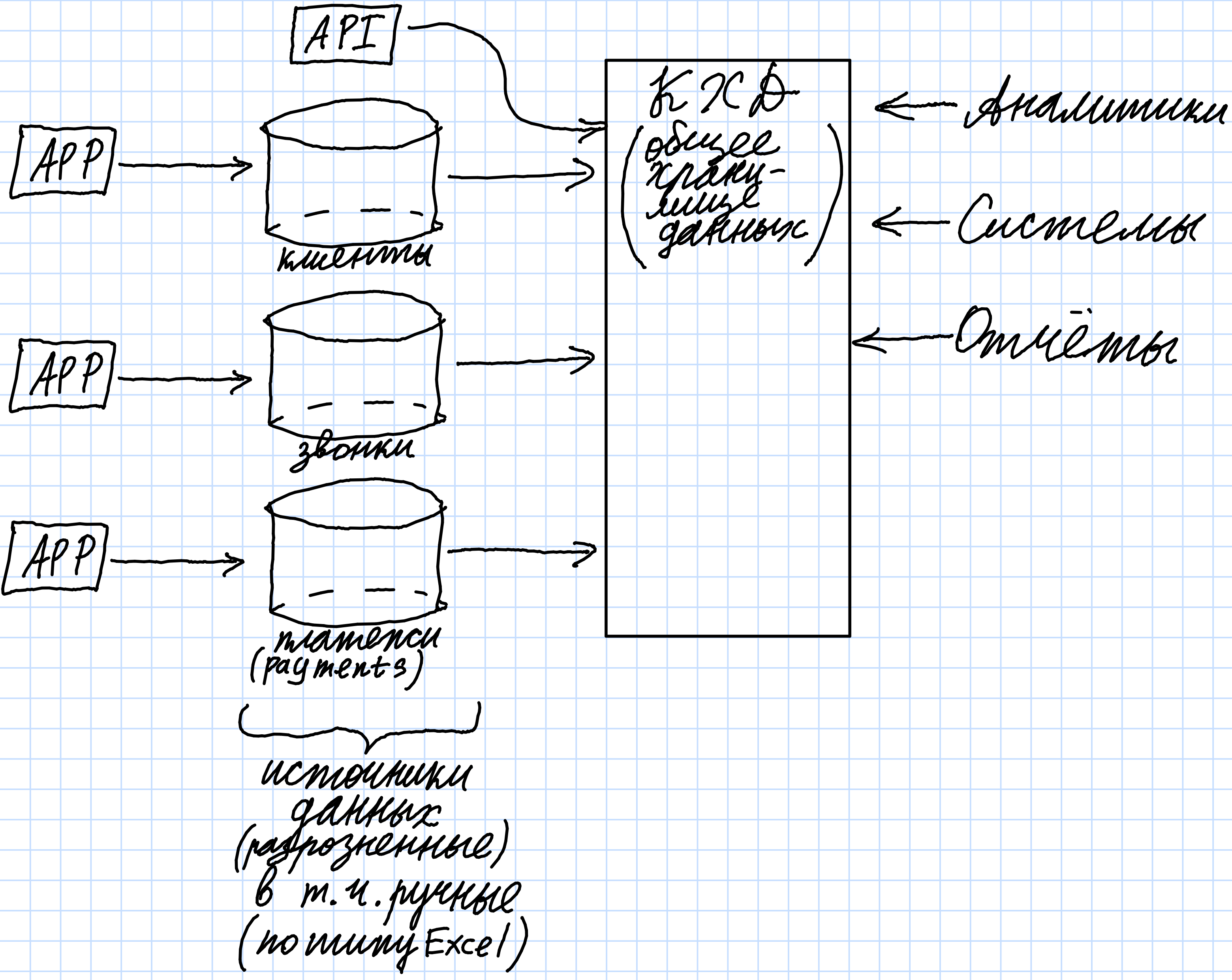
Проектирование хранилищ.

КХД — (корпоративное) хранилище данных
или
DWH — (Data) WareHouses;



Любая передача внутри хранилищ данных — ETL (Extract, Transform, Load);

Способы хранения данных:

Пространственная модель хранения данных.

оперирует 2 критериями:

— Факт — строка в таблице, кот. явл. точечным соб-ем (нет продолжительности)

— Измерение

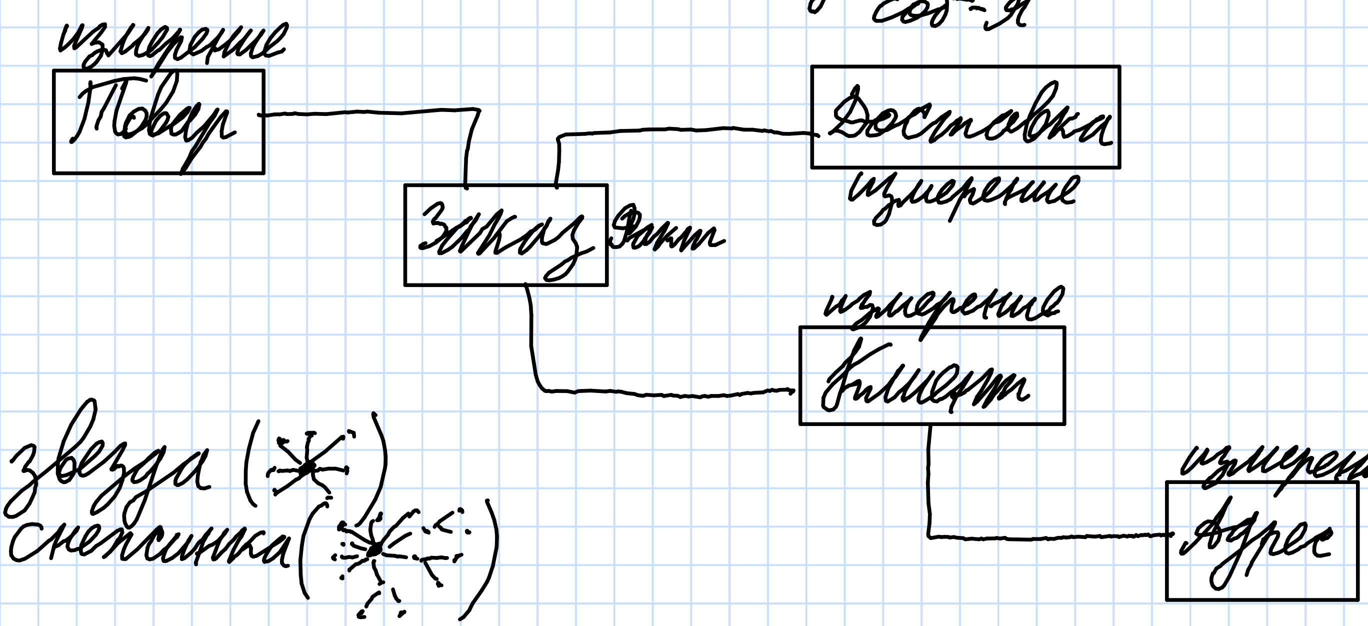
Все события в системе делятся на эти 2 категории;

Факт — транзакционное соб-е.

↳ у него есть много измерений

Измерение — мера соб-я (= справочник)

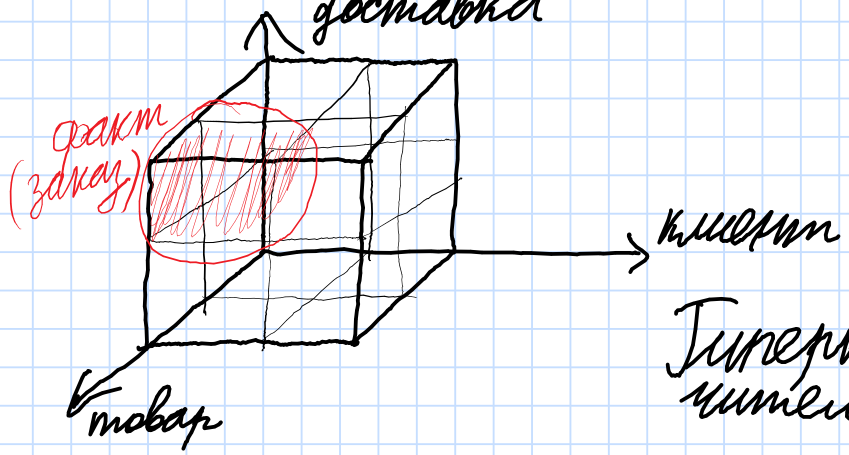
могут иметь под собой точечные соб-я



звезда ( )
снежинка ( )

Измерение — сборище звёздочек или снежинок;

OLAP — куб (гиперкуб)



Excel умеет визуализировать гиперкубы (отчёты в разных разрезах)

Гиперкуб построен исключительно на звёздочках;

Исторически сложились 2 модели хранилищ данных:

1 Модель хранилища данных по Инмону:

Данные должны быть структурированы;

КХД – единая (·) „правды";

Данные должны быть

– чистыми;

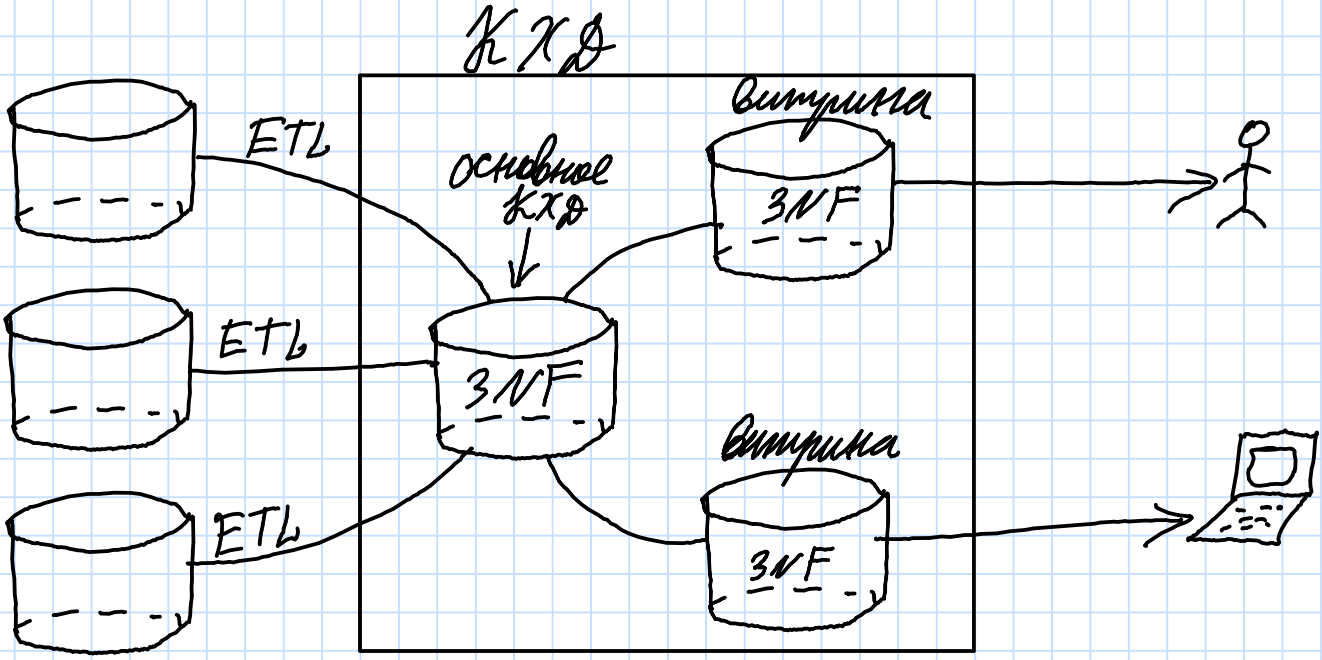
– достоверными;

– непротиворечивыми;

– подчиняться 3NF;

3NF должна лежать в основе всех хранилищ

3NF:

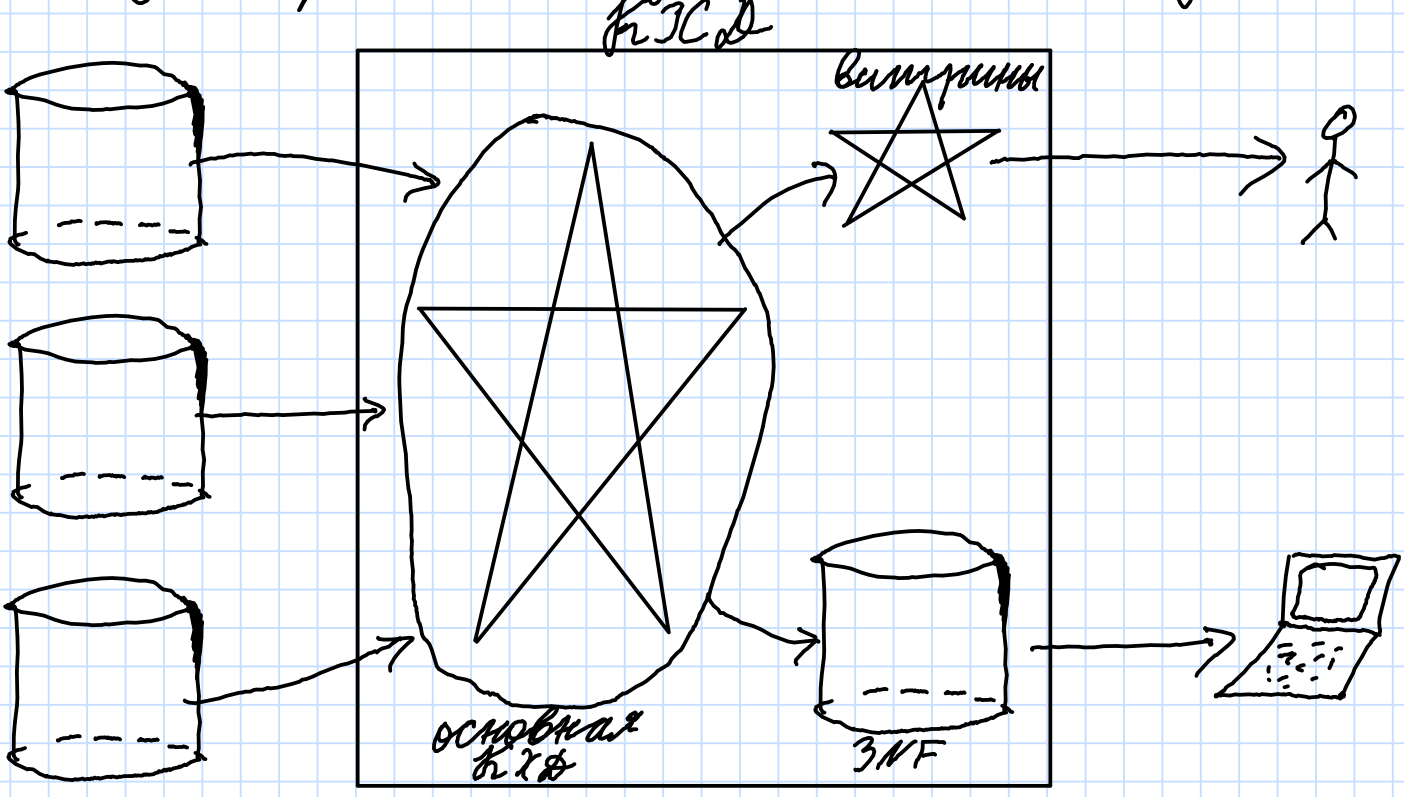


⊕ Чистота данных (все данные красивые);
⊕ V хранилища (данные неизбыточные);
⊕ Гибкость (относительно легко добавлять данные);
⊖ много join'ов (очень много времени);

Колоночные MPP системы очень тяжело джойнятся (по типу Klickhouse)

⊖ время начальной разработки (2-3 мес. вместо 8 часов);

② Модель хранилища данных по Кимбаллу :



Здесь чаще всего не 3NF, а декомпозиция

→ Она предполагает предварительную нормализацию ⟹ время разработки не сильно уменьшится;

(+) время запросов

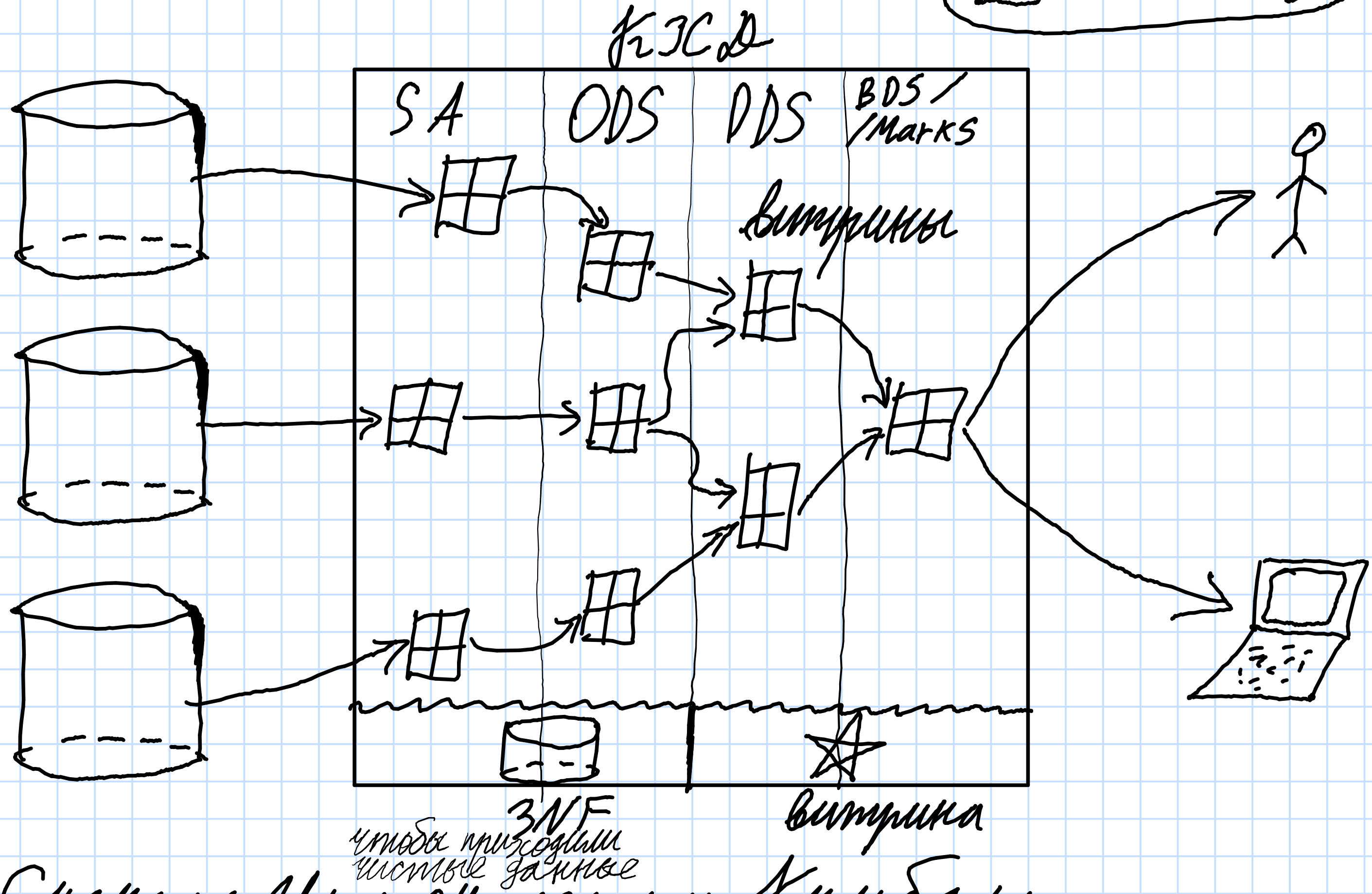
(+) звёздочка выглядит приятнее (с т.з. пользователя), чем 3NF;

(−) Аномалии (данные не очень чистые);

(−) V хранилища (избыточность данных);

Обе методологии не используются ⟹
⟹ <u>гибридный подход</u>;

⊞ – таблица



Сначала Инмон, потом Кимбалл;

Слои:

1) Staging - предподготовленные данные;

2) ODS – операционные данные (operation data stage);

SA - сырые данные (JSON, инкременты и т.д.);

ODS – по сути те же источники, только у нас в хранилище;

→ нужен для выч-я агрегаций;

DDS – Detail Data Stage; BDS – Buisness Data Stage;

Data Vault

3 сущности:

— Хаб — имеет отн-е к факту, в кот. чаще всего помещается набор ключей;
Это ключ: натур. (суррогат.) + источник + время появления, технические поля

(Напоминает усложнённый факт)

— Сателлит — ключ (ссылка на хаб) + атрибуты + ... + тех. поля
Сателлит и хабы — one to many;

(Напоминает измерение)

— Ссылка — связь между хабами + тех. поля
Это всегда связь many to many;

При построении бизнес-слоя (BDS) будем использовать эти 3 сущности;

Якорная модель хранения

4 объекта: (факт свершения соб-я)

— Якорь — суррогат. ключ + технические поля;

— атрибуты — сур. ключ. якоря + поля + тех. поля;

— связь — только many to many;

— узел: атрибуты, код (справочник);

В Data Vault упор на натур. ключи, а Якорная модель — на суррогатные;

Натуральные ключи VS суррогатные ключи :

— - - - очень долгий и непонятный разгон - - -

— сур. ключ должен быть создан при помощи Хеш-функции ;