Движок поиска и Q&A по видеоархиву — Build Blueprint

1 — Референс-архитектура: две половины, два правила

Правило 1	Индексируем один раз, читаем по запросу — архив никогда не попадает в модель.
Правило 2	Ни одного ответа без цитаты на исходный клип и таймкод.
Ингест	Чанки по сценам -> кадры -> ASR+OCR+VLM -> embedding -> индекс (таймкод + права).
Запрос	Разбор -> гибридный ретрив -> реранк -> ответ из чанков -> цитата с таймкодом.

2 — Строить или покупать (2026): арендуем модели и вектор-БД, строим остальное

```
АРЕНДА/ХОСТ  embedding-модель + модель-ответчик (VLM/LLM) + ASR/OCR/описания — самые быстрые части
ОБЛАКО       вектор-БД (Pinecone / Qdrant / Weaviate / Milvus) — готовая инфра, не изобретать
СТРОИМ       экстракция + гибридный ретрив + цитаты + governance = ваш продукт
ВАЖНО        Twelve Labs закрыли Marengo 2.7 (30 Mar 2026) — потому embedding-модель и абстрагируют.
```

3 — Поле моделей 2026 (проверьте перед запуском)

```
EMBEDDING (дорогая замена): Gemini Embedding 2 (мультимод. 3072-dim) | Cohere Embed 4 | Marengo 3 | Nova
  open: SigLIP 2 | Jina v5 | Qwen3-Embedding  — смена АННУЛИРУЕТ все векторы -> переиндексация.
ОТВЕТ (дешёвая замена): Gemini 2.5 | Claude | GPT | open LLaVA / Qwen-VL — читает лишь неск. чанков.
Храните извлечённый текст (ASR/OCR/описания) рядом с векторами — переиндексация пойдёт из текста.
```

4 — Двухстрочная модель затрат (держите на разных статьях бюджета)

```
РАЗОВАЯ ИНДЕКСАЦИЯ  ~$2.50 / час видео (ASR ~$0.36 + описания и embedding ~$1-2) + ~центы/мес хранение.
ЗА ВОПРОС          ~$0.01-0.02 (embed + вектор-поиск + реранк + генерация по неск. чанкам).
Индексировать 1,000 ч единожды ~= $2,500 фикс., затем ~1-2c за вопрос.
Long-context: 1 час ~= 946,800 токенов x $2.50/M = ~$2.37/вопрос; весь архив НЕ влезает.
```

5 — Порядок сборки: каждая веха даёт рабочий инструмент

```
1 Строка поиска с цитатами (текстовый индекс + клипы со ссылкой на точную секунду).
2 Q&A с заземлением и воздержанием (ответ лишь из чанков; цитата на каждое) — второй, НЕ последний.
3 Гибридный ретрив + реранк (добавить ключевые слова; чинит промах ретрива).
4 Governance (права + приватность + аудит).   5 Масштаб, переиндексация, оценка ретрива.
```

6 — Запуск + governance (инженерный контекст, не юридический совет)

- [ ] Каждый чанк несёт источник + таймкод с ингеста, чтобы каждый ответ мог цитировать (Правило 2).
- [ ] Embedding-модель за одним интерфейсом + сохранён извлечённый текст: снятие модели — плановая переиндексация.
- [ ] Гибридный ретрив (семантика + ключевые) + реранк — ловятся и точные имена/коды, и смысл.
- [ ] Права фильтруются на этапе РЕТРИВА по текущим правам спрашивающего — не только на индексе.
- [ ] Поиск по личности — повышенный риск (GDPR + EU AI Act, биометрия) — часто решение НЕ строить.
- [ ] Аудит-лог каждого вопроса + клипа; ретрив оценивается по вопросам с известным ответом, не по гладкости.