

DATA COOKER ETL

Трансформация и аналитика больших данных быстрее и проще*

* По процессу Lean ETL, в сравнении с классическим подходом Big Data

Оцифровка и накопление данных

Сегодня почти все предприятия собирают данные. Много данных.
Собирается и хранится почти всё.



Зачастую это совершенно разные хранилища данных.
Миллионы записей. Терабайты информации.
И очень хаотичная структура.



- Данные можно и нужно обрабатывать и анализировать, чтобы лучше понимать, оптимизировать свои процессы, своих клиентов, и их поведение.
- Специфичные наборы данных можно продавать сторонним компаниям.



Создаётся подразделение, которое занимается ETL-решениями.

ETL процесс в управлении хранилищами данных, который включает в себя: извлечение данных из внутренних источников; их трансформация и очистка в соответствии с бизнес-требованиями; и их загрузка в целевое хранилище данных.

Проблема

Стоимость владения ETL-решением с использованием классического подхода Big Data сегодня — **дорого**.

Предприятия накапливают исторические и аналитические данные в Data Lake, что приводит к удорожанию извлечения знаний из-за необходимости постоянной актуализации схемы всех накопленных данных

Высокая стоимость оптимизации ETL-решений под большой объём данных

*Для внедрения **качественного** частного ETL-решения необходима proof of concept реализация алгоритмики сначала на Python/R, а затем оптимизация на Java/Scala/Python Spark*

Проблема → **Дорого** → Обоснование

Стоимость владения решения с использованием классического Data Lake:

*Холодное хранилище
данных (AWS S3)*

19 920
руб./мес.

+

*Постоянно работающий кластер
Hadoop с хранилищем данных*

139 123
руб./мес.

+

*Администрирование
Hadoop кластера*

60 000 руб./мес.,
160 чел-часов

*Затраты на запуск
программы*

0

*Кластер работает
постоянно*

+

*Трудозатраты аналитика на
реализацию нового алгоритма
обработки данных*

80 000 руб./мес.,
в среднем 20–80 чел-часов

+

*Трудозатраты инженера для
реализации нового ETL*

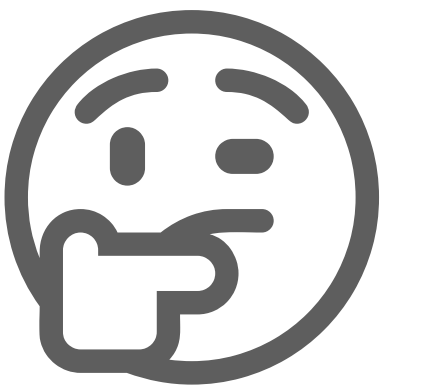
160 000 руб./мес.,
в среднем 40–80 чел-часов

= Итого: 459 043 руб./мес. или примерно 18 новых процессов в год за 5 508 516 руб.

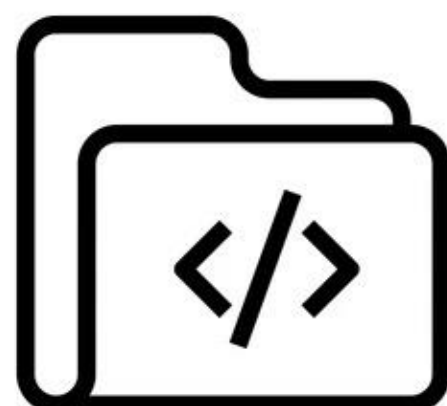
Стоимость одного процесса: **306 028 рублей**

Проблема → **Дорого** → Что делать?

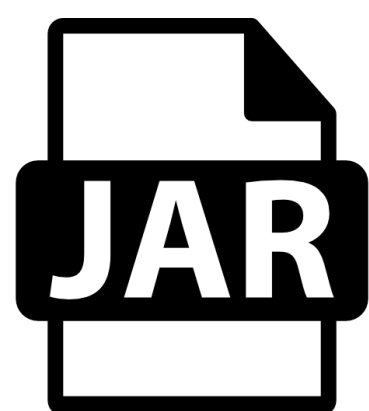
1. Оптимизация реализации алгоритмов обработки данных, реализованных в ETL-процессах → *увеличивает стоимость поддержки решения*
2. Подбор Data Lake, оптимизированного под потребности ETL-процессов → *сопровождается высокой стоимостью и большими трудозатратами для внесения данных в хранилище*
3. Использование высокоуровневых средств, уменьшающих затраты на разработку → *увеличивает затраты на проведение расчётного процесса и увеличивает «зоопарк» используемых библиотек*
4. Использование низкоуровневых средств снижает затраты на проведение расчётного процесса → *увеличивает затраты на разработку*



Решение → ETL как код и вычисления по запросу



Нами была реализована **библиотека алгоритмов трансформации данных**, которая позволяет обращаться к любому имеющемуся набору данных из хранилища, вне зависимости от его формата, и не требующее ведения централизованного каталога со схемой данных каждого набора.



Data Cooker ETL — это консольное приложение, вызываемое по запросу, что позволяет легко встроиться в конвейер любого процесса ETL. Постоянно работающий кластер Spark не требуется, также можно использовать любой оркестратор процессов и CI/CD сервис по выбору.

+

*Отказ от Data Lake, и
использование только холодного
хранилища данных*

+

*Для подтверждения гипотез:
отказ от двойной реализации
на Python/R, а затем на Spark*

Решение → Data Cooker ETL → Обоснование

Стоимость владения решения с использованием Data Cooker ETL:

Холодное хранилище
данных (AWS S3)

19 920
руб./мес.

+

Постоянно работающий кластер
Hadoop с хранилищем данных

0
Не требуется

+

Администрирование
Hadoop кластера

0
Не требуется

Затраты на запуск
программы

450 руб.
(среднее значение)

+

Трудозатраты аналитика на
реализацию нового алгоритма
обработки данных

160 000 руб./мес.,
в среднем 8–24 чел-часов

+

Трудозатраты инженера для
реализации нового ETL

80 000 руб./мес.,
в среднем 10 чел-часов

= Итого: 267 120 руб./мес. или примерно **112** новых процессов в год за 3 205 440 руб.

Стоимость одного процесса: **28 620 рублей** (дешевле в **10+** раз!)

Решение → Data Cooker ETL → Кейс использования

1. Сервисная компания, предоставляющая услуги подготовки аналитических отчётов и данных для дальнейшего использования в медиа-планировании и предсказательных моделях
2. **До внедрения** Data Cooker ETL: стоимость одного ETL процесса \$12 000 — 45 000, мощность производства — 15 новых процессов в год
3. **После внедрения** Data Cooker ETL: для разработки процесса необходим только инженер с квалификацией «аналитик данных». В результате стоимость одного ETL составила \$20—150, и мощность производства с учётом простоев между проектами — 63 процесса в год силами одного аналитика данных

Решение → Data Cooker ETL → Стоимость

Передаётся заказчику бесплатно

0 руб. Приложение BD Cooker
Выполняемый JAR-файл

0 руб. Библиотека ГОТОВЫХ
алгоритмов
В составе приложения

0 руб. Исходный код
По запросу на GitHub

0 руб. Документация
Для реализации собственных
алгоритмов в подключаемых
JAR-файлах

Дополнительные услуги

300 т.р. Обучение
Услуга обучения вашей команды по
конфигурированию, работе с ETL
конструктором и оркестратором. Курс 40
часов, проводится онлайн

2 млн. руб. Внедрение у заказчика
Услуга внедрения решения на
предприятие заказчика. Срок
выполнения 2 месяца

155 т.р. Очистка и подготовка
Услуга по очистке и подготовке больших
массивов данных в рамках проекта
конкретного заказчика. Срок выполнения
2 рабочих дня.

ООО «Готовим Данные» → Услуги

Услуги по очистке и подготовке данных

В случае, если мы предоставляем сервис обработки данных на собственной инфраструктуре, конечным результатом является:

- ▶ Поставка аналитического отчёта в виде структурированного файла в формате TSV
- ▶ Поставка документации с описанием структуры данных

Внедрение на стороне заказчика

В случае, если мы производим развёртывание в инфраструктуре клиента, в комплект поставки входят:

- ▶ Исполняемый JAR-архив с программой Data Cooker ETL
- ▶ Документация по запуску расчётного ETL процесса
- ▶ Документация по конфигурированию ETL процессов
- ▶ Обучение аналитика данных со стороны заказчика
- ▶ Создание одного ETL процесса вместе с вашими специалистами
- ▶ Помощь и консультации по развёртыванию и настройке необходимой инфраструктуры

— ?!

— **Perfecto!**

© ООО «Готовим Данные»

Презентация не является публичной офертой

По вопросам и запросам пишите в Telegram:

https://t.me/data_cooker_etl