S7 Airlines

Enterprise Data Platform

Инфраструктура данных как полигон для проверки бизнес-гипотез

Андрей Жуков, С7 Техлаб

Что здесь происходит

Кто мы такие

Авиакомпания S7, или Сибирь

- Второй перевозчик в России
- В 2020 – первый
- Пережили этот год, стали сильнее

Кто мы такие

C7 Техлаб

- Подразделение S7, которое занимается инновационными продуктами
- Код пишет даже CEO
- В подготовке презентации участвовал даже экономист :)

S7 Airlines

Обо мне

Привет! Я Андрей.

CTO в S7 Техлаб.

Техлид платформы данных.

Характер дата-инженерный, женат.

О чем мы рассказываем

- Как в С7 появилась и развивает платформа данных.
- Как мы обнаружили, что получился Data Mesh
- Что у нас получилось, что нет и какие классные штуки появились в процессе

Кому мы рассказываем

- Тем, кто еще только строит платформу данных и хочет узнать о победах и поражениях других
- Тем, кто уже ее сделал и ищет новые идеи для развития
- Тем, кто хочет понимать все эти странные слова от бизнеса и архитекторов

Как построен наш рассказ

Три акта с двумя антрактам

- Введение и интроспекция
- Погружение и действо
- Эпилог и заключение

Часть первая. Введение

Зачем нам все это понадобилось

Как мы задумались о новой архитектуре

- Архитектура данных заточена преимущественно под отчетность
- Задачи «перекидываются через стенку» командами
- Поиск и получение доступа к данным занимает недели
- Слой мгновенных данных отсутствует

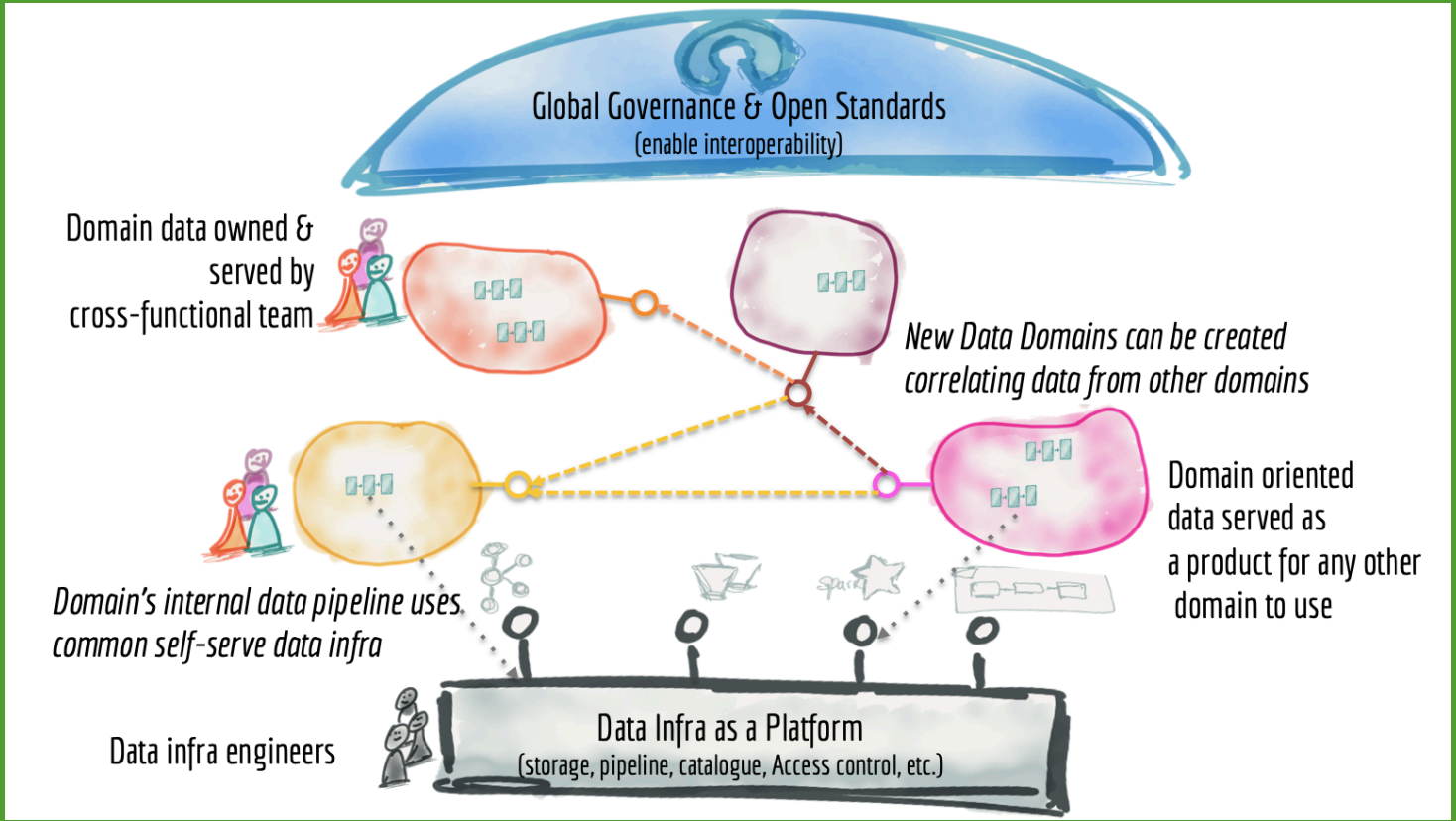
Первые шаги и ошибки

- Внешние люди плохо знают ваш домен (но показывают красивые картинки)
- Собрать и описать сразу все данные не вышло
- Сначала вместо Data Lake получился Data Swamp

Как мы случайно пришли к Data Mesh

S7 Airlines

Что такое Data Mesh



https://martinfowler.com/articles/data-monolith-to-mesh.html

Главные свойства данных по Data Mesh

- Удобство и легкость обнаружения (discoverable)
- Наличие уникального адреса (addressable)
- Доверие к данным (thrustworthy)
- Понятное описание семантики и синтаксиса данных (self-describing)
- Интегрируемость данных и стандарты, определяемые на уровне всей организации (inter operable)
- Безопасность данных и контроль доступа к ним (secure)

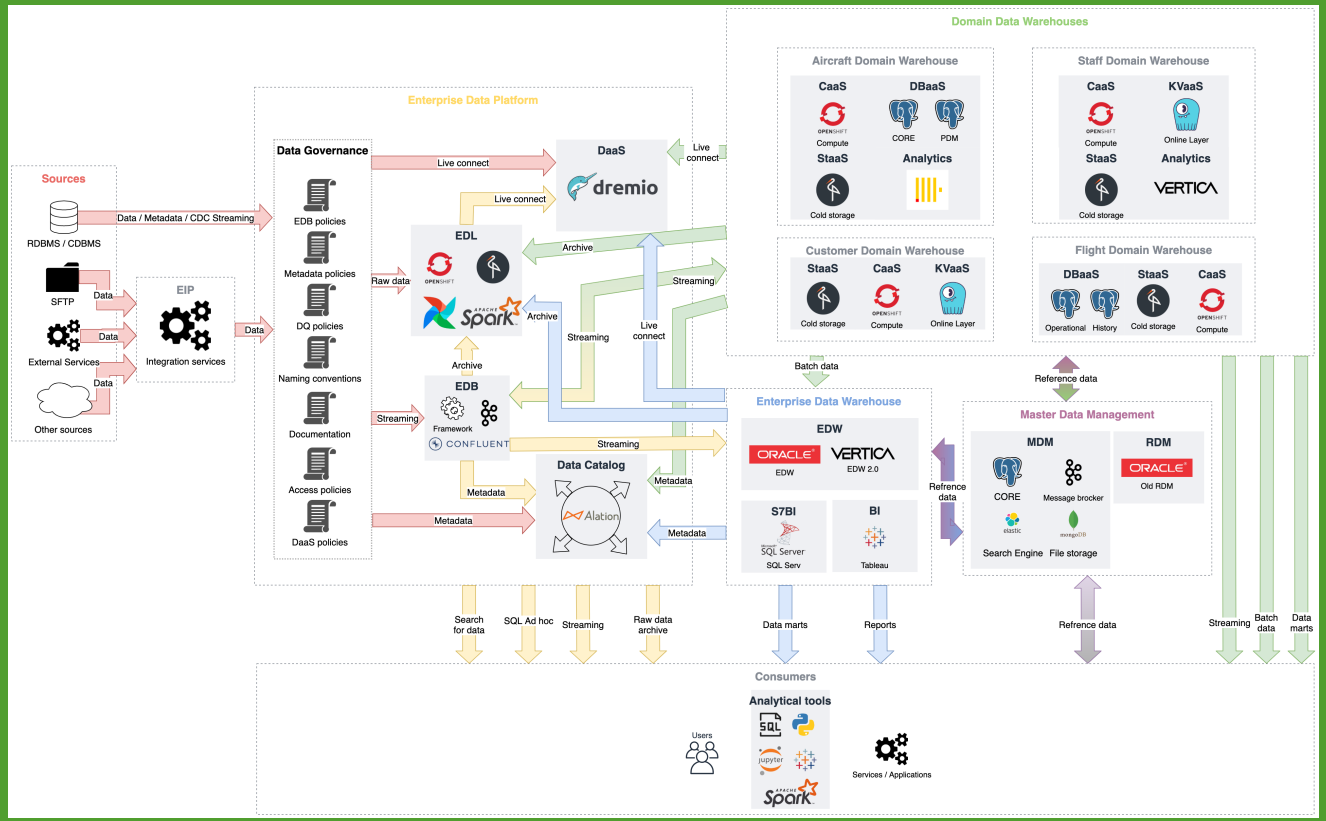
Как мы задумались о доменах

- Смена парадигмы от систем к данным: дата-центричность
- Уже созданные продукты начали создавать вокруг себя инфраструктуру данных, минуя Data Swamp
- Мы просто стали много думать:)

Как домены привели нас к Data Mesh

- От специализированных команд к кросс-функциональным
- От продуктов к платформам и к доменным хранилищам
- От систем к инфраструктуре

Что получилось (не пугайтесь)



S7 Airlines

Часть вторая.
Как мы строили, строили

Важные штуки

DWH – это вам не легаси!

Важные штуки

BI – наш главный друг!

Важные штуки

MDM – то, чему мы доверяем

Важные штуки

Интеграционная шина – то, что мы любим

Важные штуки

EIP

Integration services

Enterprise Data Warehouse

EDW

ORACLE

EDW

VERTICA

EDW 2.0

S7BI

Microsoft SQL Server

SQL Serv

BI

Tableau

Master Data Management

MDM

CORE

Message Broker

elastic

mongoDB

Search Engine

File storage

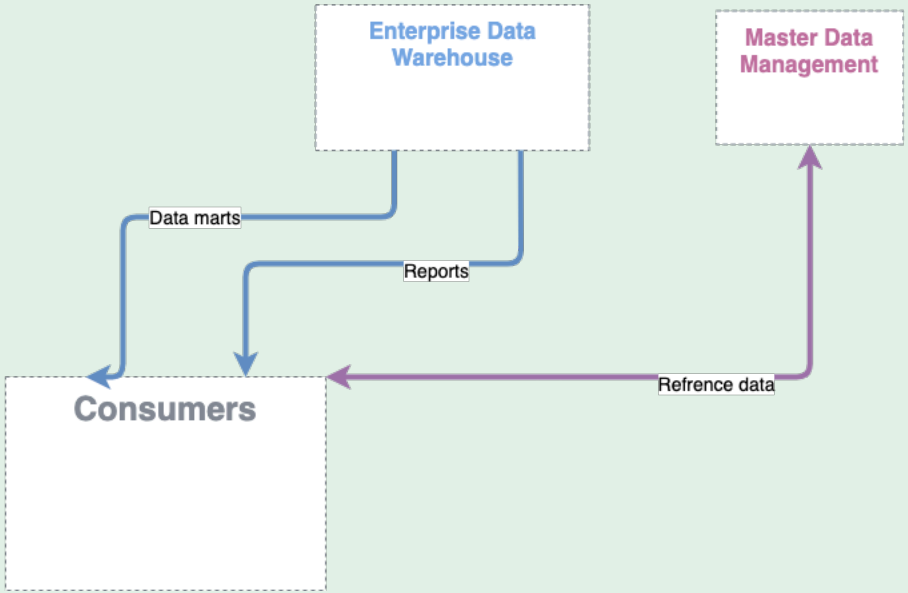
RDM

ORACLE

Old RDM

Важные штуки

EIP


Enterprise Data Warehouse
Master Data Management
Data marts
Reports
Refrence data
Consumers

Инфраструктура

*aaS как стиль жизни

- Контейнеры как сервис
- Хранилище как сервис
- Базы как сервис
- И другие

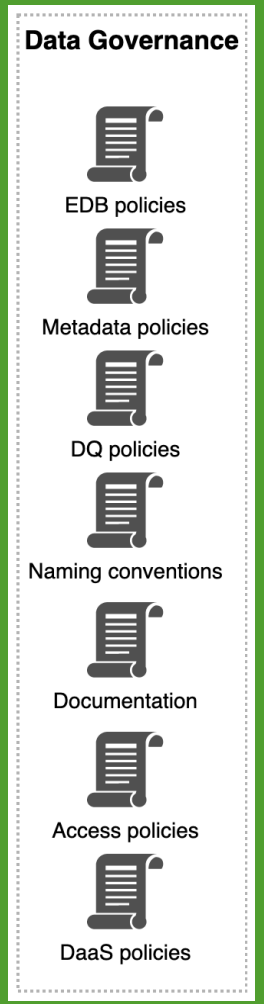
Инфраструктура

*aaS как стиль жизни

Платформа

Data Lake

- Управление данными
- Конвейеры данных
- Шаблоны и политики



S7 Airlines

Data Lake

Полученный опыт

- Это инструмент, а не самоцель
- Это должно быть удобно
- Это должно быть описываемо

Data Bus

- Real-time данные
- Транспорт данных
- Контроль схем

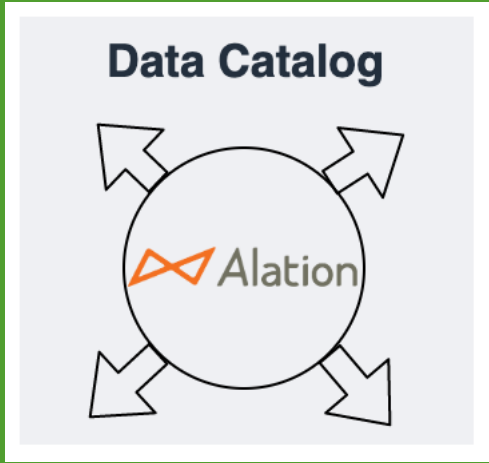
Data Bus

Полученный опыт

- Держите баланс между свободой и хаосом
- Не злоупотребляйте CDC
- Это тоже должно быть описываемо

Data Catalog

- Данные о данных (метаданные)
- Жизненный цикл данных (data lineage)
- Поиск и распространение



Data Catalogue

Полученный опыт

- Думайте о метаданных сразу
- Думайте о пользователях
- Не увлекайтесь!

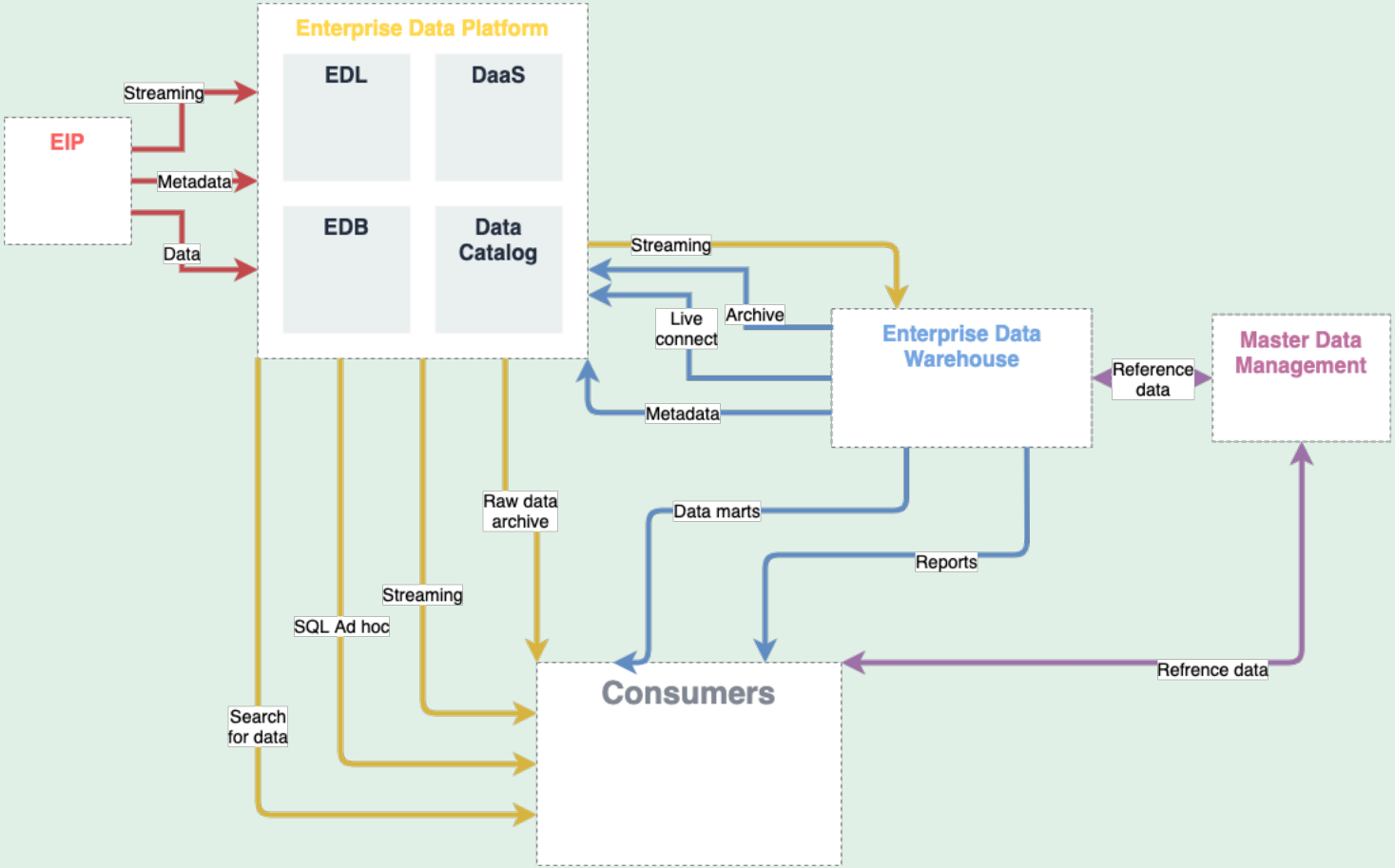
Data Federation/Virtualization

- Единая точка для анализа данных
- Максимально простой интерфейс
- Интеграция с BI

Data Federation/Virtualization

Полученный опыт

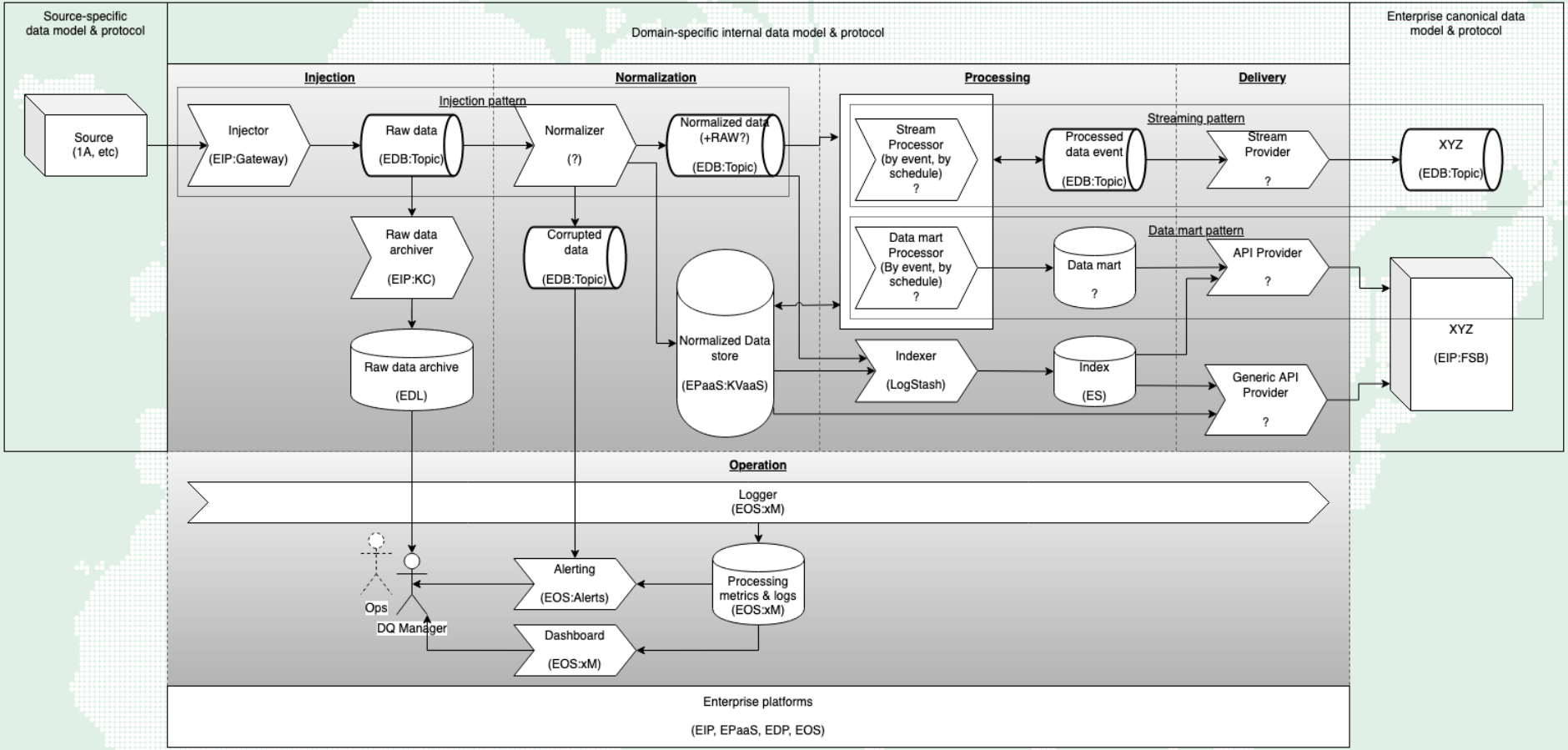
- Заботьтесь о пользователях
- Не используйте для интеграции со всем подряд :)


Enterprise Data Platform
EDL
DaaS
EDB
Data Catalog
EIP
Streaming
Metadata
Data
Streaming
Live connect
Archive
Metadata
Enterprise Data Warehouse
Reference data
Master Data Management
Raw data archive
Data marts
Reports
Streaming
SQL Ad hoc
Search for data
Consumers
Refrence data

DDW

DDW

- Все данные компании разделены на домены.
- DDW – главный за домен


Source-specific data model & protocol
Domain-specific internal data model & protocol
Enterprise canonical data model & protocol
Injection
Normalization
Processing
Delivery
Injection pattern
Source (1A, etc)
Injector (EIP:Gateway)
Raw data (EDB:Topic)
Normalizer (?)
Normalized data (+RAW?) (EDB:Topic)
Stream Processor (by event, by schedule) ?
Processed data event (EDB:Topic)
Streaming pattern
Stream Provider ?
XYZ (EDB:Topic)
Raw data archiver (EIP:KC)
Corrupted data (EDB:Topic)
Data mart Processor (By event, by schedule) ?
Data mart ?
Data mart pattern
API Provider ?
Raw data archive (EDL)
Normalized Data store (EPaaS:KVaaS)
Indexer (LogStash)
Index (ES)
Generic API Provider ?
XYZ (EIP:FSB)
Operation
Logger (EOS:xM)
Ops
DQ Manager
Alerting (EOS:Alerts)
Processing metrics & logs (EOS:xM)
Dashboard (EOS:xM)
Enterprise platforms
(EIP, EPaaS, EDP, EOS)

S7 Airlines

S7 Airlines

Это наш борт

S7 Airlines

Это наш борт

Он совершает перелет

S7 Airlines

Это наш борт

Он совершает перелет

Сколько здесь доменов?

S7 Airlines

Это наш борт

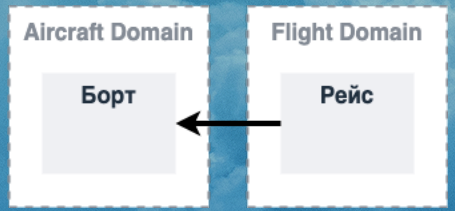
Flight Domain

Рейс

Один?

S7 Airlines

Это наш борт



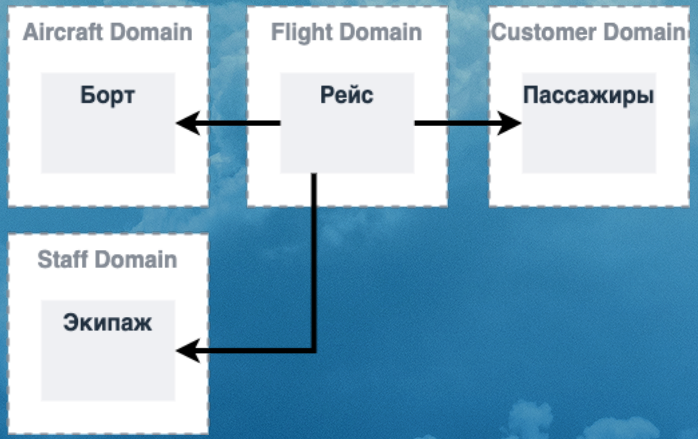
Парочка?

S7 Airlines

Это наш борт

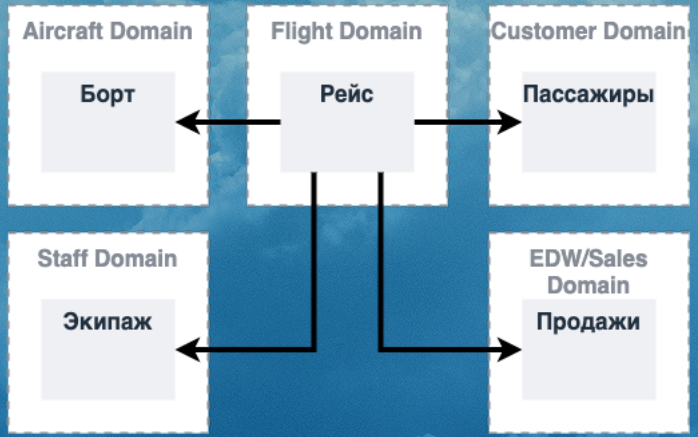
S7 Airlines

Это наш борт



S7 Airlines

Это наш борт



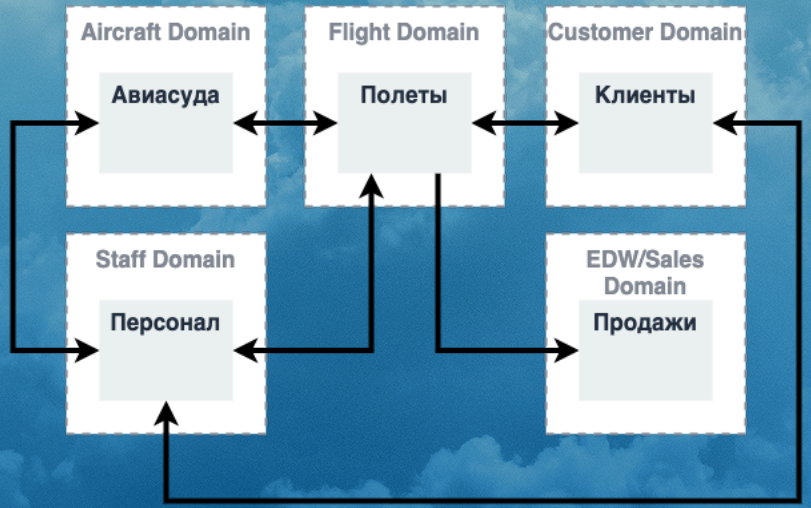
Все!

А кто главный?

Где границы?

S7 Airlines

Это наш борт



Становится только хуже

S7 Airlines

Надо разруливать!

DDW

1. Объект хранится в том DDW, к предметной области которого он относится

DDW

2. Атрибуты объекта находятся в том хранилище, частью которого является этот объект.

DDW

3. При определении принадлежности объекта к тому или иному DDW прежде всего анализируется его бизнес-суть, а не реализация работы с ним конкретной в системе/продукте.

DDW

4. Связь между объектами должна храниться в том хранилище, при создании объекта которого, появляется связь.

DDW

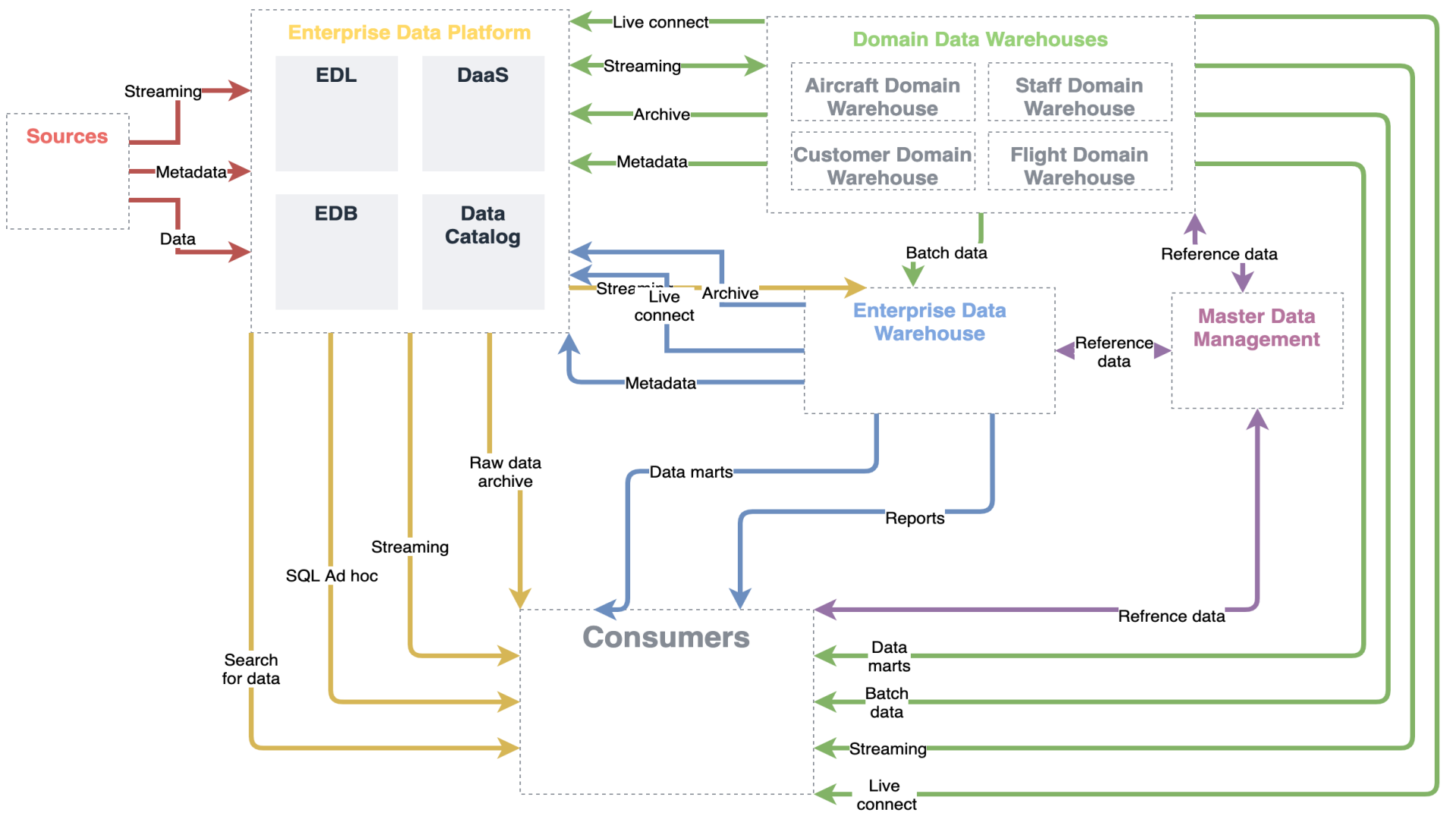
5. Для определения ID сущности DDW с которой необходимо хранить связь, обращение идет по уникальному набору данных системы, которая является первоисточником связи.

DDW

Полученный опыт

- Data Owner, Data Steward, Data Officer – это действительно важно
- Границы размыты – не рубите с плеча
- Не паникуйте!

И что же в итоге?


Enterprise Data Platform
EDL
DaaS
EDB
Data Catalog
Sources
Streaming
Metadata
Data
Live connect
Streaming
Archive
Metadata
Domain Data Warehouses
Aircraft Domain Warehouse
Staff Domain Warehouse
Customer Domain Warehouse
Flight Domain Warehouse
Batch data
Reference data
Live connect
Archive
Enterprise Data Warehouse
Master Data Management
Reference data
Metadata
Raw data archive
Data marts
Reports
Streaming
SQL Ad hoc
Refrence data
Consumers
Search for data
Data marts
Batch data
Streaming
Live connect

Часть третья.
Что у нас получилось (и что нет).

Истории со счастливым концом

Истории со счастливым концом

Как мы добро бизнес-пользователям наносили

S7 Airlines

Истории со счастливым концом

Как один сервис рекомендаций перешел от нескончаемого поиска к бесконечному потреблению

Истории неудач

Истории неудач

Сказ о том, как мы даталейку проектировал

Классные штуки

Классные штуки

Как мы Spark и OpenShift женили

S7 Airlines

Классные штуки

О декларативных дагах в Airflow

Классные штуки

Как повернуть пайплайны к пользователям лицом

Заключение

Любовь и благодарности

- С7. Лучшая авиакомпания!
- С7 Техлаб. Лучшие коллеги!
- Команды EDP, DDW, DWH, всех наших *aaS. Вы котики!
- Антон Пилипенко. Спасибо за помощь и комментарии!
- Паша Финкельштейн и Даша Буланова. Спасибо за советы по докладу!
- Слушатели. Вы это выдержали!

Советы вдохновившимся

- Не торопитесь. Платформа от вас не убежит.
- Решайте не таски в бэклоге, а задачи бизнеса.
- Смотрите на пользователей. Они подскажут путь.

Алло! Мы ищем таланты!

tg: @katie_vol

tg: @flymemary

S7 Airlines

S7 Airlines

А на этом всё!