

Анализ данных

Лекция 1

Технологии анализа данных. Хранилище данных

Гедранович Ольга Брониславовна,
старший преподаватель кафедры ИТ, МИУ
volha.b.k@gmail.com

31.08.2017

2

Вопросы лекции

- Системы поддержки принятия решений.
- Базы данных.
- Организация хранилища данных.
- Очистка данных.

31.08.2017 3

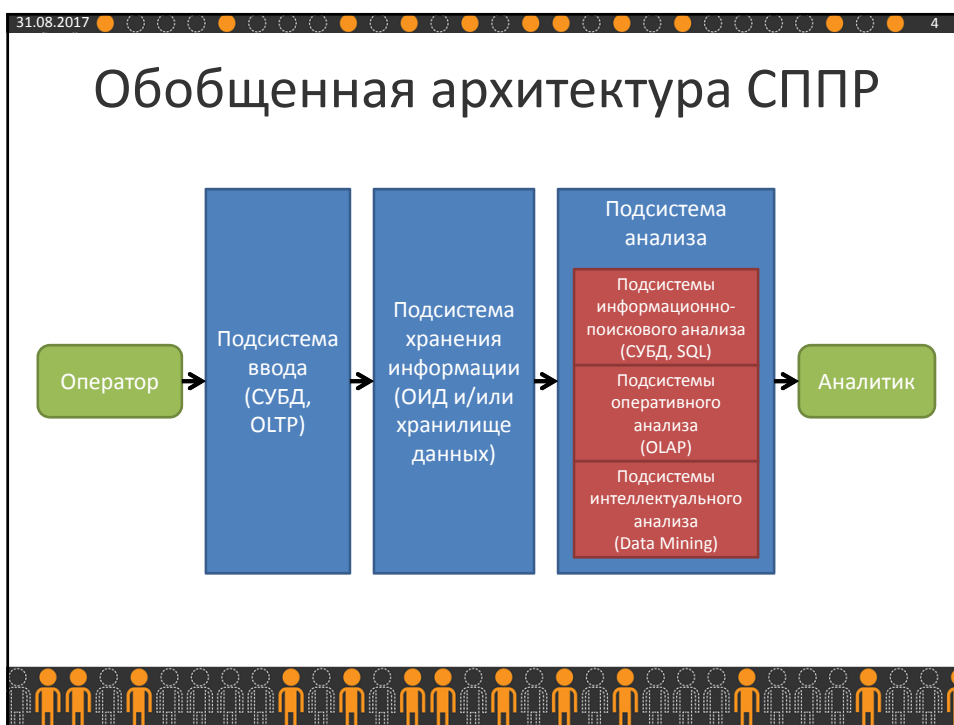
СППР

Системы поддержки принятия решений (СППР, DSS, Decision Support System) — это системы, обладающие средствами ввода, хранения и анализа данных, относящихся к определенной предметной области, с целью поиска решений.

Три основные задачи, решаемые в СППР:

- ввод данных;
- хранение данных;
- анализ данных.

31.08.2017




31.08.2017 5

Базы данных

База данных (БД) является моделью некоторой предметной области, состоящей из связанных между собой данных об объектах, их свойствах и характеристиках.

Система управления базами данных (СУБД) — совокупность программных и лингвистических средств общего или специального назначения, обеспечивающих управление созданием и использованием баз данных.

MS Access, FoxPro, Oracle, MS SQL Server, PostgreSQL, MySQL, SQLite, Neo4j.




31.08.2017 6

Реляционные БД

12 правил Кодда для реляционных БД:

0. Основное правило (Foundation Rule).
1. Информационное правило (The Information Rule).
2. Гарантированный доступ к данным (Guaranteed Access Rule).
3. Систематическая поддержка отсутствующих значений (Systematic Treatment of Null Values).
4. Доступ к словарю данных в терминах реляционной модели (Active On-Line Catalog Based on the Relational Model).
5. Полнота подмножества языка (Comprehensive Data Sublanguage Rule).
6. Возможность изменения представлений (View Updating Rule).
7. Наличие высокоуровневых операций управления данными (High-Level Insert, Update, and Delete).
8. Физическая независимость данных (Physical Data Independence).
9. Логическая независимость данных (Logical Data Independence).
10. Независимость контроля целостности (Integrity Independence).
11. Независимость от расположения (Distribution Independence).
12. Согласование языковых уровней (The Nonsubversion Rule).


Эдгар Кодд (Edgar Codd), 1985 год, журнал Computer World



31.08.2017 7

Нормализация БД

- БД имеет 1-ю НФ (нормальную форму), если каждое значение, хранящееся в ней, неразделимо на более примитивные (неразложимость значений);
- БД имеет 2-ю НФ, если она имеет 1-ю НФ, и при этом каждое значение целиком и полностью зависит от ключа (функционально независимые значения);
- БД имеет 3-ю НФ, если она имеет 2-ю НФ, и при этом ни одно из значений не предоставляет никаких сведений о другом значении (взаимно независимые значения) и т. д.




31.08.2017 8

OLTP-системы

Транзакция — последовательность операций над БД, рассматриваемых СУБД как единое целое. Транзакция переводит БД из одного целостного состояния в другое.

OLTP (Online Transaction Processing) — обработка транзакций в реальном времени. Способ организации БД, при котором система работает с небольшими по размерам транзакциями, но идущими большим потоком, и при этом клиенту требуется от системы минимальное время отклика.


OLTP-системы предназначены для ввода, структурированного хранения и обработки информации (операций, документов) в режиме реального времени.



31.08.2017 9

Хранилище данных


Хранилище данных (Data Warehouse) — предметно-ориентированный, интегрированный, неизменяемый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений.

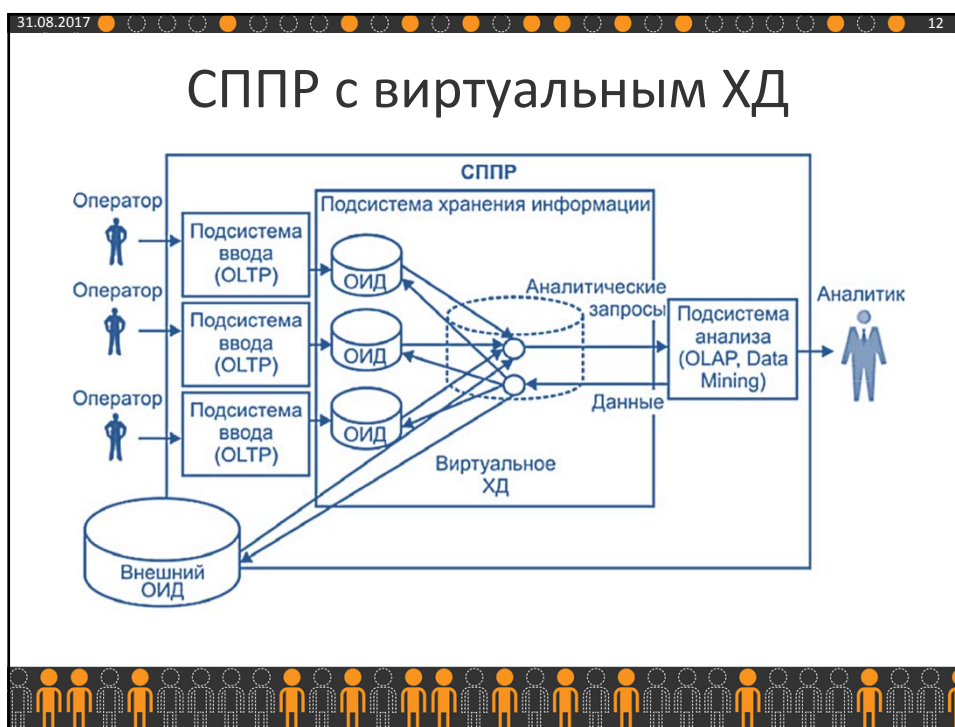
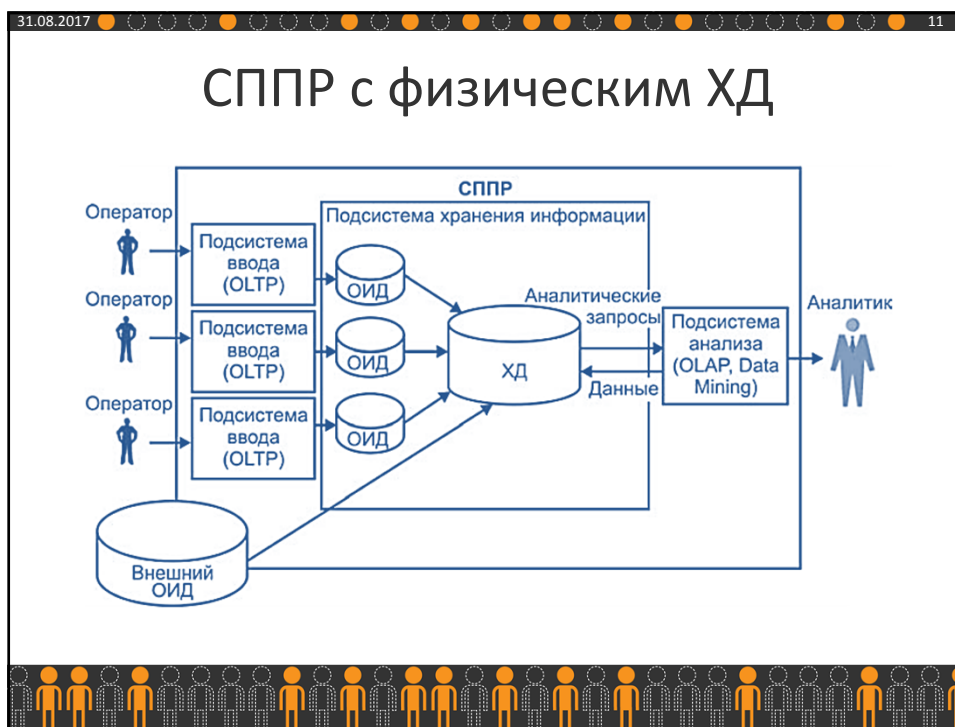


31.08.2017 10

Свойства ХД

- Предметная ориентация
- Интеграция
- Поддержка хронологии
- Неизменяемость






31.08.2017 13

Основные проблемы создания ХД

- необходимость интеграции данных из неоднородных источников в распределенной среде;
- потребность в эффективном хранении и обработке очень больших объемов информации;
- необходимость наличия многоуровневых справочников метаданных;
- повышенные требования к безопасности данных.




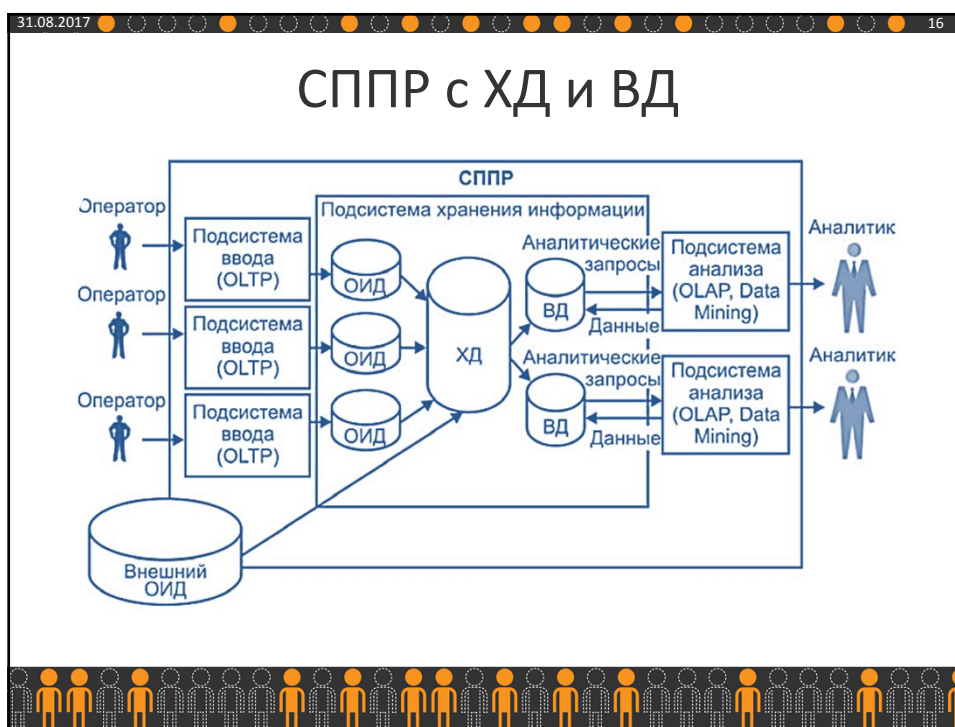
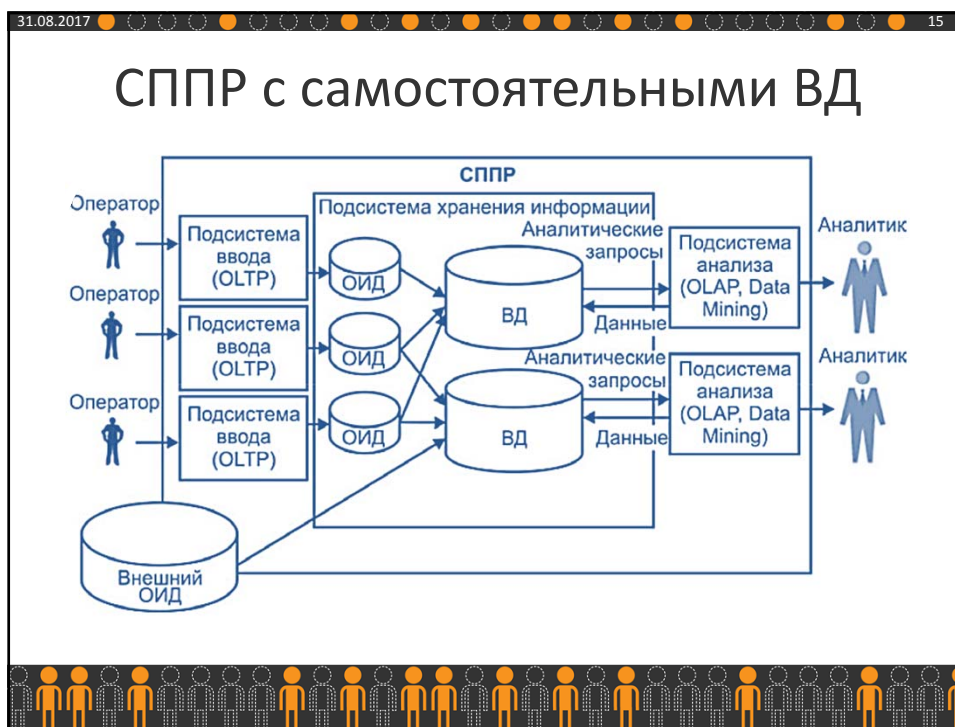
31.08.2017 14

Витрина данных

Витрина данных (Data Mart) — срез хранилища данных, представляющий собой массив тематической, узконаправленной информации, ориентированный, например, на пользователей одной рабочей группы или департамента.

Другие названия: хранилище данных специализированное, киоск данных, рынок данных.





31.08.2017 19

ETL

ETL (Extract, Transform, Load — «извлечение, преобразование, загрузка») — процесс в управлении хранилищами данных, включающий:

- извлечение данных из ОИД;
- трансформацию и очистку данных, для соответствия потребностям бизнес-модели;
- загрузку данных в ХД.

31.08.2017 19

31.08.2017 20

ETL: Transform

Преобразование данных:


- обобщение данных (aggregation);
- перевод значений (value translation);
- создание полей (field derivation);
- очистка данных (cleaning).

31.08.2017 20

31.08.2017 21

Очистка данных

- уровень ячейки таблицы:
 - орфографические ошибки (опечатки);
 - отсутствие данных;
 - фиктивные значения;
 - логически неверные значения;
 - закодированные значения;
 - составные значения;
- уровень записи;
- уровень таблицы БД:
 - нарушение уникальности;
 - отсутствие стандартов (дублирующиеся или противоречивые записи);
- уровень БД;
- уровень множества БД.




31.08.2017 22

Очистка данных

Этапы:

- выявление проблем в данных:
 - профайлинг;
 - Data Mining;
- определение правил очистки данных;
- тестирование правил очистки данных;
- непосредственная очистка данных.



31.08.2017 23

Основная литература

- Анализ данных и процессов : учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. — 3-е изд., перераб. и доп. — СПб : БХВ-Петербург, 2009. — 512 с.
- Пирогов, В.Ю. Информационные системы и базы данных: организация и проектирование : учеб. пособие / В.Ю. Пирогов . — СПб : БХВ-Петербург, 2009. — 528 с.

