IBM Watson

Введение

При всех достоинствах современных поисковых систем, у них есть еще и существенный недостаток: мы вынуждены вводить поисковые запросы в виде ключевых слов, чтобы потом самостоятельно перерывать кучу веб-страниц в поисках нужной информации. Притом не факт, что мы эту информацию найдем. И альтернативы пока нет никакой.

А представит альтернативу корпорация IBM; новый суперкомпьютер Watson, работа над которым велась в течение последних трех лет, не только понимает вопросы, заданные «нормальным человеческим языком», но и отвечает на них в соответствующей форме. Оснастив эту машину базой данных с энциклопедическими знаниями, авторы проекта получили поистине беспрецедентный источник информации.

История

IBM Watson – новый виток эволюции суперкомпьютеров IBM:

- 1944 г. Mark I. Первый программируемый арифмометр.
- 1952 г. IBM 701. Первый компьютер на вакуумных трубках.
- 1954 г. IBM NORC. Вычисление числа Пи с 3089 знаками за 13 минут.
- 1955 г. IBM STRETCH. Конвейер, предвыборка, расслоение памяти.
- 1991 г. IBM + Thinking Machines. Идея о массивно-параллельных компьютерах.
- 1997 г. IBM Deep Blue. Шахматный суперкомпьютер.
- 2000 г. IBM ASCI White. Впервые система IBM - № 1 в Top 500.
- 2004 г. IBM Blue Gene/L. Blue Gene — проект компьютерной архитектуры, разработанный для создания нескольких суперкомпьютеров и направленный на достижение скорости обработки данных, превышающей 1 петафлопс.
- 2008 г. IBM Roadrunner. Был самым производительным суперкомпьютером в мире в 2009 году
- 2010 г. IBM Watson.

История Watson началась в 2006 году, когда Дэвид Феруччи, старший менеджер отделения IBM по семантическому анализу, занялся тестированием одного из самых мощных суперкомпьютеров компании, занимавшего одну из верхних строчек 500 самых производительных машин мира. Феруччи решил попробовать, насколько эффективно машина будет справляться с задачами, поставленными "естественным языком", и предложил ей ответить на 500 вопросов, заданных в уже состоявшихся программах Jeopardy! Результаты оказались катастрофическими: по сравнению с живыми игроками, машина недостаточно быстро "нажимала на кнопку" (то есть была готова к ответу), а в случае, когда она всё-таки могла конкурировать с людьми, количество правильных ответов не превышало 15%.

Феруччи заинтересовался причинами такого поведения суперкомпьютера и в итоге в 2007 году смог убедить руководство IBM дать ему команду из 15 человек и от 3 до 5 лет на создание эффективной автоматической системы, способной отвечать на неформализованные вопросы.

Такая система пригодилась бы всевозможным колл-центрам, справочным и любым другим службам, обслуживающим клиентов. У IBM уже был успешный опыт создания машины, способной поспорить с интеллектом человека – речь идёт о суперкомпьютере Deep Blue, который в 1997 году победил чемпиона мира по шахматам Гарри Каспарова. Эта победа сделала большую рекламу IBM, но коммерческого применения подобной установке найти так и не удалось. В случае же с системой автоматических ответов на вопросы коммерческий потенциал вполне очевиден.

Принципиальное отличие Watson от Deep Blue заключается в том, что если шахматный автомат имеет дело со строго логическими правилами игры, то машина, распознающая "естественную речь", сталкивается с куда более сложными правилами языка и многочисленными искажениями и отклонениями от них.

Но самая большая сложность заключается в том, что люди, сами того не осознавая, общаются в рамках своего культурного и социального контекста. В разговорной речи полно намёков, аллюзий и коннотаций, отсылок к неким

общим для конкретной общественной среды фактам, понятиям и явлениям. В их числе и религиозные представления, и политические убеждения, и всевозможные произведения искусства – от книг и картин до кинофильмов и компьютерных игр.

Для эффективной обработки подобной информации используются статистические алгоритмы, позволяющие путём анализа самых разнообразных документов устанавливать связь разных понятий друг с другом. Проще говоря, она определяет, какие слова чаще всего употребляются вместе. К примеру, "Кремль" чаще связан со словами "Россия", "Москва", чуть реже с "Казань", "Нижний Новгород", ещё реже – с "собор", "икона"" и т.п. Хотя эти алгоритмы известны давным-давно, полноценно применять их стало возможно лишь в последнее десятилетие – после кардинального роста производительности вычислительной техники и снижения стоимости накопителей для хранения огромных массивов данных.

Уотсон

Watson — суперкомпьютер фирмы IBM, оснащённый системой искусственного интеллекта, который был создан группой исследователей под руководством Дэвида Феруччи. Его создание — часть проекта DeepQA. Основная задача Уотсона — понимать вопросы, сформулированные на естественном языке и находить на них ответы в базе данных.

Watson представляет собой программный комплекс, который работает на кластере из 10 стоек по 9 стандартных серверов IBM Power 750 на базе процессоров POWER7 и обеспечивает обработку естественной речи, поиск информации, моделирует рассуждения и реализует технологии машинного обучения для ответов на вопросы.

Эта система получила известность после победы в игре "Jeopardy!", в рамках которой игроки соревнуются, кто быстрее ответит на вопрос, заданный на естественном английском языке.

Система Watson, названная в честь основателя корпорации IBM Томаса Уотсона (Thomas J. Watson), была построена группой ученых IBM, которые стремились тем самым решить сложнейшую задачу – создать компьютерную

систему, способную на уровне человека отвечать на вопросы, изложенные на естественном языке, причем быстро, точно и достоверно.

Понимание языка	**Легко**	**Сложно**
Понимание смысла вопроса	**Легко**	**Сложно**
Оценка уверенности в ответе	**Легко**	**Сложно**
Широта кругозора	Ограничена объемом памяти >1000 тб+ способность к обобщению	Ограничена объемом памяти 4 террабайт
Вычислительная мощность	Мозг Весит ~1300г, 20Ватт, не требует охлаждения	Суперкомпьютер 2880 ядер POWER7, 80КВт +охлаждение
Определение оптимальных ставок	**Медленно и неточно**	**Быстро и точно**
Эмоции	**Да**	**Нет**

Естественная для человека задача

Watson может	Watson не может (пока)
Выбирать вопрос	Слышать
Произносить ответ	Видеть
Нажимать на кнопку	Отвечать на аудио и визуальные вопросы (исключены из игры)
Получать вопрос в письменном виде в то время как люи читают с экрана	Связь с внешним миром (подсказки)

Джепарди!

Jeopardy! – это шоу-викторина, охватывающая широкий спектр тем, таких как история, литература, политика, наука, искусство и сфера развлечений. Участие в игре Jeopardy! – чрезвычайно сложная задача для компьютера, поскольку машинный интеллект изначально не понимает естественный человеческий язык. Более того, Jeopardy! – это настоящий

вызов для компьютерной системы из-за быстроты, с которой соперники должны давать правильные ответы на вопросы, а также из-за того, что для поиска точного ответа нужно анализировать содержащиеся в вопросах трудноуловимые ассоциации, скрытые значения, иронию, загадки и другие лингвистические и интонационные нюансы.

В Jeopardy! игроки должны принимать решения, основываясь на своей уверенности, что им точно известен правильный ответ. Иными словами, здесь нужно делать то, в чем традиционно силен человек, а не компьютер.

Формат викторины «Jeopardy» является исключительно трудным, поскольку предлагаемые участникам подсказки вынуждают их анализировать тонкие смысловые оттенки, учитывать иронию, разгадывать загадки и преодолевать другие сложности, т.е. заниматься теми видами деятельности, которых люди традиционно выполняют лучше, чем компьютеры.

Watson – это аналитическая вычислительная система, которая специализируется на анализе естественного человеческого языка и очень быстро выдает точные ответы на сложные вопросы.

Watson демонстрирует настоящий прорыв в понимании компьютером естественного языка – реального языка, на котором общаются и обмениваются информацией люди, а не машинного языка, специально разработанного или закодированного для компьютеров.

Команда Феруччи загружает в память IBM Watson миллионы всевозможных документов – учебники, энциклопедии, справочники, художественную и религиозную литературу. Для анализа вопросов одновременно используется более сотни алгоритмов, предлагающих сотни возможных решений. Затем другие алгоритмы оценивают достоверность потенциальных ответов, отсеивая невозможные в силу объективных причин (например, несоответствия даты события и лет жизни действующих лиц) и маловероятные. Чем больше будет получено одинаковых ответов, тем выше вероятность, что они правильны – в процессе игры, на табло выводится рейтинг из нескольких самых вероятных ответов, помимо чаще всего встречающегося.

Он полностью автономен, то есть он не подключен к Интернету или web-поиску. Чтобы ответить на вопрос, Watson перебирает примерно 200 миллионов страниц сведений на естественном языке, содержащихся в его памяти (что эквивалентно миллиону книг), чтобы найти точный ответ. На это у него уходит меньше 3-х секунд, при этом он может «объяснить», почему его ответ правильный.

Используемая Watson технология понимает задаваемый вопрос, анализирует миллионы блоков информации, хранимой во внутренней памяти, и выдает максимально точный ответ, руководствуясь найденными фактическими данными.

Перед шоу Jeopardy! вся информация, которой будет располагать Watson – в виде энциклопедий, справочников, книг, киносценариев и многого другого – загружается в системную память. Во время игры Watson, подобно другим участникам, «копается» в информационных недрах всего, что она «вычитала и выучила», чтобы связать уникальный смысловой язык, содержащийся в вопросах викторины, со знаниями, загруженными в ее память, и уверенно находить правильные ответы.

В чем заключается победа Уотсона в игре?

Для стороннего наблюдателя событие 14 февраля 2011 года, когда машина одержала победу над двумя сильнейшими игроками в интеллектуальной телевизионной игре Jeopardy! ("Рискуй!"), может показаться фантастикой. Но если «препарировать" победителя — систему Watson — на составные части и проследить взаимосвязи между ними, то увиденное представится совершенно прозаично: годы работы сотен ученых плюс огромные вложенные средства, а в результате — вопрос-ответная система на архитектуре UIMA.

В отличие от поединка Deep Blue с Гарри Каспаровым, который за его бессмысленность справедливо назвали битвой человека с паровым катком, выигрыш компьютера Watson в телевизионной игре Jeopardy! имеет колоссальное значение как для будущего вообще, так и для развития класса аналитических систем в частности.

Попутно надо заметить, что "Своя игра", ведущая происхождение от Jeopardy!, заметно эволюционировала и стала явно интереснее своего предка, и было бы занятно посмотреть, как Watson сыграл бы против наших соотечественников. Пока он этого не может — в оригинальной игре проще стратегия, и она формальнее. На Jeopardy! разработчики Watson продемонстрировали возможность создания работающей в режиме реального времени системы ответов на вопросы, сформулированные на естественном языке, с использованием накопленной базы знаний, хранящей неструктрированные данные также на естественном языке.

Собственно игровой момент имеет очевидное значение, но важнее другое — перспектива применения такого рода систем безгранична: как утверждается, началом будет, скорее всего, медицина, а далее практически все, что угодно, в нее попадает практически любая область человеческой деятельности.

В СМИ игру Watson часто пытаются представить как самостоятельный феномен, на самом же деле перед нами побочный результат серьезной академической работы, с богатой предысторией, а участие в Jeopardy! стало для нее потрясающей маркетинговой кампанией.

История создания искусственного интеллекта

История искусственного интеллекта насчитывает несколько попыток создания машин, способных отвечать на задаваемые им вопросы, и исторически первой была вопрос-ответная система, разработанная Робертом Симмонсом в Техасском университете в 1969 году.

Следующая значимая работа была связана с именем Дугласа Лената и его компанией Cycorp, где была создана база знаний Cyc, которая действительно работала, но ее приходилось составлять вручную.

Библиотека знаний Cyc. Во всех опубликованных хрониках работ по искусственному интеллекту можно найти упоминания о работах по проекту Cyc, который сравнивают с компьютером HAL, описанным Артуром Кларком в «Космической одиссее 2001 года». Само существование Cyc не представляло никакого интереса до лета 2002 года, когда ситуация радикально изменилась — библиотека знаний Cyc стала доступна всем.

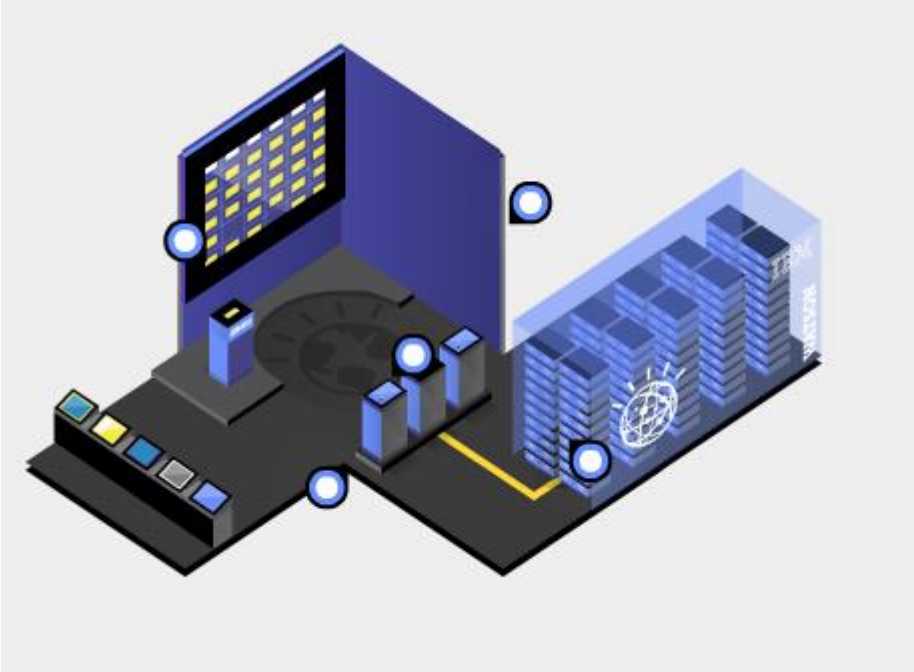
С 2002 года ведутся работы по проекту Halo, который можно рассматривать как развитие Cyc, но под патронажем агентства DARPA (среди участников — Стэнфордский исследовательский институт и несколько ведущих американских и европейских университетов). Своим путем к построению машины, отвечающей на вопросы, пришел Стивен Вольфрам, создатель программного продукта Mathematica.

Ближайшим предком Watson можно считать систему AQUAINT (Advanced Question Answering for Intelligence), созданную под патронажем Национального института стандартов (NIST). В IBM оценили ее перспективность и продолжили ее развитие, начав с того, что адаптировали этот проект под свои технологии, в результате получился практический вариант AQUAINT под именем PIQUANT (Practical Intelligent Question Answering Technology). Наследником PIQUANT является проект OpenEphyra, осуществленный IBM совместно с Университетом Карнеги-Меллона. От Watson он отличается возможностью поиска ответа в Web.

После шахматной эпопеи "Deep Blue против Каспарова", привлекшей к себе всемирное внимание, IBM требовалось что-то новое, и тогда выбор пал на систему PIQUANT, которая на тот момент могла отвечать только на 35% заданных вопросов, тратя на каждый десятки минут. Но скрытый потенциал был, и тогда в 2006 году была образована группа из 15 человек, которой поручили за пять лет довести систему «до ума».

В 2008 году Watson начал вполне успешно играть в Jeopardy!, а в заданный срок вышел и на телевизионный экран. Сверхплановой добавкой оказался созданный модным дизайнером Джошуа Дэвисом аватар, который способен изображать "эмоции" в процессе обдумывания системой Watson своих ответов.

Обзор Системы



Общий вид IBM Watson

Панель ответов.

Три лучшие ответа на каждую подсказку выведены на экран в панели ответа. Вместе с нимм показывается уровень уверенности. Если главный ответ превысит порог уверенности, то Watson подаст звуковой сигнал в ответ.

Аватар.

Призван отражать на сцене то, что происходит внутри системы Watson. Этот аватар соединен проводом с сервером, чтобы реагировать на уровень обработки Watson, например, изменение цвета, указывая на уверенность в ответ и отвечает ли он правильно или неправильно.

Ключи (подсказки).

Так как Watson не может видеть или услышать, система получает каждую подсказку в цифровой форме. Затем, тысячи алгоритмов Уотсона начинают анализировать подсказку и перерывать находки документов на естественного языке для ответа.

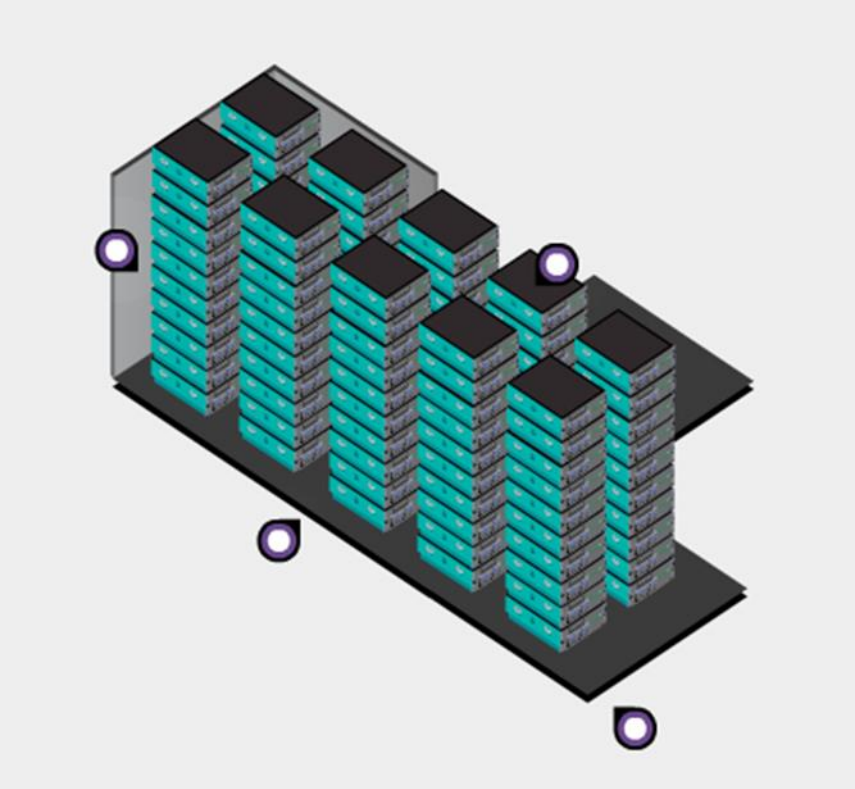
3 секунды.

Команда Watson обнаружила что в среднем, Опасность уровня чемпиона! соперник гудит через 3.5 секунды после того, как подсказка будет показана. Средние числа Watson меньше чем 3 секунды за подсказку, заключительную скорость для соревнования.

(Команда Уотсона обнаружила что в среднем, на уровне чемпиона соперник подает звуковой сигнал через 3,5 секунды после того, как подсказка была показана. Среднее время у Уотсон меньше чем 3 секунды.

Гудок

Watson использует тот же ручной зуммер как и любой другой соперник. Вместо того, чтобы использовать руку, чтобы подать звуковой сигнал, Watson снабжен механическим устройством, чтобы нажать кнопку.



Вычислительная часть

75%

В ходе предварительной серии спарринг-матчей, Уотсон был загружен всего на 75% от его полных вычислительных ресурсов.

2 часа.

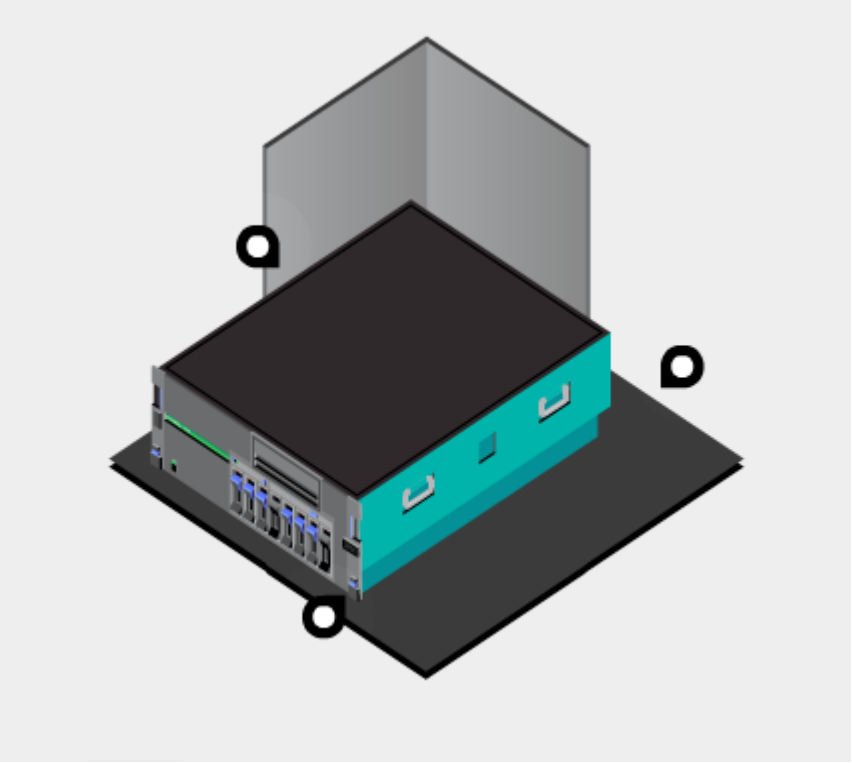
Компьютер с одноядерным процессором занимает более двух часов, чтобы выполнить глубокую аналитику необходимую, чтобы ответить на одну из подсказок в Jeopardy! Уотсон делает это менее чем за три секунды.

90 серверов.

Система, приводящая Watson в действие, состоит из 10 серверных стоек и 90 серверов IBM Power 750 на базе процессоров POWER7.

2880 процессорных ядер.

Вычислительная мощность состоит из 2880 процессорных ядер.



Серверы

Приложения.

Процессоры POWER7 внутри POWER 750 разработан, чтобы обработать приложения с интенсивными вычислениями и приложения обработки транзакций – от погодных моделирований, и банковских систем, к конкуренции против людей в Джепарди!

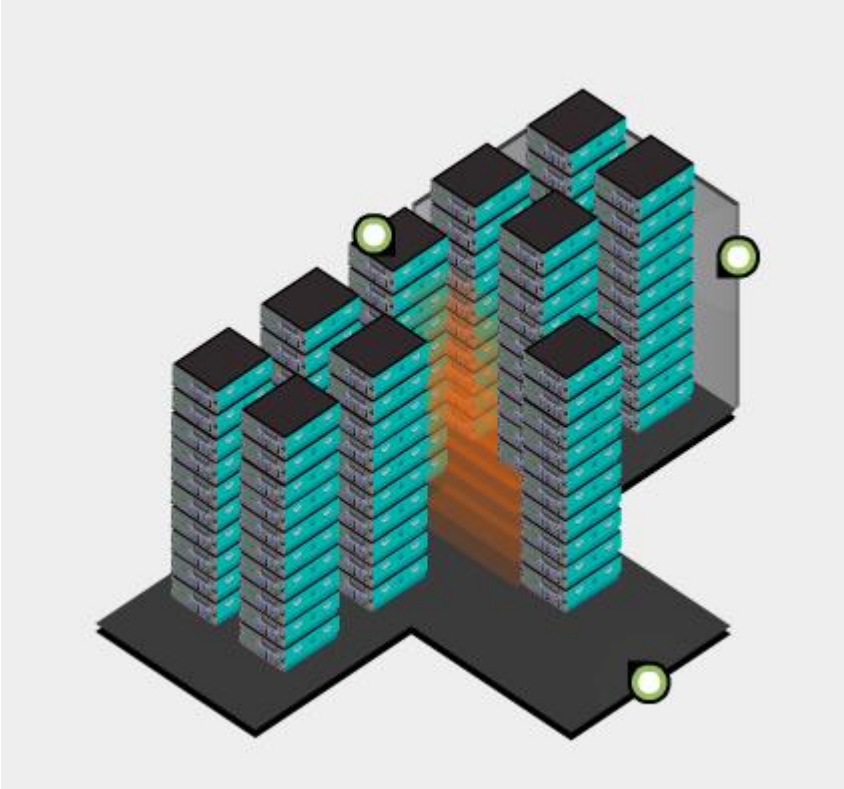
Байты по сравнению с мозгами.

Каждый сервер Power 750 размером 17,5 x 44 x 73 см. и весит 120 кг. Средний человеческий мозг меры 9,3 x 13,9 x 16,7 и весит около 15 кг.

100x.

Аппаратные средства, в сто раз более мощнее, чем Темно-синий, суперкомпьютер IBM, который побеждал самого великого шахматиста в мире в 1997.

(Аппаратные средства, что полномочия Уотсон в сто раз более мощны, чем Deep Blue, суперкомпьютер IBM который победил самого великого шахматиста в мире в 1997.



Архитектура взаимодействия подсистем

10 x 10 x 10 x 10.

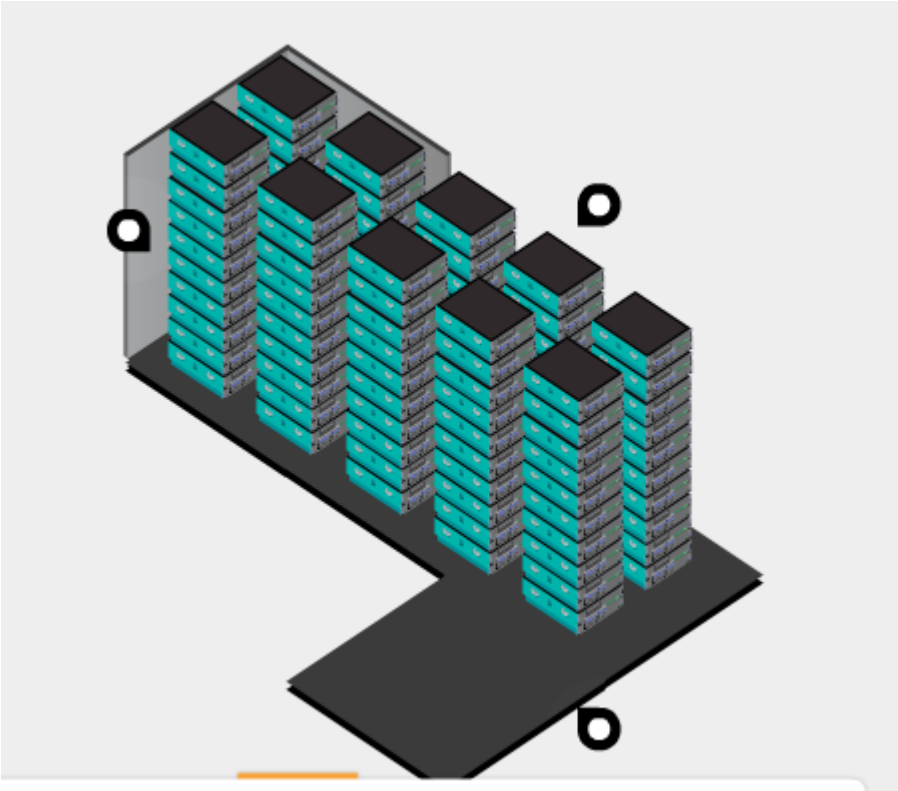
Watson использует Apache UIMA (Unstructured Information Management Architecture (Архитектура управления неструктурированной информации)), чтобы масштабировать обработку естественного языка параллельно на свои процессоры POWER7, позволяя Watson выполнить тысячи аналитических вычислений одновременно через кластер серверов, чтобы ответить на каждый вопрос как можно быстрее.

Unstructured Information Management Architecture

Watson разработан согласно архитектуре управления неструктурированной информацией – UIMA, если коротко. Эта архитектура программного обеспечения - стандарт для того, чтобы разрабатывать программы, которые анализируют неструктурированную информацию, такую как текст, аудио и изображения.

Системные факты

IBM сотрудничает с некоторыми другими компаниями для создания стандарта UIMA. Код UIMA передан Apache Software Foundation, где он теперь с открытым исходным кодом и доступен для всех. На сайте uima.apache.org можно скачать UIMA.



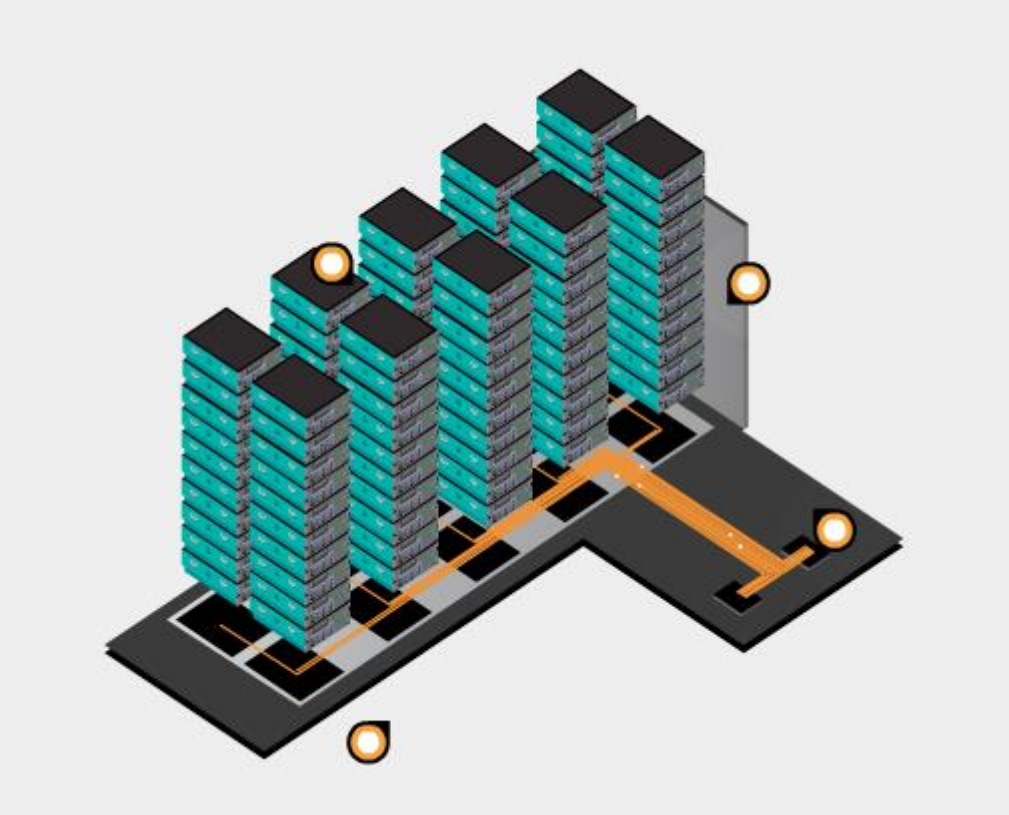
Загрузка ресурсов

Применение POWER7.

Университет Райс использует оптимизированную рабочую нагрузку системы на основе POWER7, чтобы проанализировать первопричины рака и других болезней. Для исследователей, это не просто более универсальный сервер, это гигантский скачок к пониманию рака.

1 к 1 тысяче.

Уотсон оптимизирован для ответа на каждый вопрос как можно быстрее. Та же система может быть оптимизирована, чтобы ответить на тысячи вопросов в кратчайшие сроки. Такая масштабируемость делает Уотсон весьма привлекательной для бизнес-приложений.



Питание ВС

Системные факты.

Power 750 был первым четырех сокетным для серверов этого класса. Все Power Systems включают EnergyScale ™ технологию для снижения энергопотребления и обеспечения возможности управления и настройки энергопотребления.

2880 ядер.

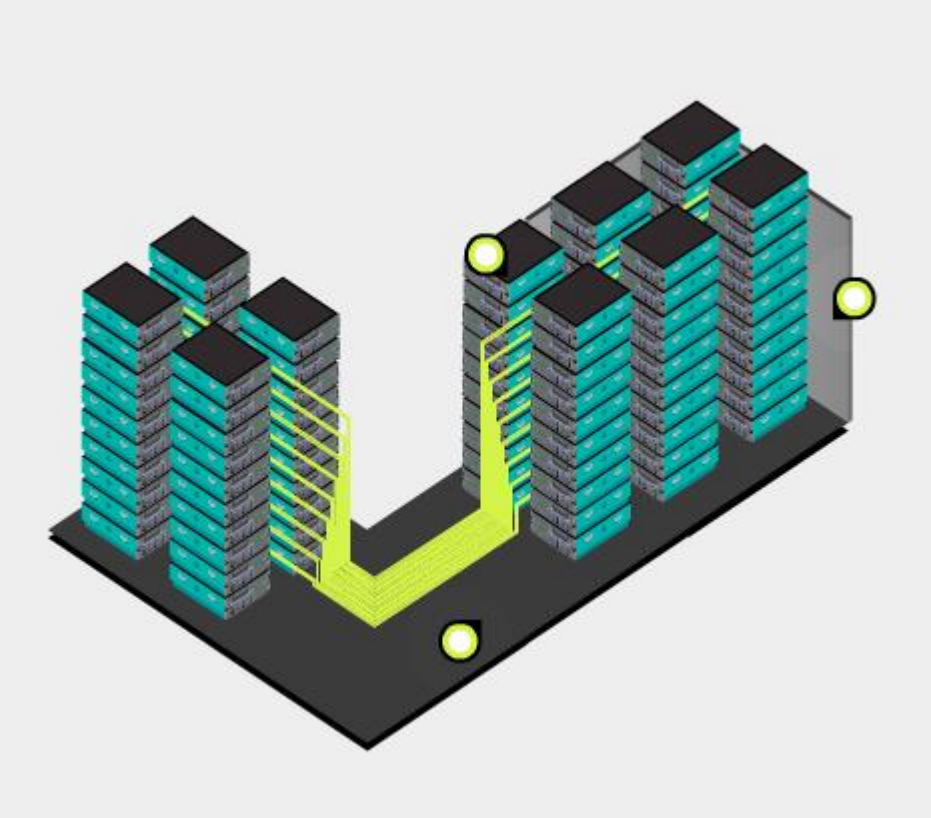
В прошлом способ ускорить обработку состоял в том, чтобы ускорить процессор. Это расходовало больше энергии и вырабатывало больше тепла. Watson масштабирует его вычисления более чем 90 серверов, каждый с 32 ядрами POWER7, достигающими 3,55 ГГц. Это обеспечивает большую производительность и потребляет меньше энергии.

Что за несколько ватт.

Глубокая аналитика Уотсона обеспечивается менее чем за три секунды и требует немного энергии. Уотсон потребляет меньше киловатт энергии в течение игры Jeopardy!, чем типичная прачечная за весь день.

5,000 мВт.

В 2005 центры обработки данных в США использовали приблизительно 5,000 мегаватт энергии. Это - примерно эквивалентно ежегодному выводу пяти электростанций.



Коммутационная сеть

9000 X.

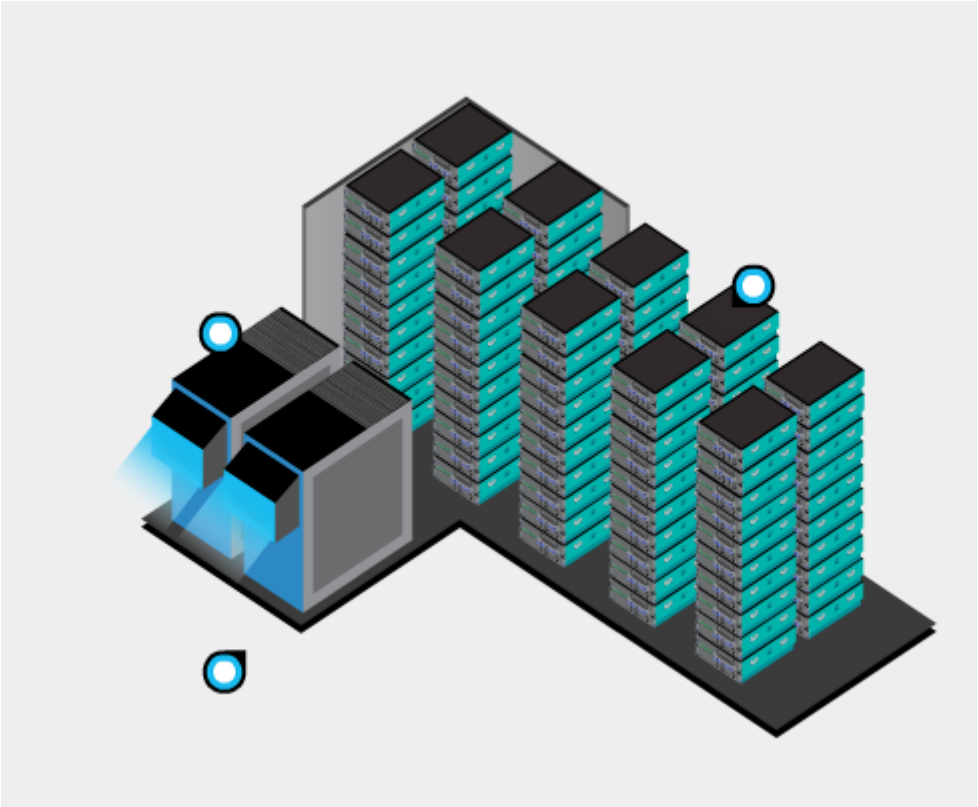
Сеть, которая соединяет серверы системы, может обработать 90 x 10 миллиардов бит в секунду. Напротив, типичная Сеть Ethernet для дома оценена в 100 Мбит/с, или в 90 раза медленнее, чем в Watson.

10 Гб скорость.

Уотсон не подключен к Интернету. Тем не менее, серверы системы соединены между собой 10 Gigabit Ethernet сетью.

Скорость + ум.

Watson генерирует сотни возможных ответов, оценивает каждый одновременно и сужает ответы к лучшему выбору в приблизительно то же время, что и человек-чемпион игры.



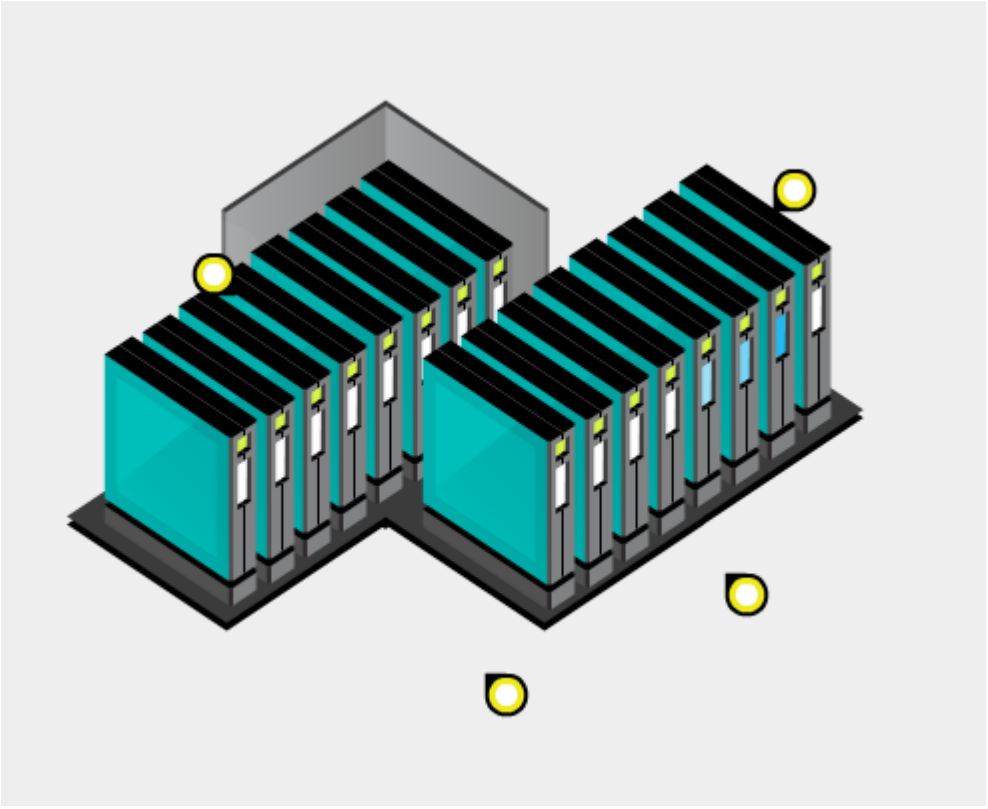
Охлаждение системы

18°C.

Два 20-тонных модулей кондиционирования воздуха, регулирующие температуру серверной Watson, достаточно на столько, чтобы охладить комнату приблизительного размера в одну треть размера футбольного поля.

Факт центра обработки данных.

До 50% энергии, расходуемой современным центром обработки данных, идет на воздушное охлаждение.



Хранилище данных

Тера-мозг.

Теоретические оценки объема памяти человеческого мозга изменяются между одним и 1,000 терабайт.

Это все?

500GB знаний, не кажутся достаточными, чтобы конкурировать в Jeopardy! Но подумайте: Уотсон хранит документы, в основном, на натуральном языке - который требует гораздо меньше памяти, чем изображения, видео и аудио файлов на персональном компьютере.

500 гигабайт.

На 500GB диска содержится всю информация, чтобы Watson мог конкурировать в Джепарди! Этот размер данных эквивалентен приблизительно 200 миллионам печатным страницам текста.

Платформа

Уотсон состоит из 90 серверов Power7 750, каждый из которых содержит по 4 восьмиядерных процессора POWER7. И всего имеется 2880

процессорных ядер, способных выполнять 80 триллионов операций с плавающей запятой в секунду (80 терафлопс. Суммарная оперативная память Уотсона более 15 терабайт памяти). 4 терабайта дискового пространства.

Система имела доступ к 200 миллионам страниц структурированной и неструктурированной информации объемом в 4 терабайта, включая полный текст Википедии. Во время игры Уотсон не имел доступа к интернету.

IBM ежегодно инвестирует в исследования и разработки приблизительно 6 млрд. долларов.

Power 750

Power 750 Express – сервер с 1-4 сокетами и поддержкой до 32 ядер, использующий 6- и 8-ядерные процессорные модули POWER7. Сервер IBM Power 750 Express обеспечивает высочайшую производительность, поддерживает операционные системы AIX, IBM i и Linux®, а также технологию виртуализации PowerVM. Если вам нужна надежная и эффективная платформа для консолидации серверов либо мощный сервер приложений или баз данных, Power 750 Express с сертификатом ENERGY STAR с успехом справится с поставленными задачами, используя технологии, обеспечивающие заказчику конкурентное преимущество.

Сервер IBM Power 750 Express Преимущества

Преимущества	Описание
Исключительная производительность POWER7	• Ускорение доступа к данным, сокращение времени отклика • Возможность обрабатывать больше рабочих нагрузок на меньшем числе серверов; снижение затрат на поддержку IT-инфраструктуры за счет сокращения количества серверов и количества лицензий на ПО
Технология виртуализации PowerVM	• Позволяет легко добавлять рабочие нагрузки по мере роста бизнеса • Полное использование потенциала системы и сокращение расходов на инфраструктуру путем консолидации рабочих нагрузок на ОС AIX, IBM i и Linux

Преимущества	Описание
	• Возможность обработки непредсказуемых пиковых нагрузок с помощью совместного доступа к ресурсам
Технология Active Memory Expansion	• Повышает производительность имеющихся ресурсов сервера
Функции надежности, готовности и удобства обслуживания (RAS)	• Поддерживают непрерывность работы работы приложений, позволяя заказчикам сосредоточиться на развитии бизнеса
Панель Light Path Diagnostics	• Позволяет легко и быстро диагностировать проблемы с оборудованием
Соответствует требованиям ENERGY STAR	• Снижение энергопотребления и тепловыделения
IBM Systems Director Active Energy Manager с технологией EnergyScale	• Значительное и динамическое повышение энергоэффективности, сокращение расходов на электроэнергию за счет инновационных функций управления энергопотреблением • Поддержка непрерывного выполнения бизнес-процессов в условиях ограниченных энергоресурсов

Сервер IBM Power 750 Express :: Характеристики

Варианты конфигурации	
Процессорные модули POWER7 – один на процессорной карте	6-ядерный с частотой 3,3 ГГц или
	8-ядерный с частотой 3,0 ГГц или
	8-ядерный с частотой 3,3 ГГц или
	8-ядерный с частотой 3,55 ГГц(*)

Сокеты	От 1 до 4
Кэш-память 2-го уровня (L2)	256 КБ на ядро
Кэш-память 3-го уровня (L3)	4 МБ на ядро
Память	от 8 ГБ до 512 ГБ памяти DDR3 в модулях RDIMM
	Технология Active Memory Expansion
Твердотельные накопители (SSD)	До восьми приводов SFF
Дисковые накопители	До восьми приводов SFF Serial Attached SCSI (SAS)
Емкость дисков	До 2,4 ТБ
Отсеки для носителей	Slimline для DVD-RAM
	Половинная высота для ленточных накопителей или съемных дисков
Разъемы для PCI-адаптеров	Два разъема PCI-X 2.0; три разъема PCI Express 8x
Стандартные адаптеры ввода-вывода	
Встроенный адаптер Integrated Virtual Ethernet	Четыре порта Ethernet 10/100/1 000 Мбит/с или
	Два порта 10 Gigabit Ethernet (GbE)
Интегрированный контроллер SAS	Один контроллер для SAS DASD/SSD и DVD-RAM
	Защищенная кэш-память 175 МБ (опция)
Другие встроенные порты	3 порта USB, 2 порта HMC, 2 системных порта, 2 порта SPCN
Разъемы GX (12X)	Два(**)
Компоненты расширения (опция)	
Расширение подсистемы ввода-вывода	До 4 выдвижных секций подсистемы ввода-вывода PCIe 12X
	До 8 выдвижных секций подсистемы ввода-вывода PCI-X DDR 12X

Высокопроизводительные PCI-адаптеры	8 Gigabit Fibre Channel (FC); 10 GbE, 10 Gigabit Fibre Channel over Ethernet (FCoE)
Другие поддерживаемые PCI-адаптеры	SAS, SCSI, Wide Area Network (WAN)/Async, USB, Crypto, SCSI over IP (iSCSI)
Технологии PowerVM	
POWER Hypervisor	Динамические процессоры LPAR, Virtual LAN (VLAN) (взаимодействие между разделами «память-память»)
PowerVM Standard Edition (опция)	PowerVM Express Edition и технология Micro-Partitioning с возможностью создания до 10 микроразделов на каждый процессор; несколько общих пулов процессоров; общие выделенные ресурсы (Shared Dedicated Capacity); PowerVM Lx86
PowerVM Enterprise Edition (опция)	PowerVM Standard Edition плюс Live Partition Mobility (LPM) и Active Memory Sharing (AMS)
Функции надежности, готовности и удобства обслуживания (RAS)	Система поиска и исправления ошибок IBM Chipkill Error Checking and Correction (ECC)
	Функция повторения инструкций процессора Processor Instruction Retry
	Функция восстановления на другом процессоре Alternate Processor Recovery
	Сервисный процессор для мониторинга ошибок
	Отсеки для дисков с возможностью «горячей» замены
	Блоки питания и вентиляторы с резервированием и возможностью «горячей» замены
	Динамическое перераспределение компонентов (Dynamic component Deallocation)
Операционные системы(***)	AIX
	IBM i

	Linux for POWER
Высокая готовность (HA)	Семейство IBM PowerHA
Энергопотребление	200-240 В, однофазный переменный ток
Габариты системы	Выдвижная секция стойки: 6,9" (высота) x 17,3" (ширина) x 28,7" (глубина) (175 мм x 440 мм x 730 мм); масса: 120,0 фунтов (54,4 кг)(****)
Гарантия (ограниченная)	Гарантия на один год (ограниченная) без дополнительной оплаты, девять часов в день с понедельника по пятницу (кроме праздничных дней), на следующий рабочий день, ремонт некоторых компонентов производится на месте; для остальных компонентов (в зависимости от страны) предоставляются заменяемые пользователем блоки. Доступны сервисные обновления и обслуживание по гарантии.
(*) Только конфигурация на 32 ядра (**) Каждый размещается поверх одного разъема PCI Express 8x и заменяет его. Доступные варианты конфигурации зависят от количества процессорных ядер и прочих факторов. (***) Более подробная информация о поддержке уровней ОС – в документе «Отчет о характеристиках и возможностях». (****) Масса зависит от количества установленных дисков, адаптеров и периферийных устройств.	

Архитектура системы

Когда говорят о Watson, то подразумевают систему, состоящую из трех компонентов:

- суперкомпьютера, работающего под управлением операционной системы Linux;
- связующего ПО, реализующего архитектуру UIMA (Unstructured Information Management Architecture);
- системы ответов на вопросы DeepQA, специально "заточенной" под Jeopardy!.

Центральной частью и, возможно, наиболее важной на последующую перспективу является UIMA.

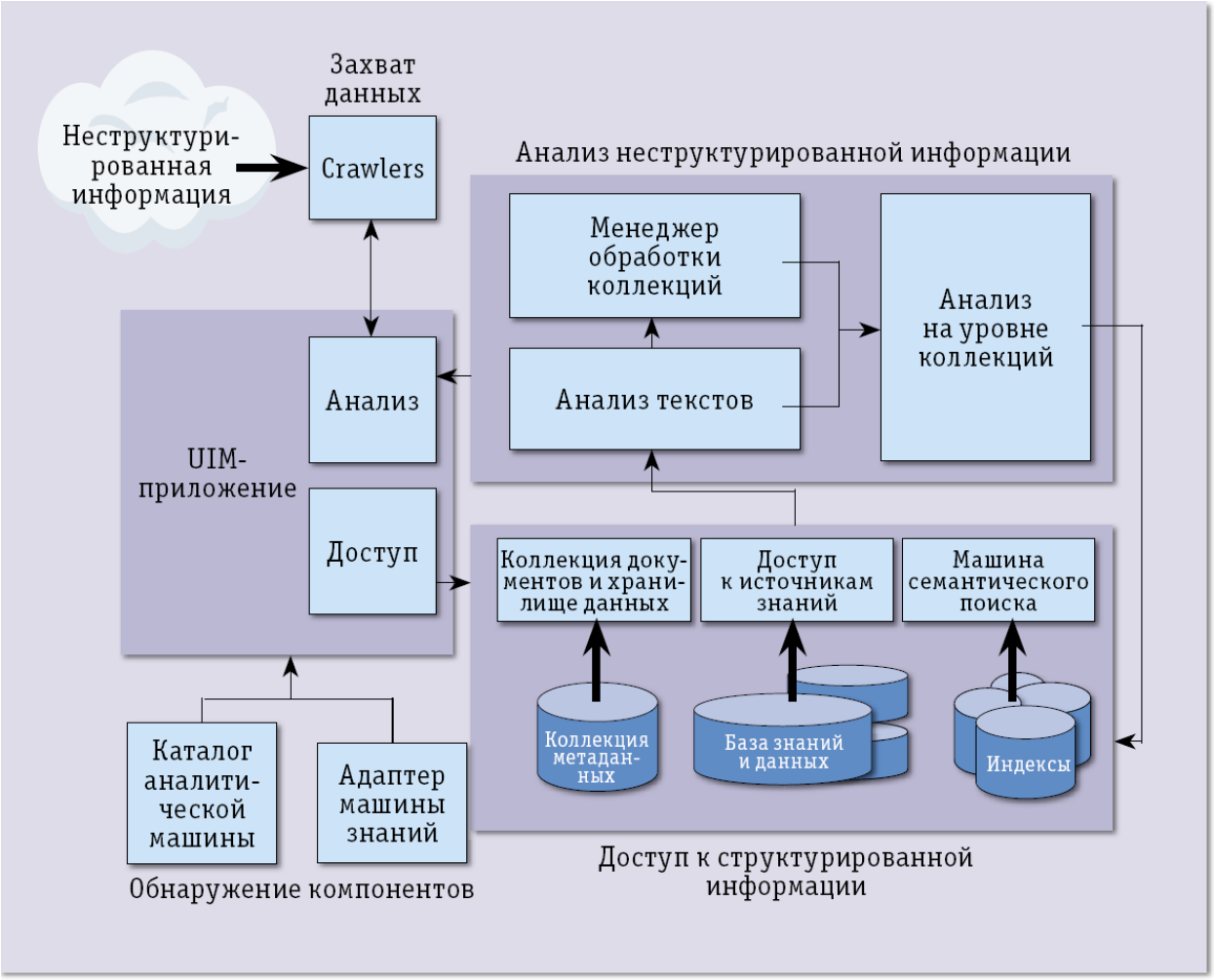
Вопрос-ответные системы (Question Answering, QA) предназначены для поиска точных ответов на вопросы, поставленные на естественном языке (Natural Language Processing, NLP). Важно подчеркнуть, что речь идет о точных ответах, человек-пользователь должен иметь возможность для однозначной интерпретации ответа, поэтому ответ может сопровождаться какой-то детализирующей или конкретизирующей информацией.

Источником сведений могут быть неструктурированные данные (книги, журналы, Web-страницы, блоги), квазиструктрированные (справочники, словари, энциклопедии, вики и ее аналоги) и базы данных. В Европе такие исследования объединяет организация Cross-Language Evaluation Forum (CLEF), а в Японии ведет рабочая группа NTCIR в рамках реализуемого Национальным институтом информатики проекта Cross-Language Information Retrieval (CLIR).

Разработка UIMA

Технология управления неструктурированной информацией (Unstructured Information Management, UIM) и соответствующая архитектура UIMA разрабатывалась в IBM Research еще с 90-х годов группой, насчитывавшей порядка 200 сотрудников. Их деятельность была сосредоточена на средствах для работы с NLP и включала поддержку диалога на естественном языке, выделение полезной информации, анализ текстов, классификацию документов, машинный перевод и вопрос-ответные системы.

Итогом стало создание связующего ПО, получившего название UIMA, которое может служить ядром для создания и внедрения распределенных аналитических машин (analysis engine), или UIM-приложений, позволяющих извлекать знания из неструктурированной информации, в том числе из текстов, аудио, видео и изображений.



Структура UIMA (рисунок выше) состоит из нескольких компонентов:

1. Захват данных (Acquisition) обеспечивает сбор документов из разных источников и формирование необходимых коллекций (collection), предназначенных для определенных приложений. Функцию захвата могут, например, осуществлять Web-пауки (web crawler), а также иные средства, какие именно, для приложений не важно, поскольку имеется специальный уровень интерфейса Collection Reader, связывающий приложения с коллекциями данных и метаданных.
2. Анализ неструктурированной информации (Unstructured Information Analysis) делится на два последовательных этапа — сначала выполняется анализ документов, а затем анализ коллекций документов. Входные документы обрабатываются текстовыми аналитическими машинами (Text Analysis Engine), в том числе трансляторами и модулями, выполняющими грамматический разбор, классификацию, обобщение. Используя входные документы, текстовые аналитические машины вырабатывают обобщенные аналитические структуры (Common

Analysis Structure). На этап анализа коллекций документы могут поступать напрямую или через промежуточный этап, на котором выполняется необходимая фильтрация и переформатирование для последующей параллельной обработки. Анализ на уровне коллекций (Collection Level Analysis) позволяет обобщить сведения, содержащиеся в коллекции документов.

3. Анализ структурированной информации (Structured Information Analysis) используется как для входных данных, поступающих в структурированной форме, так и для данных, появляющихся после анализа неструктурированной информации, где их значительная часть структурируется, с тем чтобы к ним можно было применить известные методы анализа. В результате аналитические механизмы, предназначенные для двух типов данных, оказываются охваченными общей петлей обратной связи.

В 2004 году были подведены предварительные итоги работ по созданию UIMA, и в качестве ближайшей цели была выбрана система подготовки ответов IBM Question Answering, которая к тому времени уже разрабатывалась в течение двух лет. В дальнейшем развитие UIMA продолжилось автономно от IBM.

В 2005 году правительство США спонсировало создание рабочей группы UIMA Working Group, объединившей несколько компаний и университетов, заинтересованных в создании фреймворка для решения задач NLP, в 2006 году IBM опубликовала исходные коды UIMA на портале Source Forge , а университет Карнеги-Меллона взял на себя поддержку этого репозитория. Сейчас коды UIMA доступны на сайте Apache Software Foundation .

В 2008 году был выпущен релиз Apache UIMA-AS (Asynchronous Scaleout), в котором к основной функциональности UIMA была добавлена возможность асинхронного масштабирования. Отличие этого релиза в том, что оригинальное монолитное решение Collection Processing Management заменено на решение, использующее Java Messaging Services и Apache ActiveMQ и поддерживающее обмен сообщениями. Деятельность сообщества привнесла еще одну разумную вещь — согласование Apache UIMA с Apache Hadoop. Обе эти новации были использованы при создании системы ответов DeepQA, обеспечивающей Watson способность играть в Jeopardy!.

Разработка DeepQA

Систему DeepQA разрабатывали 20 человек в течение трех лет. О значимости этой работы можно судить по тому, что ей присвоили имя основателя IBM Томаса Уотсона. Работа началась с фундаментального исследования самой игры и тактики игроков. Помимо таких очевидных задач, как генерация гипотез, сбор доказательств, анализ и численная оценка, авторам пришлось решать и специфичные задачи: улавливание иронии, обнаружение скрытого смысла и других человеческих особенностей. Поиск ответа на вопрос в игре совсем не похож на поиск данных в Web, здесь ищутся не сведения, а точный ответ, поэтому источником для поиска ответов служит собственная база данных, куда занесены и структурированные, и неструктурированные данные, собранные как в Интернете так и во множестве других источников.

Сегодня Watson уступает настоящим игрокам в том, что является системой класса NLP, то есть аудио- и видеоданные он пока не понимает.

В DeepQA используется более 100 различных методик анализа данных на естественном языке. Параллельно с главной целевой задачей, разумеется, разрабатывались и технологии широкого применения. В некотором смысле Watson все же паровой каток — в DeepQA загружено 200 млн страниц текстов, то есть он "как бы" прочел миллион книг. С таким объемом без Apache Hadoop явно не справиться, поэтому специальные программы, обеспечивающие аннотацию (в DeepQA их называют UIMA-аннотаторами), создают средствами Hadoop конструкцию MapReduce и распределяют задания по процессорам в кластере. Аннотаторы просматривают тексты и создают что-то вроде коротких рефератов, это позволяет осуществлять суждение о содержании. UIMA умеет согласовывать работу этих аннотаторов и собирать от них сведения, чтобы потом интегрировать, оценивать и тестировать.

С появлением версии UIMA-AS открылась возможность распараллеливания, и действие, которое требовало раньше два часа на одном процессоре, теперь выполняется в режиме реального времени. Кластер Watson может быть построен на процессорах Power7, ядра которых одновременно выполняют фрагменты DeepQA. В конфигурации, использовавшейся 14 февраля 2011 года, было объединено в кластер 90

Linux-серверов IBM Power 750 с 32 ядрами Power7/3,55 ГГц на каждом. Эти серверы собраны в десять стандартных стоек, укомплектованных коммутаторами и узлами ввода/вывода. Размер памяти — 16 Тбайт, производительность 80 TFLOPS. Сочетание высокой производительности ядра Power7 с памятью 512 Гбайт на ядро превращает аппаратную часть Watson в мощный инструмент для поддержки процессов, нуждающихся в большой памяти и высокой процессорной мощности. Преимущество Watson по сравнению с Deep Blue, который в свое время был собран из 30 узлов RS/6000 SP на процессорах Power2/120 МГц, состоит в том, что в последнем еще стояли 480 специальных шахматных процессоров, которые невозможно использовать ни для чего иного, а Watson собран из коммерчески доступных компонентов. Опыт его создания может быть распространен на другие приложения. Признание способности Watson понимать смысл и контекст сказанного на естественном языке, находить точные ответы на сложные вопросы может изменить представление людей о том, для чего могут быть использованы компьютеры.

И еще один важный момент, связанный с открытием кодов и их последующим использованием, — эволюция UIMA свидетельствует о рациональности подхода Open Source. Сначала была многолетняя исследовательская работа в стенах корпорации, потом стали доступны ее результаты. За время пребывания в открытом состоянии UIMA обогатилась асинхронным масштабированием Asynchronous Scaleout и поддержкой Hadoop, что существенно расширило функциональны возможности и сферу применения параллельных вычислений.

«Watson» не является нейросетевой системой в чистом виде, но использует важные и эффективные нейросетевые принципы в своей работе. Один из компонентов оценки правильности ответа определяется количеством общих слов между вопросом и предложением-гипотезой. Другой компонент основан на вычислении длины наибольшей общей последовательности между ними. Третий компонент оценки измеряет соответствие между логическими формами вопроса и найденного предложения, анализируя представление текста в виде графа, где узлы — это слова, а ребра — грамматические или семантические отношения между ними. Также учитывается контекст (принадлежность объекта к классу). В процессе обучения подбираются веса между компонентами окончательной

оценки так, чтобы максимизировать число правильных ответов на тестовом наборе вопросов.

Будущее проекта

IBM совместно с Nuance Communications планирует в ближайшие два года разработать продукт, направленный на помощь в диагностировании и лечении пациентов. Также рассматриваются возможности использования в других сферах, таких как оценка политик страхования или эффективности энергопотребления.

Помимо участия в викторине «Jeopardy» лежащая в основе системы Watson технология может быть адаптирована для преодоления реальных проблем и достижения прогресса в различных областях. Эта компьютерная система способна просеять огромное количество данных и дать точные ответы, сопровождаемые оценкой их достоверности. К примеру, эта технология может быть с успехом применена для повышения точности диагностирования пациентов в здравоохранении, для совершенствования онлайновых систем поддержки, функционирующих по принципу самообслуживания, для предоставления туристам и гражданам конкретной информации по населенным пунктам, для повышения качества поддержки клиентов по телефону, а также во многих других областях. В перспективе – новые данные о том, как работает мозг, лечение шизофрении, болезни Паркинсона и т.п.

Это обеспечит прогресс в следующих сферах:

- биологические науки и LifeSciences
- поиск новых лекарств
- новые материалы
- автомобили и самолеты
- окружающая среда и энергия
- финансы
  - оптимизация товаров
  - оценка рисков
- безопасность
  - военные исследования

- борьба с эпидемиями
- изучение природы
- Green технологии

Watson как услуга

Представители IBM сообщили весной 2012 года, что корпорация намеревается предлагать клиентам установку системы искусственного интеллекта Watson на базе частных и гибридных облаков.

Cloud computing (англ. Cloud — облако; computing — вычисления) — «облачные вычисления» — концепция «вычислительного облака», согласно которой программы запускаются и выдают результаты работы в окно стандартного веб-браузера на локальном ПК, при этом все приложения и их данные, необходимые для работы, находятся на удаленном сервере в интернете. Компьютеры, осуществляющие cloud computing, называются «вычислительным облаком». При этом нагрузка между компьютерами, входящими в «вычислительное облако», распределяется автоматически.

Облачные системы могут применяться как для обработки и анализа данных, необходимых для Watson, так и для работы самой Watson при хранении данных на платформе клиента. Но варианта с размещением Watson на серверах заказчика компания не предлагает — система будет продаваться только в виде сервиса. В IBM даже ввели для этого новый термин: WaaS — «Watson как услуга».

Другие применения

Медицина

Первым пользователем Watson уже стала медицинская страховая компания WellPoint; данный проект был развёрнут на инфраструктуре, представленной IBM, и его функцией является консультирование в сфере медицины и оценки рисков. В рамках проекта стало понятно, что диагностирование заболеваний не сильно отличается от процесса поиска ответов на вопросы викторины Jeopardy. В своем медицинском приложении «Watson» выдает перечень ответов, наряду с вероятностью их верности. Во время конкурса Jeopardy суперкомпьютер считал верным ответ, если

вероятность его верности составлял более 80 процентов, в случае с медициной считается, что существует достаточно большая вероятность того, что пациент поражен не самым очевидным видом заболевания, симптомы которого проявляются, поэтому Watson в списке ответов выдает результаты и с значительно меньшим процентом вероятности.

Способность понимания естественного языка дают «Watson» возможность оперировать совершенно новой для него категорией информации - неподтвержденными данными. Такие данные не являются абсолютно истинными, но в некоторых случаях являются исключительно полезными для постановки правильного диагноза. Суперкомпьютер способен самостоятельно "серфить" по просторам Интернета, по крупицам выискивая медицинские данные, которыми он пополняет свой банк данных.

Во время соревнования, проходившей в клинике Кливленда (Cleveland Clinic), «Watson» набрал в два раза больше баллов за точность выставления диагнозов, чем две команды опытных и уважаемых медиков-кардиологов клиники В ходе подготовки к соревнованию в «Watson» были введены тексты большого количества медицинских журналов, учебников, примеров и других данных. Это дало суперкомпьютеру самые обширные в мире знания в самых различных областях медицины.

Система IBM ”Watson” также начала использоваться в одном из ведущих мировых центров исследования рака – в онкологическом центре Memorial Sloan-Kettering в Нью-Йорке. ”Watson” будет анализировать огромное количество научных данных для ответа на вопросы о природе и лечении рака и предоставлять самую последнюю и актуальную информацию по этому вопросу.

Финансы

В начале марта 2012 г. также стало известно о заключении сделки между IBM и крупнейшим финансовым конгломератом Citigroup по созданию системы анализа ситуации на рынках. При этом связь с «Watson» будет осуществляться через интернет, а обработка информации будет осуществляться в облаке. Суперкомпьютер Watson будет работать на компанию Citigroup в виде удаленного сервиса облачных вычислений, что означает, что сам суперкомпьютер будет находиться на площадке компании IBM, а не в вычислительном центре Citigroup. Отныне IBM вообще не

планирует размещение системы на серверах предприятий-заказчиков - «Watson» будет предоставляться исключительно как сетевая услуга. IBM даже был введён специальный термин WAAS (Watson as service).

Согласно сообщению представителей Citigroup, суперкомпьютер Watson будет «анализировать текущие потребности покупателей, обрабатывать финансовую и экономическую информацию из различных источников, анализировать данные, поставляемые клиентами, что позволит поднять на совершенно иной качественный уровень область цифровых банковских и финансовых операций».

Вышесказанное, по всей видимости, подразумевает, что суперкомпьютер Watson будет постоянно заниматься анализом миллионов страниц всевозможной и разноплановой информации, предоставляя результаты специалистам компании Citigroup в удобном для восприятия виде. Уже сейчас специалисты Citigroup проводят операции по обучению искусственного интеллекта суперкомпьютера тонкостям финансового дела и специфическому жаргону, используемому на Уолл-стрит.

Аналитики IBM уверены, что данная услуга может получить широкое распространение, поскольку Watson может быть адаптирован для применения в различных областях. Наиболее эффективным он становится после периода обработки данных клиента и обучения на его задачах. В настоящее время, согласно заявлению генерального менеджера IBM Watson Solutions Маноха Саксены (Manoj Saxena), взаимодействие с «Watson» проходит в виде письменного диалога, в котором система задаёт дополнительные вопросы и запрашивает необходимую информацию, после чего выводит построенную логическую цепочку и правильный, с точки зрения машины, ответ.

Государственные службы безопасности

Директор ЦРУ Дэвид Петреус впервые публично оценил исключительную полезность новых «бытовых» технологий для шпионажа.

Все больше личных и бытовых устройств подключаются к интернету: от телевизора до навигационной системы автомобилей. Недавно глава самой крупной спецслужбы в мире Дэвид Петреус назвал новые технологии, широко распространяющиеся среди простых граждан,

«трансформационными» и имеющими огромное влияние на работу разведчиков.

Дэвид Петреус заявил, что новые онлайн-устройства - это сокровищница данных для его ведомства. Если раньше шпионам приходилось заниматься опасной слежкой и пытаться поставить прослушивающие устройства во всех местах, где бывает интересующий разведку человек, то сегодня ситуация кардинально изменилась. С появлением «умных домов», геопривязанных данных (из фото в социальных сетях, смартфонов, навигаторов и т.д.) ЦРУ в режиме реального времени может получать множество полезных данных: от местоположения человека, до тайной съемки через камеру мобильника или ноутбука.

В будущем возможности разведки в данной области еще больше вырастут, так как ожидается широкое распространение систем дистанционного управления и мониторинга, сенсорных сетей, чипов радиочастотной идентификации, серверов данных, встроенных в холодильники и даже кухонные комбайны и т.д. Вся эта электроника имеет большую вычислительную мощность и может выполнять нужные разведчикам «побочные» задачи. Большие надежды Петреус связывает с облачными сервисами, суперкомпьютерами и квантовыми компьютерами, которые смогут обрабатывать огромное количество информации, которую распространяет современный «продвинутый» пользователь.

Петреус подчеркнул, что бытовые устройства «меняют наши представления о тайне, личной информации и секретности». Так, ЦРУ имеет много юридических ограничений в отношении шпионажа за иностранными и особенно американскими гражданами. Но сбор данных вроде геолокации, интернет-статистики и т.п. – это «серая зона», где можно действовать свободно. Производители оборудования собирают огромное количество данных, и правительству очень легко следить за людьми через распространенные устройства вроде телефона или PlayStation.