



ГЕНАІ ДЛЯ БИЗНЕСА НА AWS



КТО МЫ

qCloudy - команда сертифицированных инженеров с многолетним опытом проектирования, разработки и внедрения.

Мы строим

- гибкие
- масштабируемые
- отказоустойчивые
- cost-effective

решения в публичном облаке с применением AI/ML и поддерживаем их 24/7



НАШИ УСЛУГИ



Консалтинг

- Облачная стратегия развития ИТ-инфраструктуры



Аудит и оценка

- Оценка ИТ инфраструктуры
- Оценка Data инфраструктуры
- Оценка AI/ML решений
- Оценка квалификации сотрудников
- Оценка качества архитектуры



Разработка и поддержка облачных решений

- Миграция в облако
- Оптимизация облачных расходов
- Построение гибридного облака
- Аварийное восстановление облака
- Cloud DevOps
- Разработка Cloud Native приложений
- Озеро данных и хранилища данных
- DataOps
- Генеративный ИИ
- Компьютерное зрение
- MLOps/FMOps
- Комплексное управление облачной инфраструктурой



Обучение

- AWS Workshops



КЕЙС #1 ЧАТБОТ С ИИ НА БАЗЕ AWS ДЛЯ ХОЛДИНГА БАЙТЕРЕК



ВЫЗОВЫ И РЕШЕНИЕ



ВЫЗОВЫ

- Снижение нагрузки на операторов поддержки
- Необходимость круглосуточного обслуживания клиентов
- Снижение операционных затрат

РЕШЕНИЕ

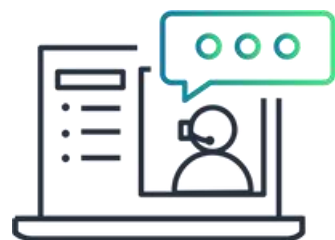
- AI-чатбот на основе Generative AI для автоматической обработки клиентских запросов на казахском и русском языках
- Few-shot prompting + RAG
- Guardrails:
 - content filtering
 - word filtering
- Cross-Region Inference Profile

AWS SERVICES USED

- Amazon Bedrock
- AWS Lambda
- AWS API Gateway
- AWS Amplify
- AWS Cognito
- AWS DynamoDB
- S3
- AWS ElastiCache



КЛЮЧЕВЫЕ РЕЗУЛЬТАТЫ



Среднее время ответа
~10с



Повышена точность и
согласованность ответов



SLA
99,9%



Доступность 24/7
и мгновенные ответы



КЕЙС #2
ИНВЕНТАРИЗАЦИЯ
СЕЛЬХОЗ ТЕХНИКИ

ВЫЗОВЫ И РЕШЕНИЕ



ВЫЗОВЫ

- Неточности при ручной инвентаризации техники
- Операционные расходы из-за частых выездов сотрудников в регионы
- Отсутствие объективных критериев оценки состояния техники

РЕШЕНИЕ

- Автоматизированная AI-система анализа фото AWS Bedrock
- Event Driven архитектура
- Serverless решение

AWS SERVICES USED

- Amazon Bedrock
- AWS Lambda
- AWS API Gateway
- AWS Cognito
- AWS DynamoDB
- S3



КЛЮЧЕВЫЕ РЕЗУЛЬТАТЫ



**с нескольких недель
до 2–3 дней**
сократилось время на
инвентаризации



Снизилась
коррупционные риски



с 3 000 до 18 000 за сезон
выросло количество
обработанных предметов



в 3 раза
сократились расходы на
инфраструктуру
в период неактивности

СПАСИБО





Александр Бернадский
Solutions Architect



Amazon in Kazakhstan



AWS Generative AI Stack

APPLICATIONS TO BOOST PRODUCTIVITY



Amazon Q Business
INSIGHTS AND AUTOMATION



Amazon Q Developer
SOFTWARE DEVELOPMENT LIFECYCLE

MODELS AND TOOLS TO BUILD GENERATIVE AI APPS

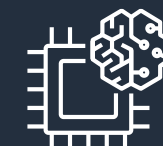


Amazon Bedrock
AMAZON MODELS | PARTNER MODELS

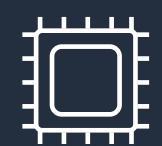
INFRASTRUCTURE TO BUILD AND TRAIN AI MODELS



Amazon SageMaker AI
MANAGED INFRASTRUCTURE



AWS Trainium
AWS Inferentia



GPUs

HIGH PERFORMANCE COMPUTE



Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

Choice of leading FMs through a single API

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and safety



Amazon Bedrock

BROAD CHOICE OF MODELS

AI21labs

Effective reasoning & rapid analysis for long context windows

JAMBA

amazon

Frontier multimodal intelligence at low-latency, Agent & RAG Applications, high-quality image & video generation

AMAZON NOVA

ANTHROPIC

Advanced reasoning & coding capabilities, including computer use skills

CLAUDE

cohere

Multimodal search & advanced retrieval powering multilingual knowledge agents

COMMAND
EMBED
RERANK

Luma

High-quality video generation from text & images

LUMA RAY 2

Meta

Advanced image & language reasoning

LLAMA

MISTRAL AI

Knowledge summarization, expert agents, & code completion

MISTRAL
MIXTRAL

poolside

Software engineering AI for large enterprises

MALIBU
POINT

stability.ai

High-quality AI image generation, easily deployable at scale

STABLE DIFFUSION
STABLE IMAGE



How we use Generative AI in AWS

Good afternoon, Aleksandr!

Ask A to Z

Favourites: Forte Return to office Submit expenses x Book Travel x Amazon Jobs (formerly J... x PeopleSoft (PeoplePorta... x MyBenefits x Employment Documents x

- All
- Org
- Local
- For you
- My Updates
- Saved

All-Amazon Global Meeting
Tuesday, March 18

Submit your questions to the S-Team.



More top stories

Sharing with you my Day 1
115 likes 24 comments

Planning some time off? Here's your out-of-office checklist.
44 likes 8 comments

How we use Generative AI in AWS

C Cedric - Productivity AI

Share conversation ▼



How can I assist you today?

Create Meeting Notes

Turn your rough notes into well structured meeting notes.



Summarize a document

Generate a summary of an Amazon internal document.



Rewrite in Amazon Writing Style

Improve your written narrative to meet Amazon guidelines.



Review an email

Improve your email using best practices.



Cedric's Knowledge

Web Insights

[Learn More](#)

Message Cedric



File Upload



Data Collections



Outputs from this tool cannot be used outside of Amazon. [See disclaimers.](#)

Total conversation tokens: 0 / 180,000



How we use Generative AI in AWS



Grzegorz Ochmanski  · 2-й

AWS for Games Senior Solutions Architect w Ama...

2 дн. · Отредактировано · 

+ Отслеживать ...

Today, I wrote a Streamlit app a chatbot frontend with Amazon Cognito login. It is running on ECS with Graviton EC2, ALB, DynamoDB for per-user chat persistence, EFS for document and image storage, and Amazon Bedrock with Claude Sonnet 3.7. I wrote maybe 50 lines of code; the rest was handled by Amazon Q Developer CLI. The app has full CI/CD and IaC (Terraform).

It supports file handling, an image gallery, full logging in CloudWatch, etc.

Took me 6 hours.

For me, this setup is much better than Cursor since I can use my trusted VS Code IDE.

<https://lnkd.in/dPG-zx5>

You still need to understand what you're building, know the syntax and logic—because sometimes it drifts into insanity, but with a reasonable chat, it gets back on track.

I give it my seal of approval!



TRY IT YOURSELF

abernads@amazon.de - 623387590579 / Admin (Not Production Account)



Search

[Option+S]



United States (Oregon) ▼

Admin/abernads-lsengard @ 6233-8759-0579 ▼



Console home [Info](#)

[Reset to default layout](#)

[+ Add widgets](#)

Recently visited [Info](#)

< 1 2 >

- Amazon Bedrock
- AWS Health Dashboard
- Athena
- AWS Glue
- S3
- Lambda
- CloudFormation

- Amazon Rekognition
- CloudWatch
- Amazon SageMaker AI
- Service Quotas
- EC2
- Infrastructure Composer
- Amazon EventBridge

[View all services](#)

Applications (0) [Info](#)

[Create application](#)

Region: US West (Oregon)

us-west-2 (Current Region) ▼

< 1 >

Name ▼ | Description ▼ | Region ▼ | Originati. ★ ▲

No applications
Get started by creating an application.

[Create application](#)

[Go to myApplications](#)

CloudShell [Feedback](#)

© 2025, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

TRY IT YOURSELF

Amazon Bedrock

Getting started

[Overview](#)

Providers

Foundation models

Model catalog [New](#)

Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

[Image / Video](#)

Builder tools

Agents

Flows

Knowledge Bases

[Amazon Bedrock](#) > Overview

Overview [Info](#)

Foundation models

Amazon Bedrock supports over 100 foundation models from industry-leading providers and emerging leaders. Select a serverless model or Bedrock Marketplace model that is best suited for achieving your unique goals.

[View Model catalog](#)

[Discover marketplace models](#)

Model spotlight

AI

Anthropic's Claude

Choose the exact combination of intelligence, speed, and cost to suit your needs. All of the latest Claude models, like upgraded Claude 3.5 Sonnet, are available in Amazon Bedrock.

[Request model access](#)



Thank you!

Almas
Moldakanov

almasmol@amazon.de

Tymur
Sydorenko

tsydoren@amazon.pl