|  |  |
| --- | --- |
|  | **Софийски университет „Св. Климент Охридски”**  Факултет по математика и информатика |

**Семинар по проектиране и интегриране на софтуерни системи**

***Летен семестър, 2017/2018 год.***

**ЕСЕ**

***Големи данни – дали всичко е свързано с cloud computing или инфракструктура. Възможности и предизвикателства***.

Изготвил:

Евгени Стойков, СИ, ФН: 61769

Съдържание

[Кaкво са големите данни? 3](#_Toc518239164)

[Откъде идва big data? 3](#_Toc518239165)

[Защо данните са важни? 3](#_Toc518239166)

[Как ще разбера дали данните ми са "големи"? 4](#_Toc518239167)

[Защо облакът е най-добрата платформа за големи данни? 4](#_Toc518239168)

[Разлика между Big Data и Cloud Computing 4](#_Toc518239169)

[Изборът между Spark или Hadoop не е ключът към успеха на големите данни. Избирането на подходящата инфраструктура е: 5](#_Toc518239170)

[Съществени съставки за големи анализи на данни 6](#_Toc518239171)

[Създаване на еластичност и скалируемост 8](#_Toc518239172)

[Предизвикателства и възможностти 9](#_Toc518239173)

# Кaкво са големите данни?

Терминът big data се отнася до данни, които обикновено са твърде скъпи за съхранение, управление и анализ, използвайки традиционни (релационни и / или монолитни) бази данни. Обикновено такива системи са неефективни по отношение на разходите поради липсата на гъвкавост при съхраняването на неструктурирани (като изображения, текст и видео), приемане на данни с "висока скорост" (в реално време) или мащабиране за поддържане на много големи ) обеми от данни. Поради тази причина през последните няколко години се наблюдава приемането на нови подходи за управление и обработка на big data, включително системите за бази данни на Apache Hadoop и NoSQL. Обаче тези опции често се оказват сложни за управление и използванe.

# Откъде идва big data?

Докато в миналото повечето данни за клиентите биха могли да бъдат категоризирани като добре структурирани (като например банкови) транзакции, днес организациите произвеждат ежедневно масивните транзакции под формата на неструктурирани онлайн данни за клиентите при взаимодействие. Неотдавнашното появяване на "Интернет на нещата", термин, описващ глобалната мрежа от милиарди взаимосвързани устройства и сензори, предизвика експлозия в обема на данните под формата на текст, видео, изображения и дори аудио. И накрая, в някои регулирани отрасли, достъпът до данни, които в противен случай биха били архивирани, често е необходим поради съображения за съответствие.

# Защо данните са важни?

Възможността за последователно получаване на бизнес стойност от данните сега е черта на успешни организации във всяка индустрия от всякакъв размер. В някои отрасли (като например търговията на дребно, рекламата и финансовите услуги, които постоянно се включват в списъка), това е дори въпрос на оцеляване. Анализът на данните връща повече стойност, когато имате достъп до повече данни, така че организациите в различни отрасли намират големите данни за богат ресурс за разкриване на задълбочени бизнес прозрения. И тъй като моделите за машинно обучение стават по-ефективни, тъй като те са "обучени" с повече данни, машинното обучение и големите данни са силно допълващи се.

# Как ще разбера дали данните ми са "големи"?

Въпреки че много предприятия все още не са достигнали petabytes мащаб по отношение на обемите данни, възможно е данните да имат една от другите две дефиниращи характеристики на големи данни. И ако има някаква гаранция, данните ще нарастват с течение на времето - вероятно, експоненциално. В този смисъл всички "големи данни" започват като "малки данни".

# Защо облакът е най-добрата платформа за големи данни?

Cloud computing предлага достъп до съхранение, обработка и анализ на данни по по-гъвкава, еластична, рентабилна и дори сигурна база, отколкото може да бъде постигната при локално разполагане. Тези характеристики са от съществено значение за клиентите, когато обемите на данни растат експоненциално - да направят необходимите ресурси за съхранение и обработка, както и да получат стойност от тези данни. Освен това, за онези организации, които току-що започват пътуването към извършване на големи анализи на данни и машинно обучение и които искат да избегнат потенциалните сложности на големите системи за данни на място, облакът предлага начин да експериментирате с управлявани услуги (като например Google BigQuery и Google Cloud ML Engine) по разходопокривен начин.

# Разлика между Big Data и Cloud Computing

Големите данни и Cloud computing, звучат познато, да! И двете са най-трендните термини в постоянния сектор на ИТ (информационните технологии) днес. Може да мислим, че и двамата термина правят същото, но всъщност, и двамата имат свои собствени начини да работят, за да изпълняват.

Където big data представляват голям обем от данни, които практически не могат да бъдат обработвани само от една машина, независимо дали данните са структурирани или неструктурирани. Cloud Computing, от друга страна, е нещо повече от приложение, което системно не само съхранява данни и програми, използващи мрежа от отдалечени сървъри по интернет, но също така предоставя услуги като SaaS, PaaS и IaaS.

Нека да навлезем малко по-дълбоко, за да получим по-голяма яснота и на двете.

А. По-голямата информация представлява просто огромни данни, структурирани и неструктурирани, които могат да бъдат допълнително обработени за извличане на информация. Огромни обеми данни във всички видове формати през интернет всяка секунда. Фирмите съхраняват тези данни, за да бъдат анализирани по-късно, като използват инструменти, увеличават резултата и производителността.

Cloud computing не е нищо друго освен множество отдалечени сървъри, които да управляват и обработват такива огромни обеми данни. В Cloud Computing не съхранявате данни и не изпълнявате услуги на локалните сървъри, а всичко е направено за вас някъде навън. Той осигурява платформата за споделяне на компютърни устройства за изпълнение на програми. Терминът "облак" означава мрежа от компютри, в които се съхраняват данни и се изпълняват приложения.

Б. Повече данни могат да бъдат структурирани или неструктурирани, които могат да бъдат допълнително обработени за анализ, като се използва анализ на данните. Cloud computing - Това е технология, използвана за съхраняване на данни и информация на отдалечен сървър, а не на физически твърд диск.

В. Големите данни носят скрити модели и алгоритми, които се отключват с помощта на различни инструменти, налични на пазара. Тези набори от данни се анализират допълнително, за да осигурят бизнес прозрения.

Cloud computing - Това е нова парадигма за изчислителни ресурси.

Г. Big данни могат да съществуват без изчислителни облаци. Cloud computing - Облакът изисква големи данни за изчислителни ресурси.

# Изборът между Spark или Hadoop не е ключът към успеха на големите данни. Избирането на подходящата инфраструктура е:

Големите данни не са за обработка в реално време или batch processing. Това не е въпрос на нито един от тях, както подчертават анализаторът на Овум Тони Баер и други. Предвид широкия спектър от възможности и натоварвания, които съставляват успешна стратегия за големи данни, това не е изненадващо или противоречиво. По-противоречива, макар и не е изненадващо, е естеството на инфраструктурата, необходима, за да се извлече максимума от големите данни. Например шефът на науката за данни на AWS (Amazon Web Services) Мат Ууд предупреждава, че докато "анализа е пристрастяващ", това положително пристрастяване бързо се превръща в неприятно, ако вашата инфраструктура не може да се справи. Ключът към големия успех на данните, твърди Ууд, е повече от Spark или Hadoop. Той е поддържането на двете върху еластичната инфраструктура. Вицепрезидентът на корпоративната стратегия в Hortonworks Шон Конъли се съгласява, че облакът има голяма роля в анализа на големи данни. Но Конъли вярва, че най-големият фактор при определянето на мястото, където се извършва голяма обработка на данните, е "гравитацията на данните", а не еластичността. Основният двигател за големи разгръщания на данни, казва Конъли, е да разшири традиционните системи на място, като например data warehouses. В крайна сметка това води до големи организации, които да качат Hadoop и други клъстери за анализ на няколко места. Въпреки това, Connolly признава, че облакът става все по-популярен вариант за разработването и тестването на нови приложения за анализи и за обработката на големи данни, генерирани "извън четирите стени" на предприятието.

# Съществени съставки за големи анализи на данни

Докато големите клиенти на AWS варират от пъргави компании като Reddit до мащабни предприятия като Novartis и Merck, Wood предлага три ключови компонента на всяка система за анализ.

А. Един източник на истината.

AWS предоставя множество начини за съхранение на този единствен източник на истината, от S3 съхранение до бази данни като DynamoDB или RDS или Aurora за решения за съхранение на данни като Redshift.

Б. Анализ в реално време.

Ууд казва, че компаниите често увеличават този единствен източник на истината с данни от стрийминг, като кликване върху уебсайтове или финансови транзакции. Докато AWS предлага Kinesis за обработка на данни в реално време, съществуват и други опции като Apache Storm и Spark.

В. Специализирани клъстери за задачи.

Клъстерите за задачи са група от копия, изпълняващи разпределена рамка като Hadoop, но се насочват специално към специална задача като визуализация на данните.

Имайки предвид тези компоненти, Wood потвърждава, че големите данни не са въпрос на batch processing или обработка в реално време, а по-скоро за широк набор от инструменти, които ви позволяват да обработвате данни по многостранен начин:

Това не е въпрос за Spark или Hadoop. Това е въпрос на "и" не "или". Ако използвате Spark, това не бива да изключва използването на традиционните MapReduce в други области, или Mahout. Можете да изберете правилния инструмент за работа, вместо да се опитваме да поставим квадратно купче в кръгла дупка.

Както Wood го вижда, "обработката на данни в реално време играе определена роля, но тя е добавка към голямата екосистема за данни".

Това отразява нещо, което създателят на Hadoop Doug Cutting каза в интервю, в отговор на въпрос дали стрийминг или обработката на данни в реално време ще изместят опции като Hadoop:

Не мисля, че ще има гигантска промяна към стрийминг. По-скоро стриймингът сега се присъединява към пакета от възможности за обработка, които имат хора на тяхно разположение. Когато се нуждаят от интерактивен BI, те използват Impala; когато се нуждаят от фасетирано търсене, те използват Solr; и когато се нуждаят от анализи в реално време, използват Spark Streaming и т.н. Хората ще продължат да извършват и ретроспективни партидни анализи. Един зрял потребител на платформата вероятно ще използва всички тези.

Connolly от Hortonworks вижда подобно бъдеще. Hadoop се възползва от предприятията като начин за разширяване на data warehouses и улесняване на анализа в съществуващите силози за приложение при драстично по-ниски разходи. Но тъй като клиентите стават по-усъвършенствани, нови източници на данни, нови инструменти и често облакът се прибавят към сместа: Ако мислите за случаи на бизнес употреба около 360-градусовия изглед (който консолидира данните за клиентите или продуктите в различните силози), това може да е on - premise. Но машинното ви обучение и откриването на данни може да са в облака. Възможно е да имате нови набори от данни като данни за времето и данни от преброяването, които вероятно не сте имали във вашите “четири стени”, затова може да искате да смесите това с някои от съществуващите си данни, за да извършвате усъвършенствано машинно обучение.

Тъй като законите на физиката забраняват лекото движение на стотици терабайта или петабайтови данни в мрежата, Connolly казва, че клиентите ще разполагат с клъстери на Hadoop на on - premise и на различни облаци, за да могат да правят подходящите анализи там, където по-голямата част от данните са пристигнали , Неговият термин за това е "гравитацията на данните". Когато новите набори от данни - като данни за времето, данни от преброяването и данни за машините и сензорите - произхождат извън предприятието, облакът се превръща в естествено място за обработка.

# Създаване на еластичност и скалируемост

Докато мнозина погрешно вярват, че големите данни са въпрос на огромни обеми данни и пренебрегват по-често срещаните сложности, присъщи на разнообразието и скоростта на данните, дори и обемът като сложност не е толкова прост проблем, до колкото някой мислят.

По думите на Wood от Amazon предизвикателството на големите данни "не е толкова за абсолютния мащаб на данните, а за относителния мащаб на данните". Тоест, докато проект като Проекта за човешкия геном може да започне като гигабайт-мащабен проект, той бързо се появи в терабайтен и след това в петабайтов мащаб. "Клиентите ще работят за мащабите, които текущо преживяват", отбелязва Wood, но когато скалируемостта трябва да се промени, предприятията стават напълно неподготвени.

Ууд казва: "Тези, които излизат и купуват скъпа инфраструктура, установяват, че проблемът с обхвата и домейнът се променят много бързо. Докато се доближават до първоначалния въпрос, бизнесът продължава.

С други думи, "Предприятията искат платформа, която благосклонно им позволява да се преместят от един мащаб в следващ и следващ. Просто не можете да получите това, ако пуснете огромна част от промените в център за данни, който е замразен във времето. "

Като пример, Ууд показва Weather Channel, където имаше само няколко милиона места, на които да съобщаваt времето на всеки четири часа. Сега той има милиарди и актуализации на всеки няколко минути на AWS, всички със 100% up - time. С други думи, това не е само за обработка на големи данни, но и за предаване на тези данни в облак.

За Connolly на Hortonworks, гъвкавостта на облака е толкова важна, колкото еластичността. Особено в случая с машинното обучение, можете да придвижите достатъчно данни, за да може устройството за машинно обучение да работи, което ви позволява да създадете своя модел за решение в облака. Този модел ще бъде използван в по-широко приложение, което може да бъде използвано другаде.

"Облакът е чудесен за типовете: "позволете ми да докажа моята концепция” или ”позволете ми да започна някои от моите идеи". След като това стане, въпросът става: Ще това ще се промени с on – premise технология, защото това е мястото, където по-голямата част от данните би трябвало да са или ще останат в облака? "

В крайна сметка, това не е дилема между "всичко в облака" и "всичко в on-premise.".Важна е стратегията предприета за ефективна обработка на данните. Става ясно, че най-успешните големи стратегии за данни ще включват редица големи информационни технологии, които се изпълняват в облака.

# Предизвикателства и възможностти

Big Data е термин, който обхваща всички видове данни, които съществуват днес. От болничните записи и цифровите данни до огромното количество документи на правителството, което е архивирано - има повече от това, което официално знаем.

Не може да се категоризират big data под една дефиниция или описание, защото ние все още работим върху нея. Великолепното нещо на информационните технологии е, че винаги са били достъпни за технологични компании, фирми и всички видове институции.

Тогава беше нуждата от появата на изчислителни облаци, които направиха по-лесно осигуряването на най-доброто от технологиите в най-рентабилните пакети. Cloud computing не само намали разходите, но и направи широка гама от приложения, достъпни за по-малките компании.

Точно както облакът непрекъснато расте, ние също така забелязваме експлозия на информация в мрежата. Социалните медии са съвсем различен свят, където и търговците, и обикновените потребители, генерират натоварвания с данни всеки ден. Организациите и институциите също създават данни на дневна база, което в крайна сметка може да стане трудно за управление. Ето статистика за генерирането на големи данни през последните пет години;

* 2,5-те quintillion байта (2,3 трилиона гигабайта) данни се създават всеки ден.
* До 2020 г. ще бъдат създадени 40 zettabytes (43 трилиона гигабайта).
* Повечето компании в САЩ имат поне 100 терабайта (100 000 гигабайта) съхранени данни.

Тези големи обеми данни представляват предизвикателство за облачната среда.Възниква въпроса - как да управляваме и защитим същността на тези данни, а не просто да ги подреждаме? Изглежда, че облачните технологии и големите данни са идеална комбинация за това. Заедно те осигуряват едно решение, което е едновременно мащабируемо и приспособимо за големи данни и бизнес анализи. Предимството в анализа ще бъде огромно предимство в днешния свят. Всички информационни ресурси стават лесно достъпни. Всяко поле на живот може да се възползва от тази информация. Нека разгледаме тези предимства в детайли:

А. Бързина:

Традиционната инфраструктура за съхранение и управление на данни се оказва по-бавна и по-трудна за управление. Това може буквално да отнеме седмици, за да инсталирате и управлявате сървър. Cloud computing е тук и може да осигури на вашата компания всички необходими ресурси. Базата данни за облаци може да даде възможност на компания да има хиляди виртуални сървъри и да ги работи безпроблемно само за броени минути.

Б. Достъпност:

Cloud computing е прикрито благословение за компания, която желае да има актуализирана технология в рамките на бюджет. Компаниите могат да избират това, което искат, и да плащат за тях, докато съществуват. Необходимите ресурси за управление на големи данни са лесно достъпни и не струват големи пари. Преди облака компаниите инвестираха огромни суми в създаването на ИТ отдели и след това платиха повече пари, за да поддържат този хардуер актуализиран. Сега компаниите могат да хостват своите големи данни на не собствени сървъри или да плащат само за място за съхранение и мощност, която използват на всеки час.

В. Обработка на данни:

Експлозията на данните води до въпроса за обработката на данните. Само социалните мрежи генерират натоварване с неструктурирани хаотични данни като съобщения, публикации, снимки, видеоклипове и блогове, които не могат да бъдат обработени в една категория. С платформите на Big Data Analytics като Apache Hadoop могат да се обработват структурирани и неструктурирани данни. Cloud computing прави целия процес по-лесен и достъпен за малки, средни и големи предприятия.

Изпълнимост:

Докато традиционните решения биха изисквали добавяне на повече физически сървъри към клъстера, за да се увеличи мощността и пространството за съхранение, виртуалната природа на облака позволява привидно неограничени ресурси при поискване. С облака, предприятията могат да мащабират нагоре или надолу до желаното ниво на мощност и пространство за съхранение лесно и бързо.

Анализите Big Data изискват нови изисквания за обработка на големи набори от данни. Търсенето на обработка на тези данни може да се повиши или да намалее по всяко време на годината, а облачната среда е перфектната платформа за изпълнение на тази задача.

Предизвикателства пред големите данни в облака:

Точно както Big Data предостави на организациите с терабайти данни, те също така представиха въпроса за управлението на тези данни в рамките на традиционната рамка. Как да анализираме голямата сума от данни, за да вземем само най-полезните бита? Анализирането на тези големи обеми данни често става трудна задача. В ерата на високоскоростната връзка, преместването на големи набори от данни и осигуряването на необходимите данни за достъп до тях също е проблем. Тези големи набори от данни често носят чувствителна информация като номера на кредитни / дебитни карти, адреси и други подробности, което повишава загрижеността за сигурността на данните.

Въпросите за сигурността в облака са основна грижа за бизнеса и доставчиците на облаци днес. Изглежда, че нападателите са безмилостни и продължават да измислят нови начини за намиране на входни точки в системата. Други проблеми включват ransomware, който засяга дълбоко репутацията и ресурсите на компанията, атаките срещу отказ на услуга, фишинг атаките и злоупотребата с облаци. В световен мащаб 40% от фирмите са имали инцидент с ransomware през изминалата година. Както клиентите, така и доставчиците на облак имат свой собствен дял от рисковете при сключването на споразумение за облачните решения. Несигурните интерфейси и слабите приложни програмни интерфейси (API) могат да предоставят ценна информация на хакери и тези хакери могат да злоупотребят с тази информация по погрешни причини. Актът за данни също е сериозен проблем, който изисква центровете за данни да бъдат по-близо до даден потребител, отколкото доставчик. Репликацията на данните трябва да се извършва по начин, който оставя нулево място за грешка; в противен случай това може да повлияе на етапа на анализ. От решаващо значение е търсенето, споделянето, съхранението, трансферът, анализът и визуализирането на тези данни да бъдат възможно най-гладки. Единственият начин да се справим с тези предизвикателства е да приложим технология от следващо поколение, която може да предскаже проблем, преди да причини повече щети. Модели за откриване на измами, криптиране и интелигентни решения са изключително важни за борбата с нападателите. В същото време е Ваша отговорност да притежавате вашите данни и да ги предпазвате от ваша страна, докато търсите бизнес интелигентни решения, които могат да осигурят стабилна възвръщаемост на инвестициите.