

Построение системы анализа для оптовой сети

Александр Дольник

alexander.dolnik@gmail.com

Развитие практических навыков по созданию ETL процессов и проектированию хранилищ

Обзор

Виды структур хранилищ

- Темпоральные базы данных (ловушки изменений);
- Пространственные базы данных (разноформатность)
- Медленно меняющиеся измерения (3 типа)

Знакомство с кейсом (формулировка задачи)

- Уровень 0. Проектирование хранилища для анализа данных
- Уровень 1. Создание процессов наполнения БД (извлечение данных с почты)
- Уровень 2. Генерация отчетов по почте
- Уровень 3. Оптимизация проблем анализа

Медленно меняющиеся измерения

- На добавление - запись новая и её необходимо добавить в таблицу.
- На изменение – запись существует в таблице, но в каких-то полях изменились содержимое.
- На удаление - запись существует в таблице, но теперь её необходимо удалить из неё.
- Существует 3 типа (из статьи Kimball): 1, 2, 3

SCD Type 1

Статус записи	Действие
На добавление	Записи присваивается следующий по порядку уникальный идентификатор. Запись добавляется в таблицу.
На изменение	Запись изменяется.
На удаление	Никаких действий над записью не производится. Удалять запись из таблицы нельзя, потому что к ней могут быть «привязаны» фактические данные.

SCD Type 1. Пример

ID	Name
1	ИП Иванов

ID	Name
1	ООО "Иванов и Ко."

SCD Type 2

Событие	Действие
На добавление	<p>Полю ID присваивается следующий по порядку уникальный идентификатор.</p> <p>Полю EFCT_DT присваивается текущая дата (SYSDATE).</p> <p>Полю END_DT присваивается дата 01.01.2999.</p> <p>Полю IS_ACT_IND присваивается 1.</p> <p>Полю IS_DEL_IND присваивается 0.</p> <p>Запись добавляется в таблицу.</p>
На изменение	<p>Полю END_DT изменившейся записи присваивается текущая дата (SYSDATE).</p> <p>Полю IS_ACT_IND изменившейся записи присваивается 0.</p> <p>Добавляется новая запись в таблицу, у которой:</p> <p>Полю ID присваивается такой же идентификатор, как и у измененной записи.</p> <p>Полю EFCT_DT присваивается текущая дата (SYSDATE).</p> <p>Полю END_DT присваивается дата 01.01.2999.</p> <p>Полю IS_ACT_IND присваивается 1.</p> <p>Полю IS_DEL_IND присваивается 0.</p>
На удаление	<p>Полю END_DT присваивается текущая дата (SYSDATE).</p> <p>Полю IS_ACT_IND присваивается 0.</p> <p>Полю IS_DEL_IND присваивается 1.</p>

SCD Type 2

ID	NAME	EFCT_DT	END_DT	IS_ACT_IND	IS_DEL_IND
1	ИП Иванов	01.10.10	10.11.10	0	0
1	ООО "Иванов и Ко."	10.11.10	01.01.99	1	0

SCD Type 3

Статус записи	Действие
На добавление	<p>Записи присваивается следующий по порядку уникальный идентификатор (в поле ID).</p> <p>В поле NAME присваивается загружаемое значение.</p> <p>В поле NAME_OLD присваивается значение по умолчанию, например «NA».</p> <p>Полю NAME_UPD_DT присваивается текущая дата (SYSDATE) или дата по умолчанию, например 01.01.1900.</p> <p>Запись добавляется в таблицу.</p>
На изменение	<p>В поле NAME_OLD присваивается значение из поля NAME.</p> <p>В поле NAME присваивается загружаемое значение.</p>
На удаление	<p>Никаких действий над записью не производится.</p>

SCD Type 3

ID	NAME	NAME_OLD	NAME_UPD_DT
1	ООО "Иванов и Ко."	ИП Иванов	10.11.10

Знакомство с кейсом

Условие задачи: организовать инфраструктуру для хранения и анализа данных оптовой сети...

- Пусть есть сеть магазинов занимающихся реализацией товара
- Каждый магазин еженедельно направляет отчет в excel о проданном товаре на почту поставщика товара

УРОВЕНЬ 0 - Проектирование

На данном уровне происходит проектирование системы и выявление её основных требований

Требование аналитика

Точка реализации	Год	Месяц	Число	Продажи		Возврат	
				Кол-во	Стоимость	Кол-во	Стоимость

Форма типового отчета

- Форма типового отчета представлена [по ссылке](#)

УРОВЕНЬ 1

Создание процесса по автоматической обработке данных с почтового сервера

Этап 1. Скачивание в DSA

Задача: Скачать данные из автоматически обрабатываемой почты в локальное хранилище (на диск)

Решение на видео:

[C:\kettle_DSA\Video\1_Step_FileDownload_Final.avi](#)

Этап 2. Фильтрация результатов

- Демо — Move files

Этап 3. Процесс загрузки данных

Данные будем грузить в базу MySQL и генерировать отчеты

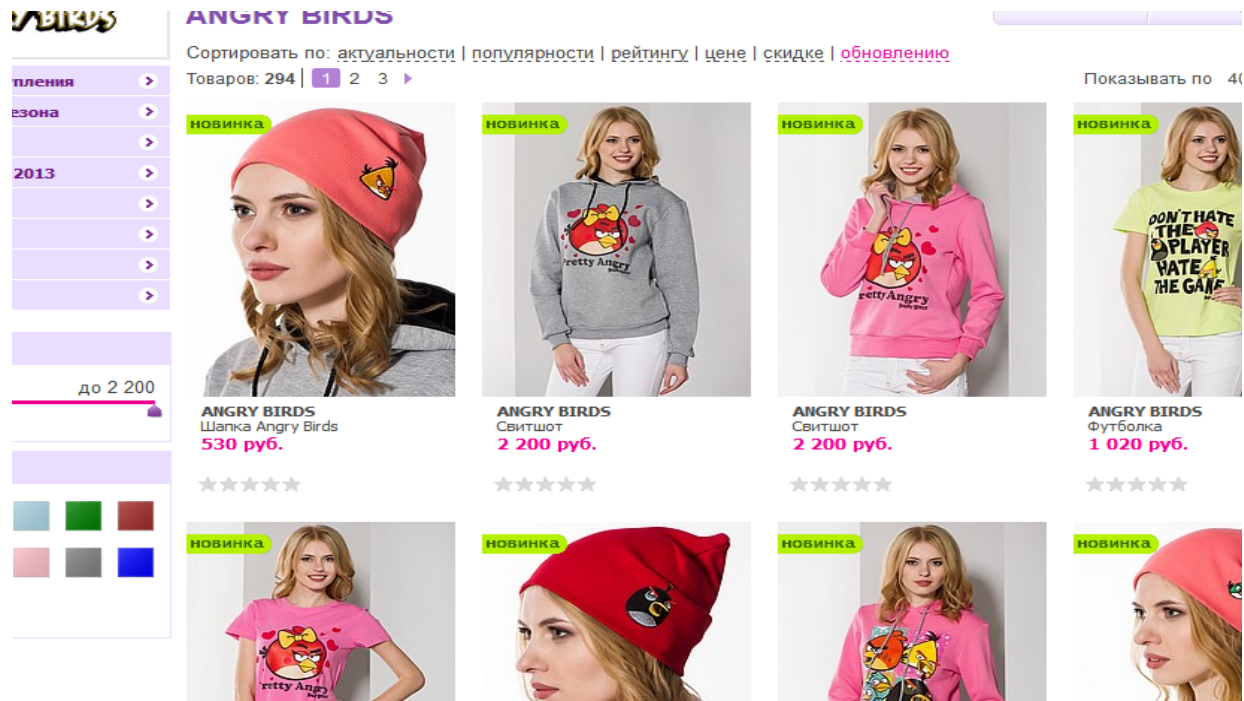
УРОВЕНЬ 2. Генерация отчетов

Необходимо сделать то, что требует аналитик и представить в красивом виде.

УРОВЕНЬ 3. Проблемы анализа

- Проанализируем полученное решение и попробуем предотвратить проблемы, которые могут быть связаны с анализом данных







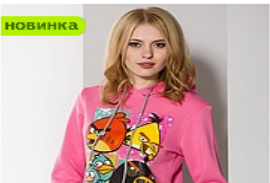

Конец или новый кейс? :)



The screenshot shows a product page for 'ANGRY BIRDS' merchandise. The page features a navigation sidebar on the left with filters for 'Плечения', 'Сезона', and '2013'. Below the filters is a price range slider set to 'до 2 200' and a color selection grid. The main content area displays a grid of eight items, each with a 'новинка' (new) label, a product image, a title, a price, and a star rating.

ANGRY BIRDS
Сортировать по: актуальности | популярности | рейтингу | цене | скидке | обновлению
Товаров: 294 | 1 2 3 ▶

Показывать по 40

- новинка**

ANGRY BIRDS
Шапка Angry Birds
530 руб.
★★★★★
- новинка**

ANGRY BIRDS
Свитшот
2 200 руб.
★★★★★
- новинка**

ANGRY BIRDS
Свитшот
2 200 руб.
★★★★★
- новинка**

ANGRY BIRDS
Футболка
1 020 руб.
★★★★★
- новинка**

ANGRY BIRDS
Футболка
530 руб.
★★★★★
- новинка**

ANGRY BIRDS
Шапка Angry Birds
530 руб.
★★★★★
- новинка**

ANGRY BIRDS
Свитшот
2 200 руб.
★★★★★
- новинка**

ANGRY BIRDS
Шапка Angry Birds
530 руб.
★★★★★