

# Облачные платформы для больших данных

# Облако – Cloud

- Облако: интернет на схеме сети
- Скрывает сложную инфраструктуру
- Cloud computing
- «Облачные вычисления» / сервисы / аналитика ...

# Работа с программой

- Доступ к программе
  - Локально
  - Удалённо через терминал
  - Облако
- Размещение программы
  - Установка - Install
  - Переносимая - Portable
  - Virtual Machine
  - Container / Docker
  - Оркестрация
  - Бессерверные вычисления

# Определения

- [https://ru.wikipedia.org/wiki/Ложные друзья переводчика](https://ru.wikipedia.org/wiki/Ложные_друзья_переводчика)
- <https://en.wiktionary.org/wiki/computing>
- <https://en.wiktionary.org/wiki/cloud>
- [https://en.wiktionary.org/wiki/cloud computing](https://en.wiktionary.org/wiki/cloud_computing)
- <https://www.merriam-webster.com/dictionary/cloud%20computing>
- <https://dictionary.cambridge.org/dictionary/english/cloud-computing>
- [https://en.wikipedia.org/wiki/Cloud computing](https://en.wikipedia.org/wiki/Cloud_computing)

# Сервер (Server)

- <https://en.wiktionary.org/wiki/Server>
- <https://ru.wiktionary.org/wiki/сервер>
- <https://www.merriam-webster.com/dictionary/server>
- <https://dictionary.cambridge.org/dictionary/english/server?q=Server>
- [https://en.wikipedia.org/wiki/Server\\_\(computing\)](https://en.wikipedia.org/wiki/Server_(computing))
- [https://ru.wikipedia.org/wiki/Сервер \(аппаратное обеспечение\)](https://ru.wikipedia.org/wiki/Сервер_(аппаратное_обеспечение))
- [https://en.wikipedia.org/wiki/Blade server](https://en.wikipedia.org/wiki/Blade_server)
- <https://ru.wikipedia.org/wiki/Блейд-сервер>
- [https://en.wikipedia.org/wiki/Hot swapping](https://en.wikipedia.org/wiki/Hot_swapping)
- [https://ru.wikipedia.org/wiki/Горячая замена](https://ru.wikipedia.org/wiki/Горячая_замена)

# Классы вычислительной техники

- Computer cluster
  - Кластер
  - [https://en.wikipedia.org/wiki/Computer\\_cluster](https://en.wikipedia.org/wiki/Computer_cluster)
  - <https://www.merriam-webster.com/dictionary/cluster>
- Server farm / server cluster
  - Ферма серверов / кластер серверов
  - [https://en.wikipedia.org/wiki/Server\\_farm](https://en.wikipedia.org/wiki/Server_farm)
- Data Center / Data Centre
  - Центр обработки данных (ЦОД), «датацентр»
  - [https://en.wikipedia.org/wiki/Data\\_center](https://en.wikipedia.org/wiki/Data_center)
- Supercomputer
  - Суперкомпьютер
  - <https://en.wikipedia.org/wiki/Supercomputer>
  - <https://top500.org/>
  - <https://top50.supercomputers.ru/>

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	<b>Frontier</b> - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	<b>Aurora</b> - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
3	<b>Eagle</b> - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
4	<b>Supercomputer Fugaku</b> - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899

№	Название Место установки	Узлов Проц. Ускор.	Архитектура: кол-во узлов: конфигурация узла сеть: вычислительная / сервисная / транспортная	Rmax Rpeak (Тфлоп/с)	Разработчик Область применения
1	«Червоненкис» Яндекс, Москва	199 398 1592	199: CPU: 2x AMD EPYC 7702 , 1024 GB RAM Acc: 8x NVIDIA A100 HDR InfiniBand / нд / 100 Gigabit Ethernet	21530.0 29415.17	Яндекс NVIDIA IT Services
2	«Галушкин»  Яндекс, Москва	136 272 1088	136: CPU: 2x AMD EPYC 7702 , 1024 GB RAM Acc: 8x NVIDIA A100 HDR InfiniBand / нд / 100 Gigabit Ethernet	16020.0 20636.1	Яндекс NVIDIA IT Services
3	«Ляпунов» Яндекс, Москва	137 274 1096	137: CPU: 2x AMD Epyc 7662, 512 GB RAM Acc: 8x NVIDIA A100 HDR InfiniBand / нд / 100 Gigabit Ethernet	12810.0 20029.19	NVIDIA Inspur IT Services
4	«Кристофари Нео» SberCloud (ООО «Облачные технологии») , СберБанк, Москва	99 198 792	99: CPU: 2x AMD EPYC 7742, 2048 GB RAM Acc: 8x NVIDIA A100 HDR InfiniBand / 10 Gigabit Ethernet / 200 Gigabit Ethernet	11950.0 14908.6	NVIDIA SberCloud (ООО «Облачные технологии») Облачный провайдер
5	«Кристофари» SberCloud (ООО «Облачные технологии») , СберБанк, Москва	75 150 1200	75: NVIDIA DGX-2 CPU: 2x Intel Xeon Platinum 8168 24C 2.7GHz, 1536 GB RAM Acc: 16x NVIDIA Tesla V100 EDR Infiniband / 100 Gigabit Ethernet / 10 Gigabit Ethernet	6669.0 8789.76	SberCloud (ООО «Облачные технологии») NVIDIA Облачный провайдер



# «Облако»

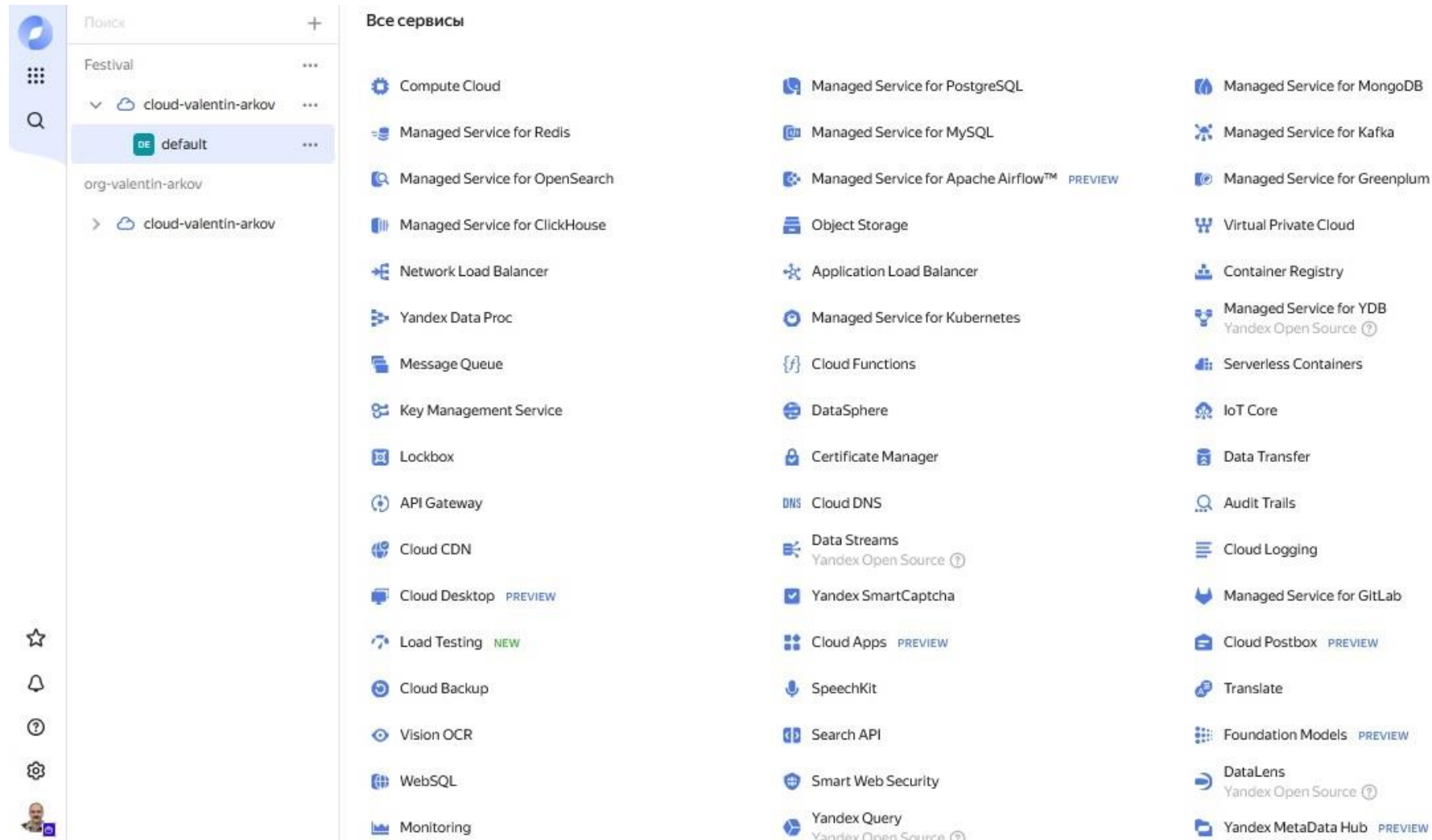
- Облачные вычисления (Cloud Computing)
  - Модель предоставления вычислительных сервисов через интернет по требованию
- Облачное хранилище данных (Cloud Storage)
  - Хранение больших объемов данных на удаленных серверах, которые доступны через интернет
- Облачная платформа для обработки данных (Cloud Data Platform)
  - Инструменты и сервисы для обработки и анализа больших данных
  - Amazon EMR, Google BigQuery и Microsoft Azure HDInsight
  - Экосистема Hadoop и Spark для распределенных вычислений и анализа данных в облаке
- Облачные сервисы для анализа данных (Cloud Analytics Services)
  - Готовые решения для анализа больших данных
  - Инструменты для машинного обучения, визуализации данных и бизнес-аналитики

# Облачные вычисления

- «Революция в ИТ-инфраструктуре»
- Модель предоставления вычислительных ресурсов через интернет
- Доступ к вычислительной мощности, хранилищам данных и сервисам по требованию
- Аренда ресурсов по мере необходимости
  - Серверы, хранилища, базы данных, сети, программное обеспечение
- Оплата за фактически использованные ресурсы
- «Вычислительный супермаркет»

# Облако Яндекс

<https://yandex.cloud>



# Преимущества облачных вычислений

- Гибкость / масштабируемость / эластичность
  - Легкое увеличение или уменьшение ресурсов
- Экономичность
  - Оплата только за используемые ресурсы (TCO)
- Доступность
  - Работа из любой точки мира
- Надежность
  - Высокий уровень отказоустойчивости (SLA)
- Безопасность
  - Профессиональные сотрудники, оборудование и ПО
    - <https://yandex.cloud/ru/blog/posts/2022/04/cloud-computing>

## «Облачные» риски



Sorry, you have been blocked

You are unable to access databricks.com

tableau.com is not available in your region.

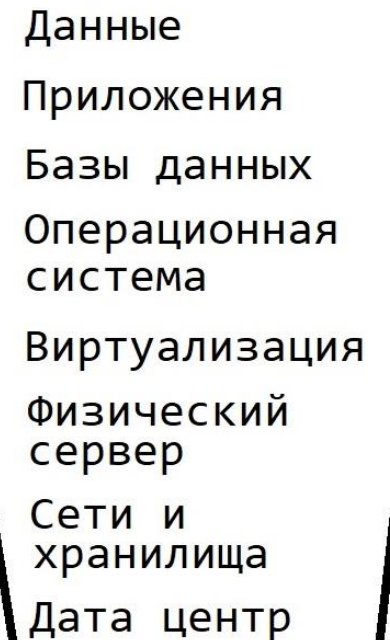
IBM

К этому контенту  
больше нет доступа

This content is no longer  
available

# Модели облачных сервисов

- Локальный ЦОД клиента / On-premise
- Colocation / Bare metal server / Выделенный сервер
- IaaS (Infrastructure as a Service):
  - Виртуальные серверы, хранилища, сети
- PaaS (Platform as a Service):
  - Разработка и развертывание приложений
- SaaS (Software as a Service):
  - Готовое программное обеспечение через интернет
- XaaS (Everything as a Service):
  - Широкий спектр услуг и ресурсов через интернет



Данные  
Приложения  
Базы данных  
Операционная  
система  
Виртуализация  
Физический  
сервер  
Сети и  
хранилища  
Дата центр

# Виды облачных решений

- Конфиденциальность & надежность
- Публичное облако / Public Cloud
  - Облачные службы для внешних клиентов
- Частное облако / Private Cloud
  - Собственная инфраструктура организации
- Гибридное облако / Hybrid Cloud
  - Комбинация частного и публичного облака
- Мульти-облако / Multi Cloud
  - Несколько облаков / провайдеров

# Рейтинги провайдеров

- Cloud Service Provider: Поставщик облачных услуг
- Gartner: Magic Quadrant
  - Analytics and Business Intelligence Platforms
  - <https://cloud.google.com/blog/products/data-analytics/2024-gartner-magic-quadrant-analytics-and-business-intelligence>
- CNews Analytics: Крупнейшие поставщики / игроки / компании
  - <https://www.cnews.ru/analytics/rating>



Magic Quadrant for Analytics and Business Intelligence Platforms



# CNews Analytics: Крупнейшие поставщики IaaS в России 2022

№ 2022	№ 2021	Компания	Выручка IaaS в 2022г., Ртыс. с НДС	Выручка IaaS в 2021г., Ртыс. с НДС	Рост выручки 2022/2021, %	Доля IaaS в общей выручке в 2022 г., %	"Платформы виртуализации"	Дата-центры
1	2	Ростелеком *	18 500 000	8 779 200	110,70%	н/д	VMware, Hyper-V, OpenStack (Tionix)	DataLine OST/Nord, Москва-I/II/III, Москва 4,1, М9, М9.PLUS, Удомля, СПб, Екб, Нск
2	3	Cloud.ru (1)	15 044 529	7 839 765	91,90%	81%	VMware, KVM, OpenStack	IXcellerate, Сколково, DataPro, 3data, ММТС-9
3	5	Selectel (2)	8 483 276	4 974 237	70,50%	87%	VMware, OpenStack-KVM	Цветочная 1-2, Дубровка 1-3, Берзарина, DataPro Moscow One, Nextremum (Новосибирск)
4	4	МТС (2)	5 700 000	4 800 000	18,70%	н/д	VMware	ЦОДы МТС в Москве (вкл. Авантаж), СПб, Нск, Владивостоке, Н.Новгороде и др.; DataSpace, Xelent, Ahost.
5	8	Яндекс.Облако	5 210 353	1 906 693	173,30%	56%	QEMU-KVM	н/д
6	6	.*	5 068 577	4 223 814	20,00%	17%	VMware, KVM	н/д
7	7	OnCloud	4 823 280	2 200 000	119,20%	56%	VMware, KVM	IXcellerate, DataSpace
8	1	Softline (3)	3 236 332	10 007 646	-67,70%	4%	VMware, Hyper-V, собственная разработка	IXcellerate, DataLine OST/ NORD, Xelent, RTCloud, «МегаФон» Екб, PS.KZ, STACK 24, «ТТЦ Останкино»
9	9	M1Cloud (4)	2 214 300	1 619 760	36,70%	н/д	VMware, Hyper-V, Рустэк, OpenStack-KVM	M1, DataPro
10	11	ITglobal.com (5)	2 158 910	1 255 180	72,00%	84%	VMware, vStack	DataSpace, IXcellerate, beCloud, AM2 Equinix, Kazteleport, NJ3, TOR3, Star of Bosphorus Data Center (Стамбул)
11	-	Beeline cloud	1 500 000	1 150 000	30,40%	91%	VMware, freebsd   bhyve	Oxygen, Linxdatacenter, ЦХД (Ростелеком), ЦОД Останкино, Beeline, Key Point

# Примеры облачных провайдеров

- Зарубежные

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)
- IBM Cloud
- Oracle Cloud
- Alibaba

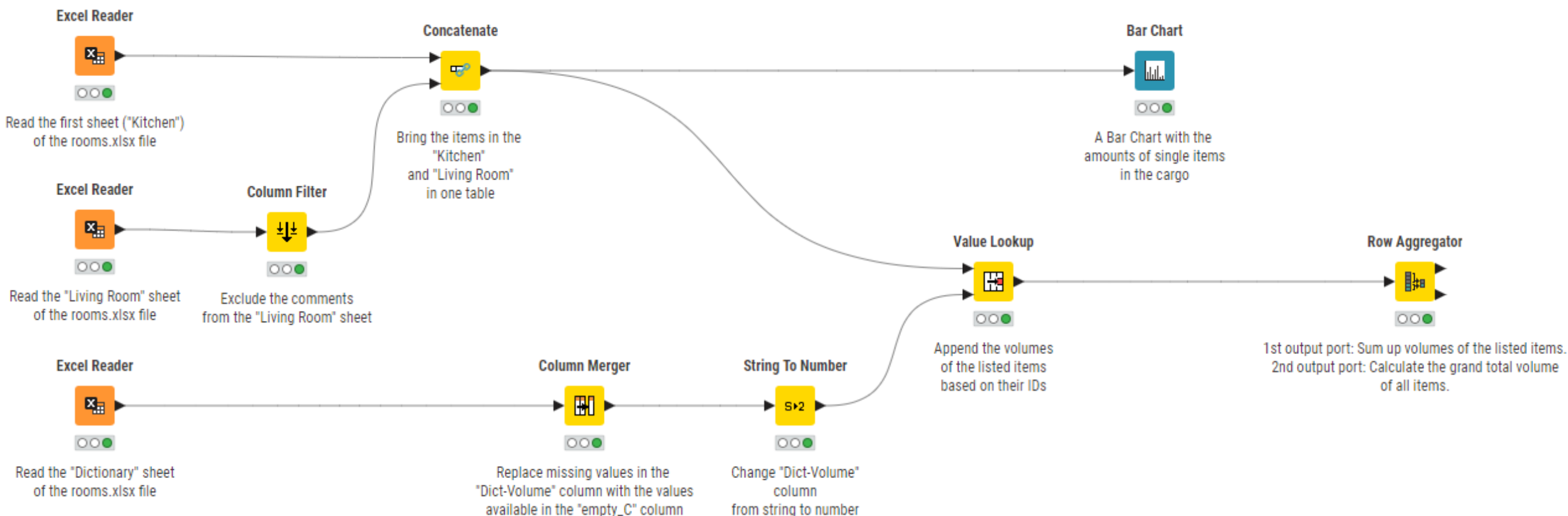
- Отечественные

- SberCloud / Cloud.ru
- Yandex Cloud
- VK / Mail.ru
- Selectel
- MTS
- Softline

# Определение облачных вычислений

- Национальный институт стандартов и технологий США (2011)
  - модель для обеспечения повсеместного, удобного сетевого доступа по требованию к общему пулу настраиваемых вычислительных ресурсов (например, сети, серверы, системы хранения данных, приложения и сервисы), которые могут быть быстро предоставлены и освобождены с минимальными эксплуатационными затратами или обращениями к провайдеру.
- NIST SP 800-145. The NIST Definition of Cloud Computing
  - Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.
  - <https://csrc.nist.gov/pubs/sp/800/145/final>

# KNIME: Визуальный конвейер обработки



<https://www.knime.com/>

# Облако Яндекс

- <https://yandex.cloud/ru/>
- Что такое облачные вычисления
  - <https://yandex.cloud/ru/blog/posts/2022/04/cloud-computing>
- Пробный период и грант
  - <https://yandex.cloud/ru/all-offers>
- Обучение и сертификация
  - <https://yandex.cloud/ru/training>
- Yandex DataLens
  - <https://datalens.yandex.cloud/>
- Основы работы с DataLens
  - <https://yandex.cloud/ru/training/datalens>

# Облако VK

- VK Cloud
  - <https://cloud.vk.com/>
  - [https://ru.wikipedia.org/wiki/VK \(компания\)](https://ru.wikipedia.org/wiki/VK_(компания))
  - [https://ru.wikipedia.org/wiki/VK Cloud](https://ru.wikipedia.org/wiki/VK_Cloud)
  - <https://cloud.vk.com/blog/>
- Cloud Big Data
  - <https://cloud.vk.com/bigdata/>
- Соглашение об уровне услуг
  - [https://ru.wikipedia.org/wiki/Соглашение об уровне услуг](https://ru.wikipedia.org/wiki/Соглашение_об_уровне_услуг)
- Cloud Native DIY
  - <https://cloud.vk.com/cloud-native-diy/>

# Платформа Snowflake

- Cloud data warehouse / Data warehouse as a service (DWaaS)
  - Интеграция с Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP)
  - Мульти-облако и многокластерная архитектура
- Аналитические запросы: SQL, Python, Java, Scala
  - <https://www.snowflake.com>
  - <https://streamlit.io/>
  - [https://en.wikipedia.org/wiki/Snowflake Inc.](https://en.wikipedia.org/wiki/Snowflake_Inc.)
  - [https://en.wikipedia.org/wiki/Data lake](https://en.wikipedia.org/wiki/Data_lake)
  - <https://cloud.google.com/discover/what-is-a-data-warehouse-as-a-service>
  - <https://quote.rbc.ru/news/article/5f5b90ed9a794716158e12d1>
  - <https://dzen.ru/a/Y5e8zAwGWSflyXzE>
  - <https://www.vedomosti.ru/technology/articles/2020/10/20/843823-osnovatel-snowflake>
  - <https://www.finam.ru/publications/item/snowflake-oblachnoe-xranilishe-dannyx-soblaznivshee-uorrena-baffeta-20200914-210500/>



# Многоуровневое хранение данных

- Иерархическое / ярусное хранение
  - Tiered Storage / Storage Tiering / «Тиринг»
  - Хранение данных на различных уровнях (память, SSD, HDD) в зависимости от частоты доступа и требуемой производительности
  - Оптимизация использования ресурсов и снижение затрат на хранение
- «Горячее» хранилище
  - Кэширование: Оперативная память и SSD для часто запрашиваемых данных
- «Тёплое» хранилище
  - Регулярный доступ: Менее активные данные на HDD дисках
- «Холодное» хранилище
  - Архивирование: Редко используемые данные - на медленные и дешевые устройства хранения (ленты, облачные хранилища)

# Типы облачных хранилищ данных

- **Файловые**
  - Иерархия папок и файлов
  - Аналог традиционных файловых систем
  - Протоколы SMB (Server Message Block) и NFS (Network File System)
  - Google Диск / Яндекс.Диск / HDFS
- **Блочные**
  - Данные разбивают на фиксированные блоки по различным серверам
  - Высокая производительность / параллельная обработка
  - Используют для виртуальных машин и баз данных
  - Amazon Elastic Block Store (EBS) / Microsoft Azure Blob Storage
- **Объектные**
  - Объекты в «бакетах»: данные + метаданные + идентификаторы
  - Большие объемы неструктурированных данных (изображения, видео и резервные копии)
  - Amazon S3 (Simple Storage Service) / Google Cloud Storage / Object Storage API
  - Интерфейс веб-сервиса: `https://<имя_бакета>.s3.<регион>.amazonaws.com/<путь_к_объекту>`

# Internet of Things (IoT) Интернет вещей

- «Сеть устройств» вместо «сети людей»
  - Датчики, программное обеспечение и обмен данными через интернет
  - Умные дома, автомобили, промышленные датчики, носимые устройства
- IoT-платформы
  - Управление устройствами (Device Management)
  - Сбор и обработка данных (Data Collection and Processing)
  - Аналитика и визуализация (Analytics and Visualization)
  - Безопасность (Security)
  - Интеграция с другими системами (Integration)
- Search Engine for the Internet of Everything
  - <https://www.shodan.io/>

# MPP-системы

- Massively Parallel Processing (MPP)
- «Массивно-параллельные системы»
- Системы с массовой параллельной обработкой
  - Множество процессоров (вычислительных узлов)
  - Вычислительная система с распределённой памятью
  - Параллельная обработка
  - Масштабируемость, производительность, гибкость
- Обработка больших объёмов данных
  - MPP-архитектуры в системах управления базами данных (СУБД)
- Классы параллельных систем
  - <https://parallel.ru/computers/classes.html>

# MPP-СУБД

- PostgreSQL
  - <https://www.postgresql.org/>
  - <https://postgrespro.ru/>
- Greenplum
  - <https://greenplum.org/>
- Arenadata DB, QuickMarts, Hadoop, Postgres
  - <https://arenadata.tech/>
- ClickHouse (Яндекс)
  - <https://clickhouse.com>
- Pangolin (Сбер)
  - <https://pangolin.sbertech.ru/>
- NoSQL: "Not only SQL"
  - <https://en.wikipedia.org/wiki/NoSQL>
  - [https://en.wikipedia.org/wiki/Data\\_orientation](https://en.wikipedia.org/wiki/Data_orientation)
  - [https://ru.wikipedia.org/wiki/Столбцовое\\_хранение](https://ru.wikipedia.org/wiki/Столбцовое_хранение)