

Базы данных

А1. Введение в базы данных



Московский государственный технический университет
имени Н.Э. Баумана

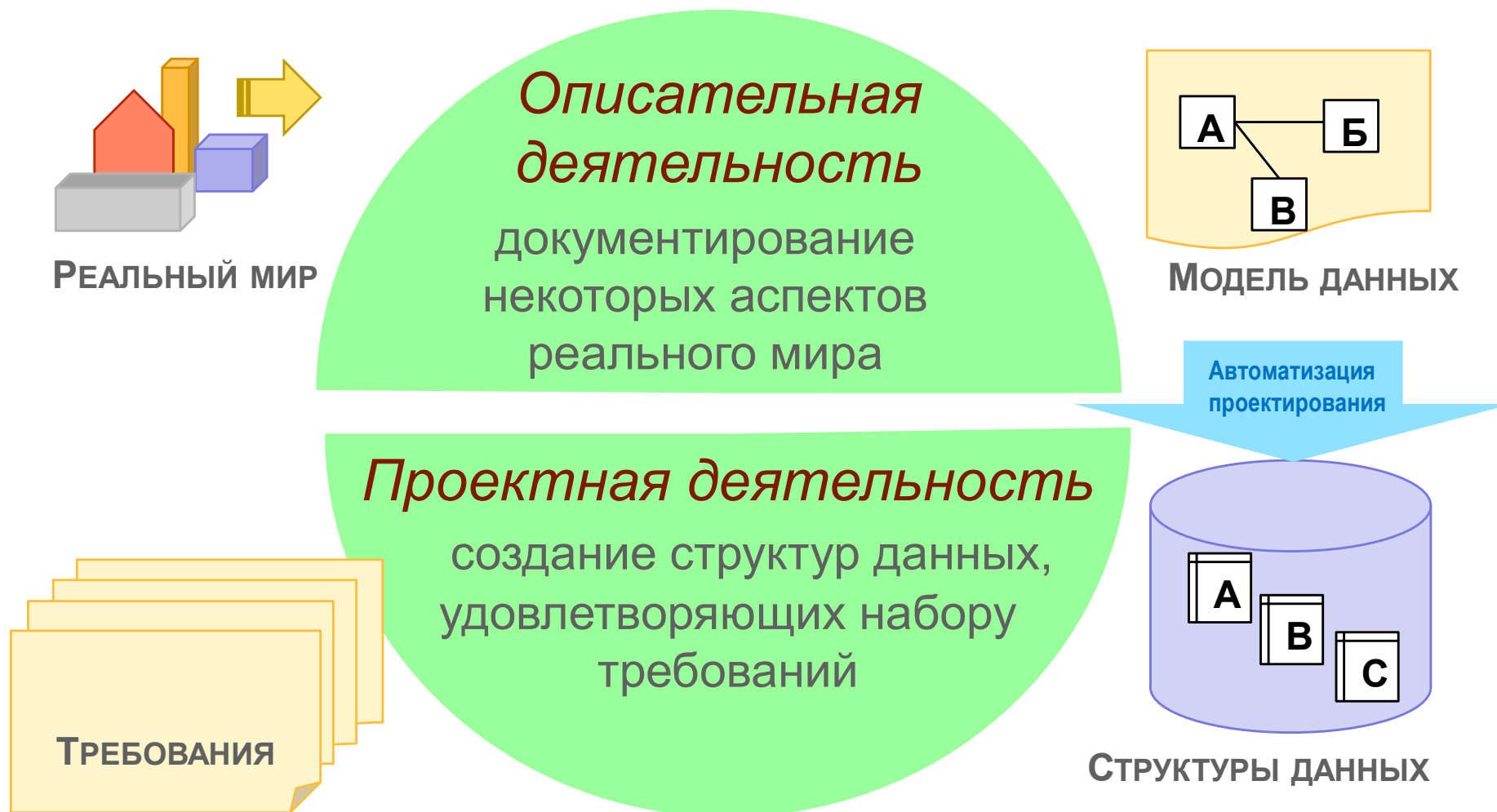
Факультет ИБМ

Фев 2025 года

Москва

Артемьев Валерий Иванович © 2025

От моделей к проектированию данных



Виды информационных систем

Информационная система (ИС) – комплекс программно-технических средств *сбора, хранения, актуализации и обработки информации* для поддержки определённого вида деятельности (*предметной области*).

Классы ИС по типу хранимой и обрабатываемой информации

Документальные системы

(текст, мультимедиа, дескрипторы и полнотекстовой поиск, гипертекст): ИПС и библиотечные системы, электронный документооборот, сайты и порталы

Геоинформационные системы

(топология, электронная карта) – аналитические системы и ситуационные центры с пространственно-географическим компонентом (транспорт, ЖКХ, месторождения, архитектурное наследие и т.п.)

Фактографические системы

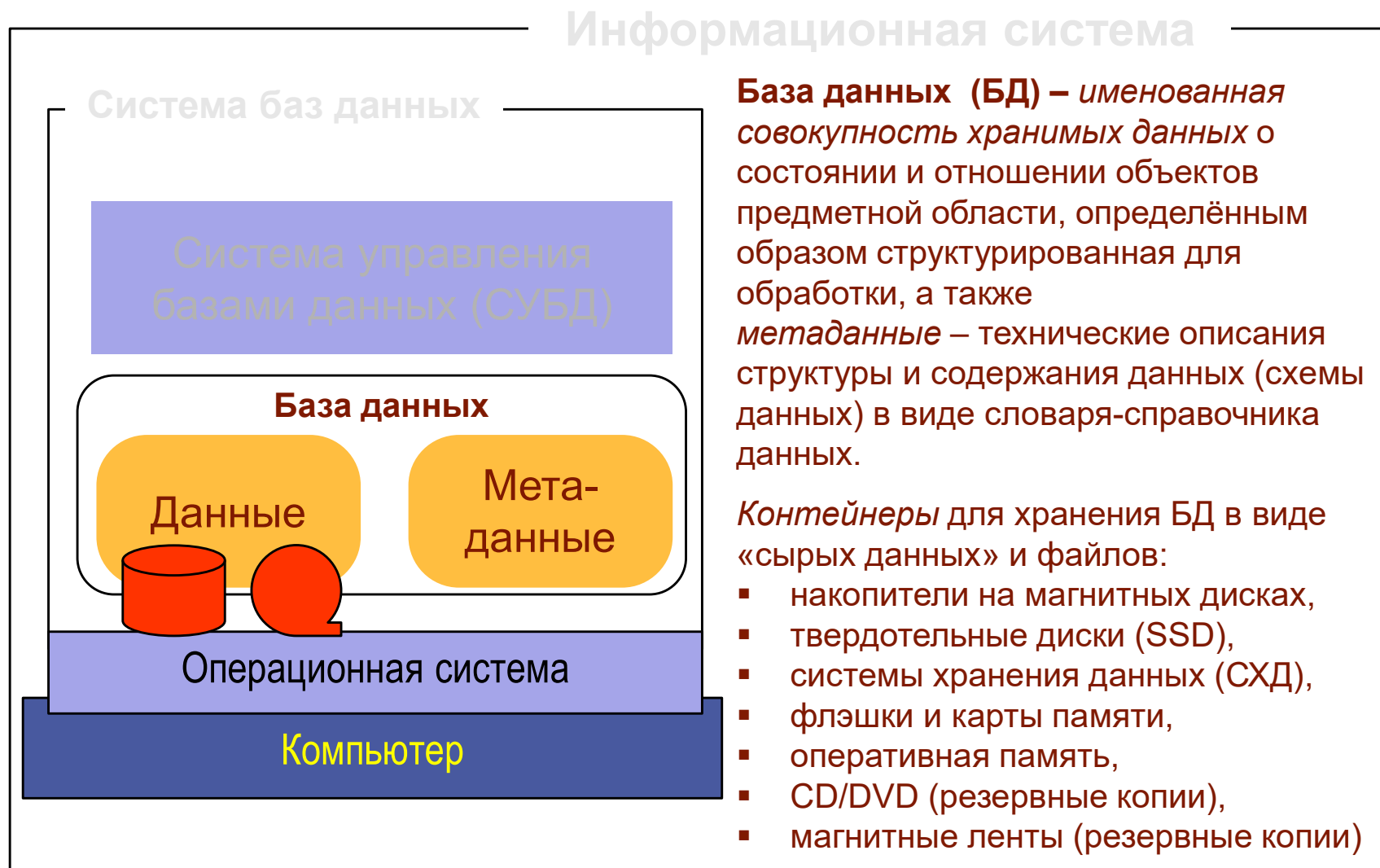
(записи о фактах, БД) – учётно-операционные системы и системы поддержки принятия решений (аналитические системы)

Системы искусственного интеллекта

(временные ряды, языковые модели и т.п.) прогнозные модели, рекомендательные и предписывающие системы, LLM-системы

Всё чаще начинают преобладать **гибридные ИС**

Понятие базы данных



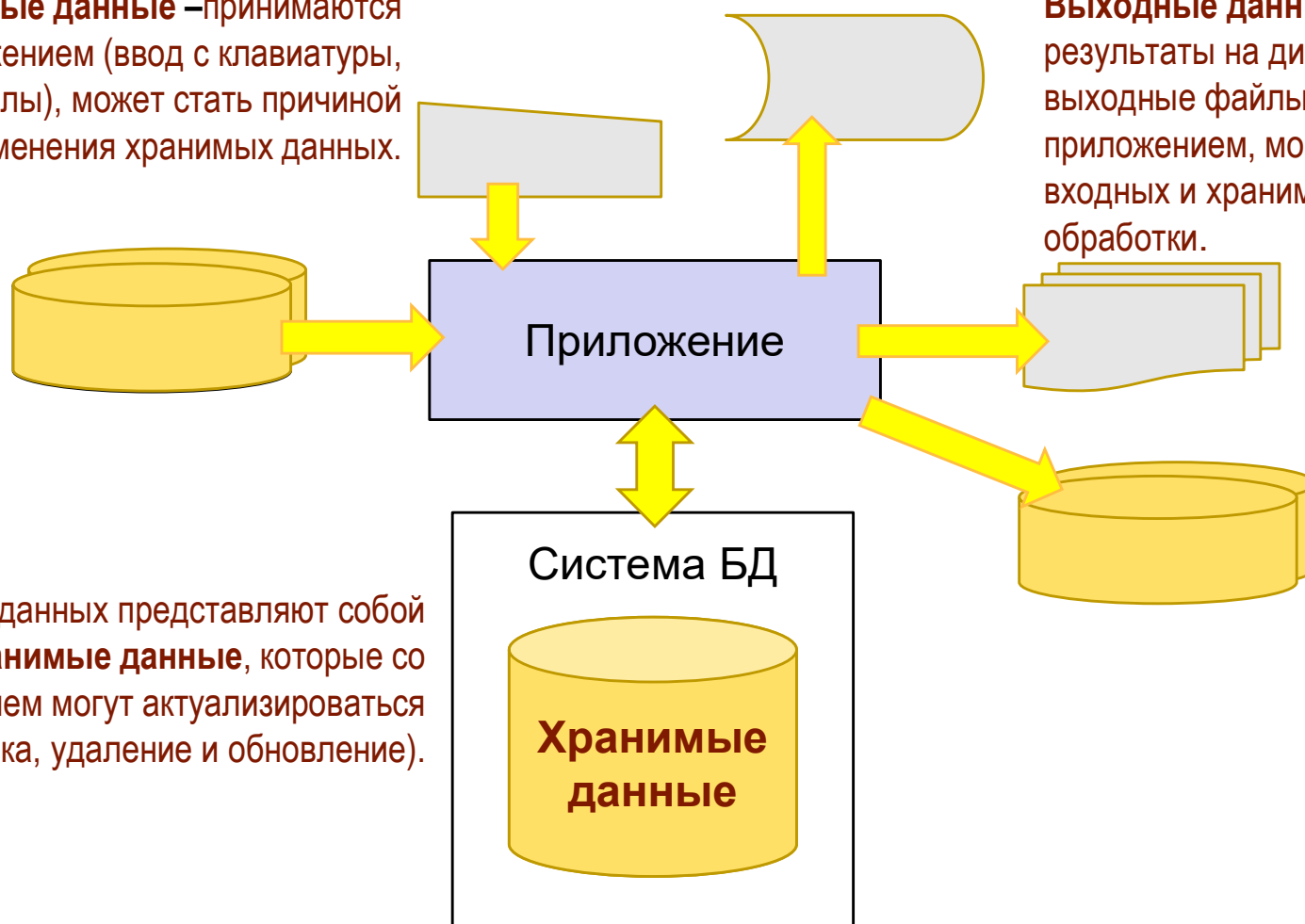
Наличие метаданных позволяет говорить о «самодокументированности» БД

Входные, выходные и хранимые данные

Входные данные – принимаются приложением (ввод с клавиатуры, файлы), может стать причиной изменения хранимых данных.

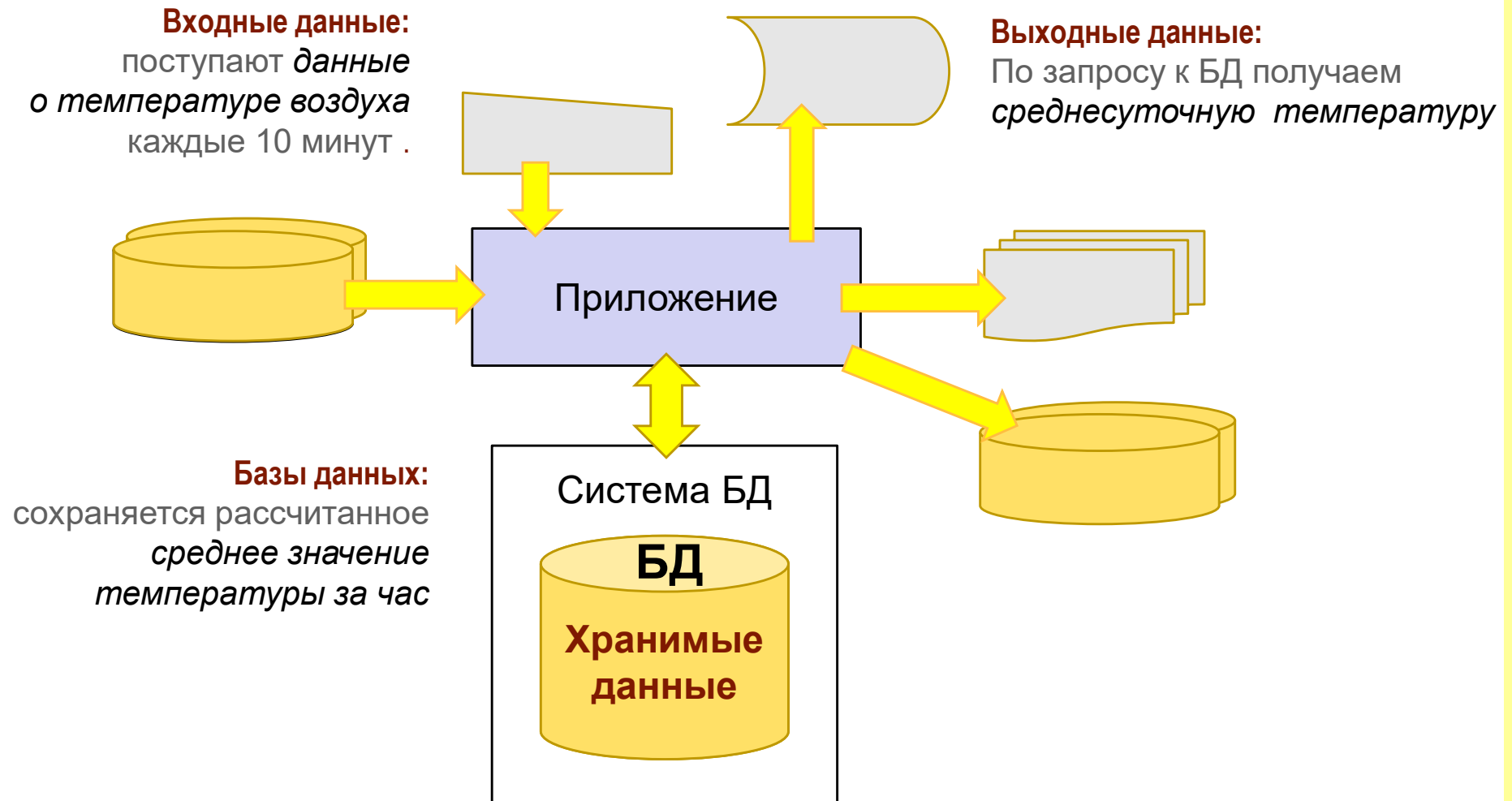
Выходные данные – сообщения, результаты на дисплей / печать, выходные файлы, выдаваемые приложением, могут получаться из входных и хранимых данных после их обработки.

Базы данных представляют собой **хранимые данные**, которые со временем могут актуализироваться (вставка, удаление и обновление).



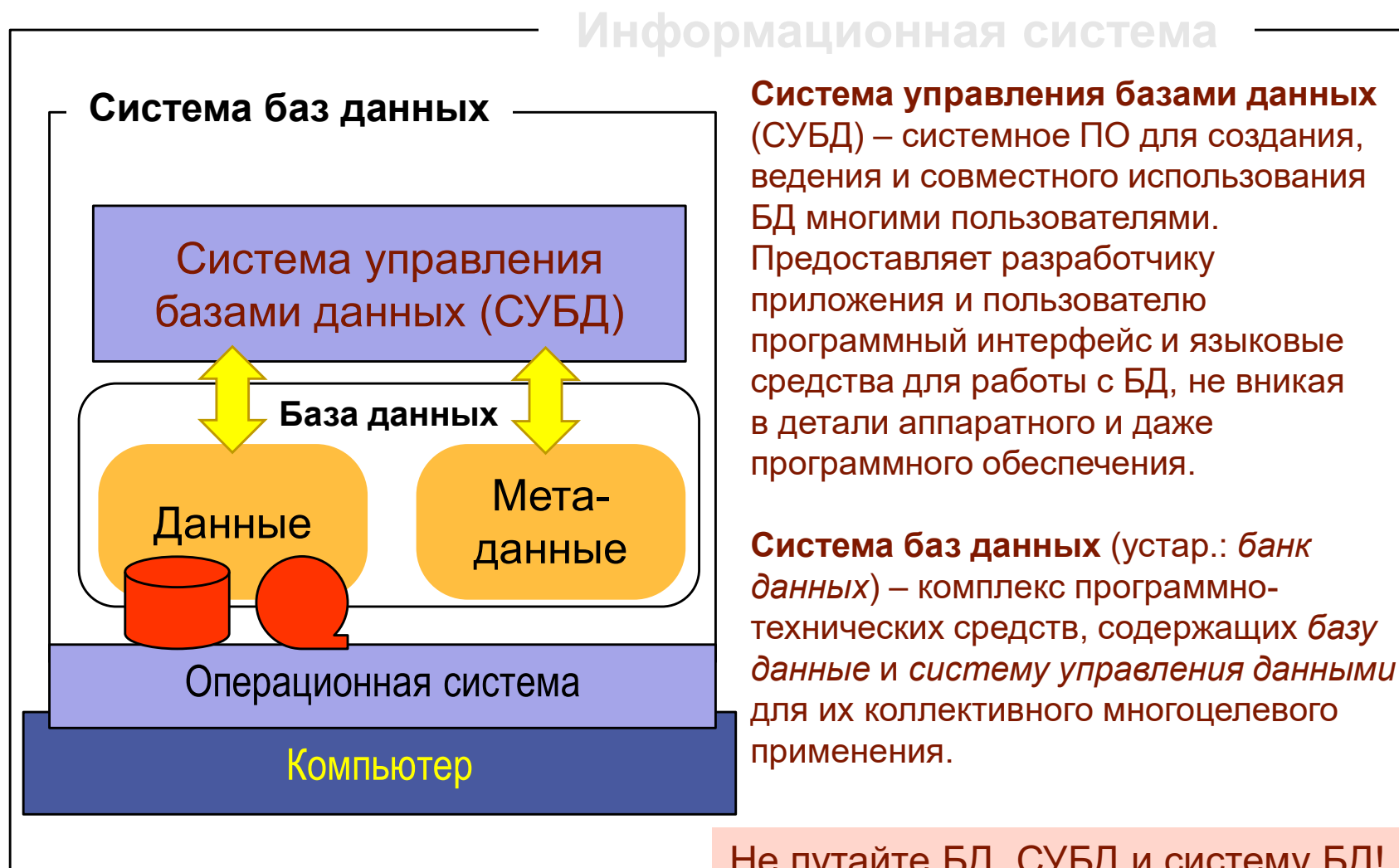
В общем случае транзитные (входные и выходные) данные не являются частью БД.

Пример различия входных, выходных и хранимых данных



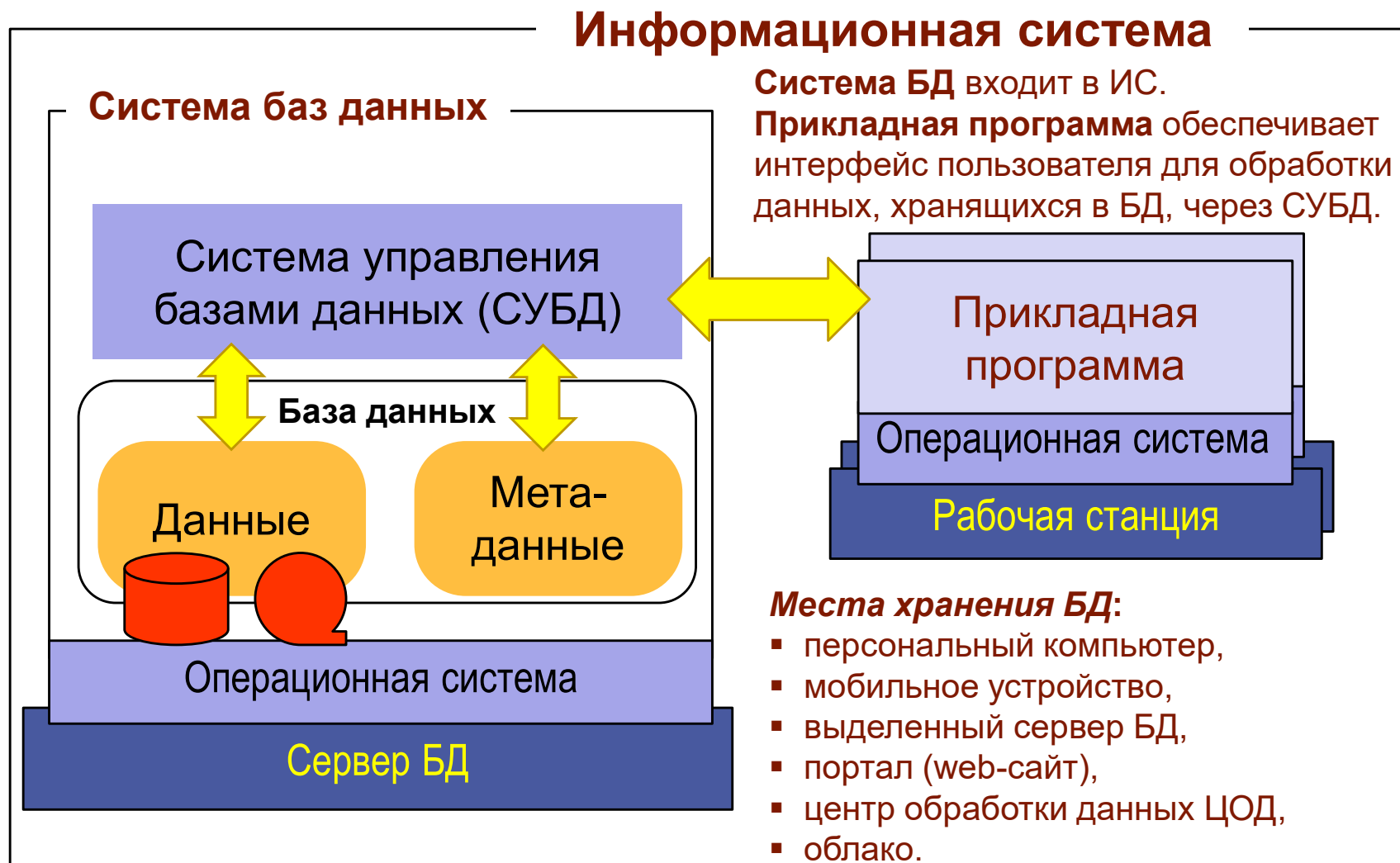
В этом случае хранимые значения будут отличаться
и от входных и от выходных данных.

Понятие СУБД и системы баз данных



БД без СУБД и прикладного программного обеспечения «не живут»!

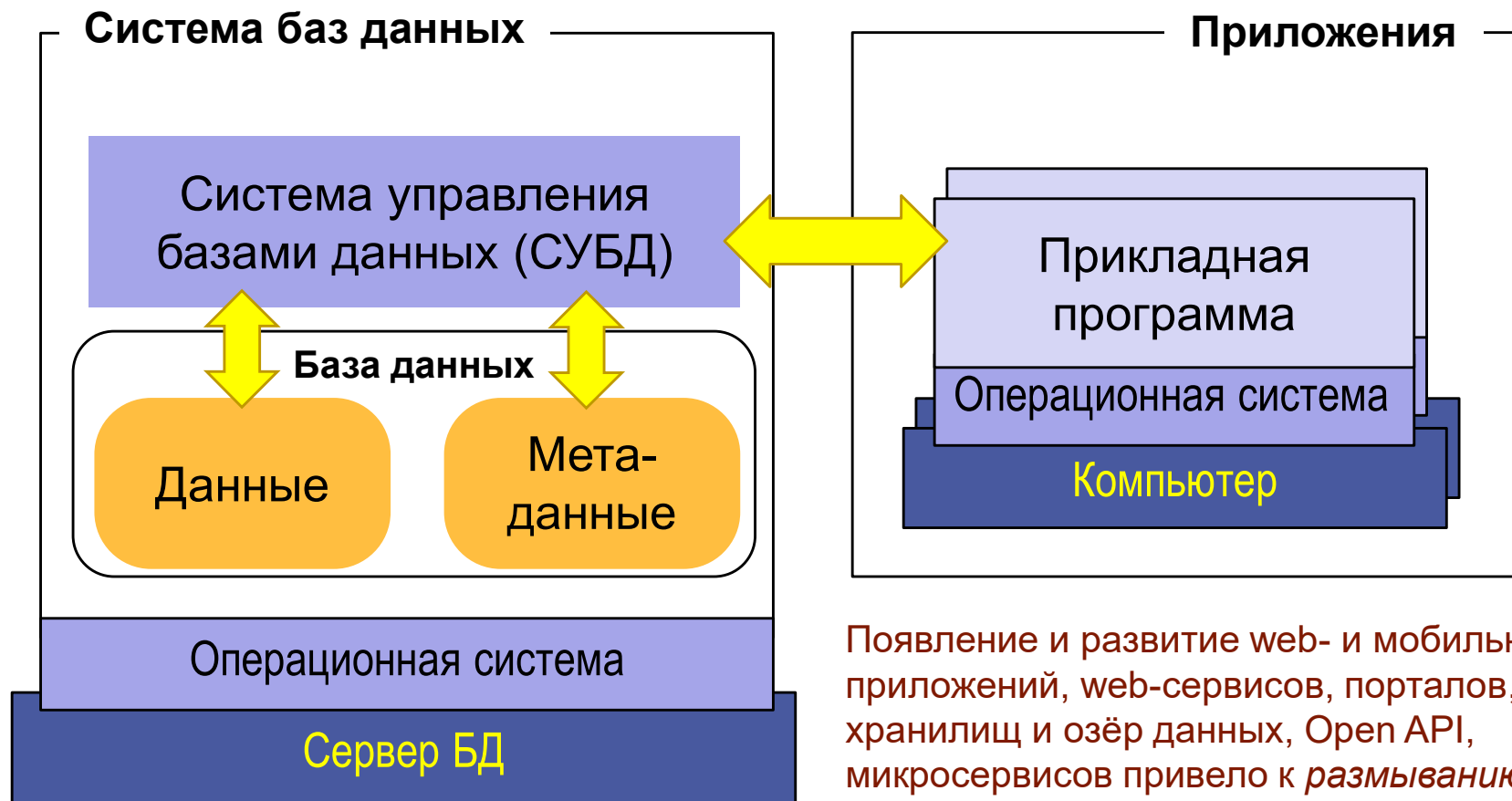
Место системы баз данных в ИС



Ориентация на приложение: раньше система БД являлась частью ИС

Современный взгляд на отношение приложений и системы БД

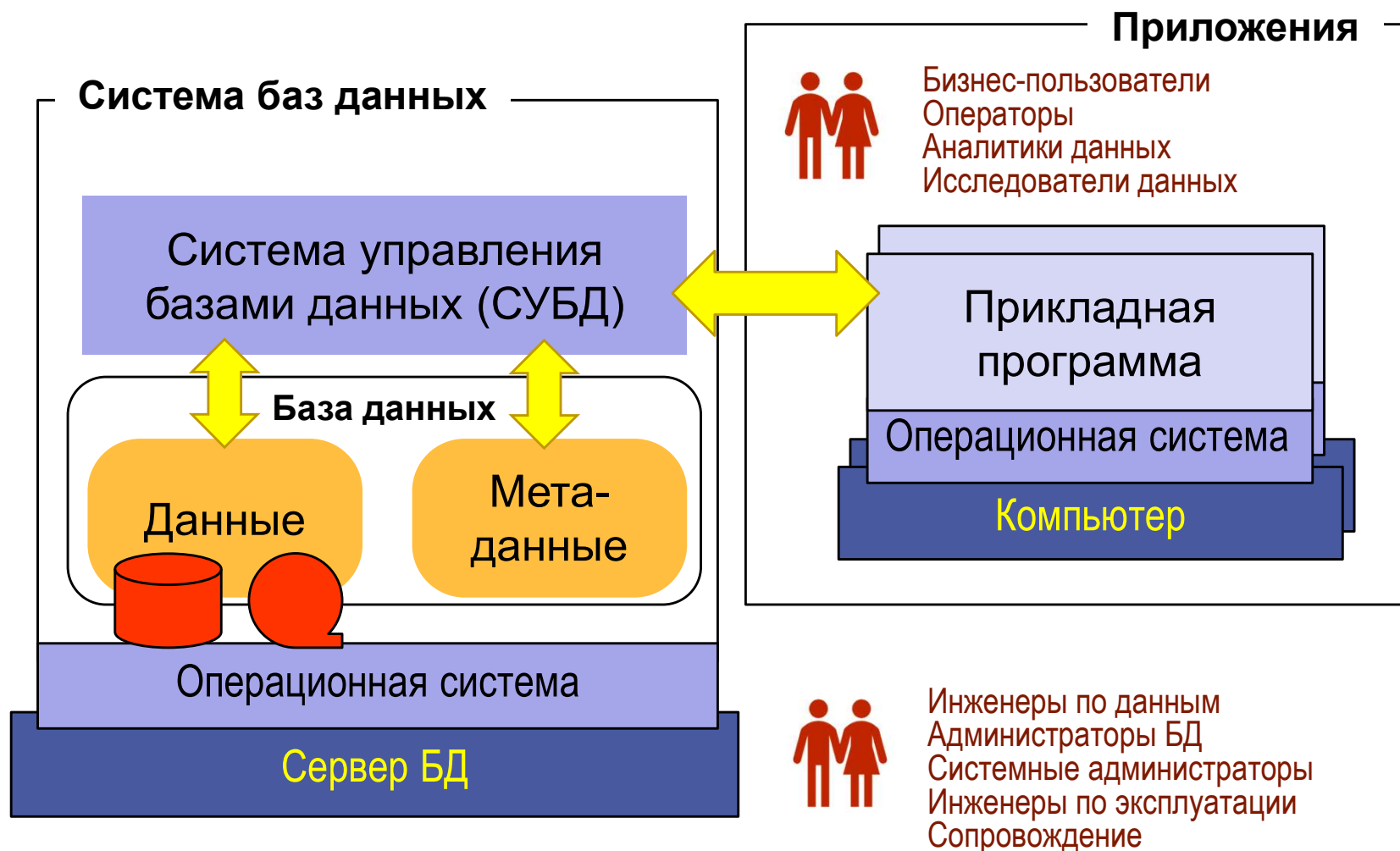
Информаци~~он~~ная система



Появление и развитие web- и мобильных приложений, web-сервисов, порталов, хранилищ и озёр данных, Open API, микросервисов привело к *размыванию понятия ИС и отделению систем БД.*

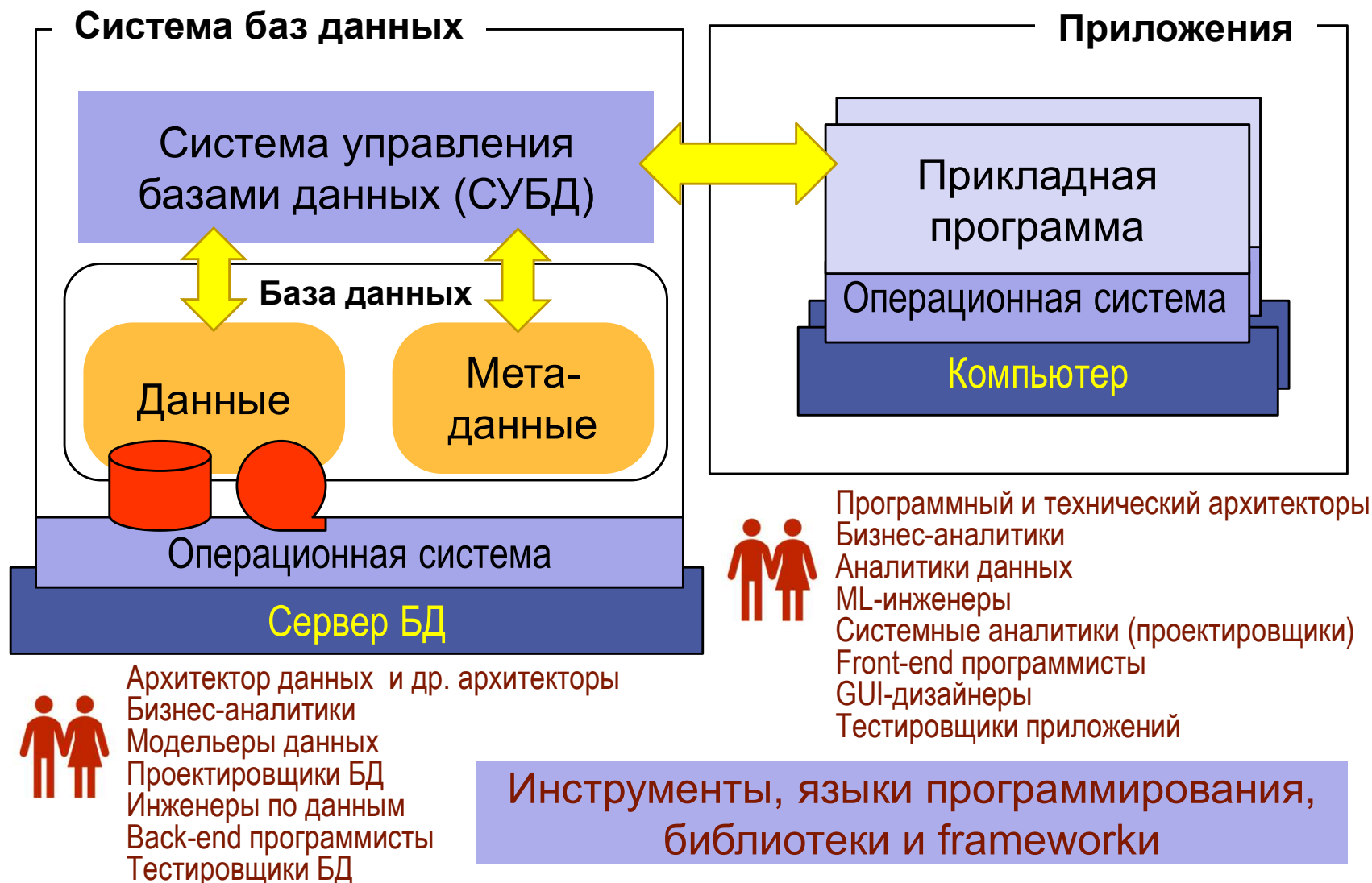
Датацентричный подход: обособление и обобществление систем БД

Роли в системе БД и приложении



Новые роли: аналитики и исследователи данных, инженеры по данным

Роли при создании БД и приложения



Новые роли: аналитики и исследователи данных, инженеры по данным

Бумажные носители данных



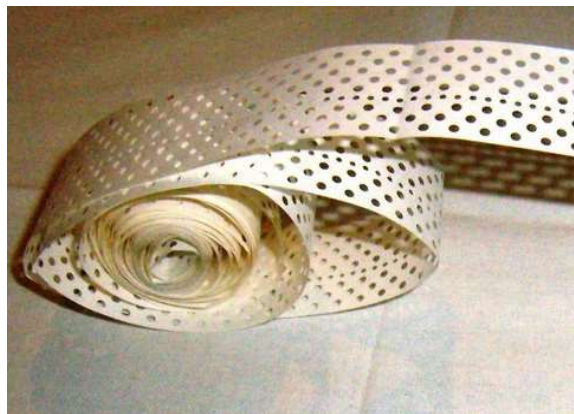
Перфокарты:
Ж.М. Жаккард,
ткацкий станок

Ленты перфокарт:

Б. Бушон,
Ж.Б.Фалькон,
Ж. де Вокансон
ткацкие станки и
игрушки-роботы

XVIII в.

1804



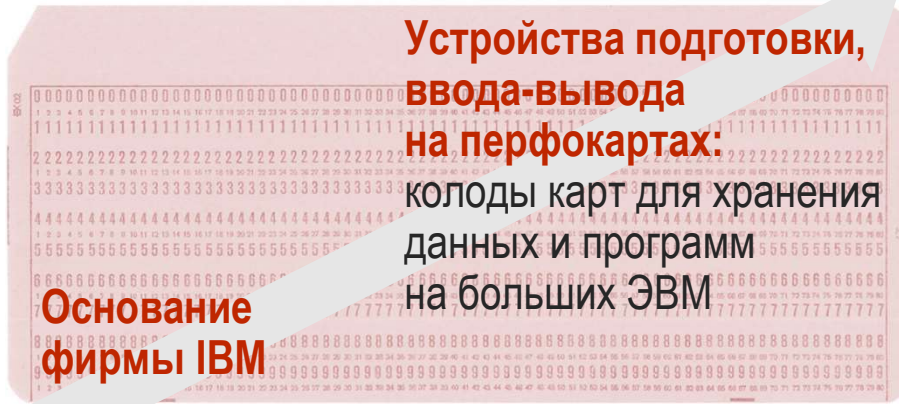
**Перфоленты
телетайп,**
ограниченное
применение
для ЭВМ

Транспортные карты,
учёт работы, шарманка

1846

**Табуляторы
перфокарт:**
Г. Холлерит,
перепись
населения США,
22 т карт, 1 Гб

1890



**Устройства подготовки,
ввода-вывода
на перфокартах:**

колоды карт для хранения
данных и программ
на больших ЭВМ

Основание
фирмы IBM

**Счётные, сортирующие
и суммирующие табуляторы:**
бухучёт, инвентаризация,
расчёт зарплаты,
планирование и т.п

1924

1950-е – 1980-е годы

1960-е – 1990-е годы

Электронные носители данных

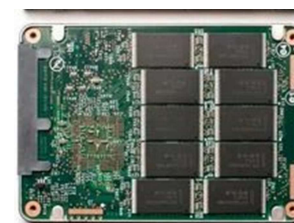


Дисковые массивы:
SAN/FC, NAS

DVD-диски
Электронные архивные системы
(кассеты 12 Tb)

SSD, Израиль

Флэш-память,
карты памяти



CD-диски

Дисковые массивы:
RAID

Гибкий магнитный диск

Магнитная кассета:
Phillips

Жёсткие магнитные диски:
IBM

5 млн символов
сейчас 20 Tb

Магнитная лента:
ЭВМ UNIVAC I

300 Mb

Магнитный барабан:
расширение ОП

Магнитная плёнка:
Ф. Пфлюмер



1928

1932

1951

1956

1963

1969

1970-е

1980-е

1984

1995

Кон. 1990-х

Иерархия систем хранения данных



Исторически сложились 2 направления для работы с данными на ЭВМ

	Виды обработки	Объёмы данных	Структуры данных	Области применения
1	Ввод-вывод и временное хранение данных, <i>сложные расчёты, двоичная арифметика</i>	небольшие	простые	Дешифрация, вооружения, физика, космос, математика, инженерия и т.п.
2	Постоянное хранение, поиск, вставка, обновление, удаление данных и <i>простые расчёты, десятичная арифметика</i>	относительно большие	средние по сложности и сложные	Перепись населения, бухгалтерский учёт, кадровый учёт, платежи, страхование, библиотеки и т.п.
3	Накопление и аналитическая обработка данных	большие	сложные	Принятие решений, анализ и исследование данных, прогнозирование, машинное обучение, ИИ

Программы–предшественники СУБД



Операции ввода-вывода в FORTRAN

PUNCH / PRINT,
OPEN / CLOSE
READ/ WRITE TAPE,
REWIND, BACKSPACE

1957



COBOL – язык программирования
обработки бизнес-данных

1959

IBM

BSAM, BISAM, BDAM – последовательный, индексно-последовательный и прямой методы доступа к наборам данных в OS/360

IBM RPG – генератор отчётов
Цикл обработки записей данных, переквалификация операторов табуляторов

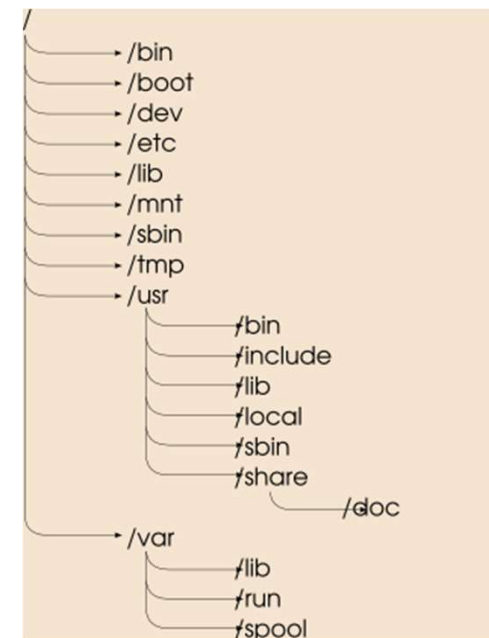
RPG

1964

1960-e

Файловые системы
ОС CTSS и UNIX

Утилиты копирования и сортировки данных



1970-e

Язык программирования обработки бизнес-данных COBOL



1959 г.

«Естественный английский» язык программирования.

Описание *структуры данных, десятичные числа.*

Основной язык разработки приложений для *сфер экономики, финансов, гос. управления до начала 1970-х годов.*

До появления магнитных дисков *преобладала последовательная обработка сортированных наборов данных на магнитных лентах.*

Развивается и используется по сей день.

Много унаследованных приложений.

70% банковских транзакций в США.

Создатели: Комитет **CODASYL**, группа фирм, рук. **Грейс Хоппер**, контр-адмирал ВМС США.

LOOP.

READ IN-FILE, AT END GOTO EOD.

ADD AMOUNT TO SUM.

GOTO LOOP.

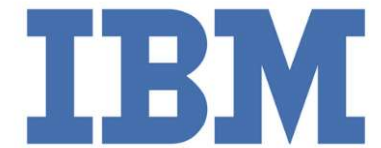
EOD.

MOVE SUM TO TOTAL-AMOUNT.

CLOSE IN-FILE.



Методы доступа к данным и наборы данных в IBM OS/360



BSAM – базисный последовательный метод доступа к данным

QSAM – последовательный метод доступа к данным с очередями

BISAM – базисный индексно-последовательный метод доступа к данным

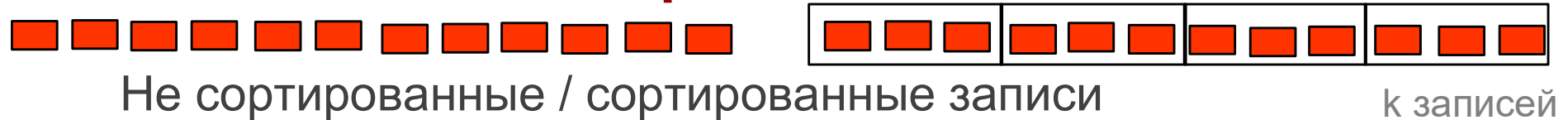
QISAM – индексно-последовательный метод доступа к данным с очередями

BPAM – библиотечный метод доступа

BDAM – прямой (произвольный) метод доступа к данным

Последовательный метод доступа к данным

Последовательная обработка данных



Поиск данных

Сортированный	Сблокированный	Число чтений
нет	нет	N
да	нет	N/2
нет	да (k)	[N/k]
да	да (k)	[N/k]/2

Обновление записей – перезапись файла и корректур

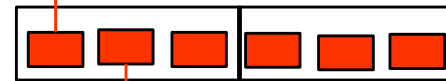
Число чтений / записи $(N+dN)*2$ или $([N/k] + [dN/k])*2$

Индексно-последовательный метод доступа к данным

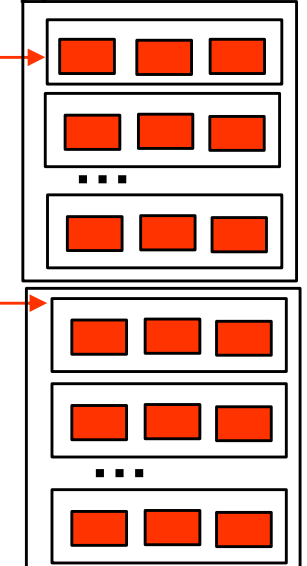
**Последовательная обработка
сортированных данных**

Индекс:

мин / макс значение ключа,
номер блока



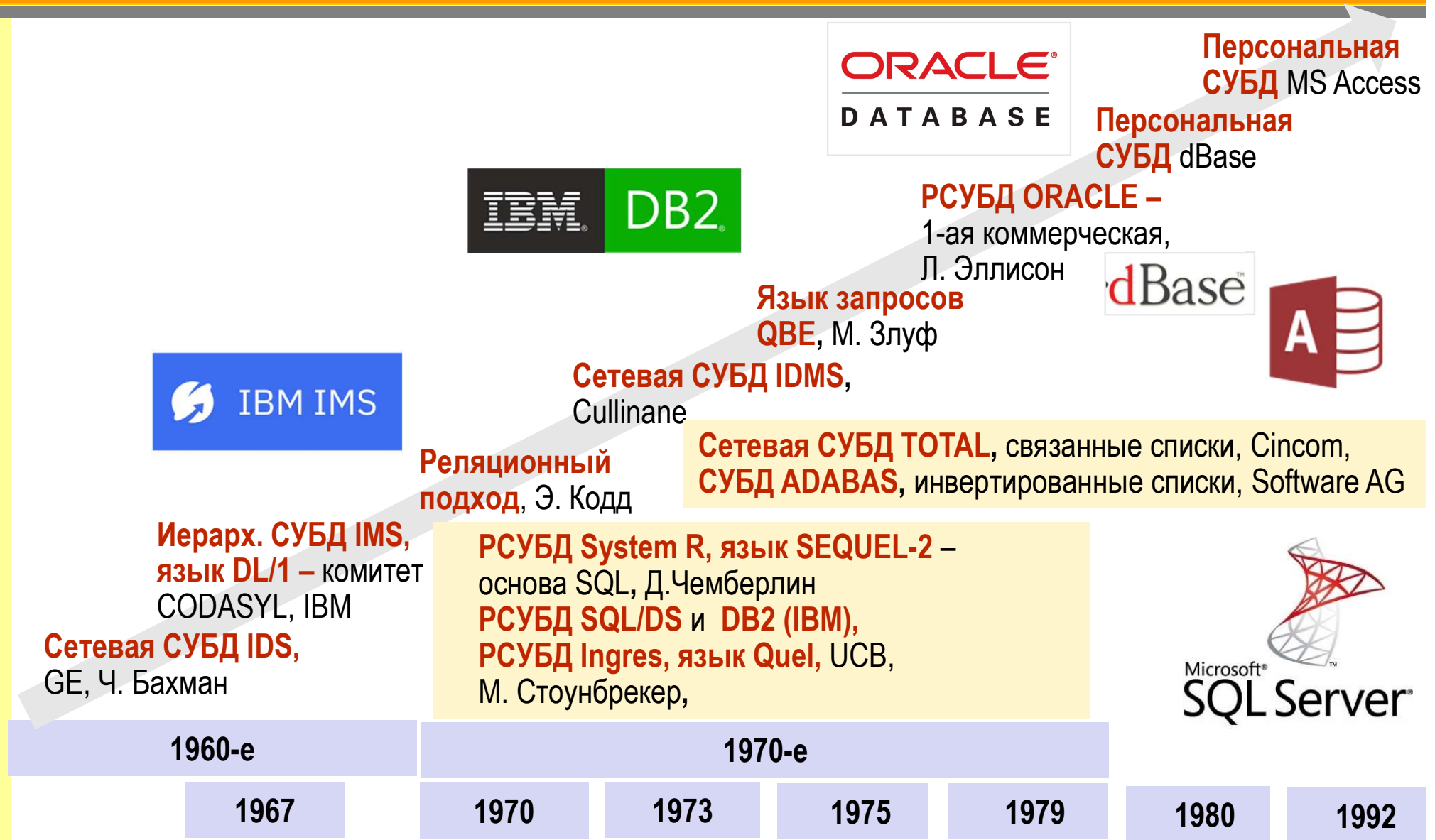
**Сегменты
данных**



Поиск данных

Сблокированный	Чтений индекса	Чтений данных
нет	$m/2$	$[N/m]/2$
да (k)	$[m/k]/2$	$[[N/k]/m]/2$

Появление систем управления БД



Развитие СУБД

Мультимодельные СУБД

1992 г. Illustra (объектно-реляционная версия Postgres) → Informix → IBM UDB

1997 г. PostgreSQL – Open Source СУБД

2000-е расширение типов данных и моделей в СУБД

Многомерные СУБД

1970 г. Express,

1993 г. OLAP, Э. Кодд, Essbase

К 2000-м OLAP встроены в СУБД ORACLE, MS SQL Server, IBM DB2

Специализированные СУБД

1980-е Tandem Non-Stop (высоконадёжная транзакционная обработка)

1990-е Teradata MPP (хранилище данных)

2000-е IBM Netezza, Oracle Exadata (хранилища данных)

СУБД класса NOSQL: начало 2000-х

Google, HDFS, MapReduce, BigTable, OpenSource, Hive, HBase, MongoDB

СУБД класса NewSQL: начало 2010-х

VoltDB, модификации MySQL, sharding, Greenplum

Облачные СУБД: SQL/NoSQL, Amazon, MS Azure, Snowflake, Oracle

Классификация СУБД

Признаки классификации:

Степень структурированности данных

Абстракция в основе модели данных

Назначение

Распределённость данных

Архитектура приложений

Характер использования

Рабочая нагрузка

Данные с разной степенью структурированности

	Структурированные данные	Полуструктурированные данные	Неструктурированные данные
Форматы	Таблицы, строки, столбцы, ключи	XML, JSON, CSV, log	Текст, картинки, аудио, видео, гипертекст, геокарты
Абстракция модели данных	Реляционная, многомерная, иерархия, сетевая	Иерархия, граф, плоская	Нет модели данных
Схема данных	Хорошо определённая фиксированная схема	Слабая схема данных	Нет схемы
Хранение данных	Реляционные СУБД, классические хранилища данных (data warehouse)	Нереляционные СУБД (NOSQL)	Файловые системы, озёра данных (data lake), облачные ХД
Методы обработки	SQL-запросы, QBE, BI, data mining, AI	Языки запросов, data mining, AI	NLP, анализ текста, LLM, распознавание образов, анализ аудио и видео

Классификация СУБД и баз данных по абстракции модели данных

- Реляционные БД (SQL)
- Иерархические и сетевые (навигационные) БД
- Поисковые (полнотекстовые) БД
- Объектные и объектно-реляционные БД
- Многомерные БД
- Темпоральные БД
- БД NOSQL (ключ-значение, документные, графовые, колоночные, семейства колонок)
- Контентные БД (изображения, аудио, видео, гипертекст)
- Мультимодельные БД
- Специализированные БД (пространственные, Native XML, векторные)

Мультимодельные СУБД

СУБД	Изначальная модель	Дополнительные модели
Oracle	Реляционная	Многомерная, графовая, документная
MS SQL Server	Реляционная	Графовая, документная
PostgreSQL	Реляционная	Графовая, документная
MarkLogic	Документная	Графовая, реляционная
MongoDB	Документная	Ключ-значение, графовая
DataStax	Семейство колонок	Документная, графовая
Redis	Ключ-значение	Документная, графовая
ArangoDB	—	Графовая, документная
OrientDB	—	Графовая, документная
Azure CosmosDB	—	Графовая, документная, реляционная

Возможности СУБД с разными моделями данных

Возможность	Реляционная	Графовая	Объектная	Семейство колонок	Документная	Ключ-значение
Поиск по ключу	+	+	+	+	+	+
Извлечение/ обновление части атрибута	+	+	+	+	+	—
Поиск по неключевым атрибутам	+	+	+	+	+	—
Составной ключ	+	+	+	+	—	—
Связи между сущностями	+	+	+	—	—	—
Стандартный интерфейс	+	±	—	—	—	—
Слабоструктурированные данные	+	—	—	+	+	+

Классификация СУБД (3)

По структурированности
данных

По абстракции модели данных

По назначению

По распределённости
данных

По доступу к БД

По характеру
использования

По рабочей нагрузке

**СУБД ОБЩЕГО НАЗНАЧЕНИЯ
(УНИВЕРСАЛЬНЫЕ)**

Специализированные СУБД

Классификация СУБД (4)

По структурированности
данных

По абстракции модели данных

По назначению

**По распределённости
данных**

По доступу к БД

По характеру
использования

По рабочей нагрузке

Сосредоточенные

- Локальные
- Централизованные

Распределённые

- Однородные
- Неоднородные
- Фрагментированные
- Тиражируемые

Классификация СУБД (5)

По структурированности
данных

По модели данных

По универсальности

По распределённости
данных

По архитектуре

По характеру
использования

По рабочей нагрузке

Встроенные

Настольные (персональные)

Сетевые (файл-серверные)

Клиент-серверные

Облачные

Облачные сервисы

Классификация СУБД (6)

По структурированности
данных

По модели данных

По универсальности

По распределённости
данных

По архитектуре приложения

**По характеру
использования**

По рабочей нагрузке

Однопользовательские

МНОГОПОЛЬЗОВАТЕЛЬСКИЕ

Классификация СУБД (7)

По структурированности
данных

По модели данных

По универсальности

По распределённости
данных

По доступу к БД

По характеру
использования

По рабочей нагрузке

**Транзакционная нагрузка
OLTP (TPC-A,B,C)**

**Поддержка принятия
решений OLAP/DSS (TPC-D)**

Мультимедийная нагрузка

Рейтинг современных СУБД

DB-Engines Ranking

424 systems in ranking, February 2025

Rank			DBMS	Database Model	Score		
Feb 2025	Jan 2025	Feb 2024			Feb 2025	Jan 2025	Feb 2024
1.	1.	1.	Oracle	Relational, Multi-model ⓘ	1254.82	-3.93	+13.38
2.	2.	2.	MySQL +	Relational, Multi-model ⓘ	999.99	+1.84	-106.67
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model ⓘ	786.87	-11.69	-66.70
4.	4.	4.	PostgreSQL +	Relational, Multi-model ⓘ	659.62	-3.79	+30.21
5.	5.	5.	MongoDB +	Document, Multi-model ⓘ	396.63	-5.87	-23.73
6.	↑ 7.	6.	Redis +	Key-value, Multi-model ⓘ	157.91	+4.55	-2.80
7.	↓ 6.	↑ 9.	Snowflake +	Relational	155.58	+1.68	+28.13
8.	8.	↓ 7.	Elasticsearch	Multi-model ⓘ	134.63	-0.29	-1.11
9.	9.	↓ 8.	IBM Db2	Relational, Multi-model ⓘ	125.44	+2.46	-6.79
10.	10.	10.	SQLite	Relational	113.82	+7.13	-3.47

**Терпения и удачи всем, кто
связан с базами данных**

Спасибо за внимание!

Валерий Иванович Артемьев

МГТУ имени Н.Э. Баумана, кафедра ИУ-5

Банк России

Департамент данных, проектов и процессов

Тел.: +7(495) 753-96-25

e-mail: viart@bmstu.ru