

Базы данных

А1. Введение в базы данных



Московский государственный технический университет
имени Н.Э. Баумана

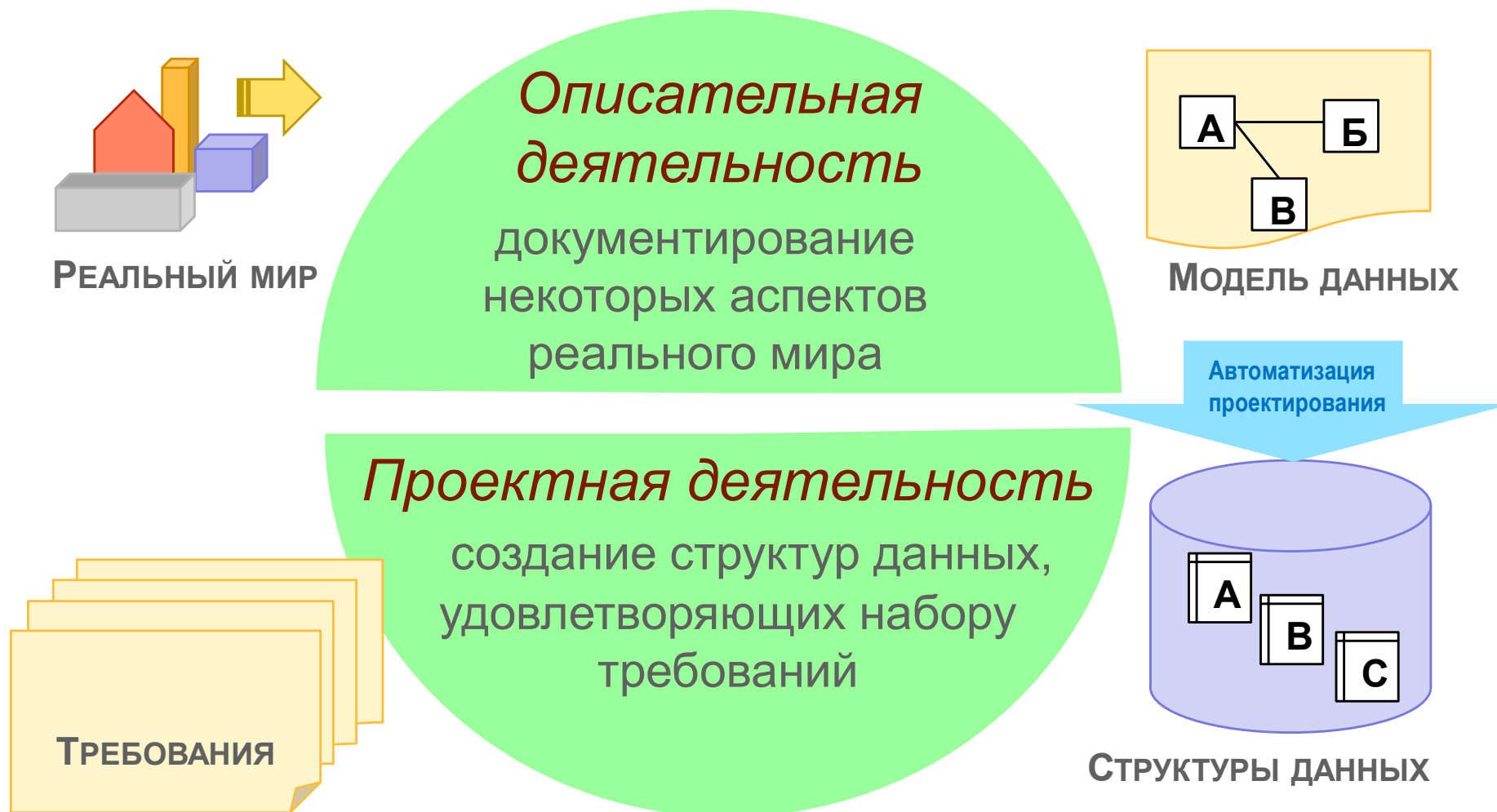
Факультет ИБМ

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

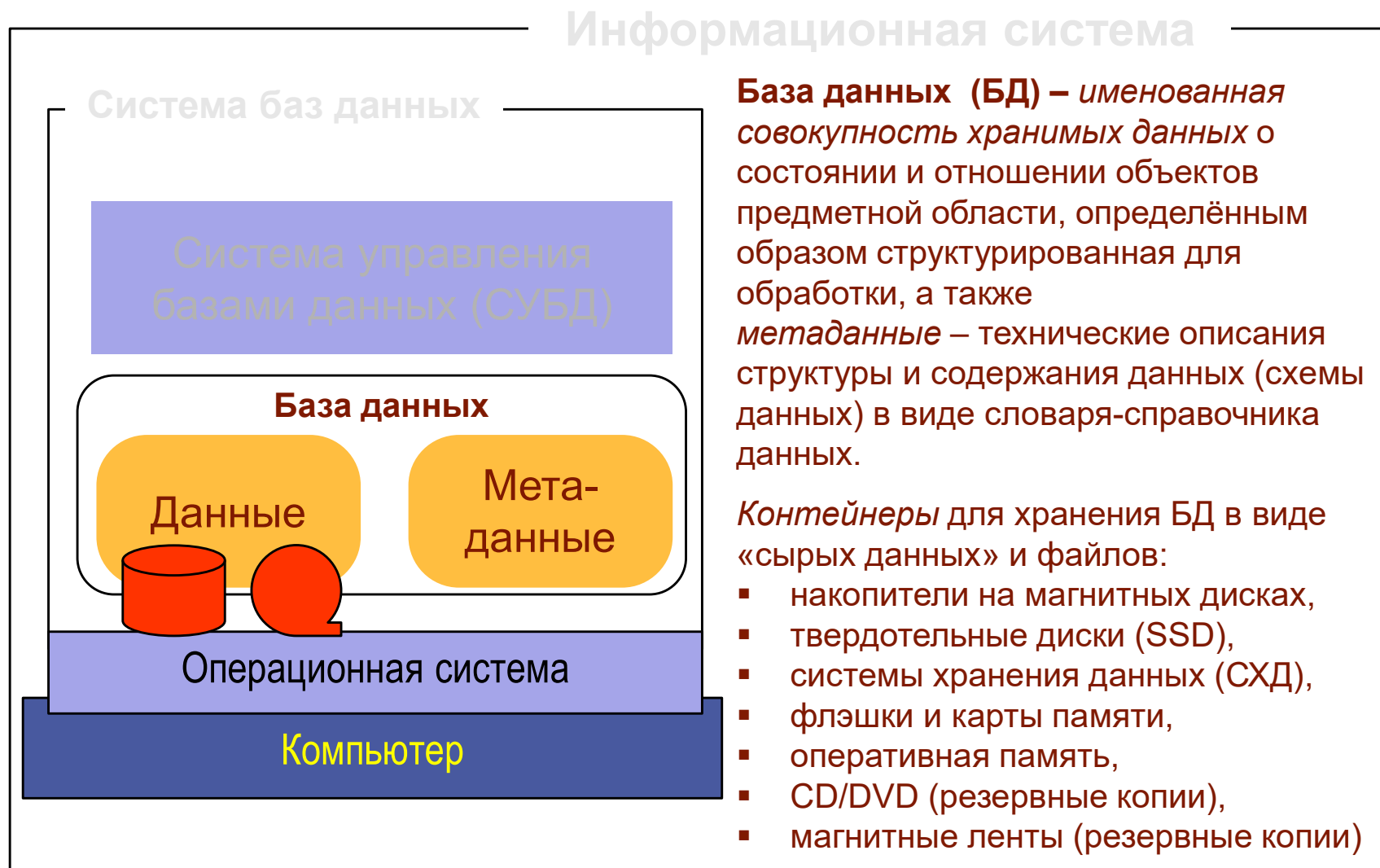
(записи о фактах, БД) – учётно-операционные системы и системы поддержки принятия решений (аналитические системы)

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



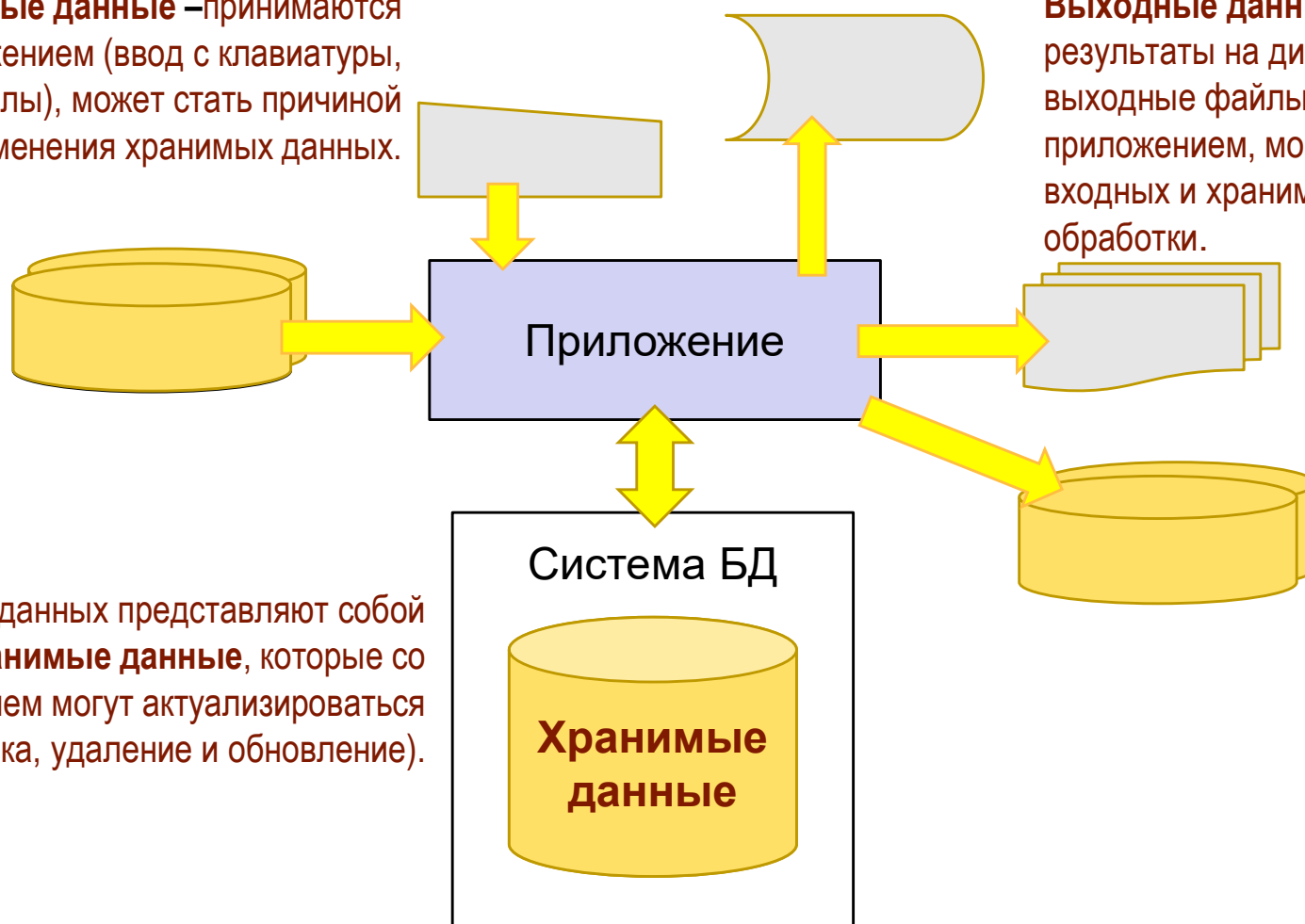
Наличие метаданных позволяет говорить о «самодокументированности» БД

Входные, выходные и хранимые данные

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

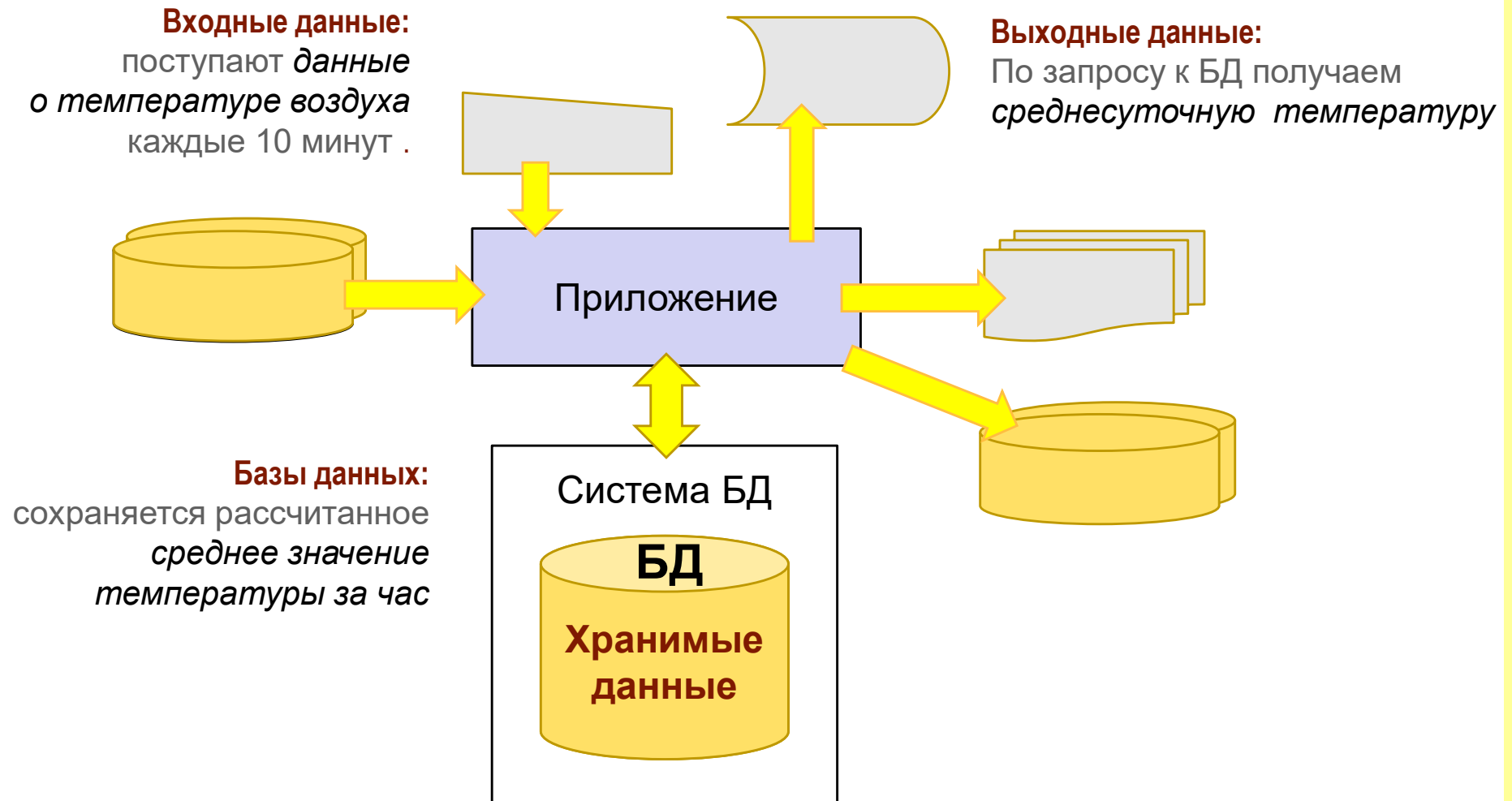
Выходные данные – сообщения, результаты на дисплей / печать, выходные файлы, выдаваемые приложением, могут получаться из входных и хранимых данных после их обработки.

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



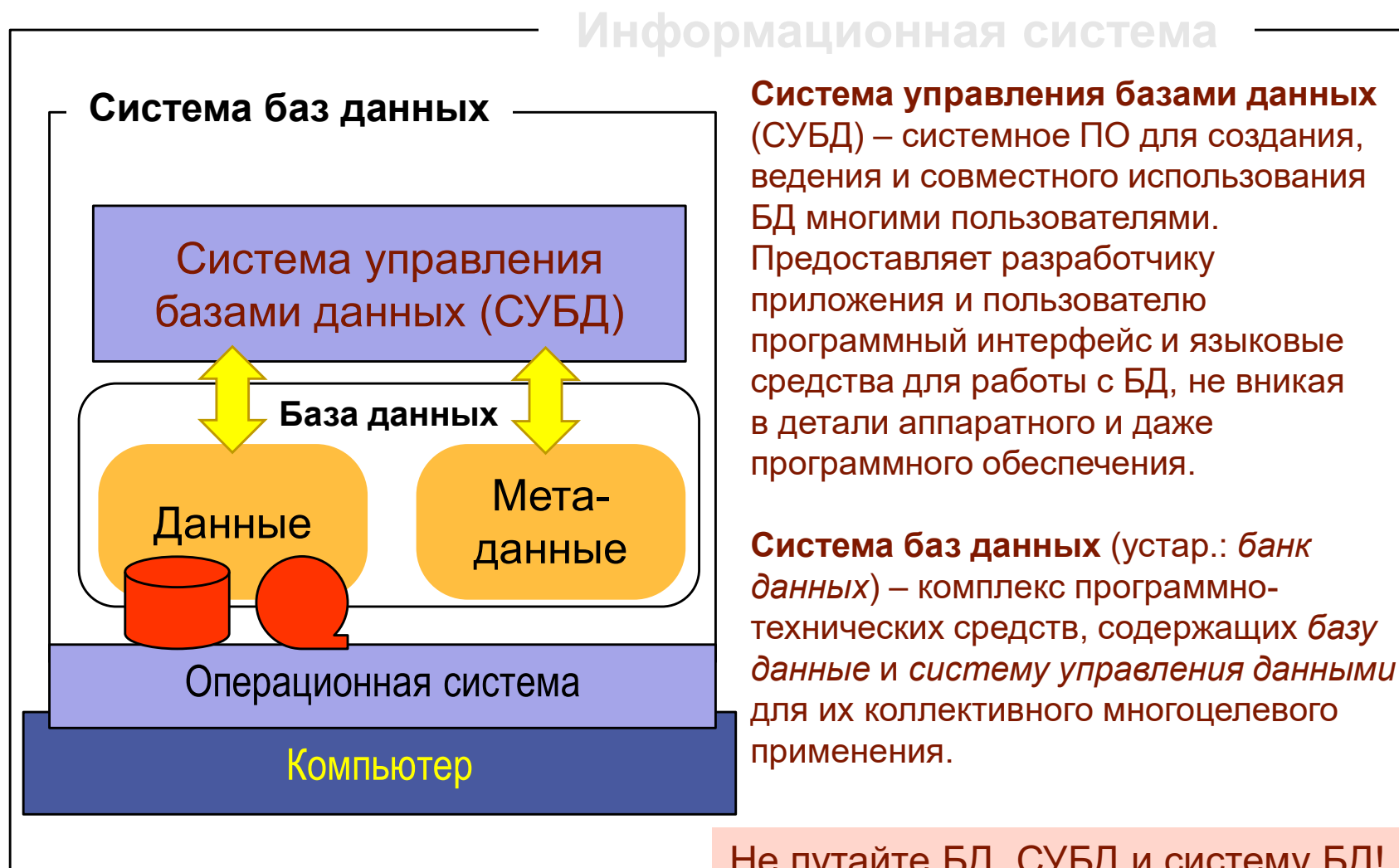
В общем случае транзитные (входные и выходные) данные не являются частью БД.

Пример различия входных, выходных и хранимых данных



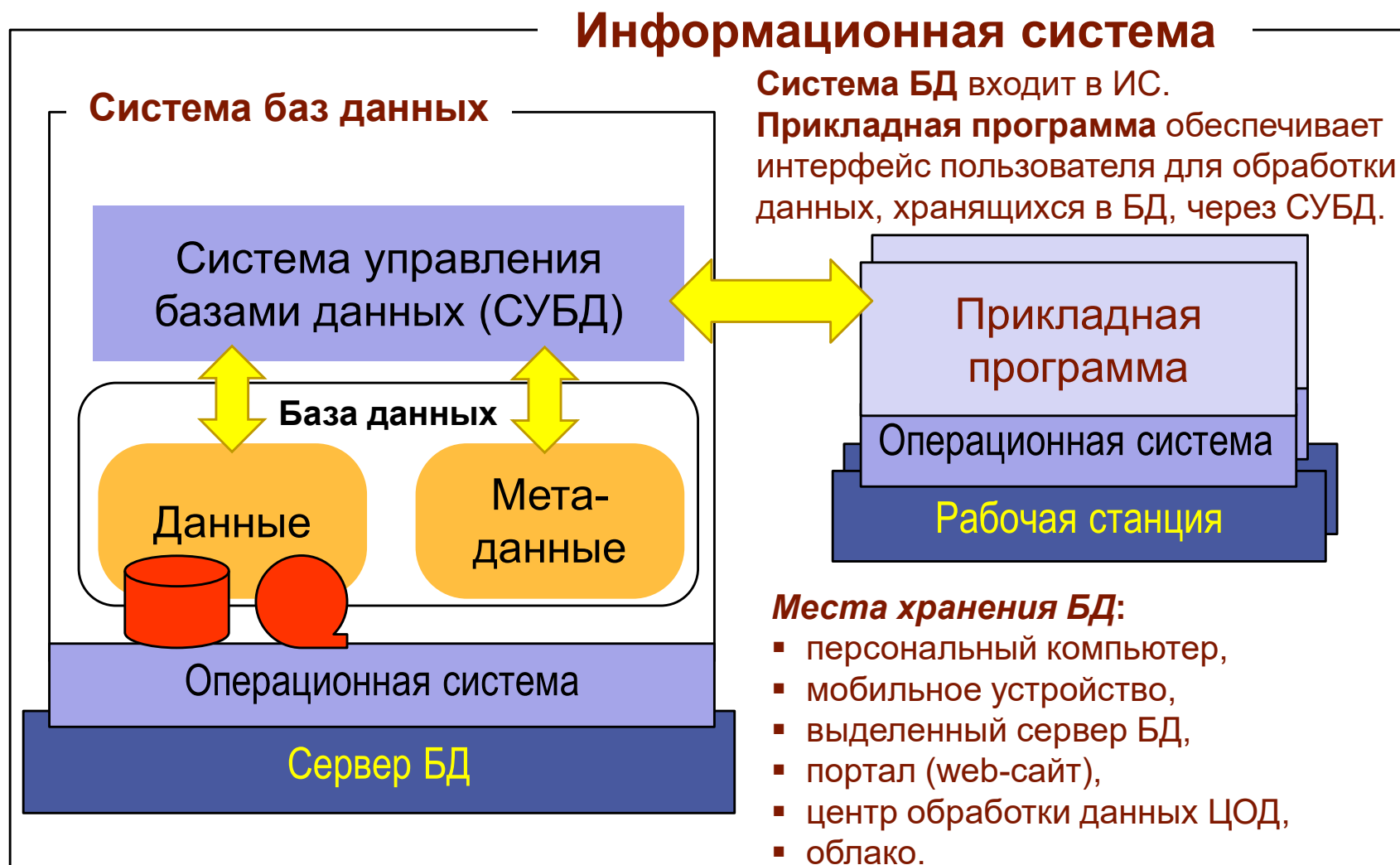
В этом случае хранимые значения будут отличаться
и от входных и от выходных данных.

Понятие СУБД и системы баз данных



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

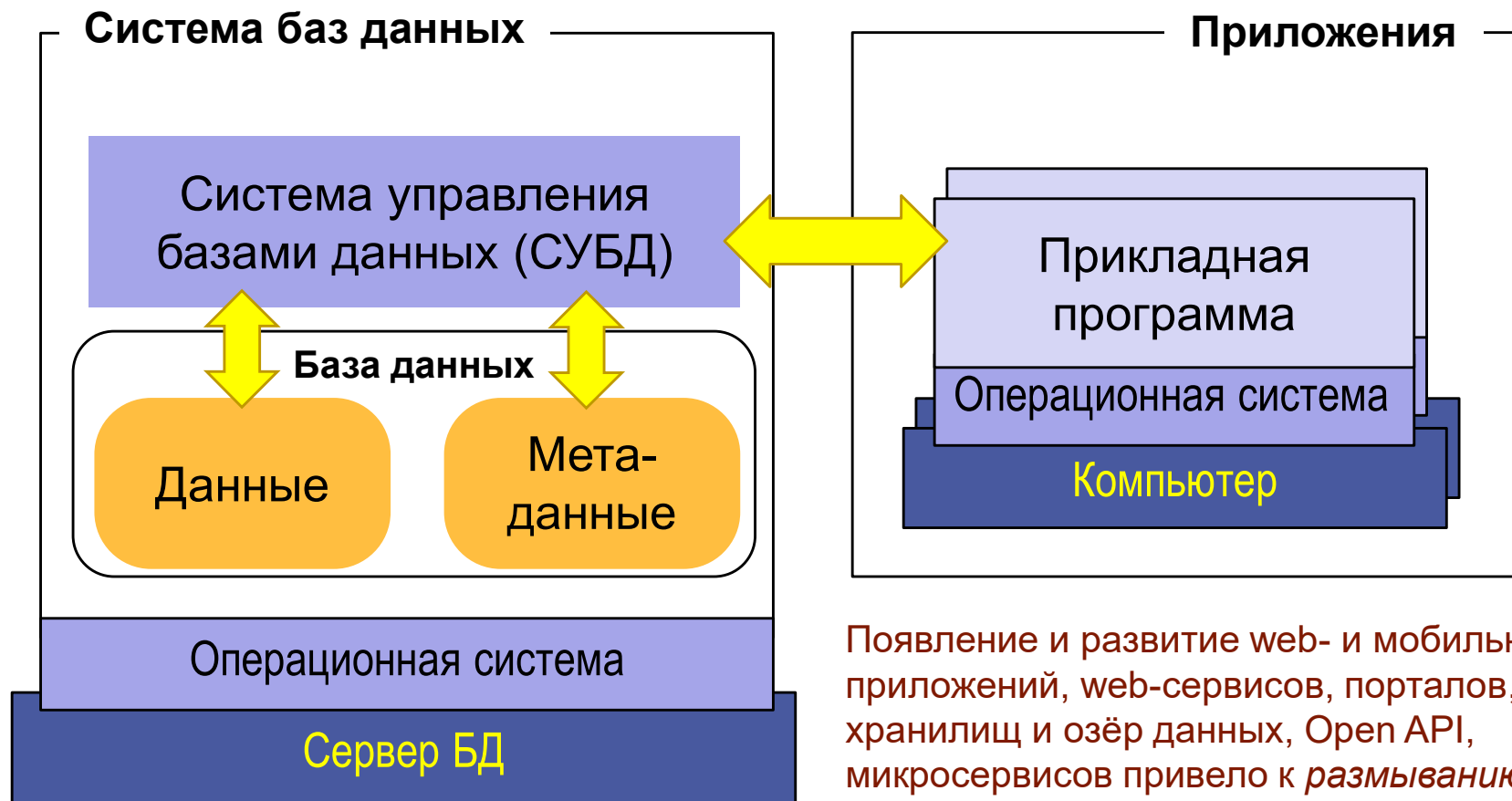
Место системы баз данных в ИС



Ориентация на приложение: раньше система БД являлась частью ИС

Современный взгляд на отношение приложений и системы БД

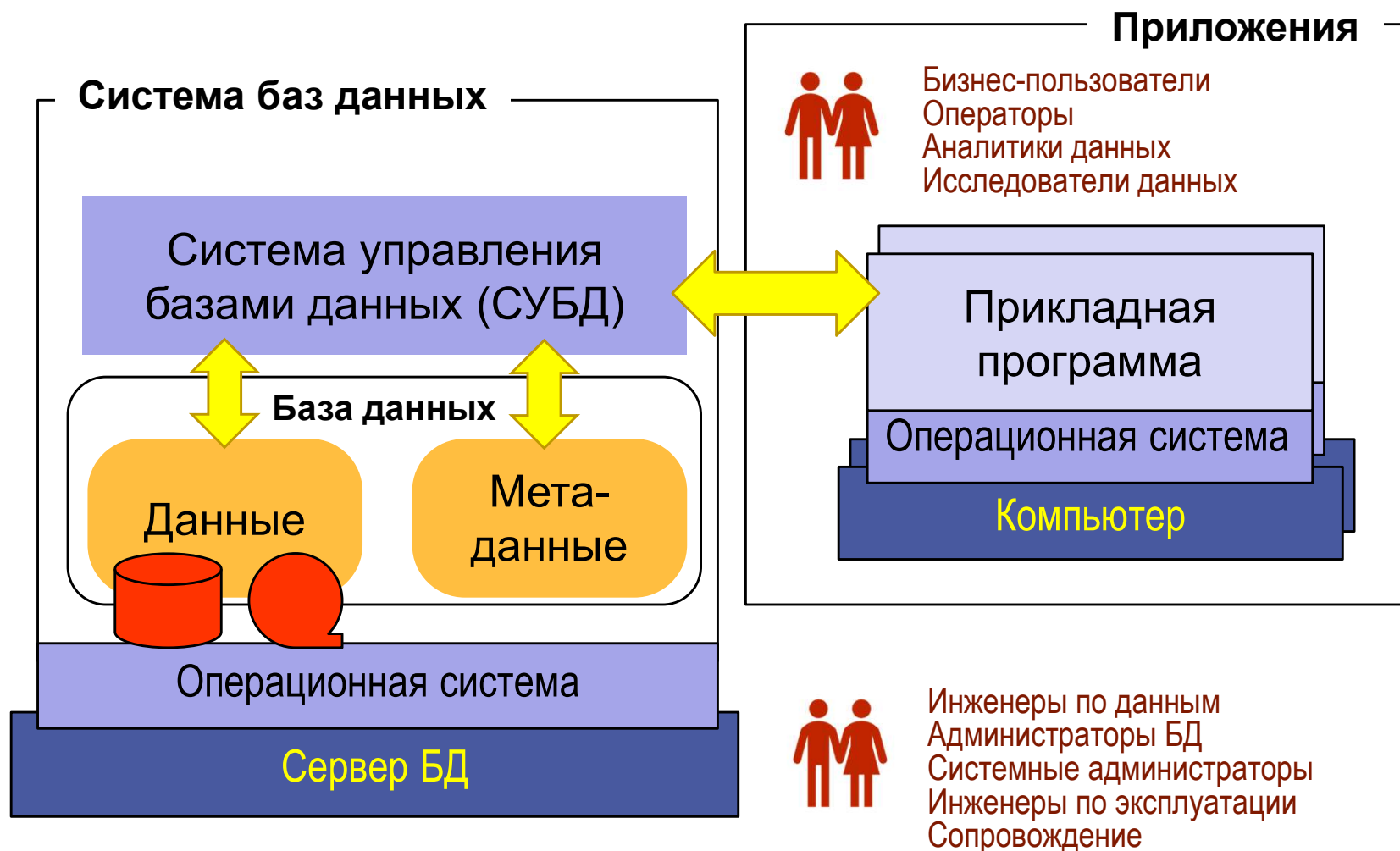
Информаци~~он~~ная система



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.*

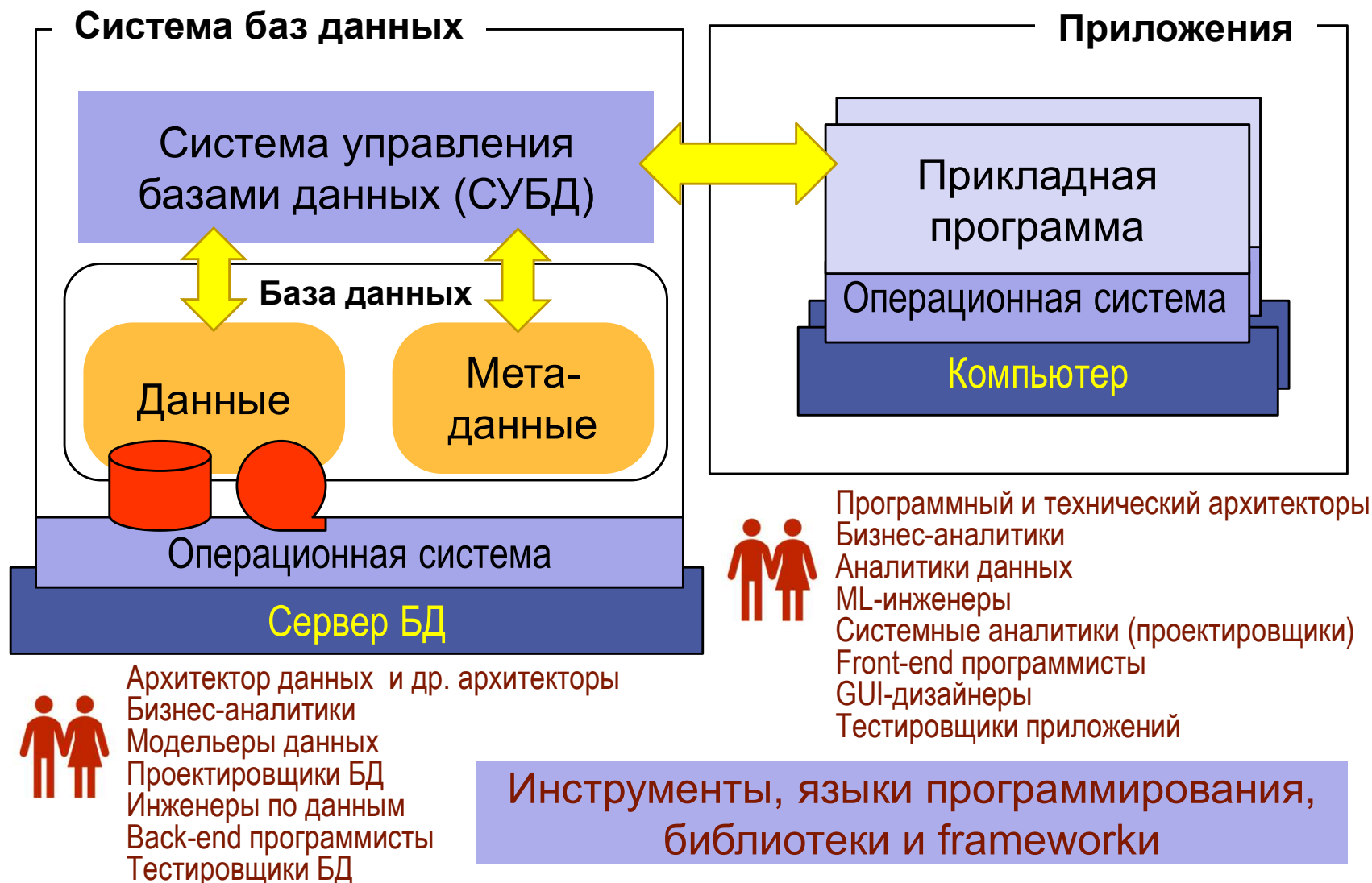
Датацентричный подход: обособление и обобществление систем БД

Роли в системе БД и приложении



Новые роли: аналитики и исследователи данных, инженеры по данным

Роли при создании БД и приложения



Бумажные носители данных



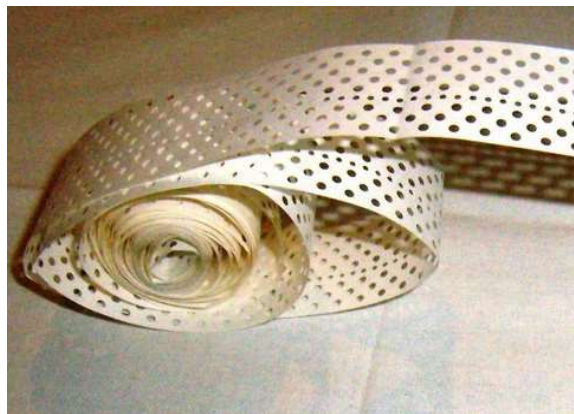
Перфокарты:
Ж.М. Жаккард,
ткацкий станок

Ленты перфокарт:

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

XVIII в.

1804



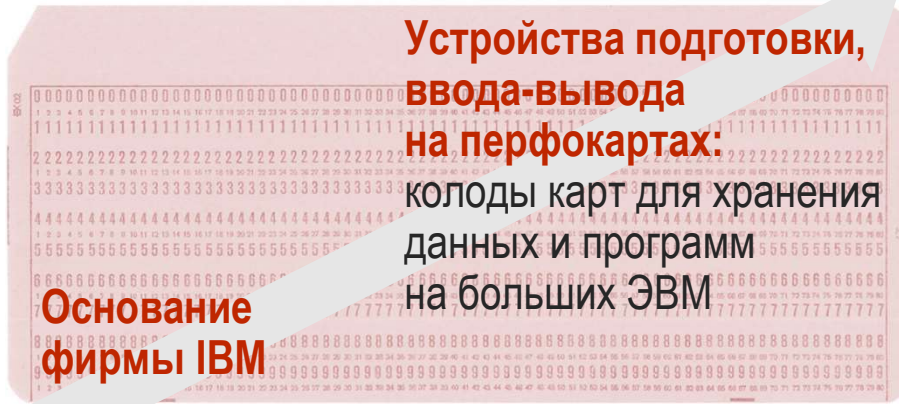
**Перфоленты
телетайп,**
ограниченное
применение
для ЭВМ

Транспортные карты,
учёт работы, шарманка

1846

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

1890



**Основание
фирмы IBM**

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

колоды карт для хранения
данных и программ
на больших ЭВМ

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

1960-е – 1990-е годы

1924

1950-е – 1980-е годы

Электронные носители данных



Дисковые массивы:
SAN/FC, NAS

DVD-диски
Электронные архивные системы
(кассеты 12 Tb)

Дисковые массивы:
RAID

SSD, Израиль
Флэш-память, карты памяти

CD-диски

Гибкий магнитный диск

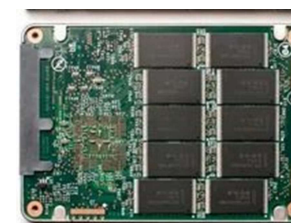
Магнитная кассета:
Phillips

Жёсткие магнитные диски:
IBM

Магнитная лента:
ЭВМ UNIVAC I

Магнитный барабан:
расширение ОП

Магнитная плёнка:
Ф. Пфлюмер



1928

1932

1951

1956

1963

1969

1970-е

1980-е

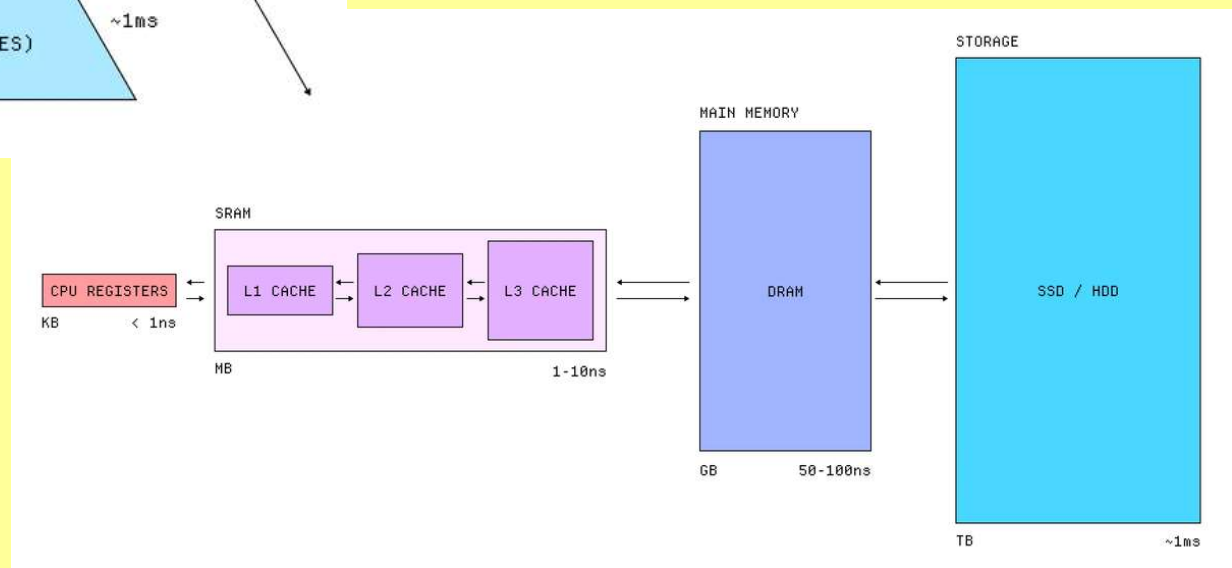
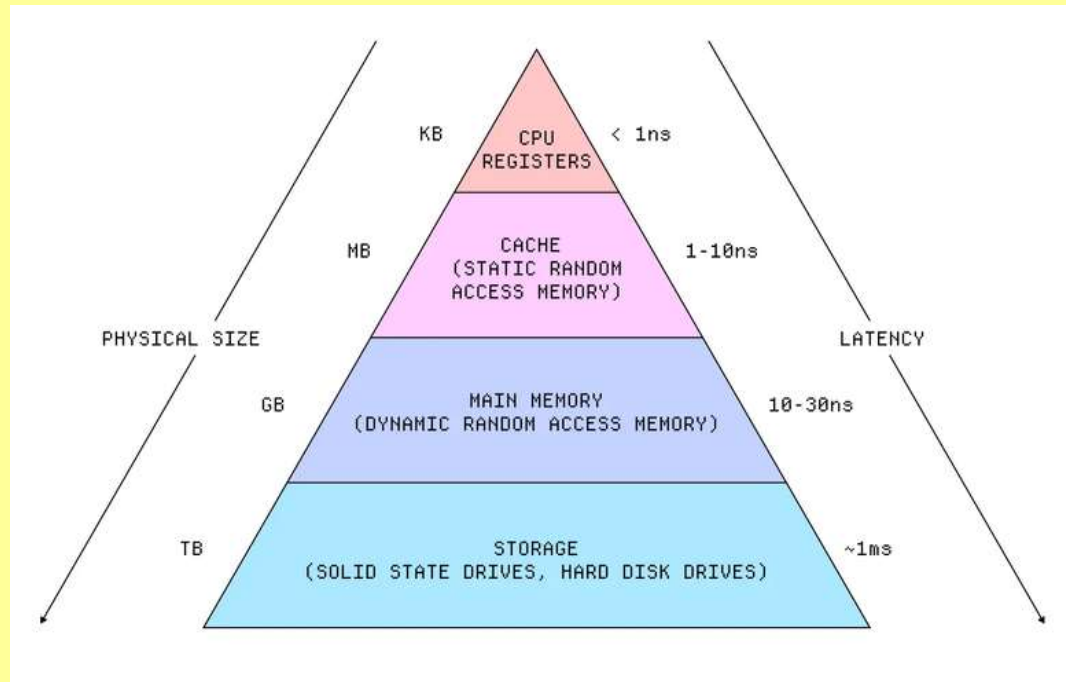
1984

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

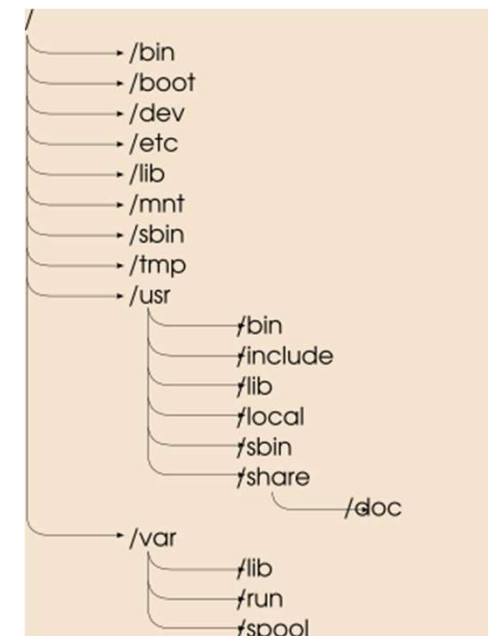
RPG

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.

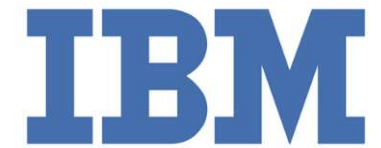
MOVE SUM TO TOTAL-AMOUNT.

CLOSE IN-FILE.



Отделение описания данных от обработки

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

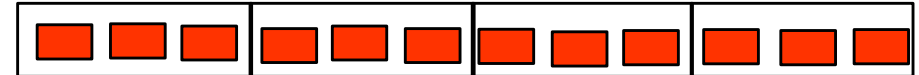
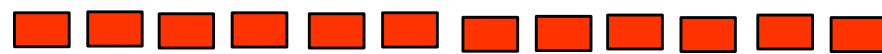
BDAM – прямой (произвольный) метод доступа к данным

Были заложены основы обработки данных
на внешних носителях

Развитие этих методов использовано в первых СУБД

Последовательный метод доступа к данным SAM

Последовательная обработка данных



Не сортированные / сортированные записи

Не сблокированные / сблокированные записи В

Форматы фиксированной F и переменной длины V

блок

b записей

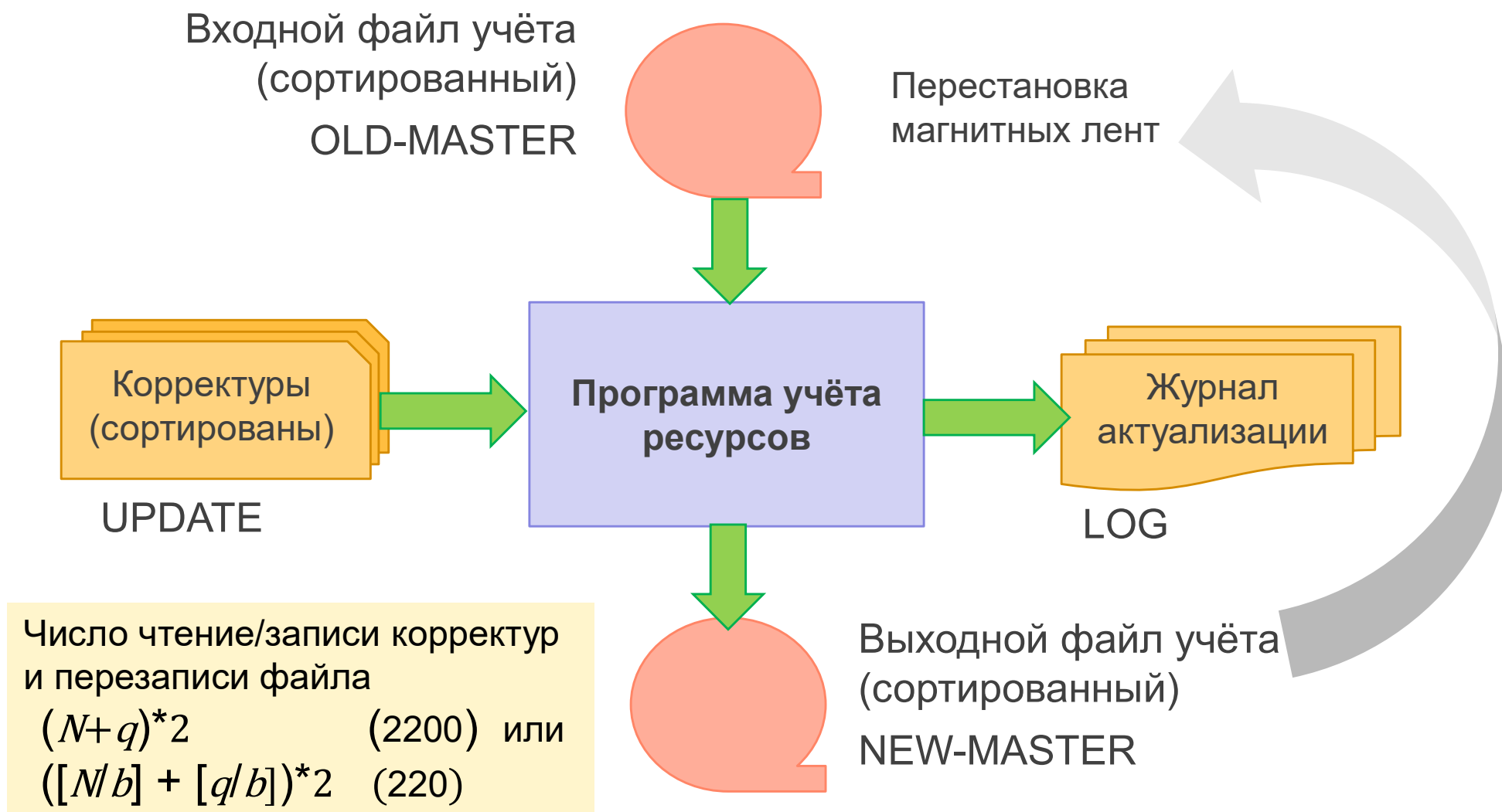
Поиск записей ($N=1000$; $b=10$; $q=100$)

Записи сортированы?	Записи сблокированы?	Ср. число чтений при поиске		Пример
		одной записи	q записей	
нет	нет	$N/2$	$q * N/2$	500; $50 * 10^3$
да	нет	$N/2$	$(N + q)/2$	500; 550
нет	да	$[N/b]/2$	$q * [N/b] / 2$	50; $5 * 10^3$
да	да	$[N/b]/2$	$([N/b] + [q/b])/2$	50; 55

Сортировка была важна при последовательном пакетном поиске.

Сблокированные записи увеличивают производительность

Последовательное (пакетное) обновление записей



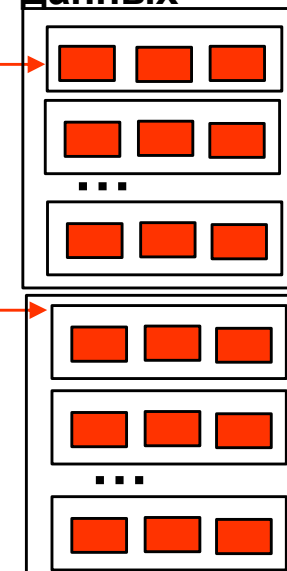
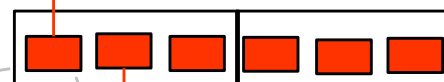
Индексно-последовательный метод доступа к данным ISAM

Комбинация последовательного и прямого доступа к *сортированным* данным

Индекс:
мин / макс значение ключа,
номер блока

Сегменты
данных

Диапазон ключей,
номер 1-ой записи
сегмента



Поиск отдельной записи ($N=1000$; $m=25$; $b=10$)

Записи сблокированы?	Ср. число чтений индекса и данных	Пример
нет	$(m + \lfloor N/m \rfloor)/2$	32,5
да	$(\lfloor m/b \rfloor + \lfloor \lfloor N/m \rfloor / b \rfloor)/2$	3,5

$$m_{\text{опт}} = \sqrt{N}$$

Изменение данных сложнее: от обновления по месту до вставки/ удаления с возможной реорганизацией сегмента и корректировкой индекса

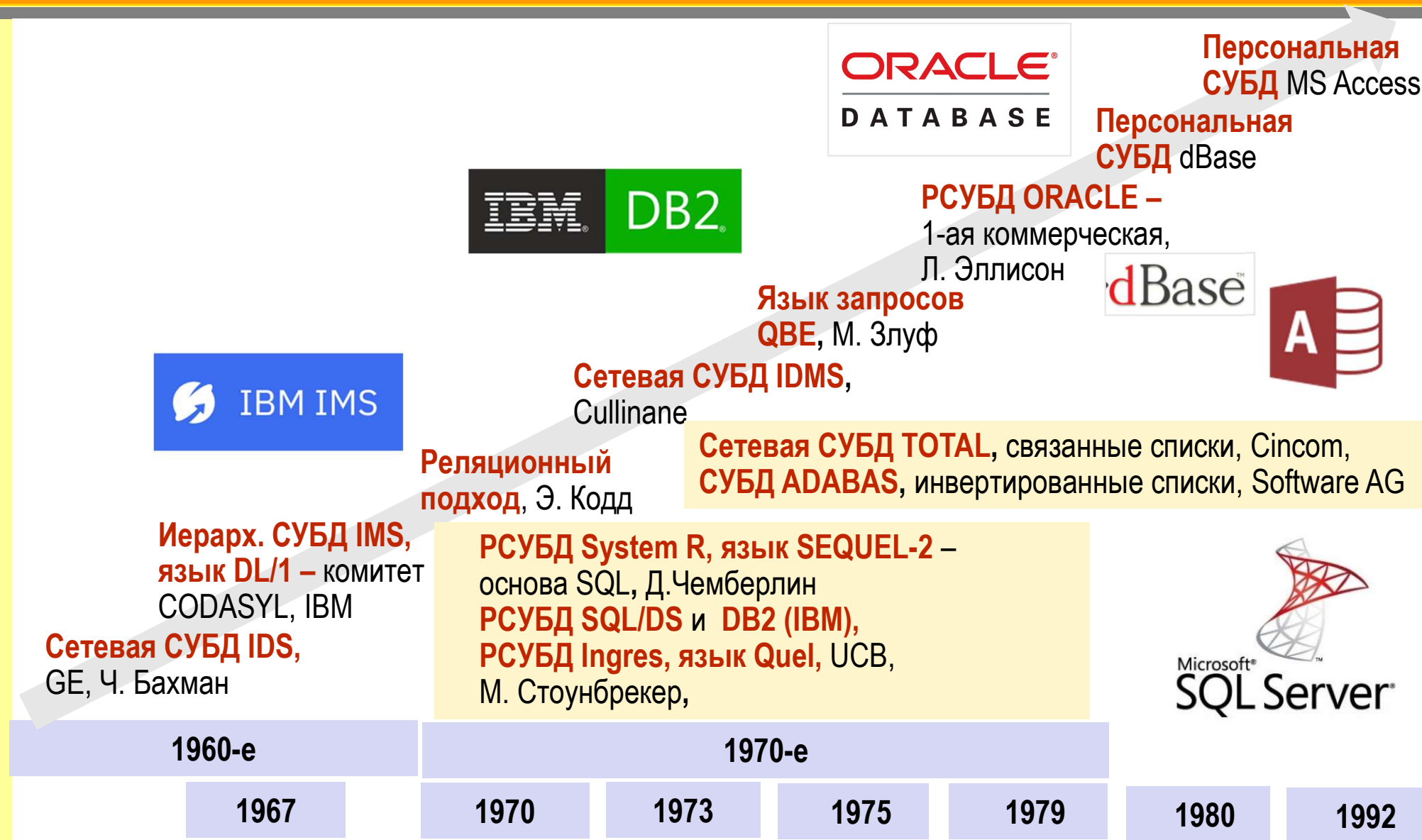
Разработан ещё для пула МЛ, важна сортировка записей.
Актуален для дисков, его модификации используется в СУБД.

Прямой метод доступа к данным

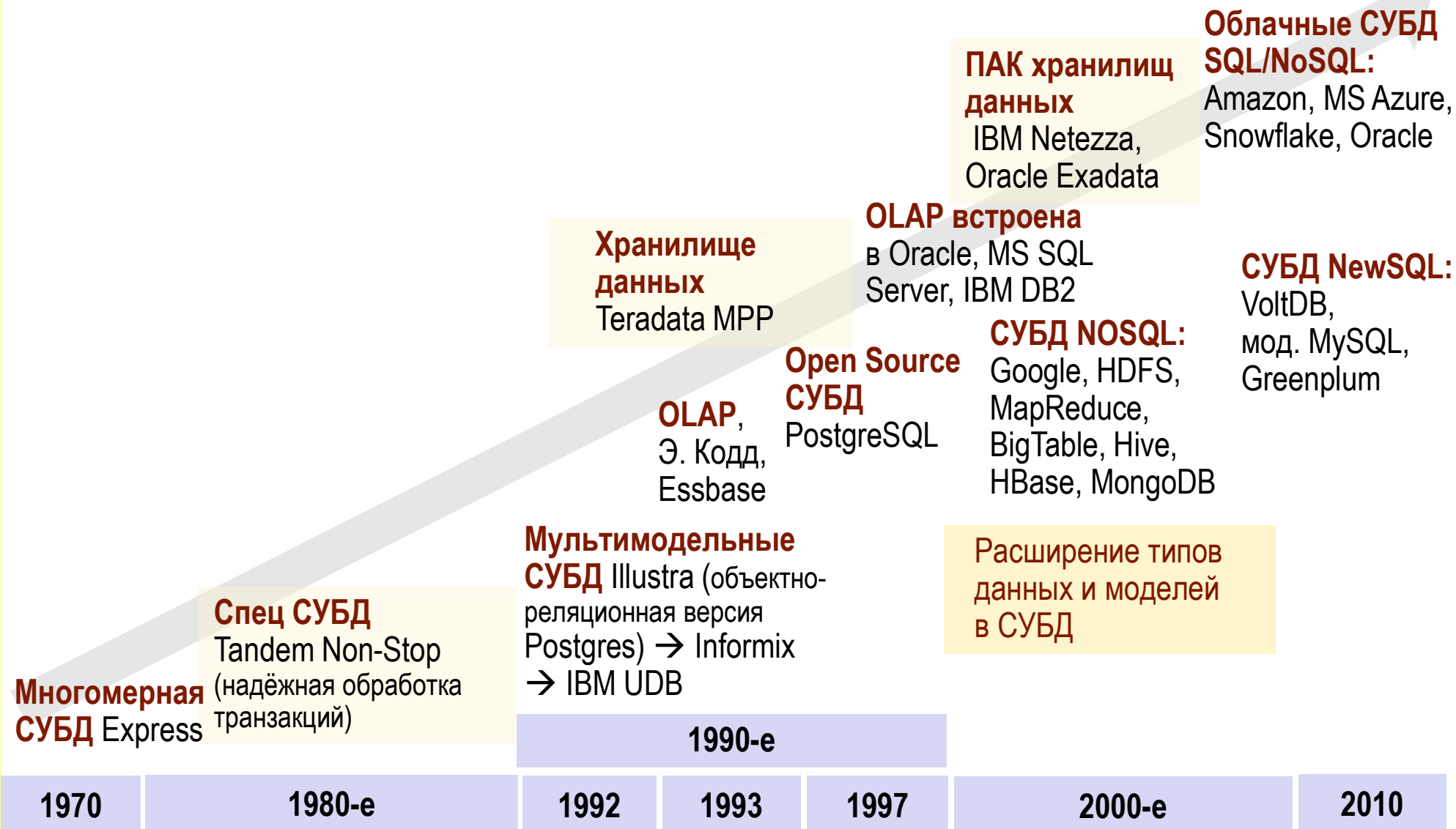
- К данным на дисках возможен прямой доступ
- Проблема соответствия ключа номеру записи
- При наличии такого прямого соответствия – получаем разреженные наборы записей
- Для плотных наборов записей используют разные способы преобразования:
 - Хеширование индексы (кластеры)
 - Древовидные индексы, как правило, В-деревья
 - Битовые индексы (по столбцам)
 - Индекс-таблица = индекс + данные



Появление систем управления БД



Развитие СУБД



Классификация СУБД

Признаки классификации



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 таблица	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

Абстракции модели данных

Иерархические БД	IBM IMS
Сетевые БД	Cullinane IDMS
Реляционные БД (SQL и NewSQL , построчные и колоночные)	Oracle, MS SQL Server, PostgreSQL, MySQL, Clickhouse, MS Access
Темпоральные (изменения во времени)	MS SQL Server 2016+, PostgreSQL расш.
Объектные и объектно-реляционные БД	Db4o, ObjectStore, ObjectDB

NoSQL	БД «ключ-значение»	Redis, Riak KV, OrientDB, Tarantool
	Документные БД	MongoDB, YDB, Енисей
	Графовые БД	Neo4J, Virtuoso, Memgraph
	БД с семейством колонок	Cassandra, HBase, Azure Cosmos DB
	Поисковые (полнотекстовые) БД	Elasticsearch, Solr, Splunk, Sphinx

Oracle, MS SQL, PostgreSQL+PostGIS, Hiperion, Oracle OLAP, MS Analytical Services Sedna, BaseX, MarkLogic InfluxDB, Prometheus, Graphite, Riak TS Kdb, Pinecone, Chroma, Milvus Jackrabbit, ModeShape	Пространственные БД Многомерные БД (OLAP, BI) Native XML БД временных рядов Векторные БД Контентные БД (изображения, аудио, видео, гипертекст)
--	--

Мультимодельные БД

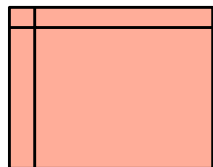
Возможности СУБД с разными моделями данных

<div> <div>Возможность</div> <div>Модель</div> </div>	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	+	—	—	+	+	+

Мультимодельные СУБД

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

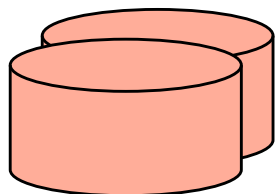
Среды хранения баз данных



Основная память

базы данных in-memory

Redis, SAP HANA, Oracle Database in-memory,
Memcached, Apache Ignite

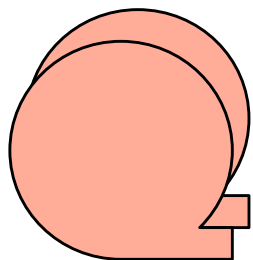


Вторичная память:

магнитные диски, SSD, СХД –

базы данных на дисках

Oracle, MS SQL Server, PostgreSQL,
MySQL, MS Access



Третичная память:

магнитные кассеты, оптические диски –

архивы и резервные копии

*не являются БД и с ними СУБД не работают,
но они сопутствуют БД и СУБД*

Универсальность применения СУБД

Определяется

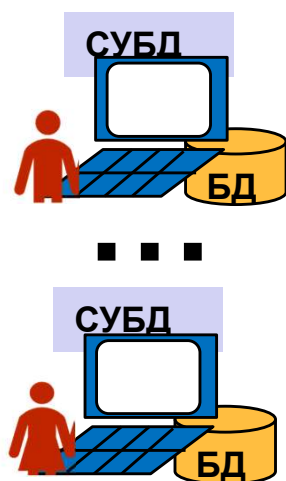
- возможностями
- мультимодельностью
- характером решаемых задач
- требованиями к безопасности и надёжности
- требованиями к производительности

Различают

- универсальные (общего назначения) СУБД
- специализированные СУБД

Коллективное использование БД

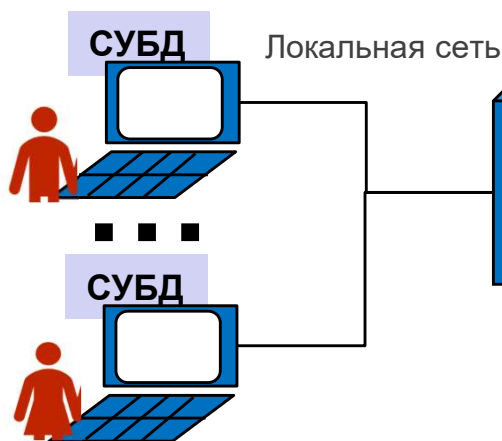
Персональные однопользовательские базы данных и СУБД



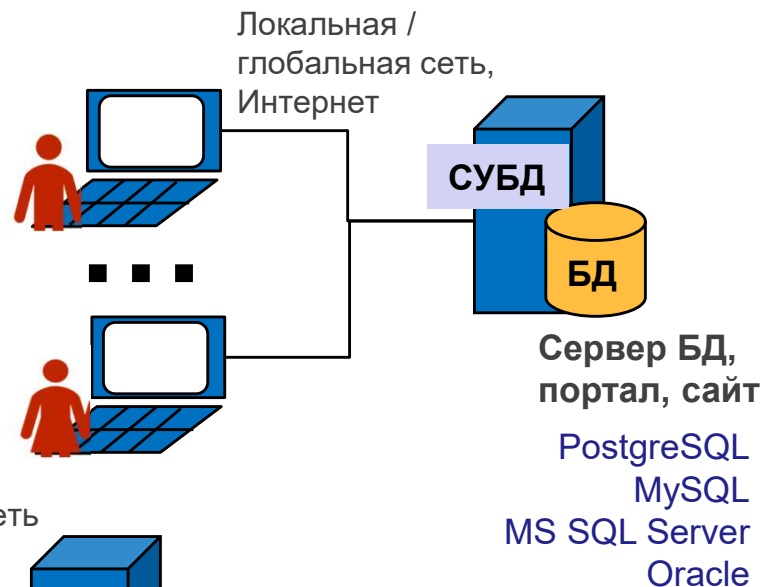
MS Access
FileMaker
Libre Office Base

Авторизация и права
доступа к файлу БД,
блокировка файла при
изменениях

ПК, ноутбуки,
планшеты,
терминалы,
рабочие станции



Коллективные многопользовательские базы данных и СУБД



Авторизация и права
доступа к БД, таблицам,
строкам и столбцам.
Целостность транзакций.
Изоляция обновлений.

Распределённые данные и СУБД

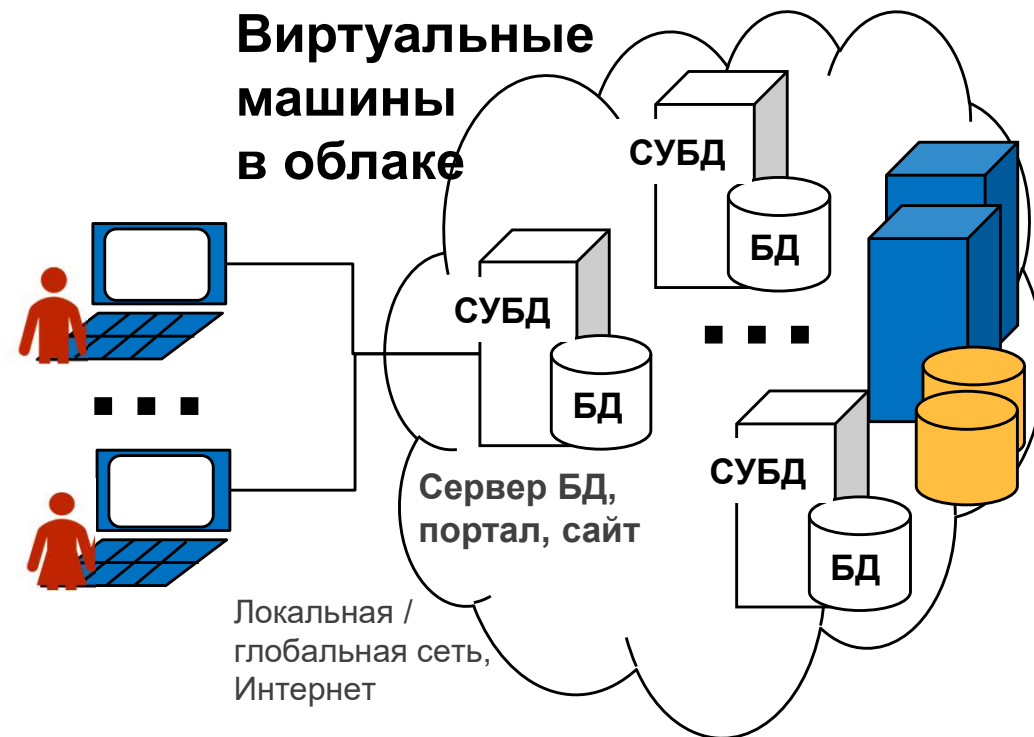
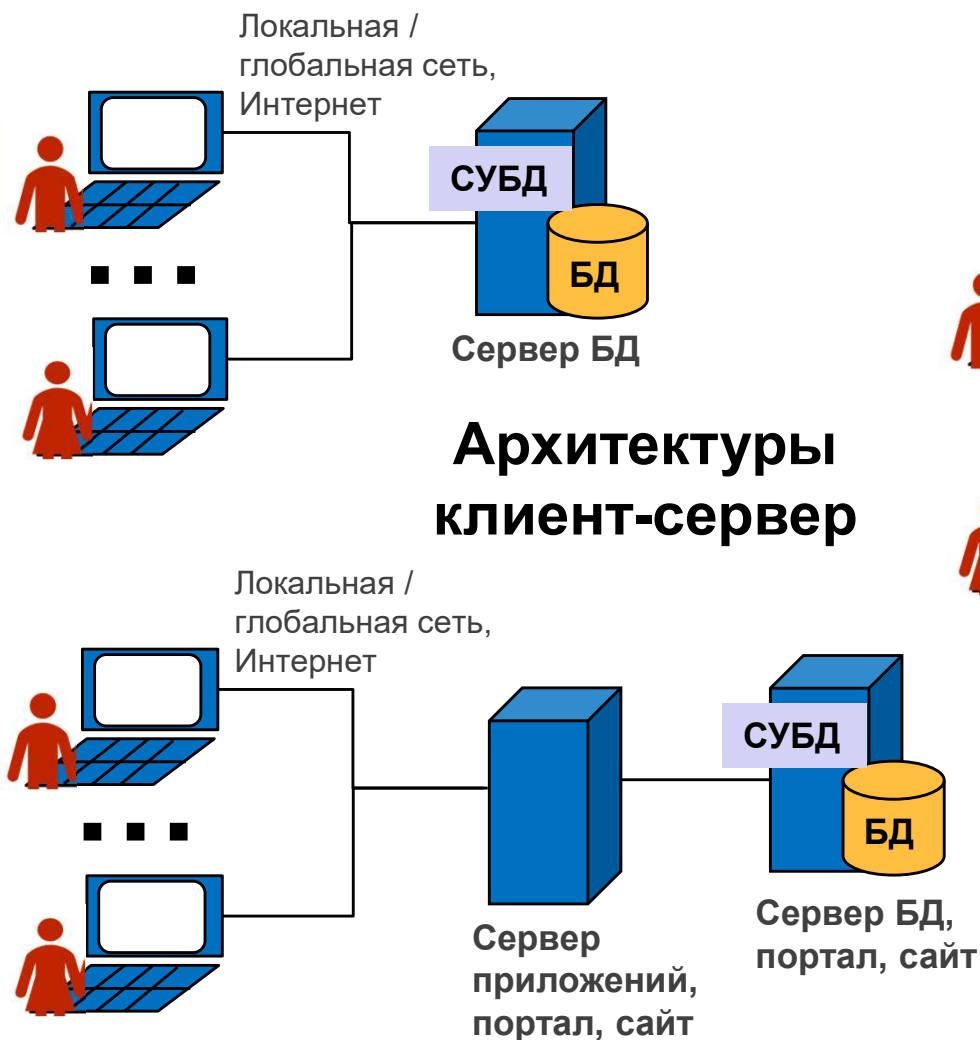
Сосредоточенные БД

- Локальные
- Централизованные

Распределённые БД

- Однородные
- Неоднородные
- Фрагментированные (шардинг)
- Тиражируемые (репликации)

Архитектура приложений



Облачные сервисы:

IaaS – инфраструктура

PaaS – платформа (СУБД)

DbaaS – СУБД + БД

Рабочая нагрузка СУБД

Транзакционная нагрузка OLTP

Высокопроизводительные HP OLTP

Высоконадёжные HA OLTP

Хранилища данных DWH

Поддержка принятия решений OLAP/DSS

Системы искусственного интеллекта AI / ML

Мультимедийная нагрузка

Рейтинг современных СУБД

DB-Engines Ranking

				429 systems in ranking, January 2026			
Rank	DBMS			Database Model	Score		
	Jan 2026	Dec 2025	Jan 2025		Jan 2026	Dec 2025	Jan 2025
1.	1.	1.	Oracle	Relational, Multi-model ⓘ	1237.34	+2.94	-21.42
2.	2.	2.	MySQL	Relational, Multi-model ⓘ	867.52	-0.97	-130.63
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model ⓘ	706.26	-16.26	-92.30
4.	4.	4.	PostgreSQL	Relational, Multi-model ⓘ	666.27	+6.84	+2.86
5.	5.	5.	MongoDB	Multi-model ⓘ	376.74	+4.46	-25.77
6.	6.	6.	Snowflake	Relational	207.79	+5.34	+53.89
7.	7.	7.	Redis	Key-value, Multi-model ⓘ	144.16	+1.68	-9.20
8.	8.	↑ 13.	Databricks	Multi-model ⓘ	141.55	+3.53	+53.70
9.	9.	9.	IBM Db2	Relational, Multi-model ⓘ	112.72	-3.04	-10.25
10.	10.	↓ 8.	Elasticsearch	Multi-model ⓘ	107.15	-2.68	-27.78

12 Open Source СУБД в 2025 году

Database Management system	Data model	Requirements	Best for
MySQL	SQL	High read/write throughput	Web apps, CMS, eCommerce
PostgreSQL	SQL	Complex queries, advanced indexing	GIS, financial services, analytics
MariaDB	SQL	Transactional + analytical workloads	MySQL replacement
MongoDB	NoSQL	Real-time queries, flexible schema	IoT, content management
SQLite	SQL	Lightweight, embedded use	Mobile apps, desktop tools
CockroachDB	SQL	Strong consistency	Multi-region SaaS
Redis	NoSQL	Sub-millisecond latency	Caching, real-time analytics
CouchDB	NoSQL	Offline sync, RESTful API	IoT, distributed mobile apps
Neo4j	NoSQL	Relationship traversal, deep joins	Social networks, fraud detection
FirebirdSQL	SQL	OLTP + OLAP concurrency	Embedded systems, SMB apps
OrientDB	SQL	Fast traversal, flexible schema	Identity management, recommendation engines
Cassandra	NoSQL	High write throughput, linear scalability	Time-series, event

**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