ИНГОССТРАХ

DataMesh как перспективная стратегия организации корпоративной платформы данных

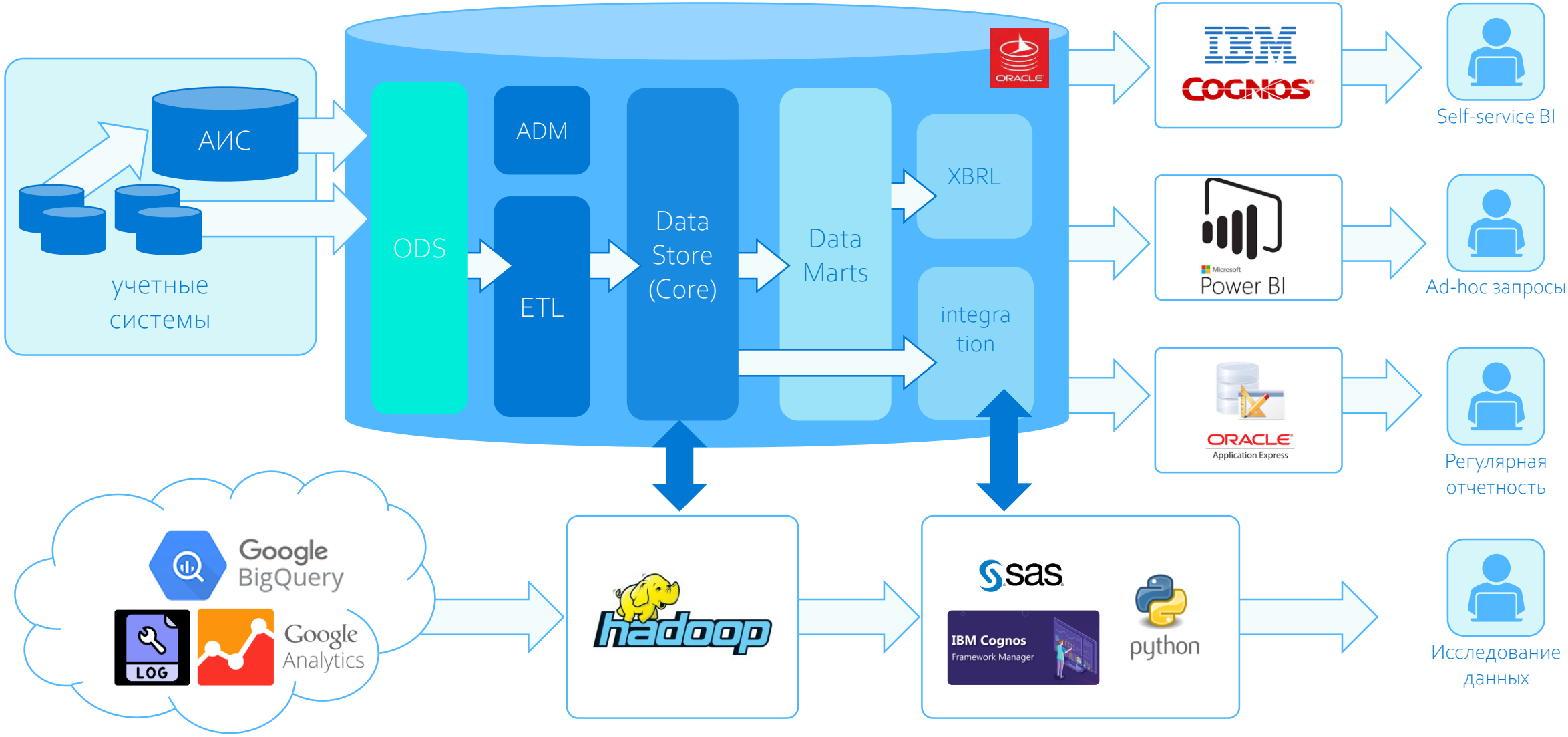
Немного истории. Этапы развития подходов к аналитике данных

ИНГОССТРАХ

Поколение	Технология	Достижения	Проблемы
0	Прямые запросы к операционным источникам данных (OLTP)	Примитивная аналитика	Непрофильная нагрузка, низкая эффективность
1	Хранилища данных (DWH)	Модель данных. Историчность. Специализация	Только структурированные данные
2	Озера данных (DataLake)	Объемы хранения Неструктурированные данные	Болото данных
3	? Есть кандидаты (LakeHouse, DataFabric, RDS, DataMesh)	?	

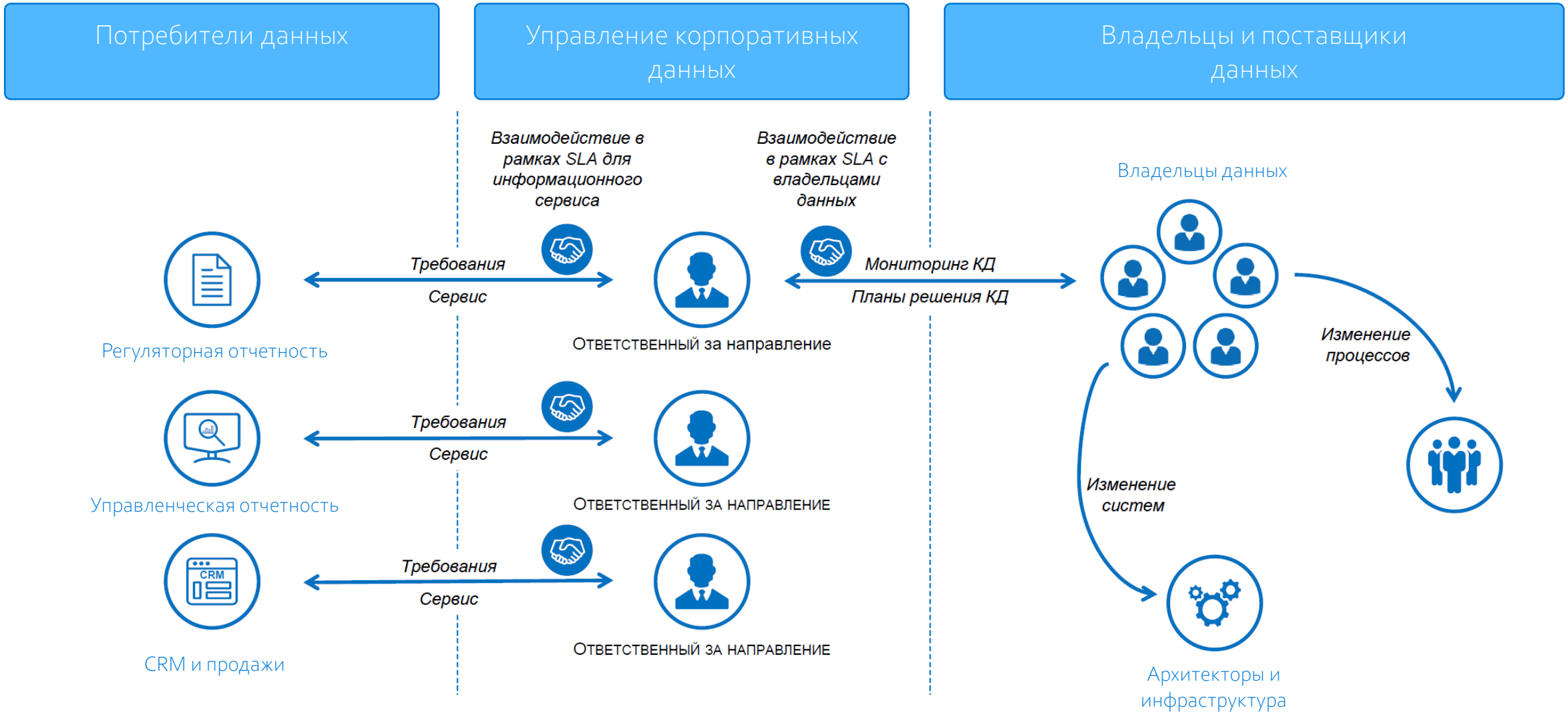
Исходный вариант в Ингосстрах перед трансформацией — платформа

ИНГОССТРАХ



Принципы организации работы с корпоративными данными

ИНГОССТРАХ



Проблемы традиционного подхода к работе с данными

ИНГОССТРАХ

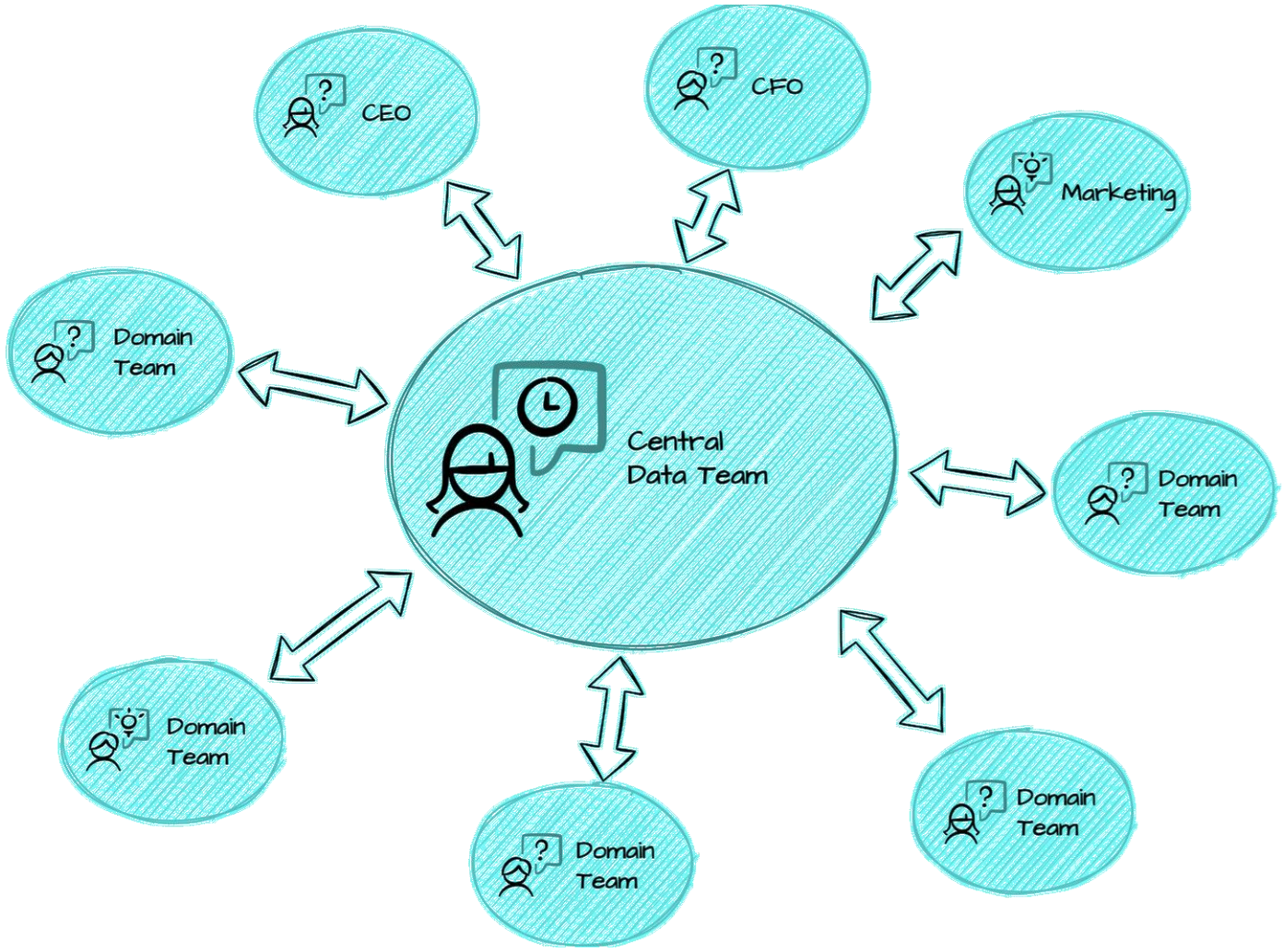
Традиционный подход: всеми процессами обработки данных в компании занимается специализированная команда (**DataTeam)**

Плюсы:

- единый центр компетенций и экспертизы принцип «единого окна»
- единая архитектура данных и подходы к обработке
- как правило, более высокая скорость обработки и поставки данных

Проблемы:

- DataTeam может стать узким местом, возникают конфликты за ее ресурс между доменными командами.
- С ростом объемов данных растет риск появления «болота данных»
- Нет «биллинга» для доменных команд (ресурсы «бесплатные»)



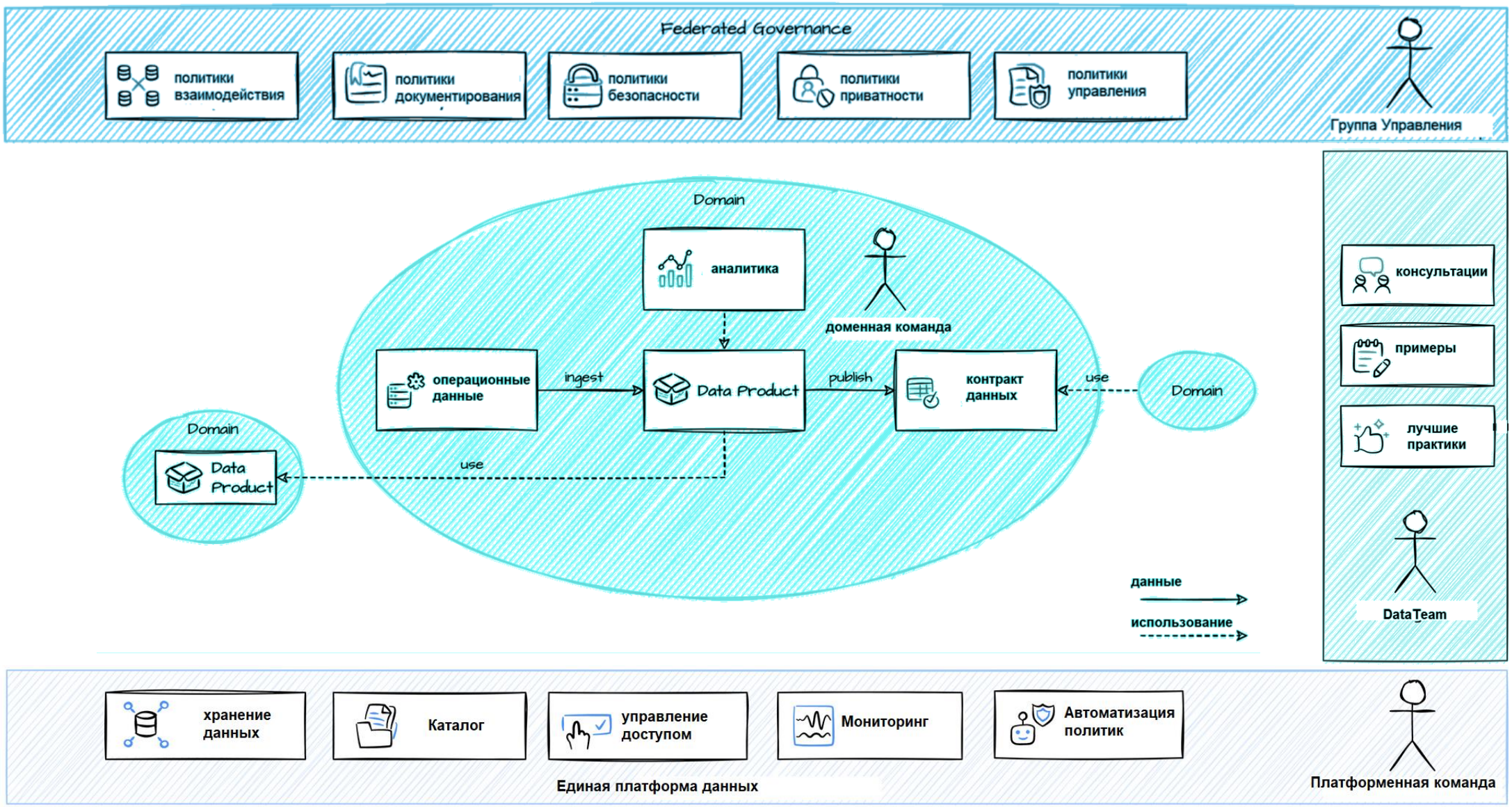
Принципы реорганизации для эффективной работы

ИНГОССТРАХ

	владение доменом	данные как продукт	self-service платформа	федеративное управление
Доменная (DDD) организация	Контекст данных привязан к домену	Продуктовое мышление	Независимость от домена	Связь контекстов
Люди	Доменные команды	DataProduct	DataTeam	«Гильдия»
Технологии	Операционные данные и аналитика вместе	Стандартные интерфейсы	Платформа данных	Управление данными

Архитектура DataMesh

ИНГОССТРАХ



DataProduct

ИНГОССТРАХ

Определение	DataProduct – это автономный технический компонент, содержащий все данные, описания и, возможно, средства обработки достаточные для полноценной работы. Можно рассматривать это как микросервис, но в в области обработки данных.
Формы предоставления	DataProduct может поставляться в различных формах, например: • Витрина данных в реляционной СУБД • Таблица BigTable • Файл Parquet в виде блока S3 хранилища • Дашборд в BI инструменте, например, Power BI или Cognos • Модель машинного обучения в формате ONXX • Готовый отчет в «печатаемом» формате типа PDF Во всех вариантах DataProduct дает пользователю информацию о: • происхождении и методике получения данных • показателях качества и чистоты данных • частоте обновления и SLA поставки
Ответственность	Продуктом владеют доменные команды. Команда отвечает за все операции на протяжении всего жизненного цикла продукта. Команда должна обеспечить регулярный мониторинг, проверку актуальности источников и соблюдение показателей качества данных. Часть действий может передаваться на обслуживание DataTeam, но ответственность за методологию и качество несет доменная команда.

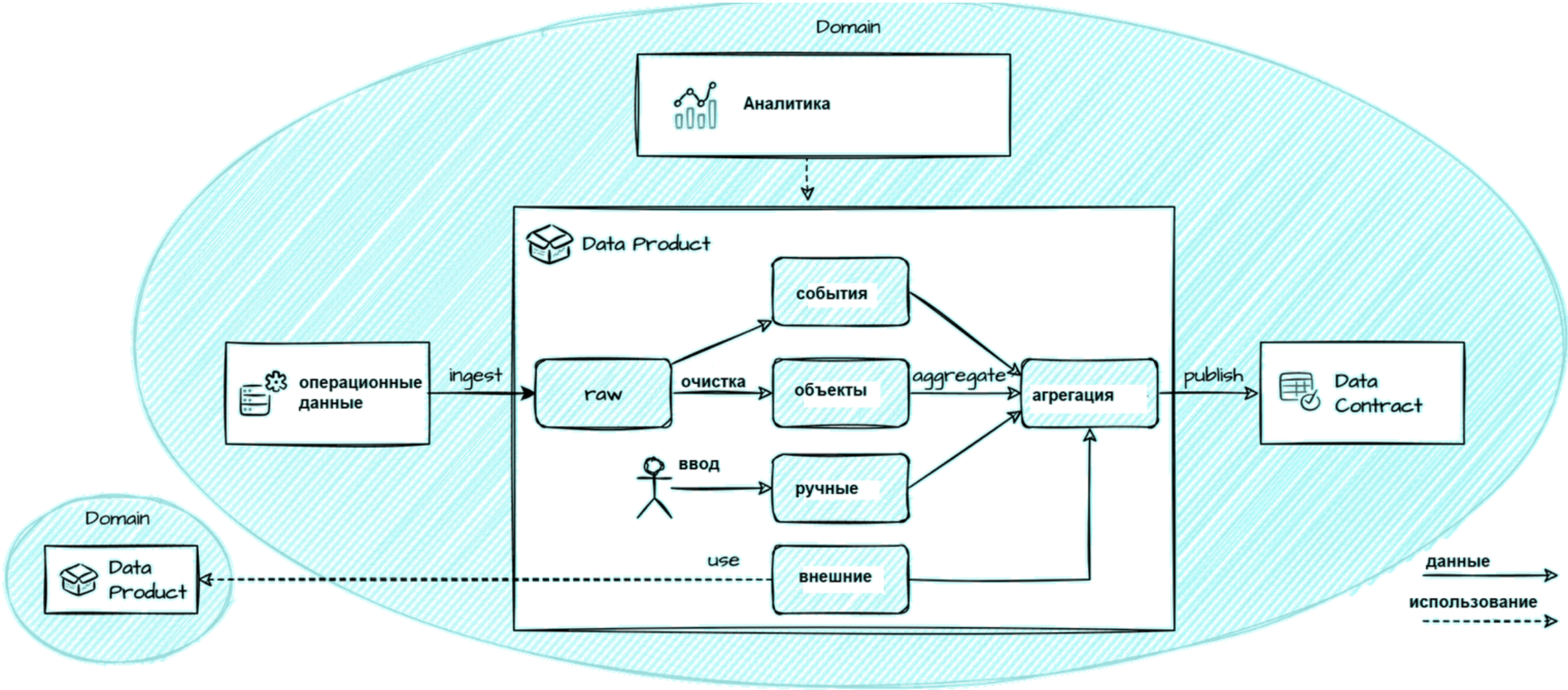
Федеративное управление

ИНГОССТРАХ

Стратегические цели	Федеративное управление предполагает формирование кросс-командной «гильдии» специалистов по данным, которые обеспечивают согласованное управление данными в рамках всей компании. DataTeam отвечает за выработку глобальных политик в рамках всей компании, а представители в доменных командах (можно называть это ролью DataOfficer) обеспечивают функционирование команд согласованное с политиками организации
Политики	Политики – это «правила игры» в DataMesh Политики обеспечивают возможность разным командам согласованно пользоваться продуктами данных в рамках всей организации Политики определяют: Форматы данных • Необходимую степень документирования • Правила безопасности • Ролевую модель доступа • Жизненный цикл продуктов
Пример	Global Policy1: Data Format: CSV; standard: RFCxxx; delimeter: comma Global Policy2: Location: S3 bucket Global Policy3: Data Discovery описание должно включать: • Наименование доменной команды, наименование продукта и владельца продукта • Частоту обновления и дату последнего обновления • Точку входа (адрес по которому расположены данные) • Модель данных Global Policy4: Acces Control: Role-based ACL

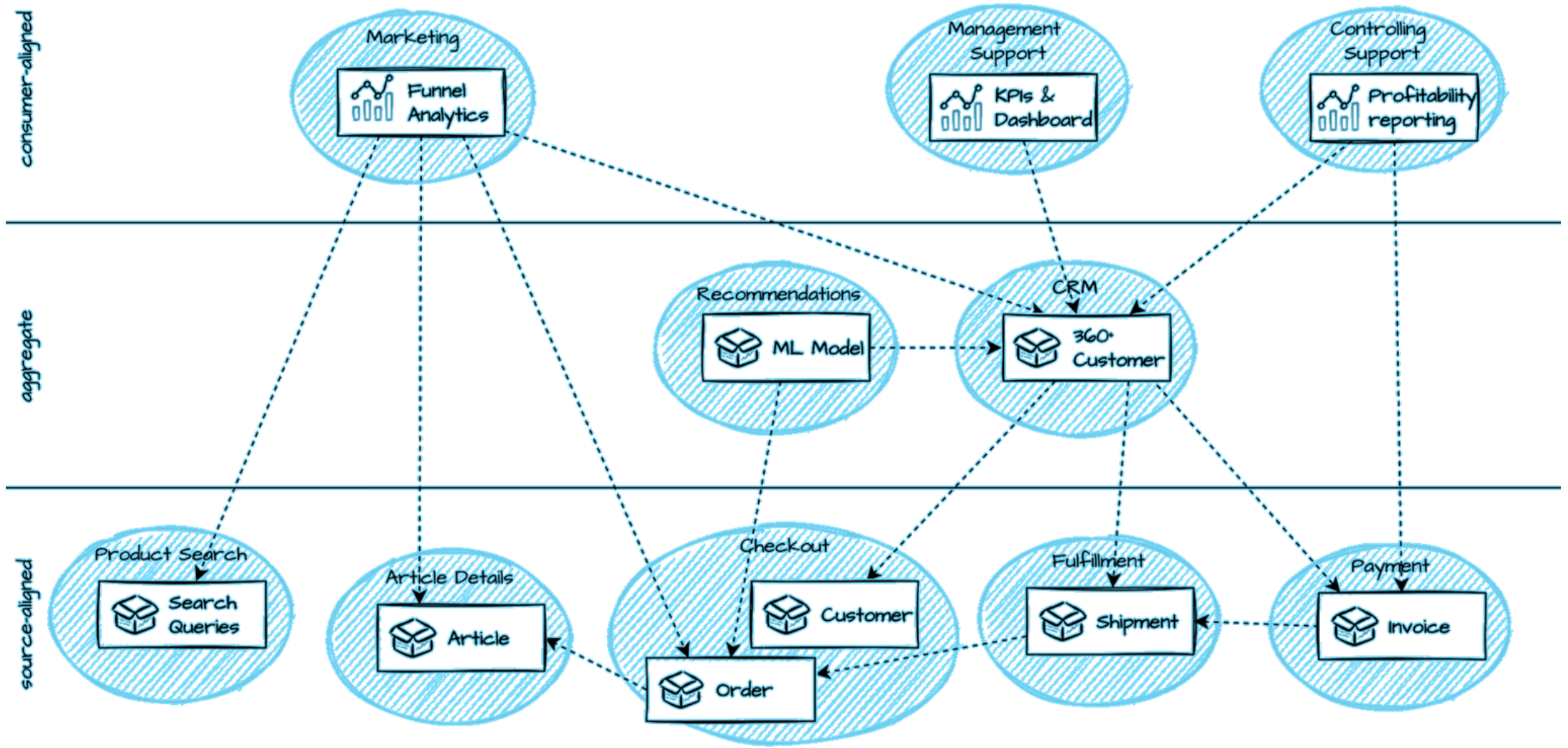
Трансформации при переходе к DataMesh

ИНГОССТРАХ



Верхнеуровневая диаграмма сетки продуктов в DataMesh

ИНГОССТРАХ



Что получилось сделать

- Выделены две команды (ML и Фин. блок) работающие на собственных мощностях
- Команды публикуют свои результаты в общем пространстве
- Отработан механизм выделения ресурсов новым командам
- «Биллинг» ресурсов выделяемых командам

Что в планах и проблемы

- Нет общей платформы (выделение ресурсов командам пока немасштабируемый процесс)
- Проблемы скорости обмена данными в децентрализованной среде. Принципиально не хотим централизованное «мега-озеро», в поиске решения
- Ищем подходящий инструмент для каталога данных
- Ищем более удобный инструмент централизованного управления доступом к данным