MODULE FOR QUERYING AND MINING FROM BIG DATA

Pavel Janečka

Master Degree Programme (2nd), FIT BUT

E-mail: xjanec11@stud.fit.vutbr.cz

Supervised by: Martin Hlosta

E-mail: ihlosta@fit.vutbr.cz

Abstract: With amount of data needed to be processed increases, the term "Big Data" becomes more and more important. This paper explains utilization of the large-scale data in terms of querying and knowledge discovery from data. It briefly introduces Malware Analysis System and its relationship to the module for querying and mining from Big Data. I studied Big Data approaches and chose Hadoop as the most convenient environment. I designed the module and implemented it with respect to the restrictions such as the need of serialization. The biggest asset of this module is that it allows the Malware Analysis System to tackle Big Data.

Keywords: Big Data, Large-scale Data, Knowledge Discovery, Data Mining, Hadoop, Hive, Mahout, Malware Analysis System, Map Reduce, Apache Thrift

1. ÚVOD

V dnešní době se již nepozastavujeme nad produkty informačních technologií, jako jsou sociální sítě, elektronické obchody či všudypřítomné internetové reklamy. Čas od času banka upozorní klienta, že s jeho účtem bylo podezřele nakládáno, potenciálnímu zákazníkovi leckterý e-shop nabídne zboží na základě chování ostatních zákazníků, samozřejmostí jsou také internetové vyhledávače. Každý z těchto příkladů silně staví na pojmu „Big Data" [1] (též Large-scale data, Rozsáhlá data), nezměrném množství dat, které není možné uchovávat ani zpracovávat konvenčními metodami.

Big Data lze zjednodušeně popsat pomocí tzv. „principu 3xV", což znamená Volume, Velocity, and Variety (objem, rychlost a rozmanitost). Zjednodušeně charakterizuje Big Data: objemem rozumíme jejich již zmíněné velké množství, rychlost se velmi blízce dotýká problematiky proudů dat a jejich zpracování (známé též jako „Complex Event Processing") a rozmanitost poukazuje na fakt, že data, která agregujeme, často pochází z různých zdrojů a jsou tudíž uspořádána v různých formátech, často dokonce nepopisují úplně stejné domény.

U Rozsáhlých dat není možné dodržovat kritérium ACID (z OLTP systémů, atomicita, konzistence, izolovanost, neměnnost). Namísto něj se objevuje tzv. „CAP teorém" (konzistence, dostupnost, zotavení – v angličtině consistency, availability, partition tolerance), který tvrdí, že nelze zároveň plně uspokojit tyto tři požadavky, nýbrž stojí proti sobě a splňujeme jeden na úkor ostatních.

Samotné uchování Rozsáhlých dat je pozoruhodný koncept, charakteristické je distribuované uložení, ovšem skutečnou hodnotu z dat získáme, zpracujeme-li je a dostaneme např. informace nebo znalosti. Proto je důležité zabývat se otázkou, jak s rozsáhle škálovatelnými daty pracovat a jak z nich získávat znalosti. To je také hlavním cílem této práce, vytvořit modul, jehož přínosem bude zpřístupnění distribuovaných dat a dolování z nich, a to tak, aby jej bylo možno využít v Malware Analysis System [2].

1.1. MALWARE ANALYSIS SYSTEM

Malware Analysis System (MAS) je flexibilní rozšiřitelný systém umožňující provádět rozsáhlé analýzy nad statickými i proudovými daty. Systém vznikl na Fakultě Informačních Technologií

VUT v Brně, je stále vyvíjen a výsledek této práce je do něj začleněn. Tato část se zabývá jeho stručným popisem, silně staví na [2].

Systém MAS je koncipován jako klient-server. Tato forma umožňuje systém škálovat a lépe udržovat jeho stabilitu. Klientskou část lze chápat jako webové rozhraní nebo externí aplikace využívající API systému MAS, serverová část, obsahuje, dolovací algoritmy, řízení systému, zprostředkování komunikace s externími entitami, získávání dat pro dolovací úlohy. Podporuje řadu datových zdrojů (např. relační databáze, OLAP kostky), umožňuje inkrementální získávání znalostí, zisk dodatečných dat pro analýzy a další. Možnost zisku dodatečných dat se jeví jako vhodné místo pro začlenění modulu, jímž se zabývá tato práce.

MAS je implementován na platformě .NET a jako takový zavádí důležité omezení pro tuto práci, totiž že výsledný modul musí být implementován v jazyce rodiny .NET.

2. ŘEŠENÍ

Aby bylo možno navrhnout a následně implementovat modul, musel jsem nejprve nastudovat problematiku práce s rozsáhlými daty. Na základě získaných znalostí jsem zvolil paradigma Map Reduce a platformu Hadoop [3]. Učinil jsem tak proto, že Hadoop je ve srovnání s alternativami, jako jsou např. MongoDB nebo CouchDB, rychlejší (pokud je řeč o Map Reduce) a spolehlivější, zatímco jiné alternativy nejsou vhodné z hlediska licenčních podmínek nebo byla jejich podpora ukončena (LINQ to HPC).

Díky mému průzkumu jsem určil klíčové části ekosystému Hadoop, které budou tvořit základ modulu. Jedná se o variantu datového skladu pro Big Data – Hive, který poskytuje HiveQL, což je podmnožina jazyku SQL – a sadu dolovacích knihoven Mahout. Oba nástroje spolupracují s HDFS, tedy distribuovaným souborovým systémem Hadoopu.

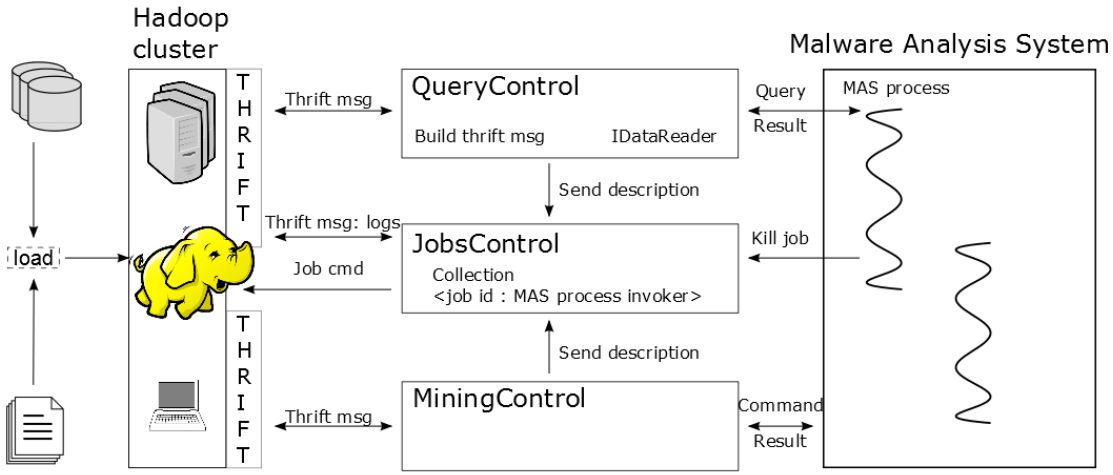
Volba ekosystému Hadoop s sebou ale přinesla i překážku. Hadoop je implementován v jazyce Java, což vyvolává rozpor s omezením způsobeným systémem MAS na implementaci modulu v jazyce platformy .NET. Aby bylo možné tento problém překlenout, nastudoval jsem možnosti propojení jazyků pomocí serializačních formátů. Na základě toho jsem zvolil serializační protokol Apache Thrift, který na rozdíl od dostupných alternativ (existují i placená komerční řešení) poskytuje oficiálně podporované rozhraní pro C# a přitom co do výkonu nijak nezaostává (což neplatí o technologii Hadoop Streaming, kterou je rovněž možno využít v C#). Tím je zajištěna komunikace mezi Javou ze strany Hadoopu a jazykem C# použitým v modulu.

2.1. ARCHITEKTURA MODULU

Modul se skládá ze tří hlavních částí, WCF služeb pro dotazování (QueryControl), dolování (MiningControl) a služby pro řízení a správu Map Reduce úloh (na které se dotazování v Hive a dolování v Mahoutu transformují) - JobsControl. Služby zprostředkovávají komunikaci mezi klientem, nejčastěji MAS procesem, a Hadoop clusterem. Služba pro řízení a správu dále komunikuje s ostatními dvěma službami.

Služba JobsControl slouží zejm. v případě, kdy je třeba zastavit právě probíhající Map Reduce úlohy (např. když byl proces čekající na výsledek úlohy změněn a již úlohu nepotřebuje), přičemž úlohu může ukončit pouze spouštějící proces nebo uživatel. Služba obsahuje kolekci identifikátorů Map Reduce úloh a identifikátory jejich spouštějících procesů. JobsControl navíc umí z popisu úlohy přes Thrift rozhraní zjistit identifikátor úlohy.

Základní funkcionalitou služby QueryControl je vyslání dotazu klientem přes Thrift rozhraní a poskytnutí výsledku dotazu ve formě implementace rozhraní IDataReader. Při vyslání dotazu do Hive je třeba evidovat identifikátor vzniklé Map Reduce úlohy. Tento identifikátor nelze získat skrze Thrift přímo, je třeba jej zpětně získat z popisu dotazu a logů v Hive, což zajišťuje JobsControl.



Obrázek 1: Architektura modulu.

3. ZÁVĚR

Práce seznámila čtenáře s pojmem „Big Data“, nezbytným pro modul tvořený v jejím rámci, a velmi stručně představila Malware Analysis System včetně jeho propojení s modulem. Nastudoval jsem a zpracoval problematiku Rozsáhlých dat, díky čemuž jsem zvolil platformu Hadoop jako základ pro modul. Implementoval jsem rozhraní pro komunikaci s Hadoopem dle navržené architektury a umožnil tak nejen dotazování nad distribuovanými daty, ale i řízení těchto dotazů. Přínosem modulu je zpřístupnění jinak neuchopitelných Rozsáhlých dat systému MAS, například pro řízení procesů nebo analýz. Část modulu MiningControl je v současnosti ve vývoji.

REFERENCE

Tento příspěvek vznikl za podpory výzkumného záměru MSM0021630528, grantů TA0101085 a FIT-S-11-2.

REFERENCE

[1] DUMBILL, Edd. *What is big data*: An introduction to the big data landscape. In: O'Reilly: Strata [online]. 2012 [cit. 2013-03-03]. Dostupné z: <http://strata.oreilly.com/2012/01/what-is-big-data.html>

[2] KUPČÍK, Jan, Hruška, Tomáš. *Towards Online Data Mining System for Enterprises*, In: Proceedings of the 7th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2012), Wrocław, PL, SciTePress, 2012, p. 187-192, ISBN 978-989-8565-13-6.

[3] WHITE, Tom. *Hadoop: the definitive guide*. 3rd ed. Sebastopol: O'Reilly, 2012, xxiii, 657 s. ISBN 978-1-449-31152-0.