ОБЛАЧНИТЕ ТЕХНОЛОГИИ ЗА РЕАЛИЗАЦИЯ НА ПРИЛОЖЕНИЯ С ГОЛЕМИ ДАННИ

Станислава Клисарова-Белчева
Пловдивски университет „Паисий Хилендарски“, Пловдив

Резюме

Комбинацията големи данни и облачни изчисления е поредното технологично предизвикателство за ИТ специалистите и бизнеса. В доклада се представя един възможен модел за съхранение и обработка на големи данни в облак. Основният фокус е в създаването на разпределена среда за работа с големи данни. Предложената рамка включва централизирана уеб услуга за управление на услуги, обслужващи различни видове бази данни.

Ключови думи: големи данни, облачни технологии, нерелационни бази данни.

Key words: big data, cloud technologies, non-relational databases.

JEL: C55, C8, D8.

Увод

През последните години облачните технологии и големите данни предизвикват нарастващ интерес в бизнеса и сред разработчиците на софтуер. В доклада се разглеждат характеристиките на големите данни, предимствата и недостатъците на облачните изчисления и се предлага модел с централизирана уеб услуга, която служи за преразпределение на информацията, за съхранение и обработка на големи данни. Оценени са предимствата и недостатъците на подобен модел, като се поставят основи за разработване на разпределени системи от услуги.

1. Големи данни

1.1. Определение

„Големи данни“ представляват колекция от данни, чиито обем и сложност правят управлението и обработката им трудна задача при използване на традиционните методи за обработка и съхранение (многомерни масиви, текстови файлове и/или релационна база данни). Въпреки че терминът големи данни се свързва често с обема данни, той има и друго значение. С големи данни се обозначава и технологията, включваща инструменти и процеси, чрез която организациите и компаниите се справят с нарастващия обем информация.

„Големите данни“ от една страна са подход за придобиване, анализиране и представяне на необходима информация, получена от данните, а от друга страна те са и набор от базови технологии, които позволяват правилното използване на вече получената информация.

Релационните бази данни преобладават при избора на модел за съхранение на финансови, производствени, лични и други видове данни. Софтуерните компании, които разработват системи за управление на релационни бази данни (СУРБД), постоянно развиват алгоритмите за управление на информацията с цел да отговорят на нарастващите изисквания на потребителите за работа с големи обеми от данни. Напредъкът в информационните и комуникационни технологии, цифровизацията на търговията, както и разпространението на социалните медии са предпоставка за генериране на огромни количества информация с нарастващ темп. Според наскоро проведено проучване [1] над 90% от данните в световен мащаб са генерирани през последните 2 години. Разнообразни сензори, сайтовете на социални мрежи и др. генерират огромен поток от информация, като цифрови снимки, видеоклипове, покупки, продажби, GPS координати и др.

Нарастването на обема данни в световен мащаб и иновациите в областта на технологиите са предпоставка събраната информация да се използва за вземане на икономически обосновани решения.

1.2. Основни характеристики на големите данни

Една от най-важните характеристики на големите данни е техният обем. Именно обемът на данните дава наименованието на този термин. Счита се, че големи са данните с обем, по-голям от 1 терабайт.

Друга основна характеристика е структурата им или по-точно липсата на такава. Структурата на големите данни се изменя с бързи темпове. Това се поражда от факта, че информацията се генерира от различни и разнообразни източници – социални мрежи, видеоканали, GPS данни и др., поради което структурата им е изключително динамична.

Многообразието на източници на информация води не само до неструктурираност, но и до възможност да липсва достоверност – данни които са съмнителни, генерирани са от софтуер или имат несигурен произход.

1.3. Основни характеристики на СУБД

Прогресът в областта на базите данни се дължи в голяма степен на релационния модел, предложен от д-р Едгар Код през 60-те – 70-те години на XX век. По същото време е създаден и езикът за манипулиране на данните SQL (Structured Query Language – език за структурирани заявки). Бързото изменение в структура на големите данни силно затруднява прилагането на релационния модел. Усъвършенстването му води до възникването на нови модели, често наричани NoSQL (Not only SQL). Едно от основните изисквания към тези бази е да са устойчиви на разделяне.

Изискванията към системите за управление на данни могат да се разделят:

- Съгласуваност (Consistency (**C**)) – съвместимост на данните. Всички потребители имат достъп до едни и същи данни по едно и също време.
- Достъпност (Availability (**A**)) – възприема се, като гаранция, че всяко запитване към СУБД ще генерира резултат. Ако времето за получаване на резултата е много голямо, то това може да се приеме като невъзможност за отговор или недостъпност на данните.
- Устойчивост за разделяне - възможност за разделяне между множество сървъри (Partition tolerance (**P**)) – системата може да се разпредели върху няколко независими сървъри и дори при загуба на някой сървър или пакети софтуерът ще продължи да работи.

Брюър представя своята CAP теорема през 2000 г. В нея той доказва, че е невъзможно една разпределена компютърна система да осигури едновременно и трите важни изисквания (C, A и P) (Фиг. 1). В доклад [8] на Нанси Линч, се дискутира всяко едно от изискванията в теоремата на Брюър.

Системите за управление на база данни (СУБД) често са класифицирани в зависимост от подхода на организация и начините за съхранение на информацията. Най-разпространени са релационните СУБД, в чиято основа е релационният модел. Тук данните се съхраняват във вид на релации. Двете основните характеристики, които осигуряват този вид база данни (БД) са достъпност и съгласуваност.

Разпределени бази данни са тези, които могат да бъдат разделени и репликирани. Към този тип се отнасят новаторските NoSQL решения. Те осигуряват свойството устойчивост на разделяне.

Главната задача на NoSQL е преодоляване на ограниченията на релационния модел. Основният проблем на релациите се забелязва с увеличението на количеството информация. Колкото се увеличава размерът, толкова нараства и времето за достъп до въведената информация.

Фиг. 1. CAP теоремата на Брюър

Теоремата на Брюър става основополагаща при избора на бази данни за реализация на софтуерни приложения.

1.4. Модели за съхранение и приложение на „Големи данни“

Прилагането на иновативни технологии при проектирането и разработването на бизнес софтуер е в центъра на успеха на организациите. Съвременните фирмени приложения предоставят възможности както за бизнес анализ на класически БД, така и за изследване на големи данни. Ето защо изборът на подходяща база данни се превръща в основна задача пред проектантите на фирмените информационни системи.

Със своите характеристики NoSQL СУБД са подходящи за съхранение и манипулиране с големи данни. Според архитектурата на обработка и съхранение на данните те се класифицират на документно ориентирани, ключ-стойност, колонно ориентирани, граф БД и др.

Бизнес анализът на големите данни подпомага устойчивото развитие на организациите и бизнеса, като предоставя разнообразни количествени оценки - на ръст, тренд на пазара и др.

2. Облачни технологии

2.1. Определение

Бързото развитие на интернет технологиите, превърна в реалност построяването на разпределени компютърни системи, осигуряващи синхронизация и цялостност на централна база данни. Подобни системи се изграждат с помощта на релационни бази данни, като синхронизацията се постига посредством услуги (сървиси) за репликация. Намаляването на разходите и увеличението на качеството и мощността на предлаганите услуги води до възникване на нова технология за съхранение и изчисления, наречена Cloud Computing (облачни изчисления, облачни

технологии или накратко облак). В облака ресурсите се предлагат като комунални услуги, които се отдават под наем, освобождават се или се наемат в зависимост от изискванията на софтуера и потребностите на компаниите. Софтуерните разработчици все по-често пренасят основни бизнес процеси и функции на организациите върху облачни платформи.

Облачните технологии представляват съвкупност от две основни тенденции в компютърните технологии [3]:

- ИТ ефективност, при която мощността на съвременните компютри се използва по-ефективно чрез силно мащабируем софтуер и хардуерни ресурси;
- Бизнес гъвкавост, позволяваща ценови модел при който се заплаща само за употребеното.

Ефективността и гъвкавостта се определят от възможността ресурсите да се конфигурират динамично. Колкото ресурси са необходими, толкова се заявяват и усвояват, след което се освобождават.

В доклад на Маккензи [4] се твърди, че „Облакът е съвкупност от хардуерно базирани услуги, които предлагат изчислителна мощност, мрежови услуги и хранилище на данни, като всички параметри могат да бъдат много еластични“.

Облакът може да се дефинира и като паралелна и разпределена изчислителна система, която представлява съвкупност от свързани помежду си виртуализирани компютри.

2.2. Характеристики и приложение

Приложението на облачните изчислителни модели е конкурентен инструмент заради възможността за бързо разгръщане на паралелна пакетна обработка. Интензивните изчислителни бизнес анализи и мобилни интерактивни приложения изискват от софтуера да отговаря в реално време на потребителските запитвания.

В свое изследване [7] Vaquero заявява „Облакът е като голям басейн за лесно използваеми и достъпни виртуализирани ресурси (хардуерни платформи и/или услуги). Тези ресурси могат да бъдат динамично преконфигурирани с цел да се приспособят към променливо натоварване, което позволява също така и оптималното използване на ресурсите.“

2.3. Видове облачни услуги

Най-често срещаните видове услуги са:

- Инфраструктура като услуга (IaaS) – както подсказва името осигурява цялостно компютърно оборудване, което може да е физическо или по-често виртуално, пространство за съхранение, защитни стени и др. като Windows Azure, Google Compute Engine.
- Платформа като услуга (PaaS) – обикновено включва операционна система, развойна среда, база данни, уеб сървър и др.
- Софтуер като услуга (SaaS) – бизнес приложения като услуга. В повечето случаи се използва като софтуер по заявка. Доставчикът осигурява софтуера, като поема отговорност за инсталация, конфигуриране и необходимите ъпдейти.

Освен тези основни услуги, облакът може да предостави и специфични услуги, например база данни като услуга, информация като услуга, тестове като услуги и др.

2.4. Предимства на облачните технологии

Предимствата на облака, могат да се разделят на няколко основни категории:

- Сериозно намаляване на разходите за хардуер. Заплащане само на това което се използва и възможност за мащабиране и динамична промяна на характеристиките;
- Софтуерната поддръжка в облака се извършва от доставчика. Доставчикът на услугите отговаря за актуализациите на софтуера, неговата защита и архивиране (резервни копия);
- Осигуряване на достъп без първоначални капиталови инвестиции;
- Намалява бариерите при използване и тестване на иновационни модели.

2.5. Недостатъци на облачните технологии

Въпреки многото позитиви и предимства на облачните технологии в решенията на компаниите, разработващи софтуер, съществуват и някои съществени недостатъци. В доклад [3] на Аврам недостатъците са класифицирани по следните признаци:

- Сигурност и поверителност – не е установено все още дали доставчиците на услуги в облака осигуряват адекватна защита на информацията;
- Свързване и достъп – пълният потенциал на използване на облачните услуги изисква високоскоростен достъп за всички потребители;
- Надеждност – бизнес приложенията изискват висока надеждност и не търпят сривове в работата си. Предпоставка за приемането би било, ако се създаде регистър на надеждност;
- Оперативна съвместимост – критична се оказва преносимостта на данните между частният и публичният облак.

Възможността за отдаване под наем на ресурсите намалява значително себестойността на услугата, което е рентабилно при изграждането на бизнес софтуер.

3. Големи данни и облачни технологии

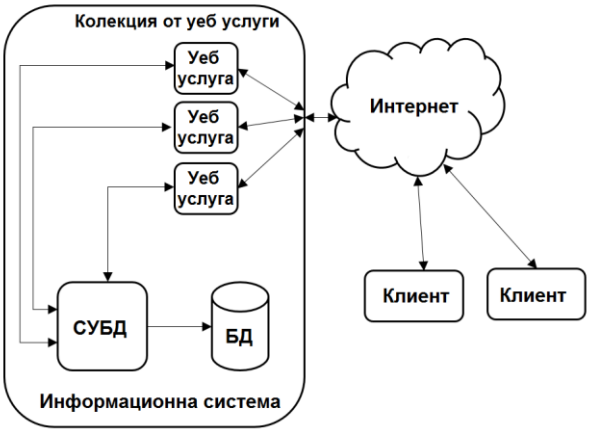
3.1. Комбинация на големи данни и облачни технологии

Реализацията на решения, съхраняващи и обработващи големи данни, изискват използване на NoSQL СУБД и множество компютри организирани в клъстер. Подобна архитектура изисква сериозни инвестиции, както за хардуер и софтуер, така и за екип, който да поддържа и развива инфраструктурата. Освен това ресурсите постоянно се амортизират. Предимството на реализацията на подобен модел е в това, че ресурсът може да се мащабира по всяко време, т.е. налице е възможност за съхранение и обработка на нарастващ обем данни. Използването на по-маломощни компютърни конфигурации може да намали разходите в сравнение с изграждането на скъп център за работа с данни. Тази идея се подкрепя и от основните възможности за скалиране и разпределение на NoSQL базите данни.

Използването на облачните технологии в частта си SaaS ще намали началното капиталовложение и ще предостави възможност за реализация на иновационни идеи. Освен това ще могат да се създават тестови установки без предварителна инвестиция за изграждането на инфраструктура от скъпоструващи центрове за работа с данни. Основната идея е облачните ресурси да се използват за съхранение и обработка на големи данни, като се внедри модел от уеб услуги за разпределяне и управление на входните данни, пресмятане на агрегати стойности и обобщения на данните. Облачните технологии и „големите данни“ са взаимно допълващи се технологии [6]. Основната задача при прилагането на големите данни в облака е изборът на архитектура за съхранение и методи за обработка и извличане на полезна информация.

3.2. Уеб услуги (Web services)

Уеб услугите представляват софтуерни системи за комуникация посредством функции и команди между взаимосвързани в компютърна мрежа устройства. Те представляват стандартни интерфейси между приложенията. Уеб услугите са форма на разпределена информационна система. Според [2] уеб услугите се опитват да решат редица ограничения при проектирането на разпределени системи, като различия в структурата на базите данни, достъп до ресурси и други.



Фиг. 2. Обща схема на информационна система с уеб услуги

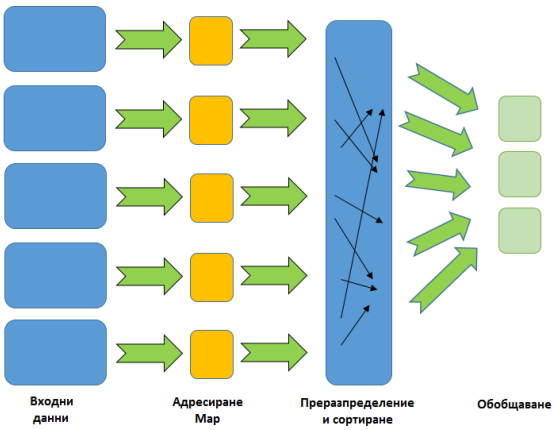
Уеб услугите служат за комуникация между информационната система и клиентите посредством стандартни съобщения. SOAP (Simple Object Access Protocol) и REST (Representational State Transfer) са два от начините за комуникация с уеб услугите. В SOAP комуникацията обменът на данни се извършва посредством обмен на XML съобщения. REST от своя страна осигурява алтернатива, при която предаването на данни използва уеб адрес и стандартни команди, като GET, POST, PUT, DELETE.

Уеб услугите позволяват независимост от платформата на приложенията, които ги използват (C#, Java и др.). Това предимство, подпомага в оперативна съвместимост между хетерогенни приложения, както и са лесни за реализация.

3.3. Модел на разпределена система с уеб услуги

Внедряването и разработката на корпоративни информационни системи за управление на околната среда (CEMIS) [5] може да се осъществи чрез набор стандартизирани уеб услуги. Подобни проекти, които включват обработка на големи данни, се реализират изцяло като архитектура, ориентирана към услуги (SOA – service oriented architecture).

Предимството на комбинацията големи данни и облак е в основата на модел, който комбинира уеб услуги за работа с данни и услуги за управление на големи данни. По този начин информацията, получена от различните източници, се съхранява независимо в различни бази данни в облака. Новият модел предлага сървис ориентирана система, съставена от колекция услуги за управление на различните бази данни. Реализацията на подобен модел включва разпределение и достъп до необходимата информацията. За тази цел е удачно да се използва моделът Map Reduce.



Фиг. 3. Модел Map Reduce

Map Reduce е вграден успешно в голяма част от NoSQL СУБД и предоставя възможност за разпределена обработка на големи данни. Основните фази на модела са три:

- Адресиране;
- Преразпределяне;
- Обобщаване.

Адресирането служи за разпределение на входната информация като всеки запис се обработва самостоятелно. При адресацията всеки запис получава ключ и стойност. Получената информация се сортира и преразпределя, за да се достигне до обобщението й.

Колекцията от уеб услуги е в основата на изграждането на разпределена система за работа с разнородни източници на данни - големи и структурирани данни. При необходимост облакът може да увеличава или намалява използваните ресурси, необходими за изграждането и поддържането на разпределена система от уеб базирани услуги.

За управлението на създадената колекция от уеб услуги се грижи централна уеб услуга, която комуникира с останалите чрез унифицирани команди в стандарт UDDI (Universal Description Discovery and Integration) и механизъм за взаимодействие [9].

Постъпването на данни от различни източници активира централната уеб услуга, която „пренасочва“ към услуги, носещи отговорност за съответния поток информация. Всяка услуга от колекцията се грижи както за съхранението и извличането на данни, така и за уеб услуги от тип актьори, които обслужват пресмятането на агрегатни стойности на входните данни. Предварително изчислените агрегати ускоряват извличането на резултатите и осигуряват бърз отговор от всеки поток при поискване. Тази колекция включва и сървисите за поддръжка на агрегатите.

Разпределението на входните данни подпомага съхранението на поверителни данни в частен облак или централен сървър, който може да предложи защитата им.

Централната уеб услуга може да осигури и данни за анализи за различни софтуерни системи и мобилни приложения. Обединението на уеб услугите в клъстер от услуги, подпомага унификацията на достъпа до разнотипни БД.

3.4. Предимства на модел от услуги и бази данни върху облачни ресурси

Улесненатa разработка на нови услуги, функции и възможности, както и намаленитe разходи за инфраструктура за работа с големи данни, е основно предимство на предложения модел за работа в облака.

Общодостъпността на облака позволява споделянето на обобщени резултати и анализи с външни организации в реално време.

4. Заключение

Въпреки несигурността на информацията от големите данни, ползите от приложението им при предвиждане и оценка на бизнес процеси оказват съществено влияние върху разработването на бизнес софтуер. Инвестицията в инфраструктура за големи данни може да се намали значително при прилагането на комбинация с облачните технологии. Изграждането на разпределена система от големи данни предоставя възможност за създаването на услуги, подобряващи ефективността на приложението им.

Литература

1. Ahmed Elragal. (2014). *ERP and Big Data: The Inept Couple*, Procedia Technology 16, 2014, p. 242–249.
2. Gustavo Alonso, Fabio Casati, Harumi Kuno, Vijay Machiraju, *Web Services: Concepts, Architectures and Applications*, ISBN 3-540-44008-9.
3. Maricela-Georgiana Avram. (2014). *Advantages and challenges of adopting cloud computing from an enterprise perspective*, 7-th ICIE ((INTER-ENG 2013). Procedia Technology 12 (2014), p. 529–534.
4. McKinsey & Co. (2009). *Clearing the Air on Cloud Computing*, Technical Report.
5. Tariq Mahmoud, Barbara Rapp, Sebastian van Vliet.(2013). *Web Service Integration within Next Generation CEMIS*, Procedia Technology 9, p. 282–290.
6. Wenhong Tian, Yong Zhao. (2015). *2 – Big Data Technologies and Cloud Computing Optimized Cloud Resource Management and Scheduling*.
7. Vaquero LM, Rodero-Merino L, Caceres J, Lindner M. (2009). *A break in the clouds: Towards a cloud definition*, SIGCOMM Computer Communications Review, 39: p. 50 55.
8. Seth Gilbert, Nancy A. Lynch. *Perspectives on the CAP Theorem*, http://groups.csail.mit.edu/tds/papers/Gilbert/Brewer2.pdf.
9. Preeti Marwahaa, Hema Banatib, Punam Bedic. (2013). *UDDI Extensions for Temporally Customized Web Services*, ICCI: Modeling Techniques and Applications (CIMTA).

IMPLEMENTATION OF CLOUD COMPUTING IN BIG DATA APPLICATIONS

Stanislava Klisarova-Belcheva
Plovdiv University Paisii Hilendarski, Plovdiv, Bulgaria

Abstract

The combination of big data and cloud computing is another technical challenge for IT professionals and business today. This paper presents a possible approach for storage and processing big data in cloud environment. The focus is on how to create a distributed framework for working with big data. The proposed model includes a central web service and auxiliary services for each database source.