



Семантические данные & Семантические хранилища данных

2021.10.08

Содержание



**Семантические
данные**



**Семантические
хранилища
данных**



Онтологии



Базы знаний



**NoSQL базы
данных**



Часть 01

Семантические данные

Семантические данные

“У меня есть мечта о сети [в которой компьютеры] смогут анализировать все данные в сети - контент, ссылки и транзакции между людьми и компьютерами. «Семантическая сеть», которая делает это возможным, еще не появилась, но, когда она появится, повседневные механизмы торговли, бюрократии и нашей повседневной жизни будут управляться машинами, общающимися с машинами. “Интеллектуальные агенты”, которых люди веками рекламировали, наконец-то материализуются.”

(Berners-Lee, 1999)

Семантическая Сеть – это «сеть данных», которая облегчает машинам понимание семантики, иначе говоря смысла, информации в Интернете. Она расширяет сеть веб-страниц, понимаемых человеком, вводя метаданные о страницах и об их отношениях друг к друг, которые понимают компьютеры, активируя автоматизированных агентов для более разумного доступа в Интернет и выполнения задач по поручениям пользователей

Бернерс-Ли в настоящее время является директором Консорциума Всемирной паутины (W3C), который курирует разработку стандартов семантической паутины. С 2013 года деятельность семантической сети была отнесена к деятельности Сети Данных.

Семантические данные

Знание конкретной области задачи, в которой программа должна решать свои проблемы, было более важным как источник силы для компетентного решения проблем, чем используемый метод рассуждений.

Подобласти представления знаний и приобретения знаний, которые позже были названы инженерией знаний, были созданы для предоставления методов и приемов представления человеческих знаний в понятной для машины манере.

Семантические данные

Семантика, наука (значения, смысла)

Семантическая технология предоставляет машинно-понятные (или лучше поддающиеся машинной обработке) описания данных, программ и инфраструктуры, позволяя компьютерам размышлять об этих артефактах.

Итак, что на самом деле означает машинная семантика?

```
<person>  
  <name>Sir Tim</name>  
  <phone number>01-444444</phone number>  
</person>
```

Семантические данные

Логика

С алгоритмической точки зрения реализация систем логического мышления ясно демонстрирует, насколько сложной разрешимостью и сложностью является управление.

- Логика высказываний
- Логика предикатов первого порядка
- Логика предикатов второго порядка
- Логика описания

Семантические данные

Семантические веб-языки

- HTML

```
<META name = "Author" lang= "fr" content = "Arnaud Le Hors">
```

- Extensible Markup Language (XML)

- Resource Description Framework (RDF)

```
<rdf:type, rdf:type, rdf:Property>
```

- RDF schema (RDFS)

```
rdfs:class, rdfs:subclass, rdfs: subproperty, rdfs:domain, and rdfs:range
```

- Web Ontology Language OWL (OWL Lite, OWL DL, and OWL Full)

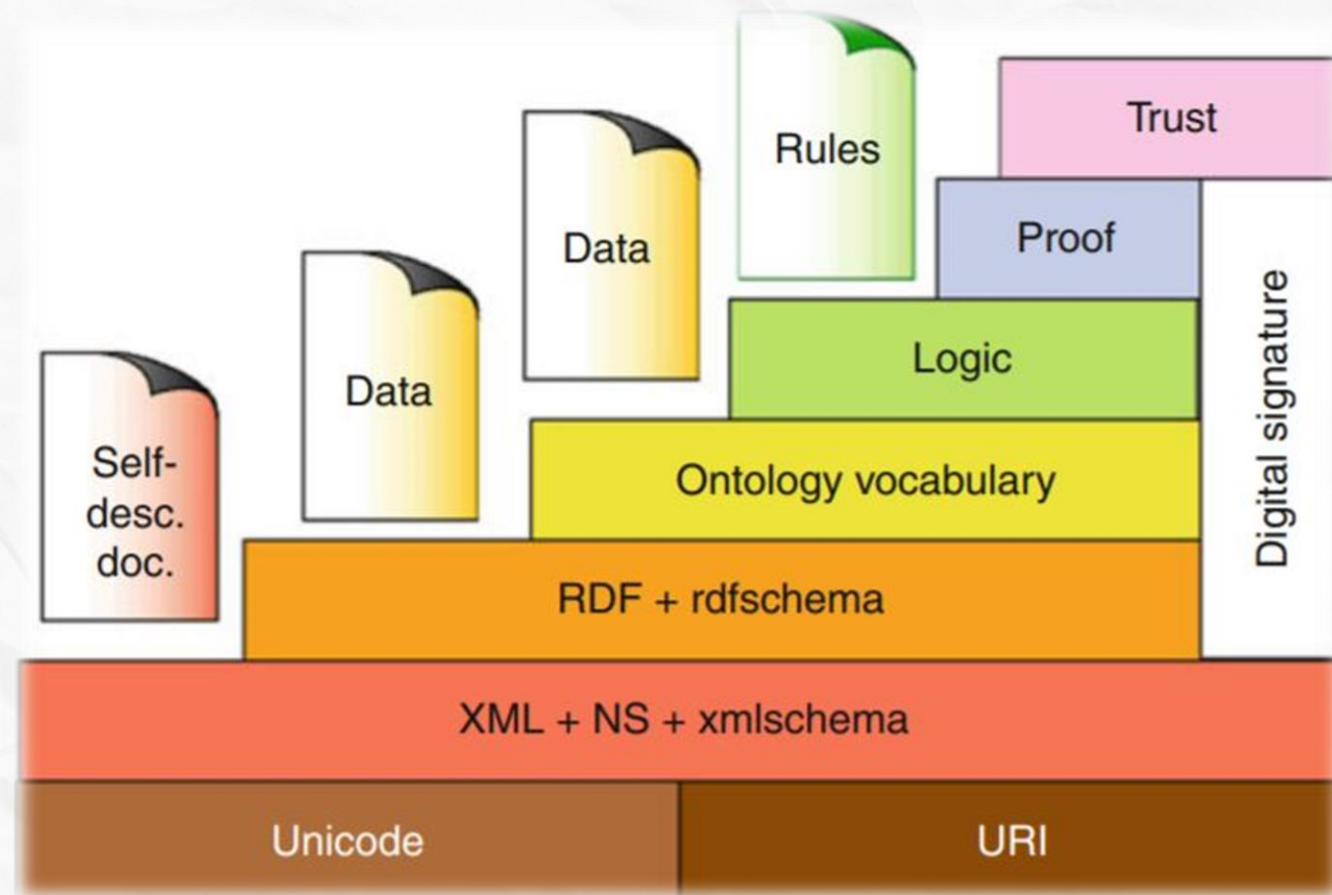
```
OWL Lite, OWL DL, and OWL Full
```

```
inverse, transitive, symmetric, or functional
```


Семантические данные

Вавилонская башня

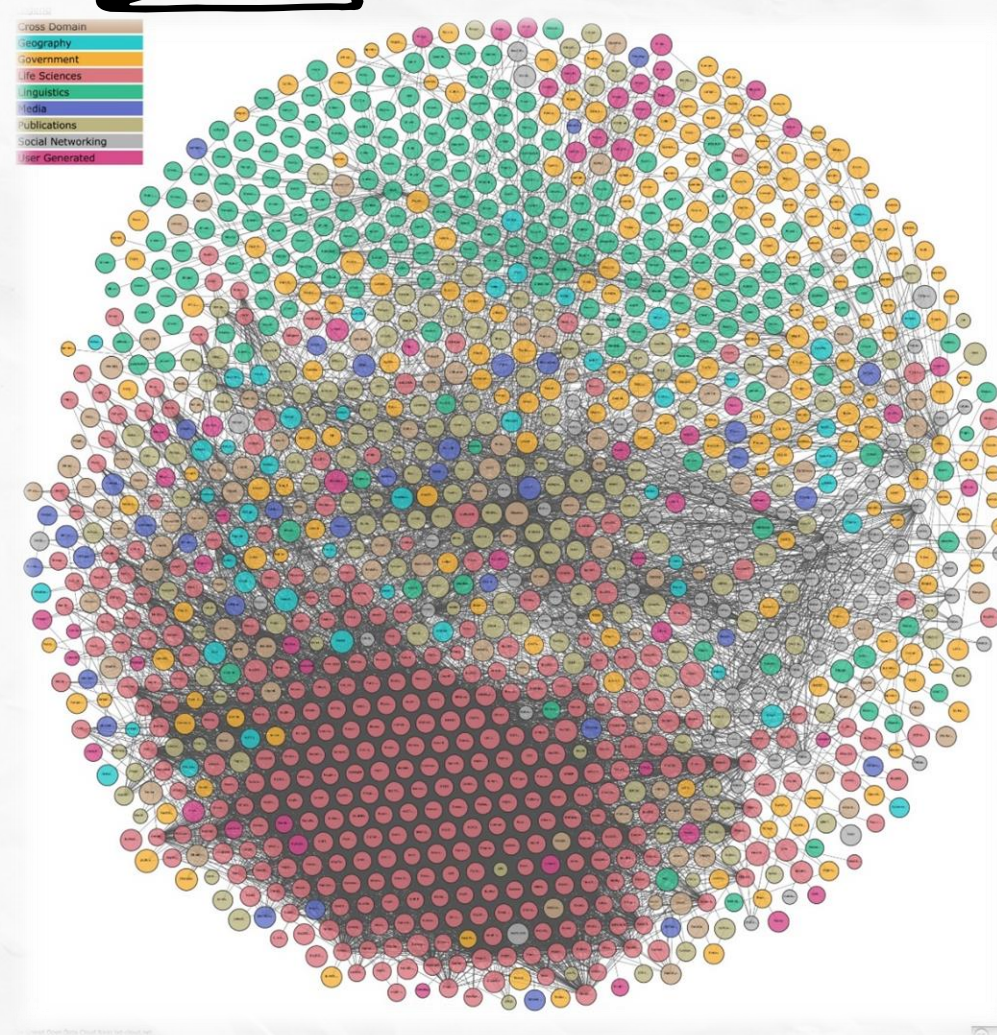
Модель взаимодействия открытых систем (модель OSI) является продуктом усилий Международной организации по стандартизации в области взаимодействия открытых систем. Это способ разделения системы связи на более мелкие части, называемые уровнями. Уровень - это набор концептуально похожих функций, которые предоставляют услуги вышележащему слою и получают сервисы от нижележащего уровня.



Семантические данные

Семантическая Сеть как база данных

- Какие конкретные методы базы данных (например, секционированные хэши, таблицы столбцов) наиболее применимы к высокопроизводительному хранилищу RDF?
- Как структурировать тесты для крупных репозиторий? В том числе какие размеры правильные?
- Когда и где следует обращаться к рассуждениям? Например, материализация (предварительное вычисление и сохранение предполагаемых троек) - дорогостоящий процесс, который может не способствовать достижению желаемых результатов.



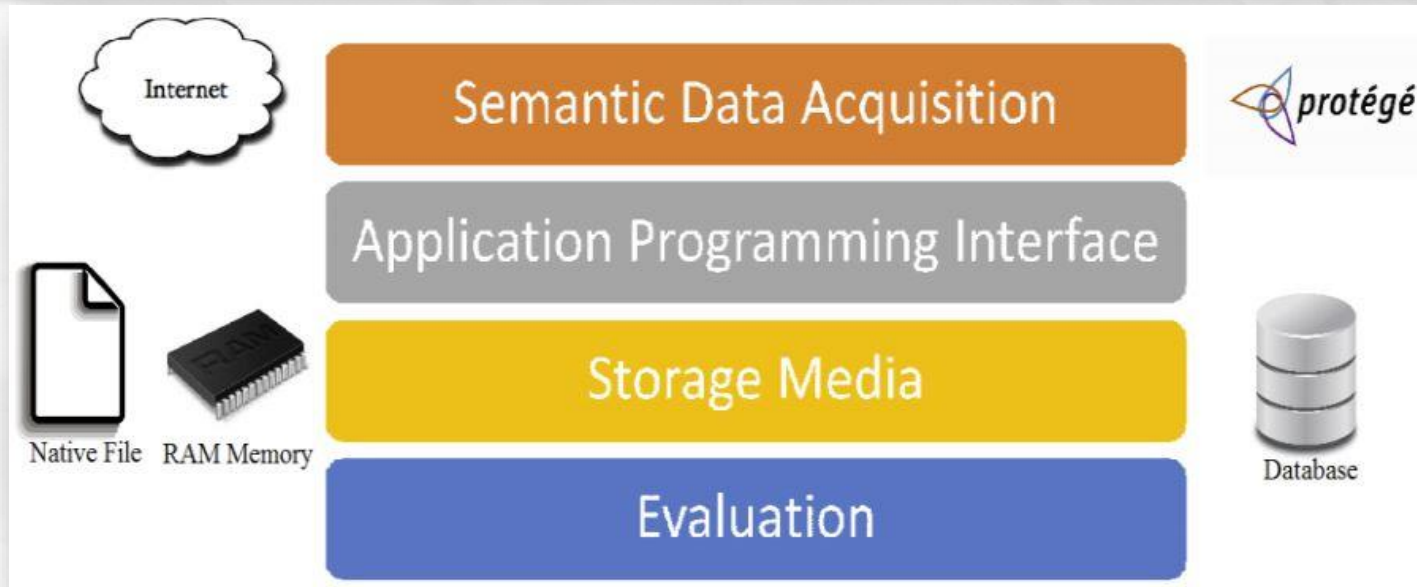


Часть 02
Семантические
хранилища данных

Семантические хранилища данных

Семантические хранилища данных

Хранение и поиск информации - важные функции информационных систем (ИС). Эти функции ИС реализовывались на протяжении десятилетий благодаря зрелости технологии реляционных баз данных. В последние годы возникла концепция семантической информационной системы (SIS) как ИС, в которой информация представлена с явной семантикой, основанной на ее значении, а не на ее синтаксисе, чтобы обеспечить ее автоматическую и интеллектуальную обработку компьютерами.



Семантические хранилища данных

Подходы к хранению семантических данных

Для хранения онтологии или семантических данных используются три подхода, а именно:

- В памяти
- Собственные или файловые системы
- Базы данных

Семантические хранилища данных

Программные платформы для хранения семантических данных

Чтобы обеспечить хранение и запрос семантических данных, было разработано несколько платформ. Самые популярные из этих платформ: AllegroGraph, Jena, Open Anzo, Minerva и Sesame.

Platform	License	Operating system	Type of Storage
AllegroGraph	Commercial/Free	Linux	Native
Jena	Free/Open Source	Windows/Linux	Memory, Native, RDB
Sesame	Free/ Open Source	Windows/Linux	Memory, Native, RDB
Open Anzo	Free/ Open Source	Windows Linux	RDB
Minerva	Free	Windows/Linux	RDB



Часть 03

Онтологии

Онтология

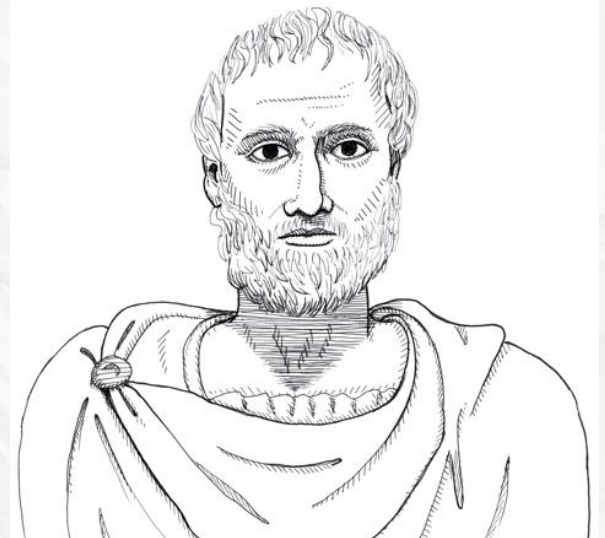
Онтология в философии

философская дисциплина - раздел философии, который имеет дело с природой и организацией реальности

- Наука о бытии (Аристотель, Метафизика, IV, 1)

Пытается ответить на вопросы:

- Что характеризует бытие?
- В конце концов, что происходит?
- Как следует классифицировать вещи?



Онтология

Онтология в философии

В философии онтология - это изучение бытия или существования. Его цель - выяснить, какие сущности и типы сущностей существуют:

- Что существует?
- Существование - это собственность?
- Что такое объект?
- Существуют ли нефизические (абстрактные) объекты?
- Как следует классифицировать вещи?

Aristotle's ontology:

Supreme genus:

Substance

Differentiae:

material

immaterial

Subordinate genera:

Body

Spirit

Differentiae:

animate

inanimate

Subordinate genera:

Living

Mineral

Differentiae:

sensitive

insensitive

Proximate genera:

Animal

Plant

Differentiae:

rational

irrational

Species:

Human

Beast

Individuals:

Socrates

Plato

Aristotle

etc.

Онтология

Онтология в информатике

Онтология - это инженерный артефакт

- Он состоит из определенного словаря, используемого для описания определенной реальности;
- И набора явных предположений относительно предполагаемого значения словаря.
 - (Почти всегда включая то, как следует классифицировать понятия.)

Онтология

Онтология в информатике

Онтология описывает формальную спецификацию определенной области следующим образом:

- Общее понимание интересующей области
- Формальная и машинно-управляемая модель интересующей области

"An explicit specification of a conceptualisation"

[Tom Gruber 1993]

Онтология

Онтология в науках

- Биоинформатика
 - Онтология генов, Онтология белков, Данные об экспрессии генов микрочипов и т.д.
- Медицина
 - Онтология Систематизированной номенклатуры медицинских клинических терминов (SNOMED CT): основная терминология, включающая более 364 000 концепций здравоохранения; более 984 000 описаний; 1,45 миллиона семантических отношений.
- • Лингвистика
- Интеграция баз данных
- Дизайн пользовательского интерфейса
- Фрактальное индексирование
- ...

*Pericardium is-a Tissue and containedIn . Heart
Pericarditis is-a Inflammation and hasLocation . Pericardium
Inflammation is-a Disease and actsOn . Tissue
Disease and hasLocation . containedIn . Heart is-a
HeartDisease and NeedsTreatment*

Онтология

Семантические онтологии в BBC

Запущенный в середине 1990-х годов веб-сайт BBC был ориентирован на поддержку вещательные бренды, такие как Top Gear, а также сайты, относящиеся к конкретному домену: новости, продукты питания, садоводство и т.д.

Веб-служба BBC является одним из самых посещаемых веб-сайтов и крупнейшим в мире новостным сайтом. По состоянию на 2007 год он содержал более двух миллионов страниц.

Основное внимание было уделено отдельным, автономным HTML-микросайтам, которые не связаны друг с другом, и другим источникам данных в Интернете

трудно найти все, что BBC опубликовала о каком-либо конкретном объекте

Мы не можем перейти со страницы о музыканте на страницу со всеми программами, в которых играл этот артист, к их биографии и т.д.

Онтология

Семантические технологии на BBC – [BBC онтологии](#)

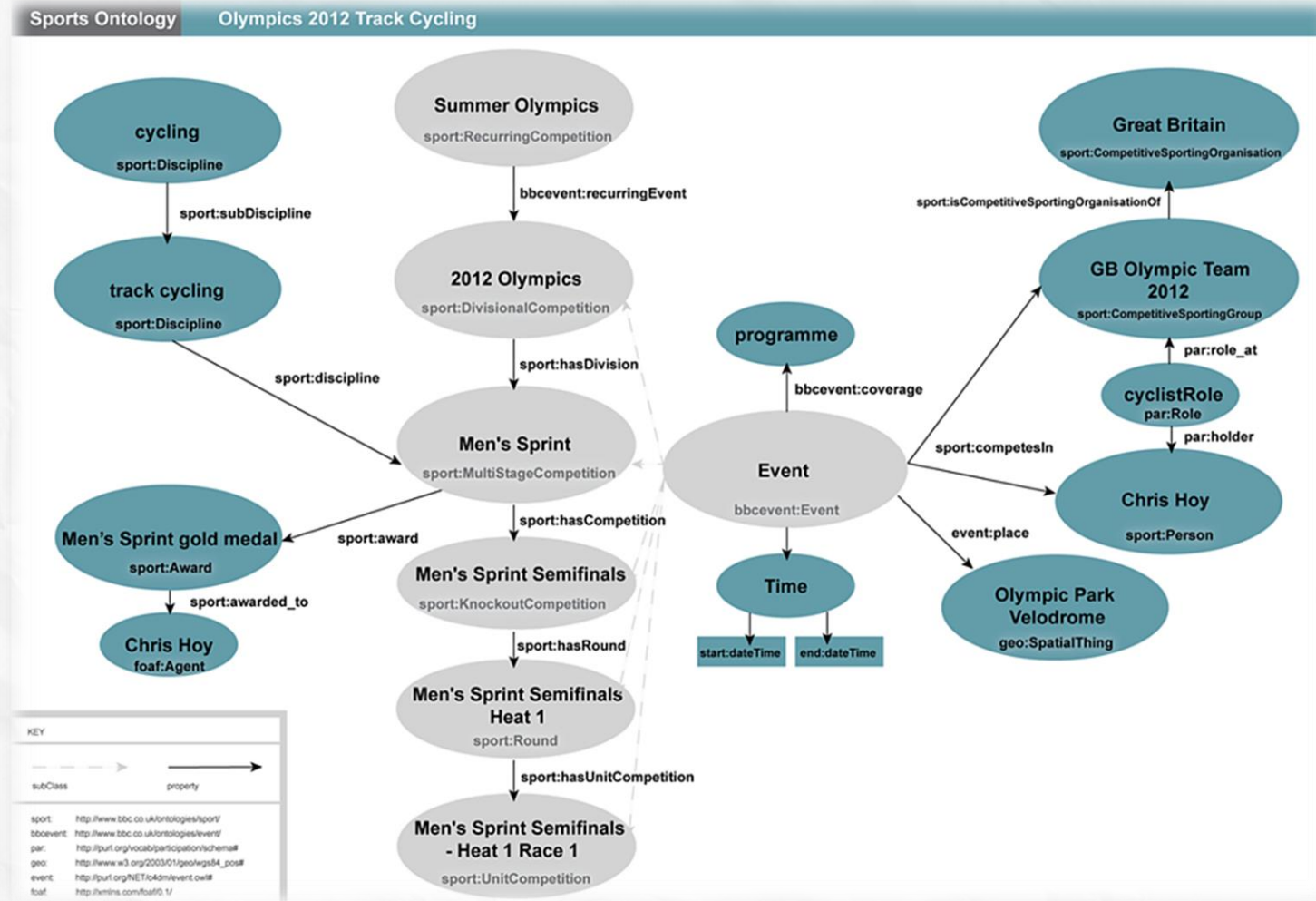
предоставляет доступ к онтологиям, которые BBC использует для поддержки приложений, ориентированных на аудиторию, таких как [BBC Sport](#), [BBC Education](#), [BBC Music](#), [News projects](#) и другие. предоставляет доступ к онтологиям, которые BBC использует для поддержки приложений, ориентированных на аудиторию, таких как.

- BBC Ontology
- Business News Ontology
- CMS Ontology
- Food Ontology
- ...

Онтология

Семантические технологии на BBC – Sport Ontology

Спортивная онтология - это простая облегченная онтология для публикации данных о спортивных соревнованиях.





ЧАСТЬ 04

База знаний

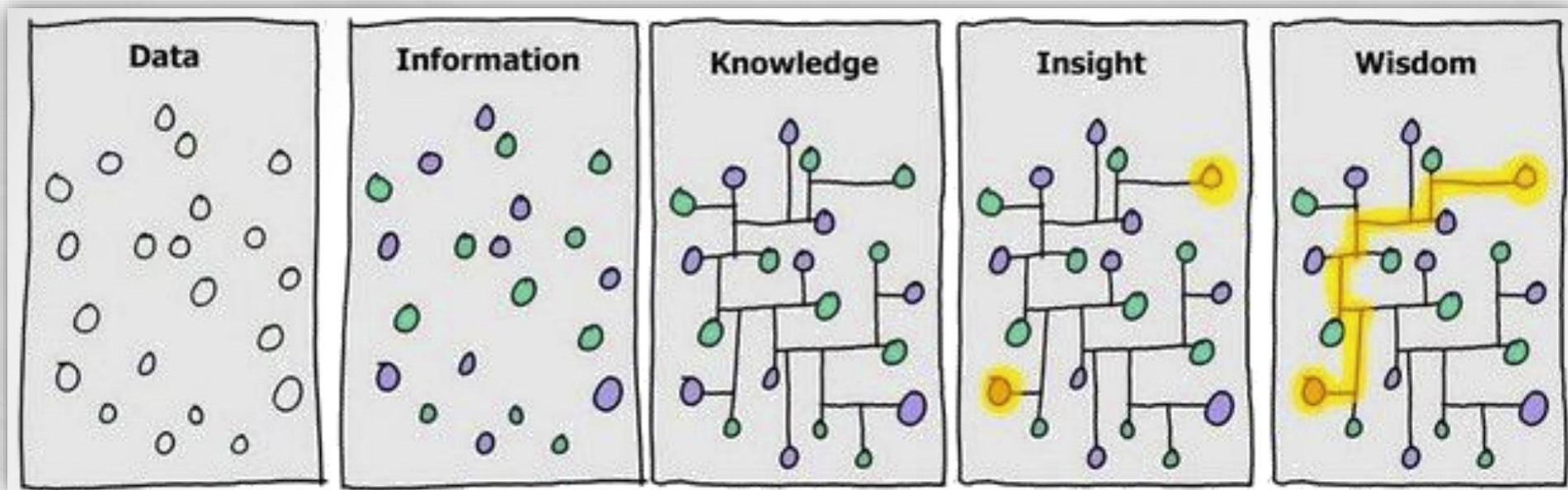
Базы знаний

База знаний (**KB**) - это технология, используемая для хранения комплексных, структурированной и неструктурированной информации используемой компьютерной системой. Первоначально этот термин использовался в связи с экспертными системами, которые были первыми системами, основанными на знаниях.

Всеобъемлющая семантически организованная машиночитаемая коллекция универсально значимых или специфичных для предметной области сущностей, классов и фактов (атрибутов, отношений)

- Плюс пространственные и временные измерения
- Плюс свойства и правила здравого смысла
- Плюс контексты сущностей и фактов, например, текстовые/визуальные доказательства
- Плюс...

Базы знаний



База данных: плоское представление данных (табличное)

База знаний: “Идеальным представлением базы знаний является объектная модель (**онтология**) с классами, подклассами и экземплярами.”

--Wikipedia

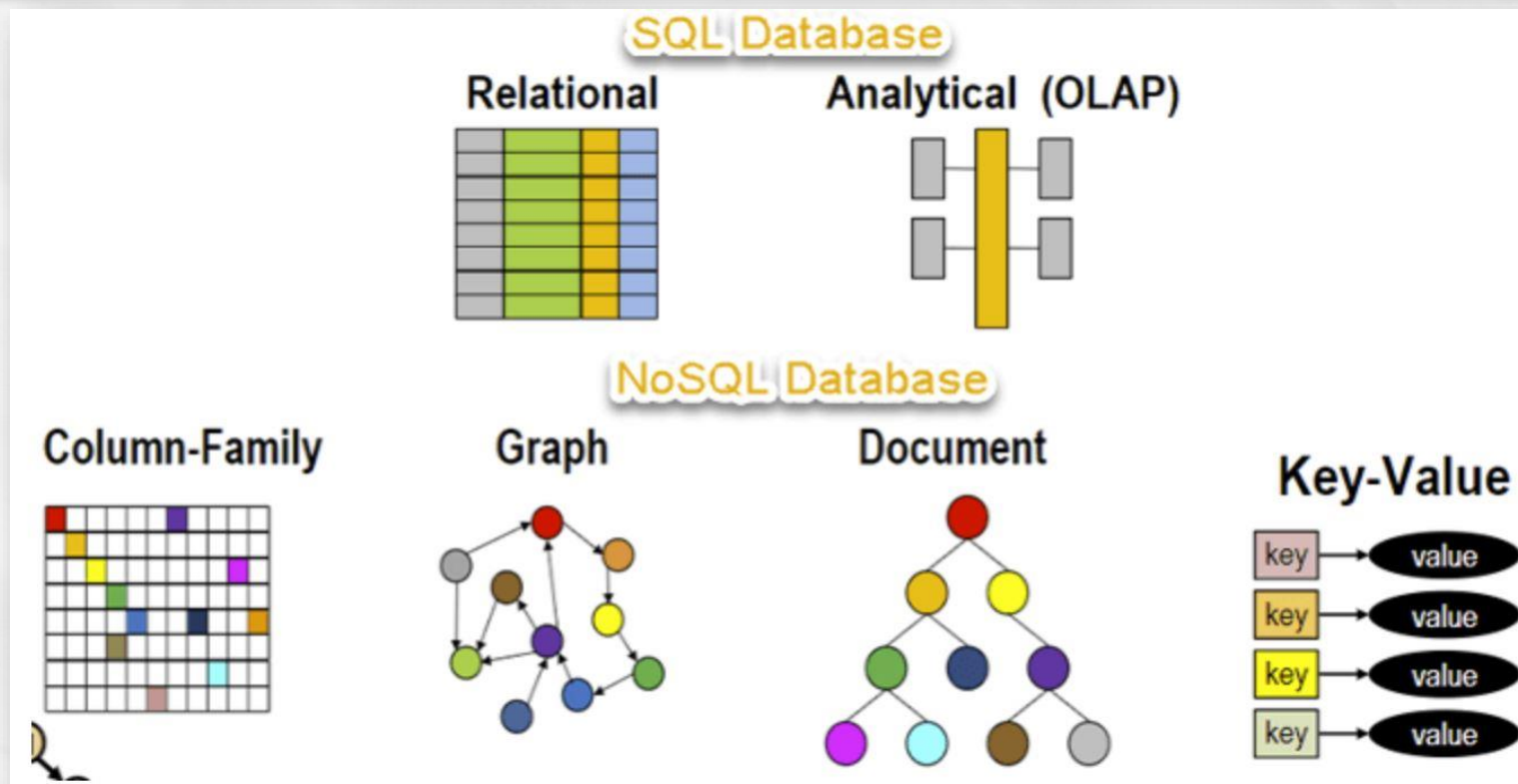


ЧАСТЬ 05

NoSQL база данных

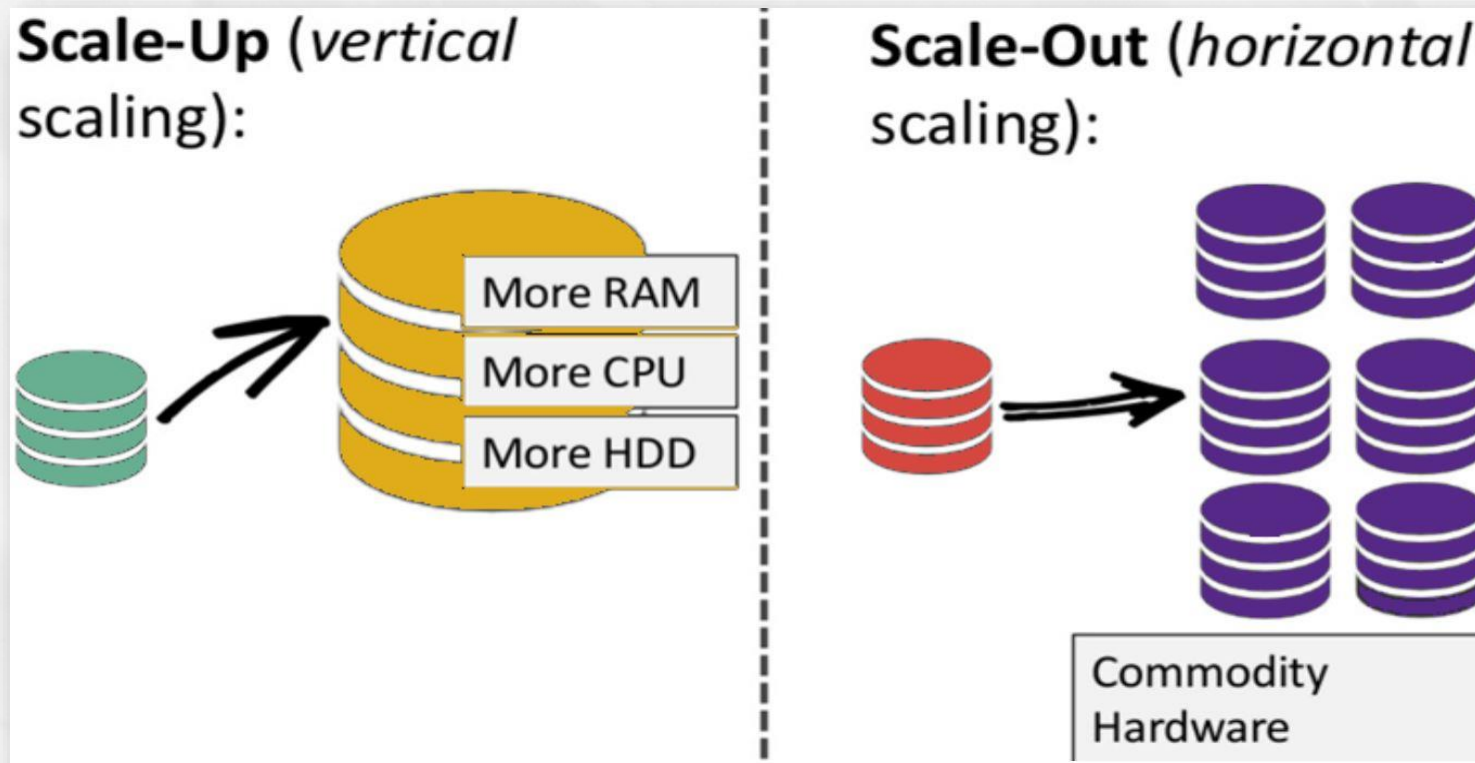
NoSQL база данных

NoSQL база данных предоставляет механизм хранения и извлечения данных, который моделируется иными способами, чем табличные отношения, используемые в реляционных базах данных.



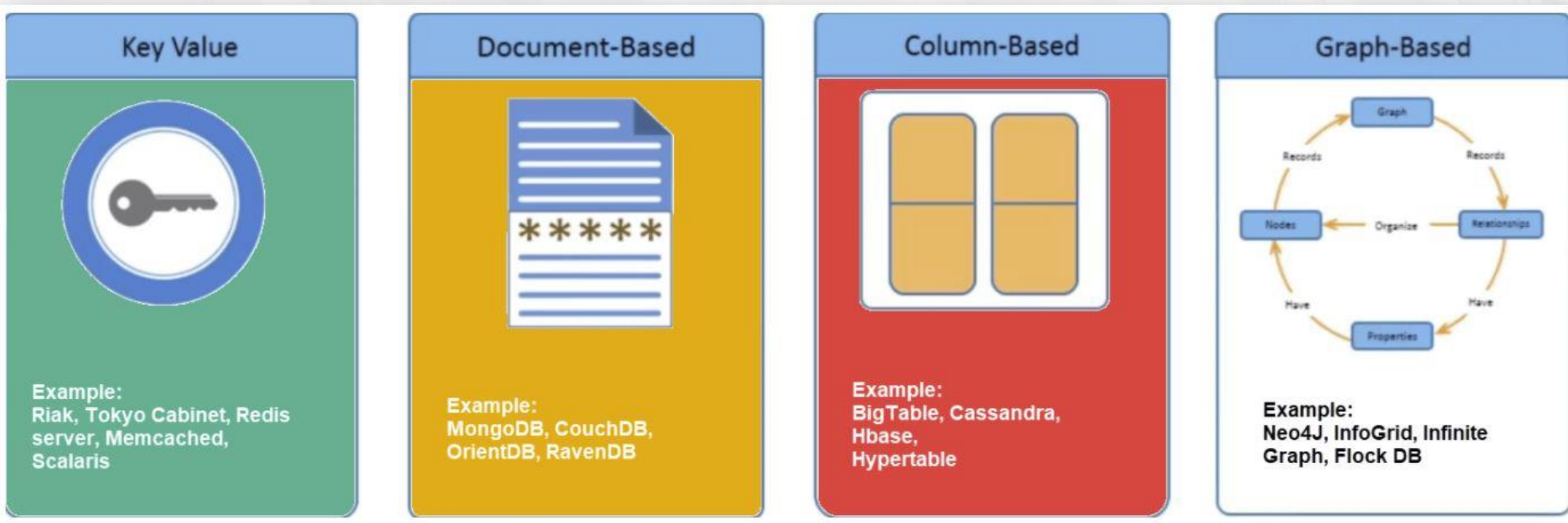
NoSQL база данных

База данных NoSQL не является реляционной, поэтому она масштабируется лучше, чем реляционные базы данных, поскольку они разработаны с учетом веб-приложений



NoSQL база данных

NoSQL базы данных в основном подразделяются на четыре типа: пара ключ-значение, семейство столбцов, графовая и документоориентированная.



NoSQL база данных

CAP теорема также называется теоремой Брюера. В ней говорится, что распределенное хранилище данных не может предоставлять более двух из трех гарантий

- Согласованность (Consistency)
- Доступность (Availability)
- Устойчивость к разделению (Partition Tolerance)

NoSQL Database

Многие базы данных NoSQL были разработаны молодыми технологическими компаниями, такими как Google, Amazon, Yahoo и Facebook, чтобы обеспечить более эффективные способы хранения контента или обработки данных для огромных веб-сайтов. Некоторые из наиболее популярных баз данных NoSQL:

- Apache CouchDB
- Apache Cassandra
- MongoDB
- Redis
- Elasticsearch



СПАСИБО

2021.10.08

mantx@itmo.ru