ОНЛАЙН-ОБРАЗОВАНИЕ

Не забыть включить запись!

Меня хорошо слышно && видно?

Напишите в чат, если есть проблемы!

Ставьте + если все хорошо

Правила вебинара

Активно участвуем

Задаем вопрос в чат или голосом

slack

Off-topic обсуждаем в Slack #канал группы или #general

Вопросы вижу в чате, могу ответить не сразу

OTUS

Занятие 25. Case studies. Кейсы компаний.

Задорожный Антон

Цели вебинара

После занятия вы сможете:

1. познакомиться с некоторыми кейсами аналитики больших данных
2. вынести для себя уроки из чужих ошибок
3. вдохновиться для новых проектов ☺

Краткое содержание

- КХД в крупном банке
- Аналитика телеком данных
- Планирование закупок в энергетической сфере
- Операционный datalake для банка и LC&I
- Streaming платформы в крупном банке
- Deep learning на производстве
- Datalake в фармацевтической компании

КХД в крупном банке

Проект

- Крупный банк в Восточной Европе
- Программа КХД с замахом на все источники данных (только крупных источников – 300 типов систем)
- Первые два проекта с суммарной длительностью около года интегрируют 20+ типов источников
- Сложные бизнес-правила реконсиляции данных
- Инструментарий:
  - СУБД: Teradata
  - ETL: Informatica
  - BI/Reporting: BusinessObjects

Поговорим о сложности («трудности»)

Сложность («трудность») – это размер усилий, которые необходимо затратить на решение задачи

Как вы думаете, от чего зависит сложность типичные задачи построения аналитической системы?

- Физически консолидировать данные в одном месте («скачать данные в одно место»)
- Интегрировать данные («сделать из двух таблиц клиентов одну»)
- Историзовать данные («понять какая запись – это новый клиент, а какая – изменение определенного атрибута»)
- …

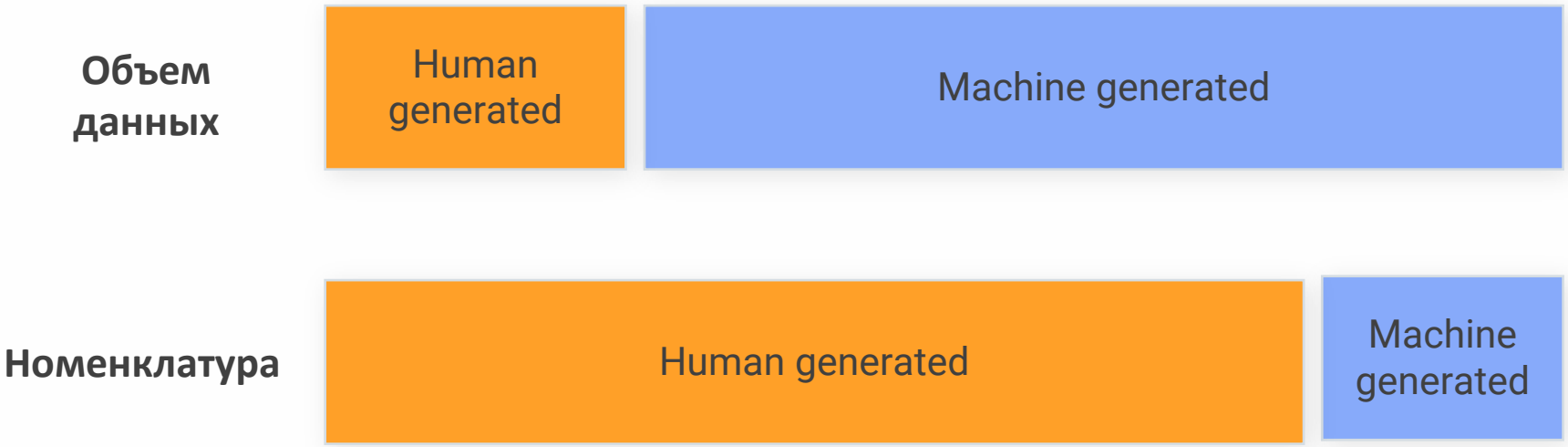
- Объем данных: гигабайт в начальной загрузке, ежедневно, в секунду (в пике)
- Номенклатура: количество источников, сущностей, связей, таблиц, колонок

Два основных источника сложности

OTUS

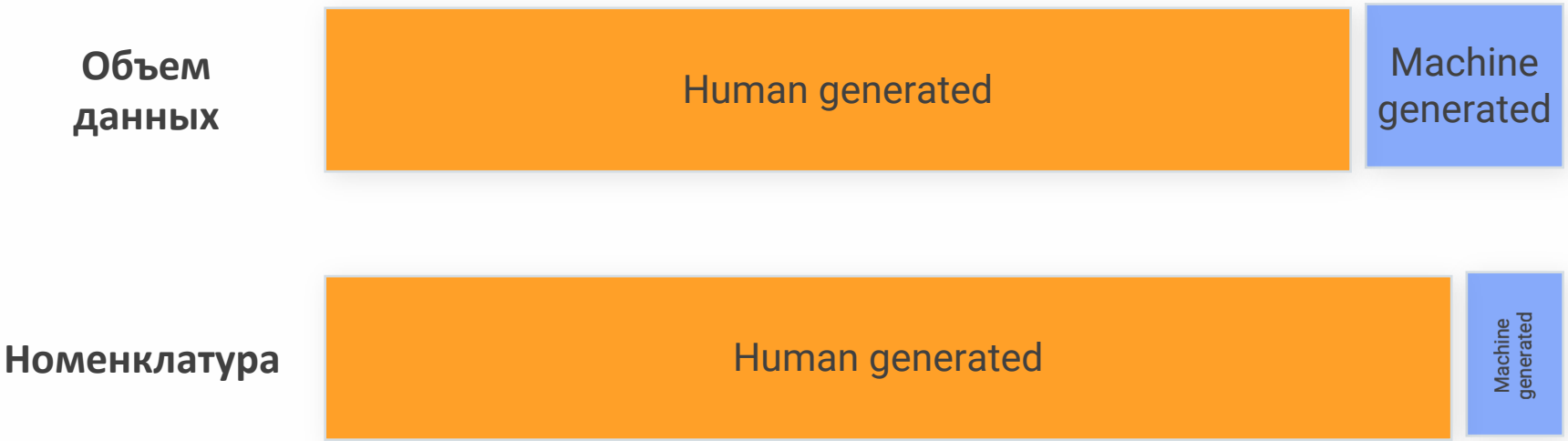
В классической компании все данные в источниках можно поделить на два класса:

1. Human generated – данные, которые создаются людьми, и входят в бизнес-процесс: анкеты абонентов в телекоме, описание проблемы в хелп-деске, операции с банковским счетом и проч
2. Machine generated – данные, которые сгенерированы машиной как фиксация события внутри системы: CDR в телекоме, запись в логе веб-сайта или мобильного приложения, показатель с сенсора



Уроки кейса

- Отдельно оценивать сложность консолидации и интеграции
- Разделение задач консолидации и интеграции по разным задачам (или командам) – путь к успеху
- Знайте (и умейте работать) с ограничениями вашего ETL инструмента
- Автоматизируйте стандартные операции: получение данных, интеграции
- Для масштабного внедрения нужны Continuous Delivery и метаданные с самого начала

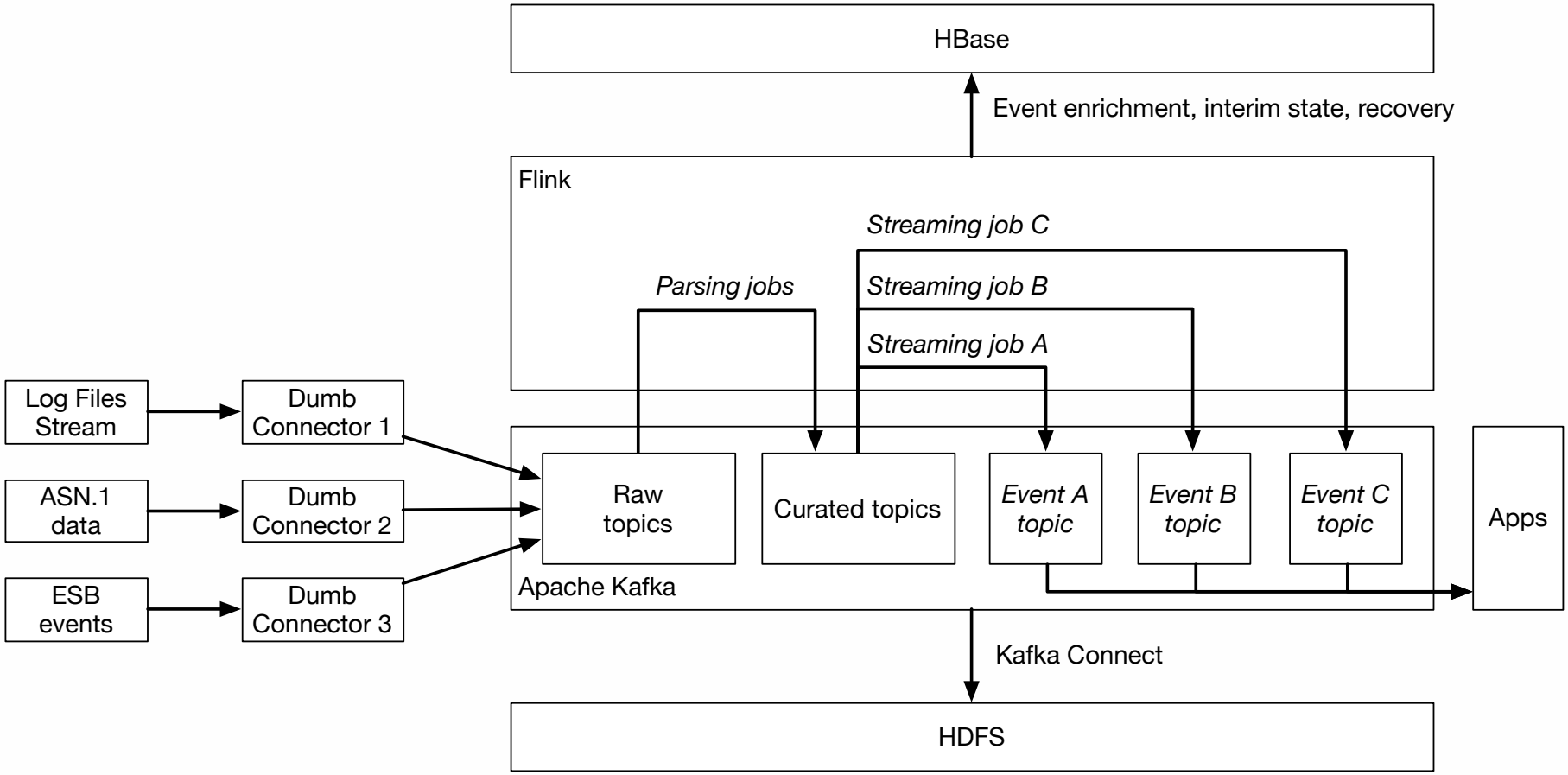


Аналитика телеком данных

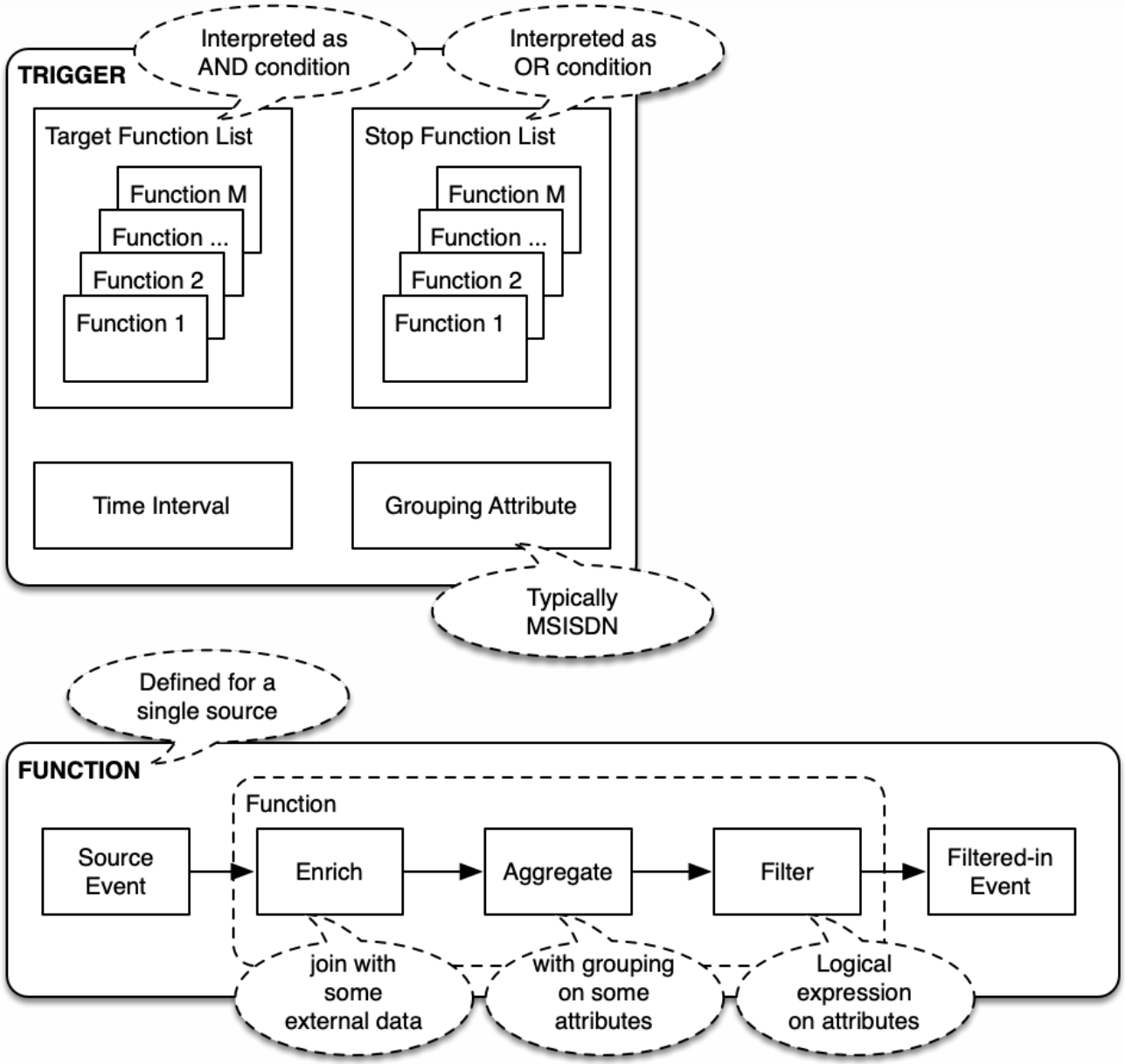
Проект

- Крупный мобильный оператор
- Программа Big Data сервисов для внутренних и внешних кейсов
- Платформа операционной аналитики, Machine Learning as a Service, с поддержкой streaming аналитики и CEP (Complex Event Processing)
- Инструментарий:
  - Hadoop, HBase
  - Kestrel -> Kafka, RabbitMQ
  - Storm -> Flink
  - MapReduce -> Spark, Impala

Архитектура



Complex Event Processing



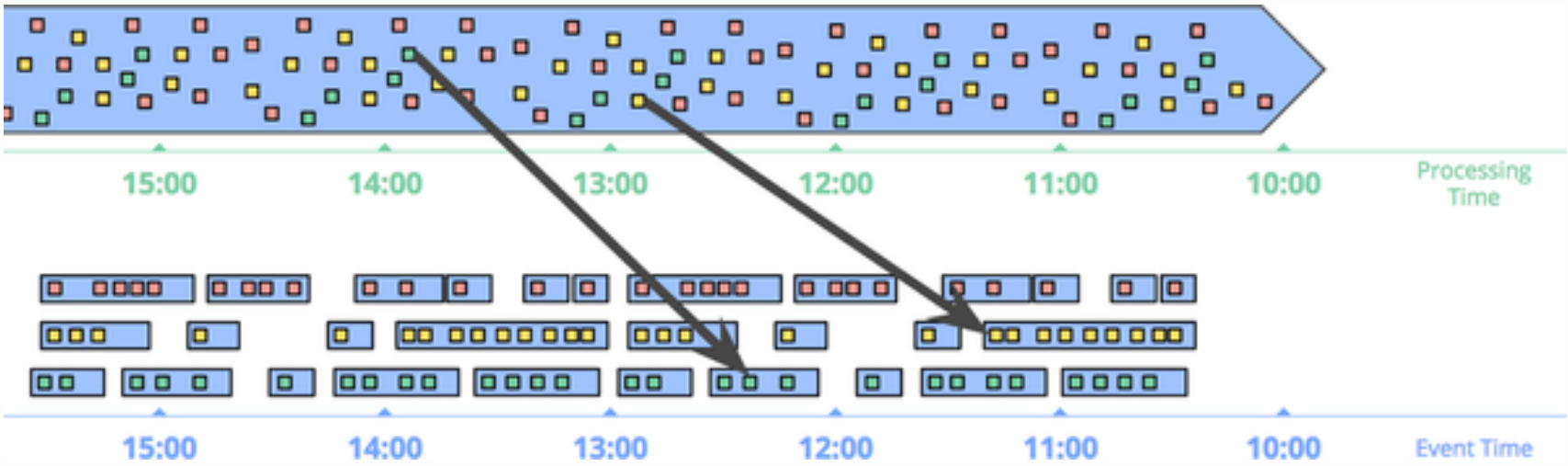
Использование

- Внутренние сервисы: финансы, сеть, маркетинг
- Внешние сервисы: финансы, smart city, реклама

- На пике активности оператор отчитался что 1% годовой выручки приносили Big Data сервис, речь о новом бизнесе, без учета улучшений в маркетинге и обслуживании клиента

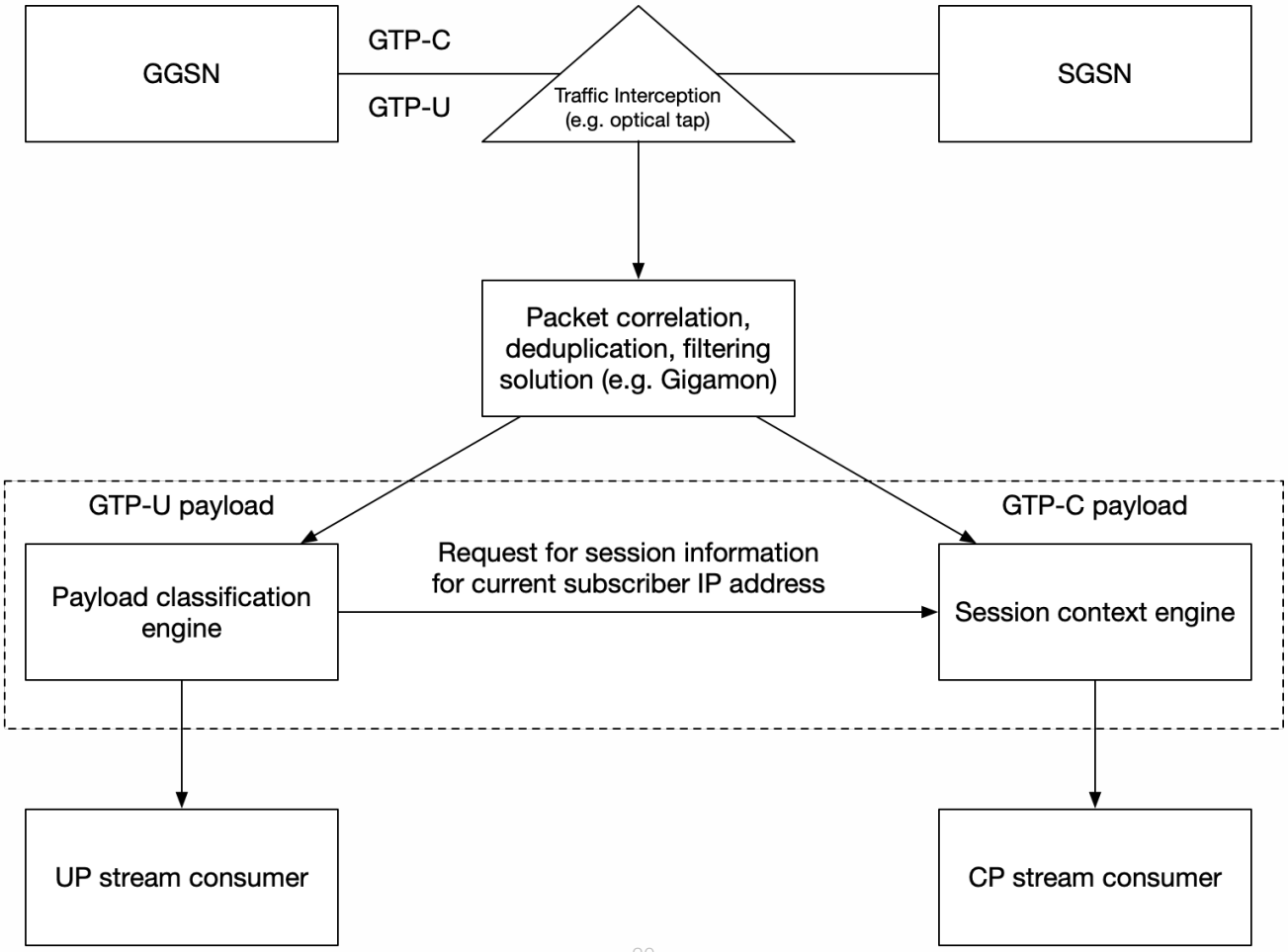
Сессонизация

- Группировка событий по определенным бизнес-правилам
- Сущность «сессия» крайне важна для целого ряда метрик, оценивающих вовлеченность пользователя, качество сервиса и других аспектов взаимодействия
- Сессонизацию сложно делать на больших данных в батче, так как она требует джоина по условию больше или меньше



Streaming join

- Join «очень большой таблицы» с другой «очень большой таблицей»
- Bucketed Tables могут помочь, но все-равно медленно
- На помощь Streaming Join



Уроки кейса

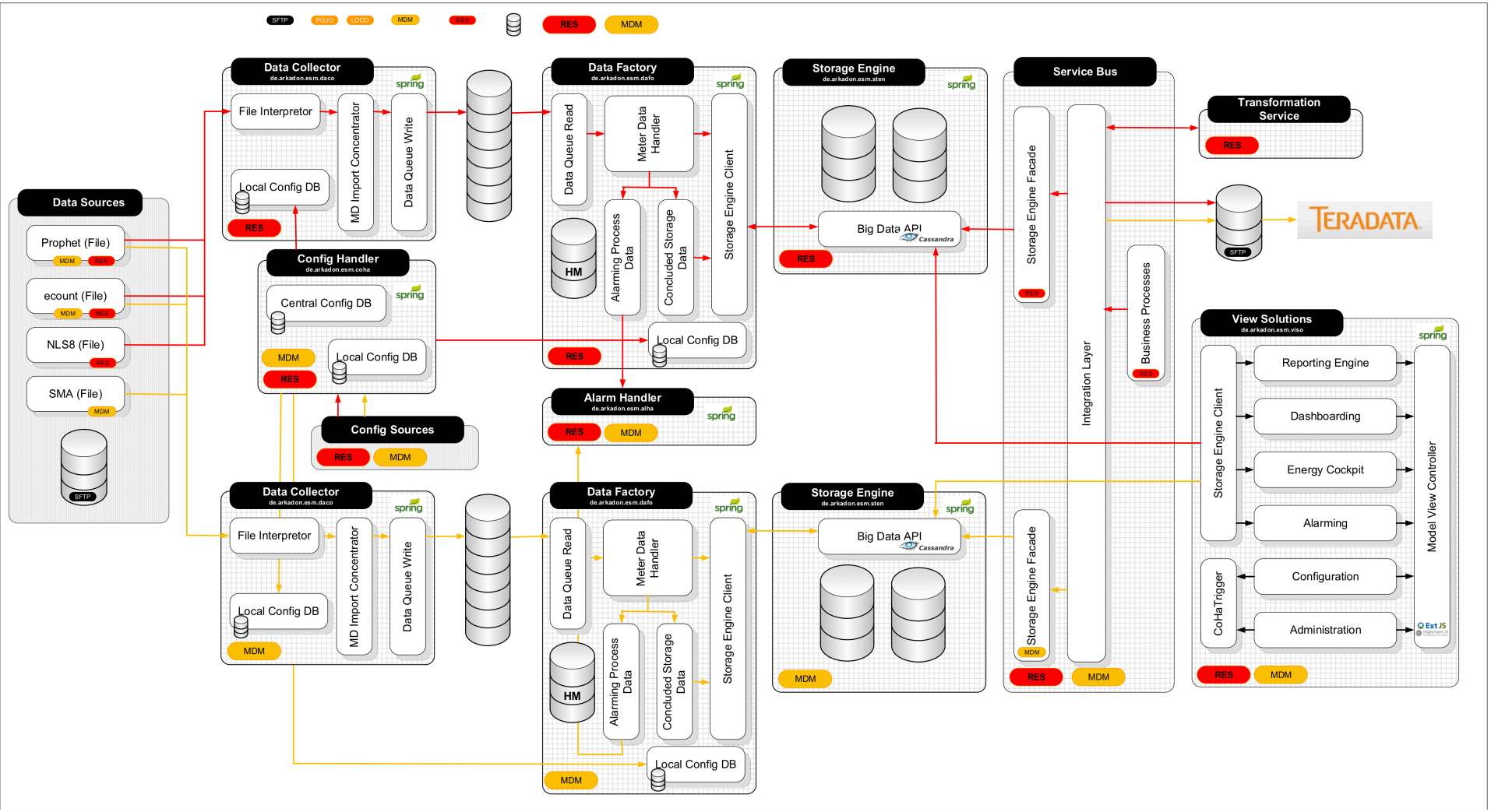
- Делая решение – думайте и о платформе
- Streaming может быть не усложнением, а упрощением конкретного решения
- Качество должно быть таким же ориентиром как производительность или скорость выкатки

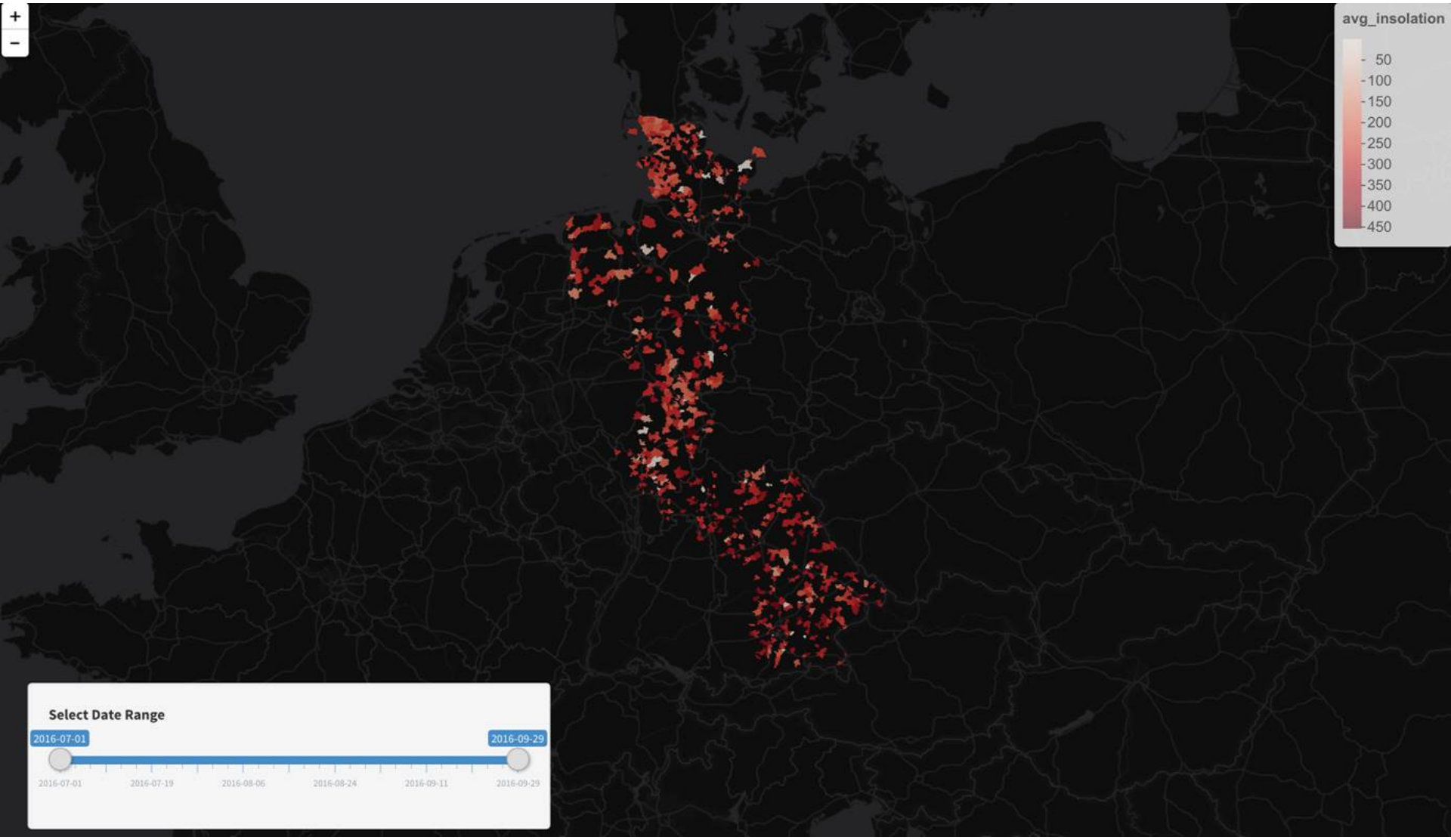
Планирование закупок в энергетической сфере

Проект

- Распределительная компания в Европе обязана покупать всю энергию от воспроизводимых источников (ветер, солнце, гидро)
- При этом энергия ветра и солнца очень нестабильны, и для компенсации компания вынуждена закупать энергию на спотовом рынке
- Чем точнее предсказаны спрос (и предложение на рынке RES), тем точнее может быть сделана ставка, сохраняя деньги компании
- Инструменты:
  - Cloudera Hadoop
  - Cassandra
  - Spark
  - R, Shiny

Архитектура




+
−
avg_insolation
50
100
150
200
250
300
350
400
450
Select Date Range
2016-07-01
2016-09-29
2016-07-01
2016-07-19
2016-08-06
2016-08-24
2016-09-11
2016-09-29


Forecast
Performance
Benchmark
SMA
Zeithorizont:
2015-01-
bis
2016-11-
Ist-Datum:
2016-11-01
Provider
Emsys
Enercast
Meteologica
Meteocontrol
PV - Forecast
emsys
enercast
meteocontrol
meteologica
3000
2500
2000
1500
1000
500
11:00
11:30
12:00
12:30
13:00
13:30
14:00
14:30
15:00
15:30
16:00
16:30
17:00
17:30
EMSYS
21
ENERCAST
20
METEOLOGICA
25
METEOCONTROL
19

Уроки кейса

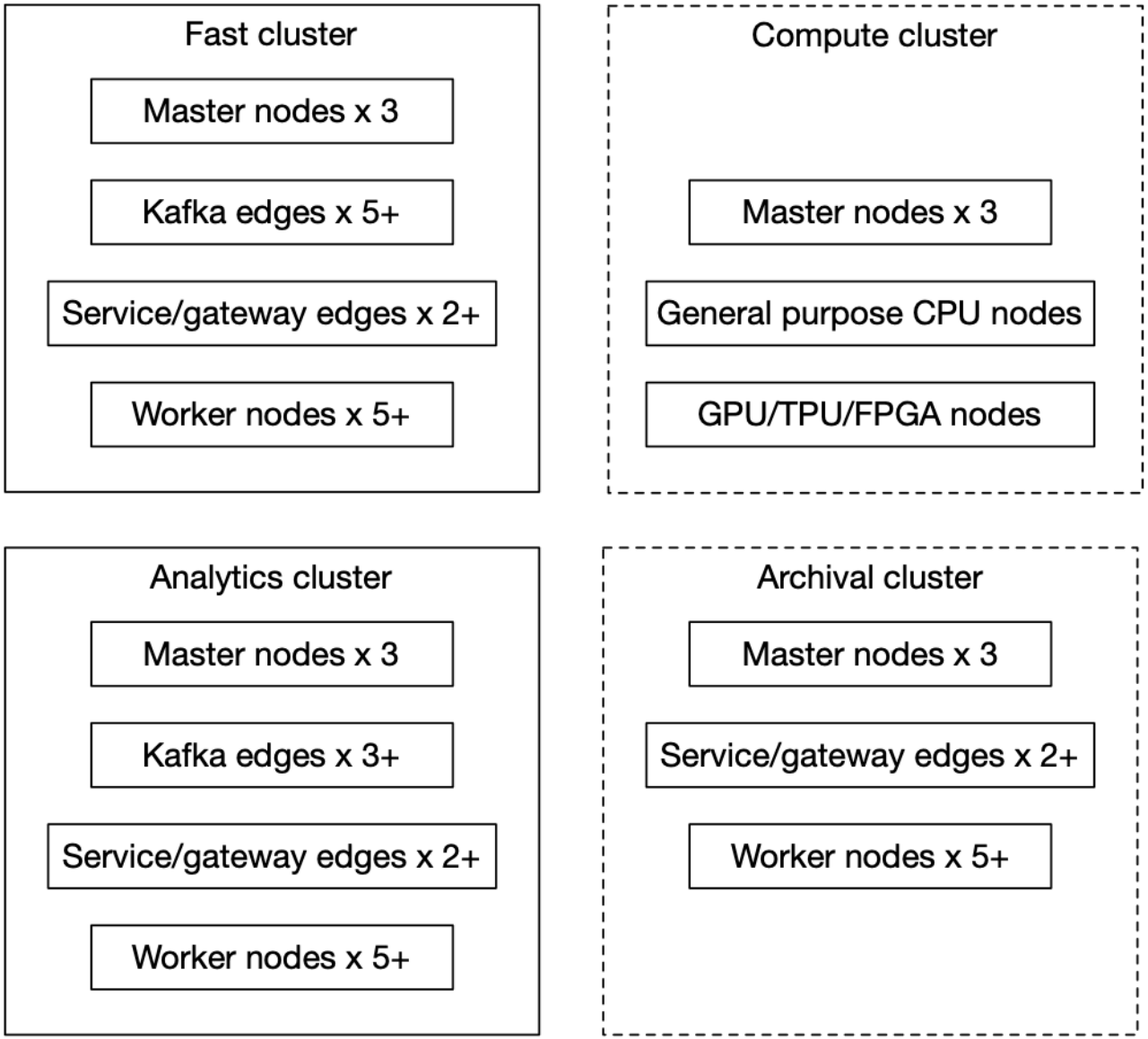
- Работа с временными рядами требует отдельного навыка от аналитиков
- Сенсорные данные нужно агрегировать
- Для кейсов сопряженных с физическим миром важны внешние данные, они стоят денег

Операционный datalake для банка и LC&I

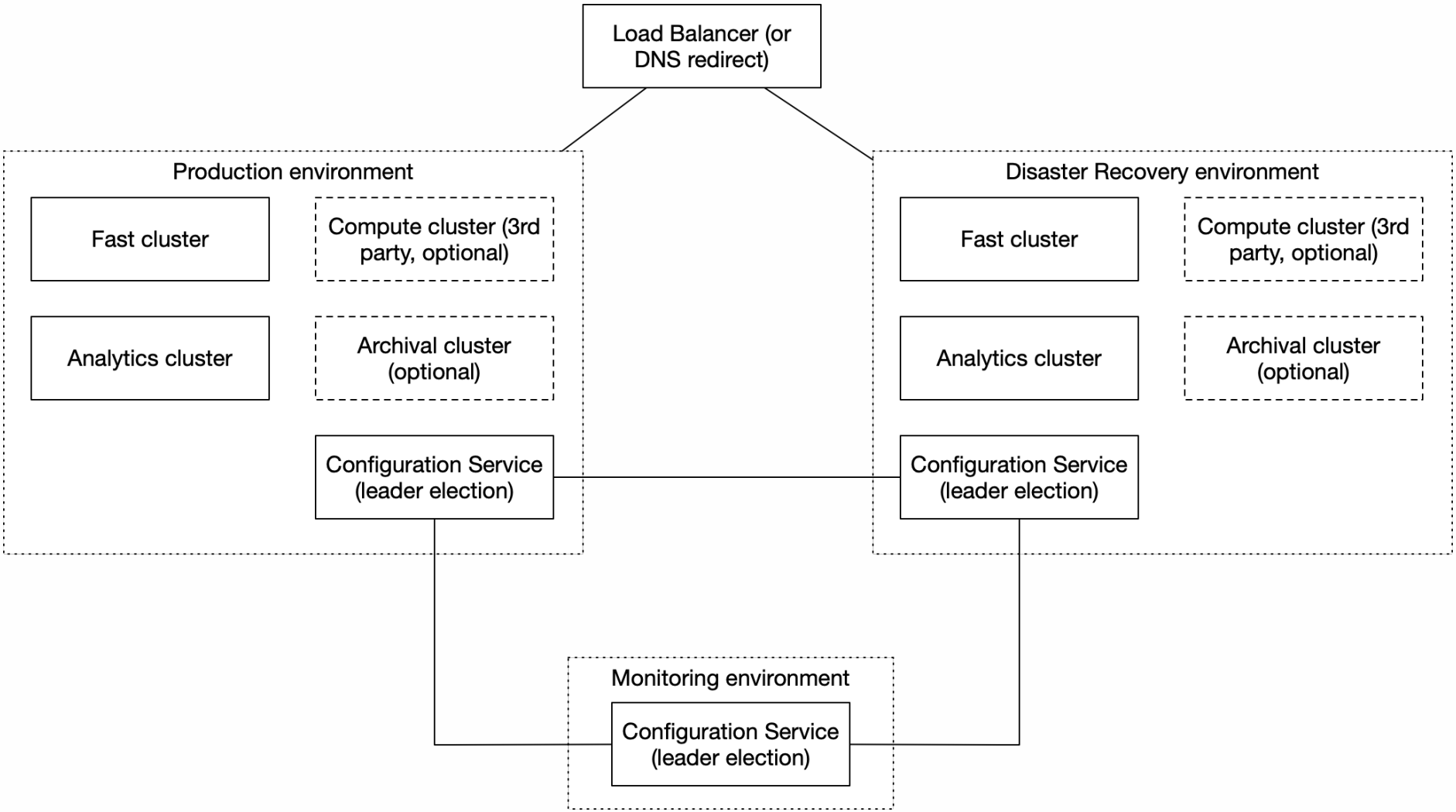
Проект

- Крупный универсальный банк в Северной Европе
- Сильная Data Science практика
- Один из бизнесов – портфельное управление Large Corporations & Institutions
- Datalake не только решает задачи аналитики, но и работает бэкендом для многих операционных сервисов
- Инструментарий:
  - Hortonworks Hadoop
  - Confluent Kafka
  - Open Shift, Nvidia GPUs

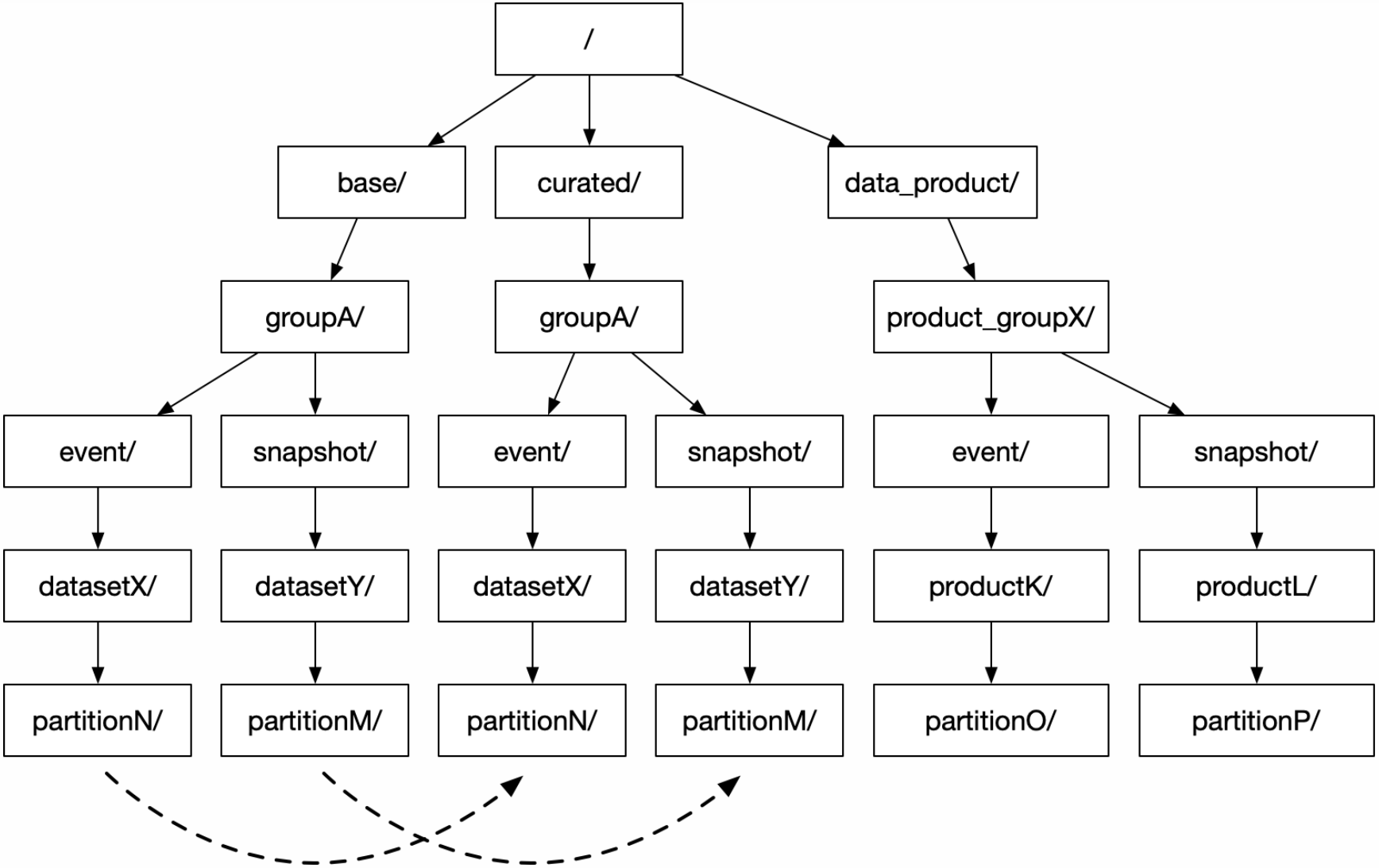
Архитектура



Датацентры


Load Balancer (or DNS redirect)
Production environment
Fast cluster
Compute cluster (3rd party, optional)
Analytics cluster
Archival cluster (optional)
Configuration Service (leader election)
Disaster Recovery environment
Fast cluster
Compute cluster (3rd party, optional)
Analytics cluster
Archival cluster (optional)
Configuration Service (leader election)
Monitoring environment
Configuration Service (leader election)

Структура данных в HDFS



Три столпа аналитики портфельного управления

1. Обновление данных: для качественной аналитики нужно иметь онлайн обновление рыночных данных
2. Complex Event Processing: возможность детектировать события по определенной бизнес-логике
3. Microbatch/batch on-demand: возможность запускать быстрые пересчеты по триггерам

Для CEP важным подспорьем является использование хэш-структур данных:

- HyperLogLog реализует count distinct
- Count-min sketch реализует groupBy,count
- И несколько других

Эти структуры можно сохранять, комбинировать, использовать также в батчевой аналитике

Уроки кейса

- Необходимость использовать свое железо не означает отсутствие эластичности и отказоустойчивости
- … и свое железо очень быстрое!
- Data locality не нужна (для пакетной аналитики)
- Определенные задачи можно решать гораздо быстрее если пожертвовать точностью
- Если на раннем этапе создать хорошую модель безопасности, то внедрение может идти гораздо быстрее и спокойнее

Streaming платформы в крупном банке

Проект

- Крупный банк в США
- Отдельная организация (со своим ВП) отвечает за стриминг, как основу бизнеса
- https://www.confluent.io/kafka-summit-SF18/building-an-enterprise-streaming-platform-capital-one
- Инструментарий:
  - Confluent Kafka
  - Spark
  - …

Creating a Public Stream via DevExchange

How to create a stream during Design Time

- Stream Details
- Steam Name
- Teams and Owners
- Tags and Metadata

- Schema Definition
- Schema Validations
- Client ID
- Client Secret
- Provision to Dev
- SDK Config

- Data Rules
- Data Security

- Schema
- Schema Validations (Gypsum)
- Datawise Rules
- Data Security

- Client ID
- Client Secret
- Provision to QA
- SDK Config

- Client ID
- Client Secret
- Provision to Prod
- SDK Config

Provisioning occurs during Design, QA and Prod

SDK Features

Feature Description	SCALA	PYTHON	SPARK
Creation of PEM files or Java Keystore files from DX credentials	✓	✓	✓
Data Validation against registered Avro schema	✓	✓	✓
Avro conversions	✓	✓	✓
Execution of data rules and watermarking	✓	✓	✓
Auto fill Enterprise Envelope fields	✓	✓	✓
Secondary cluster support for Producers	✓	✓	-
Tracking Consumer positions between clusters	✓	✓	-

Уроки кейса

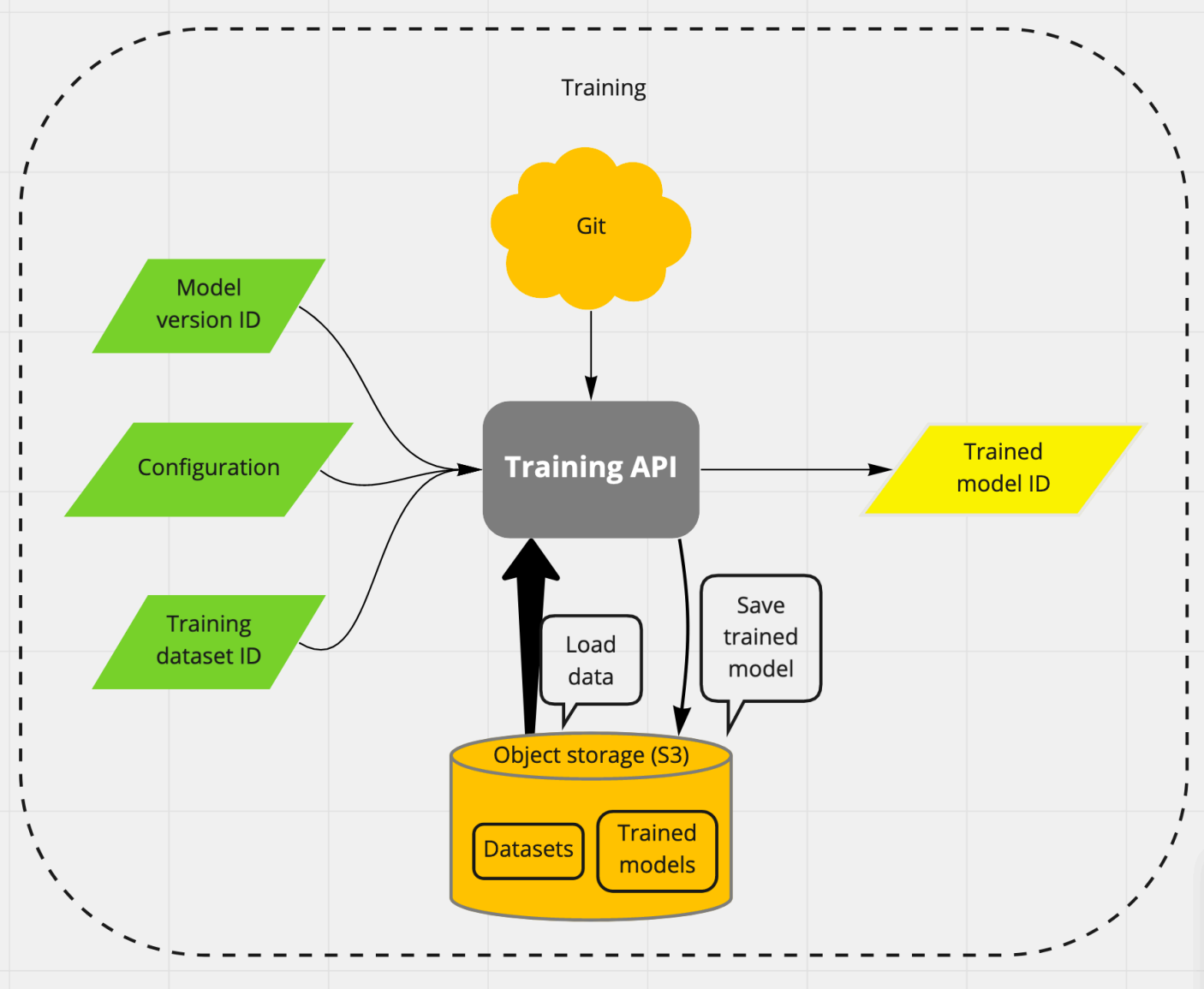
- Платформа как внутренний продукт со своим SDK
- Можно реализовать требования «на клиенте» – не стоит этим пренебрегать
- Помогает быть отдельной структурой внутри компании

Deep learning на производстве

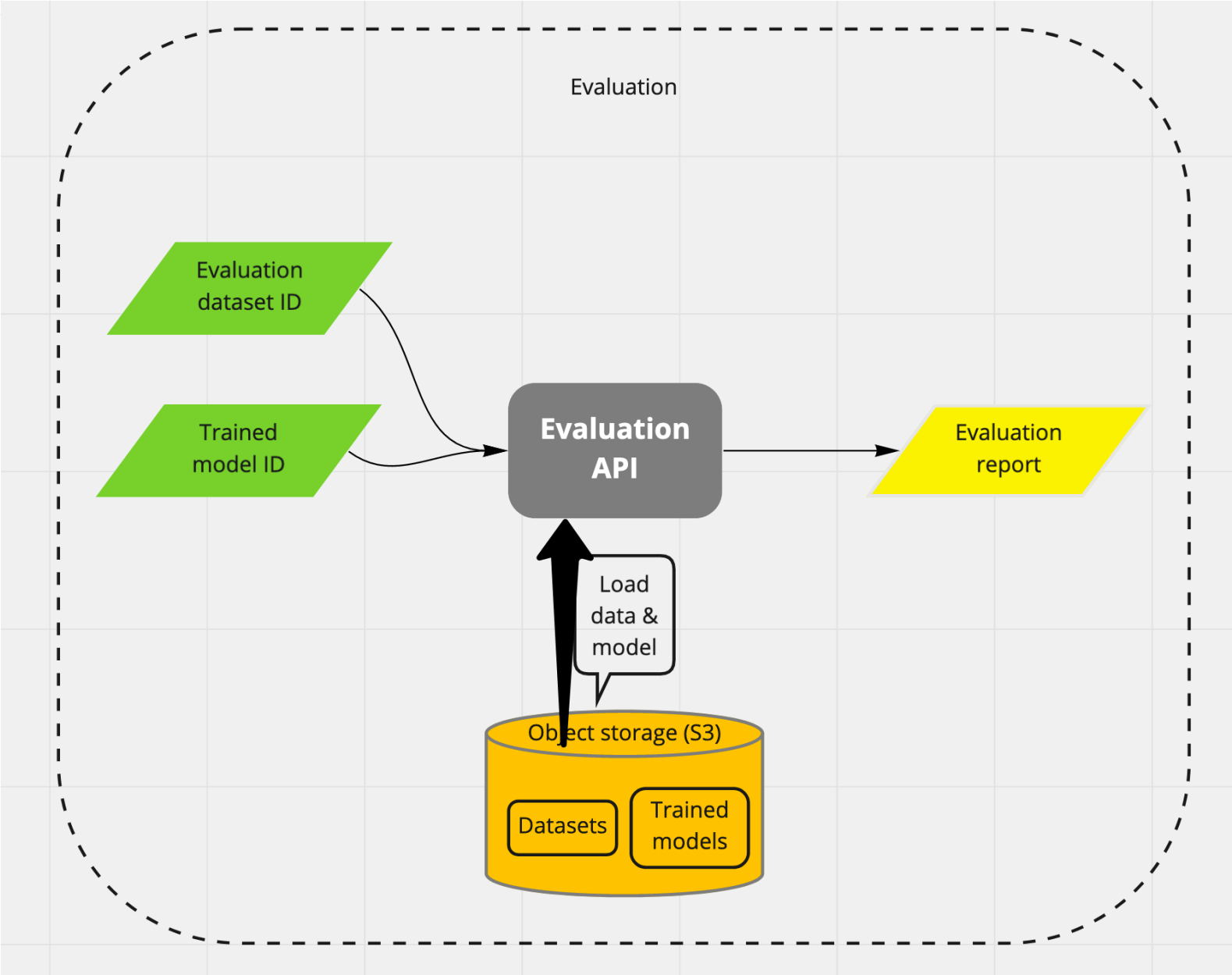
Проект

- Автопроизводитель
- Точечная сварка дает сбои, нужна аналитика которая обнаруживает плохие швы по фотографии
- Deep learning модель обнаруживает плохие швы на основе данных собранных ультразвуковым исследованием
- Инструменты:
  - AWS стэк
  - Seldon
  - Kubeflow

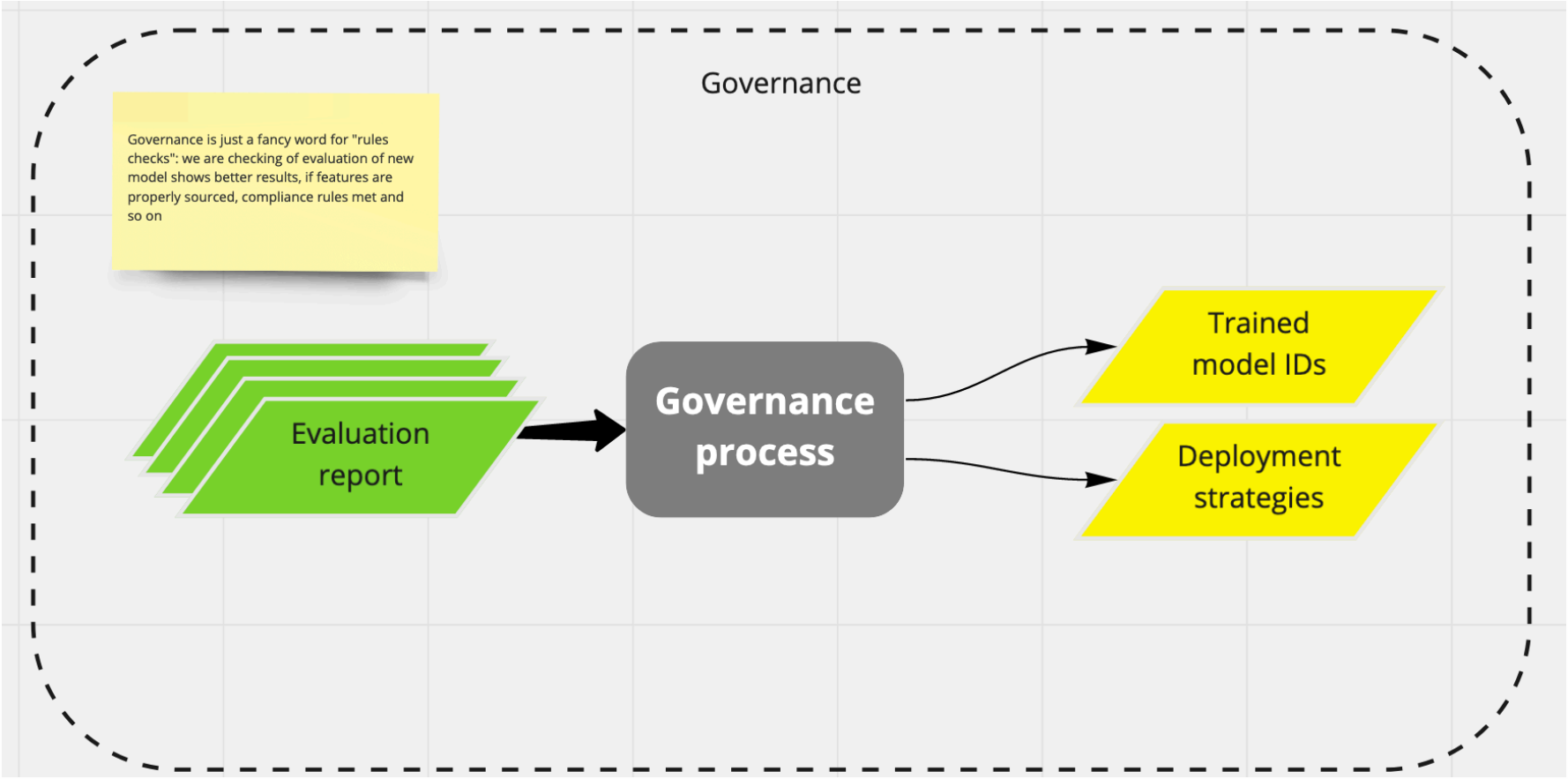
Обучение


Training
Git
Model version ID
Configuration
Training dataset ID
Training API
Trained model ID
Load data
Save trained model
Object storage (S3)
Datasets
Trained models

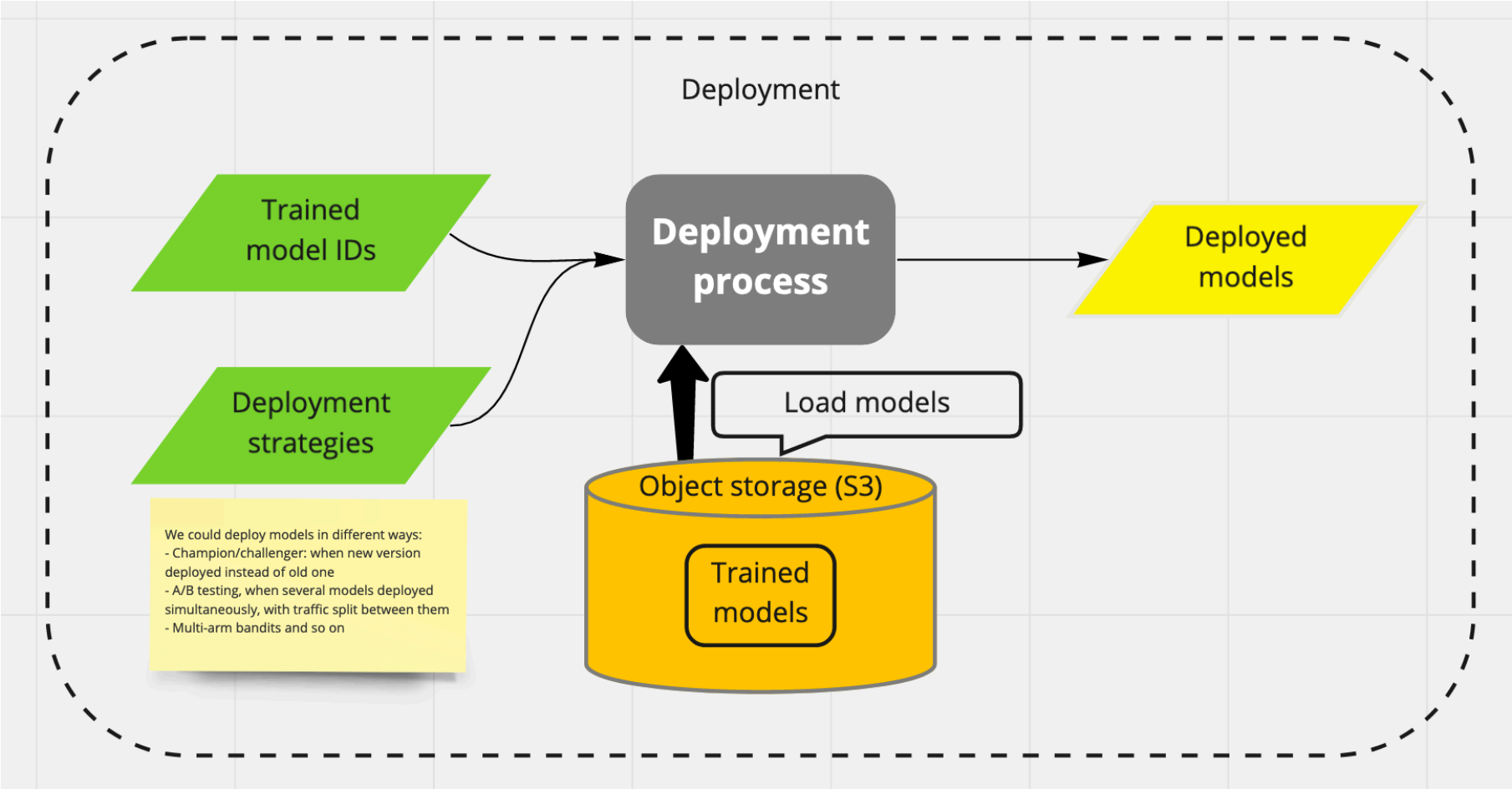
Оценка


Evaluation
Evaluation dataset ID
Trained model ID
Evaluation API
Evaluation report
Load data & model
Object storage (S3)
Datasets
Trained models

Управление


Governance
Governance is just a fancy word for "rules checks": we are checking of evaluation of new model shows better results, if features are properly sourced, compliance rules met and so on
Evaluation report
Governance process
Trained model IDs
Deployment strategies

Развертывание



Уроки кейса

- Machine learning это процесс параллельный дата/ПО инжинирингу, и должен быть формализован
- Метаданные это ключ к успеху

Datalake в фармацевтической компании

Проект

- Компания которая разрабатывает и производит лекарства и медицинское оборудование
- Американская, с присутствием в Европе: HIPAA + GDPR + …
- Datalake(s) на AWS
- Инструментарий:
  - EMR
  - S3

Изоляция ресурсов

- Данные хранятся в S3, структурированные по владельцу
- Владелец управляет набором IAM ролей
- Специальное ПО Tenant Manager отслеживает какой группе пользователь относится, и создает динамическую Security Configuration
- Пользователю предоставляется API которое позволяет запустить джоб или кластер только с подключенной Security Configuration

Анонимизация данных

- Замена ключей суррогатными / хэшами с фиксированной солью или сокрытие их через REVOKE не является анонимизацией, это маскирование или псевдоанонимизация
- Анонимизация это отрыв персоны от события, и требует удаления ключа как такового
- Существует два семейства моделей анонимизации:
  - k-анонимность
  - дифференциальная приватность

Семейство k-анонимность

- Определение: массив данных называется k-анонимным, если для каждого подмножества атрибутов существует, для любой строки существует k-1 строка с таким же значением атрибутов

- Позволяет рассуждать о защите персональных данных в отторгаемых массивах данных

- Реализуется группировкой записей и перекодированием (генерализацией) значений атрибутов

- Развитие темы:
- p-чувствительная k-анонимность
- l-разнообразие
- t-близость

- Все модели k-анонимности подвержены атаке связывания

Дифференциальная приватность

- Определение (упрощенное): рандомизированный запрос M к набору данных D считается дифференциально приватным, если для любой пары D и Di (получен удалением одной строки из D) невозможно идентифицировать удаленную строку
- Реализуется при помощи следующих механизмов:
- Сэмплирование исходного набора данных
- Добавление шума со сходным распределением для маскирования исходных значений

Уроки кейса

- Аналитика на анонимных данных возможна
- Развитая модель безопасности и ролей, связанная с инфраструктурной автоматизацией позволяет избежать строительства большого количества разрозненных песочниц

Следующий вебинар

Тема: Бонус. Дальнейшее развитие Hard skills + Soft skills.

Понедельник 2019.09.09 в 20.00

Ссылка на вебинар будет в ЛК за 15 минут

Заполните, пожалуйста, опрос о занятии

Спасибо
за внимание!

ОНЛАЙН-ОБРАЗОВАНИЕ