Počet dat na internetu se v dnešních dnech pohybuje v řádech tisíců exabytů, což je téměř nepředstavitelné množství dat různých formátů a informačních hodnot. Pouze zlomek těchto dat však poskytuje relevantní informace, které mohou být analyzovány a lze je využít. I tento zlomek je však obrovským množstvím dat, které je třeba umět zpracovat a provádět nad nimi operace.

Definice

Jsou to právě tato hodnotná data, o které mají společnosti zájem a která jsou často označována pojmem BigData. Velikost těchto dat však není jediným, co je specifikuje, je to i rozličnost jejich formátů a způsobu, jakým je třeba taková data zpracovat, rychlost jejich růstu a především neschopnost běžných softwarových nástrojů takováto data spravovat a zpracovávat v rozumných časových intervalech.

Vznik a historie

S velkými objemy dat pracují firmy typu telekomunikačních operátorů již dlouho a poměrně snadno. To co je však nutné podotknout je, že tato data jsou centralizovaná, velmi dobře strukturovaná a snadno zpracovatelná a tak nejsou problémy s jejich zpracováním. Navíc většinou docházelo k jejich zpracování postupně a pomalu.

Příchod sociálních sítí do internetových vod, spolu se stoupajícím zájmem o tyto produkty, však přinesl obrovské objemy špatně, či dokonce vůbec, strukturovaných dat, která se nedají snadno zpracovat a jsou velmi distribuovaná. Navíc se přidal tlak na efektivitu a rychlost zpracování a běžné softwarové nástroje tak přestaly stačit.

Vznikaly tak postupně různé nástroje umožňující zpracování BigData, z nichž nejpoužívanějším zástupcem byl Hadoop. Tato aplikace je open source frameworkem pro zpracování, analýzu a správu nestrukturovaných dat o velkých objemech. Její výpočetní vrstva, založená na technologii MapReduce od firmy Google, přistupuje k datům uložených na různých úložištích v různých formátech a zpracovává je na více uzlech najednou. Výsledky se pak spojí, uloží např. do formátu HDFS a odtud mohou být načtena pro analýzu.

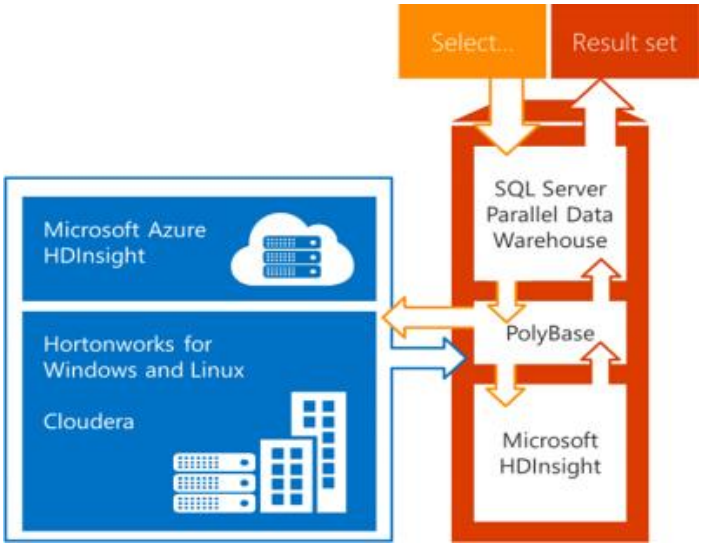
Současnost

V současné době patří mezi klasické projekty pracující s BigData především analýzy sociálních sítí a jejích uživatelů, analýzy textů či práce s daty při televizních přenosech. Práce s BigData je však stále velmi finančně náročná a tak tyto projekty provádějí především velké společnosti.

Postupem času se ukázalo, že Hadoop je sice skvělým nástrojem pro zpracování BigData, avšak jeho specializace na BigData způsobuje vytvoření dvou odlišných oblastí zpracování dat ve společnosti, jedné pro data strukturovaná a druhé pro ta nestrukturovaná.

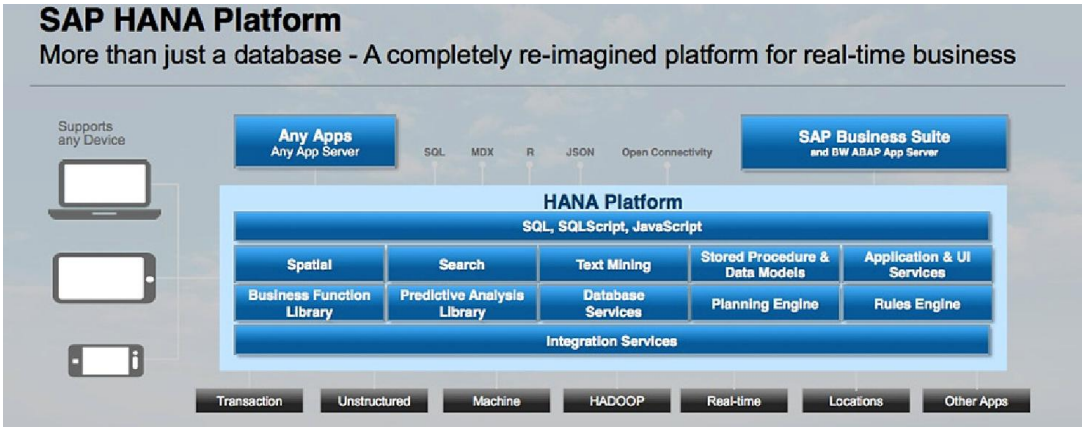
Trendem posledních let tak je vývoj a využívání nástrojů a struktur, které jsou schopné zpracovávat jak data nestrukturovaná, tak ty strukturovaná a sjednocují tak tyto dvě oblasti. Zároveň musí takové nástroje splňovat vysoké nároky na dostupnost dat a rychlost jejich zpracování. V následujících odstavcích budou představeny některé z takovýto nástrojů využívané velkými korporacemi.

Prvním představeným zástupcem je platforma APS vytvořená společností Microsoft, která v sobě spojuje nástroje Hadoop a SQL Parallel Data Warehouse a umožňuje tak pracovat s BigData klasickými nástroji pro správu SQL serverů. Její strukturu lze vidět na obrázku 1.



Obrázek 1 - Microsoft APS

Společnost SAP přistupuje k dané problematice s důrazem na zpracování jakéhokoliv typu dat v reálném čase a pro tuto příležitost představila nástroj SAP Hana. SAP Hana je databázový systém podporující velké množství typů a struktur dat a umožňuje jejich snadné zpracování a předání dalším nástrojům. Obrázek 2 ilustruje, jaké vrstvy tento databázový systém obsahuje.



Obrázek 2 - SAP Hana

Další velká společnost, IBM, přistupuje ke zpracování BigData podobným způsobem, jako společnost SAP. Její architektura Watson Foundations rovněž dokáže zpracovávat jak data strukturovaná, tak BigData a navíc, narozdíl od SAP Hana, ještě podporuje prediktivní analýzu. Podobu této architektury si lze prohlédnout na obrázku 3.

Obrázek 3 - Watson foundations

Co přinese budoucnost

I v současnosti jsou nástroje pro zpracování BigData, a práce s tímto typem dat celkově, velmi finančně náročné, avšak postupně se objevuje čím dál více dostupnějších řešení a v budoucnu tak lze očekávat výrazné zlevnění těchto technologií.

Samotná BigData jsou stále velmi nestrukturovaná a pouze část z nich nese validní informace a proto se v budoucnu dá očekávat zaměření na kvalitu informací namísto kvantity a nové způsoby "čištění" dat od nepodstatných informací.

Stejně tak důležité bude vyvinout nástroje, které dokážou s BigData pracovat, aniž by ohrožovala integritu a čistotu původních dat a zároveň budou reagovat ve velmi krátkých intervalech, ideálně real-time, jelikož BigData se začne využívat v oblastech typu trenérských analýz během sportovního utkání, kde je třeba umět získaná data zpracovat a zobrazit velmi rychle.

Zdroje

http://www.systemonline.cz/business-intelligence/big-data.htm

http://www.systemonline.cz/business-intelligence/big-data-od-velkych-ocekavani-k-praktickemu-vyuziti.htm

http://www.cmswire.com/cms/big-data/bigger-better-faster-stronger-the-future-of-big-data-027026.php

Doporučené čtení

Milan Nikl - Hadoop.pdf|Zobrazit podrobnosti