

Что такое большие данные?



DESIGNER.HIPSTER.COM

"Those 'Big Data' jokes just... never get old"

Что такое большие данные?

Data is the new oil

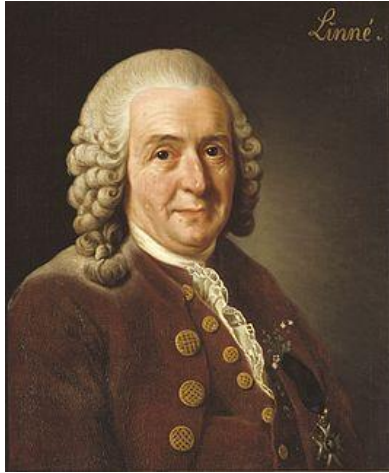


Для чего нужны большие данные?



История больших данных

Предыстория



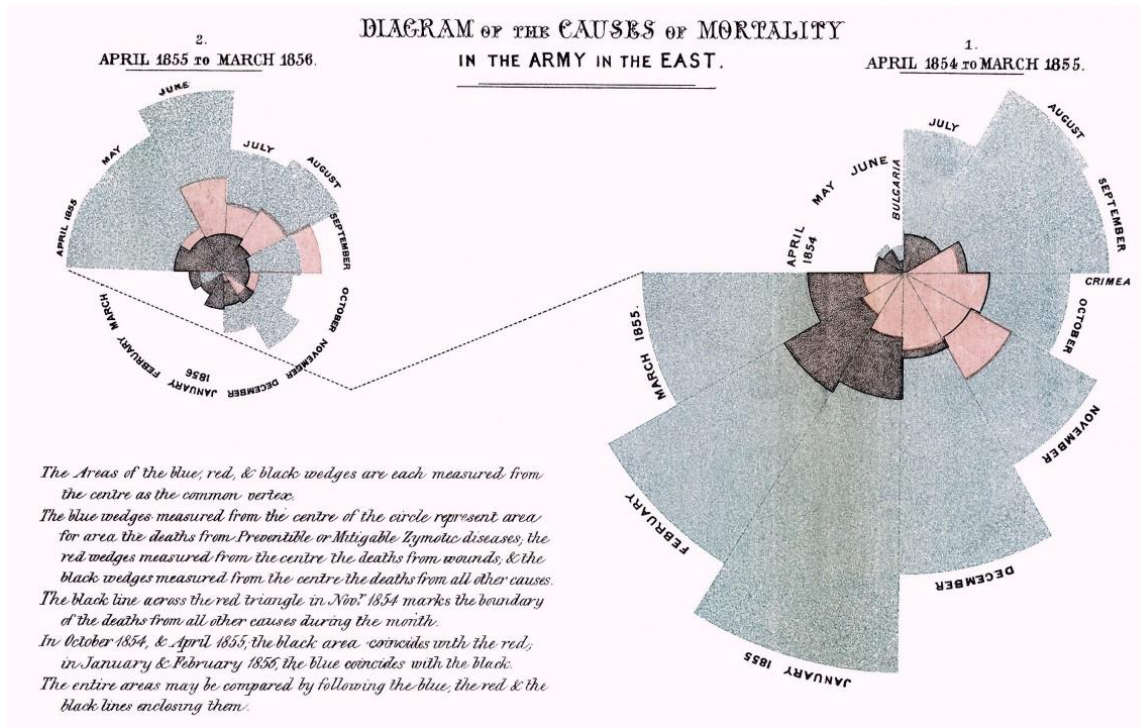
В 1820 – 1840 годах впервые в истории начали появляться большие наборы числовых данных. Этот процесс называли “лавиной чисел”.

Один из первых источников – биологические данные.

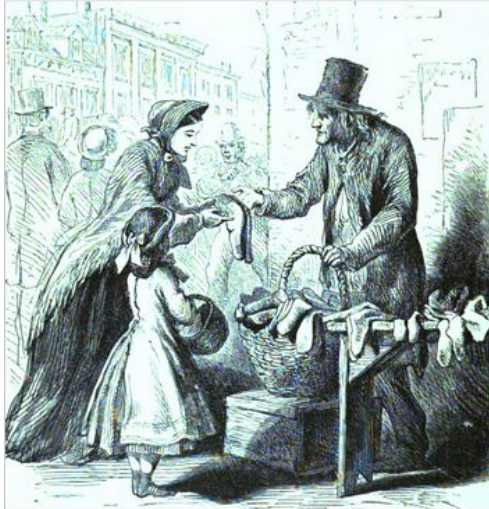
Повлиял Карл Линней (1707-1778), создавший таксономию растений и животных.

Развивались библиотечные технологии, появлялись картотеки. Росли данные, собранные в социологических переписях. Появлялись обширные геологические, антропологические данные.

Визуализация данных



«Диаграмма причин смертности в армии на Востоке» работы Флоренс Найтингейл.



*Cyclopædia of Commercial
and Business Anecdotes*

Richard Miller Devens, Sinclair Hamilton
Collection of American Illustrated Books

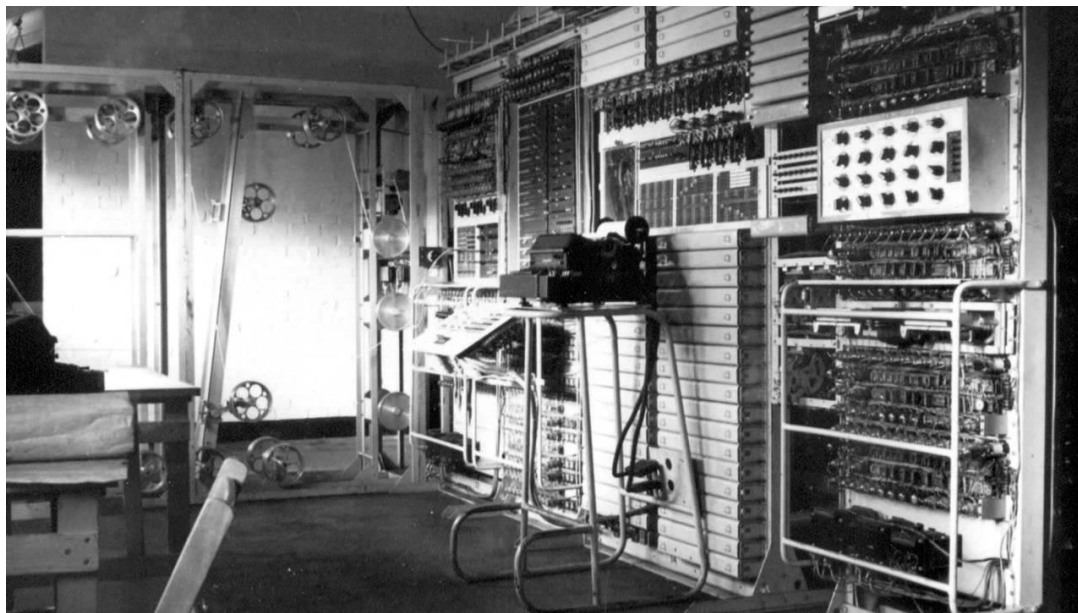
В 1865 году появился термин
Business Intelligence. Его впервые
употребил профессор Ричард
Миллер Девинс.

Под этим термином он
подразумевал использование
анализа
данных для успеха в бизнесе.

Табулятор – первое устройство для обработки больших объемов информации. Было изобретено Германом Холлеритом в 1881 году. Оно использовалось для обработки перфокарт с данными о переписи населения США (1890) и России (1897). В случае ручной обработки данных потребовалось бы несколько лет.



Первые электронные устройства, осуществлявшие анализ данных, появились во время Второй мировой войны. Они поначалу служили для дешифровки сообщений противника. На рисунке – британская машина для дешифровки Colossus.



Первые хранилища данных появились в 1950-х годах. Этот ленточный накопитель компьютера Bendix G-15 относится примерно к 1956 году. Стоимость компьютера составляла \$60,000 (\$500,000 на современные деньги).



Блок UNIVAC 1540,
использовался в середине
1960-х годов,
весил около 1000 фунтов и
имел два семидорожечных
ленточных накопителя на 7
мегабайт.
Предназначался для работы
с мейнфреймом модели
1219-B.



Sony SMC-70
Первый компьютер (1982),
принимавший 3,5-дюймовую
гибкую дискету (1.44Mb,
поначалу – 720kb),
выпущенную в 1981 году.



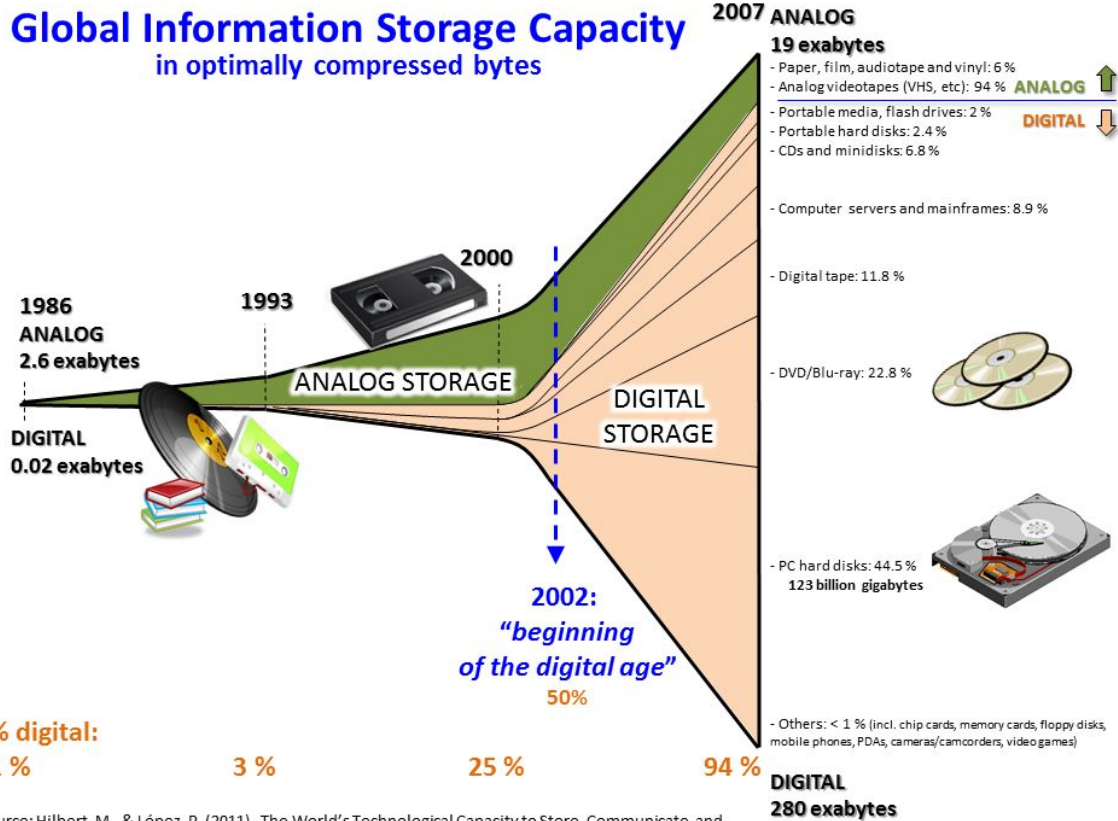
Первый CD (компакт-диск)
появился в 1982 году, а
первый CD-R впервые был
напечатан в 1988
компаниями Philips и Sony.



Современный этап (1993 - 2018 гг.)

- Появились новые понятия: машинное обучение, наука о данных, глубокое обучение
- мощность компьютеров стала достаточной для анализа данных
- для обучения нейронных сетей стали использовать графические процессоры (ускорение обучения в несколько раз)
- Появилось множество данных за счет распространения Интернета
- В 2010-х – развивается новый источник данных – мобильный Интернет

Оцифровка данных



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

Единицы информации

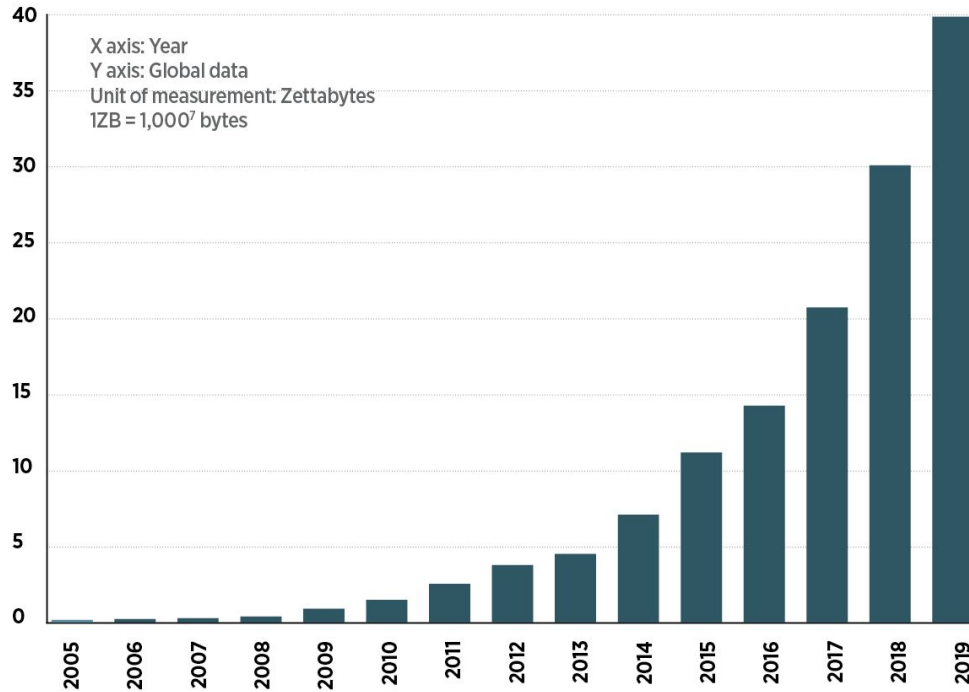
Medida	Simbologia	Equivalencia	Equivalente en Bytes
byte	b	8 bits	1 byte
kilobyte	Kb	1024 bytes	1 024 bytes
megabyte	MB	1024 KB	1 048 576 bytes
gigabyte	GB	1024 MB	1 073 741 824 bytes
terabyte	TB	1024 GB	1 099 511 627 776 bytes
Petabyte	PB	1024 TB	1 125 899 906 842 624 bytes
Exabyte	EB	1024 PB	1 152 921 504 606 846 976 bytes
Zetabyte	ZB	1024 EB	1 180 591 620 717 411 303 424 bytes
Yottabyte	YB	1024 ZB	1 208 925 819 614 629 174 706 176 bytes
Brontobyte	BB	1024 YB	1 237 940 039 285 380 274 899 124 224 bytes
Geopbyte	GB	1024 BB	1 267 650 600 228 229 401 496 703 205 376 bytes

1 Pb – 1 Петабайт (1024Tb)



Рост больших данных

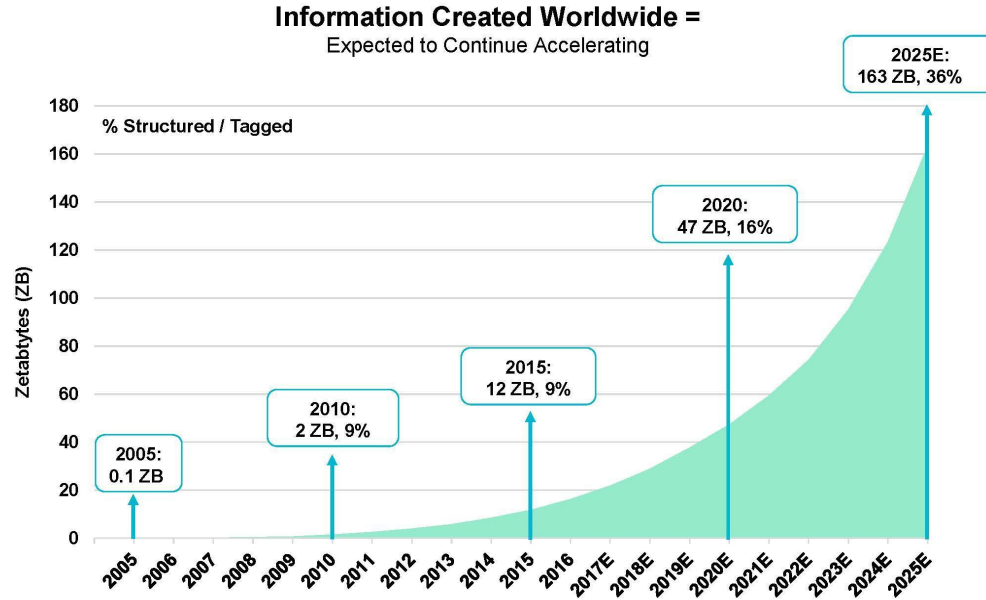
DATA GROWTH



Note: Post-2013 figures are predicted. Source: UNECE

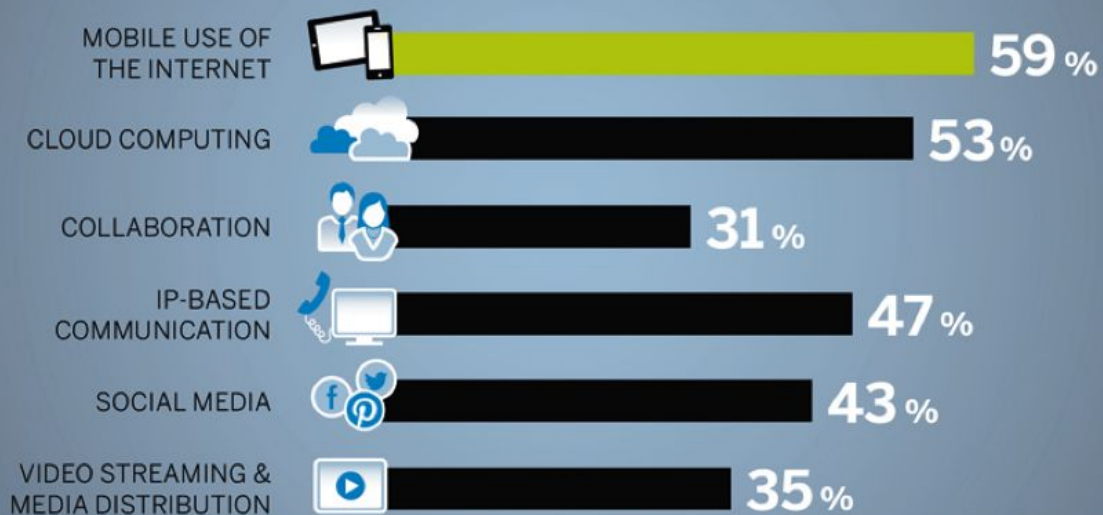
Рост больших данных

Data Volume Growth Continues @ Rapid Clip...
% Structured / Tagged (~10%) Rising Fast...



Факторы роста больших данных

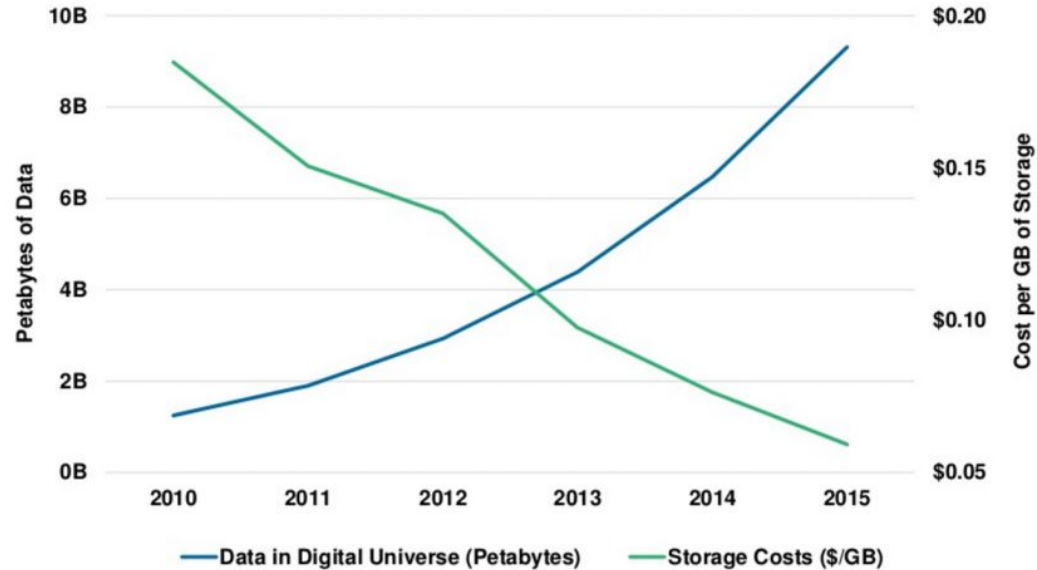
WHAT ARE THE MOST IMPORTANT FACTORS DRIVING THE GROWTH OF DATA GLOBALLY?



<http://blog.3clogic.com/topic/reporting>

Падение стоимости носителей данных

Data in Digital Universe vs. Data Storage Costs, 2010 – 2015



Рынок больших данных

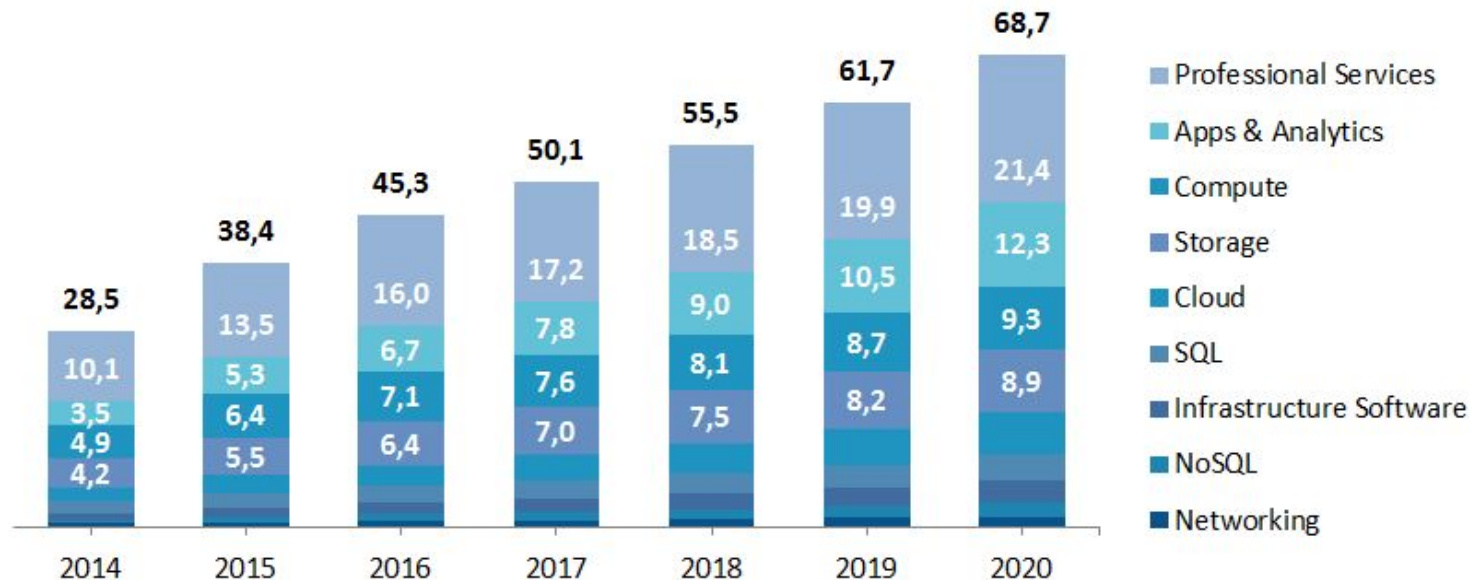
Объем рынка Big Data Германии (млн евро)



Источник: *Experton Group*

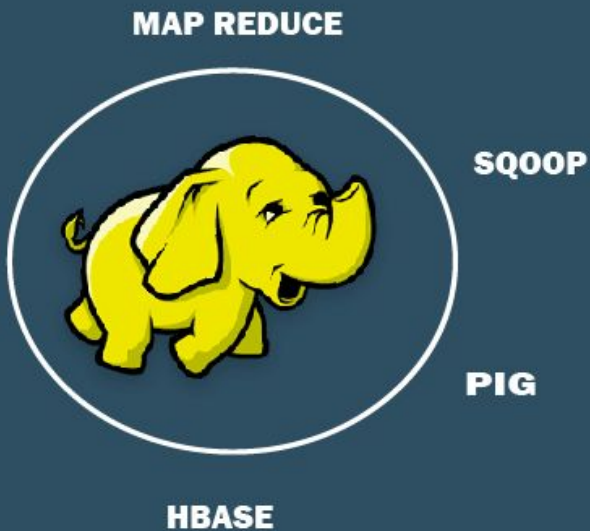
Рынок больших данных

Объем рынка Big Data по подтипам (млрд долл. США)



Источники: Wikibon, IPOboard

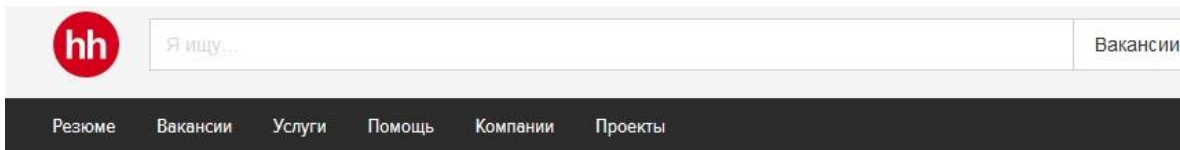
Инженер больших данных



Spark

CERTIFIED
BIG DATA
HADOOP
EXPERT

Пример вакансии



Data Engineer

от 160 000 руб. на руки

Star-staff ✓

Москва



Откликнуться



Требуемый опыт работы: 1–3 года
Полная занятость, удаленная работа

Ищем **Data Engineer** в международную компанию. Основной проект - разработка платформы для интеграции участников дистрибуции игрового контента: игровых платформ, магазинов, банков и т.д.

Задача: "с нуля" разработать систему для сбора, обработки и хранения статистики по поведению пользователей, информации, связанной с авторизацией, финансовыми операциями, etc. Например, для выявления аномального поведения пользователей.)

Пример вакансии

Задача:"с нуля" разработать систему для сбора, обработки и хранения статистики по поведению пользователей, информации, связанной с авторизацией, финансовыми операциями, etc. Например, для выявления аномального поведения пользователей.)

Какие навыки помогут работать у нас:

- Опыт в построении нагруженных серверных решений
- Практический опыт работы с Kafka и Kafka Streams или KSL
- Знание Go и/или Java
- Опыт работы с колоночными базами **данных**. (например, Yandex Clickhouse)
- Практический опыт работы с задачи агрегации и обработки **данных** из разных источников

Будет плюсом:

- Опыт работы с Elasticsearch, Mongo, PostgreSQL
- Опыт работы с Kubernetes
- **Опыт работы с HDFS/Hadoop**
- Практический опыт в использовании kubernetes/облачной инфраструктуры

Что мы предлагаем:

- Работа в **большом** международном проекте по дистрибьюции игр
- Возможно работать удаленно или в офисе (м. Дмитровская)