Лекция – 1 данные и информация

Литература

1. От хранения данных к управлению информацией / EMC. - СПб.: Питер, 2010. - 544 с.
2. Введение в облачные вычисления. Курс Интернет-университета информационных технологий. http://www.intuit.ru/department/se/inclouds/
3. Когаловский М.Р. Перспективные технологии информационных систем. -М.: ДМК Пресс, 2008. - 288с. - http://www.iqlib.ru/
4. Марц Н., Уоррен Дж. Принципы и практика построения масштабируемых систем обработки данных в реальном масштабе времени: Учебник. – Вильямс, 2017. – 368 с.
5. Леонов В. Google Docs, Windows Live и другие облачные технологии. - М.: Эксмо, 2012.-304с.
6. Сенько А. Работа с BigData в облаках. Обработка и хранение данных с примерами из Microsoft Azure. – СПб.: Питер, 2019. - 448 с.
7. Гольдштейн Б.С. Инфокоммуникационные сети и системы. - СПб.: БХВ-Петербург, 2019. – 208 с.
8. Лонг Джош, Бастани Кеннет. Java в облаке. Spring Boot, Spring Cloud. Cloud Foundry. – СПб.: Питер, 2019. – 624 с.
9. Скиена Стивен С. Наука о данных: учебный курс.: Пер с англ. – СПб.: ООО «Диалектика», 2020. – 544 с.

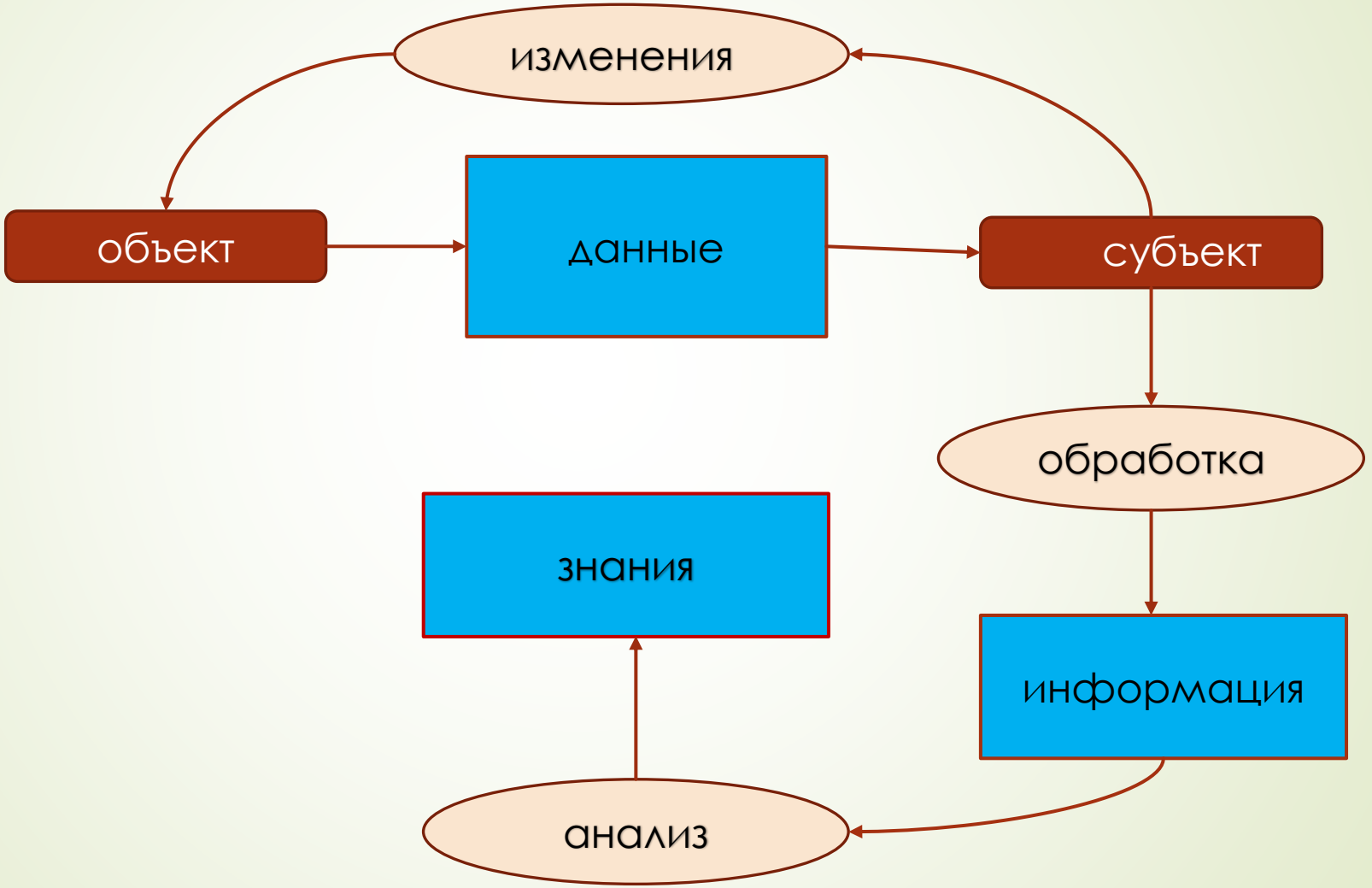
Содержание дисциплины

- Два раздела :
  - Управление данными – данные и информация, жизненный цикл информации и управление жизненным циклом, изучение эволюции систем хранения и передачи данных, структура и функционирование центров обработки данных (ЦОД), виды виртуализации, виртуализованный ЦОД.
  - Облачные технологии – основные технологии облачных вычислений, модели облачных услуг, модели развертывания облака, инфраструктура облака, безопасность в облаке, этапы перехода к облаку.

Данные - набор «сырых» фактов, из которых могут быть сделаны те или иные заключения

- Информация – данные, которым придается некоторый смысл (интерпретация) в конкретной ситуации в рамках некоторой системы понятий. Информация извлекается субъектом из соответствующих данных.
- Знания - совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача, полученное из данных
- Рост цифровой информации привел к так называемому «информационному взрыву»

Формирование информации и знаний



Классы хранения данных

	Характеристика
Базы данных (БД)	именованная совокупность данных, отображающая состояние объектов и их отношения в рассматриваемой предметной области. Характерной чертой баз данных является постоянство: состав и структура данных, необходимые для решения тех или иных прикладных задач, обычно постоянны и стабильны во времени
Система управления базами данных (СУБД)	совокупность языковых и программных средств, предназначенных для создания, ведения и совместного использования БД многими пользователями.
Банк данных (БнД)	система специально организованных данных, программных, языковых, организационных и технических средств, предназначенных для централизованного накопления и коллективного многоцелевого использования данных. база данных и система управления ею (СУБД)
База знаний (БЗ)	представляет собой совокупность БД и используемых правил, полученных от лиц, принимающих решения (*ЛПР*)

Характеристики информационного взрыва

- Мы живем в мире моментальных ответов на запросы и команды
  - Информация чаще всего необходима именно в том месте и в то время, когда она была запрошена
- Возрастает зависимость от быстрого и надежного доступа к информации
- Организациям необходимо хранить, защищать, оптимизировать и использовать информацию, чтобы:
  - Получать конкурентное преимущество
  - Получать новые возможности для развития бизнеса

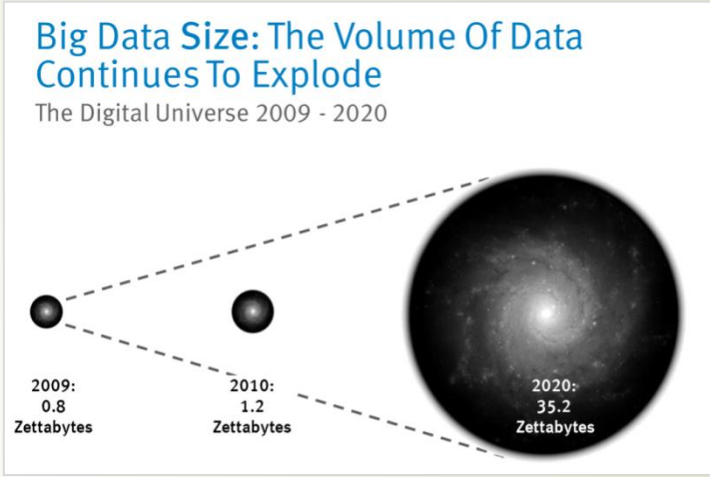
БОЛЬШИЕ ДАННЫЕ (BIG DATA) -

НАБОРЫ ДАННЫХ, РАЗМЕРЫ КОТОРЫХ НЕ ПОЗВОЛЯЮТ ТРАДИЦИОННЫМ ПРОГРАММНЫМ РЕШЕНИЯМ ПОЛУЧАТЬ, ХРАНИТЬ, УПРАВЛЯТЬ И ОБРАБАТЫВАТЬ ИХ ЗА ПРИЕМЛЕМОЕ ВРЕМЯ

- Включают в себя как структурированную, так и не структурированную информацию, сгенерированную различными источниками
- Аналитика больших объемов данных в реальном времени требует новых средств и техник, которые обеспечивают:
  - Высокую производительность
  - Платформы массово-параллельной архитектуры (Massively parallel processing (MPP))
  - «Продвинутые» методы аналитики
- Аналитика больших объемов данных предоставляет возможность транслировать большие объемы данных в правильные решения

Основные характеристики больших данных

1. **Объем данных Data Volume**
   - Возрастёт в 29,3 раза с 2010 по 2020 год (от 1,2 зетабайт до 35,2 зетабайт)
2. **Сложность обработки Processing Complexity**
   - Изменение структуры данных во времени
   - Use cases warranting additional transformations and analytical techniques
3. **Структура данных Data Structure**
   - Возрастает многообразие структур данных для поиска и исследования Greater variety of data structures to mine and analyze



Copyright © 2011 EMC Corporation. All Rights Reserved.

Три "V" и три принципа работы с большими данными

- Набор признаков V V V (volume, velocity, variety — физический объём, скорость прироста данных и необходимости их быстрой обработки, возможность одновременно обрабатывать данные различных типов) был выработан компанией Meta Group в 2001 году с целью указать на равную значимость управления данными по всем трём аспектам.

- В дальнейшем появились интерпретации с четырьмя V (добавлялась veracity — достоверность), пятью V (viability — жизнеспособность и value — ценность), семью V (variability — переменчивость и visualization — визуализация). Но компания IDC, например, интерпретирует именно четвёртое V как value (ценность), подчеркивая экономическую целесообразность обработки больших объёмов данных в соответствующих условиях.

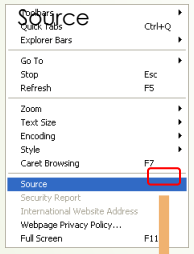
Четыре основных типа структур данных

Структурированные / Structured Data

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1

Полу-структурированные / Semi-Structured Data

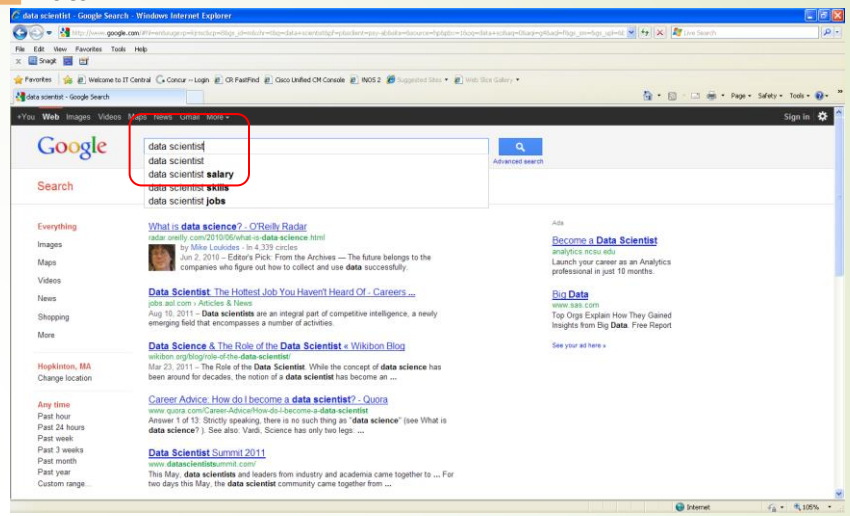
View → Source




```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-trans
<html xmlns="http://www.w3.org/1999/xhtml">

        <head>
                <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
                <META name="y_key" content="859b4020e1c9acec">
                        <link rel="canonical" href="http://www.emc.com/index.htm" />
                                <META NAME="verify-v1" CONTENT="yiZt9V0P4eV0jFdIPeVVIfRP32g4qtwFE0I2UvTMfSU
                <title>EMC - Data Recovery, Cloud Computing, and Storage Hardware</title>
                <META NAME="description" CONTENT="EMC is a leading provider of storage hardware solutions th
 data recovery and improve cloud computing." />
                                <META NAME="keywords" CONTENT="emc,network storage,data recovery,information manage
 software,nas storage,information protection,information management" />
                <!-- Start :stylehseet includues -->
<link rel="stylesheet" href="/_admin/css/styles.css" />
<link rel="stylesheet" href="/_admin/css/styles_nav.css" />
<!--[if IE]>
```

Квази-структурированные / Quasi-Structured Data



http://www.google.com/#hl=en&sugexp=kjrmc&cp=8&gs_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&sclient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aqi=g4&aql=f&gs_sm=&gs_upl=&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=d566e0fbd09c8604&biw=1382&bih=651

Неструктурированные / Unstructured Data

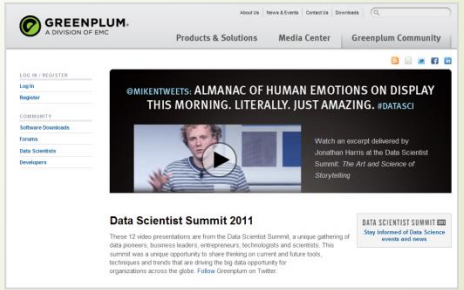
The Red Wheelbarrow, by William Carlos Williams

so much depends
upon

a red wheel
barrow

glazed with rain
water

beside the white
chickens.



Data Repositories, An Analyst Perspective

Островные данные
Изолированные хранилища

- Таблицы и небольшие БД
- выборка зависит от аналитика

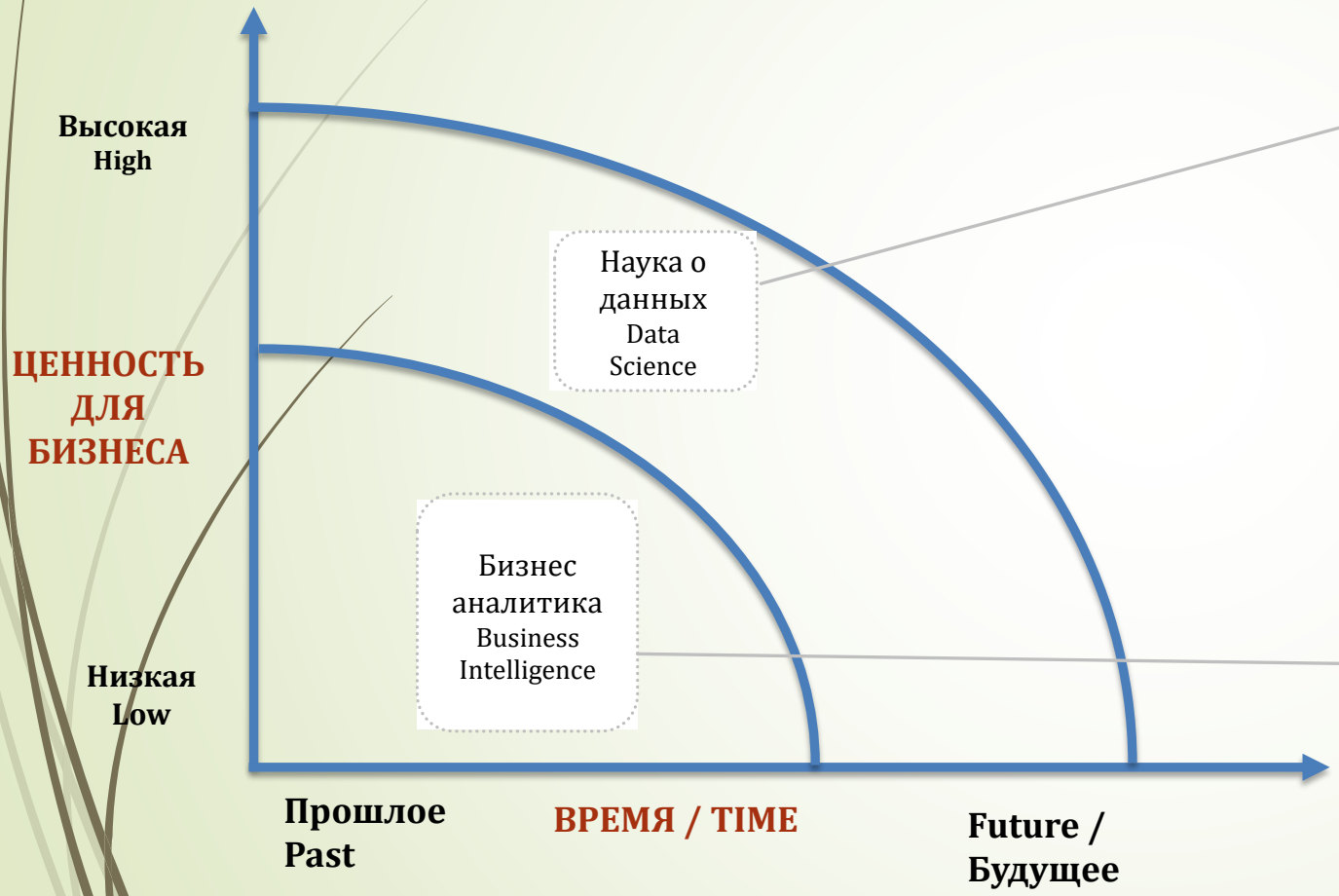
Хранилища данных
Централизованное целевое хранение

- Поддержка BI (бизнес аналитика) и отчетности, но ограничение глубокого анализа
- Зависимость аналитика от технических специалистов
- Большие временные затраты на извлечение данных из многих источников

Аналитическая песочница
Данные различных форматов и источников оптимизированные для анализа

- Обеспечивает высокопроизводительную бизнес аналитику при обработке БД
- Reduces costs associated with data replication into "shadow" file systems
- *“Analyst-owned” rather than “DBA owned”*

Аналитические подходы отвечающие потребностям бизнеса
Бизнес аналитика в сравнении с Наукой о данных



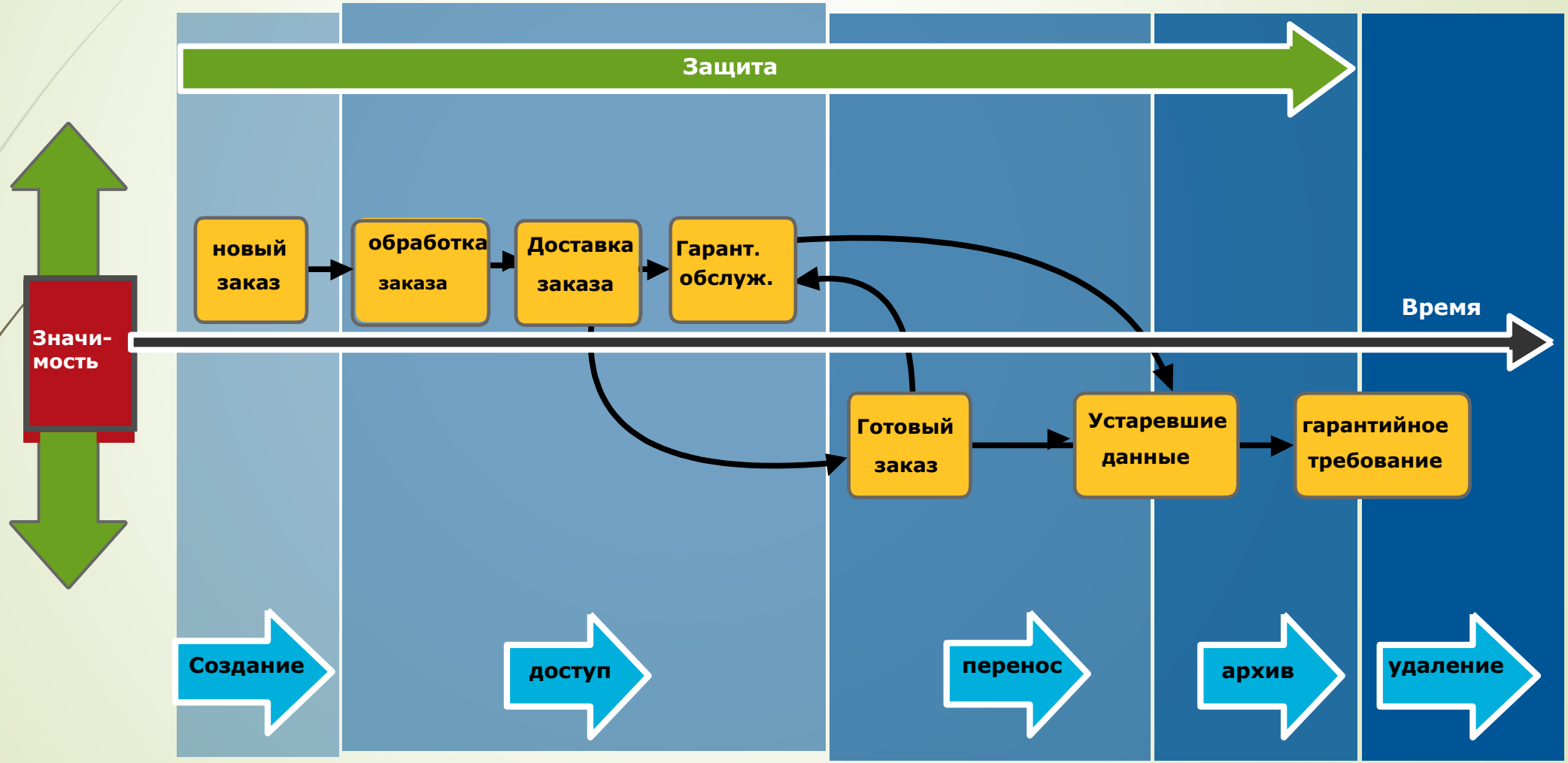
Предсказательная аналитика и добыча данных / Predictive Analytics & Data Mining (Data Science)	
Типовые техники & типы данных	• оптимизация, предсказательное моделирование, forecasting, статистический анализ • структурированные/неструктурированные данные, различные типы источников, very large data sets
Основные вопросы	• What if…..? • What's the optimal scenario for our business ? • What will happen next? What if these trends continue? Why is this happening?

Бизнес-аналитка / Business Intelligence	
Типовые техники & типы данных	• Standard and ad hoc reporting, dashboards, риски, queries, details on demand • Структурированные данные, traditional sources, manageable data sets
Основные вопросы	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

Проблемы управления информацией

- **Взрывной рост объема информации**
- **Возрастающая связность информации**
- **Изменение ценности информации для бизнеса**

Изменение ценности информации: жизненный цикл бизнес-информации



Необходимость управления ЖЦИ

- Обеспечивает эффективное использование имеющихся ресурсов
- Упрощает менеджмент
- Упрощает создание резервных копий и восстановление данных
- Помогает обеспечивать соответствие требованиям законодательства
- Снижает общую стоимость СХД

Управление жизненным циклом информации

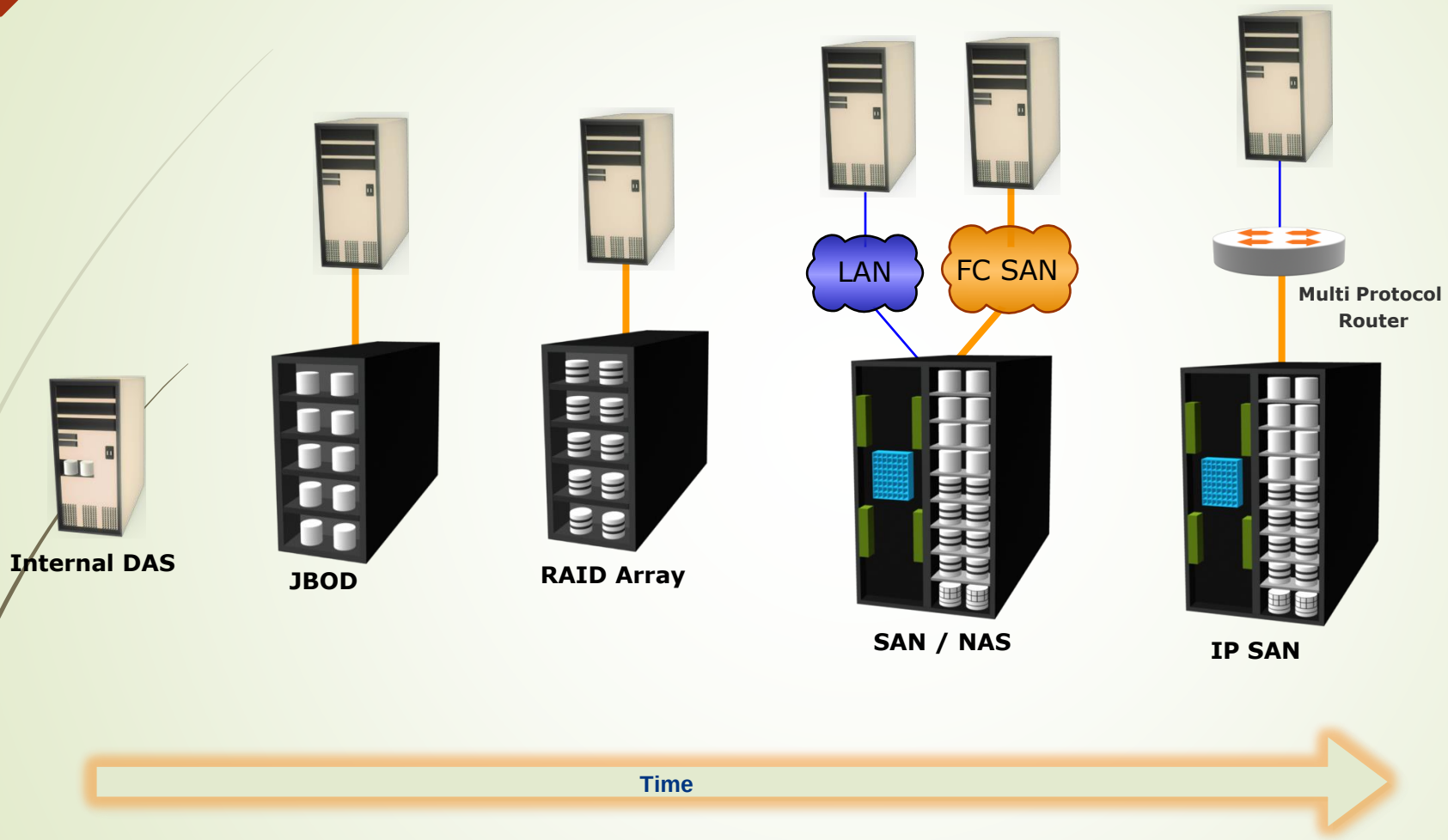
(ЖЦИ), основанное на выборе политики инфраструктуры в зависимости от объема и особенностей информации

Системы хранения данных

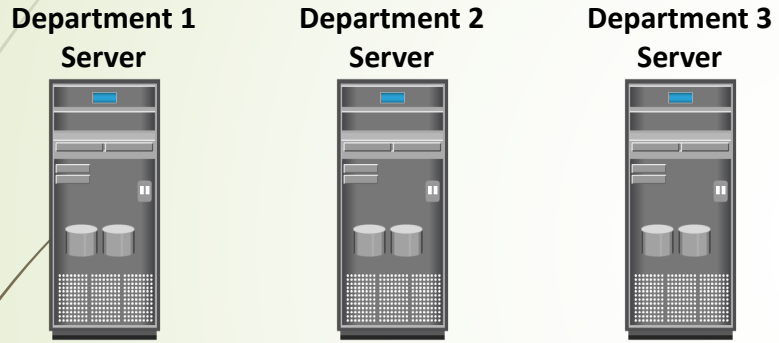
Системы хранения данных (Storage) СХД

- Хранят данные, сгенерированные отдельными пользователями и организациями
  - Обеспечивают доступ к данным для дальнейшей обработки
- Примерами устройств хранения данных являются:
  - Карта памяти в мобильном телефоне или цифровой камере
  - DVD - диск, CD-диск
  - Дисковые устройства
  - Дисковые массивы
  - Ленты

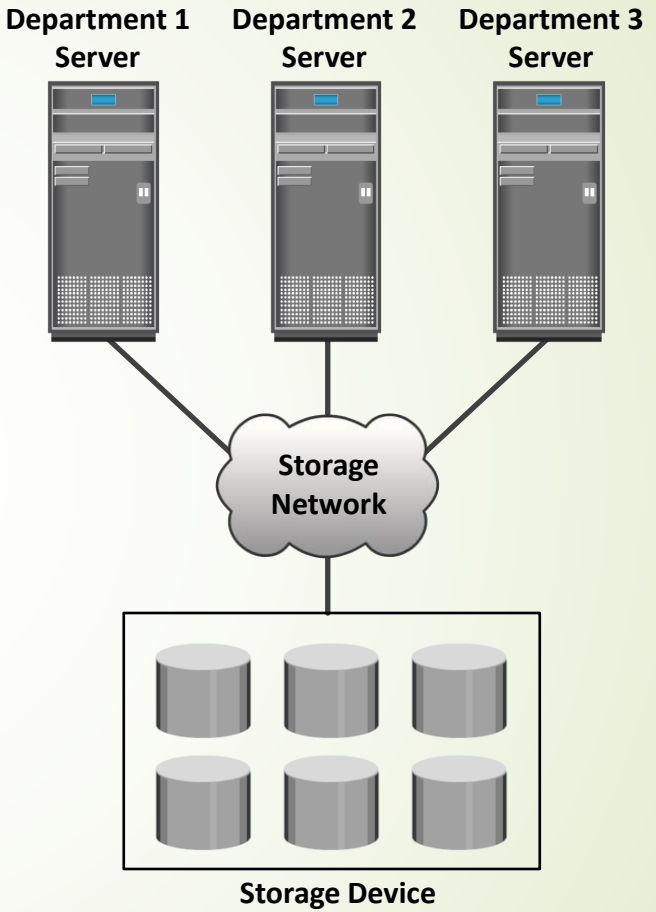
Технологии хранения и эволюция архитектур



Эволюция архитектур хранения данных



Серверно-ориентированная (Server-centric) архитектура хранения данных

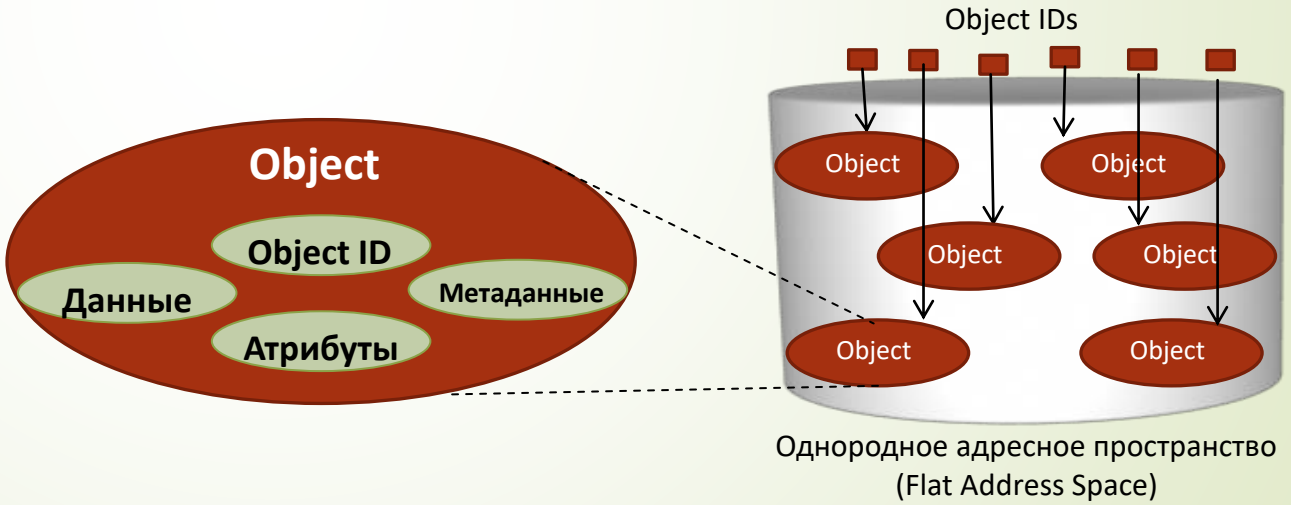


Информационно – ориентированная (Information-centric) архитектура хранения данных

Объектные (Object Based) и унифицированные (Unified) системы хранения данных

Объектное хранение данных

- Объектное хранение данных совмещает данные и метаданные для создания «объекта»
- При объектном хранении данные сохраняются в однородном адресном пространстве (flat address space)
- Каждый объект определяется уникальным ID (Object ID) с использованием хэш-функции



- При объектном хранении данных взаимодействие по http используется в качестве стандартного интерфейса
  - SOAP и REST являются протоколами, которые используются при взаимодействии на объектном уровне в облаке
- Simple Object Access Protocol (SOAP) используется для взаимодействия между участниками в распределенном окружении
  - Использует фреймворк Extensible Markup Language (XML)
- Representational State Transfer (REST) используется для получения информации от сайта, посредством просмотра страниц сайта

Почему объектное хранение?

- Рост объема неструктурированных данных
  - SAN обладает высокой масштабируемостью и поддерживает доступ к данным на блочном уровне
    - Не подходит для разделения доступа к данным
  - NAS является хорошим вариантом для приложений, которым необходимо разделять доступ к данным
    - Ограниченная масштабируемость ввиду иерархической структуры
  - Объектный подход потенциально исключает ограничения SAN и NAS
    - Высокая масштабируемость с возможностями разделяемого доступа к данным

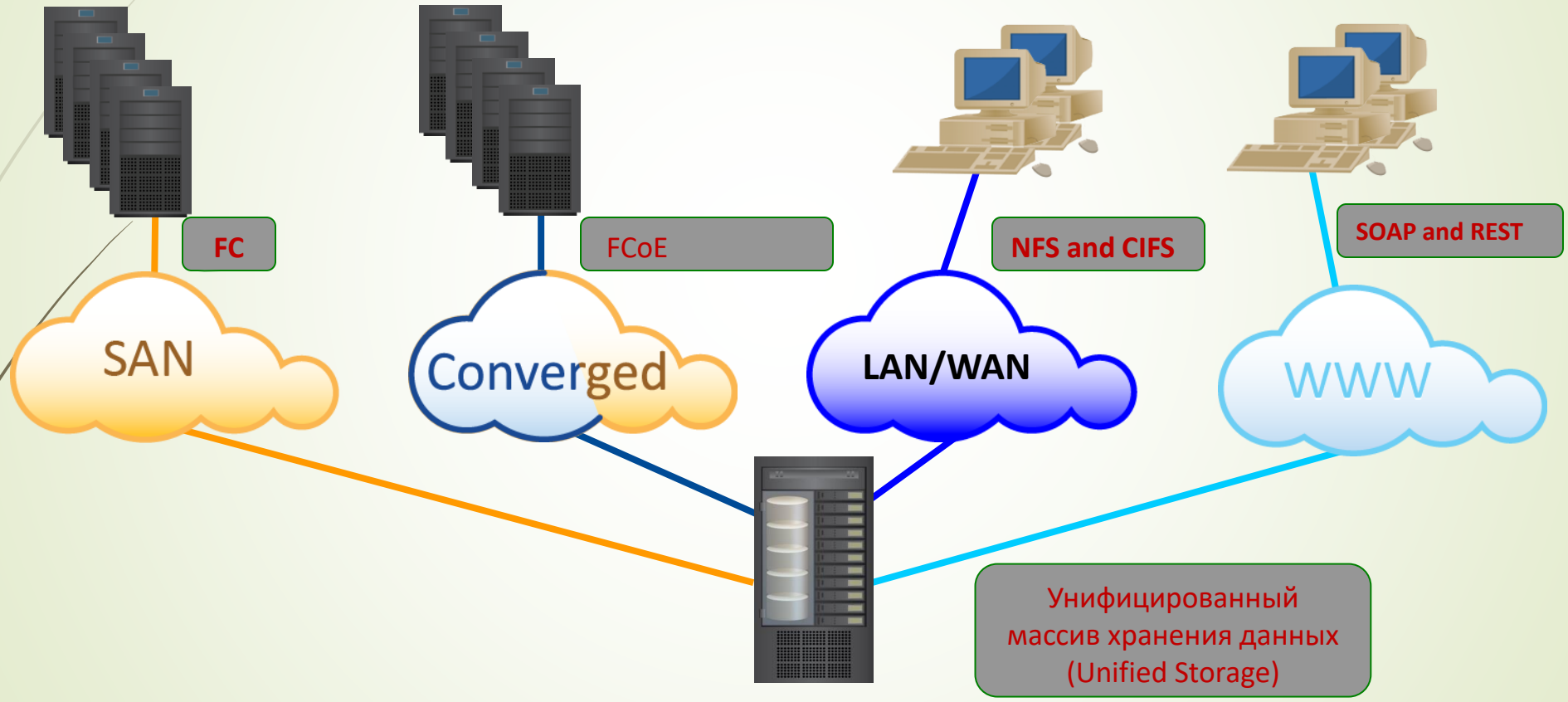
Copyright © 2011 EMC Corporation. All Rights Reserved.

Преимущества объектного хранения данных

- Автоматизирует и упрощает процесс управления инфраструктурой хранения данных
- Гарантирует целостность данных
- Гарантирует соответствие политикам и пригодность для аудита
- Обеспечивает упрощенную миграцию данных
- Реализует самовосстановление
- Позволяет реализовать интеллектуальную репликацию
- Позволяет гибкую масштабируемость

Унифицированное хранение данных

- Обеспечивает консолидированный интерфейс управления для NAS, SAN, iSCSI, FCoE, и объектных технологий

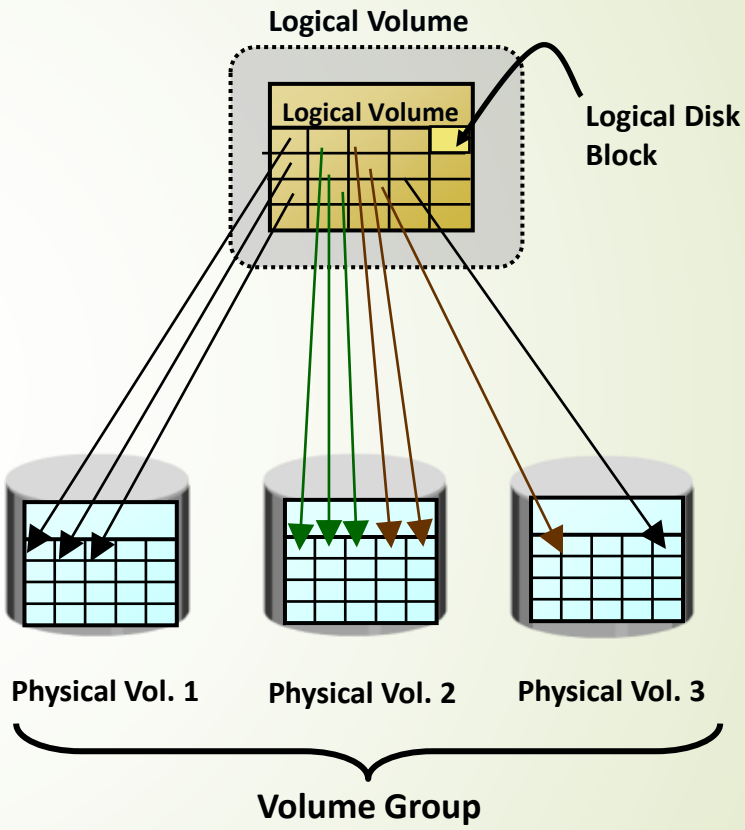


Преимущества систем унифицированного хранения данных

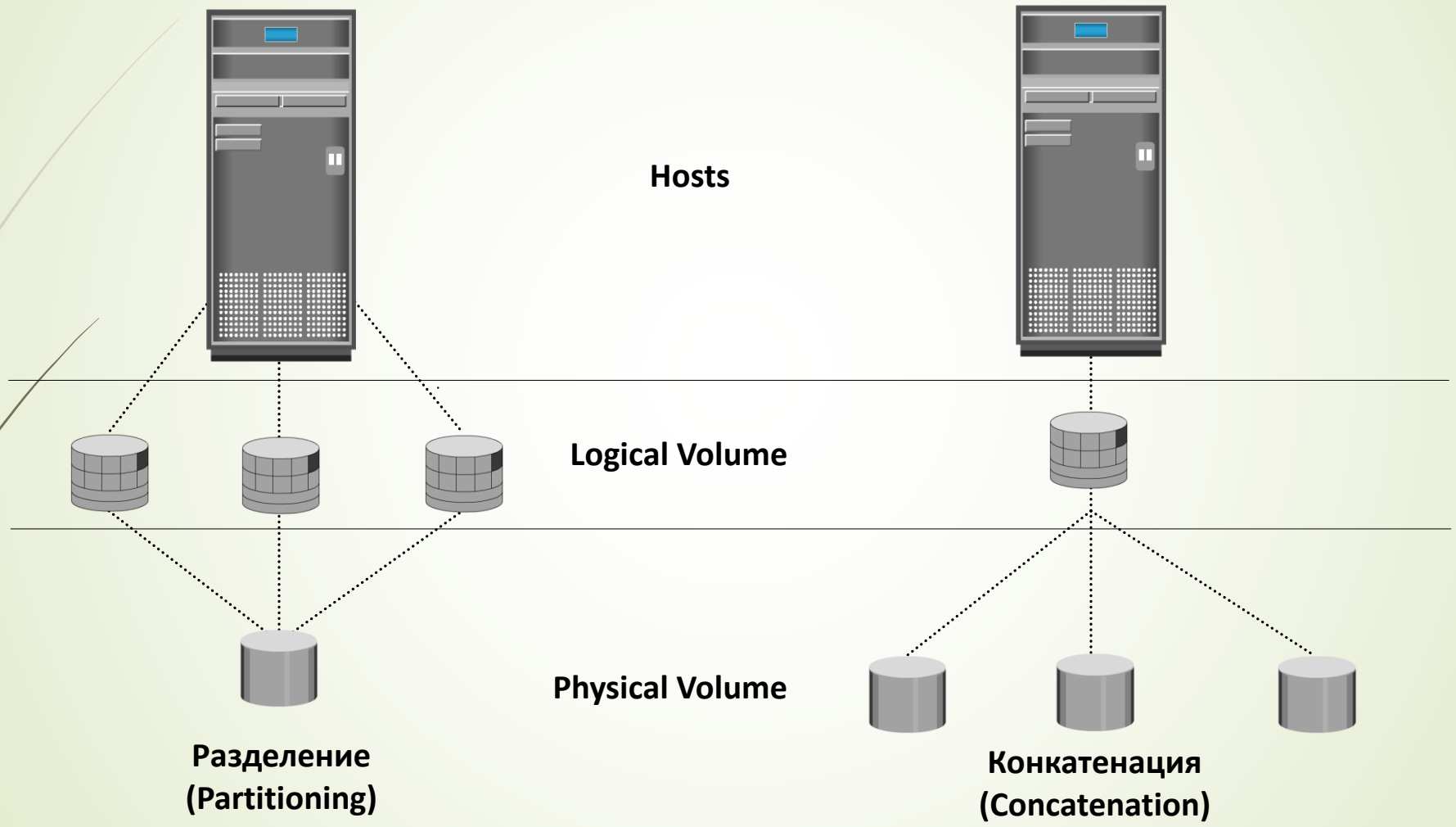
- Обеспечивает консолидированное мультипротокольное хранение
  - Файловое: NFS, CIFS
  - Блочное: iSCSI, FC, FCoE
  - Объектное: REST, SOAP
- Упрощает администрирование
  - Интегрированный интерфейс управления
- Сокращает стоимость хранения, энергопотребления и охлаждения, а также занимаемое пространство
- Обеспечивает высоко масштабируемую архитектуру

Менеджер логических томов - Logical Volume Manager (LVM)

- Отвечает за создание и контроль логических устройств хранения на уровне хоста
  - Физический вид устройства хранения конвертируется в логический вид
  - Логические блоки данных отображаются на физические блоки данных
- Один или несколько физических томов формируют группу томов (Volume Group)
  - Менеджер логических томов воспринимает группу томов и управляет ей как единым целым
- Логические тома создаются из группы томов



Пример менеджера логических томов: Разделение (Partitioning) и Конкатенация (Concatenation)



RAID - Техника, которая объединяет несколько дисковых устройств в логическую единицу (RAID set) с целью обеспечения улучшенной защиты и/или производительности

- Дисковое устройство обладает ограниченной производительностью ввиду наличия механических компонент
- Отдельно взятый диск обладает заранее определенным предположительным временем жизни, измеряемым в среднем времени между сбоями (Mean Time Between Failures, MTBF):
  - Например, если MTBF диска = 750 000 часов, и в массиве находится 1000 таких дисков, MTBF всего массива = 750 часам (750 000/1000)
- RAID был придуман для того, чтобы справиться с данным проблемами

ИТОГИ

Основные понятия, раскрытые в данной лекции:

- Данные и информация
- Типы данных
- Большие данные
- Жизненный цикл информации
- Эволюция архитектур хранения данных
- LVM и RAID