Generátor Big Data

Vedoucí: Doc. RNDr. Irena Holubová, Ph.D.
Kontakt: holubova@ksi.mff.cuni.cz
Předpokládaný počet členů týmu: 5

Motivace

Big Data je pojem, který označuje datové množiny velkého rozsahu, rychlosti nárůstu, různorodosti a nejisté věrohodnosti (v angličtině tzv. "4 V", tedy **v**olume, **v**elocity, **v**ariety a **v**eracity), které vzhledem k jejich vlastnostem není možné zpracovávat pomocí konvenčních metod a technologií. Současný boom v návrzích efektivních metod pro správu a zpracování Big Data ovšem naráží na problém kde taková data získat. Nejčastější důvody zahrnují:

- *omezení přístupu k datům*, kdy firma, která data vlastní, je typicky nemůže nebo nechce dát (zdarma) k dispozici a
- *velikost dat*, kdy data, která máme k dispozici, nejsou tak velká, jak se z počátku zdálo nebo to zatím nedovolují podmínky (např. nemáme doposud dostatečné množství senzorů nebo jiných zařízení).

Cíl projektu

Cílem projektu je navrhnout a implementovat generátor Big Data. Nástroj bude implementován jako obecný framework, který bude možné pomocí nových plug-inů dále rozšiřovat a modifikovat. Generátor bude pracovat zhruba v následujících krocích:

1. *Analýza vstupních dat*: Předpokládáme, že uživatel jako vstup poskytne existující netriviální množinu ukázkových dat, která chce "nafouknout" do kategorie Big Data.
2. *Vytvoření popisu (gramatiky) vstupních dat*: Pro každý formát dat (relační, XML, JSON, CSV, ...) bude existovat samostatný plug-in, který data zanalyzuje a vytvoří jejich gramatiku.
3. *Uživatelsky příjemná vizualizace a editace gramatiky*: Vygenerovanou gramatiku může uživatel upravit (např. zobecnit, zpřesnit), doplnit, popř. i vytvořit zcela novou (pokud nemá ukázková data). Vizualizace bude možná v textové i grafické podobě (např. obdoba grafické vizualizace Backus–Naurovy formy specifikace SQL), s podporou highligtingu, popř. jiných prvků zpřehledňujících práci s gramatikou.
4. *Parametrizace výstupu*: Uživatel dále specifikuje množství požadovaných dat, umístění dat, clusterování dat podle zadaných parametrů (specifické pro daný vstupní formát) apod.
5. *Efektivní (paralelní) generování výstupních dat*: Generátor dat bude nabízet možnost sledování průběhu generování, možnost přerušení generování + modifikace parametrů apod.

Výsledný SW by měl být při defaultním nastavení jednoduše použitelný (nenáročný uživatel zadá pouze cestu k vstupním datům, umístění výstupních dat a jejich velikost), ale současně použitelný i pro náročnější uživatele, kteří chtějí generování dále ovlivňovat pomocí vlastních nastavení. Díky plug-inům bude možné doplnit podporu dalších formátů, resp. využít již

existujících modulů. Při návrhu bude kladen důraz na maximální obecnost a znovuvyužitelnost implementovaných modulů.

Předpokládaný postup práce

Fáze I. SW projekt

V rámci SW projektu bude navržena architektura frameworku, vstupy/výstupy, rozhraní jednotlivých plug-inů atd. Dále bude implementováno jádro frameworku, uživatelsky příjemná vizualizace dat, podpora uživatelských vstupů a jednoduché metody odvozování gramatiky a generování dat pro alespoň 3 různé jednodušší typy formátů.

Fáze II. Diplomové práce

V rámci případných navazujících diplomových prací (jejich realizace může probíhat při nebo po realizaci SW projektu) bude možné např. navrhnout a implementovat sofistikovanější metody pro analýzy netriviálních typů formátů (stromové, semi-strukturované, grafové, ...), efektivnější metody generování dat, podporu neobvyklých formátů jako je např. NASA formát FITS apod.

Literatura

1. *BigDataBench, A Big Data Benchmark Suite*. ICT, Chinese Academy of Sciences. http://prof.ict.ac.cn/BigDataBench
2. Zijian Ming, Chunjie Luo, Wanling Gao, Rui Han, Qiang Yang, Lei Wang, Jianfeng Zhan: *BDGS: A Scalable Big Data Generator Suite in Big Data Benchmarking*. CoRR abs/1401.5465 (2014). http://arxiv.org/abs/1401.5465
3. Sherif Sakr Mohamed Gaber: *Large Scale and Big Data: Processing and Management*.
4. Tilmann Rabl: *Big Data Generation*. Middleware System Research Group, University of Toronto.
5. Transaction Processing Performace Council (TPC): http://www.tpc.org/
6. http://fits.gsfc.nasa.gov/