ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta elektrotechnická
Katedra telekomunikační techniky

Podvody (fraudy) v telekomunikačním provozu

Frauds in Telecommunication Traffic

Vedoucí práce: Ing. Pavel Bezpalec, Ph.D.
Diplomant: Bc. Jan Nguyen

Praha, květen 2018

Čestné prohlášení

Prohlašuji, že jsem zadanou diplomovou práci zpracoval sám pod vedením vedoucího práce a používal jsem literaturu v práci uvedenou. Dále prohlašuji, že nemám námitek proti půjčování nebo zveřejňování mé diplomové práce nebo její části se souhlasem katedry.

V Praze dne 20.5.2018

.....
Bc. Jan Nguyen

Abstrakt

Tato diplomová práce se zabývá problematikou podvodného chování (angl. frauds) v telefonním provozu a návrhy postupů na jejich detekci. Jako hlavní komponentou řešení byl použit open-source aplikační rámec Hadoop. Byly provedeny detekce různých typů fraudů na modifikovaných CDR záznamech. Na základě získaných grafický výstupů bylo vyhodnoceno, zda jsme schopni tyto fraudy v telefonním provozu detekovat. Pro získání grafických výstupů byly použity skripty psané v jazyce Python a grafický editor yEd. Ze získaných výsledků jsme byli schopni určit podezřelé aktivity s vlastnostmi charakteristickými pro námi zvolené typy fraudů, z čehož vyplývá, že naše návrhy řešení pro detekci fungují a simulace byly úspěšné.

Klíčová slova

Fraud, CDR, Hadoop, Hive, Python

Abstract

This master thesis deals with a topic of frauds in telephone traffic. It designs possible solutions for fraud detection. The main component of the solution is the open source framework Hadoop. Fraud detection were done on modified CDRs. Evaluation based on graphic outputs made the detection of frauds in the telephone traffic possible. The graphic outputs were obtained from the scripts written in Python and graphical editor yEd. Based on the graphical outputs we were able to identify suspicious activity with characteristics of the chosen type of frauds. This means our solution for frauds detection works and the simulations were successful.

Keywords

Fraud, CDR, Hadoop, Hive, Python

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Nguyen** Jméno: **Jan** Osobní číslo: **392790**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra telekomunikační techniky**

Studijní program: **Elektronika a komunikace**

Studijní obor: **Komunikační systémy a sítě**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Podvody (fraudy) v telekomunikačním provozu

Název diplomové práce anglicky:

Frauds in Telecommunication Traffic

Pokyny pro vypracování:

Analyzujte typy podvodného chování v telefonním provozu, tzv. fraud. Zaměřte se na historické souvislosti, jejich příčiny a možné finanční dopady. Dále se seznamte se strukturuo záznamů o hovorech z telefonních ústředen. Na základě metod pro vytěžování informací z velkého objemu dat navrhněte možné postupy pro odhalování podvodného chování.

Seznam doporučené literatury:

[1] White, T.: Hadoop: The Definitive Guide, 4th Edition. O'Reilly Media. April 2015. ISBN: 978-1-49190-163-2.
[2] Berka, P.: Dobývání znalostí z databází. Praha: Academia, 2003. ISBN 80-200-1062-9.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Pavel Bezpalec, Ph.D., katedra telekomunikační techniky FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **04.01.2018** Termín odevzdání diplomové práce: _____

Platnost zadání diplomové práce: **30.09.2019**

Ing. Pavel Bezpalec, Ph.D.	podpis vedoucí(ho) ústavu/katedry	prof. Ing. Pavel Ripka, CSc.
podpis vedoucí(ho) práce		podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

23.2.2018 _____ _____

Datum převzetí zadání Podpis studenta

Poděkování

Děkuji vedoucímu práce Ing. Pavlu Bezpalcovi, Ph.D. za velmi užitečnou metodickou pomoc a cenné rady při zpracování této diplomové práce. Chtěl bych také poděkovat dlouholetému kamarádovi Ing. Filipovi Sušánkovi, kolegovi Viktorovi Kalivodovi za cenné připomínky k práci.

Dále bych chtěl poděkovat celé mé rodině, přítelkyni a přátelům za podporu, trpělivost a povzbuzování při psaní této práce.

Obsah

Seznam obrázků

Seznam tabulek

Seznam příloh

Skript zajištující parsování CDR záznamu:

- parsing.py

Skripty zajištující automatizaci analýzy CDR záznamu:

- spust.py
- automatizace.hql

Skripty na vykreslení výsledků z analýz:

- casovydiagram.py - vykreslení Grafu počtu hovorů v závislosti na časovém období
- statistikavolani.py - vykreslení Grafu provolaného času na každé telefonní číslo
- kdo_komu_kolikrat_volal_02.py - vykreslení Grafu počtu opakujících se hovorů

Skripty na vykreslení výsledků z analýz (produkční CDR záznam):

- CALLS_repeating_chart.py - vykreslení Grafu počtu opakujících se hovorů
- PBX_repeating_chart.py - vykreslení Grafu počtu opakujících se hovorů mezi PBX
- CALLs_repeating_percent_all_GRAPH.py - vykreslení Grafu celé komunikace

Skripty na vykreslení a uložení výsledků z návrhů detekce fraudů:

- wangiri.py - Wangiri Fraud
- highcost.py - detekce volání do drahých destinací
- premium.py - Premium Service Rate Fraud
- robot.py - detekce robota

Soubory s Hive příkazy pro realizaci detekci jednotlivých fraudů:

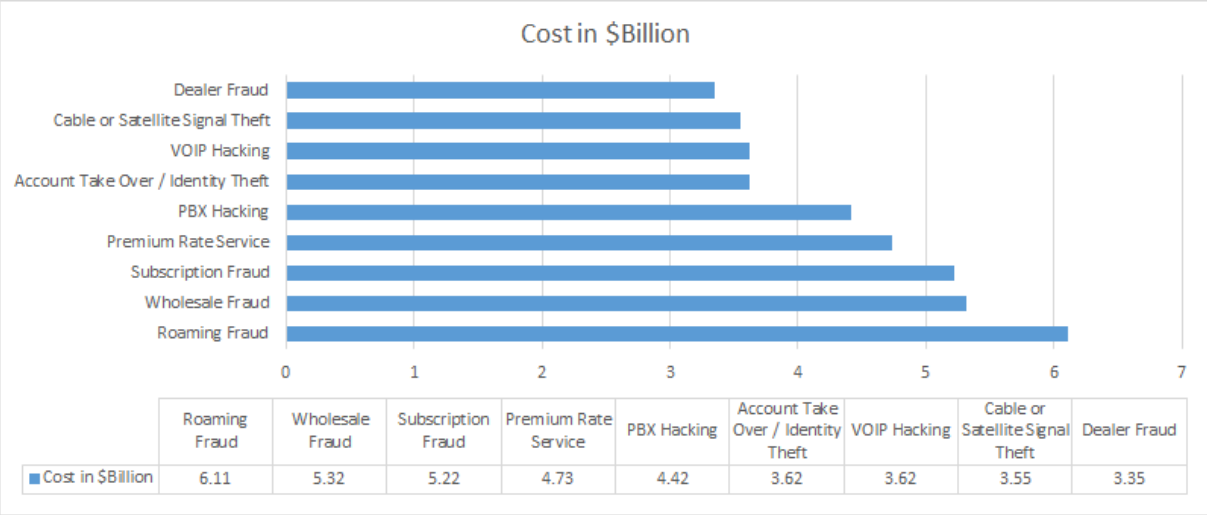
- basic_CDR_analysis.txt - analýza produkčního CDR záznamu
- premium_and_service_fraud.txt - detekce Premium Rate Service Fraud a volání do zón s nejvyšším tarifem
- wangiri_fraud.txt - detekce Wangiri Fraudu
- robot_fraud.txt - detekce robota

1 Úvod

Tato diplomová práce se zabývá problematikou podvodného chování (angl. frauds) v telefonním provozu. Na základě konzultací s panem vedoucím byly stanoveny následující cíle práce. Hlavní náplní bylo studium různých typů fraudů a jejich základních vlastností, následně návrh způsobů jejich detekce s využitím metod pro vytěžování informací z velkého objemu dat. Součástí práce bylo také seznámení se strukturou záznamů o hovorech z telefonních ústředen. Hlavním úkolem práce bylo navrhnout a realizovat možné postupy pro odhalování různých typů fraudů s využitím open-source aplikačního rámce Hadoop a dále vytvořit skripty na získání grafických výstupů.

Fraudy v telekomunikačních sítích jsou v dnešní době velice aktuální tematikou a pro operátory představují velký problém. Telekomunikační fraud lze obecně definovat jako jev, kdy dochází ke zneužití různých služeb operátora. Fraudem způsobené finanční škody jsou odhadovány ročně na 35 - 40 miliard amerických dolarů. V přepočtu se jedná o 3 - 8% z celkových výnosů operátora, v některých zemích jsou tato čísla vyšší. [3]

S rychlým pokrokem technologií se mění i typy a způsoby provedení fraudů. Fraudy jsou dnes sofistikovanější a z technického hlediska náročnější na realizaci. Je proto velmi důležité sledovat technologické trendy, abychom získali přehled a věděli jak se jednotlivým fraudům bránit. Na níže uvedeném Obrázku 1.1 je graf ze zdroje [47], kde jsou zachyceny finanční ztráty operátorů telekomunikačních služeb způsobené uvedenými typy fraudů.



	Roaming Fraud	Wholesale Fraud	Subscription Fraud	Premium Rate Service	PBX Hacking	Account Take Over / Identity Theft	VOIP Hacking	Cable or Satellite Signal Theft	Dealer Fraud
Cost in $Billion	6.11	5.32	5.22	4.73	4.42	3.62	3.62	3.55	3.35

Obr. 1.1: Finanční ztráty způsobené fraudy

Na trhu existují řady systémů sloužících k detekci fraudů, ať v oblasti telekomunikací, bankovnictví či leasingových společností. Téměř všechny systémy jsou postaveny na vytěžování informací z velkého objemu dat, tedy na Big Data technologiích. Nad těmito systémy jsou dále postavena interaktivní rozhraní pro obsluhu. V této práci byl z Big Data technologií zvolen aplikační rámec Hadoop jako hlavní jádro technické realizace. Hadoop obsahuje sadu open-source komponent určených právě na analýzu a zpracovaní velkých dat. Díky tomuto aplikačnímu rámci je možné postavit detekční systém, který lze nasadit do reálného provozu.

Práce dle jejího obsahu lze rozdělit na tři hlavní části: teoretickou, praktickou a vyhodnocení. První část práce se zabývá teorií fraudů, Big Data, záznamy z ústředen (CDR) a krátce čtenáře seznamuje s metodami na detekci fraudů a existujícími systémy na detekci fraudů.

Druhá část obsahuje čtyři podkapitoly. První a druhá podkapitola popisuje instalaci a přípravu prostředí Hadoop ve virtualizačním programu VirtualBox, ve kterém byly provedeny ana-

lýzy CDR záznamů, a to konkrétně analýza počtu provolaných vteřin na každé telefonní číslo, analýza počtu hovorů v daném časovém období a analýza počtu opakujících se hovorů v daném časovém období. Třetí podkapitola popisuje návrh řešení pro zpracování CDR záznamů, který prakticky využívá řešení a analýz z předchozí podkapitoly, ovšem již na reálném Hadoop klastru a produkčních datech.

Poslední podkapitola obsahuje návrhy možných postupů pro odhalování fraudů. Jsou zde popsány jednotlivé detekce fraudů, které byly provedeny na modifikovaných CDR záznamech. V rámci vyhodnocení jsou shrnuty a diskutovány dosažené výsledky.

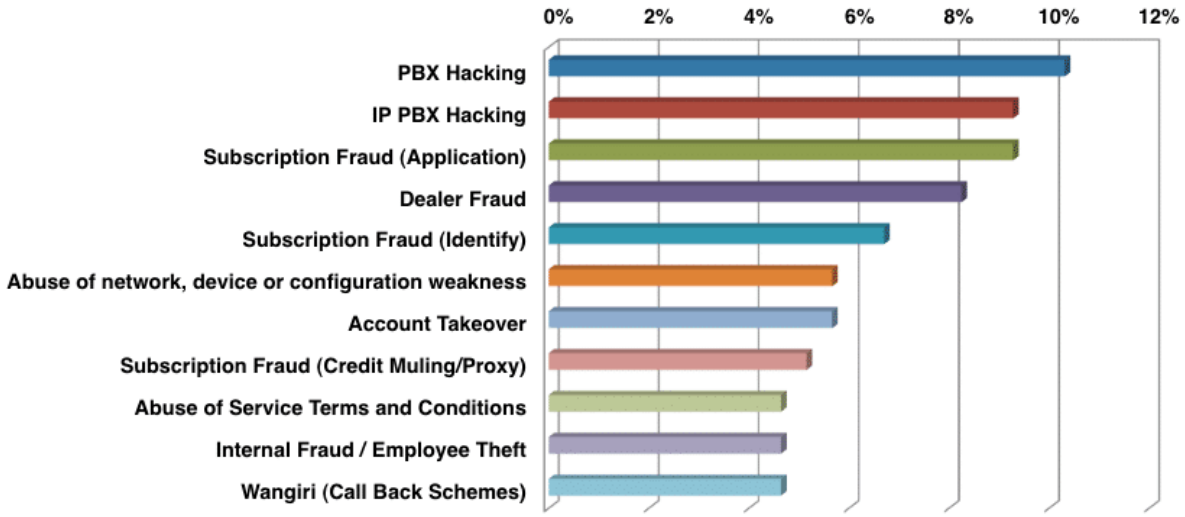
2 Teoretický rozbor

Jako každá oblast průmyslu, i telekomunikační průmysl čelí řadě hrozeb, které ohrožují nejen samotné operátory, ale také jejich koncové zákazníky. Tyto hrozby mají za hlavní následek snížení příjmů operátorů a s tím související zhoršení reputace. Obecně můžeme tyto hrozby klasifikovat do dvou skupin: hrozby vycházející z chyb technologie a hrozby kriminálního charakteru.

Do skupiny hrozeb vycházejících z chyb technologie patří např. technické selhání pobočkové ústředny. Jedním ze selhání může být to, že po ukončení hovoru jednou ze stran zůstane hovor nerozpojen. Konkrétní případ tohoto selhání je uveden v materiálech firmy Telefonica O2. Jednalo se o situaci, kdy zákazník uskutečnil hovor do zahraničí, který trval téměř devět dní. Účet za tento hovor se vyšplhal na více než 59 000Kč.[1]

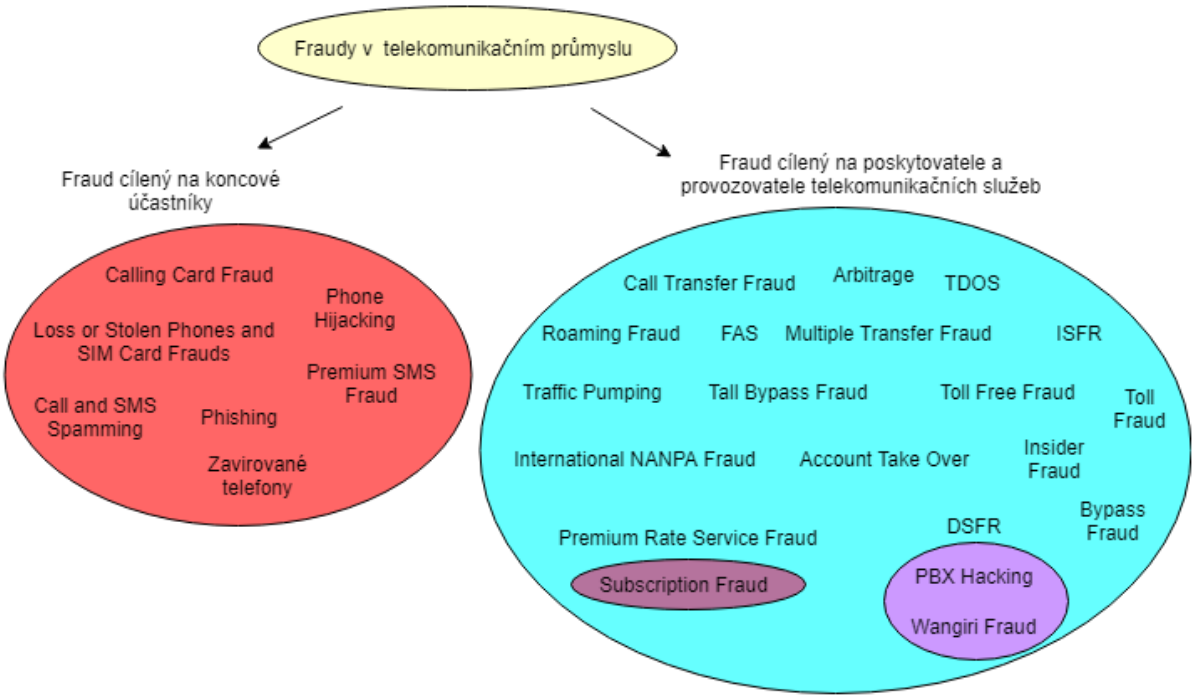
Hrozby kriminálního charakteru ohrožují jak telekomunikační společnosti, tak i jejich zákazníky. Zmíněná skupina se obecně označuje termínem „fraud“, v českém překladu podvod či zneužití. Dle zdroje [2] je termín fraud definován jako neoprávněné využívání některých služeb zákazníkem. Do této velké skupiny řadíme aktivity, které mají za následek snížení zisků operátorů, poškození firemní značky a především ztráty důvěry zákazníků. S vývojem komunikačních technologií se vyvíjí také metody a způsoby realizace fraudů. [1] [3]

Motivací útočníků realizujících fraudy může být například finanční úspora za volání, zisk informací, skrytí identity, přímý zisk, poškození operátora či konkrétního uživatele nebo získaní sociální prestiže. Útočníkům hrozí při zjištění jejich identity legislativní postih. [4] Na uvedeném Obrázku 2.1 ze zdroje [44] jsou na grafu zobrazeny nejčastější typy vyskytujících se fraudů a jejich celkové procentuální zastoupení.



Obr. 2.1: Typy fraudů a jejich výskyt

V následujících částech této kapitoly jsou obecně popsány typy a principy různých fraudů. Jejich dělení a struktura se v různých zdrojích mohou mírně lišit. V této práci bylo zvoleno pro jednodušší přehled rozdělení fraudů dle jejich cílení. Anglické názvy nebudou překládány, jelikož ne všechny názvy mají plnohodnotný český ekvivalent. Často se zde objevují termíny jako podvodník a útočník. Tyto dva pojmy se obecně významově liší, ovšem v této práci se termíny rovnají a představují pojem zastřešující jak člověka, který je schopen prolomit zabezpečení pobočkové ústředny, tak i člověka, který je schopen získat metodami sociálního inženýrství citlivé informace oběti. Na níže uvedeném Obrázku 2.2 je zobrazeno použité rozdělení fraudů v této práci.



Obr. 2.2: Rozdělení fraudů

2.1 Fraudy

2.1.1 Fraudy cílené na koncové účastníky

Tato skupina fraudů může být ovlivněn majiteli telefonů a zahrnuje všechny typy fraudů založených na odcizení osobních údajů k telefonním účtům. Jako příklad můžeme uvést ztrátu či odcizení telefonu nebo SIM karty. Nejlepší způsob jak bojovat proti tomuto typu fraudu je vzdělávat uživatele o základních pravidlech bezpečnosti jako jsou např. dostatečně dlouhá hesla. Níže jsou uvedeny typy fraudů spadající do této skupiny. [5]

Calling Card Fraud

Jde o typ fraudu, který může být proveden mnoha způsoby. Jeho základem je podvodný telefonní hovor s cílem získání citlivých informací. Podvodník může např. vystupovat jako zaměstnanec telefonního operátora snažící se různými metodami sociálního inženýrství získat citlivé údaje. Další možností zjištění hesla je nahlížet oběti přes rameno.

Uživatel se v tomto případě může bránit tak, že bude vyžadovat, aby se podvodník autorizoval a tyto získané údaje si ověřit. Další doporučení platí stejně jako u používání kreditních karet, kdy je třeba zadávat heslo tak, aby jej ostatní neviděli. [5]

Loss or Stolen Phones and SIM Card Frauds

Tento typ podvodu je poměrně nebezpečný z pohledu koncového účastníka. Podvodník získá přístup do telefonu uživatele, kde může zjistit veškeré přihlašovací údaje nejen do sítě, ale také do e-mailu, elektronického bankovnictví či sociálních sítí. Podvodník může napáchat velké finanční škody, např. voláním na prémiová čísla, která jsou zpoplatněna mnohem vyšší sazbou než běžné hovory nebo využitím placených služeb. Podvodník se dokonce může pokusit ukrást identitu oběti.

Mobilní telefony je třeba chránit dostatečně dlouhými hesly nebo využitím biometrického způsobu autentizace - snímání otisků prstů. Dále se doporučuje neukládat do mobilního telefonu všechny přístupy např. do banky, pošty nebo sociálních sítí. Při ztrátě telefonu je potřeba ihned zablokovat účty či změnit hesla do těchto služeb. [5]

K tomuto typu fraudu se váže SIM Cloning Fraud, jehož princip spočívá v provedení duplikátu SIM karty pomocí speciálního programu. Proces klonování může být realizován také pomocí OTA (Over-the-air programming). Detailnější popis tohoto fraudu je uveden ve vědeckém článku [36]. Podvodník může následně provádět bankovní transakce, pokud má i údaje o platební kartě oběti a autorizační SMS zprávy chodí na číslo, které má duplikovaná karta. [27]

Operátor je schopný monitorovat klonované telefony a SIM karty pomoci algoritmu Velocity Check a Collision Check, jejichž princip spočívá v porovnávání SIM karet či telefonů v závislosti na geografické poloze. V případě detekce dvou hovorů ze stejného telefonního čísla z různých geografických poloh, bude operátor upozorněn. [26]

Call and SMS Spamming

Je typ fraudu, při kterém je uživateli mobilního telefonu doručena nežádoucí textová zpráva nebo je uživatel informován prostřednictvím hovoru o speciální výhodné nabídce či službě. Call and SMS Spamming je velice nepříjemný ze dvou důvodů. Prvním důvodem je neexistence filtrů pro hovory a textové zprávy podobně jako u spamových filtrů při emailové komunikaci, a proto je poměrně obtížné takovým hovorům a zprávám předcházet. Druhým důvodem jsou finanční dopady. Obětem takového podvodu mohou být účtovány poplatky za každou přijatou textovou zprávu či hovor. [5]

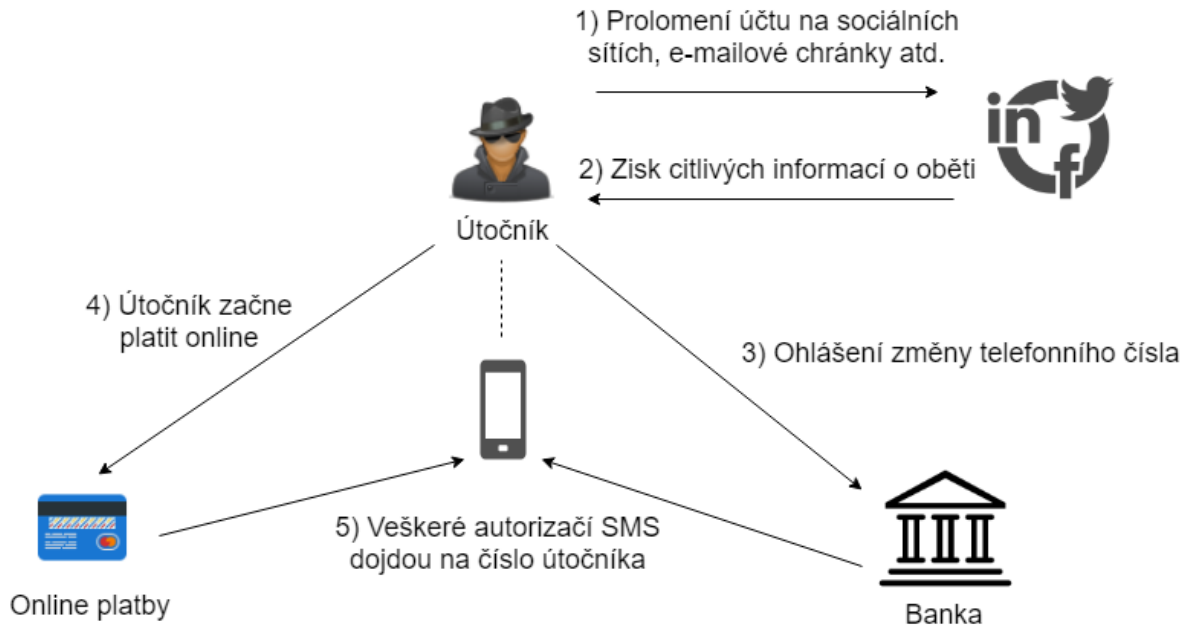
Na Google play marketu nyní existuje řada spamových aplikací jako jsou JokesPhone, Juasapp atd., které dokáží během několika minut při zadání telefonního čísla oběti provádět žertovné hovory. [49] Tyto hovory jsou velmi autentické a nejde zprvu poznat, že se jedná o automat. Hlasový záznam z tohoto hovoru je možné vzápětí sdílet přes sociální sítě. Hovory nejsou realizovány z telefonu volajícího, ale ze systémů provozovatele aplikace. První hovor je zdarma a další jsou již placené. Hovory jsou prováděny nejčastěji z čísel začínajících trojčíslím 672.

Dle zákona č. 101/2000 Sb. o ochraně osobních údajů se v tomto případě jedná o zásah do osobních práv, jelikož dochází k nahrávání hovoru, aniž by tato skutečnost byla volanému předem sdělena.[48]

Jediným a zároveň nejrychlejším způsobem, jak se bránit proti takovému fraudu je blokace všech čísel, která začínají předčíslím 672.

Phone Hijacking

Princip tohoto fraudu je poměrně jednoduchý. Útočník získá citlivé informace o oběti na Internetu, metodou sociálního inženýrství nebo prolomením hesel do emailové schránky či účtů na sociálních sítích. Informace jako datum narození, trvalé bydliště, číslo účtu či rodné číslo jsou klíčové pro tento podvod. Na Obrázku 2.3 je zachycen princip tohoto fraudu, při němž útočník získá přístup k emailové schránce a účtům na sociálních sítích oběti. Následně je proveden například hovor do banky s pokusem o změnu telefonního čísla, na které budou posílány autorizační zprávy. K úspěšnému provedení většinou postačí právě citlivé informace oběti. V případě, že je změna telefonního čísla úspěšná, může nyní podvodník provádět platby a napáchat velké finanční škody.[30]



Obr. 2.3: Princip Phone Hijacking

Phishing

Je populární forma hackingu, při kterém se podvodník vydávající se za banku, úřad či jinou instituci snaží získat osobní údaje oběti jako jméno, heslo nebo informace ke kreditní kartě. Tento

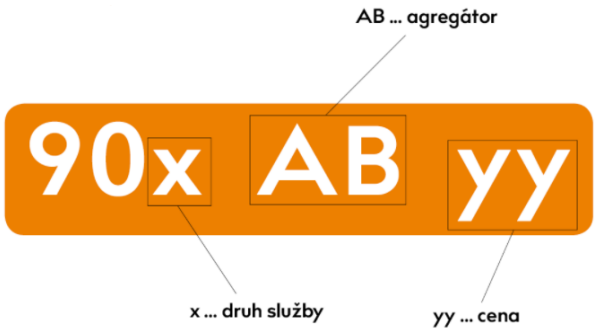
typ fraudu může být realizován prostřednictvím nejen e-mailů, ale také pomocí telefonních hovorů či textových zpráv. Někdy se také označuje jako voice phishing, zkráceně "vishing".

Zde můžeme uvést fraud SMS phishing taktéž označován jako „smshing“. Oproti emailovým spamům, u SMS zpráv nejde předem filtrovat odesílatele a uživatelé jsou účtováni za každou přijatou zprávu. Překvapivě smshing je efektivnější než e-mail phishing kvůli krátkému textu pocházejícího z telefonního čísla, což působí mnohem důvěryhodněji. [5]

Premium SMS Fraud

Premium SMS je služba, za kterou zákazník platí firmě, která službu provozuje. Platba je zákazníkovi stržena z kreditu či v rámci zúčtování mobilních služeb. Službou je zde míněn např. nákup jízdenky na MHD, hlasování v soutěžích nebo erotické služby na internetu. Samotnou platbu zprostředkuje agregátor, se kterým se firma předem dohodne. Mobilní operátor zde figuruje pouze jako zprostředkovatel, jelikož on žádné služby neprovozuje. Tím pádem nemá možnost monitorovat obsah jednotlivých Premium SMS zpráv a neví, kam tyto zprávy směřují.

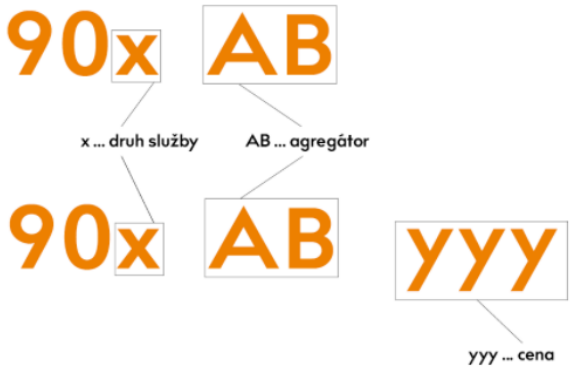
Existují obecně dva typy Premium SMS. První z nich je zpoplatněna okamžitě při odeslání. Například, odesláním SMS zprávy na číslo 9001230 Vám bude stržena částka 30 Kč. Této skutečnosti zneužívají podvodné aplikace a viry, které jsou schopny odesílat v pozadí SMS zprávy na sedmimístná čísla začínající 90X bez vědomí vlastníka. Operátoři se snaží předejít této situaci odesláním potvrzující SMS na číslo z něhož Premium SMS zpráva odeslána. Je proto nutné dávat pozor, kterým aplikacím je povolen přístup k SMS zprávám. Na Obrázku 2.4) je uvedeno, jak taková SMS zpráva může vypadat a dle čeho se dá poznat, kolik bude stát peněz.



Obr. 2.4: Premium SMS - okamžité zpoplatnění

Druhým typem jsou Premium SMS zprávy zpoplatněné až při jejich přijetí a využívají se u opakovaně účtovaných služeb, jako jsou třeba přístupy do placených sekcí různých webových stránek. Tento typ Premium SMS zpráv je objednáván přes Internet či odesláním SMS zprávy ve formátu 90X AB na pětimístné číslo. Při odeslání takovéto SMS zprávy přijde na číslo odesilatele potvrzení. Pokud na příslušné číslo není zákazníkem poslán souhlas v definovaném formátu, objednávka nebude dokončena. U objednávky přes Internet je zákazníkovi doručen potvrzující kód, který je pak zadán do webového formuláře. Cena za takovou Premium SMS zprávu se může pohybovat v rozmezí od 1 do 999 Kč. U služeb s týdenním předplatným je to 99 Kč, s měsíčním 199 Kč. Fraudy tohoto typu jsou často prováděny přes sociální sítě. Uživatelé jsou žádáni podvodníkem, který se vydává za přítele v nouzi, o telefonní čísla a následně o potvrzující kód.

Vůči tomuto fraudu se lze bránit bezplatným zablokováním veškerých plateb třetím stranám prostřednictvím Premium SMS. O zablokování lze požádat telefonního operátora.[28]



Obr. 2.5: Premium SMS - zpoplatněné až po přijetí

Zavirované telefony

Jedná se o nový fenomén, který se vyskytl v České republice. Problém byl zaznamenán u mobilních telefonů značky myPhone CLASSIC, které jsou určeny především seniorům. Zmíněné telefony obsahují škodlivý malware již od výroby, který se aktivuje po vložení SIM karty do telefonu a po připojení k datové síti. Telefon si po aktivaci stáhne soubor s aktualizovanými telefonními čísly do drahých destinací a začne tyto čísla vytáčet bez vědomí majitele. Jednalo se o destinace, jako jsou Čad, Nigérie, Malawi, Senegal.[33]

2.1.2 Fraudy cílené na poskytovatele telekomunikačních služeb

Jedná se o nejkomplexnější fraudy, při kterých jsou zneužívány slabiny operátorů pomocí metod, které budou dále popsány. Některé typy fraudů spadající do této skupiny jsou cílené jak na operátory, tak na koncové uživatele. Operátoři jsou obecně velmi zranitelní vůči telekomunikačním fraudům. Podvodníci často využívají slabého zabezpečení ze strany zákazníků operátora. Útoky jsou obvykle prováděny během svátků a víkendů. V těchto obdobích jsou sítě monitorovány méně detailně. Jako dobrý způsob obrany se jeví monitoring CDR záznamů v reálném čase se zaměřením na podezřelý provoz nebo podle konkrétního vzoru.

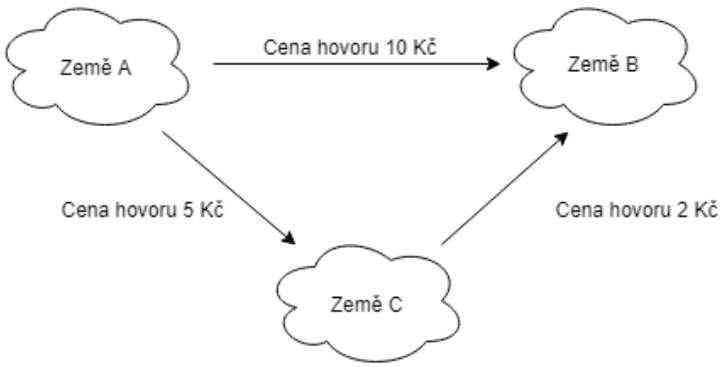
Fraudy cílené na poskytovatele telekomunikačních služeb jsou často děleny na dvě větší skupiny. Domestic Revenue Share Fraud (DSFR) a International Revenue Share Fraud (ISFR). Do skupiny DSFR patří například PBX Hacking nebo Wangiri Fraud. ISFR je bez diskuze jedním z nejvíce poškozujících podvodů. Způsob realizace je stejný jako DRSF, jen s tím, že je realizován v mezinárodním měřítku. Jsou zde platné stejné principy a postupy útočníků. [5]

Insider Fraud

Obecně se jedná o zneužívání telekomunikačních služeb ze strany zaměstnanců poskytovatele. Zaměstnanci zneužívají možnosti modifikovat nastavení účtování pro vybraná čísla. Dalším příkladem je manipulace se systémy a sítí operátora s cílem finanční ztráty operátora. [24] [25] [26]

Artbitrage

Telekomunikační arbitrážní typy fraudů jsou založeny na rozdílech v sazbách účtování mezi zeměmi. Telefonní operátoři často účtují různé sazby za propojení podle druhu hovoru nebo poskytovatele služeb. Mezinárodní hovory se nemohou zpracovávat a dokončovat prostřednictvím jednoho telefonního operátora a proto musí směrovat provoz přes další operátory za poplatek. Na níže uvedeném Obrázku 2.6 je uveden příklad. Země A má mnohem nižší sazby účtování pro zajištění hovoru do země B přes zemi C. Proto je pro zemi A výhodnější směrovat její provoz pro zemi B přes zemi C.[5]

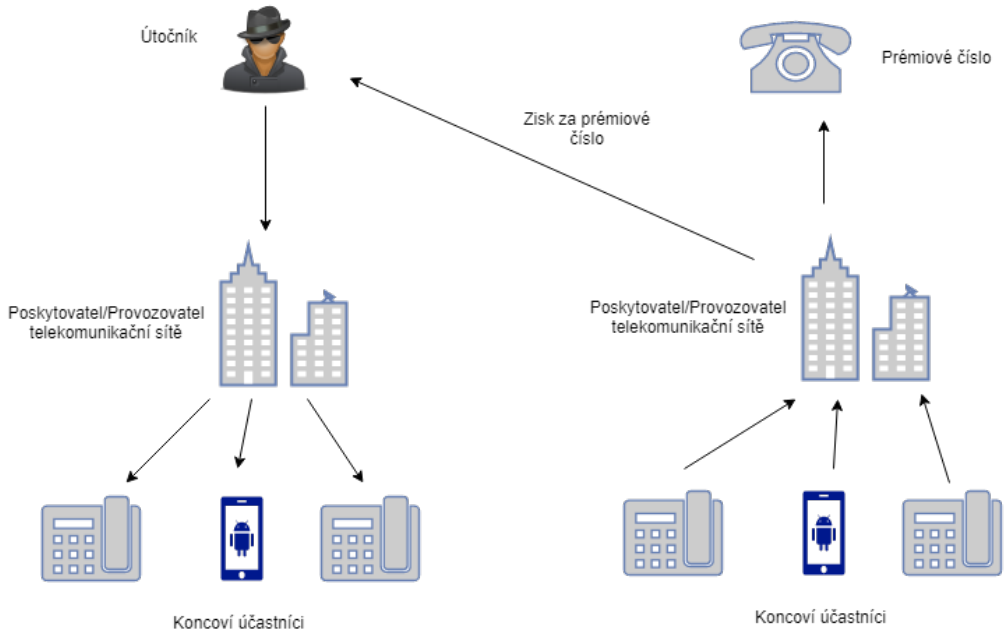


Obr. 2.6: Princip Arbitrage Fraud

Wangiri Fraud

Tento typ fraudu je také znám jako „one ring and cut“. Jeho princip je v podstatě prostý. Útočníci zanechají obětem zmeškaný hovor, a když jim oběti zavolají zpět, dovolají se často na zahraniční nebo prémiová čísla, za což jsou účtovány vysoké poplatky.

Jedním důvodem je, že mezinárodní předvolba "00"nebo "+"je u příchozího telefonního čísla často potlačena a koncový účastník neví, že se jedná o hovor ze zahraničí. Druhým důvodem je nepozornost koncového účastníka. Útočníkům mohou plynout finanční zisky z provozu prémiových čísel nebo jen úmyslně poškozují operátory bez vidiny jakéhokoliv zisku. Obecný princip tohoto podvodu můžete vidět níže.[5] [25] [51]



Obr. 2.7: Princip Wangiri Fradu

PBX Hacking

Při tomto typu útoku útočník získá přístup do slabě zabezpečené pobočkové ústředny PBX (Public Branch Exchange) a provede mezinárodní dálkové hovory nebo hovory na prémiová čísla. Útočník má dále přístup k hlasovým zprávám, je schopen odposlouchávat hovory uživatelů, měnit konfiguraci ústředny nebo změnit hesla k účtům.

Často se uvádí ještě jedna forma tohoto fraudu, a to Voice Mail Hacking. Tento útok je prakticky podobný PBX Hacking. Útočník se snaží přistoupit do hlasové schránky oběti pomocí překonáním hesla. Od chvíle, kdy útočník získal přístup k hlasové schránce, má možnost poslouchat nebo mazat hlasové zprávy.[51]

Call Hijacking

Útočník vystupující zde jako prostředník mezi operátory láká jiné poskytovatele telefonních služeb na nízké poplatky za sestavení hovorů ke koncovým uživatelům. Útočník je schopen nabízet nízké poplatky, jelikož jeho priorita není sestavování hovorů.

Ve chvíli, kdy je navázána spolupráce mezi útočníkem a nic netušícím poskytovatelem, útočník část hovorů, které mají být směrovány ke koncovým uživatelům, směruje na nějakou hlasovou

zprávu či oznámení o nedostupnosti. Útočník toto dělá záměrně a účtuje poskytovatelům služeb stejné poplatky jako v případě sestavení hovorů. [29]

Premium Rate Service Fraud

Při volání na prémiová telefonní čísla se peníze dělí mezi operátora a provozovatele. Jedná se často o hlasování v rámci soutěží, televizních pořadů, služby pro dospělé atp. Oběti těchto podvodů si neuvědomují dodatečné poplatky spojené s těmito službami, které bývají uvedeny drobným písmem pod uvedeným prémiovým číslem. [51]

Subscription Fraud

Tento podvod spočívá v založení telefonního účtu u operátora s použitím falešných informací. Pod tímto účtem útočník vykonává nelegální aktivity s motivem finančního zisku, např. volání na prémiové účty, které si sám založil u jiného operátora nebo provádí mezinárodní hovory a tím způsobí škodu operátorům. Pokud by byl tento útok proveden ve větším rozsahu z více podvodných telefonních účtů, je velmi obtížné útočníky dopadnout. [51]

Roaming fraud

Roaming je jedna z nejvýnosnějších služeb operátorů. Dopadem těchto podvodů jsou velké finanční ztráty, což může vést ke zvýšení cen mobilních tarifů a tím související nespokojenost zákazníků. Roaming je automatické připojení uživatele k návštěvnické síti ve chvíli, kdy jeho domovská síť není dostupná. Toto připojení nastane tehdy, pokud obě sítě mají uzavřenou roamingovou dohodu. Záznam o podrobnostech hovoru CDR uživatele bude doručen do domovské sítě do několika dnů, někdy i týdnů, což je příležitost pro podvodné útoky. Tento fraud nastane v případě, když uživatel využil služeb návštěvnické sítě a odmítá za ně zaplatit tvrzením nevědomosti nebo jiným důvodem.

Tento typ fraudu není již aktuální vzhledem k okolnosti, že roaming v rámci Evropy není zpoplatněný, dle nové regulace od roku 2017.[51] [6]

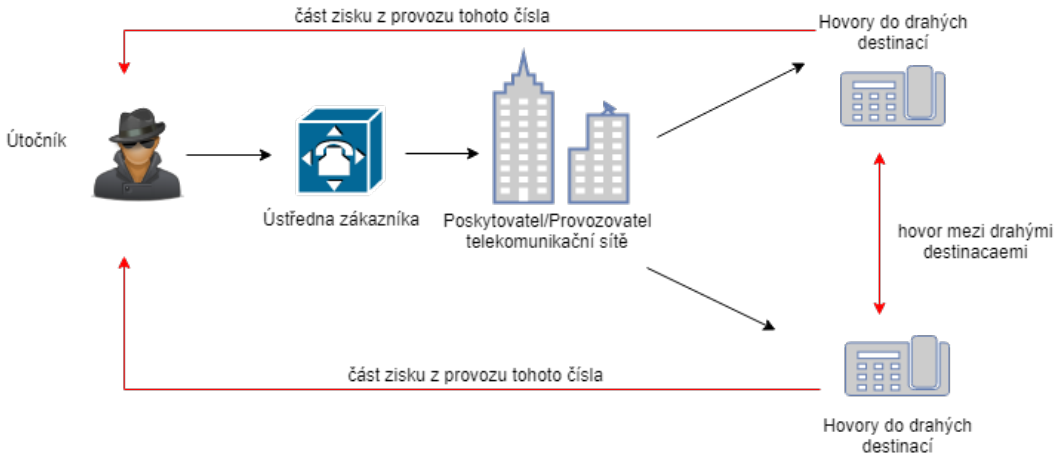
Call Transfer Fraud

Je zaměřen na uživatele využívající virtuální ústředny. První fází tohoto útoku je zisk přístupu do PBX některého zákazníka. Následně útočník může provádět mezinárodní hovory nebo hovory přímo na svoje zpoplatněné telefonní služby. Po sestavení takového hovoru útočník vynutí přesměrování hovoru mezi útočníkem a drahou destinací. V tomto případě je poškozen především operátor. Tato varianta je znázorněna na níže uvedeném Obrázku 2.8. [5]

Obr. 2.8: Princip Call Transfer Fraud

Multiple Transfer Fraud

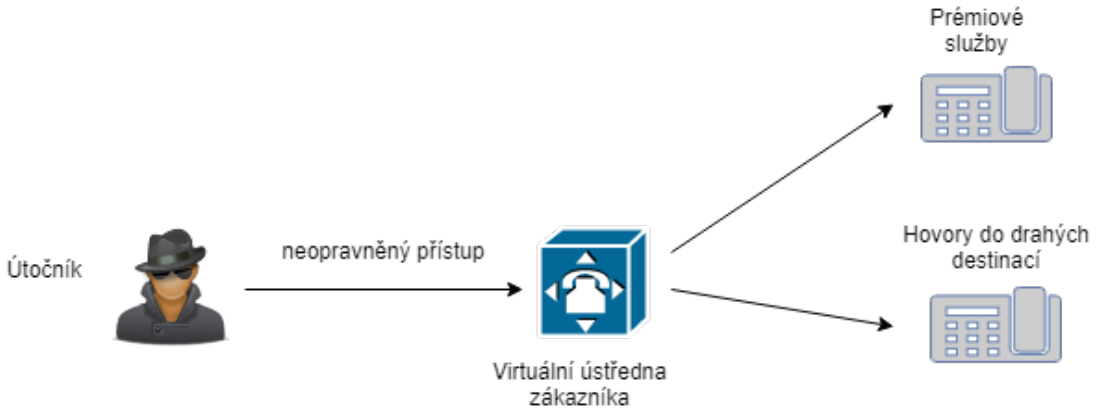
Tento fraud má dvě fáze a je podobný předchozímu. Jeho princip je znázorněn na Obrázku 2.9. Útočník získá neoprávněný přístup do ústředny a realizuje například hovory do drahých destinací. V počáteční fázi je realizován hovor mezi útočníkem a drahou destinací. V druhé fázi útočník vynutí pomocí ústředny přesměrování hovoru mezi dvěma drahými destinacemi. Když je tato fáze dokončena, může útočník zavěsit. Útočník získává finanční odměnu z provozu cílových čísel.[5]



Obr. 2.9: Princip Multiple Transfer Fraud

Call Forwarding Fraud

Tento útok spočívá v získání přístupu do korporátního PBX skrze jeho webový portál. Útočník hádá uživatelská hesla do PBX a poté co je získán přístup, nakonfiguruje PBX, aby realizoval hovory na zahraničí nebo prémiové čísla. Útočník bude volat na číslo uživatele, ke kterému získal přístup a následně bude přesměrován do zahraničí nebo na prémiové číslo.[5]



Obr. 2.10: Princip Call Forwarding Fraud

False Answer Supervision (FAS)

Princip FAS spočívá v nesprávném účtování za hovor volající straně. U prvního typu účtování nastane mnohem dříve, než volaná strana zvedne telefon a realizuje se hovor. Částka je naúčto-

vána již během vytáčení. U druhého typu je po uskutečnění hovoru přehrávána zvuková hláška s cílem udržet volající stranu co nejdéle aktivní a tím způsobit co největší škodu. [5]

International NANPA Fraud

NANPA - North American Numbering Plan Administration uvadí čísla jak v USA, tak i mezinárodní čísla Karibiku. Řada ústředen nemá informaci o číslech, která jsou uvedena na seznamu NANPA nebo LERG. To vede ke zranitelnosti operátorů vůči ISFR, protože nemohou blokovat hovory do konkrétních podvodných destinací jako je Karibik.[5]

Account TakeOver

Jedná se o typ podvodů, který je zaměřen na finanční instituty. Podvodníci se vydávají za legitimního zákazníka s cílem získání informací o jeho účtu. Jedná se o kombinaci s metodami sociálního inženýrství. Podvodník se snaží o změnu hesla nebo změnu adresy. Instituce se proti těmto útokům brání různými metodami jako ověření hlasem, rodným číslem nebo heslem.[5]

Bypass Fraud

Bypass Fraud je často také označován jako Interconnect Fraud, GSM Gateway Fraud nebo SIM Boxing. Slovo bypass můžeme volně přeložit jako obcházení. Tento fraud je velmi komplexní a vyžaduje významnou investici do technologií, které umožňují realizovat mezinárodní hovory za stejnou cenu jako národní hovory tak, že obchází účtovací systém pro mezinárodní hovory. Typicky podvodníci prodávají karty pro mezinárodní hovory. V okamžiku provedení hovoru z této karty do zahraničí podvodný operátor upraví hovor tak, že vypadá jako národní hovor.

Na Obrázku 2.11 je zobrazen princip fraudu. Účastník ze země A využije služeb podvodné firmy, která poskytuje SIM karty na dálkové hovory. Účastník realizuje ze země A volání do země B hovor, podvodná firma zajistí, aby se hovor jevil jako národní hovor, tím pádem dochází ke snížení sazby oproti mezinárodnímu hovoru.[5][29]

Obr. 2.11: Princip Bypass Fraud

Toll Fraud

Je podvod s následujícím scénářem. Podvodník překoná zabezpečení ústředny a skrze napadenou ústřednu poskytuje jako operátor hovory do drahých destinací či na prémiové služby koncovým uživatelům. Podvodník profituje na základě poskytovaných služeb.[31][5]

Telecom Denial of Service (TDOS)

Jedná se o poměrně nový fenomén mezi fraudy. Útočník profituje z pronajatých linek, ke kterým získal neoprávněný přístup. Charakteristickým rysem tohoto fraudu bývá zvýšený objem dlouho trvajících hovorů s náhodnými čísly. Cílem tohoto fraudu je vyčerpat kapacitu sítě organizace či společnosti. V případě tohoto útoku ve veřejné telefonní síti by mohlo dojít k nedostupnosti čísel policie nebo záchranné služby. Provozovatelé těchto linek velmi špatně detekují tento útok, jelikož vzorec hovorů se jeví jako normální.[5]

2.2 Fraudy v České republice

V této kapitole jsou uvedeny obecné informace týkající se fraudů a legislativy související s touto problematikou v rámci ČR. Zdrojem těchto informací byl Český telekomunikační úřad (ČTÚ), který je podle § 3 odst. 1 zákona č. 127/2005 Sb., o elektronických komunikacích a změně některých souvisejících zákonů (dále jen ZEK), ve znění pozdějších předpisů je zřízen jako ústřední správní úřad pro výkon státní správy ve věcech stanovených tímto zákonem. Dle zákona č. 40/2009 Sb. o podvodné jednání v elektronických komunikacích se svojí povahou jedná o trestnou činnost.

Dle vyjádření ČTÚ se účastníci hlasových služeb elektronických komunikací často setkávají s problémem volání z neznámých čísel, na která není možné se zpětně dovolat a pokud ano, tak se jedná zpravidla o zahraniční telefonní číslo, což vede k vysokým poplatkům. ČTÚ také zaznamenalo stížnosti ohledně zneužívání telefonních čísel se zvýšeným tarifem typu 90X i přesto, že dle novely zákona o spotřebitelském úvěru se zakazuje použití tohoto čísla pro nabídku, sjednávání nebo zprostředkování spotřebitelského úvěru.

Navíc od 1. října 2012 Asociace provozovatelů mobilních sítí (AMPS) zavedla povinnou bezplatnou informační hlásku u těchto čísel, která jsou často využívaná pro spotřebitelské úvěry či nabídky práce. [25]

ČTÚ dále uvádí, že se často vyskytuje fraud typu "one ring and cut", tedy masivní prozvánění účastníků v mobilních sítích ze zahraničních čísel. Oklamaní spotřebitelé často volají zpět a jsou směrováni na telefonní čísla v mobilních či pevných sítích s nahranou hláskou s inzertním sdělením apod. V jiných případech jsou nasměrováni na zahraniční telefonní číslo. Výsledkem jsou vysoké poplatky za provedený hovor. Kromě zneužití telefonních služeb jsou zneužívány také prémiové SMS zprávy v rámci soutěží, kvízů a televizních či on-line her. Vyskytuje se také mnoho obtěžujících volání různého druhu, což můžeme klasifikovat jako call spamming.

Pokud se účastník stane obětí fraudu a je mu naúčtován vysoký poplatek, má dle zákona § 64 odst. 7 zákona o elektronických komunikacích podat reklamaci u poskytovatele služeb, a to nejpozději do dvou měsíců ode dne dodání vyúčtování za poskytnutou službu. V případě zamítnutí reklamace se mohou účastníci obrátit na ČTÚ, který na své webové stránce poskytuje mnoho informací, jak dále postupovat. [7]

2.3 Big Data

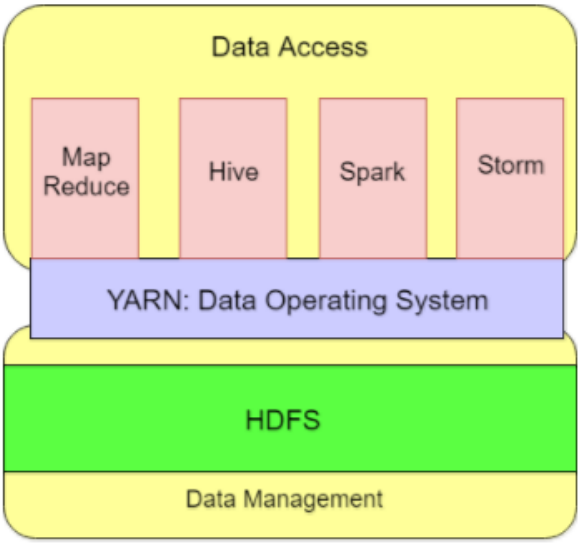
Pojmem Big Data jsou označovány techniky a technologie pro zachycování, ukládání, zpracování, distribuování a analyzování dat, jejichž velikost odpovídá řádově petabajtům (1 PB= 1024 TB) a vyšší. Pro představu, denně je generováno 2.5 exabajtů dat (1 EB = 1000 PB), což může představovat např. 530 milionů hudebních skladeb či video záznamy v HD kvalitě, které by postačily na 90 let nepřetržitého přehrávání. [13] [14]

Big Data také zahrnují data, která jsou produkována různými aplikacemi a zařízeními. Jsou to data např. z dopravních prostředků, sociálních sítí, komunikačních sítí či data z vyhledávacích engines. Zmíněná data mohou být strukturovaná (např.: relační data), nestrukturovaná (např.: PDF, Word) a částečně strukturovaná (např.: XML data). Big Data jsou často také spojována s „třemi V“ z anglického Volume (objem dat), Variety (různorodost dat) a Velocity (rychle se měnící data).[16]

Na trhu je řada technologií pro práci s Big Data od různých firem jako Amazon, IBM, Microsoft či Google. V této práci byl použit pro zpracování záznamů z ústředen aplikační rámec Hadoop.

2.3.1 Hadoop

Hadoop je open-source aplikační rámec provozovaný nad commodity hardwarem. Pojmem commodity hardware se v oblasti IT označují zařízení nebo komponenty, které jsou levné, rozšířené a zaměnitelné s jiným hardwarem.[11] Zmíněný aplikační rámec se využívá pro distribuované uložiště a distribuované zpracování velkého objemu dat. Všechny moduly v Hadoopu jsou navrženy tak, že aplikační rámec by měl být schopen automatických oprav hardwarových chyb. Hadoop rozděluje data do velkých bloků a distribuuje je přes všechny uzly nebo klastr. Poté do zmíněných uzlů pošle kód a následně se na těchto uzlech zpracovávají paralelně data. Výhodou toho přístupu je manipulace s lokálními daty, což přináší rychlost a efektivitu. [16] [8]



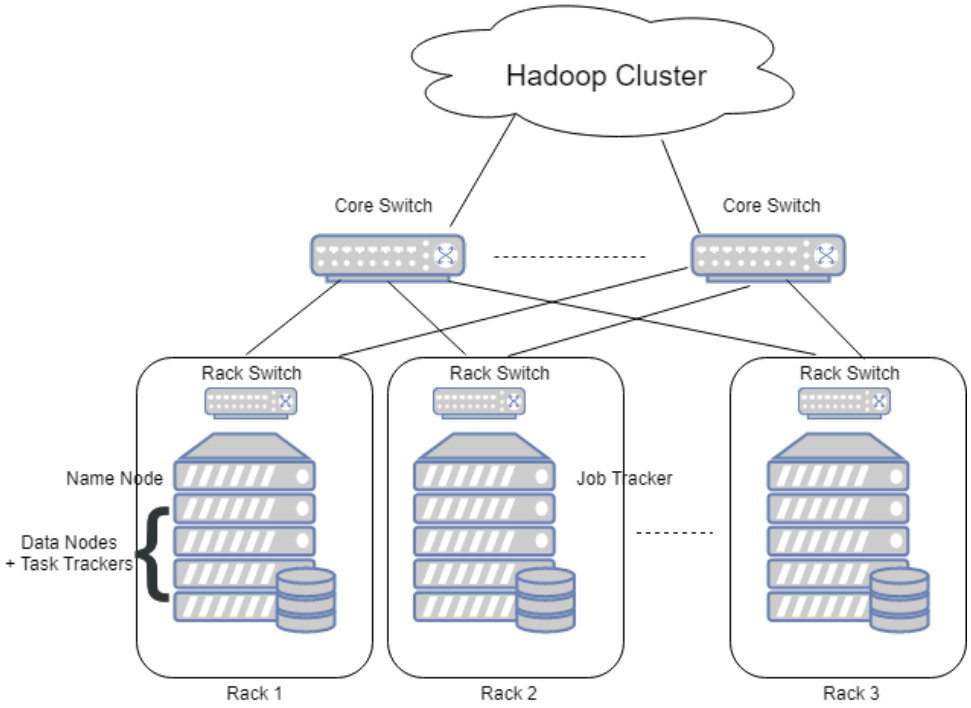
Obr. 2.12: Hadoop architektura

Na obrázku je zobrazena architektura Hadoopu. Jádrem Hadoopu je Distributed File System (HDFS) pracující jako úložná část a YARN (Yet Another Resource Manager) starající se o dostupné prostředky. Nad YARN je postaveno mnoho dalších modulů sloužících pro výpočty a jiné operace. Jako příklad jsou zde uvedeny MapReduce, Hive, Spark či Storm. Každý z těchto

modulů má jiný přístup a také se hodí na různé typy práce. [9]

Hadoop klastr

V rámci Hadoop klastru existuje Name Node a zpravidla více Data Node. Name Node je nejdůležitější prvek v celém klastru, jelikož uchovává v paměti informaci o tom, na kterých Data nodech jsou data uložena. Při výpadku nebo úmyslné destrukci tohoto uzlu data již nelze zpět rekonstruovat. Pro takový případ většinou bývá v klastru ještě Secondary Name Node. Data Node zajišťuje ukládání datových bloků (standardně 128 MB) od klientů a od ostatních Data Nodů, jelikož v rámci klastru probíhá proces replikace mezi data nody. Standardně je replikační faktor nastaven na hodnotu 3. Mimo Name Node je v racku také Job Tracker, který se stará o to, kde se která úloha spustí a řídí běh této úlohy. Dohled nad samotnou úlohou na konkrétním Data Node má na starost Task Tracker. Na uvedeném Obrázku 2.13 je popsán Hadoop klastr.[12][9]



Obr. 2.13: Hadoop klastr

2.3.2 Hive

Jedná se o standardní komponentu, která umožňuje spouštět programy v Hadoopu. Je postavena nad YARN a společně tvoří Hadoop ekosystém. Hive umožňuje pomocí jazyku Hive Query Leanguage (HQL), který je velmi podobný jazyku SQL, provádět dotazy, analýzy či výběry nad velkou tabulkou dat. Existují zde dva typy tabulek: interní a externí. Hlavní rozdílem mezi těmito tabulkami je místo, kam se data ukládají. Interní tabulka má data uložená přímo v Hive a externí mimo něj. Další rozdíl nastává v případě odstranění tabulek. Při odstranění interní tabulky se odstraní jak metadata, tak i data z Name Node. Při odstranění externí tabulky dojde pouze k odstranění metadat. [17] [18]

Hive byl zvolen pro zpracování CDR záznamů, protože jeho tabulky jsou limitovány pouze samotným hardwarem. Tato skutečnost nám umožňuje uchovávat dlouhé tabulky se záznamy hovorů, což je pro nás velmi důležité. Prováděná statistika bude vycházet ze všech dat, nikoliv

jen z dat získaných v daném okamžiku. Hive zapisuje veškeré výsledky operací na disk, a proto je pomalejší než například Spark. Ten vše zapisuje do paměti. Rychlostně je na tom Spark lépe téměř o jeden řád. Ovšem Spark zvládne úlohy s daty jen o velikosti stovek GB. Hive je schopen zpracovávat mnohem větší objemy dat, řádově v jednotkách TB a PB. Jeho pomalá rychlost je způsobena starým výpočetním paradigmatem Map-Reduce, který je ale na rozdíl od nových velmi stabilní.[10][17]

2.4 Python

Jedná se o velmi rozšířený dynamický interpretovaný jazyk, který patří do rodiny vyšších programovacích jazyků jako PHP, Java, Prolog a další. Autorem projektu Python je Guido Van Rossum. Python je flexibilní, jednoduchý, objektově orientovaný programovací jazyk, který má široké pole uplatnění. Využívá se pro tvorbu systémových a síťových programů nebo grafických rozhraní. Najde uplatnění i v oblasti databází, programování počítačových her či skriptování.

Tento programovací jazyk se stal dnes populární díky jeho rychlosti, podpoře jiných technologií, přenositelnosti mezi OS a jednoduchosti. Jedná se o open-source vyvíjený jazyk, tudíž jsou veškeré aplikace, resp. jejich zdrojové kódy volně k dispozici. Toto přináší jistou nevýhodu oproti například jazyku Java.[19] [20]

V této práci byly vytvořeny skripty v tomto jazyce s cílem získání grafických výstupů. Bylo k tomu využito vývojové prostředí PyCharm, více informací o tomto prostředí lze dohledat na oficiálních stránkách.[37]

2.5 Call Detail Records (CDR)

CDR – Call Detail Records je velmi důležitý zdroj informací z hlediska analýzy. Poskytuje telekomunikačnímu průmyslu nové možnosti jak maximalizovat příjmy. Práce s CDR záznamy je velmi náročná kvůli jejich objemu a mohou být považovány za zdroj Big Data. Tyto záznamy obecně obsahují detailní informace o telefonních transakcích, např. dobu zahájení, dobu ukončení a dobu trvání hovorů. Dále obsahují čísla volaného, volajícího, informace pro účtování atd. CDR záznamy jsou zpracovávány za účelem detekce podvodů, účtování, získávání statistik pro dimenzování sítě a jiné.[21]

Každá telefonní ústředna generuje CDR záznamy v různých formátech, bylo proto potřeba, abychom měli k dispozici dokumentaci k formátu CDR záznamu, který byl zpracováván. Dokumentace je uložena v dokumentu CDR_format_SASMEDIA, viz příloha. Níže na Obrázku 2.14 je uvedena struktura CDR záznamu. Je zde vidět např. sekvenční číslo, které nabývá hodnot od 0. do 20. bitu nebo telefonní číslo volajícího, kterému je vyhrazeno místo od 83. do 115. bitu.

Item	Explanatory text	Offset - bytes
sekv	sequential number	0
opc	Originating point code as number	20
opc_t	Originating point code as text	30
opc_gpc	Originating point code – global	39
dpc	Destination point code as number	49
dpc_t	Destination point code – as text	59
dpc_gpc	Destination point code – global	68
cic	Circuit identification code	78
Dn_a	Calling „A" number	83

Obr. 2.14: Struktura CDR záznamu

2.6 Metody na detekci fraudů

Většina metod na detekci fraudů v telekomunikačním průmyslu je dnes založena na dataminingu z CDR záznamů, které slouží jako hlavní zdroj informací. Na základě toho, co je vytěženo z dat, existují mnohé metody na detekci jednotlivých fraudů. Mezi tyto metody patří např. detekce na základě různých pravidel, pomocí neuronových sítí, strojového učení, vizualizace dat nebo detekce anomálií. Následuje stručný popis metody využívající detekce na základě pravidel a metody vizualizace dat. Tyto dvě metody byly využity v praktické části této práce.

Metoda detekce na základě pravidel je založena na hledání specifických signatur dat, o kterých zcela jistě víme, že jejich výskyt indikuje podezřelou aktivitu. Využívání této metody vyžaduje přesnost pravidel a rychlou reakci na nové fraudy. Každý nový typ fraudu znamená definování nových pravidel.

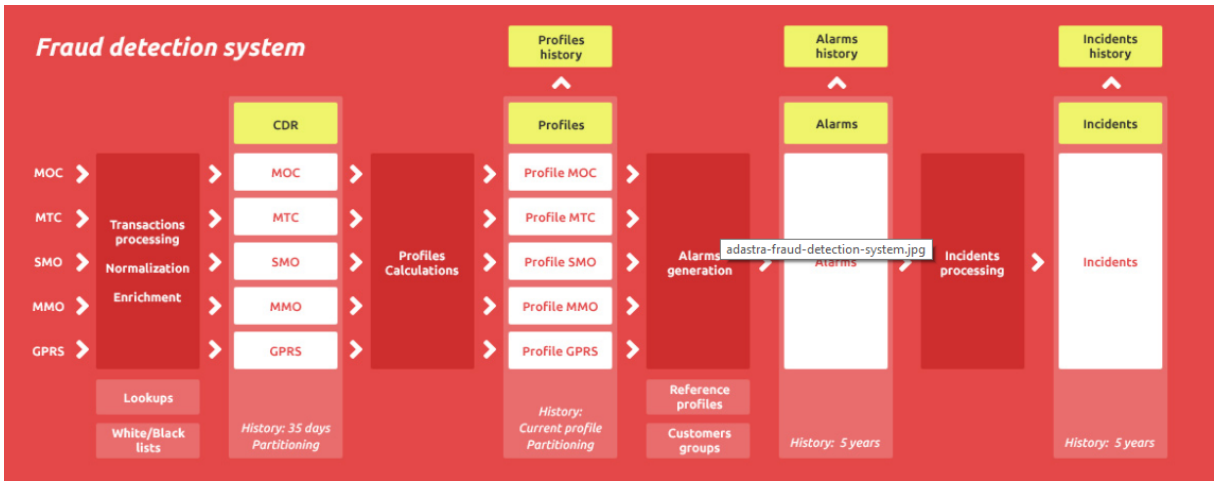
Při metodě využívající vizualizaci dat se sestavují grafy nebo vytvářejí vizualizace. Tato metoda je závislá na lidské rozpoznávací schopnosti. Na základě dané vizualizace člověk rozhoduje, zda se jedná o fraud či nikoliv. Zmíněná metoda je dynamická a více adaptivní, a proto je schopná reakce na různé měnící se techniky podvodníků.[32]

2.7 Systémy pro detekci fraudu

Dnes existuje v telekomunikačním průmyslu řada systémů na detekci fraudů. Dále je popsáno krátké seznámení se třemi z nich. Detailnější informace o jednotlivých systémech jsou dostupné na webových stránkách výrobců.

Jako první příklad je zde uveden Adastra Fraud Detection System, který zákazníkům nabízí komplexní analýzu pro vyhledávání nežádoucího vzorce chování. Systém monitoruje provozní data a na základě nich odhaluje podvody. [23]

Na Obrázku 2.15 je zobrazeno schéma systému Adastra Fradu Detection System. Je zde vidět nejprve načtení dat z CDR záznamů, ze kterých jsou následně získána metadata o odchozích (MOC), příchozích hovorech (MTC), SMS (SMO), MMS (MMO) a metadata o mobilních datech (GPRS). Získaná data jsou dále normalizována a obohacena o vypočtené hodnoty. Dále jsou systémem vypočteny profily a identifikátory, které popisují konkrétní druh chování. Na základě identifikátorů jsou stanoveny hodnoty, při jejichž překročení dojde ke spuštění alarmu. Tento návrh lze dle výrobce používat na sledování libovolných provozních dat.



Obr. 2.15: Adastra Fraud Detection System

Dalším příkladem je systém Frades založený na detekci zneužití telekomunikačních služeb, vyhodnocení a eskalovaní, čímž významným způsobem minimalizuje výši vzniklé finanční ztráty.[24]

Jako poslední příklad je uveden systém americké firmy Transnexus. Zmíněná firma díky dlouhodobému působení ve svém oboru disponuje širokou škálou detekčních systémů. Tento systém provádí analýzu CDR v reálném čase a na základě monitoringu je schopen okamžitě zastavit jakoukoliv podezřelou aktivitu. Například při detekci vysokého počtu volání do drahých destinací je systém schopen ihned tento hovor terminovat.[5]

3 Praktická část

Pro seznámení se s Hadoop ekosystémem byl připraven virtuální stroj (VM) s operačním systémem z rodiny Linux, konkrétně se jedná o distribuci Ubuntu 16.04 LTS. Konfigurace virtuálního stroje je uvedena v Tabulce 3.1. Hadoop, dle uvedeného zdroje [38], je možné také nainstalovat na operační systémy rodiny Windows. Před samostatnou instalací byl VM vybaven nejaktuálnější verzí Javy a webovým serverem Apache.

V této části práce byl nainstalován Hadoop ve verzi 2.6.1 v Local/Standalone módu, který je jednodušší na konfiguraci a zároveň nejméně náročný z hlediska hardwaru. Standalone mód se liší od zbylých dvou módu Pseudo Distributed a Fully Distributed v tom, že Hadoop v tomto módu běží jako jeden Java proces. U zbylých módů jednotlivé komponenty Hadoop běží jako oddělené Java procesy. Cílem této části práce bylo nainstalovat a zprovoznit funkční Hadoop ekosystém.

Jako řešení se zprvu jevilo také využití distribuce firmy Cloudera nebo Hortenworks, kde je Hadoop již nainstalován a vyladěn. Ovšem hardwarové požadavky těchto distribucí jsou velmi náročné. Distribuci Hortonworks nebylo možné vůbec spustit a distribuci Cloudera bylo možné spustit, ovšem práce v ní už nebyla plynulá a reakce byly pomalé. Proto byl Hadoop ručně nainstalován na běžící linuxový operační systém.

Virtuální stroj s Hadoop ekosystémem	
Komponenta	Konfigurace
CPU	2x Intel(R) Core(TM) i5-6300 CPU @ 2.40GHz 2.50 GHz
RAM	4 GB
Disk	20 GB

Tab. 3.1: Virtuální stroj s Hadoop ekosystémem

3.1 Hadoop ekosystém - příprava pracovního prostředí

3.1.1 Instalace Hadoop

Framework byl stažen z oficiálního FTP serveru [41]. Stažený soubor byl dále extrahován a přesunut do námi zvoleného adresáře. Po tomto kroku bylo potřeba upravit konfigurační soubory hdfs-site.xml, core-site.xml, mapred-site.xml, hadoop-env.sh a nastavit cestu k Javě v souboru /.bashrc. Po výše uvedené konfiguraci a nastavení jednotlivých konfiguračních souborů je Hadoop připraven ke spuštění. Jednotlivé kroky instalace jsou dostupné ze zdroje [34]. Před spuštěním bylo ovšem nutné zformátovat Name Node.

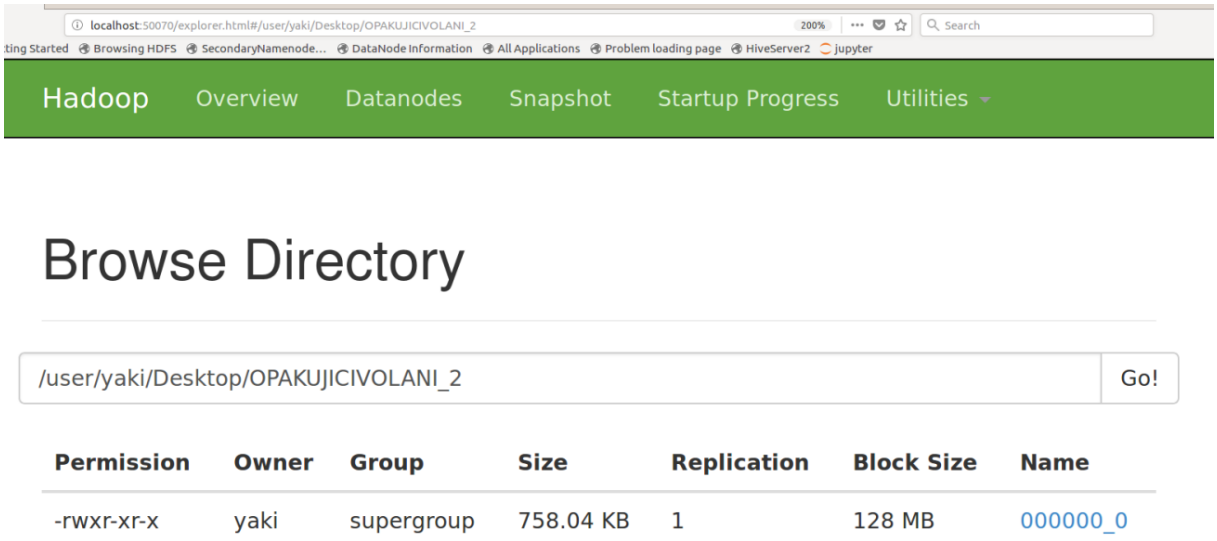
Na Obrázku 3.1 byly příkazem *jps* zobrazeny všechny běžící procesy. Po restartování VM je potřeba všechny tyto procesy spustit skriptem start-all.sh.

```
yaki@HadoopSingleNodeMicro:/usr/local/hadoop/etc/hadoop$ jps
19335 SecondaryNameNode
19033 NameNode
18169 ResourceManager
20346 Jps
19146 DataNode
18283 NodeManager
```

Obr. 3.1: Hadoop - running

Dále byla ověřena funkčnost Hadoopu přístupem na webový server, běžící na adrese localhost (http://localhost:8088/ a http://localhost:50070/). Z tohoto webové rozhraní je možné stahovat soubory z HDFS, získávat informace o stavu jednotlivých uzlů či kontrolovat log záznamy.

Na níže uvedeném Obrázku 3.2 je zobrazeno webové rozhraní nainstalovaného Hadoopu. Je zde vidět soubor, který má definovaná přístupová práva, jeho vlastník, celková velikost, a počet replikací. Jelikož je Hadoop nainstalován ve Standalone módu, neprovádí replikace. Block size je definovaná velikost, kterou Hadoop považuje za jeden blok dat.

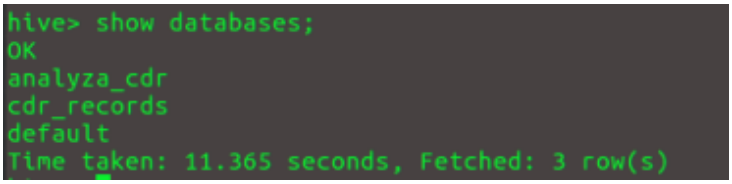


Obr. 3.2: Přístup k HDFS přes webové rozhraní

3.1.2 Instalace Hive

Hive byl stažen z oficiálního FTP serveru [40]. Po stažení bylo provedeno extrahování a stejně jako u instalace Hadoop bylo potřeba nastavit cesty v souboru /.bashrc a nastavit konfigurační soubory hive-env.sh, hive-site.xml. Jednotlivé kroky nastavení byly provedeny dle návodu ze zdroje [35]. Po výše uvedených krocích byl nainstalován Hive ve verzi 2.3.0, následně bylo možné Hive spustit jednoduše z příkazové řádky příkazem *hive*.

Na níže uvedeném Obrázku 3.3 je zobrazena příkazová řádka Hive s výpisem existujících databází. Všechny názvy databází, tabulek, funkcí či příkazů v Hive budou zvýrazněny v texu kurzívou. Ovšem ve výpisech použitých příkazů nebo obrázcích nejsou už nijak zvýrazněny. Databáze *default* je po instalaci Hive vždy vytvořena a nejde smazat. Zbylé dvě databáze *analyza_cdr* a *cdr_record*, které jsou viděny na obrázku, nám poslouží v další části. Jazyk HQL není case sensitive a každý příkaz je potřeba ukončit středníkem. Často se v práci vyskytují názvy tabulky v příkazu jako *"Vstupni_CDR"* a výstup v konzoli bude *"vstupni_cdr"*. Oba názvy jsou shodné. V příkazové řádce lze využít také automatického napovídání pomocí tabulátoru.



Obr. 3.3: Hive - výpis existujících databází

3.2 Analýza CDR záznamu

Vytvořený a nakonfigurovaný VM s Hadoop ekosystémem byl dále použit pro analýzu menšího vzorku CDR záznamu. Jedná se o vzorek veřejně dostupného záznamu ze zdroje [39]. Analýza CDR záznamu byla prováděna v následujících krocích:

- příprava dat,
- analýza počtu hovorů v závislosti na daném časovém období,
- analýza počtu opakujících se hovorů,
- analýza počtu provolaných vteřin na každé telefonní číslo.

3.2.1 Příprava dat

Nejprve byla vytvořena databáze *analyza_ cdr*, poté v této databázi byla vytvořena tabulka *cdr_ nacteni*. V příkazu pro vytvoření tabulky je potřeba definovat názvy jednotlivých sloupců a jejich datový formát. Je zde definován také oddělovač, kterým jsou jednotlivé záznamy v CDR odděleny. Další příkaz obsahuje definovanou cestu k souboru, který byl načten. Vytvořená tabulka byla následně naplněna CDR záznamem.

Hive umožňuje načítat data jak z HDFS, tak přímo z lokálního uložiště. V tomto bodě je vše připraveno pro realizaci operací nad načtenými daty. Níže jsou uvedeny použité příkazy.

```
hive> create database analyza_cdr;

hive> create table cdr_nacteni (cf string, ct string, ts string, te string,
tt string, n bigint, sts string, pr double) row format delimited fields
terminated by "," lines terminated by "\n" stored as textfile;

hive> load data local inpath '/home/yaki/Desktop/input.csv' overwrite
into table cdr_nacteni;
```

3.2.2 Zpracování dat

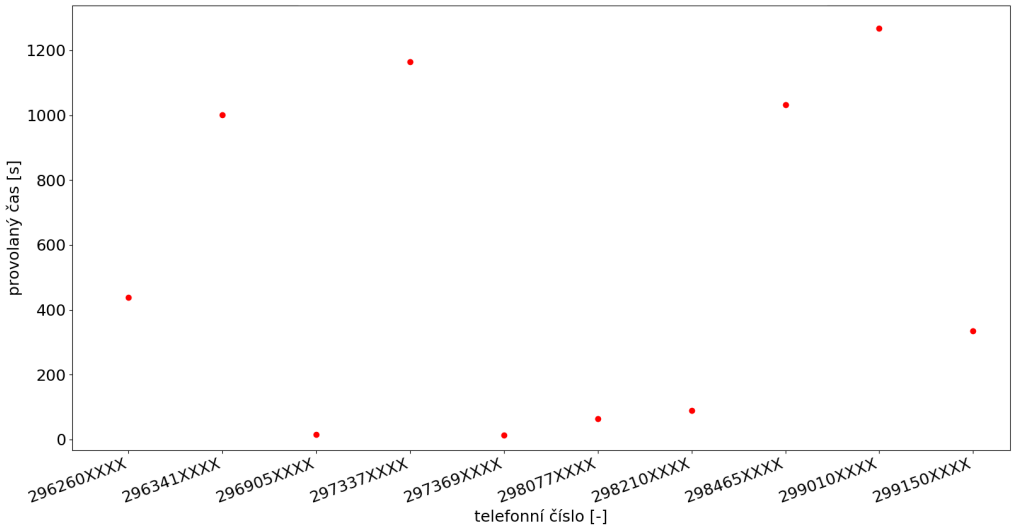
Analýza počtu provolaných vteřin na každé telefonní číslo

Pro získání grafu provolaných vteřin na každé telefonní číslo byly získány potřebné údaje, telefonní čísla a počet provolaných vteřin z tabulky *t1*. Další možností by bylo odečíst časový záznam o ukončení hovoru a zahájení hovoru. V našem případě byla použita první možnost a pomocí funkce *select()* byly získány potřebné informace. V reálném případě by bylo potřeba ještě sečíst počet provolaných vteřin pro každé číslo, které se vyskytlo v CDR záznamu vícekrát. Bylo provedeno ověření, že žádné telefonní číslo se v CDR záznamu vícekrát neopakuje. Dotaz pomocí funkce *select()* v Hive příkazové řádce je uveden níže. Získaný výsledek dotazu byl následně vykopírován z HDFS a uložen na lokální uložiště.

```
hive> insert overwrite local directory 'Desktop/statistikavolani_02' row
format delimited fields terminated by '\t ' select cf,n as count from t1
group by cf,n;
```

Na Obrázku 3.4 jsou zachyceny provolané vteřiny na každé telefonní číslo. Bylo zde provedeno maskování posledních čtyř číslic znaky XXXX. Nejedná se ovšem o kompletní graf. Ten by obsahoval tisíce záznamů a stal by se nepřehledným. Jedná se pouze o část grafu, který obsahuje maximální hodnotu provolaného času.

Je zde vidět, že maximální provolaný čas: více než 1200 vteřin bylo provoláno telefonním číslem 299010XXXX. Jedná se o graf získaný z CDR záznamu, který byl generován v krátkém časovém intervalu. Tento grafický výstup byl získán pomocí skriptu statistikavolani.py, viz příloha.



Obr. 3.4: Graf počtu provolaných vteřin na každé telefonní číslo

Analýza počtu hovorů v daném časovém období

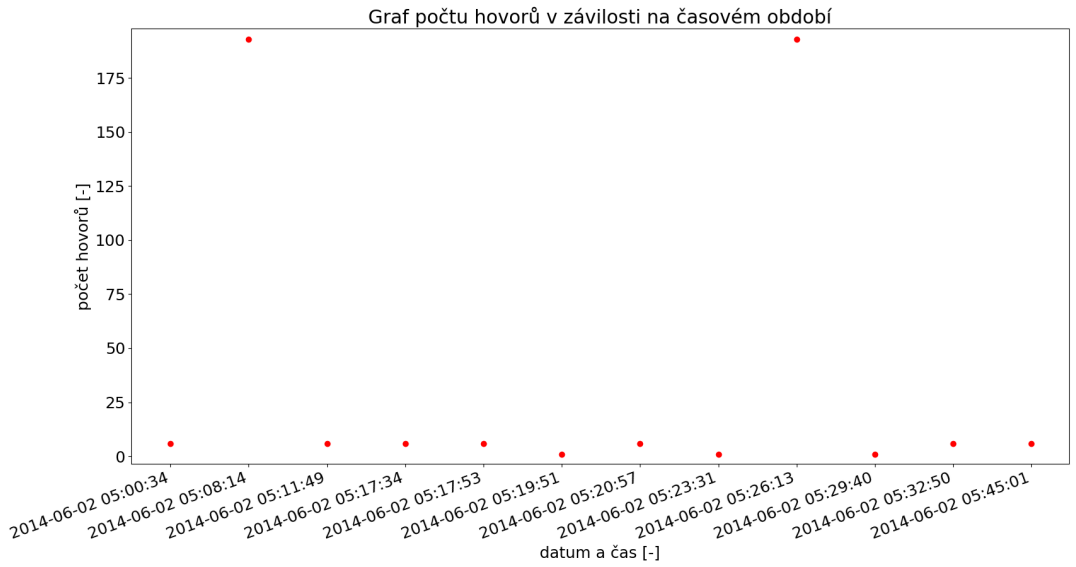
Tato analýza spočívá v získání přehledu o celkovém počtu provedených hovorů v jeden daný časový okamžik. Vzorek CDR záznamu neobsahoval žádné opakující se hovory, byly proto záměrně některé záznamy duplikovány, abychom mohli nalézt opakující se hovory. Data pro vykreslení grafu závislosti počtu hovorů v daném časovém období byla získána z tabulky *cdr_nacteni* následujícími kroky. Jako první logický krok se jeví všechny záznamy v tabulce časově seřadit.

Každá ústředna ovšem generuje různé formáty časových údajů a v našem případě bylo potřeba tento časový údaj převést do formátu, kterému Hive rozumí a následně získané informace uložit do nové tabulky *t2*.

Po převedení do správného časového formátu a seřazení byly hovory sečteny a seskupeny dle časového záznamu. Výsledek dotazu byl následně vykopírován z HDFS a uložen na lokální uložiště.

Dále jsou uvedeny Hive příkazy na vytvoření nové tabulky *t2*, konverzi času do správného formátu a následné uložení výsledků do tabulky *t2*. Poslední příkaz provedl sečtení hovorů pomocí funkce *count()* a seskupení dle časového záznamu pomocí funkce *group by*.

```
hive> create table t2 (cf string, ct string, ts timestamp, n bigint) row
format delimited fields terminated by "\t " lines terminated by "\n"
stored as textfile;

hive> insert overwrite table t2 select cf,ct,from_unixtime(unix_timestamp(ts,
"dd/MM/yyyy HH:mm:ss")),n from cdr_nacteni;

hive> insert overwrite directory 'Desktop/casovydiagram_02' row format
delimited fields terminated by ', 'select count(cf),ts from t2 group by ts;
```



Obr. 3.5: Graf počtu hovorů v daném časovém období

Na Obrázku 3.5 je vidět, kolik hovorů se v daný den a čas realizovalo. Na obrázku je zachycen pouze krátký časový interval. Můžeme si také všimnout dvou bodů na grafu, které jsou

nad hodnotou počtu opakování 175. Jedná se o hodnoty získané duplikací některých telefonních záznamů v CDR.

Uvedený graf neobsahuje ovšem všechny záznamy. Graf s více jak tisíci záznamy by se stal dosti nepřehledným. Tento grafický výstup byl získán pomocí skriptu casovydiagram.py, viz příloha.

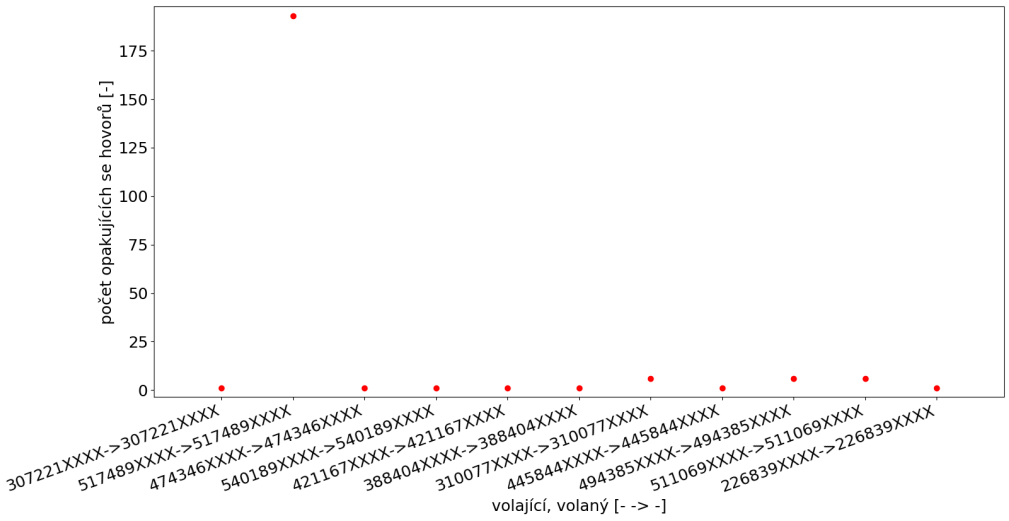
Analýza počtu opakujících se hovorů v daném časovém období

Oproti předchozí analýze zde zjišťujeme počet opakujících se hovorů mezi stejným volajícím a volaným během definovaného časové intervalu. Pro získání počtu opakujících se hovorů bylo nutné získat informace o volajících a volaných, kteří si za dané období volali vícekrát.

Z tabulky byly následně pomocí příkazů *select()* a *group by* vybrány informace o volajícím a volaném, dále byly seřazeny do skupin a následně byla pomocí funkce *count()* spočtena četnost těchto skupin. Výsledek z tohoto dotazu byl následně vykopírován z HDFS a uložen na lokální úložiště. Níže jsou uvedeny použité příkaz.

```
hive> insert overwrite directory 'Desktop/kdo_komu_kolikrat_volal_02' row
format delimited fields terminated by ',' select count(cf), cf, ct from t2
group by cf, ct;
```

Na Obrázku 3.6 lze vidět počet opakujících se hovorů v daném časovém období. Je zřejmé, že počet hovorů mezi telefonními čísly 517489XXXX a 517489XXXX přesahuje hodnotu 175. Počet opakujících se hovorů u ostatních čísel se pohybuje mnohem níže. Takovýto výskyt v reálném telefonním provozu by nás mohl upozornit, že se jedná o velmi podezřelé chování a bylo by vhodné tato dvě čísla prověřit. V našem případě je tento výskyt způsoben duplikací hovorových záznamů v našem CDR záznamu. Bylo zde opět zajištěno maskování a pro vykreslení byl použit skript kdo_komu_kolikrat_volal_02.py, který je uveden v příloze.



Obr. 3.6: Graf počtu opakujících se hovorů v daném časovém období

3.2.3 Automatizace

Aby bylo možné realizovat jednoduchý systém na detekci nebo analýzu, bylo potřeba výše zmíněné kroky z praktických důvodů automatizovat. Cílem bylo spustit skript, který by provedl všechny analýzy a následně zobrazil nebo uložil grafické výstupy. Hive umožňuje spouštět sekvenci příkazů ze skriptu a pro tento účel byl vytvořen skript automatizace.hql. Tento skript obsahuje použité hive dotazy, které mají pevně dané pořadí. Jedná se o dotazy typu vytvoření tabulek, filtrování dat, konverze do správného časového formátu, uložení výsledků do tabulek a vykopírování výsledků z HDFS.

Dále bylo zapotřebí po uložení výsledků získat grafické výstupy. Pro tento účel byl vytvořen jednoduchý bash skript, který spustí všechny tři zmíněné skripty na vykreslení grafů. Aby byly jednotlivé grafické výstupy odlišeny, byla do názvu uložených výstupů vložena časová razítka.

```
#!/bin/bash
#spusteni sady HIVE dotazu
hive -f automatizace.hql

#spusteni grafickeho vystupu
./statistikavolani.py
./kdo\_komu\_kolikrat\_volal\_02.py
./casovydiagram.py
```

Výše je uveden bash skript spust.py, který obsahuje příkazy na spuštění skriptů automatizace.hql, statistikavolani.py, kdo_komu_kolikrat_volal_02.py a casovydiagram.py Protože jsou CDR záznamy z ústředen generovány v určitých časových intervalech, je možné pomocí softwarového démona Cron skript spust.py pouštět pravidelně podle časových intervalů ústředen za podmínky, že CDR záznamy budou ukládány na lokálním stroji, na kterém běží Hadoop. V případě, že je potřeba CDR záznamy ještě kopírovat z jiného zdroje, bylo by na místě vytvořit skript, který by se např. pomocí protokolu SSH přihlásil na zmíněný zdroj s CDR záznamy a stáhnul by je na stroj, kde běží Hadoop.

Níže je uvedena část skriptu automatizace.hql s jednotlivými příkazy. Jedná se o použité příkazy z předchozích analýz v pevně dané sekvenci.

```
#HQL skript
show databases;
use ANALYZA_CDR;
drop table CDR_nacteni;
.
.
.
create table T1 (CF String, CT String, TS Timestamp, N BigInt) row format
delimited fields terminated by "\t " lines terminated by "\n" stored as
textfile;

load data inpath '/user/yaki/Desktop/konvertovani/000000_0' overwrite
into table T1;

INSERT OVERWRITE DIRECTORY 'Desktop/casovydiagram_02' row format
delimited fields terminated by ', 'SELECT COUNT(CF),ts FROM T1 GROUP BY ts;
```

3.3 Návrh řešení pro zpracování CDR záznamů

Dle zkušeností z předchozího praktického zpracování CDR záznamů bylo navrženo následující řešení, které bude zpracovávat produkční CDR záznam pomocí Hadoop klastru. Námi navržené řešení obsahuje následující kroky:

- Předzpracování CDR záznamů.
- Načtení do Hive a následné operace.
- Exportování a grafické výstupy.

V této kapitole jsou zpracovávány produkční CDR záznamy, které nejsou veřejně dostupné. Nejprve byly provedeny obdobné analýzy jako v předchozí podkapitole, čímž jsme ověřili přenositelnost řešení do reálného prostředí. Na Obrázku 3.7 je zobrazen postup zpracování. CDR záznamy budou předzpracovány skriptem do definovaného formátu. Tyto předzpracované záznamy budou dále načteny do Hive tabulky, ve které se provedou konkrétní operace. Výsledky uložíme do souborů a pomocí skriptů na vykreslení získáme jejich grafickou podobou.

Obr. 3.7: Návrh řešení pro automatické zpracování CDR záznamů

Oproti předchozí kapitole byl zde využit Hadoop klastr. V uvedené Tabulce 3.2 jsou informace o Hadoop klastru, který je součástí našeho návrhu. Je zde uvedena hardwarová konfigurace a patřičné verze jednotlivých komponent. Na tomto klastru je nainstalována distribuce Cloudera, která je v České republice nejvíce rozšířená. Na klastru je nyní nainstalován Hadoop ve verzi 2.6.0 a Hive ve verzi 1.1.0. Detailnější informace o nainstalovaném software na Hadoop klastru jsou uvedeny na oficiální webové stránce MetaCloudu. [42]

Hadoop klastr - MetaCloud	
Komponenta	Konfigurace
CPU	2x 8-core Intel Xeon E5-2630 v3 2.40 GHz
RAM	128 GB
Disk	2x 1 TB systém, 12x 4 TB data
Síťová karta	1x Infiniband 40 Gbit/s, 2x Ethernet 1 Gbit/s

Tab. 3.2: Tabulka parametrů Hadoop klastr

Na trhu existují ještě distribuce Hortonworks a MAPR. Velkou výhodou těchto distribucí je snadná instalace díky grafickému prostředí, provoz a správa přes webové rozhraní. Viz podkapitola 3.1.1, můžeme s jistotou potvrdit, že použitím výše zmíněných distribucí lze ušetřit hodně času během nasazení. Rozdíl mezi distribucemi je hlavně otázkou licencí a peněz. Distribuce Cloudera je placená a je potřeba si platit za roční licenci. Distribuce Hortonworks poskytuje bezplatné používání distribuce a platí se pouze za uživatelskou podporu. [9]

Dle vyjádření provozovatele clusteru, uživatelé mají k dispozici plný výkon, který je limitován pouze hardwarem. Pokud ve výpočetní části (YARN) zabírá jeden uživatel více jak 85% clusteru a ve frontě čekají další joby, je povolena preempce.

3.3.1 Předzpracování CDR záznamů

V této kapitole je popsán postup předzpracování CDR záznamů pomocí vytvořeného skriptu parsing.py v jazyce Python. Formát CDR záznamů je definován a detailněji popsán v dokumentu dodavatele CDR_format_SASMEDIA, ve kterém jsou uvedena přesná uspořádání bitů jednotlivých hodnot. Do záznamů bylo potřeba vložit oddělovače, v našem případě jsou to čárky, které nám umožnily v Hive tabulce načítat tyto hodnoty do jednotlivých sloupců. Dalším potřebným krokem je konverze časové značky v linuxovém formátu do formátu, který je pro nás lépe čitelný, a to konkrétně YYYY-MM-DD HH:MM:SS.

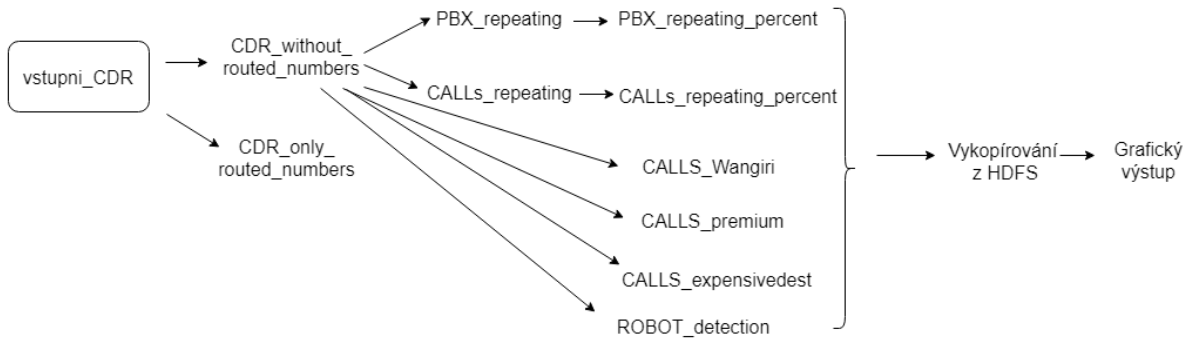
Výstupem z této části jsou CDR záznamy, které obsahují oddělovače za každou hodnotou a časové záznamy jsou převedené do správného formátu. Celý skript je uveden v příloze, zde je uvedena pouze jeho část. Hlavní operací je načtení celé řádky z původního CDR záznamu a dle uspořádání jednotlivých hodnot přidání oddělovače a následně uložení výsledku do nového souboru.

```
for input_file_name in input_file_names:
with gzip.open(input_file_path, "rb") as input_file:
for line in input_file:
l = str(line)
write_file_column(output_file, l[2:22])  # sekv
write_file_column(output_file, l[22:32])  # opc
.
.
.
write_file_column(output_file, l[8194:-2], "\n")  # CR
output_file.close()
```

3.3.2 Načtení do Hive a následné operace

V této části práce je popsáno načtení CDR záznamů do Hive a jejich následné zpracování. Na níže uvedeném schématu jsou zobrazeny kroky k získání námi požadovaných výstupů. V každém kroku byla vždy vytvořena nová tabulka, do které byly uloženy výsledky. Hive umožňuje také uložení výsledků mimo HDFS, což je také často v této práci využíváno. Proces vytvoření tabulek, následné načtení dat do nich a další operace jsou zaznamenány krok po kroku v příloze.

Na níže uvedeném Obrázku 3.8 je zobrazeno schéma zpracování. Je zde zachyceno, jak v jednotlivých fázích vznikají tabulky a z kterých dat budou naplněny.



Obr. 3.8: Schéma zpracování

Nejprve byly CDR záznamy načteny do tabulky *vstupni_CDR* a následně jsme z této tabulky vytvořili další dvě. První tabulka *CDR_only_routed_numbers* je vyplněna především záznamy hovorů, které byly dále směrovány mimo síť operátora.

Druhá tabulka *CDR_without_routed_numbers* byla naplněna zbylými záznamy, tedy hovory, které zůstaly v síti operátora. Z této tabulky byly získány informace o počtu volání mezi ústřednami a počtu volání mezi jednotlivými koncovými uživateli. Tyto informace byly uloženy do tabulek *PBX_repeating* a *CALLs_repeating*.

Tabulky *PBX_repeating* a *CALLs_repeating* obsahují velký počet záznamů a vygenerování grafických výstupů z nich by bylo výkonově a časově náročné. Abychom tomu předešli, realizovali jsme následující postup. Z tabulek *PBX_repeating* a *CALLs_repeating* byly nejprve získány maximální hodnoty počtu opakujících se hovorů, které budou představovat 100%. Zbylé hodnoty počtu opakujících se hovorů budou následně vypočteny právě vůči těmto maximálním hodnotám.

Tabulky *PBX_repeating_percent* a *CALLs_repeating_percent* obsahují kromě čísla volajícího, volaného a počtu opakujících se hovorů mezi nimi také přepočtené hodnoty na procenta. Z těchto tabulek je možné nyní definovat hraniční hodnotu, která nás bude zajímat. Například, pokud nás budou zajímat jen záznamy s počtem opakování kolem 80% vůči maximálnímu počtu opakování. Díky takové podmínce bude počet záznamů na výstupu relativně méně a vykreslení takového grafu záležitostí okamžiku.

Před samotným vykreslením je potřeba uložit výsledky na lokání stroj z HDFS. Tyto výsledky byly následně staženy a uloženy na jiný stroj a zde pomocí skriptů v jazyce python vykresleny grafické výstupy. Na Obrázku 3.9 je zachycena Hive konzole s výpisem všech tabulek, které jsme v rámci této práce vytvořili.

```
hive> show tables;
OK
calls_expensivedest
calls_premium
calls_repeating
calls_repeating_percent
calls_wangiri
cdr_only_routed_numbers
cdr_without_routed_numbers
pbx_repeating
pbx_repeating_percent
vstupni_cdr
Time taken: 0.327 seconds, Fetched: 10 row(s)
```

Obr. 3.9: Výpis tabulek v příkazové řádce Hive

Veškeré operace v Hive jsou přiloženy v souboru v příloze, kde jsou detailněji popsány. V dále uvedené tabulce je uvedeny použité příkazy v jazyce HQL.

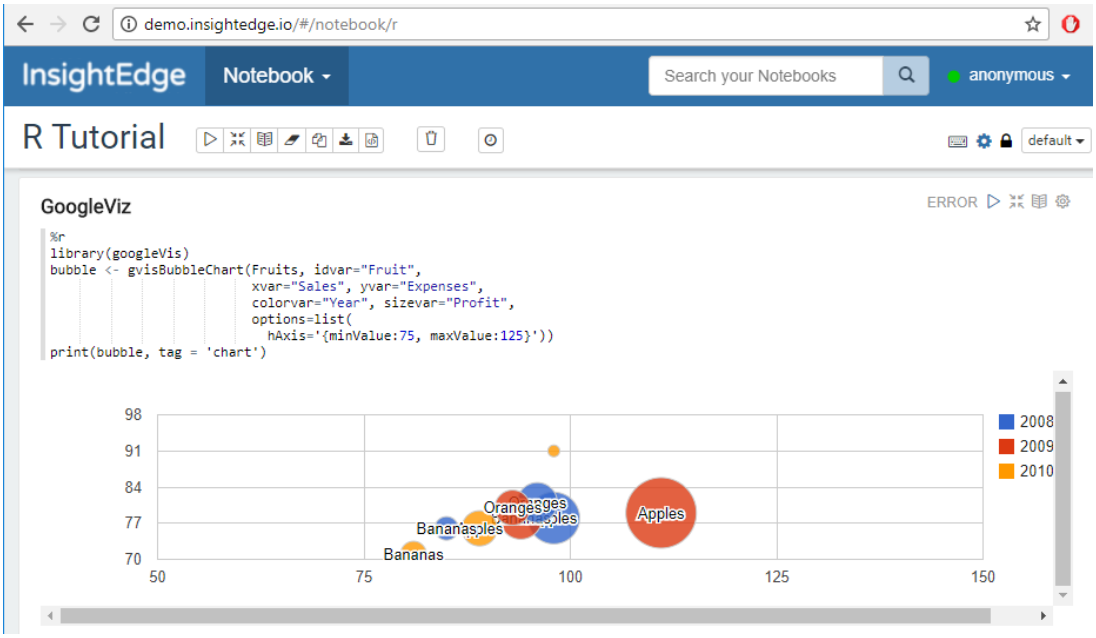
Prvním příkazem byla vytvořena tabulka *PBX_repeating* obsahující tři sloupce. První dva sloupce měly formát textového řetězce a poslední byl celočíselného typu. Druhý příkaz provedl dotaz, jehož výstupem byl seznam obsahující identifikace obou ústředen a počet opakujících se hovorů mezi nimi. Třetí příkaz provedl to samé, jen s tím rozdílem, že výsledek byl následně uložen do tabulky *PBX_repeating*.

```
hive> create external table pbx_repeating (pbx1 string,pbx2 string,opakovani
int) row format delimited fields terminated by "," lines terminated by "\n"
stored as textfile;

hive> select opc,dpc,count(opc) from cdr_without_routed_numbers group by opc,
dpc;

hive> insert overwrite table pbx_repeating select opc,dpc,count(opc) from cdr_
without_routed_numbers group by opc,dpc;
```

3.3.3 Exportování výsledků a grafické výstupy

V této kapitole byly grafické výstupy vykresleny pomocí skriptů v jazyce Python či grafického editoru yEd. Za zmínku zde stojí ještě uvést notebooky Jupyter a Zeppelin. Notebook v tomto kontextu představuje interaktivní prostředí pro spuštění například HQL dotazu či Python nebo Ruby kódu. Lze získat ihned interaktivní grafický výstup. Krom toho umožňují exportování přehledných grafických reportů pro management a pravidelné spouštění dotazů v přesně definovaných časových intervalech. Toho lze využít pro automatizaci nebo v případě potřeby rychlého zobrazení dat.

Obě distribuce Cloudera a Hortonworks disponují těmito notebooky ve formě webové rozhraní. Pokud bychom chtěli pomocí těchto notebooků zobrazit grafické výstupy s více jak několika tisíci záznamy, operace vykreslení by trvala více jak deset minut. Následná manipulace s těmito výstupy se zmíněným počtem záznamu je prakticky nereálná. Při drobných úpravách dochází často k výpadku notebooku. Kvůli těmto uvedeným omezením bylo zvoleno jiné řešení. Veškeré grafické výstupy byly vykresleny pomocí skriptů či jiných nástrojů. Na níže uvedeném Obrázku 3.10 lze vidět vykreslení grafu v jazyce ruby ve webovém rozhraní notebooku Zeppelin. Demo verzi lze vyzkoušet z uvedeného zdroje [43].

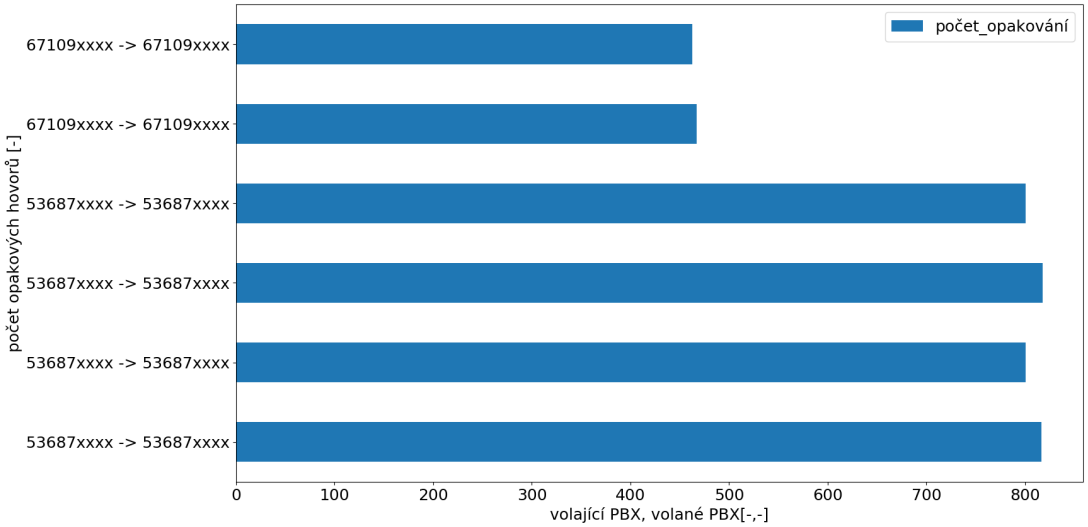


Obr. 3.10: Notebook Zeppeling

Další alternativa k získání grafických výstupů může být použití nástroje Power BI od firmy Microsoft. Tento program umožňuje připojení k HDFS a následné zpracování souborů a vytvoření grafických výstupů.

Analýza počtu opakujících se hovorů mezi ústřednami

Z tabulky *PBX_repeating_percent* byl získán grafický výstup, na kterém jsou zachyceny opakující se hovory, jejichž hodnota je vyšší než 50% proti maximálnímu počtu opakujících se hovorů. Uvedených 50% slouží pouze jako příklad, hodnota v reálném nasazení bude jistě jiná. Díky stanovení této hraniční hodnoty bude výstupní graf dynamický a vykreslené hodnoty se budou vždy vztahovat k maximální hodnotě. K vykreslení byl použit skript PBX_repeating_chart.py, jehož výstupem jsou dva soubory: obrázek ve formátu .png a soubor ve formátu .gml.

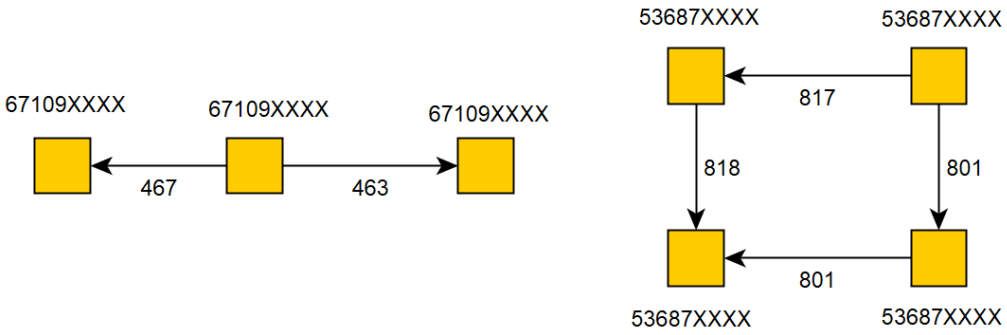


Obr. 3.11: Graf počtu opakujících se hovorů mezi ústřednami

Výše na Obrázku 3.11 je zobrazen grafický výstup počtu opakujících se hovorů mezi ústřednami v časovém intervalu 30 minut ze záznamů jedné ústředny. Oproti Obrázku 3.6 na začátku v praktické části byl nyní zvolen přehlednější způsob prezentace, a to pomocí dvourozměrného pruhového grafu.

Na ose x jsou viděny počty opakujících se hovorů mezi ústřednami a na ose y jsou číselná označení ústředen a směr hovoru. Bylo zde použito maskování, ovšem je důležité se zmínit, že se nejednalo o maskování telefonních čísel, ale

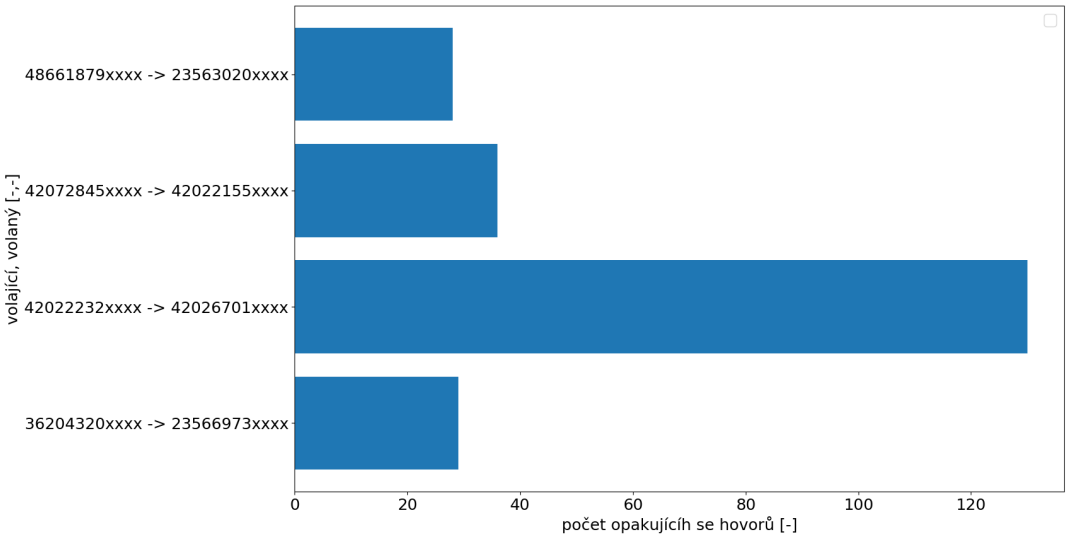
Na Obrázku 3.12 je viděn stejný výsledek v grafu s hranami. Tento grafický výstup byl získán pomocí grafického editoru yEd, který umožňuje provádět vizualizaci dat z gml souborů. Bylo použito grafické uspořádání Circular. Získaný výstup je mnohem přehlednější, jsou zde zachyceny také směry komunikace mezi jednotlivými ústřednami a jejich celkový počet.



Obr. 3.12: Graf počtu opakujících se hovorů mezi ústřednami

Analýza počtu opakujících se hovorů

Z tabulky *CALLs_repeating_percent* byl získán grafický výstup pro počet opakující se hovorů mezi koncovými účastníky. Pro vykreslení byl použit skript CALLs_repeating_chart.py, detailnější popis je v komentářích skriptu. Opět jsou výstupem dva soubory. Na prvním níže uvedeném Obrázku 3.13 jsou zachyceny opakujících se hovory, jejichž hodnota opakování je vyšší než 20% vůči maximu.



Obr. 3.13: Graf počtu opakujících se hovorů

Na druhém níže uvedeném Obrázku 3.14 je graf s uzly a hranami. K vykreslení byl použit grafický editor yEd a jeho grafové uspořádání Flowchart. Důvodem využití takového zobrazení oproti Obrázku 3.13 je především jednoduchost, snazší orientace v grafu a rychlé pochopení informací, který graf poskytuje. Je zde patrné, že nejvíce hovorů bylo provedeno mezi telefonními čísly 42022232XXX a 42026701XXX. Další důležitou informací je směr komunikace znázorněný orientovanou hranou.

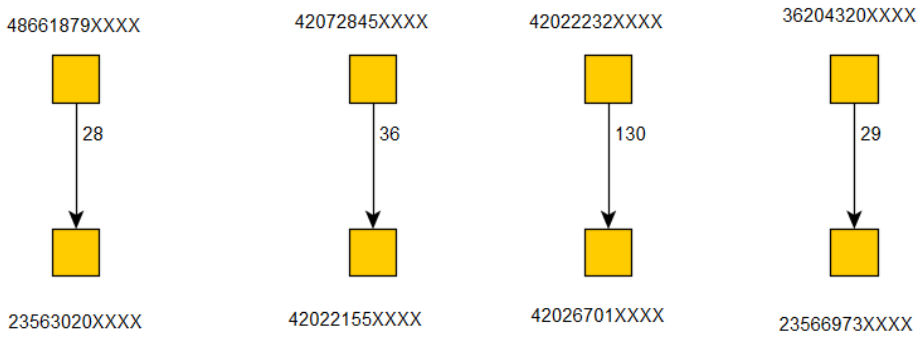
Grafická zobrazení uzlů s hranami nám mohou být velice nápomocna z hlediska detekce fraudů. Jedná se o způsob ověření, zda v daném časovém úseku nějaké telefonní číslo volalo na více telefonních čísel. Nebo zda nějaké telefonní číslo přijímalo více hovorů v jeden okamžik.

Pokud by jeden z těchto scénářů nastal, signalizovalo by to nestandardní a podezřelé chování. Takové chování není pro koncového uživatele běžné, ale mohlo by to poukazovat na robota nebo nějaký program určený k takovým úkonům.

Dalším možným využitím je pro detekci Wangiri Fraudu. U takové detekce bychom viděli uzly, ze kterých by vycházelo velké množství hran.

Na níže uvedeném Obrázku 3.15 je zachycena veškerá komunikace procházející jednou konkrétní ústřednou v jednom CDR záznamu v časovém okně 30 minut. K vykreslení byl použit editor yEd a jeho grafické uspořádání Circular.

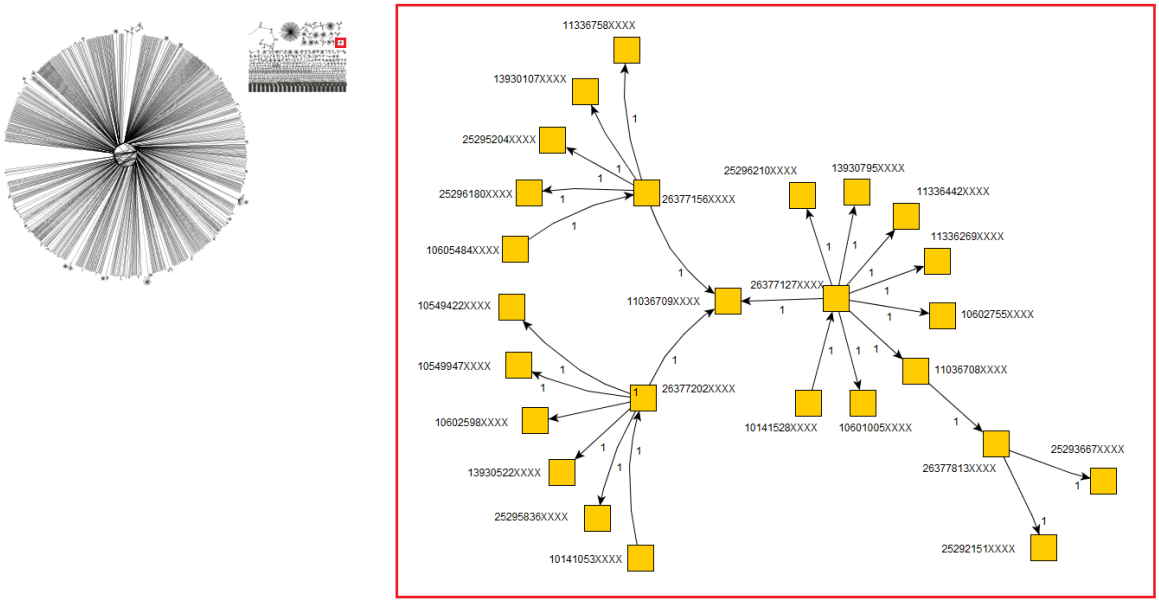
Lze zde spatřit zajímavé směry komunikace. Při přiblížení menší části grafu je viděn jeden uzel, ze kterého vychází více hran. Zmíněný uzel představuje číslo 26377127XXXX, které volalo během 30 minut na 9 dalších čísel. Jsou zde uvedeny také počty opakujících se hovorů. Takový výskyt není nijak příliš zajímavý. Ovšem v případě výskytu uzlu, ze kterého by vycházelo například 30 hran, by bylo dobré ho ověřit. Mohlo by se jednat o podezřelou aktivitu například,



Obr. 3.14: Graf počtu opakujících se hovorů

Wangiri Fraud. V případě zpracování více CDR záznamů z různých ústředen bychom získali zajímavější výskyty.

Na tomto grafu je zobrazena veškerá komunikace, bez jakéhokoliv specifického výběru dat. Například ve výše uvedených případech, kdy byly stanoveny hraniční hodnoty, se nám grafické výstupy mnohem zpřehlední a budou nám poskytnuty lepší interpretace výsledků. Tím pádem bude mít obsluha snazší rozhodování.



Obr. 3.15: Graf všech hovorů - přiblížení

3.4 Návrh možných postupů pro odhalování podvodného chování

Na základě vyjádření ČTÚ se v ČR se nejčastěji vyskytují fraudy typu one-ring and cut neboli Wangiri Fraud ze zahraničních čísel a často také dochází k zneužívání čísel 90X ke spammingu. Zneužívání čísel 90X se zvýšeným účtováním nadále existuje i přesto, že dle novely zákona o spotřebitelském úvěru je zakázáno využívání tohoto čísla k nabízení a sjednávání spotřebitelského úvěru.[7][25]

Dle těchto informací byly realizovány návrhy na detekci zmíněných fraudů. Ve všech návrzích je využíváno znalostí a zkušeností z předchozích kapitol.

V dostupném CDR záznamu z ústředny jsou telefonní čísla uvedena v mezinárodním formátu bez symbolu "+". Námi uměle vytvořené telefonní záznamy budou proto obsahovat čísla bez symbolu "+"např.: 42072043XXXX.

3.4.1 Detekce Wangiri Fraudu

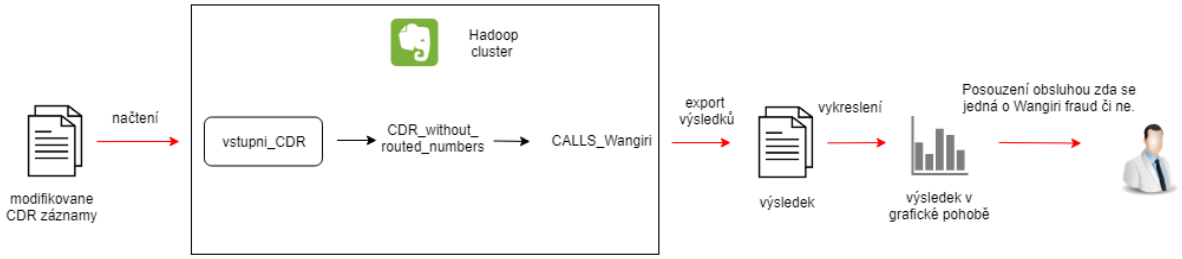
Princip tohoto fraudu je založen na masivním prozvánění telefonních čísel. Aby bylo možné detekovat takové chování, bude vycházeno z principů zpracování z podkapitoly 3.3.3. Návrh obsahuje následující kroky:

- načtení původního CDR do Hive a následné zpracování
- export výsledků a vykreslení.

Z CDR záznamu byly získány informace: volající, volaný, počet volání mezi nimi a délka hovoru. Protože se jedná jen o vyzvánění velkého rozsahu telefonních čísel, počet volání mezi čísly bude roven hodnotě 1. Délka hovoru bude rovna nule, jelikož došlo jen k prozvonění.

Každá telefonní ústředna generuje jiným způsobem své záznamy. Některé ústředny generují záznamy pouze v případě hovoru, nikoliv v případě prozvonění. Náš CDR záznam obsahuje pouze realizované hovory. Na základě těchto poznatků byla realizována simulace na detekce Wangiri Fraudu.

Pro simulaci tohoto fraudu bylo v prvé řadě potřeba vložit do CDR záznamu upravené hovorové záznamy. Bylo zvoleno jedno telefonní číslo, které prozvonilo třicet náhodných mobilních čísel začínající 42072043XXXX. Poslední 4 číslice byly vybrány náhodně tak, aby se neopakovaly. Abychom se přiblížili reálnému Wangiri Fraudu, bylo vybráno jiné telefonní číslo 25641435XXXX, což je číslo s předvolbou Ugandy, jehož poplatek činí dle O2 ceníku 169,98 Kč. Ceník je dostupný ze zdroje [50].



Obr. 3.16: Návrh postupu na detekci Wangiri Fraudu

Bylo postupováno dle schématu na Obrázku 3.16. Ovšem oproti schématu na Obrázku 3.7 byl přeskočen krok předzpracování a zpracovaní parserem. Modifikovaný CDR záznam byl nahrán do Hadoop klastru do externí tabulky *vstupni_CDR* a z ní byly následně získány tabulky *CDR_without_routed_numbers* a *CALL_repeating*. Poté byla vytvořena další tabulka *CALLS_Wangiri*, do které byly uloženy záznamy splňující podmínku počtu opakování rovno 1.

Tabulka *CALLS_ Wangiri* obsahuje nyní hovory, které se uskutečnily právě jednou. Tento výsledek byl vyexportován a uložen na počítač. Následně proběhlo grafické vykreslení.

Posledním krokem bylo samotné vyhodnocení obsluhou, zda se jedná o zmíněný fraud či nikoliv. V tomto případě se jedná o fraud.

Posloupnost HQL příkazů je uvedená v příloze. Níže je uveden jen klíčový příkaz s podmínkou, kdy počet volání mezi čísly byl roven právě hodnotě 1 a výsledek byl zároveň uložen do tabulky *CALLS_ Wangiri*. Druhý příkaz provedl stejnou operaci, ovšem s rozdílem, že data neukládá do tabulky, ale do lokálního úložiště.

```
hive> insert overwrite table ',' select * from
call\_repeating where opakovani = 1;

hive> insert overwrite local directory '/storage/brno2/home/nguyengo
/DIP/OUTPUT/callss_wangiri' row format delimited fields terminated
by ',' select * from call\_repeating where opakovani = 1;
```

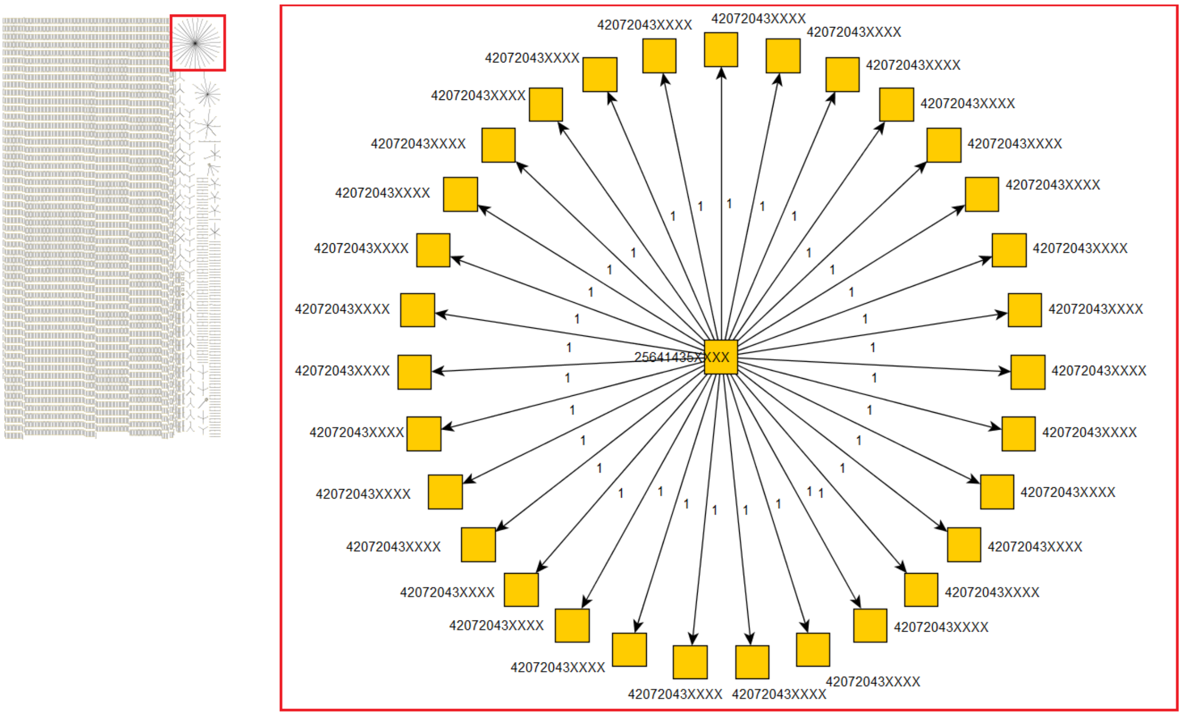
Pro ověření získaných výsledků byl proveden kontrolní dotaz na tabulku *CALLS_ Wangiri*. Níže je uveden tento dotaz s omezeným výstupem, jelikož celkové množství záznamů je 7969.

```
hive> select * from calls_repeating  where opakovani = 1;
4873383XXXX             4853739XXXX             1
4873388XXXX             4888175XXXX             1
4873389XXXX             4857776XXXX             1
.
.
.
4873389XXXX             4879020XXXX             1
4873391XXXX             4866563XXXX             1
4873398XXXX             4851618XXXX             1
Time taken: 0.315 seconds, Fetched: 7969 row(s)
```

K získání grafického výstupu byl použit grafický editor yEd a jeho grafické uspořádání Circular. Proces maskování telefonních čísel proběhl ve zmíněném editoru manuálně. Důvodem bylo sloučení posledních čtyř číslic při maskování a výsledek se tím pádem stává matoucím.

Na uvedeném Obrázku 3.17 je zobrazen náš očekávaný výsledek. Na levé straně je grafické zobrazení všech hovorů, které proběhly právě jednou. Jedná se pouze o hovory z časového intervalu 30 minut z jedné ústředny. V této části obrázku jsou viděny jednotlivé telefonní hovory znázorněné dvěma uzly a hranou seřazeny za sebou. Na první pohled si lze snadno povšimnout dvou shluků.

Po přiblížení je vidět, že největší shluk představuje právě námi simulovaný Wangiri Fraud s číslem 25641435XXXX. Jedná se o zahraniční číslo, a to konkrétně z Ugandy. Zbylá telefonní čísla byla náhodně vybraná z definovaného rozsahu. Je zde také patrné dodržení podmínky počtu volání mezi čísly rovné 1.



Obr. 3.17: Simulace Wangiri Fraudu

3.4.2 Detekce Premium Rate Service Fraud

V tomto návrhu bylo vycházeno ze skutečnosti uplatňování vyšší sazby na některá telefonní čísla začínající číslicí 9. Přehled čísel 9X s vyšší cenou pro Českou republiku dle [22] je uveden níže v Tabulce 3.3.

Předčíslí	Využití čísla
900 a 906	obchodní,inzertní a soutěžní služby
905 a 908	služby s jednorázovou cenou využívané např. pro hlasování
909	zábavné služby pro dospělé
907	přístup k veřejné datové síti a k síti Internet

Tab. 3.3: Přehled čísel 9X s vyšší cenou

Na základě těchto uvedených čísel byla získána data z naší tabulky. Samotné voláním na tyto čísla není ještě fraud. Ovšem v případě detekce, že nějaké telefonní číslo provádělo často hovory na čísla začínající 9X, je vhodné ověřit, zda dané číslo nebylo odcizeno nebo se jedná o standardní chování uživatele.

Výsledek takového jednoduchého návrhu může sloužit jako počáteční fáze detekce Premium Rate Service Fraud, Subscription Fraud, Spamming Fraudu a jiných.

Princip tohoto návrhu spočívá především ve způsobu získávání klíčových informací z tabulky *vstupni_CDR*. V rámci této podkapitoly byla provedena simulace Premium Rate Service Fraud následujícím kroky:

- modifikace CDR záznamu,
- načtení do Hive a následné zpracování
- export výsledků a vykreslení.

Nejprve do vstupního CDR záznamu byly uměle doplněny záznamy hovorů z náhodných čísel opět v rozsahu 42072043XXXX, které se realizovaly na telefonní čísla začínající číslicí 9.

V tabulce ze zdroje [22] jsou uvedeny příklady čísel, které budou v naší simulaci použity. Jedná se o čísla 900 951 234, 906 705 678, 900 509 123, 905 951 234, 908 791 234,976 951 23. Poslední číslo zajišťující přístup k Internetu nebylo zahrnuto.

Abychom dostali zajímavější výsledky, některé záznamy hovorů byly duplikovány důvodu pro získání vyššího počtu opakujících se hovorů. Pokud je v reálném prostředí detekováno volání v rámci krátkého časového intervalu na prémiové číslo třeba více jak 20 krát, je potřeba ověřit, zda se nejedná o podezřelou aktivitu.

Telefonní číslo uživatele	Čísla s vyšší cenou	Počet opakování
42072043XXXX	905951234	34
42072043XXXX	906705678	50
42072043XXXX	900509123	19
42072043XXXX	905951234	12
42072043XXXX	908791234	20

Tab. 3.4: Modifkované záznamy hovorů na prémiová čísla

V Tabulce 3.4 jsou uvedeny hovory mezi konkrétními čísly a jejich počet opakování. Telefonní čísla a počet opakování by se měly shodovat s konečnými výsledky. V této tabulce není uvedeno zbylých 25 záznamů, které uskutečnily právě jeden hovor na prémiová čísla. Dle této tabulky v končeném výstupu by mělo být detekováno, že na telefonní číslo 906 705 678 bylo uskutečněno 50 hovorů z čísla 42072043XXXX.

Jako další krok byla vytvořena nová Hive tabulka *CALLS_ Premium*, která byla následně vyplněna výsledkem z dotazu *select()* s podmínkou: první tři číslice telefonního čísla musí rovnat 900 nebo 906 nebo 905 nebo 909 nebo 908. Získaná data byla uložena do tabulky *CALLS_ Premium*. V poslední fázi byl proveden export, uložení a zobrazení výsledků. V posledním kroku bylo obsluhou rozhodnuto, zda se jedná o fraud či nikoliv.

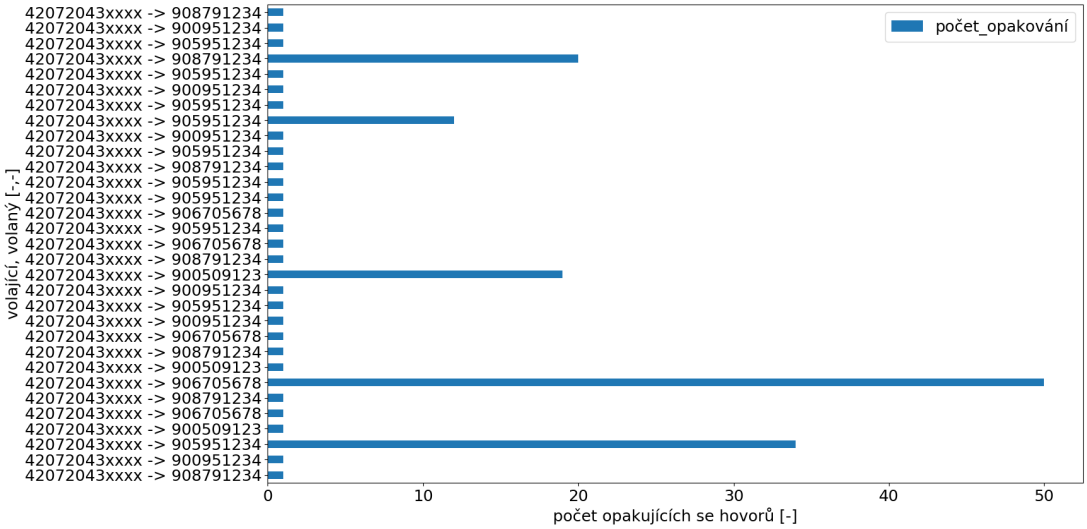
Níže je uveden HQL dotaz s využitím funkce *substring()*, která umožňuje porovnávat znaky dle definovaného intervalu v textovém řetězci. První argument v *substring()* funkci je název sloupce, ze kterého budou vybírány textové řetězce. Další dvě celá čísla definují interval znaků, které budou porovnávány. Celý příkaz realizuje uložení záznamů vyhovující podmínce do tabulky *CALLS_ premium*. Pro kontrolu je zde uveden také výstup dotazu. Výsledkem bylo 30 záznamů volajících, volaných a počtu opakování.

```
hive> insert overwrite table calls_premium select * from calls_repeating
where substring(ct,1,3)=900 or substring(ct,1,3)=906 or substring(ct,1,3)=905
or substring(ct,1,3)=909 or substring(ct,1,3)=908;

hive> select * from calls_repeating where substring(ct,1,3)=900 or
substring(ct,1,3)=906 or substring(ct,1,3)=905 or substring(ct,1,3)=909
or substring(ct,1,3)=908;
OK
42072043XXXX    908791234       1
42072043XXXX    906705678       50
42072043XXXX    900951234       1
.
.
.
42072043XXXX    905951234       12
42072043XXXX    900509123       19
42072043XXXX    908791234       1
Time taken: 0.545 seconds, Fetched: 30 row(s)
```

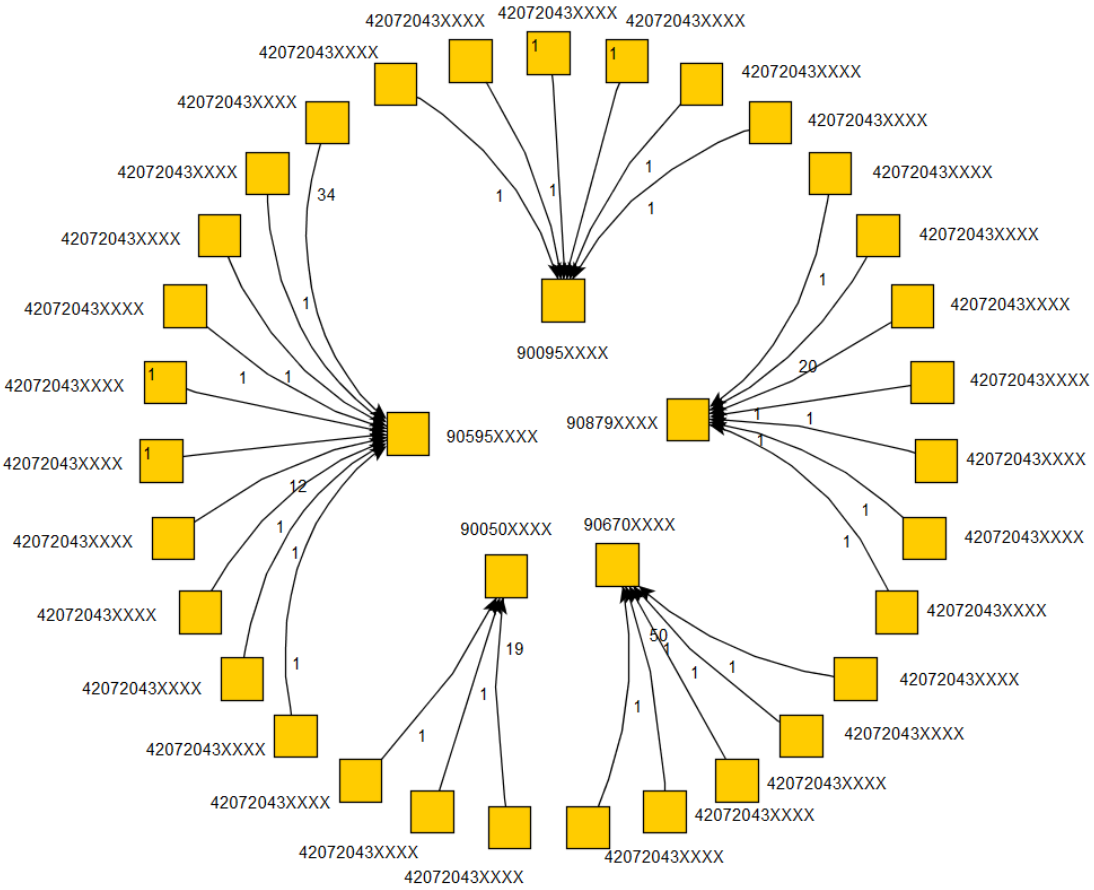
Na uvedeném Obrázku 3.18 je vidět prémiové číslo 90670XXXX, které bylo voláno nejvíce krát a to 50 krát z čísla 42072043XXXX. Další čísla z tabulky 3.4 jsou zde také zachyceny. Dle obrázku lze usoudit úspěšnost naší detekce. Příklady prémiových čísel zde nejsou maskovány, protože jsou veřejně publikované.[22]

Bylo by zde možné využít definované hraniční hodnoty opakování a realizovat tak dynamický graf. Ten by se posouval podle stanovené definované hodnoty opakování, která by byla v praxi zajímavější, stejně jako v podkapitole 3.3.3. Tímto způsobem by byly vyfiltrovány hovory opakující se právě jednou a výstupní graf by byl čitelnější. Jako důkaz, že do CDR záznamu bylo vloženo třicet záznamů, jsou zanechány všechny záznamy bez jakékoliv filtrace.

Na dalším Obrázku 3.19 je zobrazena veškerá komunikace na prémiová čísla. Způsob, jakým jsme získali hodnoty v tabulce *CALLS_ Premium* lze snadno modifikovat a aplikovat na jiné detekce. Lze spatřit hrany směřují jen na prémiová čísla, což je správně. Dále jsou zde vidět počty volání, které odpovídají vstupu.



Obr. 3.18: Graf počtu volání na prémiová čísla



Obr. 3.19: Graf komunikace na prémiová čísla

3.4.3 Detekce volání do zón s nejvyšším tarifem

Návrh detekce volání do drahých destinací je velmi podobný předchozímu. Veškeré informace o sazbách byly čerpány z online ceníku operátora O2, kde je uvedeno deset zón s příslušnou cenou za každou provolanou minutu. Tento ceník se samozřejmě mezi operátory velmi liší. Ceník firmy O2 je dostupný ze zdroje [45].

V rámci simulace detekce volání do zón s nejvyšším tarifem a ověření funkčnosti tohoto návrhu bylo vybráno pět telefonních čísel z pěti destinací. Vybrané destinace a jejich příslušné předčíslí jsou uvedeny v Tabulce 3.5.

Počet destinací lze libovoně v rámci tohoto řešení přidávat či ubírat. Bylo potřeba opět doplnit do CDR záznamu modifikované záznamy hovorů. Bylo doplněno 30 záznamů z různých telefonních čísel z našeho rozsahu 42072043XXXX.

Předčíslí	Destinace	cena s DPH [Kč/min]
98	Irán	94,37 Kč
244	Angola	113,36 Kč
86	Čína	94,37 Kč
20	Egypt	75,55 Kč
995	Gruzie	50,36 Kč

Tab. 3.5: Tabulka vybraných předčíslí a jejich cena

Nejprve byl modifikovaný CDR záznam načten do tabulky *vstupni_CDR*. Následně byla filtrací dat získána tabulka *CDR_without_routed_numbers*. Byla vytvořena nová tabulka s označením *CALLS_expensivedest* do které byla uložena data dle specifických podmínek. První podmínka: první dvě číslice telefonních čísel se budou rovnat předčíslím pro konkrétní destinaci.

Telefonní číslo uživatele	Čísla v vyšší cenou	Počet opakování
42072043XXXX	24493724XXXX	10
42072043XXXX	22423XXXX	15
42072043XXXX	98215XXXX	30
42072043XXXX	86109XXXX	25
42072043XXXX	24493724XXXX	20

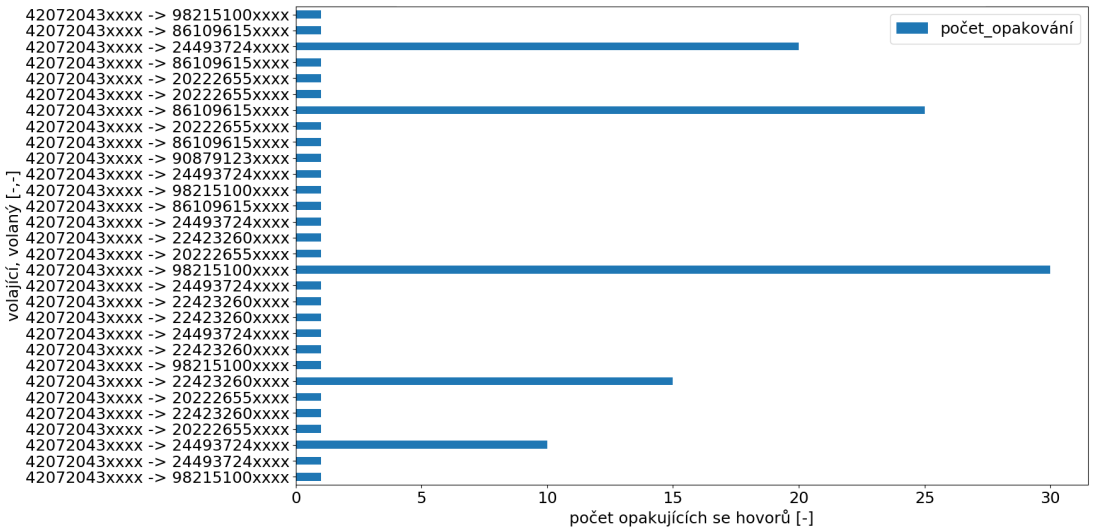
Tab. 3.6: Modifkované záznamy hovorů do drahých destinací

Oproti předchozímu návrhu je zde potřeba k první podmínce zakomponovat i druhou podmínku, a to pro první tři číslice rovnající se 244 nebo 382. Pro simulaci byly vybrány následující cílová telefonní čísla a to: 98215XXXX, 24493724XXXX, 86109XXXX, 2022265XXXX, 22423XXXX. Dále je zde uvedena tabulka s hovory, které proběhly vícekrát. Hovory uskutečněny právě jednou zde nejsou uvedené, ale měly by být vidět z grafického výstupu. V Tabulce 3.6 jsou uvedeny doplněné záznamy do CDR záznamu.

HQL dotaz je zde podobný jako v předchozím případě. Byly zde upraveny pouze intervaly znaků, které budou porovnávány. Je zde zobrazen také výstup dotazu. Výsledkem bylo 30 záznamů.

```
hive>insert overwrite table calls_expensivedest select * from CALLS_repeating
where substring(ct,1,2)=98 or substring(ct,1,2)=86 or substring(ct,1,2)=20
or substring(ct,1,3)=244 or substring(ct,1,3)=995;

hive> select * from calls_repeating where substring(ct,1,2)=98 or
substring(ct,1,2)=86 or substring(ct,1,2)=20 or substring(ct,1,3)=244
or substring(ct,1,3)=995;
OK
42072043XXXX    98215XXXX       1
42072043XXXX    24493724XXXX    1
42072043XXXX    24493724XXXX   10
.
.
.
42072043XXXX    24493724XXXX    20
42072043XXXX    86109XXXX       1
42072043XXXX    98215XXXX       1
Time taken: 0
```

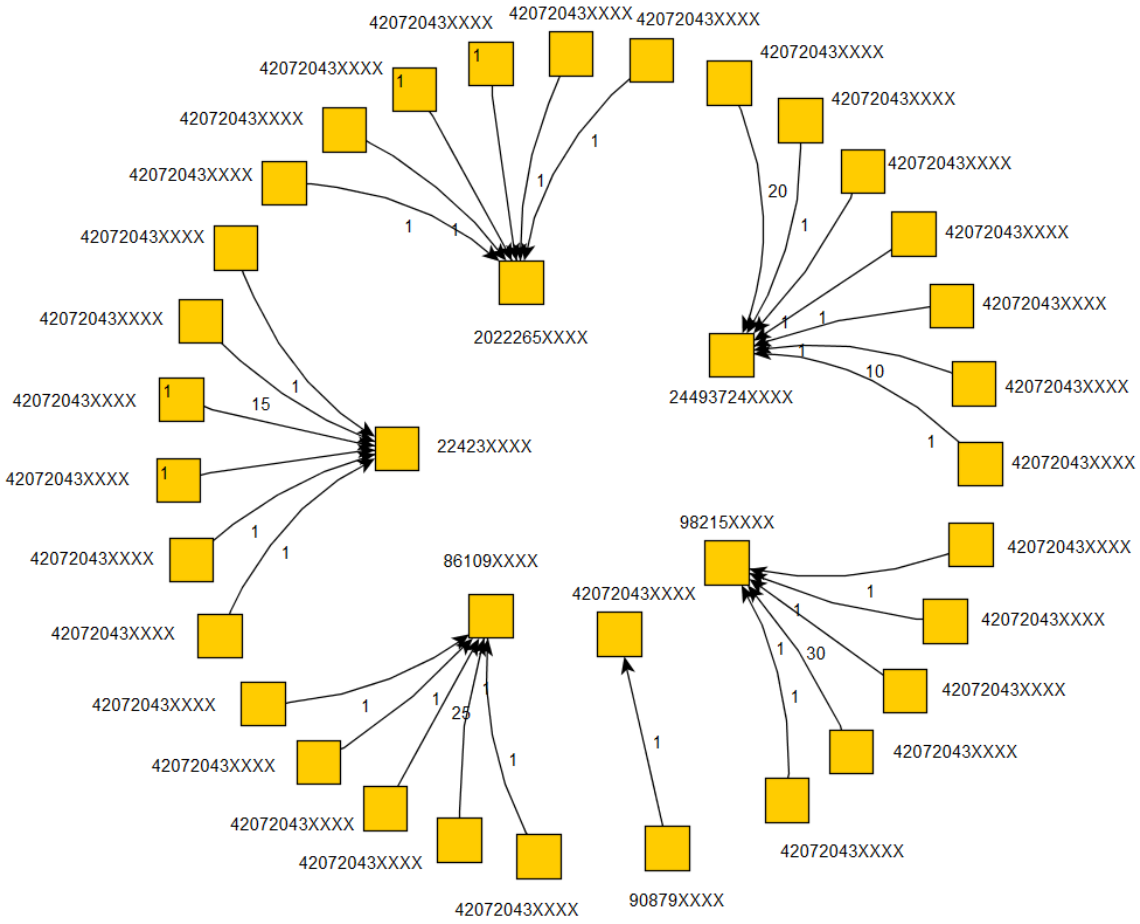
Na níže uvedeném Obrázku 3.20 jsou zobrazeny počty opakujících se hovorů do zón s nejvyšším tarifem. Grafický výsledek je shodný s vstupní Tabulkou 3.6. Nejvíce hovorů bylo prováděno z čísla 42072043XXXX do Iránu na číslo 98215XXXX.



Obr. 3.20: Graf hovorů do drahých destinací

Na Obrázku 3.21 je zobrazena jiná prezentace grafického výstupu našeho návrhu. Jsou zde zachyceny opět všechny hovory bez podmínky filtrace dat. V centru kruhového uspořádání jsou naše zvolená zahraniční čísla. Na grafu jsou také vidět počty opakujících se hovorů. Bylo zde použito grafické uspořádání Radial v editoru yEd.

Při větším množství hovorů by bylo vhodné využít definování hraniční hodnoty, která může poukazovat na podrezřelé chování. Bylo zde opět provedeno maskování telefonních čísel.



Obr. 3.21: Graf hovorů do drahých destinací

3.4.4 Detekce Call Spamming Fraud

Návrh na detekci Call Spamming Fraudu je založen na stejných principech jako návrhy na detekci Premium Rate Service Fraud a volání do zón s nejvyšším tarifem. Tento návrh nebude prakticky realizován, ale kvůli jeho aktuálnosti a především díky popularitě aplikace PhoneJokes, která je schopna provádět žertovné hovory, bude teoreticky popsán.

Za testovacím účelem byla zmíněná aplikace nainstalována a byl proveden jeden žertovný hovor. Výsledek byl překvapující. Volající opravdu nerozpoznal, že se jednalo o hlasový automat. Hovory jsou totiž velmi autentické. Žertovný hovor byl uskutečněn z čísla 672XXXXXX.

Na základě této skutečnosti bychom opět provedli HQL dotaz s upraveným intervalem znaků. Příkaz by vypadal takto:

```
hive> insert overwrite table calls_spam select * from
cdr_without_routed_numbers where substring(ct,1,3) = 672;

hive> select * from calls_spam;
```

Druhý příkaz by vypsal obsah tabulky *calls_spam*. Z výsledků z dotazu bychom následně získali grafický výstup s telefonními čísly začínající předčíslím 672.

3.4.5 Detekce SIM Card Fraud

Tento návrh je zaměřen na detekci klonovaných SIM karet. Jedná se opět o teoretický návrh. Mějme případ, kdy telefonní číslo provedlo hovor z Prahy a následně po pár minutách vytočilo další hovor ze stejného čísla, ale z Brna. Je logické, že se jedná o velmi podezřelé chování, protože jsou tyto dvě města geograficky velmi vzdálená.

Aby bylo možné detekovat takovéto chování, je potřeba, aby CDR záznamy obsahovaly informace o geografické poloze. CDR záznamy by byly načteny do Hive. Následně by se provedlo filtrování telefonních čísel, která realizovala za daný časový interval více než jeden hovor. Po této filtraci by se seřadily jednotlivé hovory podle volaného čísla. Na základě informace o geografické poloze by bylo možné nyní provádět kontrolu, zda hovory nebyly provedeny ze dvou vzdálených míst.

Bohužel CDR záznamy dostupné v rámci této práce jsou z tranzitních ústředen, kam se takové informace nepřenášejí. Pokud bychom měli k dispozici záznamy z účastnické ústředny, tento teoretický návrh by bylo možné realizovat.

V praxi existují algoritmy sloužící k monitorování klonovaných telefonů a SIM karet pomocí algoritmu Velocity Check a Collision Check, jejichž princip spočívá v porovnávání SIM karet či telefonů v závislosti na geografické poloze. V případě detekce dvou hovorů ze stejného telefonního čísla z různých geografických poloh, bude operátor upozorněn.[26]

3.4.6 Detekce robota

Tato detekce vychází z předpokladu, že běžný uživatel vlastnící telefonní číslo nedokáže provádět více hovorů v jeden časový okamžik. Koncový účastník je schopen provádět více hovorů, ale ne v jeden moment. Mějme případ vytvoření konference koncovým účastníkem. Konferencí se rozumí hovor mezi třemi a více účastníky. [46]

Do zmíněné konference jsou přidáváni další účastníci postupně. Nelze tedy realizovat telefonní konferenci tak, že koncový účastník vytočí všechny hovory v jeden časový okamžik. V případě, že by telefonní číslo dokázalo v jednom okamžiku provést například 30 hovorů, musí se jednat o robota. Pojmem robot se zde myslí specifický program, který by takové chování dokázal realizovat.

Telefonní číslo robota	Počet volaných čísel z rozsahu 42072043XXXX	Časové razítko
42077777XXXX	15	2017-06-23 23:51:34
42060222XXXX	5	2017-06-23 23:59:22
42079999XXXX	10	2017-06-24 23:40:40

Tab. 3.7: Modifkované záznamy hovorů pro detekci robota

Abychom byli schopni provést simulaci takovéto detekce, byly do CDR záznamu vloženy záznamy hovorů z telefonních čísel, které realizovaly v jednom okamžiku více hovorů. Přidané záznamy vidíme v Tabulce 3.7. Na výstupu bychom měli detekovat všechna uvedená telefonní čísla.

Detekce byla provedena následovně. Nejprve byl modifikovaný CDR záznam nahrán do Hive tabulky *vstupní_CDR*. Z ní byla následně získána tabulka *CDR_without_routed_numbers*. Pro realizaci detekce robota byly využity následující funkce v Hive a to: *count()*, *collect_list()*, *concat_ws()*, *group by*.

Funkce count provádí součet, *collect_list()* provádí ukládání hodnot do seznamu, *concat_ws()* přidává do seznamu oddělovače a *group by* umožňuje seřazení dle uvedené hodnoty. Celý dotaz je složen do dvou dílčích dotazů, které se označují jako subquerry. Níže je uveden dotaz, kterým byl získán výstup detekce robota.

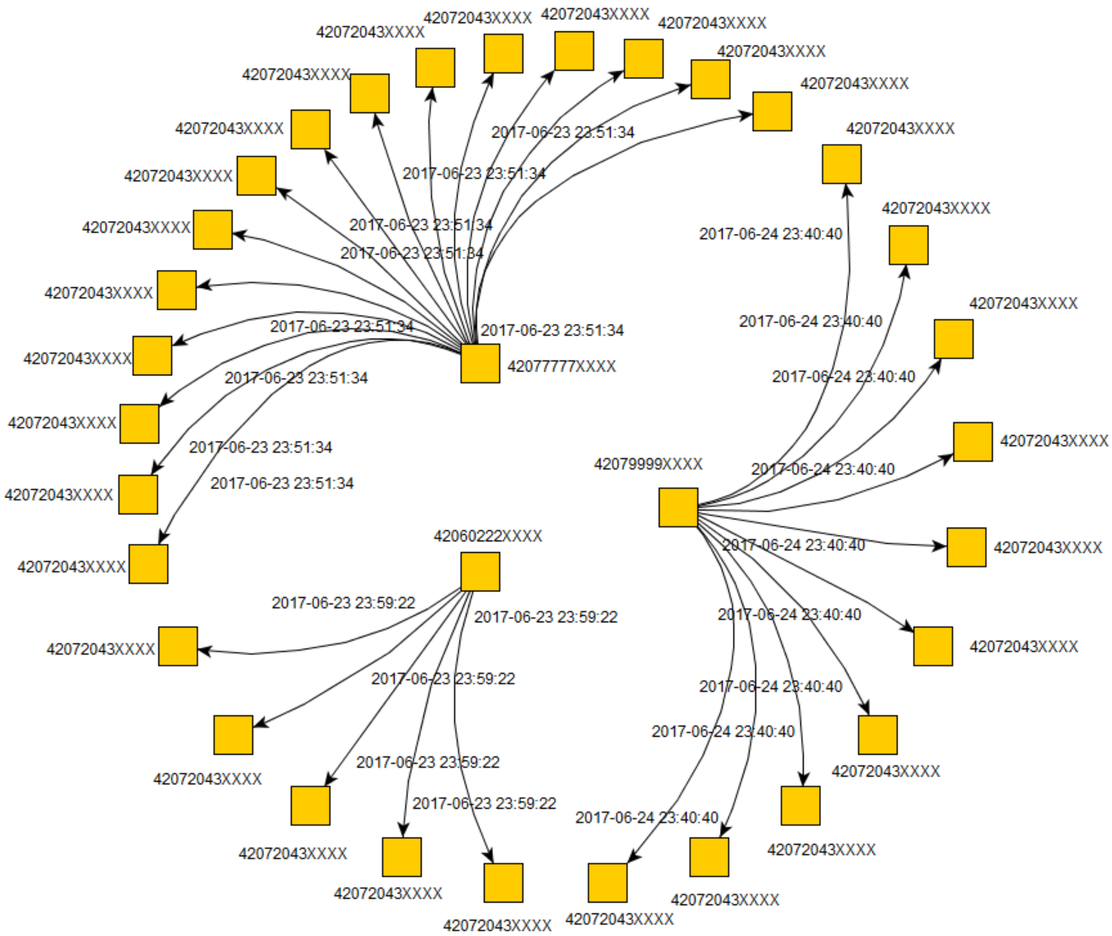
```
hive> insert overwrite local directory '/storage/brno2/home/nguyengo/DIP/
OUTPUT/ROBOT_detection'row format delimited fields terminated by ","
select M_srz,Dn_a,pocet,volana_cisla  from (select Dn_a,count(Dn_b) as
pocet, concat_ws(",",collect_list(Dn_b)) as volana_cisla, M_srz from
cdr_without_routed_numbers group by M_srz,Dn_a) s where pocet > 1 ;
```

První část dílčího dotazu provedla nejprve dotaz na tabulku *CDR_without_routed_numbers* a byly získány hodnoty: volající číslo, počet hovorů, které volající číslo realizovalo, seznam těchto volaných čísel, časové razítko. Získané hodnoty byly seřazeny dle časového razítka volajícího čísla. Nad výsledkem z prvního dílčího dotazu byl proveden druhý dílčí dotaz, který vyfiltroval telefonní čísla realizující více než jeden hovor v stejný časový okamžik. Tato podmínka může být změněna dle potřeby.

Níže je uveden výstup z Hive. Je zde vidět dotaz, doba zpracování v MapReduce, náš očekávaný výstup a celkový čas trvání. Jsou zde příslušné časové okamžiky, kdy z jednoho čísla bylo voláno na více čísel. Dále volající čísla, počet volaných čísel a seznam volaných čísel. Maskování čísel zde bylo provedeno manuálně. Výsledek byl získán z CDR záznamu, který byl vygenerován po třiceti minutách z ústředny.

```
hive> select M_srz, Dn_a,pocet, volana_cisla from (select Dn_a,count(Dn_b) as
pocet, collect_list(Dn_b) as volana_cisla, M_srz from
cdr_without_routed_number group by M_srz,Dn_a) s where pocet >1 ;

Total MapReduce CPU Time Spent: 8 seconds 510 msec
OK

2017-06-23 23:51:34 42077777XXXX 15 ["42072043XXXX","42072043XXXX",
"42072043XXXX","42072043XXXX","42072043XXXX","42072043XXXX","42072043XXXX",
"42072043XXXX","42072043XXXX","42072043XXXX","42072043XXXX","42072043XXXX",
"42072043XXXX","42072043XXXX ","42072043XXXX"]

2017-06-23 23:59:22     42060222XXXX 5 ["42072043XXXX","42072043XXXX",
"42072043XXXX ","42072043XXXX","42072043XXXX"]

2017-06-24 23:40:40     42079999XXXX 10  ["42072043XXXX","42072043XXXX",
"42072043XXXX","42072043XXXX","42072043XXXX","42072043XXXX","42072043XXXX",
"42072043XXXX","42072043XXXX","42072043XXXX"]

Time taken: 29.579 seconds, Fetched: 3 row(s)
```

Na uvedeném Obrázku 3.22 je zobrazen graf detekce robota. Je zde zobrazen přehledně směr komunikace, časová razítka a telefonní čísla volajícího a volaného. Námi zvolená čísla 42077777XXXX, 42060222XXXX, 42079999XXXX jsou ve středu grafu a z nich směřují hrany na telefonní čísla z rozsahu 42072043XXXX. Grafický výstup se shoduje s hodnotami z tabulky 3.7 z čehož vyplývá, že náš návrh detekce je úspěšný. Tento výstup byl získán pomocí grafického editoru yEd s využitím rozložení Radial. Pro přehlednost byla některá časová razítka smazána.



Obr. 3.22: Detekce robota

4 Vyhodnocení

V první části diplomové práce je popsána problematika fraudů, jejich principy a vlastnosti. Dále je zde také zmíněna aktuálnost fraudů v rámci České republiky, krátké seznámení s metodami detekce fraudů a existující komerční řešení.

Další část byla zaměřena na analýzu CDR záznamu na virtuálním stroji. Pro získání grafických výstupů z těchto analýz byly vytvořeny skripty. Je zde také popsáno provedení, jak celou analýzu zautomatizovat.

V poslední části práce byla provedena opět analýza CDR záznamu, ale již na reálném Hadoop klastru. Pracovalo se už s produkčním CDR záznamem, ze kterého byly těženy informace. Nejprve byly provedeny analýzy počtu opakujících se hovorů a počty hovorů mezi ústřednami, stejně jako v předchozí části. Následně byly navrženy postupy na detekci fraudů. Simulaci jednotlivých typů fraudů jsme realizovali modifikací CDR záznamů, na které jsme aplikovali naše detekční návrhy.

Na základě dosažených výsledků jsou námi navržené postupy na detekci fraudů funkční. Jednotlivé detekce se dají modifikovat pro další typy fraudů a různě vylepšovat.

Možné vylepšení u návrhu detekce Wangiri fraudu by bylo definování přesnějších hodnot pro filtrování dat. Na základě praktických zkušeností z dlouhodobého sledování telefonního provozu bychom dokázali lépe stanovit například hraniční hodnotu, která by signalizovala fraud. Tato myšlenka převedená do praktické realizace by znamenala možné vylepšení našeho návrhu. Toto zlepšení by spočívalo ve výběru uzlů, které by měly stupeň vrcholu větší než hraniční hodnota. Získali bychom tak ještě přesnější a přehlednější grafický výstup.

U detekce volání do zón s nejvyšším tarifem je možné rozšířit podmínku pro porovnávání předčíslí. V uvedeném návrhu bylo vybráno pouze pět předčíslí. V praxi bychom mohli zmíněnou podmínku rozšířit o více předčíslí ze zón s nejvyšším tarifem. Pro Service Rate Premium Service Fraud a Call Spamming Fraudu by bylo potřeba získat aktuální seznam prémiových čísel a čísel, ze kterých se provádějí spamové hovory. Na základě takového seznamu by bylo jednoduché jednotlivé fraudy detekovat, případně rovnou blokovat.

Detekce robota je možné vylepšit sledováním okamžiku, kdy byl volaný volán a okamžiku, kdy volaný hovor zvedl. Tento časový interval by nám mohl poskytnout indikaci toho, zda se jedná o robota či ne. Z reálného života víme, že pokud je někdo volán, hovor nezvedne ihned. Člověk nejprve registruje vyzvánění či vibrace, následně vezme telefon do ruky a hovor přijme. Proto usuzujeme, že pro robota je charakteristické to, že hovor přijme ihned. Aby tato detekce byla přesná, bylo by potřeba mít velké sady dat, ze kterých by byly napočítány doby mezi okamžikem, kdy je někdo volán a samotným zvednutím hovoru. Získané hodnoty bychom následně zprůměrovali a získali tak průměrnou hodnotu představující indikace pro konkrétní čísla. Tato hodnota by představovala průměrný čas kdy je hovor uživatelem přijat. Při znalosti takové hodnoty je možné časový interval porovnávat. V případě, že by se jednalo o časový interval mnohem nižší než průměrná hodnota, mohl by to být první signál k podezřelému chování. Ze získaných postupů lze sestavit detekční systém, který lze použít pro detekci fraudů v telefonním provozu.

Na základě výše uvedených tvrzení lze shrnout, že všechny zadané úkoly diplomové práce byly splněny. Výsledkem této práce jsou navržené postupy pro detekci konkrétních typů fraudů v telefonním provozu. Tato práce může také posloužit jako návod pro provádění analýz, různých statistik, například pro potřebu dimenzování sítí operátora.

V budoucí navazující práci bych chtěl pokračovat v tomto tématu. Chtěl bych realizovat detekční postupy využívající například napočítání profilů jednotlivých telefonních čísel, což by se velmi blížilo existujícím komerčním řešením. Dále bych se chtěl pokusit o využití strojového učení k detekci fraudů.

Literatura

[1] Kummer R.: Současné telekomunikační hrozby a boj proti nim [online].2010, dostupé z: http://www.voip-forum.cz/archiv/tk_hrozby_kummer.pdf [cit. 2017-08-20].

[2] Brabec Z.: Řízení telekomunikačních sítí a služeb [online], dostupé z: http://docplayer.cz/1034774-Rizeni-telekomunikacnich-siti-a-sluzeb-zdenek-brabec.html [cit. 2017-08-20].

[3] Implementace HP Fraud Management Systému v T-Mobile Czech Republic a.s. [online].2006, dostupé z: www.cssi.cz/cssi/system/files/all/SI_06_1_HP.pdf [cit. 2017-09-15].

[4] Troller P.: Technické aspekty realizace podvodů v telekomunikačních sítích [online], dostupé z: http://www.xphonet.com/doc/aspekty-podvodu.pdf [cit. 2017-09-10].

[5] Telecom Fraud Guide [online], dostupé z: https://transnexus.com/resources/telecom-industry-topics/telecom-fraud-guide [cit. 2017-09-30].

[6] Roaming není volání do zahraničí [online].2017, dostupé z: https://www.ctu.cz/roaming-definice [cit. 2017-10-03].

[7] Charburský M.: Dokumet ČTÚ-50 591/2017-620/I. vyř. - LhO.2017.

[8] Ghotekar N.: Analysing and Data Ming of Call Detail Records using Big Data Technology [online].2016, dostupé z: https://www.ijarcce.com/upload/2016/december-16/IJARCCE%2064.pdf [cit. 2017-10-12]. 2278-1021.

[9] Sušický M.: Architektura Hadoop Clusteru [online].2017, dostupé z: https://cw.fel.cvut.cz/old/_media/courses/b0m33bdt/b0m33bdt-2p.pdf [cit. 2017-10-25].

[10] Hučín J.: Spark [online].2017, dostupné z: https://cw.fel.cvut.cz/old/_media/courses/b0m33bdt/b0m33bdt-5p.pdf [cit. 2017-11-20].

[11] Rouse M.: Commodity Hardware [online].2013, dostupné z: http://whatis.techtarget.com/definition/commodity-hardware [cit. 2017-11-22].

[12] Dr. Laxmi Lydia E., Dr. Ben Swarup M.: Analysis of Big data trough Hadoop Ecosystem Components like Flume, MapReduce, Pig and Hive [online].2016, dostupné z: http://whatis.techtarget.com/definition/commodity-hardware [cit. 2017-11-25]. 2319-7323.

[13] Harshawarhan Bhosale S., Prof. Devendra Gadekar P.: A Review Paper on Big Data and Hadoop [online].2014, http://www.ijsrp.org/research-paper-1014/ijsrp-p34125.pdf [cit. 2017-11-25]. 2250-3153.

[14] Khoso M.: How Much Data is Produced Every Day? [online].2016, http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/ [cit. 2017-11-27].

[15] Pasčenko P.: Technologie pro velká data [online].2017, https://cw.fel.cvut.cz/old/_media/courses/b0m33bdt/b0m33bdt-1p.pdf [cit. 2017-11-25].

[16] Hadoop - Big Data Overview [online], dostupné z: https://www.tutorialspoint.com/hadoop/hadoop_discussion.htm [cit. 2017-11-30].

[17] Kratochvíl M.: Storage [online].2017, dostupné z: https://cw.fel.cvut.cz/old/_media/courses/b0m33bdt/b0m33bdt-3p.pdf [cit. 2017-11-28].

[18] Larry F.: Co je Apache Hive a HiveQL v Azure HDInsight? [online].2018, dostupné z: https://docs.microsoft.com/cs-cz/azure/hdinsight/hadoop/hdinsight-use-hive [cit. 2018-02-10].

[19] Nosrati M.: Python: An approriate language for real world programming [online].2011, dostupné z: http://waprogramming.com/download.php?download=50ae4a1125d607.12866725.pdf [cit. 2018-01-12]. 2222-2510.

[20] Sanjay Karnewar A.: Designing Python Code for Derivation of Flow Matrix [online].2015, dostupné z: http://ijarcsse.com/Before_August_2017/docs/papers/Special_Issue/ITSD2015/32.pdf [cit. 2017-12-19]. 2277 128X.

[21] Sara Elagib B., Aisha-Hassan Hashim A., Olanrewaju R. F.: CDR Analysis using Big Data Technology [online].2015, dostupné z: https://mafiadoc.com/cdr-analysis-using-big-data-technology-ieee-xplore_597b30161723ddad8e078f7d.html [cit. 2018-01-19]. 978-1-4673-7869-7.

[22] Informace pro účastníky pro volání na telefonní čísla začínající číslicí 9 [online], dostupné z: https://www.ctu.cz/cs/download/ochrana_spotrebitele/ochrana_spotrebitele_informace-ucastnici_cislo-9.pdf [cit. 2018-01-21].

[23] DETEKCE PODVODŮ [online].2016, dostupné z: http://www.adastra.cz/byznys-reseni/fraud-detection [cit. 2018-01-21].

[24] FRADES – detekce podvodného chování [online], dostupné z: http://telpro.cz/?page_id=89 [cit. 2018-01-21].

[25] Bauman B.: Zneužívání služeb v telekomunikacích: Jak se mu bránit, co je třeba znát a vědět [online].2013, dostupné z: https://cfoworld.cz/analyzy/zneuzivani-sluzeb-v-telekomunikacich-jak-se-mu-branit-co-je-treba-znat-a-vedet-2507 [cit. 2018-01-25].

[26] Mach M.: Kladivo na podvodníky aneb jak se odhalují podvody v telekomunikacích [online].2006, dostupné z: https://computerworld.cz/securityworld/kladivo-na-podvodniky-aneb-jak-se-odhaluji-podvody-v-telekomunikacich-46288 [cit. 2018-01-29].

[27] BORKAR N.: There's A New 'SIM Card Cloning ' Scam In Town, And Here's Why You Need To Be Careful [online].2016, dostupné z: https://www.indiatimes.com/culture/who-we-are/there-s-a-new-sim-card-cloning-scam-in-town-and-here-s-why-you-need-to-be-careful-256536.html [cit. 2018-02-03].

[28] Jak fungují podvody s Premium SMS a jak se jim bránit [online].2017, dostupné z: http://www.apms.cz/aktuality-a-zajimavosti/jak-funguji-podvody-s-premium-sms-a-jak-se-jim-branit [cit. 2018-02-10].

[29] The definitive guide to fraud types [online].2013, dostupné z: http://www.capacityconferences.com/assets/fraud/fraud%20types.pdf [cit. 2018-02-10].

[30] Christie S.: SCAM WARNING Fraudsters are now hijacking phones and diverting texts and calls to their mobiles to steal YOUR money [online].2017, dostupné z:

https://www.thesun.co.uk/money/3416363/fraudsters-are-now-hijacking-phones-and-diverting-texts-and-calls-to-their-mobiles-to-steal-your-money/ [cit. 2018-02-15].

[31] Toll Fraud [online].2013, dostupné z: https://support.zoho.com/portal/DocsDisplay?attachId=32e9dbbff29cfce38fd7a59afe0a20b6e612657a51fd5fbe&action=download&zgId=0224ca4b5d0b8863dc91a236738b3d70&entityId=32e9dbbff29cfce3155681d15a6c26c0e612657a51fd5fbe&portalId=0224ca4b5d0b8863e13a4987f38626734fc7cadb3e9388d8afb6719abeecaa48 [cit. 2018-02-18].

[32] Y. Kou, Chang-Tien L., S. Sinvongwattana,Yo-Ping H.: Survey of Fraud Detection Technique [online].2004, dostupné z: http://ieeexplore.ieee.org/document/1297040/ [cit. 2018-02-24]. 0-7803-8193-9.

[33] Grund J.: Levné telefony volaly potají do ciziny. Operátoři varují před neprověřenými přístroji z volné distribuce. [online].2018, dostupné z: http://www.apms.cz/tiskove-zpravy-a-stanoviska/levne-telefony-volaly-potaji-do-ciziny-operatori-varuji-pred-neproverenymi-pristroji-z-volne-distribuce [cit. 2018-03-28].

[34] PRAJAPATI V.: HOW TO INSTALL APACHE HADOOP 2.6.0 IN UBUNTU (SINGLE NODE SETUP) [online].2015, dostupné z: http://pingax.com/install-hadoop2-6-0-on-ubuntu/ [cit. 2017-11-28].

[35] Awanish: Apache Hive Installation on Ubuntu, https://www.edureka.co/blog/apache-hive-installation-on-ubuntu [online].2014, dostupné z: https://www.edureka.co/blog/apache-hive-installation-on-ubuntu [cit. 2017-11-28].

[36] Singh J., Ruhl R., Lindskog D.: GSM OTA SIM Cloning Attack and Cloning Resistance in EAP-SIM and USIM [online].2013, dostupné z: http://ieeexplore.ieee.org/document/6693458/ [cit. 2018-03-29]. 978-0-7695-5137-1.

[37] Pycharm [online].2017, dostupné z: https://www.jetbrains.com/pycharm/ [cit. 2017-12-19].

[38] Shantanu S.: Installing Hadoop-2.6.x on Windows 10 [online], dostupné z: http://www.ics.uci.edu/~shantas/Install_Hadoop-2.6.0_on_Windows10.pdf [cit. 2017-11-26].

[39] https://gist.github.com/nguyengo1/c2afc7ebabf758fe584ea1e09859b57d

[40] Odkaz na stažení Hive, dostupný z: ftp://mirror.hosting90.cz/apache/hive/ [cit. 2017-10-21].

[41] Odkaz na stažení Hadoop, dostupný z: ftp://mirror.hosting90.cz/apache/hadoop/common/ [cit. 2017-10-21].

[42] Hadoop [online].2017, dostupné z: https://wiki.metacentrum.cz/wiki/Hadoop [cit. 2018-02-14].

[43] Odkaz na vyzkoušení Zeppelin demo, dostupný z: http://demo.insightedge.io/#/ [cit. 2018-01-28].

[44] BRAS L.: Digital transformation means fraud transformation [online].2017, dostupné z: https://inform.tmforum.org/features-and-analysis/2017/02/digital-transformation-means-fraud-transformation/#prettyPhoto [cit. 2018-03-04].

[45] Mezinárodní volání - Rozdělení zemí do jednotlivých zón [online].2017, dostupné z: https://www.o2.cz/_pub/ac/da/6b/240291_1061003_Mezinarodni_volani___Rozdeleni_zemi_do_jednotlivych_zon.pdf "[cit. 2018-03-21]"

[46] Snášel J.: Polopatě jak telefonovat s několika lidmi najednou [online].2004, dostupné z: https://www.mobilmania.cz/clanky/polopate-jak-telefonovat-s-nekolika-lidmi-najednou/sc-3-a-1108728/ [cit. 2018-03-21].

[47] Fraud in the telecommunications industry [online].2014, dostupné z: http://smartipx.com/fraud-in-the-telecommunications-industry-part-1/ [cit. 2018-03-24].

[48] VEPŘOVSKÁ J.: Dejte si pozor na aplikaci, která vás napálí [online].2018, dostupné z: https://decinsky.denik.cz/zpravy_region/dejte-si-pozor-na-aplikaci-ktera-vas-napali-20180320.html [cit. 2018-03-24].

[49] Odkaz na stažení aplikace JokesPhone, dostupné zde: https://play.google.com/store/apps/details?id=com.cashitapp.app.jokesphone&hl=cs

[50] Ceník mezinárodních hovorů frimy O2, dostupné zde: https://www.o2.cz/osobni/roaming/118622-mezinarodni_hovory_z_mobilu_cenik.html [cit. 2017-03-22]

[51] Dr. Howells I., Dr. Scharf-Katz V., Stapleton P.: TELECOM FRAUD 101: Fraud Types, Fraud Methods,& Fraud Technology [online]. dostupnéz: https://www.scribd.com/document/311777864/Telecom-Fraud-101-eBook [cit. 2017-10-23].