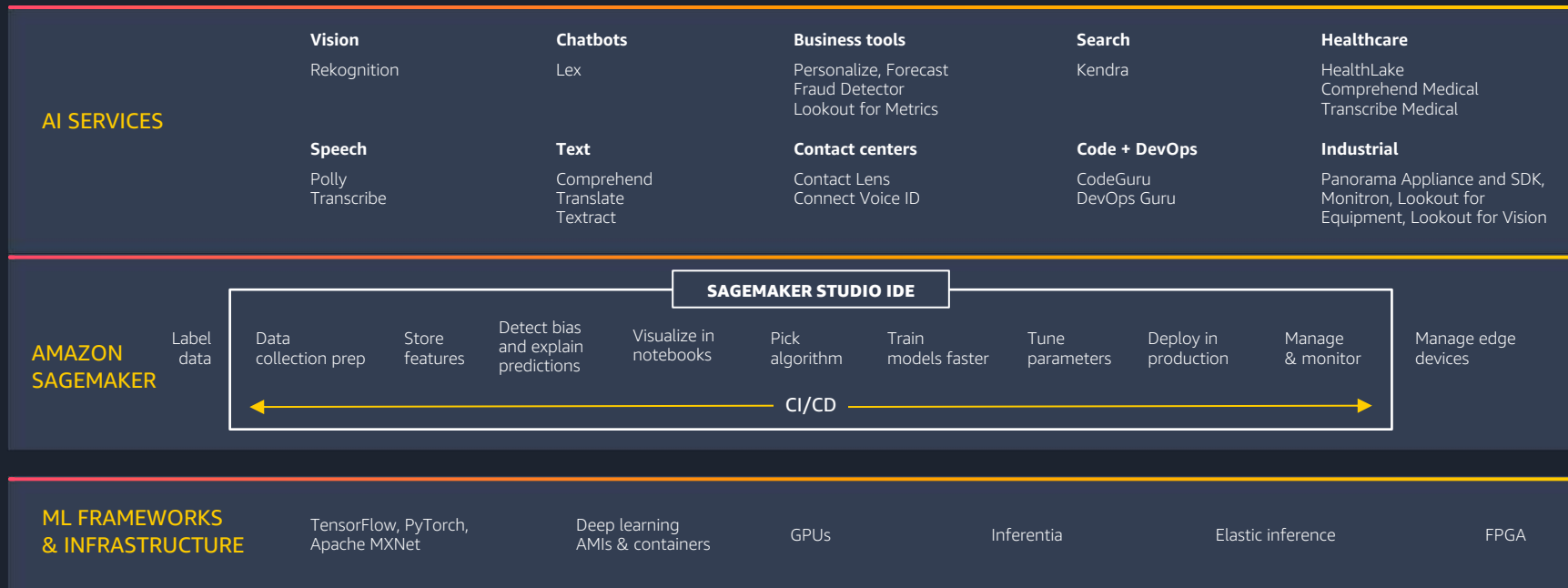
AWS Elemental Software
& Appliances



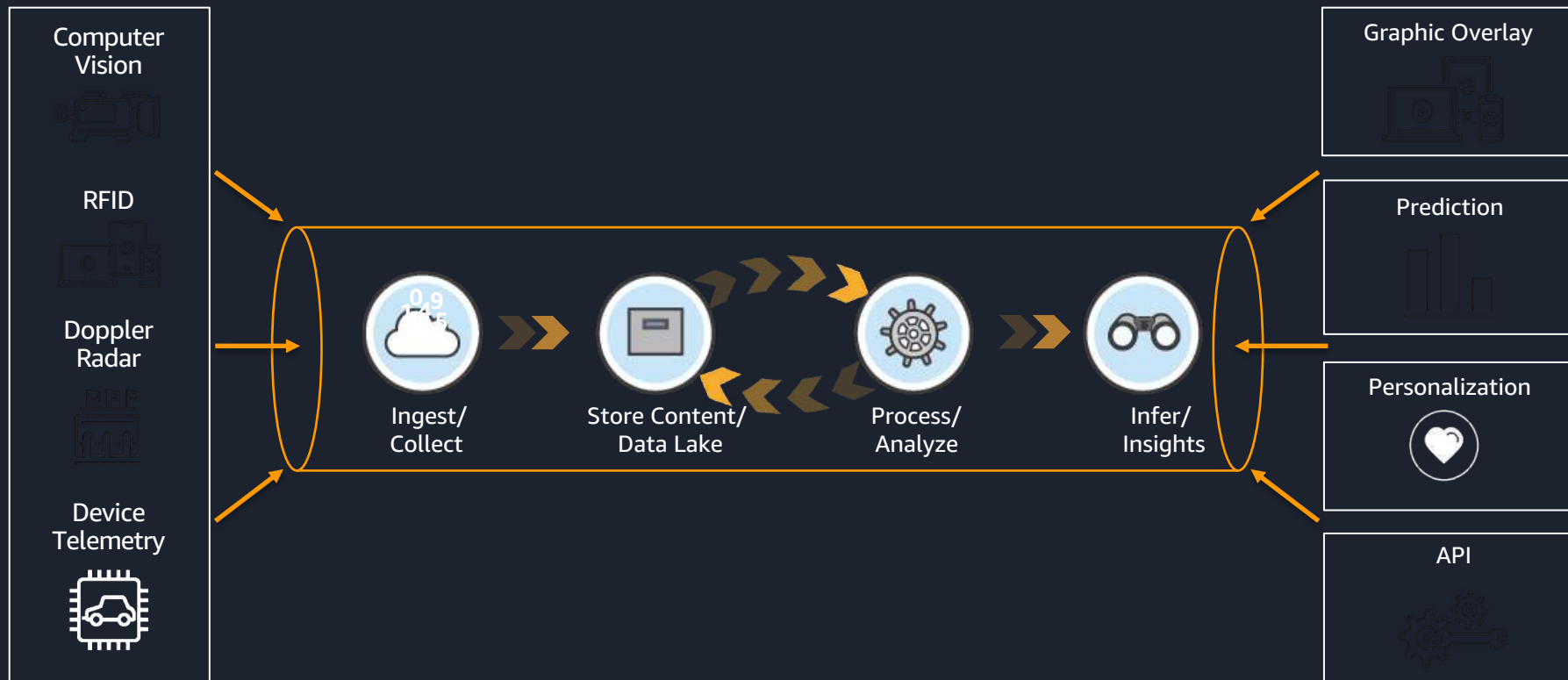
AWS Elemental
Link

The **AWS ML stack**

Broadest and most complete set of machine learning capabilities



Modern data architecture for media enrichment



PROBLEM

Wanted to build a flexible and scalable content recommendation product with a fast go to market

SOLUTION

Used Amazon Personalize to create personalized content experiences for customers on the discovery+ streaming platform

IMPACT

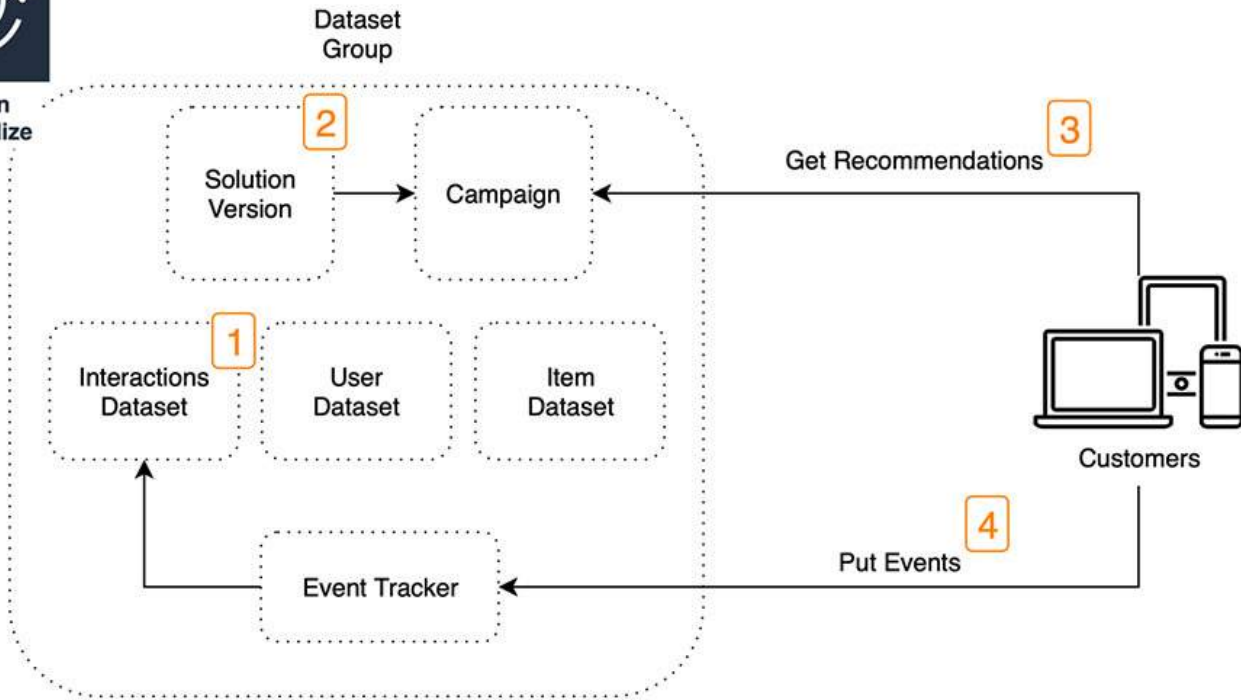
Improved customer journey in discovery+ platform
Met Discovery's desire to deliver content against time-to-market goals
Solution works with existing infrastructure

Discovery Communications, Inc.





Amazon Personalize



Use case: From manual tagging to AI-based tagging

Background

- Over 200,000 hours of content
- Only half the content is manually tagged
- How can they enrich the metadata on AWS?
- How can they unleash the value of content they already own once on AWS?

Challenge

- Large-scale video library
- High accuracy required
- Limited budget
- Ability to extract from video
- Keep up with daily increase in content

Results

- Solution developed within three weeks
- Live video frame-based analysis
- Established, searchable baseline archive
- All content is now tagged and indexed
- Over 99,000 faces indexed and searchable
- Saved ~9,000 hours a year in manual curation costs



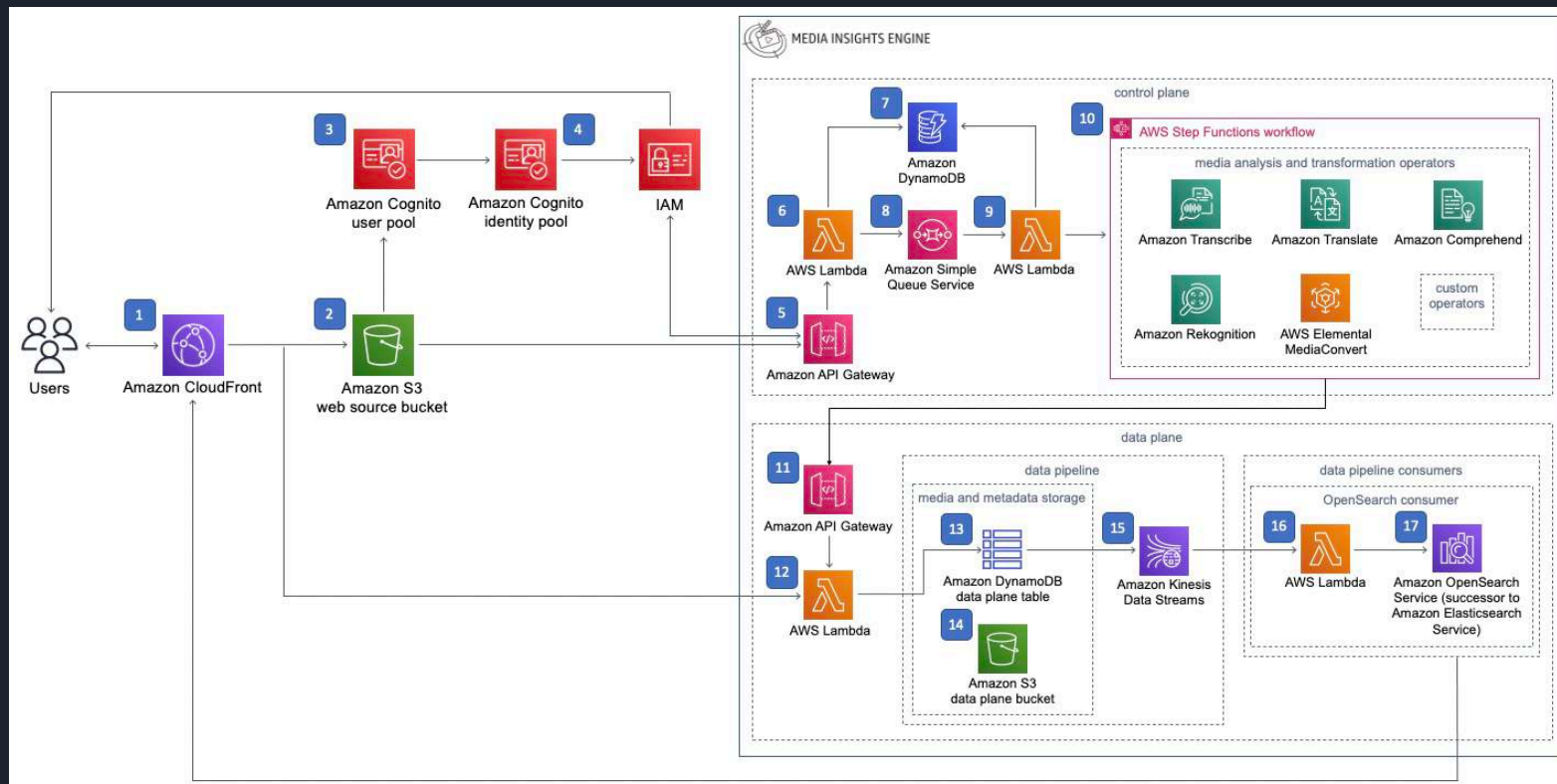
Enhanced viewer experience with metadata



- Actors
- Music
- Trivia
- Skip credits/skip intro
- Objects/locations/logos
- Content aware ad breaks



Content Analysis



Tag, filter, and redact inappropriate content



Person 99.2%

Gun 84.6%

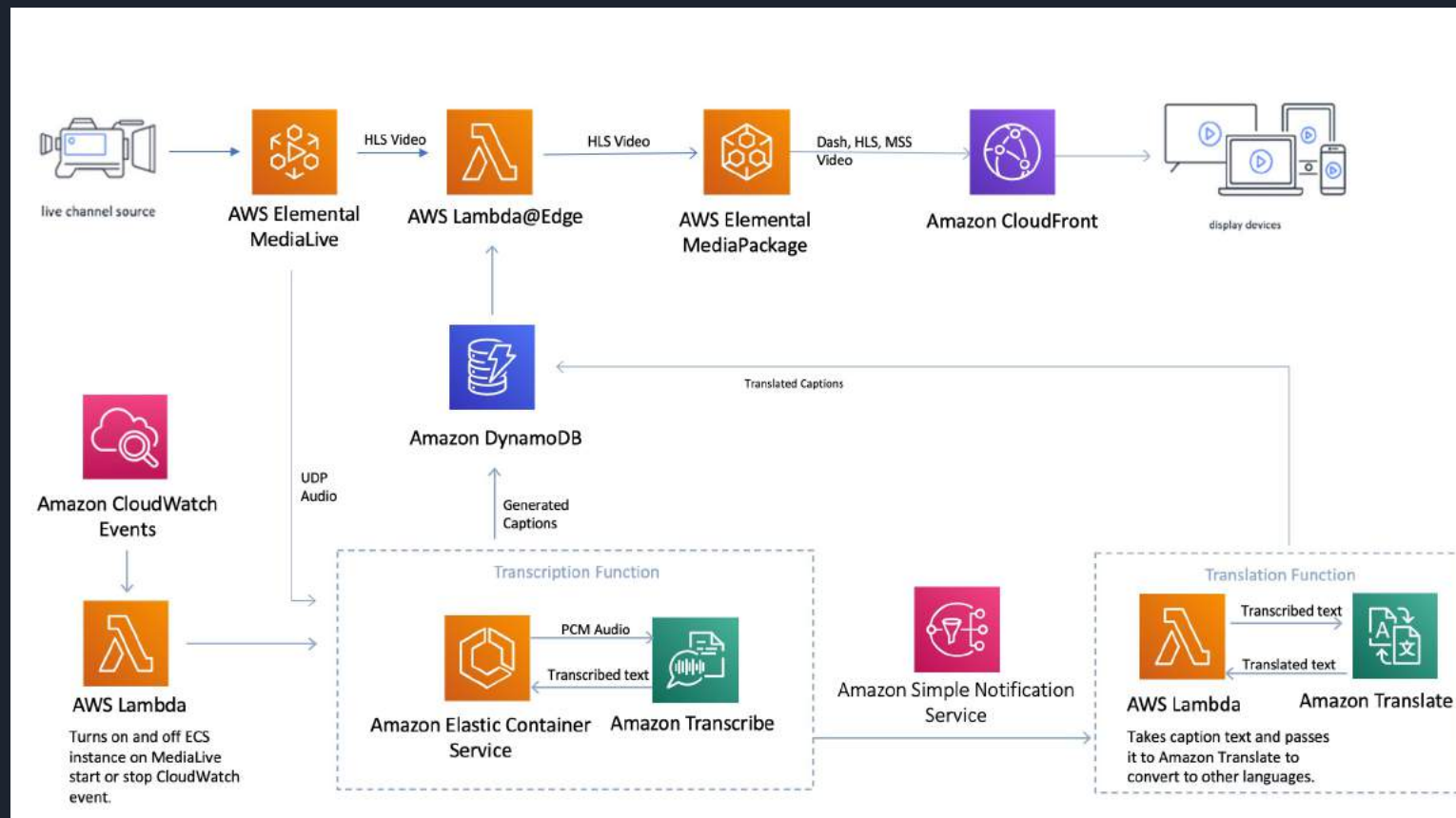
Drink 96.4% Alcohol 80.1% Wine 69.9%



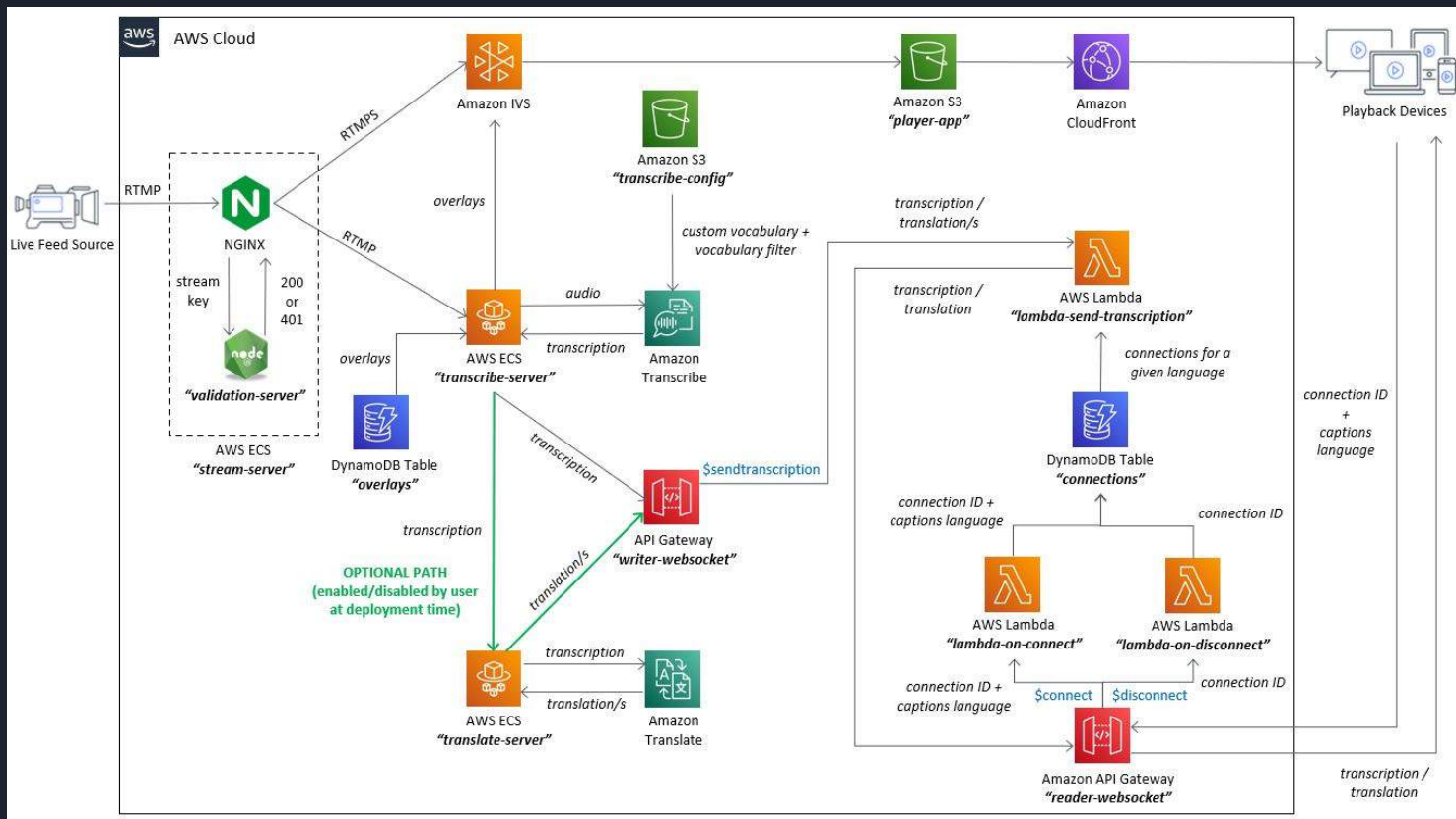
Handgun 73.5%



Live Streaming with Automated Multi-Language Subtitling



Amazon IVS Auto-captions



E-commerce interactive video experience



```
productId: 1000567894 (75.83s)
productId: 1000567894 (74.72s)
productId: 1000567893 (73.64s)
productId: 1000567893 (72.77s)
productId: 1000567893 (71.72s)
productId: 1000567893 (70.61s)
productId: 1000567893 (69.78s)
productId: 1000567892 (68.69s)
productId: 1000567892 (67.86s)
productId: 1000567892 (66.82s)
productId: 1000567892 (65.70s)
```



Eyeshadow Palette - Set of 3 Palettes

\$125 \$99



Waterproof Eyeliner and Mascara

\$27



Divine Shine - Satin Rose Lipstick set

\$35 \$28



Gloss Bomb Universal Lip Luminizer

\$19 \$15



7-in-1 daily wear palette essentials

\$103 \$95



EYE CARE SET - sparkle gloss

Event broadcasting enrichment

The screenshot displays the Sky News website interface for the event 'THE ROYAL WEDDING'. The page is divided into several sections:

- Header:** Sky News logo, search bar, and navigation menu (Home, UK, World, Politics, US, Ocean/Rescue, Science & Tech, Business, Entertainment, Offbeat, Weather).
- Main Content:** A large video player showing a scene from the wedding with David Beckham and Victoria Beckham. Below the video is a 'PROFILES' section for David Beckham, including a small profile picture and a brief biography: "David Robert Joseph Beckham, OBE is an English former professional footballer. He played for Manchester United, Preston North End, Real Madrid, Milan, LA Galaxy. Not playing..."
- ARRIVALS:** A list of guests with their names and roles, each accompanied by a small profile picture:
 - Elton John (Singer)
 - David Furnish (Elton John's husband)
 - David Beckham (Former footballer)
 - Victoria Beckham (Businesswoman)
 - Princess Beatrice (Real Family)
 - Princess Eugenie (Real Family)
 - Prince Seunso Bering (Real Family of Lorothe)
 - Princess Madeireng Seino (Real Family of Lorothe)
- Right Sidebar:** A video player showing a scene with David Furnish and Elton John. Below it is a list of guests with their names and roles, each accompanied by a small profile picture:
 - Elton John (Singer)
 - David Furnish (Elton John's husband)
 - David Beckham (Former footballer)
 - Victoria Beckham (Businesswoman)

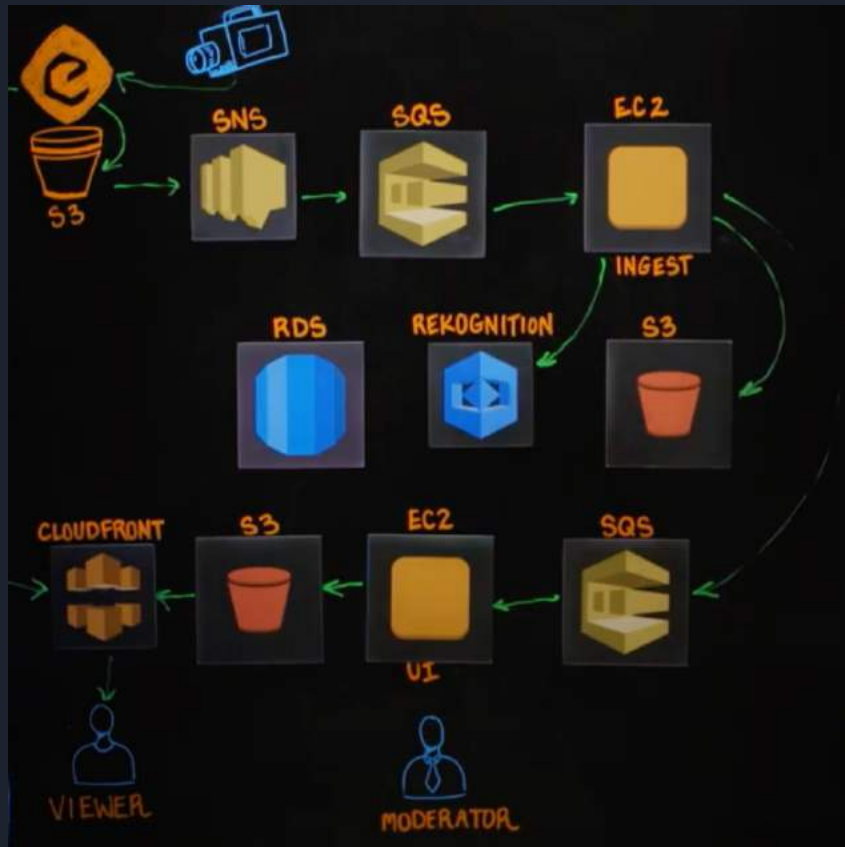
CONTENT
INNOVATION
AWARDS 2018

sky news

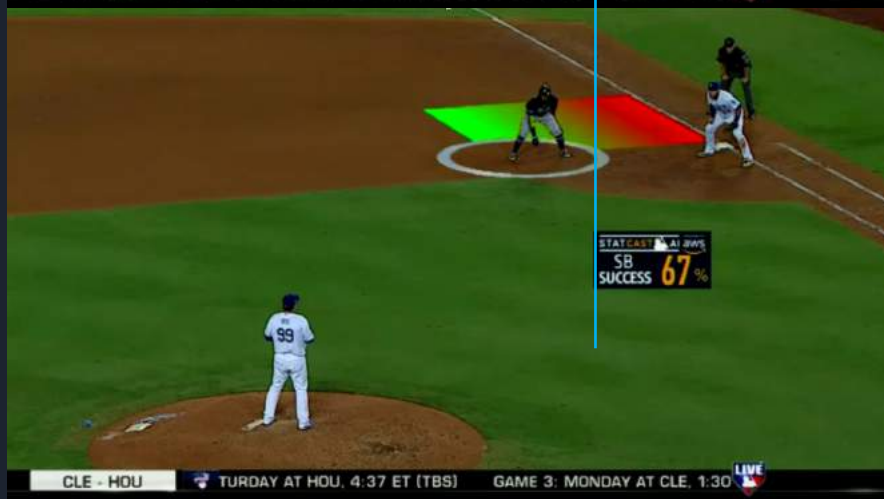
aws

UI CENTRIC

GRAYMETA



<https://youtu.be/7XmyGDMdZV4>





Prior to each pitch, MLB sends a request for inferences (predictions) to AWS

~128 milliseconds for 21 inferences



Response contains probabilities (lead distances 0-20 ft) to render graphics in real-time

PROBLEM

3TB+ data, 1,500+ hours play time per week

Needed ML solution for real-time stats

Lean team, no data science expertise

SOLUTION

Next Gen Stats (NGS)

Live data streamed to AWS from
RFID tags on players and in game ball

Data processed in 100+ steps in under 1 second

Real-time predictions

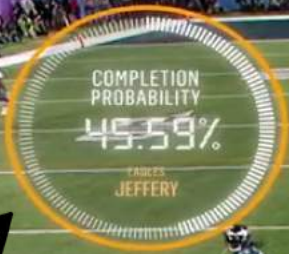
Output stats published via
APIs and on-screen graphics

IMPACT

Quickly launched 20+ stats

Sports announcers get interesting
data points to engage fans





Distance thrown

Double coverage

Separation

Formation

Time to throw

Distance to rusher



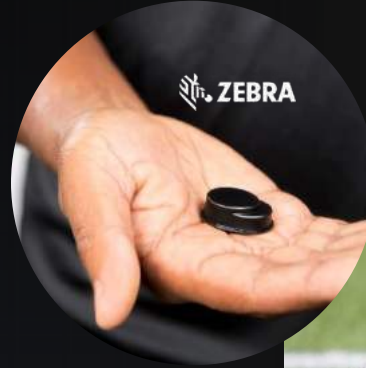
Amazon SageMaker

- Model trained with ~35k plays
- Every 10th of a second, stadium sensors gets signal from player on the field
- Predictions made in less than one second
- Pushed to broadcasters via API for in-game integration



- Real time data capture and processing
- Stadium sensors track RFID tags on players
- Location data within inches
- Speed and location of every player
- 3 TB of data captured every game week

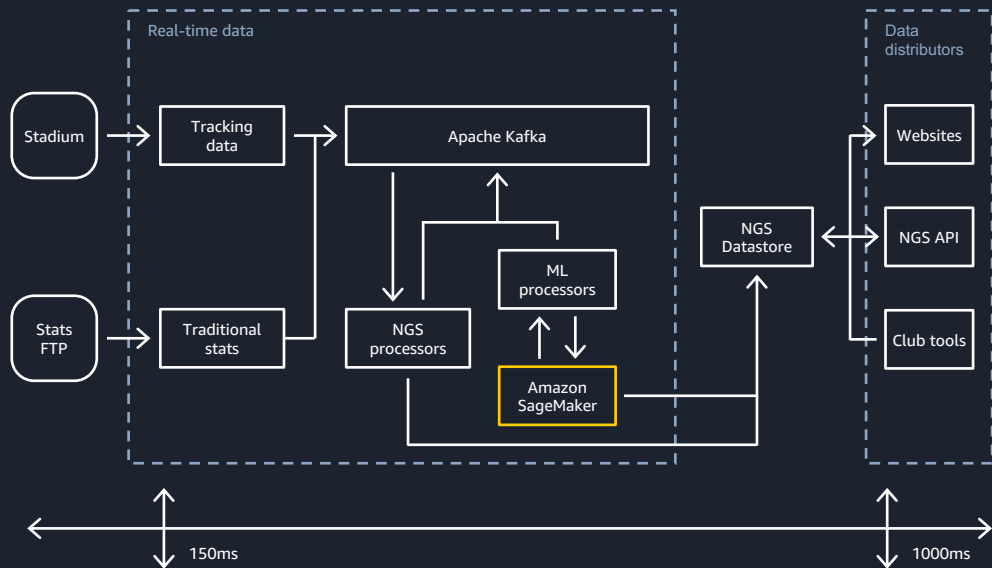
RFID Chip





SOLUTION ARCHITECTURE

Next Gen Stats



Entire solution runs on AWS

Amazon SageMaker lets ML models deployed into existing analytics pipelines

Streaming data architecture

PROBLEM

Bundesliga wanted to enhance the fan experience by providing more contextual data around the probability of scoring a goal for each shot in a match

SOLUTION

Amazon SageMaker to train a model to make predictions on event data, trained on over 40k historical data points

SageMaker Clarify implemented to help explain model logic behind individual xGoals predictions

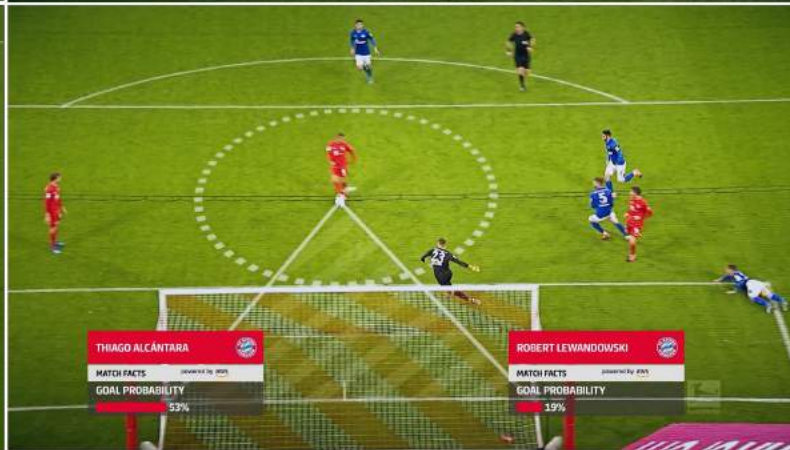
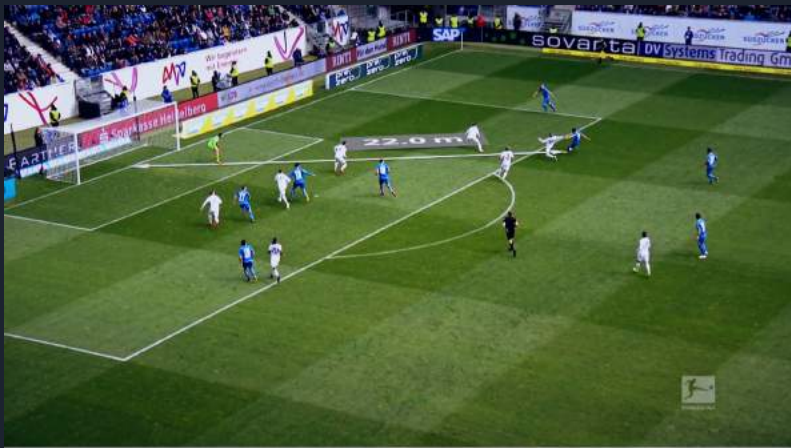
Full serverless architecture with AWS Fargate and AWS Lambda to ingest event data

IMPACT

Using SageMaker, Bundesliga is able to improve the fan experience by providing more data around each shot attempt in a match



BUNDESLIGA

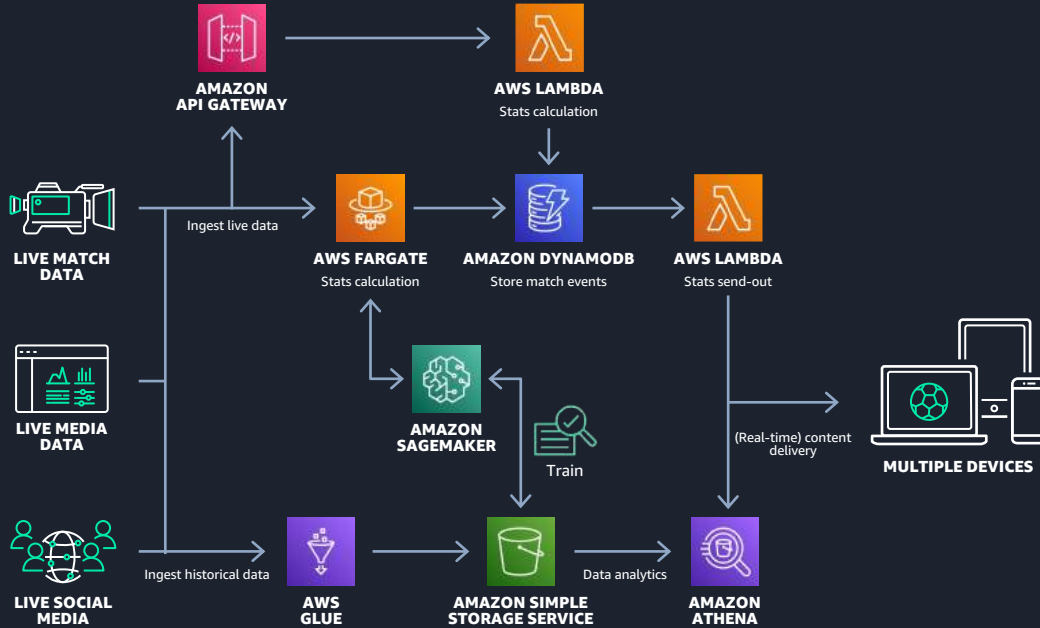


THIAGO ALCÁNTARA	
MATCH FACTS	powered by AWS
GOAL PROBABILITY	53%

ROBERT LEWANDOWSKI	
MATCH FACTS	powered by AWS
GOAL PROBABILITY	19%



Bundesliga solution architecture



- Event data and player metadata are ingested
- An AWS Lambda function is triggered to pre-process data and invoke a SageMaker endpoint to retrieve the xGoals value
- DFL broadcasts the value, and stores it in S3 to continually improve the model

AI and Machine Learning

CHALLENGE

With 120 car telemetry sensors create 3Gb of data per second, how can Formula 1 bring analytics and insights to car data that drives advances in the sport and a deeper fan experience delivered by their next generation OTT/VOD platform.

SOLUTION

- AWS Kinesis streams data into S3 in real-time
- Amazon Rekognition, Amazon SageMaker, and Amazon Transcribe analyze race data
- AWS Elemental Media Services powers next generation video platform

BENEFIT

- Deeper fan experience
- Monetize data and direct to consumer video platform

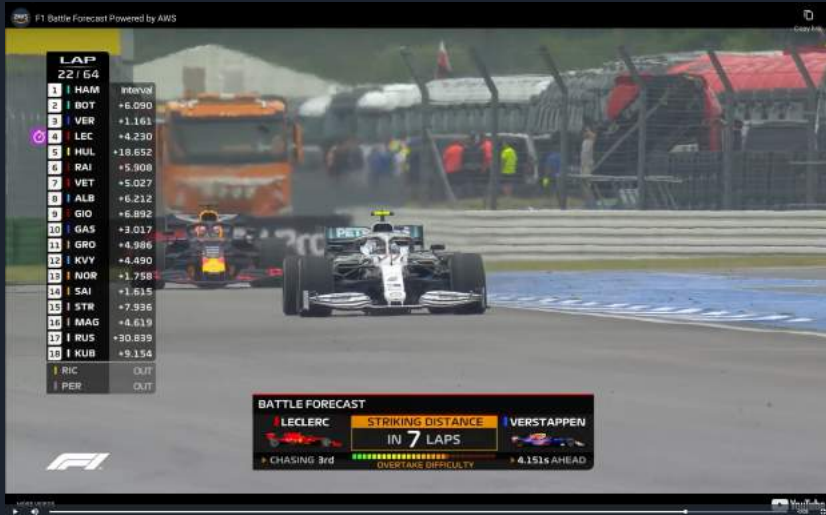


About Formula One

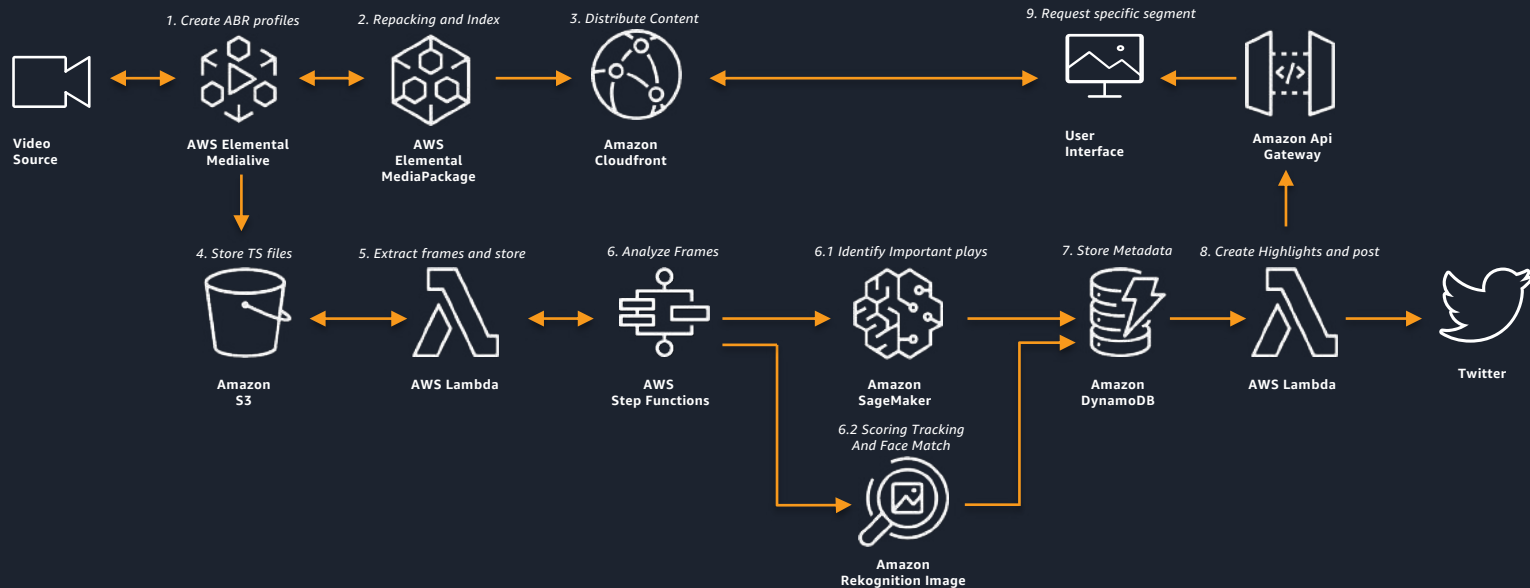
The Formula One Group is responsible for the promotion of the FIA Formula One World Championship, a series of auto racing events in 21 countries where professional drivers race single-seat cars on custom tracks or through city courses. Formula One has more than 500 million global fans watching 21 global races annually.

“ For our needs, **AWS outperforms all other cloud providers**, in speed, scalability, reliability, global reach, partner community, and breadth and depth of cloud services available. ”

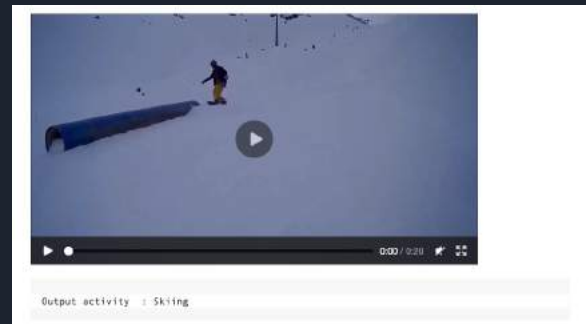
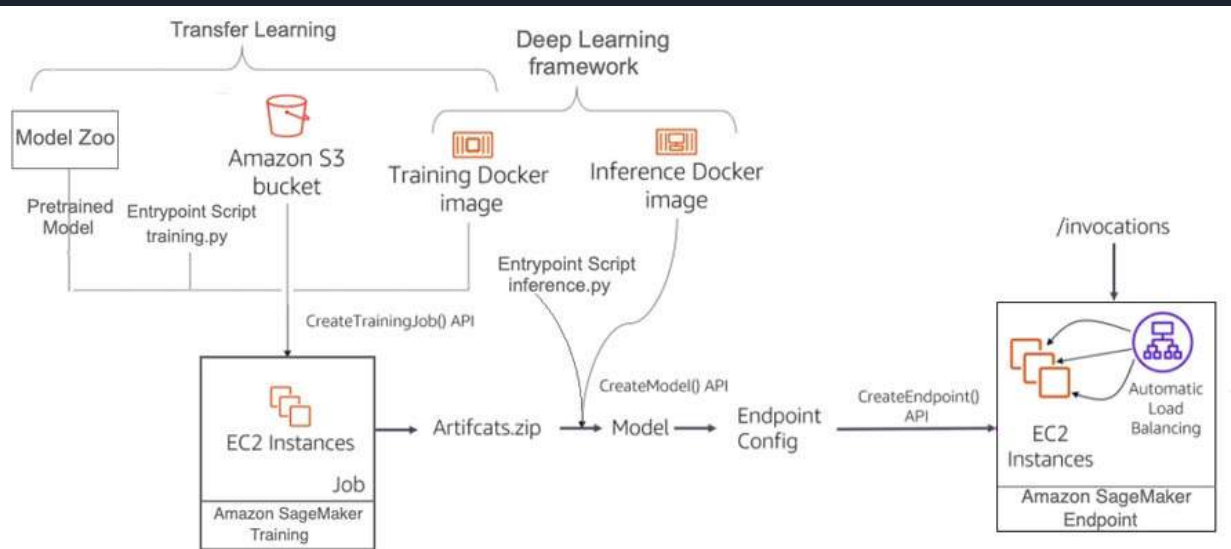
Pete Samara, Director of Innovation and Digital Technology



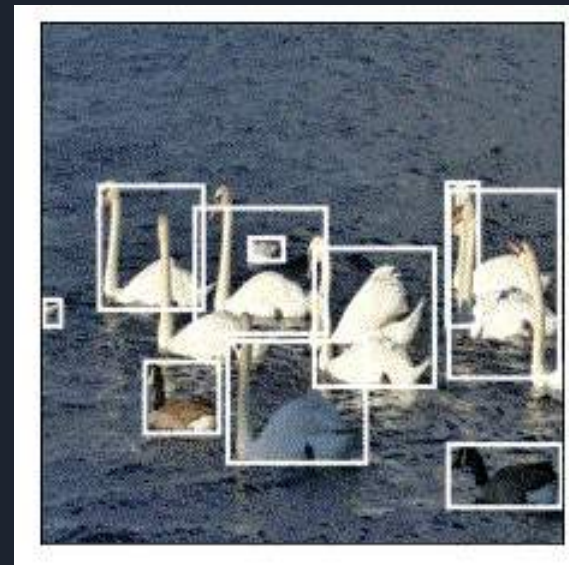
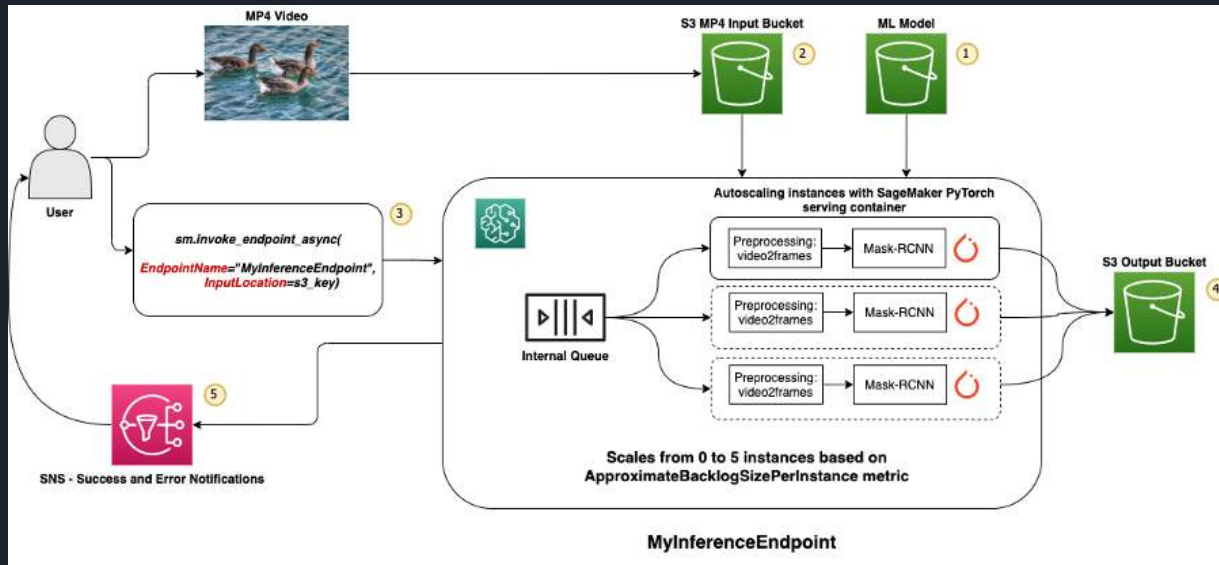
AUTOMATICALLY CREATE HIGHLIGHTS



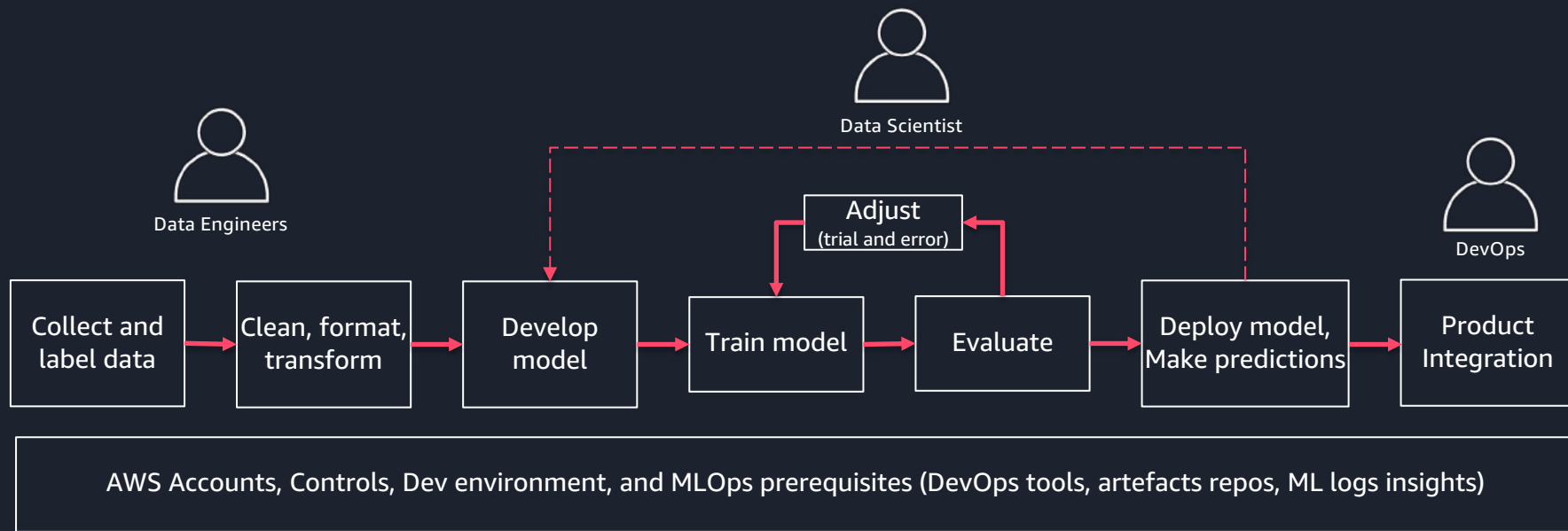
Activity detection on a live video stream



Run computer vision inference on large videos with Amazon SageMaker asynchronous endpoints



ML lifecycle | From idea to model



Cloud Engineer

Integrated Workbench

Capabilities designed specifically for ML, data preparation, experiment management, and workflows

Managed Infrastructure

Designed for ultra low latency and high throughput, automatic scaling, and distributed training

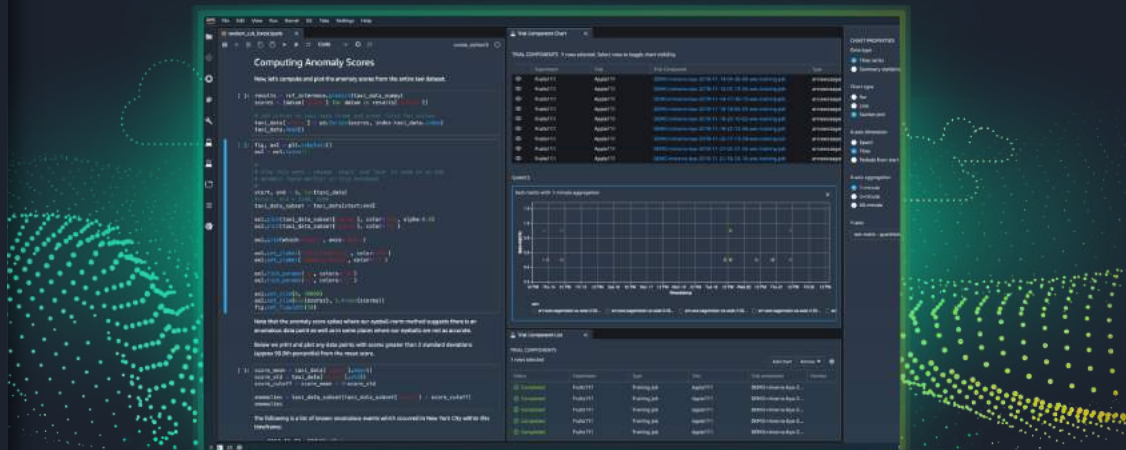
Managed Tooling

Purpose-built from the ground up to work together including auto ML, collaboration, debugger, profiler, bias analyzer, and explainability

<https://aws.amazon.com/sagemaker>

Amazon SageMaker

Most complete, end-to-end ML service



Amazon SageMaker overview

Amazon SageMaker

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 predictions

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 JumpStart **NEW**

Pre-built solutions for common use cases

TRAIN & TUNE

Managed 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 cost by 90%

DEPLOY & MANAGE

Managed Deployment

Fully managed, ultra low latency, high throughput

Kubernetes & Kubeflow Integration

Simplify Kubernetes-based machine learning

Multi-Model Endpoints

Reduce cost 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

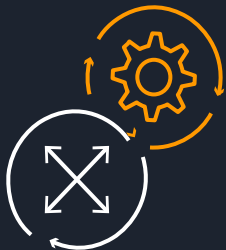
SageMaker Studio

Integrated development environment (IDE) for ML

Accurate ML models require high-quality data



Data labeling challenges



Massive scale

ML models need large, labeled datasets



High accuracy

ML models depend on accurately labeled data



Time consuming

Building training datasets takes up to 80% of a data scientist's time

Amazon SageMaker Ground Truth

Easily label training data for machine learning at scale



Fully managed
and secure



30+ labeling
workflows



Assistive
tooling,
auto-labeling,
consensus

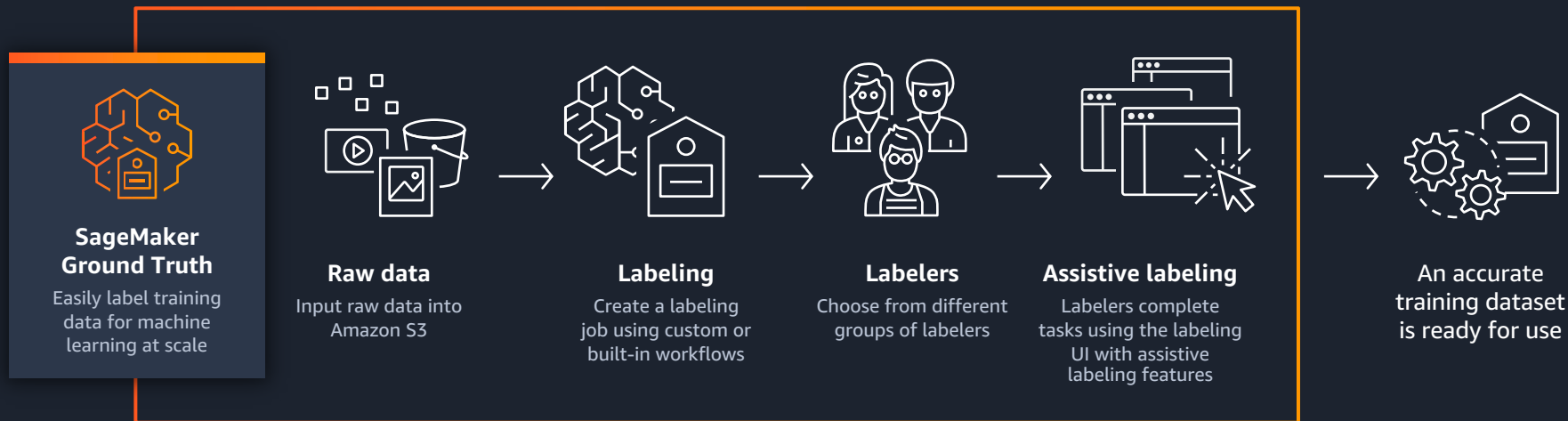


Your choice
of workforce



Pay as you go,
volume discounts

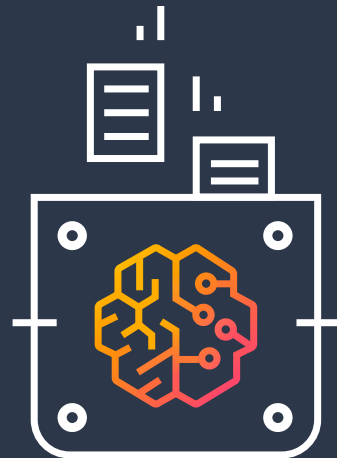
How Amazon SageMaker Ground Truth works



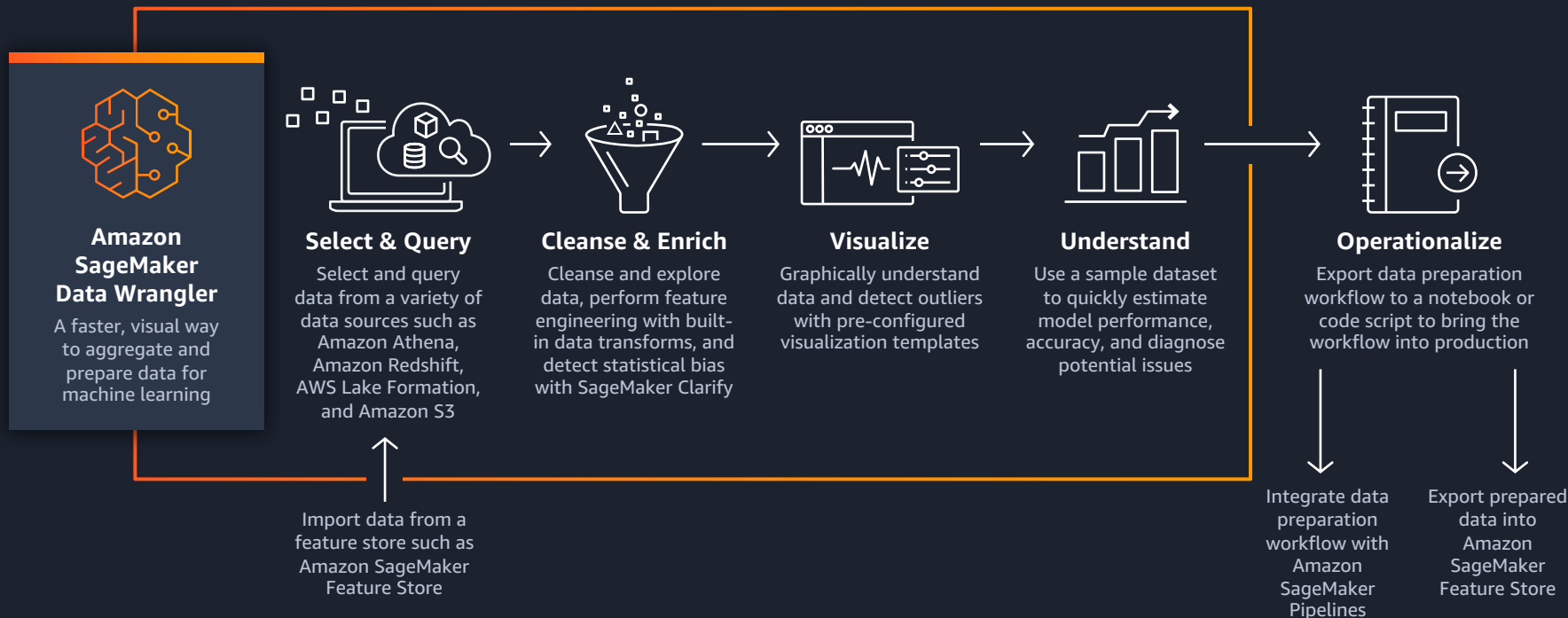
NEW

Amazon SageMaker Data Wrangler

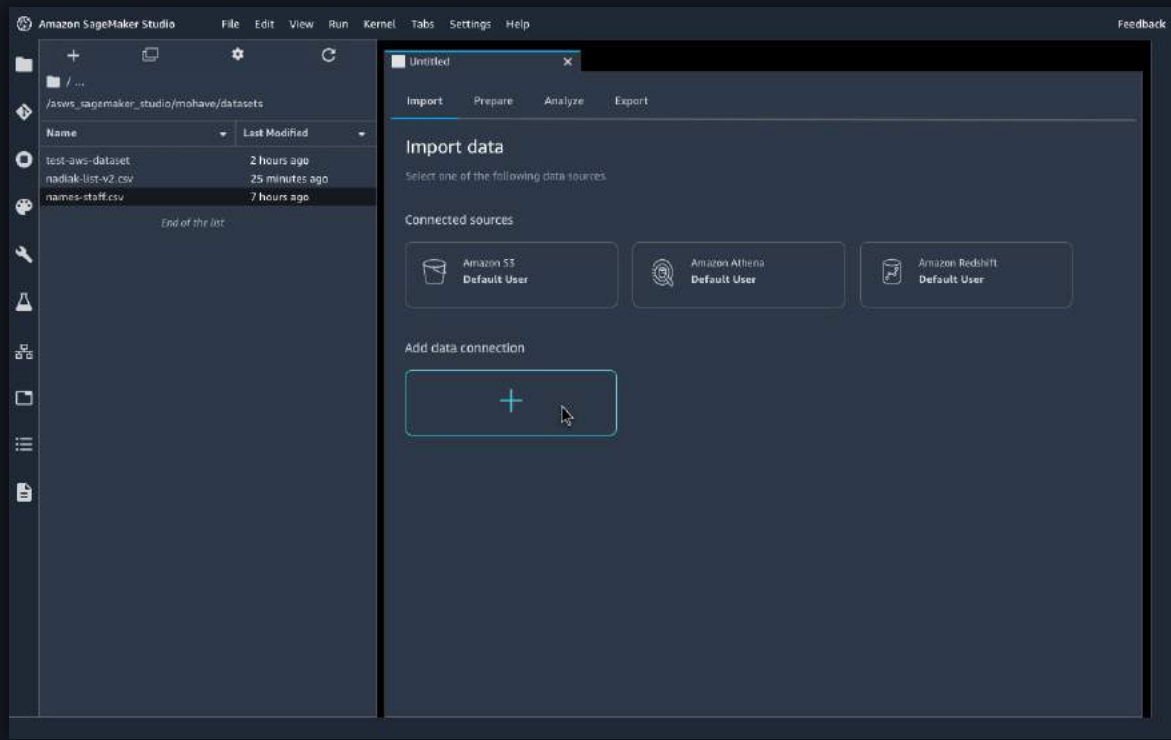
The fastest and easiest
way to prepare data for
machine learning



How SageMaker Data Wrangler Works



Quickly select and query data



Select data from Amazon Athena, Amazon Redshift, AWS Lake Formation, Amazon S3, and features from SageMaker Feature Store

Write queries for data sources before importing data over to SageMaker Data Wrangler

Import data in various file formats, such as CSV files, parquet files, and database tables directly into Amazon SageMaker

Easily transform data

Transform your data without writing a single line of code using pre-configured data transforms

Preconfigured data transforms include convert column type, rename column, and delete column

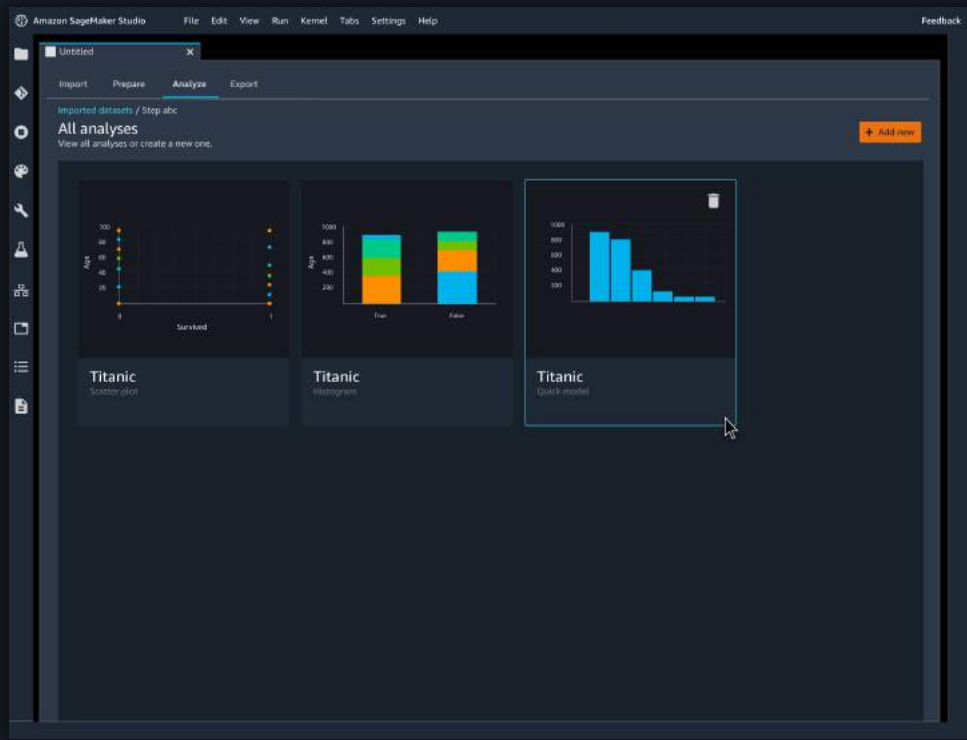
Author custom transforms in PySpark, SQL, and Pandas

Detect bias and identify dataset imbalance with SageMaker Clarify

The screenshot displays the Amazon SageMaker Studio interface. At the top, there's a menu bar with options like File, Edit, View, Run, Kernel, Tabs, Settings, and Help. Below the menu, there's a toolbar with icons for Import, Prepare, Analyze, and Export. The main area shows a data table with columns: String App, String Category, Decimal Rating, Integer Reviews, String Size, and String Installs. The table contains various app entries like 'Photo Editor Coloring bookCamera &', 'Coloring book means', 'U Launcher Lite', etc. On the right side, there's a sidebar with 'TRANSFORMS' and a list of suggested transforms such as 'Find-Replace', 'Rename column', 'Replace rare', 'Impute missing categorical', 'Impute missing numeric', 'Tokenizer', 'TF-IDF text embedding', 'Ordinal encode', 'Onehot encode', 'Drop column', 'Duplicate column', 'Flatten vector', 'Split by delimiter', 'Type conversion', and 'Custom transform'. There's also a 'PySpark' dropdown and a 'Preview' button.

String App	String Category	Decimal Rating	Integer Reviews	String Size	String Installs
Photo Editor Coloring bookCamera &	ART_AND_DESIGN	4.1	159	19M	10,000+
Coloring book means	ART_AND_DESIGN	3.9	967	14M	500,000+
U Launcher Lite	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+
Themes, Hide Apps	ART_AND_DESIGN	4.5	215644	25M	50,000,000+
Sketch - Draw & Paint	ART_AND_DESIGN	4.5	567	2.8M	100,000+
Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.4	167	5.6M	50,000+
Paper Flowers Instructions	ART_AND_DESIGN	3.9	178	19M	50,000+
Smoke Effect Photo Maker - Silhouette	ART_AND_DESIGN	4.1	36815	29M	1,000,000+
Editor	ART_AND_DESIGN	4.4	11791	55M	3,000,000+
Infinite Painter	ART_AND_DESIGN	4.7	121	3.1M	10,000+
Garden Coloring Book	ART_AND_DESIGN	4.4	19880	28M	1,000,000+
Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.4	8788	12M	1,000,000+
Text on Photo - Fontface	ART_AND_DESIGN	4.2	44829	20M	10,000,000+
Name Art Photo Editor - Focus n Filter	ART_AND_DESIGN	4.6	4326	21M	100,000+
Tattoo Name On My Photo Editor	ART_AND_DESIGN	4.4	1518	57M	100,000+
Mandala Coloring Book	ART_AND_DESIGN	3.2	55	2.7M	5,000+
3D Color Pixel by Number - Sandbox	ART_AND_DESIGN	4.7	3652	5.5M	500,000+
Art Coloring	ART_AND_DESIGN	4.5	27	17M	10,000+
Learn To Draw Kawaii Characters	ART_AND_DESIGN	4.3	194216	59M	5,000,000+
Photo Designer - Write your name with shapes	ART_AND_DESIGN	4.6	224399	31M	10,000,000+
FlipsClip - Cartoon animation	ART_AND_DESIGN	4.2	450	14M	100,000+
350 Dly Room Decor Ideas	ART_AND_DESIGN	4.6	4.6	10M	100,000+
lits Paint X			4		

Understand your data visually

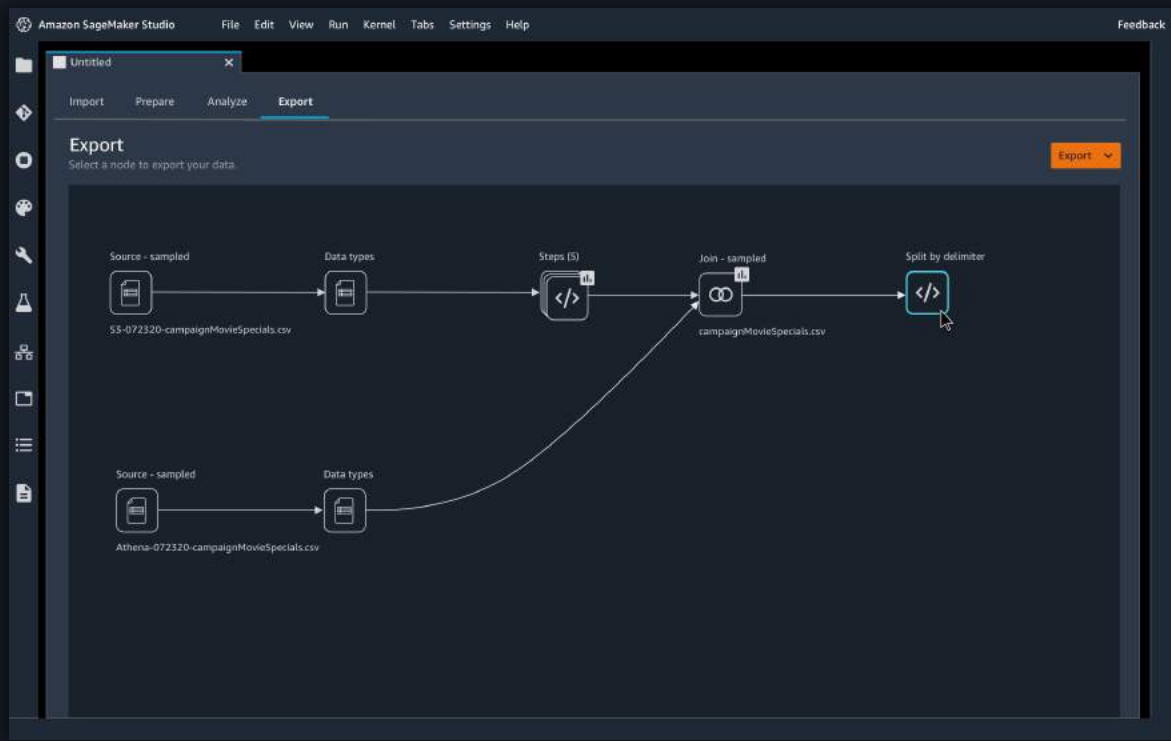


Intuitively understand your data with a set of pre-configured visualization templates

Preconfigured visualization templates include histograms, scatter plots, box and whisker plots, line plots, and bar charts

Interactively create and edit your own visualizations so you can quickly detect outliers or extreme values

Deploy data preparation workflows into production

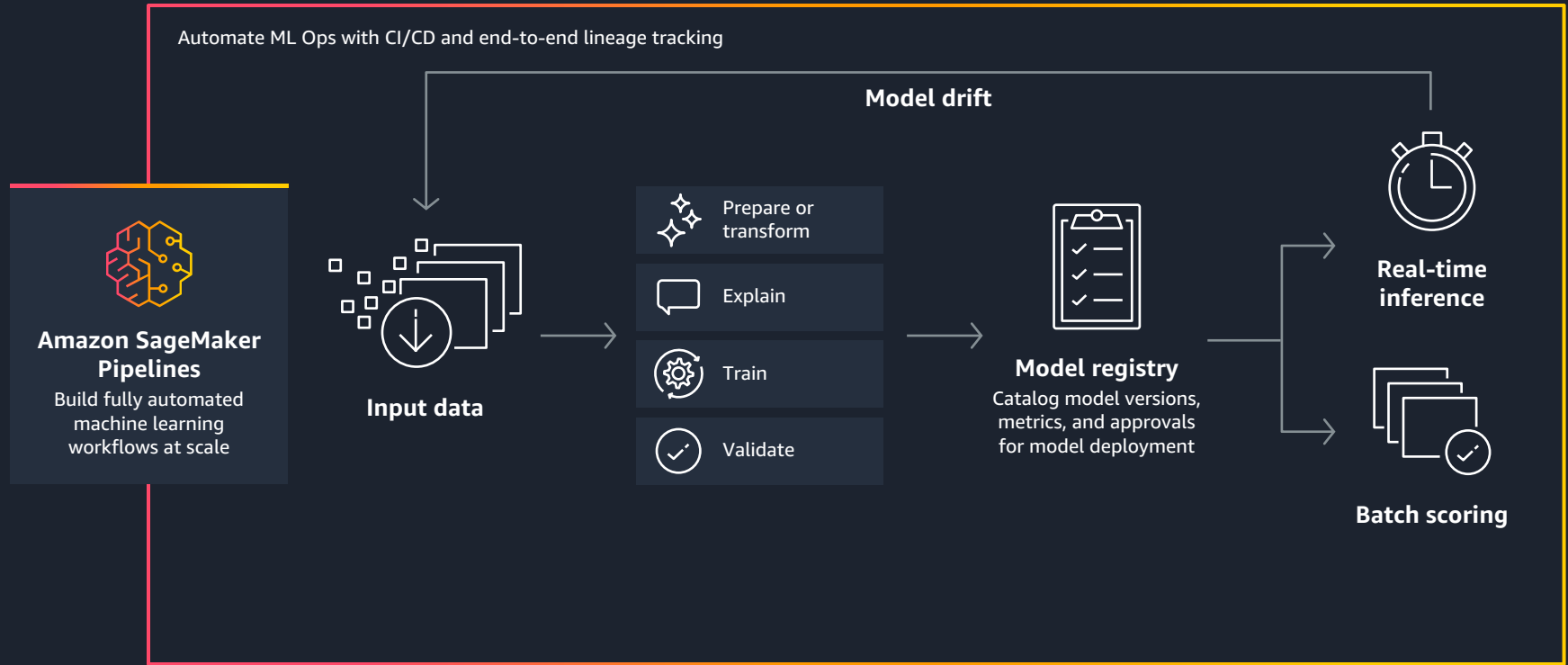


Export data preparation workflows to a notebook or Python code

Integrate your workflow with SageMaker Pipelines to automate model deployment and management

Publish created features to SageMaker Feature Store for reuse and syndication across teams and projects

Amazon SageMaker Pipelines overview



Compose and manage workflows

Create your ML workflows using the Python SDK. Define the parameters and steps

Visualize the workflows with SageMaker Studio

Include steps such as data transformations, training, debugging, and optimizing models

```
# Example processing step for feature engineering
```

```
step_process = ProcessingStep(  
    name="PreprocessAbaloneData",  
    processor=sklearn_processor,  
    outputs=[  
        ProcessingOutput(output_name="train", source="/opt/ml/processing/train"),  
        ProcessingOutput(output_name="validation", source="/opt/ml/processing/validation"),  
        ProcessingOutput(output_name="test", source="/opt/ml/processing/test"),  
    ],  
    code=os.path.join(BASE_DIR, "preprocess.py"),  
    job_arguments=["--input-data", input_data],  
)
```

```
# Example training step
```

```
step_train = TrainingStep(  
    name="TrainAbaloneModel",  
    estimator=xgb_train,  
    inputs={  
        "train": TrainingInput(  
            s3_data=step_process.properties.ProcessingOutputConfig.Outputs[  
                "train"  
            ].S3Output.S3Uri,  
            content_type="text/csv",  
        ),  
        "validation": TrainingInput(  
            s3_data=step_process.properties.ProcessingOutputConfig.Outputs[  
                "validation"  
            ].S3Output.S3Uri,  
            content_type="text/csv",  
        ),  
    },  
)
```


See pipeline execution details and metrics in real-time

Follow completed steps and monitor steps in progress

Understand the output from each step with the output logs

Monitor, change, and manage the parameters for each step

The screenshot displays the AWS SageMaker Pipeline console interface. At the top right, a status bar shows a green circle indicating the pipeline is running, with a timestamp of 2/22/2021, 10:57 AM and an elapsed time of 12m46s. Below this, a flow diagram shows the pipeline steps: PreprocessAbaloneData, TrainAbaloneModel (highlighted with a blue border), EvaluateAbaloneModel, CheckMSEAbaloneEvaluation, and RegisterAbaloneModel. A 'true' label is visible between the CheckMSEAbaloneEvaluation and RegisterAbaloneModel steps. On the right, a detailed view for the 'TrainAbaloneModel' step is open, showing a table of metrics and a file output.

Output	Logs	Information
Metrics		Value
train:rmse		1.68371
validation:rmse		2.22083
Files		S3 bucket URI
model.tar.gz		s3://sagemaker-project-p-7k5s...

Catalog and discover models in Model Registry

Model registry is a central repository of trained models

Access the model registry through SageMaker Studio or the Python SDK

The screenshot displays the Amazon SageMaker Studio interface. On the left, the 'Components and registries' sidebar shows the 'Model registry' selected. The main panel displays the 'Recommendations Model - Latin America' page, which includes a table of model versions. The table has columns for Name, Status, Step, Description, Status updated by, Modified on, and Actions. A context menu is open over the 'version 5' row, showing options like 'Open model version' and 'Update model version status...'. The status bar at the bottom indicates 'Python 3 (Data Science) | Idle'.

Name	Status	Step	Description	Status updated by	Modified on	Actions
version 6	Pending	Staging	New model with SKE...	Pending	10/10/20	⋮
version 5	Approved	Production	Model updated on 8/1...	Jen Cabro		Open model version Update model version status...
version 4	Approved	Archived	Model updated on 7/15...	Jen Cabro		
version 3	Approved	Archived	Model updated on 6/15...	Jen Cabro		
version 2	Approved	Archived	Model built on 5/15/20...	Jen Cabro	10/10/20	
version 1	Approved	Archived	Model built on 4/15/20...	Jen Cabro	10/10/20	

View and compare evaluation metrics from training step

The screenshot displays the Amazon SageMaker Studio interface. On the left, the 'Components and registries' sidebar shows a 'Model registry' with a search bar and a 'Create model group' button. Below this, a list of models is shown with columns for 'Name', 'Modified on', and 'Actions'. The main area is titled 'Comparing model versions' and shows a comparison between 'version 5' and 'version 4' for a model group named 'modelGroup-5'. The comparison is visualized through three rows of plots: 'Confusion matrix', 'Receiver operating characteristic curve', and 'PRC'. Below these plots is a table of metrics comparing the two versions.

Metric	Value	SD	Value	SD
Recall	0.25	0.25	0.25	0.25
Precision	0.25	0.25	0.25	0.25
Accuracy	0.625	0.625	0.625	0.625
Balanced accuracy	0.0	0.0	0.0	0.0
Precision best constant classifier	0.0	0.0	0.0	0.0
Accuracy best constant classifier	0.0	0.0	0.0	0.0
True positive rate	0.25	0.25	0.25	0.25
True negative rate	0.25	0.25	0.25	0.25
False negative rate	0.25	0.25	0.25	0.25
False positive rate	0.25	0.25	0.25	0.25
auc	1.0	1.0	1.0	1.0

Approve models for production

The screenshot shows the Amazon SageMaker Studio interface for a model group named 'Recommendations Model - Latin America'. The 'Versions' tab is selected, showing a table of model versions. A modal dialog 'Update model version status' is open over version 6, which is currently 'Rejected'. The dialog allows updating the status to 'Approved' and adding a comment: 'The model accuracy of this model looks good. Approved.'

Name	Status	Step	Description	Status updated by	Modified on	Actions
version 6	Rejected	Staging	New model with SML...	Jen Cabro	10/10/20	Open model version (G) Update model version status...
version 5	Approved	Production	Model updated on 8/1...	Jen Cabro		
version 4	Approved	Archived	Model updated on 7/15...	Jen Cabro		
version 3	Approved	Archived	Model updated on 6/15...	Jen Cabro		
version 2	Approved	Archived	Model built on 5/15/20...	Jen Cabro	10/10/20	
version 1	Approved	Archived	Model built on 4/15/20...	Jen Cabro	10/10/20	

Model training is important, but *deployment* drives costs

% Infrastructure Cost



Compute optimized for your Media Workflows



- Compute-intensive applications such as **video encoding, rendering**
- Up to **40% better price-performance** for general purpose, compute-intensive, and memory-intensive workloads

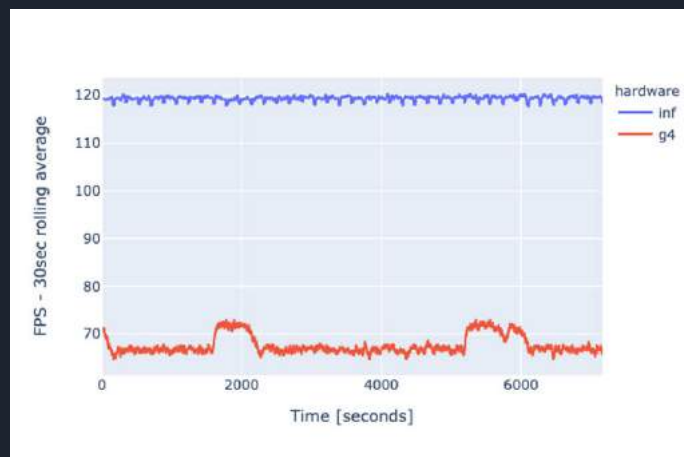
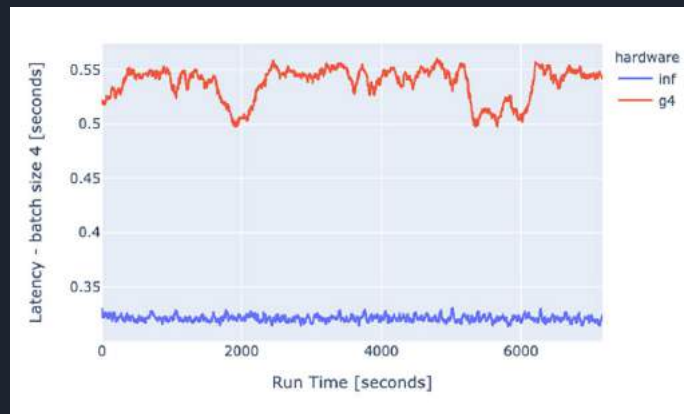


- Latest generation NVIDIA
- DL1 instances with Gaudi accelerators from Habana Labs (an Intel company)
- Inf1 instances with AWS Inferentia chips

Experiment 1 – classification (ResNet50)

Inference accelerator instance type	TPS	Latency (ms, p50)	\$/hour	Cost per 1M inferences (\$, batch size = 1)
CPU C5.xlarge	5.9	215.23	0.23	10.83
CPU-compiled C5.xlarge	12.9	121.61	0.23	4.95
GPU G4dn.xlarge	74.7	21.19	0.82	3.05
GPU-compiled G4dn.xlarge	140.4	11.28	0.82	1.62
AWS Inferentia Inf1.xlarge	304.3	4.9	0.33	0.30

Experiment 2 – detection (YOLOv4)





**Inference
pipelines**



**Multi-model
endpoints**

Getting started on your ML journey



DATA STRATEGY

Lay the foundation for transformation and innovation



START WITH THE BUSINESS CHALLENGE

Find the right use case based on the needs of your business



PARTNER FOR SUCCESS

AWS has multiple programs and training options to help you along the way

ML.aws

Thank you!