**ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE**

**Fakulta dopravní**
**Ústav letecké dopravy**

# Využití potenciálu Big Data analýzy v letecké dopravě

# Leveraging the Potential of Big Data Analysis in Aviation

Diplomová práce

Studijní program: Technika a technologie v dopravě a spojích
Studijní obor: Provoz a řízení letecké dopravy

Vedoucí práce: Ing. Eva Endrizalová, Ph.D.

**Bc. Rostislav Pšovský**

**Praha 2015**

**ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE**
**Fakulta dopravní**
**d ě k a n**
Konviktská 20, 110 00 Praha 1

**K621** ..... **Ústav letecké dopravy**

# ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení studenta (včetně titulů):

**Bc. Rostislav Pšovský**

Kód studijního programu a studijní obor studenta:

**N 3710 – PL – Provoz a řízení letecké dopravy**

Název tématu (česky): **Využití potenciálu Big Data analýzy v letecké dopravě**

Název tématu (anglicky): Leveraging the Potential of Big Data Analysis in Aviation

## Zásady pro vypracování

Při zpracování diplomové práce se řiďte osnovou uvedenou v následujících bodech:

- Úvod
- Historický vývoj množství digitálních dat
- Charakteristiky Big Data a jejich zpracování
- Technologická platforma pro Big Data
- Příklady využití Big Data v letecké dopravě
- Podpora rozhodovacího procesu na provozním dispečinku společnosti Travel Service pomocí Big Data
- Závěr

Rozsah grafických prací: dle pokynů vedoucího diplomové práce

Rozsah průvodní zprávy: minimálně 55 stran textu (včetně obrázků, grafů a tabulek, které jsou součástí průvodní zprávy)

Seznam odborné literatury:
SCHMARZO, Bill. Big data: understanding how data powers big business

DUNNING, Ted a B FRIEDMAN. Practical machine learning: a new look at anomaly detection

MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. Big data: a revolution that will transform how we live, work, and think

Vedoucí diplomové práce: **Ing. Eva Endrizalová, Ph.D.**

**Ing. Vladimír Fajt**

Datum zadání diplomové práce: **31. července 2014**
(datum prvního zadání této práce, které musí být nejpozději 10 měsíců před datem prvního předpokládaného odevzdání této práce vyplývajícího ze standardní doby studia)

Datum odevzdání diplomové práce: **30. listopadu 2015**
a) datum prvního předpokládaného odevzdání práce vyplývající ze standardní doby studia a z doporučeného časového plánu studia
b) v případě odkladu odevzdání práce následující datum odevzdání práce vyplývající z doporučeného časového plánu studia

L. S.

doc. Ing. Stanislav Szabo, PhD. MBA
vedoucí
Ústavu letecké dopravy

prof. Dr. Ing. Miroslav Svítek, dr. h. c.
děkan fakulty

Potvrzuji převzetí zadání diplomové práce.

Bc. Rostislav Pšovský
jméno a podpis studenta

V Praze dne 30. června 2015

## Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Nemám závažný důvod proti užívání tohoto školního díla ve smyslu § 60 Zákona č.121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V Praze 28.11.2015 Podpis…….

# Abstrakt

**Autor:** Bc. Rostislav Pšovský

**Název práce:** Využití potenciálu Big Data analýzy v letecké dopravě

**Rok vydání:** 2015

**Klíčová slova:** Big Data, informační revoluce, růst dat, mobilní revoluce, internet věcí, Big Data technologie, Big Data analýza, korelace, HDFS, NoSQL, MapReduce, personalizovaný marketing, profilování zákazníka, prediktivní údržba, Travel Service, provozní dispečink, podpora rozhodování

Tato diplomová práce se zaměřuje na poměrně novou a progresivní oblast analýzy velkých datových sad za účelem objevení skrytých vzorců a spojitostí k podpoře obchodních modelů leteckých společností. První část opodstatňuje nutnost využití Big Data analýzy rozborem bezprecedentního nárůstu digitálních dat v průběhu předešlých dekád. Autor také popisuje samotný fenomén skrývající se pod souslovím Big Data, výhody oproti tradičním metodám inferenční statistiky a technologickou náročnost analýzy velkých datových sad. V druhé části práce jsou rozebírány možnosti využití Big Data analýzy v odvětví letecké dopravy. Detailněji jsou popsány 3 různé aplikace: Profilování zákazníků leteckých společností na základě nových druhů dat a poskytování individuálních nabídek, zdokonalení prediktivní údržby letadel díky detekci anomálního chování ze senzorových dat a podpora rozhodovacího procesu na provozním dispečinku letecké společnosti Travel Service, a.s.

# Abstract

**Author:** Bc. Rostislav Pšovský

**Title:** Leveraging the Potential of Big Data Analysis in Aviation

**Publication year:** 2015

**Keywords:** Big Data, Information revolution, Data growth, Mobile revolution, Internet of Things, Big Data technologies, Big Data analysis, correlation, HDFS, NoSQL, MapReduce, Personalized marketing, Customer profiling, Predictive maintenance, Travel Service, Operations Control Center, Decision making support

This diploma thesis is focused on the topic of a relatively new and progressive area of large data sets analysis with the aim to discover hidden patterns and connections in order to support airlines‘ business models. The first part of the paper justifies the need of using Big Data analysis by describing the unprecedented growth of digital data during last decades. The author also describes the phenomenon of Big Data itself, its advantages over traditional inferential statistics methods and technology demands of large data sets analysis. The second part analyses the possible use of Big Data analysis in the aviation industry. Three different applications are described in detail: Airlines‘ customers profiling using new data sources and providing them with individual offers, improvement of aircraft predictive maintenance through sensor data anomalous behaviour detection and decision making support at operations control center of the company Travel Service, a.s.

## Poděkování

Rád bych touto cestou poděkoval vedoucí mé práce, paní Ing. Evě Endrizalové, Ph.D., za nasměrování při volbě tématu, vytvoření poklidné pracovní atmosféry a věcné připomínky vedoucí k úspěšnému dokončení této práce.

## Obsah

## Předmluva

Big Data analýza představuje nové, aktuální a progresivní téma. Vzhledem k tomu, že k prudkému rozvoji v oblasti velkých datových sad došlo teprve před několika málo lety, nemá zatím sousloví Big Data všeobecně zažitý český překlad, a většina českých autorů se omezuje na využívání původního anglického termínu. Nakladatelství Computer Press v českém překladu knihy Big Data [1] navrhuje český termín „veledata“, který však považuji za příliš kostrbatý a navíc nevhodný pro účely vyhledávání informací o oblasti Big Data analýzy. Z tohoto důvodu si dovoluji v této práci využívat původní sousloví Big Data a pro účel snadnější četby skloňovat slovo „Data“ dle pravidel českého pravopisu.

# Úvod

Žijeme uprostřed informační revoluce. Bezprecedentní rozmach informačních technologií má za následek generování masivních objemů dat v oblasti téměř jakékoliv lidské činnosti. Tyto masivní datové sady, které se z řádu gigabytů před více než dvaceti lety postupně vyvinuly do řádu petabytů, a v dnešní době díky rozmachu mobilních technologií a vzniku dat ze sociálních sítí až do řádu exabytů, se velmi často označují obecným souslovím „Big Data“.

Kolem konceptu Big Dat se v poslední době napříč všemi odvětvími rozmohla vášnivá diskuse. Neustálé zvyšování výpočetního výkonu počítačů a snížení nákladů na skladování dat položilo základy novým analytickým možnostem, které by mohly vyústit v novou „průmyslovou revoluci“ založenou na datech. Dle některých společností jsou data dokonce jednou z nejcennějších komodit 20. století. Síla Big Dat leží v nově nabyté schopnosti odpovídat na otázky, na které jsme do dnešní doby nebyli schopní odpovědět. Nabízí totiž alternativu ke klasickým statistickým metodám, které jsou založené na vyvozování závěrů o velké datové sadě na základě z ní pečlivě vybraného vzorku. Big Data analýza naproti tomu znamená analýzu kompletní datové sady a odhalování skrytých vzorců a souvislostí, které dříve nebylo možné kvůli nedostatku relevantních dat a výpočetního výkonu odhalit.

Letecká doprava stojí na předních místech mezi odvětvími, která mají na dosah vytěžení potenciálu Big Dat. Podnikání v oblasti letecké dopravy vždy silně spoléhalo na dostatek dat a všechny letecké společnosti tak dnes disponují masivním množstvím historických dat. Tato data jsou skladována v obrovském množstvím vzájemně nekomunikujících úložišť, která je potřeba integrovat a na základě Big Data analýzy rozpoznat skryté souvislosti, jejichž pochopení bude mít významný dopad na podporu obchodních strategií leteckých společnosti. Odvětví letecké dopravy se vyznačuje nízkými maržemi, což nutí společnosti, které v něm podnikají, neustále hledat způsob, jak se odlišit od svých konkurentů a strhnout tak poptávku na svou stranu. Implementace Big Data analýzy do firemních procesů může znamenat právě onu hledanou konkurenční výhodu. Je dokonce možné, že se do budoucna stane standardem, a neschopnost reagovat na vhledy vyvozené z analýzy velkých datových sad pro mnoho společností vyústí v nekonkurenceschopnost.

Koncept Big Dat není jednoduché pochopit. Manažeři mnohých společností sice cítí, že Big Data analýzu je potřeba urychleně zapracovat do obchodní strategie, ale často bojují s představou, jak může reálně obohatit firemní procesy. V této práci si dávám za cíl pojednat o technologických nárocích na Big Data infrastrukturu a její implementaci do současných systémů leteckých společností, ale hlavně o nástin možností využití Big Dat leteckými společnostmi, včetně podpory prodeje, údržby a rozhodovacích procesů.

# 1 Exploze dat

Okolo roku 3000 př. n. l. se v tehdejších raných civilizacích vyvinulo první písmo. Představovalo tlustou čáru mezi primitivními a pokročilejšími společnostmi. Podporovalo totiž odvěkou snahu lidí analyzovat okolní svět tím, že umožňovalo informace o něm zaznamenávat [1]. Jednalo se o první velkou informační revoluci, která položila základy vzniku vůbec prvních dat. Celkové množství uložených informací však bylo značně omezeno intelektuálními a technologickými možnostmi té doby. Základní předpoklad vzniku záznamu totiž představovala jednak znalost samotného písma a také dispozice záznamového média – hliněné destičky či papyru. Vznik písma ale beze sporu podnítil rychlý intelektuální rozvoj lidské společnosti a rozmach obchodu.

Z dnešního pohledu je téměř nepochopitelné, že forma ručního záznamu informací byla používána v takřka nezměněné podobě po několik tisíc let a rychlost reprodukce informací byla přímo závislá na zdatnosti písaře. Manipulace s informacemi byla až do začátku 15. století doménou katolické církve, která knihy schraňovala mezi klášterními zdmi a přístup k nim měla zpravidla jenom hrstka vyvolených. Vše se změnilo okolo roku 1439 s vynálezem Gutenbergova tiskařského lisu.

Tiskařský lis naprosto změnil podmínky, za kterých byly informace sbírány, ukládány, vyhledávány, posuzovány, objevovány a propagovány. Možnost rapidního šíření informací mezi široké vrstvy obyvatelstva měla okamžitý kauzativní efekt na reformaci, renesanci a vědeckou revoluci. Vznikla tak kulturní exploze zapříčiněná hlavně do té doby nebývalým množstvím slov, obrázků a diagramů, které naráz učenci měli ke své dispozici.

Reformace by patrně nikdy nemohla proběhnout, kdyby se Martin Luther nedostal k tehdejším výtiskům biblických textů a znepokojujícím variacím různého výkladu základních dogmat. Jenom díky tomu mohl Luther vyvinout svou protestantskou doktrínu, revoluční teologii, která za začala rychle šířit a vyústila v protestantskou reformaci. Italská renesance sice započala již před tiskařským lisem, jeho vynálezem však byla výrazně ovlivněna opět hlavně díky možnosti šíření myšlenek do nejzazších koutů země. Poslední z tří hlavních událostí „období tiskařského lisu“, vědecká revoluce, spočívala především v nově vzniklé ochotě přehodnocovat vědecké závěry starověkých učenců a v nově získaném sebevědomí zvažovat nové myšlenkové koncepty. Kultura ručního přepisování považovala původní učení za to nejčistší, poněvadž nebylo zatíženo chybami způsobenými ručním přepisováním a leželo z časového hlediska nejblíže počátku všech věcí. Kultura tisku znamenala změnu ke kumulativnímu vývoji vědomostí, čímž nám dovolila hodnotit výsledky minulého bádání s patřičným odstupem. Díky tisku se také zrodil vědecký sběr dat. V roce 1543 Koperník vydal své dílo *De Revolutionibus Orbitum Coelestium.* Porovnal v něm

myšlenky Ptolemaia, Aristotela a dalších a povšiml si výrazných chyb a nekonzistencí v jejich dílech, čímž započal vědeckou revoluci [2].

Vynález tiskařského lisu měl výrazný dopad na vývoj společnosti, která se nacházela v přerodu mezi středověkem a novověkem. Přinesl více dostupných informací, více názorů, informovanější diskusi a ostřejší kritiku autorit [3]. Bez jakékoliv pochyby v následujících stoletích vyvolal jednu z největších informačních revolucí v historii lidstva [2]. Od založení Konstantinopole v roce 330 n. l. do vynálezu knihtisku, tedy během 1200 let, bylo ručním přepisováním vyprodukováno zhruba 8 milionů knih. Stejné množství bylo vytištěno na tiskařském lisu jen mezi lety 1453 a 1503, tedy v horizontu pouhých 50 let. Můžeme tedy konstatovat, že během těchto 50 let vzrostl objem informací v Evropě na dvojnásobek [1].

V současnosti se nacházíme uprostřed velmi podobné informační revoluce, která má svůj prazáklad v prudkém technologickém pokroku vycházejícím z druhé světové války. Ve snaze prolomit kód Enigmy vyvinul v Bletchley britský matematik, logik a kryptoanalytik Alan Turing přístroj, později prozaicky nazvaný Turingův stroj, ze kterého se postupně vyvinuly dnešní počítače [4]. Samotná revoluce pak začala v roce 1962, kdy Paul Baran představil koncept přepojování paketů jako způsob udržení propojení vojenské velitelské a řídicí sítě v případě nukleárního útoku [2]. V roce 1969 tak vznikla vůbec první vojenská výzkumná počítačová síť nazvaná ARPANET, skládající se ze 4 počítačů zesíťovaných na univerzitní půdě [5]. Tím započal exponenciální růst počítačových sítí. Rychlé šíření bylo možné zejména díky tomu, že pokud uživatel vlastní počítač, připojení do sítě představuje pouze minimální, pokud vůbec nějaký náklad, neboť každý z uživatelů sám nese financování a rozhoduje o svých nárocích na technologickou stránku. V roce 1983 došlo k vydělení nevojenské části ARPANETU a ta se o rok později, kdy propojila prvních 1000 univerzitních a korporátních počítačů, přejmenovala na INTERNET [5]. Činnost samotného ARPANETU byla ukončena v roce 1989 [2].

S rychlým pokrokem v oblasti počítačů a počítačových sítí došlo také k exponenciálnímu růstu počítačových dat, tedy informací zpracovávaných a ukládaných počítačem. Formát těchto dat je binární, na základní úrovni se tedy jedná o soubor jedniček a nul. Díky tomu je možné tato data vytvářet, zpracovávat a ukládat digitálně. Jejich nespornou výhodou je snadný přenos v počítačových sítích a také to, že postupem času nebo opakovaným použitím neztrácí svou kvalitu [6].

Firmy, které v 70., 80. a 90. letech začaly jako první těžit z potenciálu počítačů, se zaměřovaly především na transakční data, která pro ně v té době představovala největší hodnotu. Poprvé v historii tak bylo možné detailně sledovat kdo, kdy, kde a co kupuje [7][8]. Pro účel skladování těchto dat byly vyvinuty databáze, které měly zajistit takový způsob uložení informací, který umožnil jejich snadné opětovné vyvolání. Obchodní společnosti využívaly relační databáze, které reprezentují informace

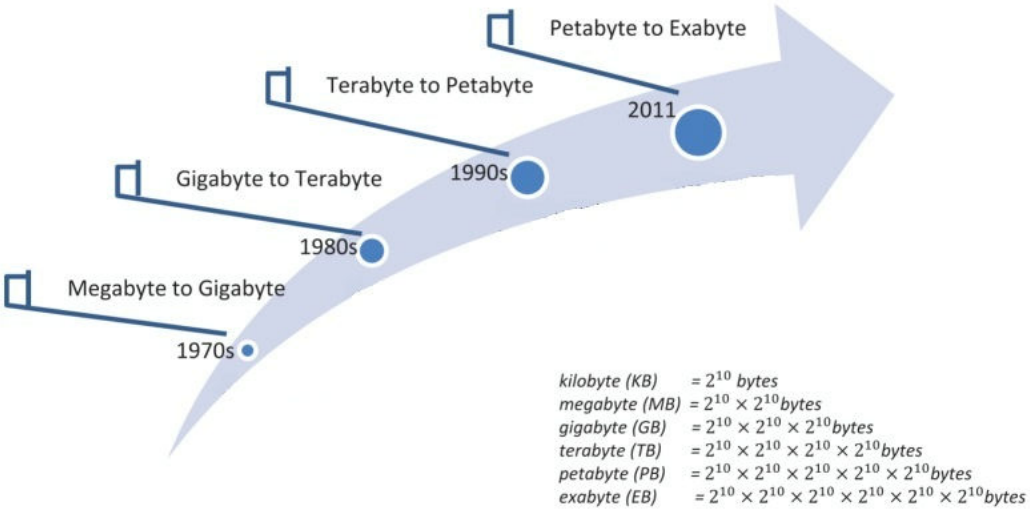
ve formě řádků a sloupců. Vzniká tedy tabulka, kde spolu data souvisí na základě společných klíčů nebo konceptů, a která splňuje podmínku schopnosti data z této tabulky zobrazit pomocí dotazování [9]. Již v 80. letech množství transakčních dat narostlo do řádu gigabytů, což představovalo obrovské nároky na tehdejší datová úložiště [10], protože standartní technologii tehdejší doby představovala magnetická páska s kapacitou 225KB, 5,25 palcová disketa s kapacitou 1,2MB a pevný disk s kapacitou 5MB. V roce 1982 byl představen pevný disk s výrazným navýšením kapacity na 1GB a o dva roky později, v roce 1985, spatřil světlo světa první CD-ROM s kapacitou 700MB [11].

Ke konci 80. let popularizace digitálních technologií vyhnala objemy dat do řádu terabytů. Takové množství dat přesahuje úložné a procesní možnosti jakéhokoliv samostatného počítačového systému. Z důvodu rozšíření možností úložišť tak vznikl systém paralelních výpočtů spočívající v distribuci dat a s nimi souvisejících úloh mezi větší množství rozličného hardwaru. Tato myšlenka vyžadovala nový typ paralelních databází, z nichž se nejvíce osvědčila tzv. architektura „shared-nothing“, ve které spolu jednotlivé počítače nesdílí žádný společný úložný prostor ani výpočetní výkon a jsou tak na sobě naprosto nezávislé [10].

Radost z vývoje paralelních databází však neměla trvat příliš dlouho. V roce 1989 tehdejší zaměstnanec Evropské organizace pro jaderný výzkum CERN, Tim Berners-Lee, vyvinul World Wide Web (WWW), původně určený k výměně informací mezi vědci, univerzitami a dalšími institucemi po celém světě. V dubnu roku 1993 zveřejnil první verzi zdrojového kódu HTML a započal tak rapidní rozvoj Webu verze 1.0 [12]. Svět byl tímto uvržen do éry internetu. Ve stejném roce se počet internetových stránek pohyboval okolo 130. V roce 1996 jich na internetu bylo k nalezení na 100 tisíc [5] a v průběhu 90. let množství internetových dat dosáhlo řádu petabytů. Rozvoj webu s sebou však nesl také jeden velký nešvar – drtivá většina dat nacházejících se na internetu je strukturovaná pouze částečně, popřípadě vůbec. Komunita okolo relačních databází byla uvržena do chaosu. Relační databáze skvěle zvládají organizaci strukturovaných dat, která zařazují do tabulek dle předdefinovaného formátování a příslušného klíče. Neexistuje však způsob, jak jejich pomocí skladovat nestrukturovaná data, pod kterými si můžeme představit např. e-mailové zprávy, které se v 90. letech staly běžnou záležitostí. Volný text nemá jasně definovanou strukturu a není ho tak možné rozkouskovat a roztřídit do jednotlivých políček tabulky. Exploze nestrukturovaných a částečně strukturovaných dat si vyžádala vývoj nových, tzv. NoSQL databází, které umožňují ukládat data bez nutnosti předdefinování formátu. Stejně tak vnikly i nové nástroje umožňující jejich analýzu [10]. Technologické základně pro zpracování velkých objemů dat bude věnována samostatná kapitola.

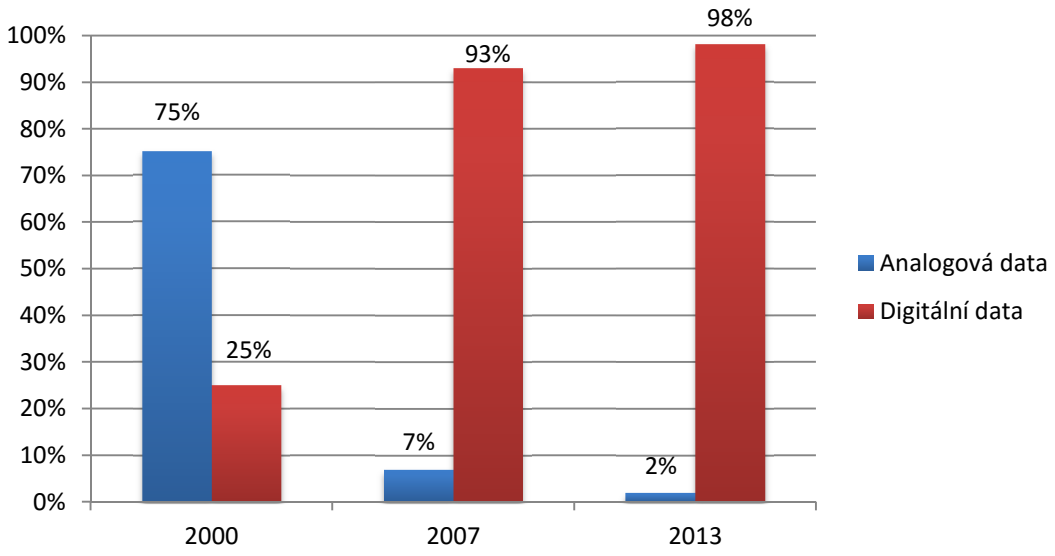
Ve světle současného vývoje se obchodní společnosti zanedlouho budou muset potýkat s množstvím dat v řádech exabytů a toto množství dat bude klást nové nároky

na infrastrukturu, neboť současná technologie je dobře uzpůsobena ke skladování a zpracovávání terabytů, popř. petabytů dat [10]. Na předních frontách této datové záplavy stojí internetové firmy, jako např. Google, který v současné době zpracovává denně více než 24 petabytů dat [1]. V době psaní této práce, tedy na začátku října roku 2015, se počet aktivních unikátních webových stránek pohybuje okolo 937 milionů [13]. Jejich množství tedy od představení jazyka HTML v roce 1993 vrostlo více než 7 milionkrát. Dynamika nárůstu množství digitálních dat je zachycena na obrázku 1.



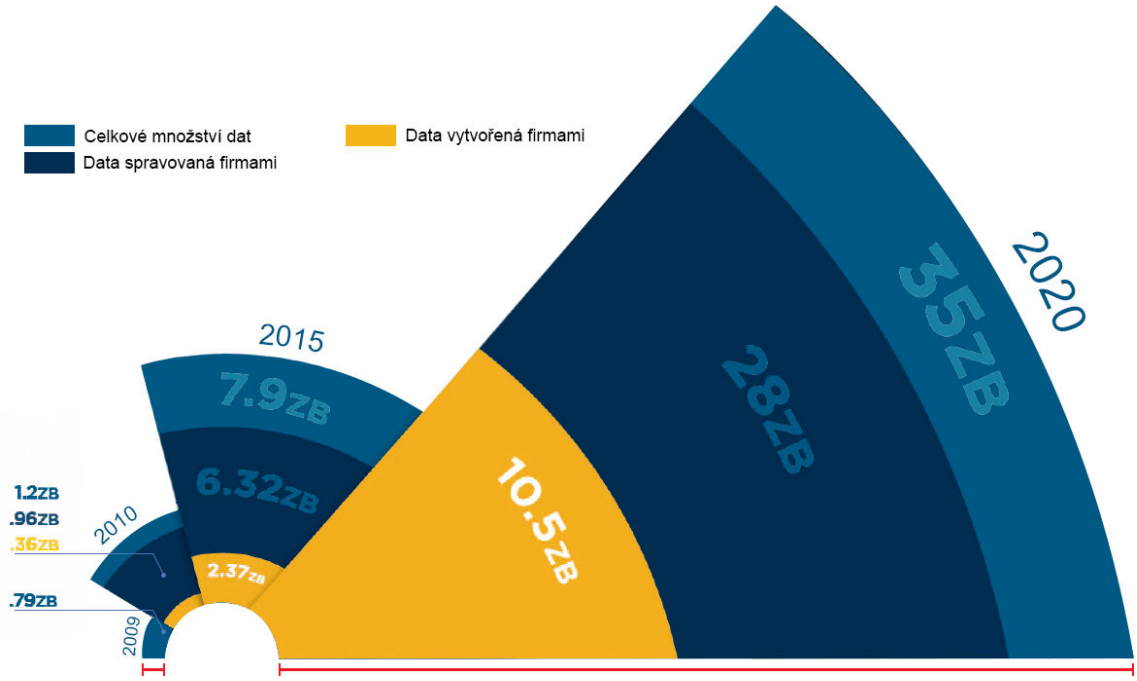

Obr. 1 *Dynamika vývoje množství digitálních dat* [10]

V nedávné době vzniklo hned několik projektů, které si daly za cíl vyčíslit množství veškerých informací, které nás obklopují. Jeden z nich byl proveden Martinem Hilbertem na University of South Carolina. Tato studie se snažila pojmout vše, co kdy bylo vytvořeno, uloženo nebo sděleno ať už v digitální nebo analogové formě. Mezi vyčíslená data tedy zahrnula nejen e-maily, obrazy, knihy, hudbu, videa a fotografie, ale také videohry, záznamy telefonních hovorů, GPS navigaci, poštovní dopisy a televizní a rozhlasové vysílání. V závěru této studie bylo vyčísleno, že v roce 2007 na světě dle odhadu existovalo na 300 exabytů uložených dat. Za pozornost jistě stojí fakt, že v roce 2000 existovalo pouze 25 % světových informací v digitální podobě. Z výše zmíněných 300 exabytů dat v roce 2007 jich naopak už pouze 7 % bylo zaznamenáno v analogové formě. Dá se tedy říct, že někdy v průběhu těchto krátkých 7 let proběhla digitální revoluce. Objem digitálních dat se zdvojnásobuje zhruba každé 3 roky, zatím co množství analogových informací stagnuje. V roce 2013 se množství uložených světových informací odhadovalo na přibližně 1200 exabytů, přičemž ty analogové zaujímaly méně než 2 % z celkového množství.




Obr. 2 *Vývoj poměru množství digitálních a analogových dat*

Dynamiku vývoje množství světových dat nám může přiblížit také přirovnání, které říká, že množství uložených informací roste 4 krát rychleji než světová ekonomika. Paralelně s tím se zvyšuje i výpočetní výkon počítačů, který roste 9 krát rychleji [1].




Obr. 3 *Rapidní růst globálních digitálních dat do řádu zettabytů* (ZB)
(Zdroj: http://assets1.csc.com/insights/downloads/CSC_Infographic_Big_Data.pdf?ref=dbc)

Nejmodernějším řešením pro ukládání dat v řádech exabytů představuje Cloud Computing. Tato technologie byla poprvé představena jako Amazon AWS (neboli

Amazon Web Services) v roce 2006. Pro obchodní společnosti to znamená, že samy nemusí vlastnit infrastrukturu pro ukládání dat. Pomocí služby Cloud Computing jsou data uložena ve vzdáleném datovém centru, neboli „datové farmě“, a uživateli jsou zpřístupněna přes internet. Kromě toho, že uživatel nemusí investovat do infrastruktury, patří mezi výhody takových úložišť přístup z jakéhokoliv zařízení připojeného k internetu, placení pouze za přesně tak velký úložný prostor, který aktuálně využívá, a několikanásobná záloha dat. Jelikož data neukládáme v úložišti, které fyzicky vlastníme, přirovnává se přistupování k těmto datům jako jejich stahování z jakéhosi imaginárního obláčku – proto Cloud Computing. Mezi aplikace přístupné běžným uživatelům internetu a založené na této technologii patří například e-mailový klient. E-maily můžeme číst z jakéhokoliv zařízení připojeného k internetu a přitom je fyzicky nestahujeme do vlastního zařízení. Dalšími cloudovými službami jsou např. komerčně známé aplikace Google Drive, DropBox nebo iCloud. Na stejné bázi funguje také nahrávání jakéhokoliv obsahu na sociální sítě – fotografie nahrané na Facebook putují do vzdáleného úložiště této společnosti a pro samotného uživatele „žijí“ v již zmiňovaném imaginárním obláčku [14][15].

Snaha domyslet, jaký dopad bude mít rapidní nárůst informací okolo nás na život jednotlivců a rozvoj společnosti, se podobá věštění z křišťálové koule. Hrubou představu si však můžeme utvořit porovnáním současné situace s předešlou informační revolucí, tedy obdobím následujícím bezprostředně po vynálezu Gutenbergova knihtisku. Tiskařský lis byl prvním skutečným komunikačním médiem typu 1:m, kdy informace poskytované jedním vysílačem mohou být nezávisle přijímány několika přijímači. V informačním věku dnešní doby jsou tyto komunikační schopnosti zastoupené počítačovými sítěmi, které tento věk definují. Díky těmto vzniknuvším paralelám mezi obdobím tiskařského lisu a dnešním informačním věkem můžeme předpokládat následující vývoj:

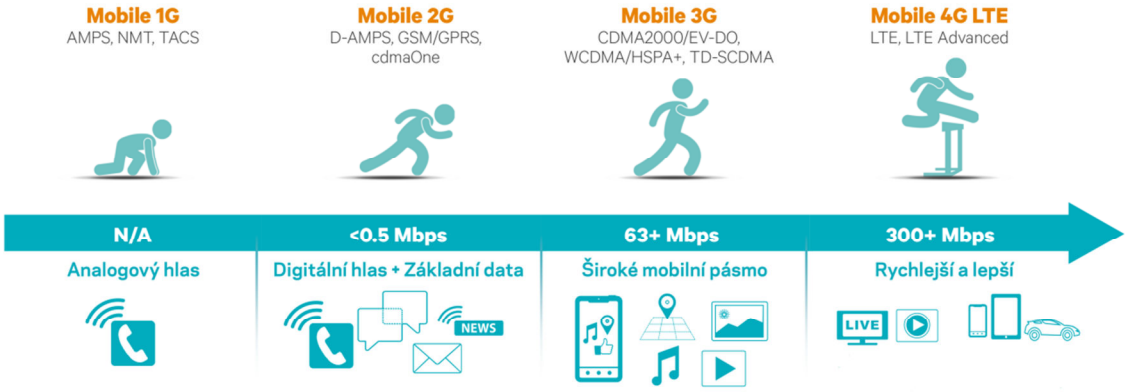
- **Společenské změny tohoto informačního věku budou stejně dramatické jako ty ve středověké Evropě**
- **Budoucnosti informačního věku budou dominovat neočekávané důsledky**. I v této době musíme očekávat nepřímé důsledky vyplývající z rapidního nárůstu přístupných informací, stejně jako byl kdysi přechod k heliocentrické soustavě nepřímým důsledkem vynálezu knihtisku.
- **Plné důsledky informačního věku se projeví za několik desítek let.** Plný dopad tiskařského lisu na společnost se projevil v průběhu následujících 100 let. Z důvodu mnohem rychlejšího tempa nárůstu informací v dnešní době můžeme očekávat zkrácení této doby na několik desítek let.

Éra internetu je v důležitých aspektech velmi podobná éře tiskařského lisu. Jelikož měl tiskařský lis hluboký dopad na své období, můžeme očekávat podobný dopad internetu na současný informační věk. Obě tyto technologie představují průlom v tom, jak spolu

lidé komunikují, jak uchovávají, aktualizují a rozšiřují vědomosti, ale také v definování vlastnictví vědomostí a jejich získávání. Nezřízený vývoj a rozmach internetu vyvolává znepokojující otázku ohledně nutnosti jeho regulace. Zkušenosti z období éry tiskařského lisu však jasně ukazují, že společnosti, které striktně potlačovaly důsledky tiskařského lisu, skončily v úpadku. Dokonce i ty Evropské země, které se snažily potlačit jen jeho negativní důsledky, tímto rozhodnutím trpěly. To silně naznačuje, že jakákoliv jeho negativa byla mnohonásobně převážena pozitivy. Z historického hlediska je tedy nejlepší cestou ponechat internet neregulovaný [2].

## 1.1 Úsvit sociálních dat

Minulou dekádu můžeme bez nadsázky nazvat desetiletím mobilní revoluce, která byla odstartována umožněním komerčního využití mobilních sítí třetí generace v roce 2003. Mobilní 3G technologie se od předchozích generací mobilních sítí lišila především tím, že veřejnosti umožnila využívat širokopásmové připojení, které ve svém důsledku znamenalo možnost připojit se k internetu z jakéhokoliv mobilního zařízení podporující tuto technologii a nacházející se v oblasti pokryté dostatečně silným mobilním signálem [16].

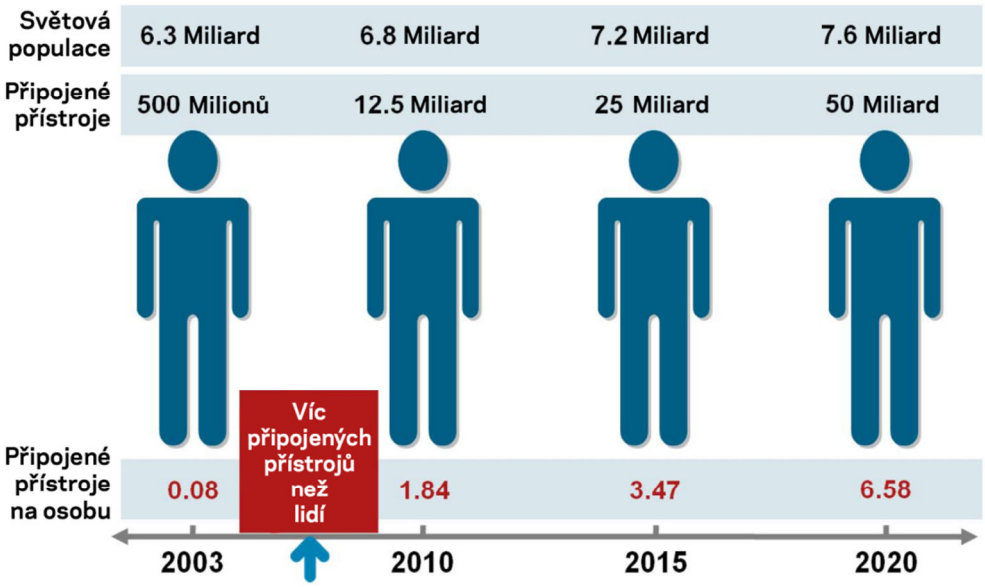



Obr. 4 *Evoluce mobilních technologií* [16]

Tato technologie naprosto změnila způsob, jakým využíváme internet a měla také přímý dopad na prudký rozvoj dalších technologií včetně nové generace „chytrých telefonů“ započaté představením první generace telefonu Apple iPhone v roce 2007. Toho se během jediného roku prodalo téměř 5 milionů [17].

Dopad 3. generace mobilních sítí můžeme demonstrovat na exponenciálním růstu zařízení připojených k internetu bezprostředně po zahájení jejich komerčního využití. V roce 2003 na Zemi žilo přibližně 6,8 miliard lidí, zatím co počet přístrojů připojených k internetu se odhadoval na 500 milionů. V roce 2010 bylo k internetu připojeno již 12,5 miliardy zařízení (tedy asi 1,84 zařízení na jednoho člověka) a toto číslo neustále roste. Odhaduje se, že počet těchto zařízení překonal hranici 7 miliard někdy v období mezi lety 2008 a 2009. Tato významná událost je velmi často nazývána jako vznik

„Internetu věcí“ [18]. Odhady budoucího vývoje počtu zařízení připojených k internetu se však výrazně liší a pochybují se mezi 25 až 75 miliardami v roce 2020 [18][19][20].




Obr. 5 *Zrod „Internetu věcí“ mezi lety 2008 a 2009* [18]

Exponenciální nárůst pozorujeme také v počtu lidí využívajících internet. Prvních 50 milionů uživatelů dosáhl internet v roce 1998, tedy 5 let po vzniku prvních webových stránek. V průběhu následujících 10 let však došlo k dvacetinásobnému nárůstu a v roce 2009 tak bylo dosaženo 1 miliardy uživatelů. Již v roce 2013 internet využívaly 2,1 miliardy lidí a toto číslo se v roce 2013 zvedlo na 2,7 miliardy. Internet v současné době využívá zhruba 40 % světové populace a z toto 60 % alespoň 3 hodiny denně [5].

Podíváme-li se na rychlost nárůstu počtu uživatelů internetu, není divu, že do období spuštění 3G mobilních sítí spadá také vznik prvních sociálních médií, která revolucionizují způsob, jakým mezi sebou lidé komunikují. Mezi ty první patří LinkedIn (2003) zabývající se našimi minulými a aktuálními kariérními zkušenostmi, Facebook (2004) mapující vztahy mezi lidmi a Twitter (2006) zachycující naše nálady [1][5]. Spojení sociálních médií, nárůstu oblíbenosti mobilních technologií a možnosti připojit se k internetu takřka kdekoliv má za následek masivní proud sociálních dat, tedy dat, která vznikají jako produkt lidské činnosti na sociálních sítích, popř. na internetu jako takovém. Tato data mají obrovský potenciál využití v oblasti marketingu, neboť tvoří základ pro jeho personalizaci.

Vyhledávač společnosti Google v prvním roce svého „života“, tedy v roce 1998, musel odpovídat na 3,6 milionu uživatelských dotazů denně. V roce 2007 těchto dotazů denně zpracovával na 1,2 miliardy a v roce 2012 dokonce už 3 miliardy [5].

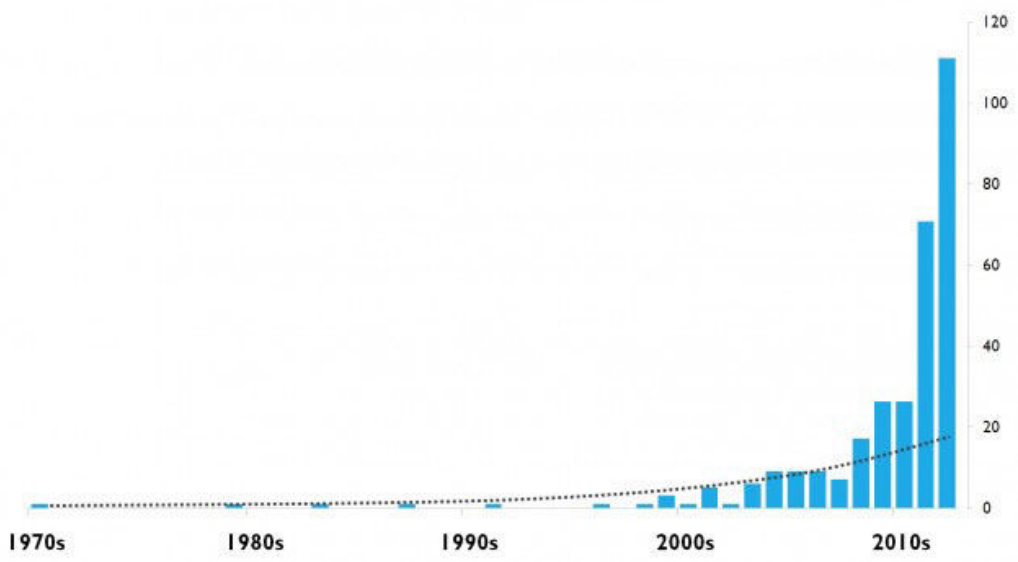
Dle aktuálních statistik z roku 2014 každou minutu uživatelé Twitteru napíší 277 tisíc tweetů, na Facebook je nahráno téměř 2,5 milionů fotografií a na serveru YouTube přibude 72 hodin videa. Každá z těchto sociálních sítí má v současnosti více než 1 miliardu uživatelů. K tomu všemu je po světě během této jedné minuty rozesláno 204 milionů e-mailů [21].

Je téměř neuvěřitelné, jak závislými na počítačích jsme se stali. Podíváme-li se na dobu před 150 lety, většina lidí produkovala pouze minimální množství dat. U některých to dokonce bylo pouze datum narození a jména rodičů, místo a datum jejich manželského sňatku a datum úmrtí [4]. Dnes lidé produkují masivní množství sociálních dat a ochotně tak vědomě či nevědomě poskytují své osobní informace třetím stranám, které si díky jejich analýze dokáží vytvořit komplexní, téměř dokonalý obraz o nás. Je-li daná firma schopná v těchto datech rozeznat obchodní příležitost, pak může znalost svých uživatelů jednoduchým způsobem přeměnit na zisk.

# 2 Big Data

Situace spojená se současným explozivním nárůstem dat a potenciálem jejich využití obchodními společnostmi je v dnešní literatuře velmi často reprezentována souslovím „Big Data“. Najít jejich přesnou a jednotnou definici ve skutečnosti není vůbec jednoduché a v mnohých případech se odbývá jednoduše uvedením indiferentního anglického termínu „buzzword“. Autoři publikací zabývajících se problematikou současné záplavy daty se tímto termínem snaží naznačit, že se jedná o horké téma na poli technologií, které je dnes velmi často skloňováno.

Bibiliometrická studie z roku 2012, která si dala za cíl vyčíslení výskytu pojmu „Big data“ ve vědeckých studiích, ukazuje, že tento pojem byl použit již v příspěvku z roku 1970 zabývajícím se zvukem v atmosféře a prostředí oceánů. Zároveň však ukazuje i výrazný růst počtu vědeckých článků s tématem Big Data počínaje rokem 2008 [22].




Obr. 6 *Meziroční nárůst množství vědeckých prací o Big Datech* [22]

Big Data můžeme popsat jako proces analýzy komplexních datových sad za účelem objevení skrytých informací, které mohou dopomoci lepšímu rozhodování, popř. které odhalí vzorce a souvislosti, které dříve nebyly známé [23].

S dnes nejpoužívanější definicí Big Dat přišla v roce 2001 firma Meta (dnes Gartner), která si jako jedna z prvních povšimla bezprecedentní rychlosti, s jakou objem dat narůstá, a také rozličnosti jejich formátu. Dnes je tato definice notoricky známá jako „3V“, neboť se skládá ze tří anglických slov vystihujících základní podstatu Big Dat a zároveň začínajících písmenem „V“. Jedná se o objem (volume), rychlost (velocity) a různorodost (variety). Vzhledem k tomu, že definice Big Dat není oficiálně stanovena, byla díky své popularitě v nekončící snaze o lepší pojetí pojmu Big Data definice firmy Meta časem rozšířena až na několik desítek „V“ [24].

Zamyslíme-li se nad výše zmíněnou definicí, můžeme si povšimnout, že vychází hlavně z toho, jaké nároky kladou Big Data na současnou výpočetní infrastrukturu. Objemy v řádech petabytů a exabytů vyžadují vysokokapacitní úložiště, která navíc musí být schopná reagovat na neustálý růst dat, a splňovat podmínku škálovatelnosti, tedy být schopná elasticky reagovat na zvyšující se objem dat. Abychom dokázali maximalizovat hodnotu dat, musíme je umět analyzovat stejně, nebo alespoň téměř tak rychle jako je rychlost jejich vzniku. Tato vlastnost Big Dat se stává kritickou hlavně v aplikacích, které jsou založené na okamžité analýze konstantního proudu dat. Nejvýznačnější z takových aplikací je detekce podvodů v online transakcích. I když většina aplikací funguje na odlišném modelu, ve kterém jsou data nejprve uložena a až poté analyzována, představuje rozměr rychlosti Big Dat vysoký nárok na přenosová média [10].

Slovní spojení Big Data je do jisté míry zavádějící. Naznačuje totiž, že jejich čirý objem je základním problémem, který musíme vyřešit. Největší výzva z hlediska jejich analýzy však ve skutečnosti vychází z posledního z oněch tří „V" – různorodosti. Dle některých odhadů je pouze 5 % všech digitálních dat strukturovaných – tzn., že je lze uložit do řádků a sloupců relační databáze [1]. Taková data si můžeme představit např. jako data v excelovské tabulce. Kdybychom se potýkali s čistě strukturovanými daty v řádu exabytů, problém objemu bychom dokázali vyřešit relativně snadno využitím dostatečného výpočetního výkonu. Dnešní data však pochází z obrovského množství zdrojů a výrazně se od sebe liší svou strukturou. Můžeme si pod nimi představit textový soubor, zvukový záznam, video, obrázky, senzorová data, počet klepnutí na webové stránky, geolokační data z GPS přijímačů, data z analýzy DNA apod. Jedná se o velké objemy dat s malou hodnotovou hustotou, jejichž analýzou můžeme dosáhnout hlubšího porozumění daného problému [10].

| **Srovnání tradičních dat s Big Daty** | | |
|---|---|---|
| | **Tradiční data** | **Big Data** |
| Objem | GB | neustále aktualizovaný (dnes TB nebo PB) |
| Rychlost generování | za hodinu, den,… | mnohem rychlejší |
| Struktura | strukturovaná | částečně strukturovaná nebo nestrukturovaná |
| Datový zdroj | centralizovaný | plně distribuovaný |
| Datová integrace | jednoduchá | obtížná |
| Datové úložiště | relační databáze | HDFS, NoSQL |
| Přístup | interaktivní | dávkový nebo v reálném čase |

Tabulka 1 *Srovnání tradičních dat s Big Daty* [10]

Moderní technologie okolo Big Dat jsou velkým tématem a v drtivé většině případů jsou vydávány za jejich samotnou definici. Tato definice ale ve skutečnosti opomíjí důležitý fakt, že Big Data znamenají především transformaci myšlení. Počítače data sice zpracovávají, ale to samo o sobě neznamená revoluci. Revoluce probíhá v samotných datech a ve způsobu, jakým s nimi pracujeme. Máme-li přístup k velkým datovým

sadám, jsme schopni provádět operace, které bychom s malými objemy dat dělat nemohli. Se změnou množství se totiž pojí změna podstaty. Jako příklad můžeme uvést rozdíl mezi fotografií a filmem. Před vynálezem videokamery jsme byli schopný zachytit pouze jednotlivé statické obrazy daného objektu. V momentě, kdy jsme byli schopni zachytit alespoň 24 snímků tohoto objektu za sekundu, se statické obrázky proměnily na film – došlo ke změně podstaty. Získali jsme tedy nástroj, který nám umožňuje pozorovat vývoj daného objektu v čase. Zvětšením rozsahu dat tedy můžeme získat informace, kterých bychom analýzou malých datových množin nemohli dosáhnout. Tím se otevírá cesta k prediktivním systémům, které pracují s velkými objemy dat. Na jejich výstupu vznikají předpovědi, které jsou základním smyslem Big Data analýzy [1].

## 2.1 Analýza základního statistického souboru

Potenciál ležící v Big Datech je pro lidi, kteří se s nimi setkají úplně poprvé, většinou náročný na představu. Není jednoduché na první pohled vidět, v čem je Big Data analýza lepší než výsledky klasických metod statistické analýzy. Základní překážka pro uvolnění potenciálu Big Dat leží totiž především ve skutečnosti, že naše myšlení pochází ze světa malých dat, tedy z takového světa, kde data představují jakousi vzácnost. Aniž si to uvědomujeme, stavíme si tím umělé bariéry, které již ve světě Big Dat neexistují. Způsob, jakým se data zpracovávají ve světě malých dat, nám nastiňuje Novovičová na příkladu předvolebního průzkumu [25]:

> „Bylo by příliš nákladné a nerealistické dotazovat se všech voličů na jejich volební preference. Statistikové, kteří si přejí odhadnout mínění celé *populace* voličů ČR, se mohou dotazovat jen pečlivě vybrané skupiny několika tisíců voličů. Taková skupina voličů se nazývá *výběr* z populace. Statistikové analyzují informace získané z výběru, aby udělali závěry o volebních preferencích celé voličské populace … *Inferenční statistika* se skládá z metod pro přijímání spolehlivosti závěrů o populaci založených na informacích získaných z výběru této populace.“

Metody inferenční statistiky vznikly v době, kdy kvůli technologickým omezením, popř. z ekonomických důvodů, nebylo možné zpracovávat velké datové sady. Jednalo se vlastně o umělé limity kladené soudobou technologií, které však byly považovány za nevyhnutelnou realitu. Statistika představuje nástroj, pomocí něhož se na základě minimálního množství dat snažíme vyvodit závěry o základní, tedy úplné datové množině. S pomocí statistiky jsme navíc schopní určit i pravděpodobnost, s jakou se takto vypočtené předpoklady o základním souboru blíží realitě.

Nasnadě je hned první problém. Novovičová [25] uvádí, že je rozhodující, „aby výběr byl reprezentativní, to znamená, že musí odrážet co možná nejvěrněji relevantní charakteristiky základního souboru, který je předmětem našeho zkoumání.“ Jak ale vybrat ze základního souboru vzorek tak, aby co nejméně zkresloval jeho realitu? V praxi se ukázalo, že abychom dosáhli co nejmenší chybovosti, je nutné při výběru

vzorku z populace usilovat o nahodilost [1]. V praxi se nejčastěji používají tzv. pravděpodobnostní výběry, které mají zabránit jednostrannosti výběru a dovolují statistikům kontrolovat nereprezentativnost daného vzorku [25]. Statistici zjistili překvapující závěr: Při určité velikosti vzorku se s jeho dalším rozšiřováním přesnost predikce již nezlepší, pokud jeho jednotlivé členy nebyly vybírány nahodile. Právě nahodilosti se však dosahuje těžko a můžou proto vznikat systematické odchylky, které vedou k vysoké chybovosti výsledků zpracování extrapolovaných dat. Tento problém se často citelně odráží např. právě ve volebních průzkumech. Při souboji Baracka Obamy a John McCaina v prezidentských volbách v roce 2008 přinášely předvolební průzkumy nepřesné předpovědi, což vhledem k vyváženosti tohoto politického boje představovalo problém. Po volbách se ukázalo, že chyba vnikla především tím, že tazatelé volali výhradně na pevné telefonní linky. Zahrnutí uživatelů mobilních telefonů přitom dokáže ovlivnit předpověď o jedno až tři procenta [1]. S odlišným přístupem k volebním průzkumům přišel blogger deníku The New York Times Nate Silver. Právě v roce 2008 ohromil celý svět, když pouze díky Big Data analýze datového souboru popisujícího zhruba 46 socio-ekonomických proměnných dokázal přesně predikovat výsledky těchto prezidentských voleb v 49 z 50 států. Při znovuzvolení Baraca Obamy o 4 roky později se mu podařilo predikovat přesné výsledky již ve všech jednotlivých státech [26].

Nakolik nízká míra nahodilosti při vybírání jednotlivých prvků ze základního souboru představuje problém, nejedná se o jediné omezení, které s sebou vzorkování nese. Při vzorkování především dochází k snížení rozlišení a ztrátě detailů. V realitě to pak především znamená, že při zkoumání daného vzorku již není možné snížit měřítko a zkoumat jednotlivé podkategorie celé populace [1].

Představme si modelovou situaci, ve které se aerolinka snaží zjistit, jak její cestující vnímají zrušení bezplatného cateringu na všech jí operovaných letech. Pro tento účel byl vyvinutý jednoduchý dotazník, k jehož vyplnění byli cestující během letu vybízeni. Při výstupu z letadla ho odevzdávali palubním průvodčím a byli přitom ponecháno na cestujících, jestli dotazník odevzdají nebo ne. Na konci sezóny pak byla provedena analýza odpovědí 2000 respondentů, během které se zjistilo, že 50 % cestujících zrušení bezplatného občerstvení vnímalo negativně, ale nebude je to motivovat k přechodu ke konkurenci, 30 % vnímalo změnu neutrálně a zbylých 20 % s touto společností už nikdy nepoletí. Při prezentaci těchto výsledků managementu společnosti vyvstaly doplňující otázky: Byla tato změna vnímána na letech delších jak 3 hodiny negativněji než na kratších letech? Byli cestující na charterových letech méně spokojení než cestující na pravidelných linkách? Kdo byl spokojený méně, cestující na Kapverdy nebo do Bulharska? Jak by spokojenost cestujících zvýšilo alespoň bezplatné podávání teplých nápojů?

V první řadě si musíme uvědomit, že cílem průzkumu bylo zjistit postoj všech cestujících dané aerolinky, tedy ze statistického hlediska celého základního souboru, k zrušení bezplatného občerstvení, a pro tento účel byl vybrán reprezentativní vzorek ve snaze odpovědět na tuto jednu konkrétní otázku. Chceme-li dodatečně odpovědět na některou z výše uvedených otázek, musíme si uvědomit, že při zkoumání jakékoliv podkategorie výběru z populace (např. cestující na letech nad 3 hodiny, cestující do Bulharska, cestující na pravidelné lince,…) se vystavujeme riziku velké chyby, neboť jsme během průzkumu v první řadě nezajistili rovnoměrný počet respondentů v každé podkategorii. Kdyby se tedy na letech na Kapverdy průzkumu zúčastnili pouze 3 cestující, zatím co na letech do Bulharska 250, je evidentní, že výsledek porovnání sentimentu těchto dvou podkategorií cestujících by bylo zatíženo velkou chybou. Poslední z výše uvedených otázek ukazuje, že výběru z populace většinou není možné dodatečně položit další otázky, které nebyly součástí samotného šetření, neboť po skončení sběru dat již těžko zjistíme, jestli by podávání bezplatné kávy nebo čaje přispělo ke zvýšení spokojenosti.

Je zde jasně vidět, že vzorek rychle ztrácí svou užitečnost v případě, kdy chceme přejít na nižší úroveň. Ztrátu detailů jednotlivých podkategorií můžeme připodobnit ke komprimaci digitální fotografie, na které chceme rozpoznat tvář člověka vpovzdálí. Díváme-li se na ni z dálky, vidíme celý obraz dobře a ostře. Začneme-li ji však přibližovat, uvidíme už jen jednotlivé pixely. Stejně jako u výběru z populace ztrácíme možnost zkoumat daný segment do hloubky, zaměřit se podrobně na daná hlediska a prozkoumat data z různých úhlů [1].

Náhodné vzorkování je přežitkem z dob málo vyspělých informačních technologií. V dnešní době, kdy disponujeme dostatkem výpočetní a úložné kapacity a špičkovými nástroji, jsme v mnohých případech schopni náhodné vzorkování nahradit Big Data analýzou, která odstraňuje většinu jeho nedostatků. Používání veškerých dostupných dat namísto jejich malé výseče nám umožňuje vysledovat podrobnosti a souvislosti, které se jinak v záplavě dat ztrácejí. Data můžeme procházet do větší hloubky a opakovaně analyzovat zcela novým způsobem, nad kterým jsme při jejich sběru ještě vůbec neuvažovali. Big Data nám umožňují zkoumat nové hypotézy na mnoha úrovních detailů, protože jsou založena na co největším množství informací a jejich analýza není zatížena rizikem rozostření. Použijeme-li všechna data, můžeme pozorovat i ty nejmenší detaily, kterých bychom si jinak u vzorku nikdy nemohli všimnout, protože by v něm prostě nebyly přítomné.

Jak již bylo několikrát řečeno, Big Data představují analýzu na úrovni kompletní datové sady. Provádíme tedy operace nad základním statistickým souborem. V „3V“ definici Big Dat figuruje na prvním místě objem (volume), který naznačuje nutnost analýzy bezprecedentně velkého množství dat. Považuji však za vhodné na tomto místě zmínit, že kompletní datová množina může, ale nemusí nutně znamenat velký absolutní rozsah

v řádu terabytů nebo petabytů. Jedná se spíše o relativní velikost v porovnání s maximálním množstvím dostupných dat. Znamená to pouze to, že se za každých okolností snažíme využít kompletní informace, které máme k dispozici [1].

## 2.2 Od přesnosti k přibližnosti

Další oblastí, ve které nás Big Data nutí přehodnotit naše dosavadní postupy, je přesnost zpracovávaných dat. Ve světě statistické analýzy, kde se snažíme vyvodit závěry o celé populaci na základě jejího vzorku, jsme museli těch pár reprezentativních čísel, která mají zásadní dopad na vypovídající hodnotu dané statistiky, kvantifikovat co možná nejpřesněji. Analyzujeme-li pouze omezený počet datových bodů, výsledná chyba se stále více zvýrazňuje s tím, čím méně přesná data používáme. Když tedy používáme pouze malé množství dat, snažíme se logicky dosáhnout jejich pokud možno co nejvyšší kvality.

Shromažďování velkého množství dat popravdě vytváří prostor pro vznik nepřesností. V tomto objemu jich vzniká dokonce tolik, že již není v našich silách všechny odstraňovat. Jedná se však o jistý druh kompromisu – připustíme-li chybovost dat, dokážeme jich shromáždit o to větší množství. V konečném důsledku na tom můžeme víc získat než ztratit.

Chyby jsou do datové množiny vnášeny různými způsoby. Častější výskyt chyb je v první řadě spojen se samotným růstem objemu dat. Pokud měříme teplotu výfukových plynů motoru tisíckrát častěji, poroste i pravděpodobnost chyb. Chybovost pak může být zapříčiněna i kombinací informací z různých zdrojů, které se liší strukturou, a nejednotností formátování, kvůli které je data před zpracováním potřeba vyčistit (například letadlo může být v závislosti na zdroji vyjádřeno jako airplane, aeroplane, aircraft, A/C, ACFT, apod.). Nepřesnosti vznikají také při získávání a zpracování dat, kdy je vezmeme a přetransformujeme na něco jiného, např. když zanalyzujeme hlasové zprávy z call centra letecké společnosti a přeměníme je na predikce prodeje letenek v následujícím roce.

Řekněme, že potřebujeme měřit napětí nosníku křídla letadla. Pokud máme na tomto nosníku jenom jediný senzor, pak potřebujeme, aby dodával přesná data a fungoval nepřetržitě. Jestliže však umístíme jeden senzor na každý čtvereční centimetr, dovolí nám to použití levnějších senzorů (pouze však do té míry, aby nezpůsobily systematickou odchylku měření). Dá se předpokládat, že některé s nich budou čas od času reportovat nepřesné hodnoty a my tak získáme více chybnou datovou sadu, než kdybychom použili jeden přesný senzor. Přestože jedno konkrétní měření nemusí být zcela správné, získáváme tím mnohem podrobnější celkový obrázek. Představme si, že nyní zvýšíme frekvenci odečtu senzorů na deset za sekundu místo původního jednoho odečtu za minutu. Klesne tak pravděpodobnost, že sekvence údajů bude souvislá. Některé údaje se mohou ztratit nebo zpozdit. Vzniklá data nám však

v celkovém měřítku mohou dát úplnější obrázek o situaci a vyváží tak vzniklé chyby. V prvním případě jsme se sice vzdali přesnosti jednotlivých datových bodů, ale mohli jsme tak provést mnohem víc měření a v důsledku toho zjistit podrobnosti, které jsme předtím nemohli pozorovat. V druhém případě jsme na úkor přesnosti zvýšili frekvenci odečtu a dokázali tak zachytit dynamiku celého procesu.

Je jasné, že data, která využíváme, nemohou být zcela chybná. Pokud se však snažíme zachytit pouze obecný trend, můžeme částečně slevit ze svých nároků na přesnost a za odměnu dostaneme možnost pracovat s mnohem větší datovou sadou. Při zpracování Big Dat se totiž spíše než přesných hodnot dobíráme pravděpodobnostních údajů. Slovy firmy Forrester: „Dvě a dvě se někdy rovná 3,9 a tento výsledek je docela dobrý [1].“ Přesnost je možné různými prostředky zvyšovat, ale v případě, kdy se snažíme vysledovat pouze obecnou závislost, to mnohdy není příliš ekonomicky výhodné. Posedlost přesností spadá do doby, kdy měl každý bod datové množiny zásadní vliv na přesnost výsledku analýzy. V dnešní době vykupujeme snížením přesnosti možnost vytvořit si mnohem komplexnější obrázek o dané situaci.

## 2.3 Post hoc ergo propter hoc

*Potom, tudíž proto.* Latinský výrok v názvu této podkapitoly mistrně vystihuje chybný proces uvažování, kterým je odjakživa zatížena lidská mysl [27]. Z důvodu nutnosti rychlého rozhodování v nebezpečných situacích se nás mozek snaží ošálit a nutí nás myslet v intencích příčinných souvislostí neboli kauzalit. Jedná se o metodologickou a logickou chybu argumentace, neboť předpokládáme, že jedna událost má svou příčinu v události jí předcházející jen proto, že po ní bezprostředně následuje [28].

Pro vysvětlení rozporu naznačeného v předchozím odstavci uveďme příklad letecké společnosti poskytující charterový servis do exotických letovisek, která se dočkala dodávky svého prvního letadla typu Boeing 787. Po jeho dodání v měsíci květnu vzrostl zisk letecké společnosti v následujících měsících o 300 %. Manažeři si mnou ruce, pochvalují si, jak jim moderní letoun dopomohl ke zvýšení zisku, a plánují doobjednávku dalších dvou letadel stejného typu. Dopustili se při tom klasické chyby – přisoudili nárůst zisku dodávce letounu jen proto, že následoval bezprostředně po něm. Ve skutečnosti se za zvýšením zisku skrývá nástup letní sezóny, která tradičně začíná v měsíci květnu, ve kterém bylo čirou náhodou dodáno i nové letadlo.

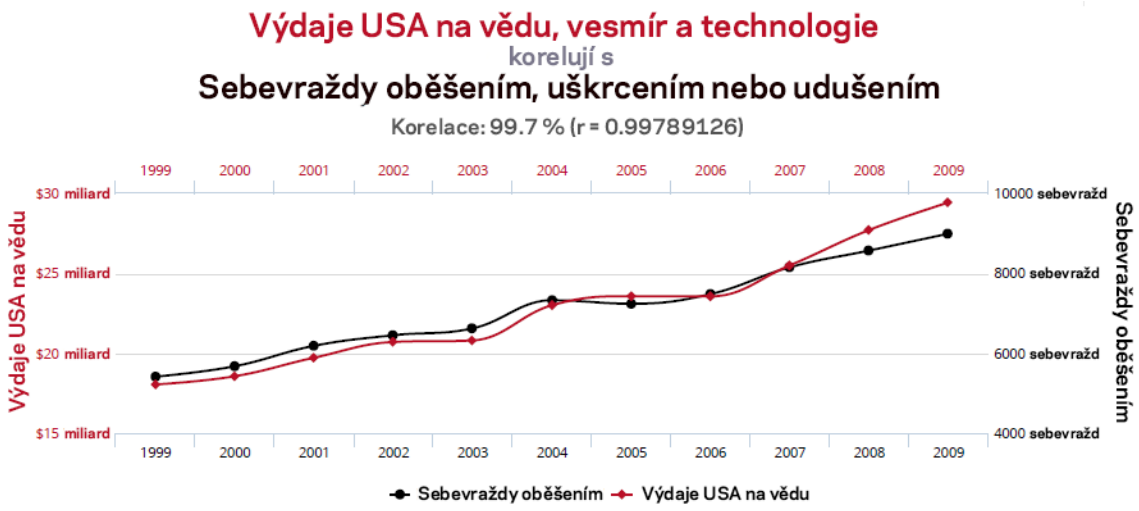
Jak vidíme v předchozím příkladu, v mnohých případech představuje takové rychlé kauzální myšlení pouze poznávací zkratku a přestože je to v rozporu se selským rozumem, nijak neobohacuje naše poznání světa. Ukazuje, že chceme věřit tomu, že každý jev má svou příčinu a na základě toho se snažíme najít smysl v tom, co se právě děje.

Je nutné si uvědomit, že při vyhledávání vzorů a souvislostí pomocí Big Data analýzy se nesnažíme dokazovat kauzální vztahy. To znamená, že neumíme se stoprocentní

jistotou potvrdit, že jev A byl příčinou jevu B. Pokud vidíme, že jev A je nějakým způsobem svázaný s jevem B, říkáme, že spolu korelují. Jinými slovy se pomocí korelace snažíme kvantifikovat statistický vztah mezi dvěma datovými hodnotami. Pokud spolu dvě veličiny silně korelují, znamená to, že když se bude jedna z nich měnit, bude ji s největší pravděpodobností následovat i druhá. Zároveň však spolu dvě veličiny mohou silně korelovat jen z toho důvodu, že jsou obě navázané na tzv. skrytou veličinu [25]. Jako úsměvný příklad korelace můžeme uvést zhoršování výkonnostních charakteristik motorů v závislosti na počtu pasažérů, kteří mají oblečené tričko s krátkým rukávem. Cestující v krátkém tričku nejspíše sami o sobě nedokážou ovlivnit výkonnost motorů – pouze se přizpůsobují vysoké venkovní teplotě, skryté veličině, která má zároveň negativní vliv na výkon motorů.

Silné korelace jsou pro nás velmi důležité obzvláště z důvodů vzájemné zastupitelnosti korelujících veličin. Je-li pro nás obzvláště obtížné získat data o veličině A, ale zároveň víme, že s ní veličina B, kterou měřit umíme, silně koreluje, dokážeme prostřednictvím veličiny B stanovit závěr o veličině A. Při Big Data analýze to pro nás představuje obrovskou výhodu, poněvadž už nemusíme zkoumat přesné vnitřní principy daného jevu. Stačí nám pro něj najít jev zástupný. Díky tomu jsme při pozorování současného stavu schopni predikovat stav dané veličiny v budoucnosti.

Korelace jsou pro nás v tomto směru skvělým nástrojem, ale přesto je nutné mít neustále na zřeteli, že udávají pouze pravděpodobnost, že spolu nějaké dva jevy souvisí. Ve skutečnosti se může jednat pouze o tzv. „zrádnou nahodilost“, jak je vtipně naznačeno na obrázku níže.




Obr. 7 *Nahodilá korelace mezi výdaji USA na vědu a sebevraždami oběšením*
(Zdroj: http://www.tylervigen.com/spurious-correlations)

Korelace byly samozřejmě využívány již před érou Big Dat, ale pouze v omezené míře. Zkoumání korelací je dle statistických metod založeno na potvrzování a vyvracení hypotéz. Na počátku zkoumání stojí hypotéza, která předpokládá, že jev A by snad

mohl být zástupnou veličinou pro jev B. Můžeme se například domnívat, že zpoždění letu silně koreluje s ochotou cestujících letět s danou aerolinkou znovu v budoucnosti. Jedná se o pouhopouhou teorii. Nejprve shromáždíme data o zpoždění letů této aerolinky a poté provedeme průzkum mezi jejími cestujícími. Následně provedeme korelační analýzu těchto dvou datových množin. V případě, že spolu zpoždění a ochota cestovat s danou společností silně korelují, prohlásíme, že se jedná o zástupné jevy. V opačném případě zavrhneme původní hypotézu, zvolíme novou zástupnou veličinu a proces opakujeme.

Na první pohled je vidět, že se jedná o velmi zdlouhavý proces zjišťování metodou pokusů a omylů a rozšiřování našeho poznání tímto způsobem vede pouze k velmi pomalému pokroku. Vzhledem k časové náročnosti navíc není příliš výhodné touto metodou zkoumat jiné než lineární korelace, což pro nás znamená obrovské omezení, protože korelace mezi veličinami bývají často velmi komplexní.

Testování korelace pomocí hypotéz představuje v dnešní době přežitek. Dnes, kdy disponujeme dostatečným výpočetním výkonem a velkými datovými množinami, již nevybíráme zástupné veličiny jednu po druhé, ale identifikujeme je na základě vysoce sofistikovaných algoritmů. Zástupné veličiny se nám tak zjeví automaticky provedením Big Data analýzy. Výsledky budou nejen méně zkreslené, ale navíc je obdržíme mnohem rychleji.

Predikce založené na korelacích jsou jádrem Big Dat. Například data spotřebitelského marketingu byla identifikována jako zástupná veličina k analýze vzorků krve a moči, pozdní placení složenek jako indikace probíhajícího rozvodu nebo zájem o zdravý životní styl a kupovaní neparfémovaného pleťového mléka jako doprovodný jev těhotenství. Stejně jako kauzalitu nelze ani korelace se stoprocentní jistotou dokázat. Umíme je pouze potvrdit s vysokým stupněm pravděpodobnosti. Jednou z výhod korelace je také možnost jejího vyjádření pomocí matematického zápisu. To v případě kauzality není možné. Kauzální vztahy lze dokázat jedině experimentálně navozením velmi přesných laboratorních podmínek, ve kterých jsme schopní příčinu daného jevu uměle vyvolat nebo naopak potlačit. Vznik, popř. absence následného jevu je pak indikací kauzálního vztahu. Experimentální zkoumání kauzalit však může vyvolávat obtížné etické otázky. Jak by asi vypadal kauzální experiment, zkoumající, proč těhotné ženy kupují neparfémované pleťové mléko? V těchto případech je užitečná právě korelace. Nejen že jsme korelační analýzu schopni provést relativně levně a rychle, ale zároveň dostáváme i návod kde máme začít se zkoumáním kauzálních vztahů, pokud bychom je snad z nějakého důvodu potřebovali znát. Pomocí korelační analýzy se sice nedozvíme odpověď na otázku PROČ spolu nějaké jevy souvisí, ale v drtivé většině případů nám při rozhodování stačí vědět pouze to, že spolu tyto jevy vůbec nějak souvisí. Ve skutečnosti nás tedy vlastně nezajímá *proč.* Docela nám stačí, když víme *co*.

## 2.4 Uvolnění skryté hodnoty datových sad

Vnímání dat jako pouhého doprovodného jevu probíhajících transakcí se začíná vyvíjet a pomalu ale jistě začínají být považována za samostatně obchodovatelný artikl. S tím, jak postupně mizí mnohá základní omezení sběru, uchovávání a analýzy velkých datových sad, dochází k tomu, že za cenná jsou považována úplně všechna data sama o sobě. Jsou tím myšlena i taková data, která se do dnešní doby zdánlivě nevyplatilo uchovávat. Naše technologické možnosti dospěly do stádia, kdy jsme zpravidla schopní zachycovat a uchovávat velké objemy informací velmi levně. Z dob, kdy se firmy z důvodu omezené úložné kapacity a její vysoké ceny musely rozhodovat, která data jsou důležitá a vyplatí se je uchovat a která je potřeba smazat, přecházíme do stavu, kdy se velmi často můžeme rozhodnout data si ponechat. Za posledních 50 let vzrostla kapacita úložišť 50 milionkrát a zároveň náklady na ně klesaly zhruba na polovinu každé dva roky.

Okamžitá hodnota dat, která firmy shromažďují, je většinou zcela evidentní. Obchodní společnosti shromažďují transakční data za účelem podpory prodeje, výzkumné ústavy generují data v naději objevní nových poznatků o určitém fenoménu, letecké společnosti získávají data od svých partnerů, aby si dokázaly utvořit ucelenější obrázek o aktuálním dění v provozu, apod. Důvody, které podněcují firmy ke sběru těchto dat, se nazývají primární aplikace. Jinými slovy se jedná o účel, za kterým jsou data sbírána.

V předchozí kapitole jsem zmínil, že jednou z vlastností dat je i to, že se opakovaným použitím neopotřebovávají. Jedná se o zásadní vlastnost, která hraje ve světě Big Dat velkou roli. Znamená to totiž, že když data využijeme ke svému primárnímu účelu (tedy když si např. vytvoříme povědomí o okamžité situaci v leteckém provozu z dat o pohybech letadel), neztratí tím na své hodnotě. Tato data můžeme opakovaně využít ke stejnému účelu a co víc, s trochou fantazie je můžeme využít k účelům, které při jejich sběru nebyly zjevné. Ukazuje to, že hodnota dat je mnohem větší než pouze ta hodnota, kterou jsme získali při jejich primární aplikaci.

Hodnotu dat můžeme připodobnit k ledovci plovoucímu oceánem. Primární aplikace dat představuje jeho špičku, která ční nad hladinou a není možné ji přehlédnout. Potenciál dat je však představovaný tou částí, která se skrývá pod hladinou, a uvidíme ji pouze v případě, že ji tam budeme předpokládat. Tento potenciál představuje sekundární aplikaci využití dat. S čím více různými způsoby různého využití jedné datové sady dokážeme přijít, tím víc bude časem její hodnota růst. Podívejme se například na historická data z globálních distribučních systémů. Tato data, která primárně sloužila k prodeji letenek, je možné díky prediktivní analýze proměnit v předpovědi vývoje cen letenek v budoucnosti, což představuje sekundární využití. Sekundární aplikace si vyžadují především otevřenou mysl, inovativní přístup ke zpracování dat a moderní sadu metod a nástrojů, se kterými přichází nová generace statistiků. V minulosti byla data po dosažení cíle, pro který byla nasbírána, vymazána

nebo minimálně přesunuta do archivu. Dnes již víme, že v sobě stále schovávají potenciální hodnotu, kterou je možné vhodnými prostředky uvolnit. Mezi tyto prostředky řadíme především opakované využití a kombinace datových množin.

Skrytá hodnota jedné datové množiny se někdy plně projeví až v momentě, kdy je zkombinována s jinou množinou, která často může popisovat diametrálně odlišný jev. Jako příklad vezměme studii, ve které bychom chtěli pomocí Big Data analýzy zjistit, jaký vliv má kosmické záření, kterému jsou lidé vystavení během letu, na vznik nádorových onemocnění. Předpokládejme, že bychom byli schopní získat seznamy pasažérů všech českých leteckých společností za posledních 10 let a zároveň data o onkologických pacientech v České republice za stejné období. Objem těchto dat by byl tak velký, že by se jistě tradičními metodami nedal analyzovat. V tomto případě bychom jistě přistoupili k Big Data analýze a prozkoumali, nakolik spolu tyto dvě datové sady korelují.

Z hlediska marketingu představují skrytou hodnotu i takzvané datové zplodiny. Jedná se o data ve formě tzv. digitální stopy, kterou po sobě zanechává uživatel svou činností v prostředí internetu. Mezi tyto informace patří doba, kterou uživatel strávil čtením dané stránky, trajektorie pohybu kurzoru myši po obrazovce, počet kliků myši, slova zadávaná do vyhledávače, apod. Tato data, která byla dlouhou dobu přehlížena, mají nedocenitelnou hodnotu, protože jsou základem tzv. „učení z dat“ (angl. machine learning). Opouští-li uživatel stránku příliš brzo, indikuje v systému, že nenašel to, co potřeboval. Vysoký počet kliků na dané stránce může odkazovat na špatnou strukturu internetové stránky, na které se těžko hledají informace a uživatel se k nim musí „proklikat“. Tvůrci tak dostávají zpětnou vazbu o tom, co se uživatelům líbí a co by se naopak mělo zlepšit. Firma, která je schopná využít potenciál datových zplodit, může nabýt výraznou konkurenční výhodu.

# 3 Technologie stojící za Big Daty

Jedná z mnohým definic Big Dat říká, že se jedná o novou generaci technologií a architektur navržených k tomu, aby bylo možné ekonomicky extrahovat hodnotu z velmi velkých datových množin skládajících se z rozmanitých dat, umožněním jejich vysokorychlostního záznamu, objevování a analýzy [10]. Skutečně, definice Big Dat se velmi často odráží od jejich technologické základny, neboť se jedná o revoluční nástroje, které přináší nový způsob práce s datovými množinami. Základní znalost nových technologií v oblasti zpracování velkých datových množin je nutný pro pochopení výzev, které s sebou Big Data přináší. Mezi tyto technologické výzvy můžeme zařadit následující problémy:

- Rozmanitost různých zdrojů dat a samotná velikost datových množin představují vysoké nároky na jejich sběr a integraci
- Big Data systémy musí ukládat a spravovat tyto nashromážděné masivní a heterogenní datové sady a zároveň poskytovat funkční a výkonnostní garanci ve smyslu jejich rychlého načítání, škálovatelnosti (tzn. rozšiřování a smršťování systému v závislosti na aktuálním množstvím dat v něm, aniž by to mělo vliv na jeho výkonnost) a ochrany soukromí
- Big data analýza musí efektivně dolovat informace z masivních datových sad na různých úrovních v reálném čase, popř. v téměř reálném čase, včetně modelování, vizualizace, predikce a optimalizace, aby z nich vyplynuly hodnoty nutné k zajištění usnadnění rozhodovacího procesu a zisku dalších výhod

Výše popsané technologické výzvy nás nutní k hluboké revizi současných systémů řízení báze dat, tedy databází, počínaje principem architektury až po detaily jejich implementace. Tradiční databáze a analytické systémy, převážně založené na konvenčních relačních databázových systémech, jsou nedostačující k řešení předeslaných výzev kladených Big Daty. Neshoda mezi tradičními databázemi a objevujícím se Big Data paradigmatem vychází především z toho, že relační databáze představují velkou podporu pro strukturovaná data, ale pro nestrukturovaná téměř žádnou. Jelikož využívají nákladný hardware, nevyhovují ani podmínce škálovatelnosti, neboť s jejich rozšiřováním začne téměř okamžitě rapidně růst křivka nákladů na hardware [29].

Pro Big Data systémy byla s ohledem na tyto výzvy navržena různá řešení, jejichž prozkoumání si dává tato kapitola za cíl.

## 3.1 Architektura Big Data sytému

**Hodnotový řetězec** Big Data analýzy se skládá ze 4 základních fází:


*Vznik dat* je fází, ve které se data rodí. Jak již bylo řečeno dříve, výraz Big Data má za úkol indikovat velkou, různorodou a komplexní datovou množinu generovanou z různých distribuovaných datových zdrojů, kterými mohou být například senzorová data, video, kliknutí na webové stránce, a další dostupné digitální zdroje.

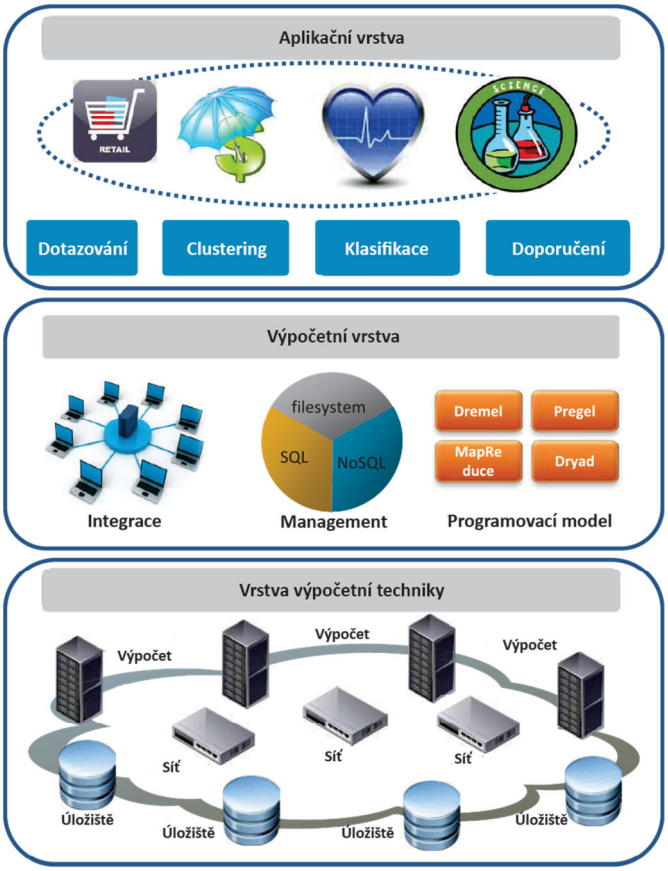
*Získání dat* odkazuje na proces obstarávání informací a dělí se dále na sběr, přenos a předzpracování dat. Sběr je zastoupen technologií určenou pro sběr dat, prostřednictvím které získáváme hrubá data ze specifických oblastí jejich produkce. Po nasbírání hrubých dat potřebujeme vysokorychlostní přenosový mechanizmus pro přenos dat do příslušného úložiště odpovídajícího specifické analytické aplikaci. Sesbírané datové sady mohou ve skutečnosti obsahovat velké množství nesmyslných dat, která zbytečně zabírají úložný prostor a ovlivňují následnou analýzu. Například redundance je typickým doprovodným jevem sběru senzorových dat a o její minimalizaci se pokoušíme prostřednictvím předzpracování dat.

*Uložení dat* znamená trvalé uložení a spravování velkých datových sad. Úložný systém může být rozdělen na 2 části: hardwarovou infrastrukturu a správu dat. Hardwarová infrastruktura představuje elastický soubor zdrojů informační techniky. Měla by splňovat podmínku škálovatelnosti a schopnost dynamické rekonfigurace z důvodu jejího použití v různých typech aplikačního prostředí. Software datové správy se rozprostírá nad hardwarovou infrastrukturou za účelem „údržby“ velkých datových sad. Navíc k tomu musí být zajištěna funkce rychlého dotazování a vytvořeny programovací modely a rozhraní, umožňující rychlou analýzu a interakci s daty.

*Analýza dat* využívá analytické metody nebo nástroje k průzkumu, transformaci a modelování dat za účelem extrahování hodnoty. Nově vznikající analytický výzkum může být rozdělen do šesti technologických oblastí: analýza strukturovaných dat, analýza textu, analýza multimédií, analýza webu, síťová analýza a analýza mobilních technologií.

Jednotlivé fáze hodnotového řetězce budou rozpracovány jako samostatné podkapitoly.

Alternativu k hodnotovému řetězci Big Data analýzy tvoří tzv. **vrstevnatý model**.




Obr. 8 *Vrstevnatý model Big Data architektury* [10]

V nejspodnější vrstvě vrstevnatého modelu leží *vrstva výpočetní techniky*. Nad ní stojí *výpočetní vrstva* skládající se z různých datových nástrojů, které pracují nad informační technikou. Mezi typické datové nástroje ve sféře Big Dat řadíme integraci dat, správu dat a programovací model. Integrace znamená získávání dat z rozličných zdrojů a jejich seskupení do unifikované formy včetně procesu předzpracování. Správa dat neboli data management odkazuje na mechanizmy a nástroje poskytující stálé uložení dat a jejich vysoce efektivní správu, kterými jsou například distribuované souborové systémy a datová úložiště založená na SQL a NoSQL databázích. Programovací model implementuje logiku abstraktní aplikace a usnadňuje datovou analýzu.

Na vrcholu modelu stojí *aplikační vrstva*. Ta využívá rozhraní poskytované programovacím modelem k implementaci různých funkcí datové analýzy včetně dotazování, statistické analýzy, clusteringu a klasifikace. Poté použije kombinaci základních analytických metod k vytvoření různých aplikací souvisejících s daným odvětvím, kterým může být např. doprava, zdravotnictví, obchod, apod.

## 3.2 Získání dat

Jak již bylo zmíněno dřív, hodnotový řetězec Big Dat systému začíná samotným vznikem dat. Jelikož byl vznik a nárůst digitálních dat podrobně rozpracován na začátku

této práce, dovoluji si ho na tomto místě přeskočit a pokračovat rovnou druhým bodem hodnotového řetězce – *Získáním dat.*

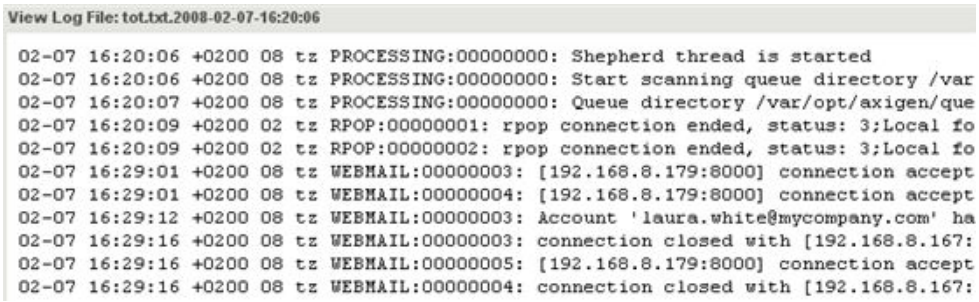
Získání dat má v hodnotovém řetězci funkci agregace informací v digitální formě pro budoucí uložení a analýzu. Skládá se ze tří kroků – sběru, přenosu a předzpracování. Přitom není přesně určeno, zda předzpracování dat proběhne před, anebo až po přenosu dat do úložiště.


A. *Sběr dat*

Sběr dat je proces získávání hrubých dat z objektů reálného světa. Tento proces musí být dobře navržený, protože nepřesný sběr by mohl mít dopad na následnou proceduru datové analýzy a nakonec by vedl k nesprávným výsledkům. Metody sběru dat přitom nejsou závislé čistě na fyzických charakteristikách zdroje dat, ale také na předmětu zamýšlené analýzy. Důsledkem toho vzniká mnoho různých metod sběru dat. Jedněmi z nejčastějších jsou sběr pomocí senzorů a takzvaných souborů typu „log".

Senzory jsou využívány především v případech, kdy potřebujeme změřit fyzickou kvantitu určité veličiny a převést ji na čitelný digitální signál za účelem jejího zpracování. Na základě veličin, které senzory měří, se dělí na senzory: akustické, vibrační, chemické, tlakové, termální, měřící vzdálenost, počasí, elektrický proud, atd. Pomocí drátových nebo bezdrátových sítí jsou signály ze senzorů přenášeny do sběrného bodu.

Soubory typu log jsou generovány systémy obsahujícími zdroj dat, aby mohl být proveden záznam aktivit v specifikovaném datovém formátu pro následnou analýzu. Jsou využívány téměř všemi aplikacemi, které běží na digitálních přístrojích. Například webový server v nich běžně zaznamenává všechna kliknutí, přístupy a další atributy vytvořené každým uživatelem webové stránky. Na rozdíl od senzorů samotných můžou být logy považovány za „software plnící funkci senzoru".


View Log File: tot.txt.2008-02-07-16:20:06

```
02-07 16:20:06 +0200 08 tz PROCESSING:00000000: Shepherd thread is started
02-07 16:20:06 +0200 08 tz PROCESSING:00000000: Start scanning queue directory /var
02-07 16:20:07 +0200 08 tz PROCESSING:00000000: Queue directory /var/opt/axigen/que
02-07 16:20:09 +0200 02 tz RPOP:00000001: rpop connection ended, status: 3;Local fo
02-07 16:20:09 +0200 02 tz RPOP:00000002: rpop connection ended, status: 3;Local fo
02-07 16:29:01 +0200 08 tz WEBMAIL:00000003: [192.168.8.179:8000] connection accept
02-07 16:29:01 +0200 08 tz WEBMAIL:00000004: [192.168.8.179:8000] connection accept
02-07 16:29:12 +0200 08 tz WEBMAIL:00000003: Account 'laura.white@mycompany.com' ha
02-07 16:29:16 +0200 08 tz WEBMAIL:00000003: connection closed with [192.168.8.167:
02-07 16:29:16 +0200 08 tz WEBMAIL:00000005: [192.168.8.179:8000] connection accept
02-07 16:29:16 +0200 08 tz WEBMAIL:00000004: connection closed with [192.168.8.167:
```

Obr. 9 *Příklad struktury souboru typu log*

(Zdroj: https://www.axigen.com/docs/60/View-Log-Files_387.html)

B. *Přenos dat*

V momentě, kdy hrubá data nashromáždíme, je musíme přenést na infrastrukturu datového úložiště za účelem následné analýzy. Proces přenosu můžeme v základu rozdělit do dvou úrovní – přenos po páteřní síti a přenos v rámci datového centra.

Páteřní sít poskytuje na lokální úrovni, popř. na úrovni internetu vysokorychlostní přenosové médium k přemístění velkým datových sad z místa jejich vzniku do datových center. Rychlost přenosu je závislá na samotném fyzickém médiu a metodách managementu propojení. Fyzická média jsou většinou složená z mnoha optických vláken svázaných dohromady pro zvýšení kapacity. Management propojení navíc řeší způsob, jakým bude signál přenášen skrz fyzické médium. V této oblasti se nejčastěji používá princip vlnového multiplexu (WDM). Jedná se o technologii, která multiplexuje větší množství optických nosných signálů v jediném optickém vlákně použitím různých vlnových délek laserového paprsku. Pomocí této technologie je v dnešní době běžně dosahováno přenosových rychlostí okolo 40 Gb/s. V blízké budoucnosti se však počítá se zrychlením až do řádu Tb/s.

Poté, co data dorazí po páteřní síti do datového centra, jsou dále distribuována v jeho rámci za účelem přizpůsobení jejich umístění, zpracování, apod. Tento proces se nazývá přenos v datovém centru a vždy se pojí s architekturou sítě datového centra a transportním protokolem. Jejich přesné parametry však přesahují rámec této práce a případným zájemcům o tuto problematiku doporučuji studium příslušné literatury [10].

C. *Předzpracování dat*

Z důvodu rozličnosti datových zdrojů mohou mít sesbírané datové množiny různé úrovně kvality ve smyslu šumu, redundance, konzistence, apod. Přenášení a ukládání hrubých dat by s sebou neslo zbytečné náklady. Na straně poptávky po datech pak může stát i konkrétní požadavek na úroveň jejich kvality. Z těchto důvodů byly navrženy techniky předzpracování dat k zlepšení jejich kvality.

i. Integrace. Techniky integrace dat mají za úkol sjednotit data nacházející se v různých datových zdrojích a poskytnout nám unifikovaný pohled. Tato technika původně vychází z tradičních databází a skládá se ze dvou různých přístupů – metody datového skladu a metody virtualizace dat.

Metoda datového skladu se často schovaná pod zkratkou ETL neboli extraction (extrahování), transformation (přeměna) a loading (načtení). Extrahování představuje připojení k zdroji informací a sběr nezbytných dat pro zpracování analýzy. Transformace znamená aplikaci řady pravidel na extrahovaná data za účelem jejich převedení do standartního formátu.

Konečně fáze načtení znamená importování extrahovaných a transformovaných data do infrastruktury datového úložiště.

Pomocí metody virtualizace dat vytváříme virtuální databázi pro dotazování a agregaci dat z různých datových zdrojů. Virtuální databáze neobsahuje samotná data, ale pouze metadata o těchto datech a jejich umístění.

ii. Čištění. Jedná se o proces identifikace nepřesných, nekompletních a bezpředmětných dat a poté jejich doplnění nebo odstranění za účelem zvýšení kvality datové sady. Proces čištění dat se skládá z následujících kroků: definování a určení druhů chyb, vyhledávání a identifikace chybných dat, opravy chyb, dokumentace chybných dat a druhů chyb a modifikace procedury vstupu dat do systému k snížení chyb v budoucnosti. Čištění dat je vnímáno jako základ pro udržování konzistence a aktuálnosti dat. Zároveň je nezbytné pro následnou analýzu, protože zlepšuje její přesnost. Jelikož však závisí na komplexním vztahovém modelu, nese s sebou výpočetní náročnost a generuje zpoždění v systému. Proto musí být vždy ustanovena rovnováha mezi výsledným zlepšením přesnosti analýzy a komplexností čisticího modelu.

iii. Eliminace redundance. Datové sady trpí problémem opakování a nadbytečnosti dat, který nazýváme redundance. Redundance zbytečně navyšuje nároky na přenos dat a způsobuje nevýhody pro datová úložiště, jako plýtvání úložným prostorem, nekonzistenci dat, sníženou spolehlivost a poškození dat. I snaha o snížení redundance s sebou nese zvýšení nároků na výpočetní výkon.

## 3.3 Uložení dat

Subsystém datového úložiště Big Data platformy organizuje sesbírané informace ve vhodném formátu za účelem analýzy a extrahování hodnoty. Z tohoto důvodu subsystém datového úložiště musí poskytovat tyto dvě služby: musí poskytovat trvalé a spolehlivé uložení dat a zabezpečit škálovatelné přístupové rozhraní pro dotazování a analýzu obrovského množství dat. Z toho vyplývá, že subsystém datového úložiště můžeme rozdělit na úložnou infrastrukturu a data management.

### *3.3.1 Úložná infrastruktura*

Úložná infrastruktura slouží k fyzickému uskladnění nashromážděných dat a může na ni být nahlíženo z několika perspektiv. S ohledem na technologie ji můžeme dělit na paměť typu RAM (nestálý typ paměti, ze kterého se data smažou při odpojení od zdroje elektrické energie), magnetické disky a jejich shluky (pevné disky, které jsou primárními komponentami moderních úložných systémů) a nemechanická paměťová média (paměťová média typu flash disk, která neobsahují žádné mechanické komponenty). Rozdílné výkonnostní metriky těchto zařízení mohou být využity pro stavbu škálovatelného a vysoce výkonného subsystému pro ukládání Big Dat.

Na úložnou infrastrukturu může být nahlíženo i z pohledu síťové architektury. Z tohoto pohledu ji dělíme na úložiště přímo připojená k serverům (Direct Attached Storage – DAS), datová úložiště na síti (Network Attached Storage – NAS) a síť externích zařízení připojených k serverům (Storage Area Network – SAN). Pro Big Data aplikace se jako nejvhodnější jeví síťová architektura typu DAS, ve které jsou datová úložiště přímo napojená na servery bez nutnosti mezilehlé datové sítě, dodávající úlohám nadbytečnou komplikovanost.

### *3.3.2 Data management*

Data management se obecně zabývá otázkou vhodného způsobu organizace informací s ohledem na efektivitu jejich zpracování. Můžeme ho rozdělit na tři základní vrstvy: Souborový systém, databázové technologie a programovací modely.

Souborový systém

Pevný disk se na základní úrovni skládá ze sektorů očíslovaných 0, 1, atd. Bez dalšího zásahu by každý z nich představoval jeden velký shluk dat. Operační systémy proto vytváří systém adresářů dělících úložný prostor na jednotlivé soubory, kterým přiřazuje jméno. Zároveň spravuje volný úložný prostor, který nechává k dispozici pro nově vzniklé soubory [30]. Každý dokument, prezentace, obrázek, hudba, video, databáze nebo e-mail jsou tedy samostatným souborem, který na úložném zařízení zabírá určité množství místa a dochází tak k fragmentaci úložného zařízení [30][31]. Adresářovou strukturu a metody organizace a dělení nazýváme souborovým systémem [30].

V oblasti Big Dat jsou využívány takzvané distribuované souborové systémy, mezi které řadíme především Google File System (GFS) a jeho open-source derivát Hadoop Distributed File System (HDFS). Jedná se o škálovatelné souborové systémy (jejich kapacita může být jednoduše navýšena připojením dalšího běžně dostupného hardwaru) navržené pro vysoce distribuované datové aplikace [10]. GFS nejprve velké datové sady rozdělí do menších částí o velikosti 64MB a označených unikátním identifikátorem, aby tak snížil nároky na přenos souborů mezi různými lokalitami, kde jsou tyto části uloženy. Počítače jsou organizovány do clusterů, z nichž se každý může skládat ze stovek až tisíců počítačů. Zajímavý pak může být především fakt, že Google v tomto ohledu nevyužívá sofistikované technologie budoucnosti, ale běžně dostupný hardware. Počítače v rámci clusteru plní různé funkce: funkce hlavního serveru, koordinujícího aktivity v rámci clusteru; klientů, manipulujících s daty; a úložišť skladujících 64 megabytové části, z nichž je každá replikována, aby se v systému nacházela alespoň třikrát a na různých počítačích, čímž je dosaženo potřebné redundance [32].

Databázové technologie

Databáze použitelné v oblasti Big Data musí splňovat několik základních podmínek, mezi které patří podpora různých datových formátů a jednoduché replikace,

jednoduché prostředí pro programování aplikací, konzistence a podpora velkých datových sad [10]. Již dříve bylo naznačeno, že tradiční databázové systémy jsou pro Big Data nevhodné, protože vyžadují, aby data byla strukturovaná a bez chyb. Nestačí proto data jednoduše jen uložit, ale musíme je navíc rozdělit na jednotlivé záznamy ve formě polí. Indexy takových databází jsou navíc definovány předem a nelze do nich ukládat jiná data, než pro která byly určeny.

Standardem se pomalu ale jistě stávají tzv. NoSQL databáze. Mezi jejich nesporné výhody oproti konvenčním databázím patří podpora velkých objemů strukturovaných, částečně strukturovaných i nestrukturovaných dat, rychlá iterace, podpora objektově orientovaného programování a efektivní škálovatelná architektura založená na běžně dostupném nenákladném hardwaru.

Nejblíže k tradičním databázím mají NoSQL databáze typu „key-value". V tradiční databázi řádky reprezentují klíče (např. „B738" reprezentující typ letadla), zatím co každý sloupec obsahuje konkrétní vlastnost přiřazenou klíči (rozpětí křídel, výkonnost, kapacita,…). Zadám-li pak příslušný klíč do vyhledávání, zobrazí se mi všechny jeho atributy uložené ve sloupcích. V NoSQL databázi typu „key-value" najdeme sloupce pouze dva. Jeden obsahující klíč a druhý shluk veškerých atributů. Dále existují dokumentové databáze, které přiřazují klíče komplexním datovým strukturám ve formě dokumentů bez předdefinované struktury a podporují aplikace, které vzhledem ke své komplexnosti nemohou být řešeny pomocí „key-value" databází. Pro dotazování velkých datových sad byly vyvinuty sloupcově orientované databáze, které hodnoty ukládají ve sloupcích místo řádků. Posledním typem NoSQL databází jsou Grafové databáze, které ukládají informace o sítích, jako např. síť přátel uživatele Facebooku [10][33].

Programovací modely

Přes všechny početné výhody oproti relačním databázím s sebou NoSQL databáze nesou i řadu nevýhod. Mezi ty nejmarkantnější patří omezená podpora dotazování a analytických operací. Programovací modely snižují výkonnostní propast mezi relačními a NoSQL databázemi tím, že nám umožňují provádět operace nad velkými množinami dat [10]. MapReduce je nejpoužívanějším programovacím modelem pro jednoduchý vývoj aplikací, které paralelně zpracovávají velké množství dat na velkém clusteru (tisíce počítačů) běžného hardwaru spolehlivým způsobem [34]. Uživatel pomocí něho definuje dvě funkce nazvané `map()` a `reduce()`. Funkce `map()` vezme vstupní hodnotu, vrátí ji zpět spolu s klíčem ve formě key:value páru a sdruží hodnoty dohromady podle stejného klíče. Funkce `reduce()` vezme výstup funkce `map()` a zredukuje ho na požadovanou výslednou hodnotu. Jelikož funkce `map()` k určení výstupní hodnoty potřebuje pouze vstupní hodnotu, může být aplikována paralelně na různé hodnoty [35]. Rámec MapReduce a HDFS běží na stejném setu počítačů, což nám umožňuje provádět operace s daty přímo v místě,

kde jsou uložena. MapReduce tedy rozkládá úkoly na menší části, které jsou paralelně řešeny na různých serverech a umožňuje tak zpracovávání neomezeně velkého množství dat [36].

Mějme například na úložištích v daném počítačovém clusteru uloženou datovou sadu obsahující informace o odečtech senzorů měřících teplotu výstupních plynů motoru. Ta je rozkouskována mezi jednotlivými uzly tohoto clusteru. Pokud chceme zjistit, kolikrát teplota výstupních plynů dosáhla 500°C, zadáme prostřednictvím klientského počítače hlavnímu serveru clusteru příkaz, aby aplikoval na datovou sadu funkci `map()`, která je definovaná tak, aby pokaždé, když narazí na hodnotu 500°C, vrátila key:value pár 500:1, a na konci procesu funkce `map()` key:value pár se sdruženými hodnotami 500:[1,1,1,…,1]. Tento výstup použijeme jako vstup do funkce `reduce()` a jako výsledek dostaneme počet událostí, kdy teplota výstupních plynů motoru dosáhla 500°C.

## 3.4 Analýza dat

Posledním a nejdůležitějším článkem hodnotového řetězce Big Dat je analýza dat, jejímž cílem je extrahování cenných hodnot, návrh závěrů nebo podpora rozhodování. Analýza dat se zabývá informacemi získanými pozorováním, měřením nebo experimentováním s fenoménem, který je předmětem zájmu. Jejím cílem je extrahování co největšího množství informací relevantních uvažovanému předmětu zájmu, jehož podstata se může případ od případu výrazně lišit. Stejně tak se může lišit i účel analýzy. Mezi důvody k provedení analýzy patří např.:

- Extrapolace a interpretace dat a určení jejich využití
- Ověření správnosti dat
- Poskytování rad a podpora rozhodování
- Diagnóza a vyšetření důvodu vzniku chyby a
- Predikce budoucího vývoje

Z hlediska hloubky analýzy můžeme analýzu dat dělit na 3 podskupiny: deskriptivní analýzu, prediktivní analýzu a preskriptivní analýzu.

Deskriptivní analýza se zabývá historickými daty a snaží se zjistit, co se stalo. Například může být provedena regrese k nalezení určitých trendů v datových sadách a použita vizualizace reprezentující data snadno pochopitelným způsobem. Deskriptivní analýza se většinou pojí s oblastí Business Intelligence, tedy snahou pochopit minulý vývoj obchodní společnosti.

Prediktivní analýza se zaměřuje na předpověď budoucích pravděpodobností a trendů. Prediktivní modelování například využívá statistických technik jako lineární a logistické regrese k pochopení trendu a predikci budoucích důsledků, a data mining k extrakci skrytých vzorců, které poskytují vhled do situace a předpovědi.

Preskriptivní analýza se zabývá rozhodováním a efektivitou. Kupříkladu se využívá simulace k analýze komplexních systémů za účelem dosažení náhledu na chování systému a identifikace problémů. Využity mohou být i optimalizační techniky k nalezení optimálního řešení za daných omezení.


Obr. 10 *Fáze analýzy dat* [39]

## 3.5 Integrace Big Data technologií se stávají architekturou

Letecké společnosti mají v zásadě 4 možnosti integrace Big Data technologií se současným hardwarem, softwarem a databázemi k vytvoření nákladově efektivní Big Data platformy:

- Datová „jezera“
  Představme si, že letecká společnost působí v několika různých zemích (Např. v případě společnosti Travel Service, a. s. to jsou Česká republika, Slovensko, Polsko a Maďarsko) a chce provést analýzu zákaznických dat z těchto různých zemí. Pro tento účel se jeví jako účelné vytvoření tzv. datového jezera „rozbitím“ jednotlivých úložišť a převedením dat na jediné místo. To přinese především možnost čilého reagování na obchodní výzvy. Prozkoumávání dat se tím značně zjednoduší, umožní se rozhodování v reálném čase a otevře se také možnost minimalizace nákladů na vlastnění IT infrastruktury.
- Rozšíření datového skladu
  Tato možnost je vhodná v případě, že má aerolinka zájem na manipulaci s velkým množství dat rozličného formátu, ale zároveň většinu rozhodnutí zakládá na datech ze stávajícího datového úložiště, které spravuje již mnoho let. Stávající infrastrukturu je tak potřeba rozšířit o Big Data platformu, která však současný stav nijak negativně nenaruší. Většina leteckých

společností zabývajících se Big Daty volí právě tuto variantu z důvodu její schopnosti výrazně zlepšit provozní efektivitu.

- Přístupné archivy

  Přístupné archivy jsou využívány jak pro analytické účely, tak i kvůli nutnosti souladu z různými regulacemi ohledně skladování dat. Některá nařízení vyžadují, aby data byla aktivně skladována po dobu pěti, deseti a někdy až dvaceti let pro případ nutnosti vykázat postupování v souladu s předpisy. Taková data jsou pak skladována v archivech, jejichž provoz je extrémně nákladný, neboť musí zároveň poskytovat možnost aktivního vyhledávání v datech, která pro nás už sama o sobě nemusí představovat velkou hodnotu. Big Data platforma nám umožňuje, aby data nebyla aktivně indexována, a pracuje pouze s meta-daty těchto dat, čímž minimalizuje náklady na jejich skladování.

- Modernizace platformy

  Mnohé tradiční systémy jako např. ticketing, nástroj cenotvorby, jsou založené na téměř 50 let staré IT architektuře. S tím se pojí značné náklady na informační infrastrukturu a zároveň se stabilita systému snižuje s nárůstem dat a rozsahem analýzy, kterého se dnes aerolinky snaží dosáhnout. Z toho důvodu bude dříve či později nutné provést modernizaci platformy. Big Data jsou vysoce moderním nástrojem k dosažení analytických potřeb leteckých společností, kterých současné platformy nejsou schopné.

# 4 Big Data v letecké dopravě

Data byla v odvětví cestovního ruchu vždy vnímána jako důležité aktivum a společnosti podnikající v oblasti cestovního ruchu stanuly mezi prvními, které data dokázaly využít k obchodním účelům. Letecké společnosti se staly průkopníky v oblasti analýzy cenové optimalizace, která byla úspěšně přejata i v hotelnictví, a pomocí dat také optimalizovaly detaily plánování posádek a tratí. Odvětví letecké dopravy bylo také jedním z prvních, kdo začal využívat výhody věrnostních programů [37].

Všechny společnosti podnikající v oblasti letecké dopravy v historii produkovaly velké množství dat a toto množství neustále roste. Podle ženevské organizace Air Transport Action Group (ATAG) je na celém světě denně přepraveno 8,6 milionu cestujících prostřednictvím téměř 100 tisíc letů, což ročně představuje asi 35 milionů letů, během kterých vznikají biliony datových bodů [38]. Každá činnost spojená s provedením jednotlivých letů za sebou nechává datovou stopu. V souhrnu jsou tak vytvářeny stovky terabytů nebo petabytů strukturovaných transakčních dat v konvenčních databázích a neméně velké množství dat nestrukturovaných [37]. Velká letecká společnost denně vyprodukuje více než 3 exabyty dat a toto číslo se zdvojnásobuje každé 3 až 4 roky. V každý okamžik tak velká letecká společnost musí být schopná uchovat víc než jeden terabyte dat [39]. Globální distribuční systém společnosti Sabre dokáže během jednoho dne vyprodukovat až 7 terabytů transakčních dat [40]. Stejný scénář se odehrává i na straně výrobců letadel. Např. se odhaduje, že společnost Boeing disponuje asi 100 petabyty dat, ve kterých má v úmyslu najít korelace za účelem poskytování lepších služeb leteckým společnostem a potažmo cestujícím [41]. Samotná letadla dnes v závislosti na vysoké míře instrumentace mohou produkovat na průměrném letu 500 až 1000 gigabytů dat [42].

Přecenit transformační potenciál Big Dat v letecké dopravě je téměř nemožné. Big Data jsou v současnosti pravděpodobně největší příležitostí pro subjekty podnikající v oblasti letecké dopravy k využití měnící se struktury dat a maximalizaci jejich užitku. Nesou v sobě potenciál zefektivnění obchodu a zároveň i zkvalitnění zážitku z cestování.

Společnosti v letecké dopravě dnes stojí na důležité křižovatce vytyčené Big Daty, která hrají důležitou roli ve vytváření efektivnějšího a personalizovaného zážitku přinášejícího prospěch pro jak letecké společnosti, tak i cestující. Z potenciálu Big Dat však zatím těží pouze hrstka společností. Jiná odvětví již oblast cestovního ruchu ve využití dat k analytickým účelům předběhla. Vyjma několika málo online firem poskytujících služby v letectví se kupříkladu pouze několik málo společností zaměřuje na poskytování cíleného a individualizovaného cestovatelského zážitku. Big data představují možnost transformace odvětví letecké dopravy a první společnosti již podnikají kroky k jejich brzkému využití k podpoře firemních procesů. Big Data se mohou stát nejvlivnější inovací od doby zavedení online rezervačních systémů.

Předpokládá se, že Big Data mohou v oblasti letecké dopravy přispět k podpoře rozhodovacího procesu, vzniku nových produktů a služeb, zlepšení vztahů s cestujícími a zlevnění a zrychlení procesů. Průkopníci mezi leteckými společnosti již dnes Big Data využívají k optimalizaci Revenue Managementu, podpoře distribučních systémů, optimalizaci vnitřních procesů a stimulaci finanční výkonnosti.

V této kapitole se zaměříme na oblast personalizace produktu, prediktivní údržbu a konečně podporu rozhodovacího procesu provozního dispečinku letecké společnosti Travel Service pomocí Big Dat.

## 4.1 Personalizace

Odvětví, která produkují a ukládají velké objemy zákaznických dat, mají největší příležitost vytvořit relevantní obchodní hodnotu pomocí agregace a analýzy těchto dat a patřičné reakce na její výstupy. Z tohoto pohledu odvětví letecké dopravy velmi pravděpodobně disponuje velkou příležitostí v oblasti personalizace a obchodu, dost možná větší příležitostí, než jakou dnes v tomto ohledu mají ostatní odvětví. Letecké společnosti sedí na terabytech vlastních zákaznických dat obsahujících detaily o nákupních návycích cestujících, historická data z odbavovacího procesu, preferencích palubních služeb, pohybu cestujících v prostorách letišť, interakcích se zákaznickým centrem, a další mnohé podkategorie dat. S tím, jak se zákazníci setkávají s personalizací produktů v jiných odvětvích, jako např. v bankovnictví, maloobchodu nebo stravování, očekávají to stejné od společností podnikajících v oblasti cestovního ruchu a tento standart budou vyžadovat už velmi brzo.

Tradiční styčná místa pro cestující s leteckými společnostmi představují už dlouhá léta cestovní agentury, palubní průvodčí a pracovníci zákaznických center. V dnešní době ale nosí cestující svůj osobní komunikační prostředek přímo u sebe – chytrý telefon. Nějaké mobilní zařízení dnes u sebe nosí 97 % cestujících, a z těchto zařízení 81 % představují chytré telefony. Tento komunikační prostředek se stává zcela běžnou záležitostí a dává nám příležitost k shromažďování dat v průběhu všech fází interakce s cestujícími a získání úplného obrázku o nich. Kromě toho, že o cestujícím získáme představu v rozsahu 360 stupňů, jsme navíc schopní vybavit zaměstnance v předních liniích styku s cestujícími správnými informacemi pro poskytnutí optimálního zážitku z cestování. Samotní cestující jsou dokonce ochotní poskytovat svá data leteckým společnostem. 72 % z nich potvrdilo, že jsou ochotní sdílet s aerolinkou svá geolokační a osobní data v případě, že to bude mít pozitivní dopad na jejich zážitek z cestování.

Prodej vedlejších produktů leteckým společnostem generuje globálně zisk ve výši 30 miliard amerických dolarů ročně a tato hodnota se meziročně zvyšuje. Jelikož tento prodej může tvořit u některých společností až 40 % celkového zisku, nedá se na něj již déle nahlížet jako na pouhou vedlejší obchodní činnost. Úspěšný obchodní model můžeme definovat jako poskytování správného produktu, prezentovaného správnému

zákazníkovi ve správný čas. Letecká společnost by v ideálním případě měla být schopná personalizace nabídky a balíčků služeb skrz globální distribuční systémy, stejně tak jako i prostřednictvím webu a mobilní platformy. Personalizací vedlejších služeb může zisk leteckých společností z vedlejší činnosti vzrůst o zhruba 22 %. Pro středně velkého dopravce to může v horizontu pěti let představovat nárůst zisku v celkové hodnotě 163 milionů amerických dolarů. Jedná se o významnou obchodní příležitost, kterou vedení leteckých společností neberou na lehkou váhu.

Klíčem k úspěchu je správné přiřazení produktů a služeb k cestujícímu, který z nich bude mít optimální užitek. K tomu je nutné, aby letecké společnosti analyzovaly svá data o každém z jejich zákazníků, a uplatnily systémy automatizovaných pravidel pro poskytování personalizovaných nabídek a proaktivního zákaznického servisu. Koncept personalizace znamená schopnost reagovat na specifické požadavky každého jednoho cestujícího.

V rámci společností, jejichž business model předpokládá vysokou četnost styku se zákazníkem, letecké společnosti zaostávají ve schopnosti vyvozovat závěry z dat. Množství zákaznických dat přístupných aerolinkám v posledních pěti letech vykazuje exponenciální nárůst. Letecké společnosti se proto snaží odvrátit od tradičního modelu založeného na transakčních systémech a zavést modely integrující různé druhy dat včetně těch, které poskytují informace o aktivitě cestujících na sociálních sítích. Konkurenční boj se začíná zaměřovat na udržení stávajících cestujících generujících nejvyšší zisk a zvyšování podílu na trhu prostřednictvím personalizace produktů. Dopravci neschopní obohatit svůj business model o schopnost hledání souvislostí v datech mohou velmi brzy přijít o konkurenční výhodu.

Letecké společnosti neustále hledají způsob, kterým by se odlišily od konkurence. Rychlost, jakou jsou schopné přeměnit masivní množství dat na hodnotné vhledy a dále tyto vhledy do následných činností, může znamenat významnou odlišnost od konkurence. V současné době se již na lineární přístup k datům nedá spoléhat – není vhodné strávit šest měsíců sběrem dat, dalších několik měsíců jejich analýzou a ještě déle vyvozováním důsledků z nich. Data si vyžadují čilost a schopnost reakce v řádu minut a hodin, nikoliv týdnů.

Díky přechodu od diagnostické datové analýzy zaměřující se na historická data k prediktivnímu nebo dokonce preskriptivnímu modelu, se před leteckými společnosti otevírá možnost vzniku konkurenční výhody zvýšením provozní rychlosti založeným na personalizaci cestovního zážitku v reálném čase.

Metody životního cyklu zákazníka a segmentace trhu se pomalu ale jistě stávají zastaralými pozůstatky éry nedostatku dat a technologií potřebných k pochopení komplexních zákaznických postojů, chování a návyků na individuální úrovni. Zákaznickou loajalitu je možné získat uchopením a personalizací nabízených linek,

obchodního partnerství, cateringu, zábavy, zákaznického servisu, vedlejších produktů, technologických novinek atd. Letecké společnosti, které předčí očekávání cestujících prostřednictvím těchto příležitostí, nabydou schopnost rychlé absorpce, interpretace, personalizace a poskytování konzistentního produktu, který si zákazníci zapamatují, budou za něj ochotní zaplatit a v budoucnu jej mnohokrát znovu využijí [39].

*Nároky na informační infrastrukturu*

První firmy, které se pustily do implementace analýzy masivního objemu nestrukturovaných a rychle vznikajících dat, byly internetové firmy. Dá se říct, že společnosti jako Google, eBay, LinkedIn a Facebook byly vystavěné kolem konceptu Big Dat od samého začátku. Jelikož disponovaly masivními objemy dat v novém a méně strukturovaném formátu, jako například clickstream data, soubory webových serverů typu log, vztahy na sociálních sítích a výsledky řízených experimentů, neměly jinou možnost než zavést nové technologie a manažerské přístupy [37].

Ve světě dat se letecké společnosti potýkají s výzvami, které mnohá jiná odvětví řešit nemusí. Základní problém vychází z toho, že zákaznická data leteckých společností rostou zhruba od poloviny minulého století. Odvětví letecké dopravy vytvořilo systémy a procesy, nejprve v analogové a později digitální éře, k zákaznické podpoře mnohem dříve, než své služby začal nabízet jakýkoliv první internetový obchod. Z tohoto důvodu leží v technologických základech leteckých dopravců, kteří v odvětví podnikají již desítky let, původní tradiční systémy jako ticketing, databáze záznamů z cestovních kupónů a systémy zpracovávající zprávy z procesu odbavení, na kterých se v průběhu času stavělo. Letecká doprava je odvětvím, ve kterém jsou na produkty uvalené jen velmi nízké marže, což často zabraňuje rozsáhlým investicím do techniky. Manažeři leteckých společností si již dávno uvědomili, že jejich zákazníci jsou velmi citliví na cenu a právě proto se rozhodli zaměřit na snižování nákladů, minimalizmus a zvyšování efektivity. Investice do technologické infrastruktury a modernizace datových technologií tím byly značně zpomaleny. Můžeme to dokázat například tím, že velkým krokem kupředu ve zpracování dat leteckých společností bylo představení Airline Control Programu společnosti IBM mezi lety 1960 a 1970, který je do dnešní doby ještě stále součástí operačních systémů některých leteckých společností. S tím, jak se vyvíjí zákaznická data, však bude potřebná adopce nových technologií na zpracování dat, které umožní personalizaci v reálném čase [39].

Dalším nepříjemným důsledkem původní datové architektury je fragmentace relevantních klíčových dat napříč mnohými rozličnými odděleními společnosti [37]. V ohledu využití zákaznických dat jsou letecké společnosti v přirozené nevýhodě. Data jsou uložena v izolovaných úložištích, což s sebou nese další vrstvu komplexnosti ve snaze unifikace a normalizace dat napříč systémem [39]. Například data letecké společnosti popisující zážitek cestujících z letu, jsou roztroušena mezi provozní úsek, databáze zavazadel, věrnostních programů a stížností a vnější zdroje, ze kterých

můžeme jmenovat např. sociální média. Abychom moli efektivně rozhodovat o tom, jakým způsobem budeme podporovat a nabízet naše produkty cestujícím a napravovat chyby našich služeb, musíme všechny informace agregovat do jediného úložiště a aplikovat na ně stejnou sadu algoritmů. S tím se pochopitelně váže i patřičná investice [37]. Návratnost investice do technologií pro zpracování dat se však dnes jeví mnohem pozitivněji než dřív. Jako příklad může uvést významné snížení nákladů na úložnou jednotku. Ty se snížily ze zhruba 18,95 USD v roce 2005 na asi 0,66 USD za gigabyte v roce 2015. Jedná se o 96 procentní snížení v horizontu posledních deseti let. Podle průzkumu společnosti IBM zaznamenaly společnosti, které implementovaly obchodní strategii založenou na analýze dat, průměrně o 49 % vyšší růst zisku než jejich konkurenti, kteří tak neučinili [39].

Vytvoření integrovaného zdroje zákaznických dat však není jenom nákladné, ale všeobecně náročné bez ohledu na to, jak velký rozpočet máme k dispozici. Jednak narážíme na problematiku ochrany soukromí cestujících a také na to, že jeden cestujících může v různých systémech vystupovat pod velkým množstvím různých identit. Proto je velmi těžké kombinovat data ze sociálních sítí s daty z vnitřních transakčních systémů.

Dalším výsledkem dlouhodobého extenzivního využívání informačních systémů u velkých tradičních dopravců je, že architektury Big Data technologií budou muset koexistovat spolu s již existujícím hardwarem, softwarem a databázemi, protože tradiční nástroje a data, která obsahují, jsou stále potřebné a budou dále užitečné k analýze a zlepšování provozu a vztahů s cestujícími. Technologie pro Big Data ve své čisté podobě jsou vhodným řešením pro nově založené společnosti a čistě internetové firmy, ale u stávajících velkých společností lze v blízké budoucnosti očekávat hybridní prostředí, což povede k výzvám v oblasti architekturní koheze informačních technologií a efektivní funkčnosti jak nových, tak i starých systémů. Dříve zmíněné datové technologie společnosti IBM z 60. let 20. století navíc nejsou schopné podporovat open-source prostředí pro analýzu Big Dat, jako např. Hadoop, což je komerčně využitelný softwarový rámec popsaný v kapitole o Big Data technologiích, včetně distribuovaného souborového systému (HDFS) a platformy MapReduce [37].

Moderní transakční a provozní systémy pomalu nahrazují tradiční systémy, ale manažeři leteckých společnosti o transakční data přestávají jevit zájem. Letecké společnosti na čele trhu ční nad svými konkurenty především díky analýze zákaznických dat a znalosti svých cestujících. Vyvíjející se objem, struktura a další atributy dat, která s aerolinkami cestující sdílí, však způsobují nesnáze. Správa a integrace dat z různých datových zdrojů je globálně největší výzvou pro všechny profesionály v oblasti zákaznické analýzy. Letecké společnosti se pokoušely o agregaci dat po desítky let a doufaly, že se jim podaří vytvořit jediný zdroj absolutní pravdy o jejich cestujících. Vzhledem k tomu, že data jsou ve většině společností roztříštěna mezi 20 různých

datových úložišť a u velkých dopravců dokonce 50 nebo víc, představuje vytvoření jediného zevrubného pohledu na zákazníka vysoce komplexní a nákladný projekt. Některé společnosti zaznamenaly úspěchy vytvořením jediného systému správy zákaznických dat, ale mnohdy tak vniklo pouze další izolované datové úložiště vyžadující pravidelnou re-kalibraci, čímž docházelo k větší desorganizaci, než jaká panovala bez něho.

*Profil zákazníka*

Mnozí obchodníci mají velké štěstí, pokud pro ně existuje jeden nebo dva styčné body se zákazníky. Naproti tomu letecké společnosti mají privilegium pozorovat zákazníky v různých fázích a po prodlouženou dobu na jakési imaginární zákaznické cestě hodnototvorným procesem letecké společnosti. Styčné body nepředstavují pouze unikátní příležitosti k poskytnutí zákaznických služeb, ale také i podstatnou příležitost ke sběru dat, ze kterých můžeme vyčíst mnohé o jednotlivých zákaznících na individuální úrovni. Důležitým prostředím pro komunikaci aerolinek s cestujícími představuje letiště. Roste zde jak množství obchodních příležitostí, tak i možnost technologické podpory. Internet věcí představuje nové možnosti obohacení zážitku cestujících z letišť a inteligentní zařízení neustále připojená k internetu se stanou dalším rozhraním pro sběr zákaznických dat. Všechny digitální styčné body s leteckou společností, včetně mobilních zařízení a internetu věcí, budou ve stavu neustálého generování nových dat. Pro účely personalizace tato data musí být nejprve normalizována a agregována.

V ideálním případě by měl tato data načíst a zpracovat centrální rezervační systém, který si tak vytvoří široký obrázek o cestujícím a jeho pohybu. Usnadnění personalizovaného dialogu s cestujícím v průběhu jeho cesty hodnotovým řetězcem letecké společnosti, vyžaduje bezchybně integrovanou technologii, která v hodnotovém řetězci pracuje v reálném čase. Program zákaznického zážitku založený na rozličných technologických řešeních vyžaduje ke spuštění nákladnou integraci a pravidelnou re-kalibraci pro dosažení aktuálnosti. Ideálním řešením pro tuto komplexní obchodní výzvu je bezešvé prostředí zákaznických dat, které se pojí se všemi systémy zapojenými v nákupu, rezervacích, odbavení a dalších krocích. Tato bezešvá technologie propojená s rezervačním systémem a obsahující zákaznické profily, by mohla sdružit všechny systémy produkující data a odsunout problémy s integrací dat do minulosti.


Obr. 11 *Cesta zákazníka napříč hodnototvorným procesem letecké společnosti* [39]

Pokoušení se o personalizaci cesty zákazníka bez obsáhlých dat by bylo pouhým automatizovaným věštěním z křišťálové koule. Po dlouhou dobu nebyla technologie schopná umožnit leteckým společnostem získat úplný náhled na své zákazníky. Kombinace transakčních, provozních, sociálních, behaviorálních a dalších mnoha zdrojů dat v kontextu reálného času se pro letecké společnosti stane klíčovým odlišujícím prvkem. Je nutné získat holistický náhled na zákazníka, nezbytný pro vytvoření integrovaného zážitku, což vyžaduje obecné chápání problematiky Big Dat napříč leteckou společností a kooperaci mezi jednotlivými úseky.

Profil zákazníka je klíčovým prvkem potřebným k realizaci obchodní strategie, v jejímž středu stojí zákazník. Mnohé letecké společnosti nejsou schopné identifikovat jednotlivé zákazníky, pokud nejsou členy jejich věrnostního programu. Koncept profilu zákazníka to umožňuje především díky tomu, že dává možnost vytvořit úplné profily všech cestujících. Zároveň neustále roste i vliv sociálních médií na preferenci určité značky a nákupní rozhodování. V případě jakékoliv interakce, kdy zákazník cokoliv veřejně prohlásí o letecké společnosti, by tato informace měla být zahrnuta do jeho zákaznického profilu, protože cokoliv, co o nějaké značce lidé říkají svým přátelům, má velký vliv na jejich budoucí nákupní rozhodnutí.

Profil zákazníka je ideálně pohled v rozsahu 360 stupňů obsahující vyčerpávající seznam atributů o zákazníkovi. Na jeho začátku figurují deskriptivní atributy jako kontaktní informace, demografická a psychografická segmentace, zájmy a uvedené preference. Důležitá je také hodnota daného zákazníka. Úplný a přesný pohled na dlouhodobou potenciální hodnotu by mohl být určen pomocí indikátorů jako celkový zisk, příspěvek k marži, historie nákupu vedlejších produktů a informace z věrnostních programů. K tomu navíc hraje svou roli i historie interakcí letecké

společnosti se zákazníkem. Od osobního kontaktu přes zákaznická centra až po příspěvky v sociálních médiích, pomáhá zpracování interakcí v reálném čase udržovat informační povědomí o aktuálním statusu a spokojenosti zákazníka a dovoluje tak ušít následující interakci na míru dané situaci.

Nákupní chování zákazníka je relevantní jak na individuální úrovni, tak i na úrovni celého trhu, kdy bereme v úvahu několik zákazníků a provádíme analýzu na základě vybraných atributů za účelem objevení nových trendů a příležitostí.

Zaměření se na obchodní strategií postavenou kolem zákazníka a založenou na jeho profilu, dovoluje leteckým společnostem vidět potřeby jednotlivých zákazníků při nabídce produktů a správě zisku. Pokud například prodej vedlejších produktů na lince provozované mezi dvěma konkrétními letišti, v nějakém regionu nebo celkové produktové kategorii klesá, může být spuštěn systém upozornění, která dokáží identifikovat základní příčinu poklesu dříve, než se rozšíří na úrovni trhu. Začne-li například na určité lince klesat poptávka po připojení k Wi-Fi do takové míry, že se změní historické hodnoty jejího prodeje, může být provedena analýza příčin tohoto trendu a navržena sada opatření proti jeho rozšíření na celý trh.

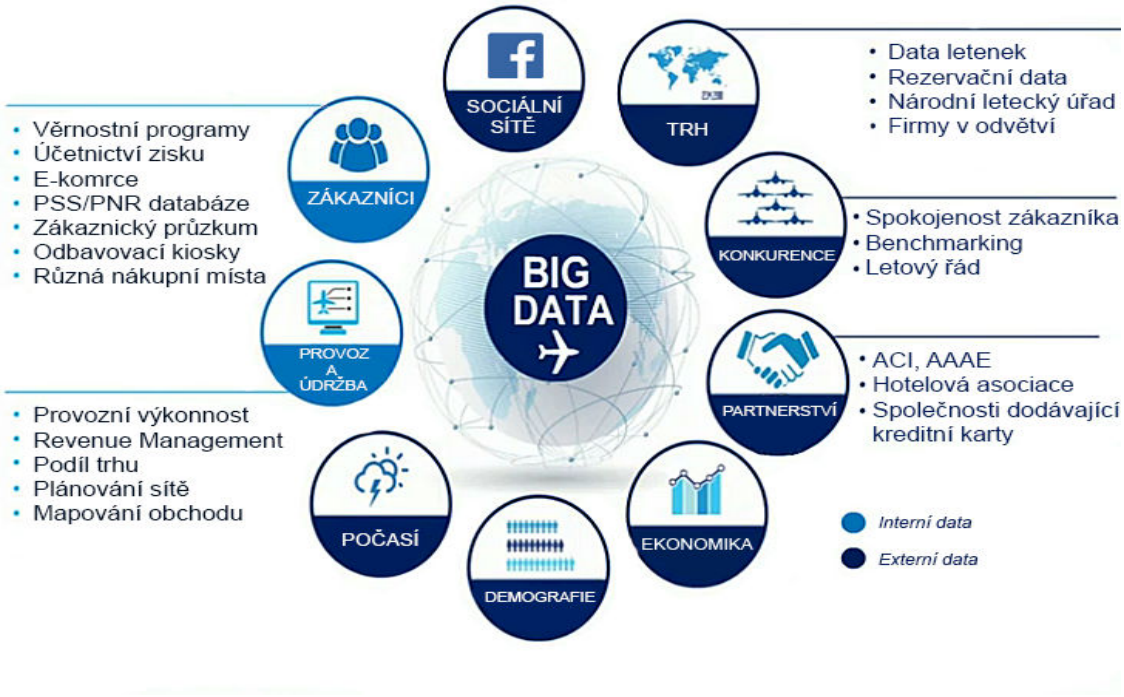
Nová generace obchodních technologií zaměřených na zákazníka v jejich středu bude založena na schopnosti ovlivnění nákupního rozhodování zákazníka a jeho budoucí zájmy s ohledem na jeho minulou činnost pomocí prediktivního odvozování. Využitím databáze zákaznických profilů a shlukováním zákazníků do skupin napříč jejich mnohými dynamickými atributy, mohou být algoritmy přizpůsobené tak, aby předpovídaly chování zákazníka a pomohly tak dosáhnout unikátních obchodních potřeb letecké společnosti [39].

*Personalizace v distribuci*

Distribuce byla po dlouhou dobu doménou, ve které se využívají počítače, data a analýza na nejvyšší úrovni. Proto se dá předpokládat, že v počáteční fázi budou mít Big Data největší dopad právě v oblasti distribuce. Některé distribuční nabídky jsou již do jisté míry personalizované, např. na základě věrnostního statusu nebo historie nákupního chování, ale většina nabídek dnes konkrétním požadavkům cestujících odpovídá jen velmi vzdáleně [37]. Naším cílem není zahltit zákazníka informacemi, ale poskytnout mu v čase nákupu právě ty informace, které by mu v té chvíli mohly připadat zajímavé. Big Data nám dávají skvělou možnost ve chvíli, kdy si zákazník prohlíží nás produkt, zjistit, co ho zajímá, a jak na jeho nákupní požadavek co nejlépe odpovědět. Zároveň je však třeba zdůraznit, že analýzu provádíme pouze za účelem poskytování relevantních informací a nikdy s úmyslem cenové diskriminace [43].

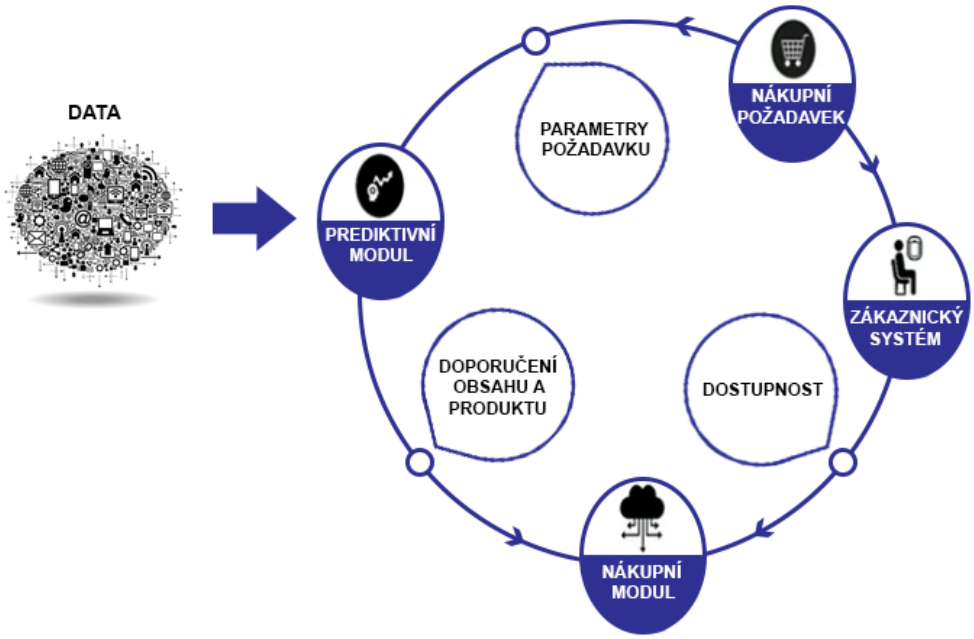
Prvním krokem pro personalizaci je identifikace relevantních datových množin, nad kterými provedeme Big Data analýzu za účelem nalezení skrytých trendů a korelací. V závislosti na cíli analýzy mohou být shromažďována interní a externí data

o zákaznících společnosti, jejím provozním stavu, data popisující situaci na trhu, srovnávací data o konkurenčních společnostech, ekonomická data, počasí, atd. V dalším kroku je nutné integrovat veškerá zákaznická data a na základě jejich Big Data analýzy identifikovat jednotlivé zákaznické profily, které se stanou základem pro budoucí behaviorální cílení.




Obr. 12 *Příklady zdrojů pro Big Data analýzu využitelných leteckou společností* [43]

Po identifikaci zákaznických profilů jsme již schopní zpracovávat nákupní požadavky na personalizované úrovni. Nákupní požadavek putuje v reálném čase jednak do zákaznického systému a jeho parametry také do prediktivního modulu. Zákaznický systém vrátí dostupnost požadavku, zatím co prediktivní modul určí, které produkty a služby by mohly být pro tohoto konkrétního zákazníka zajímavé. Tyty dvě informace se spojí v nákupním modulu a výsledek je zákazníkovi prezentován prostřednictvím distribučního systému, který používá [43].



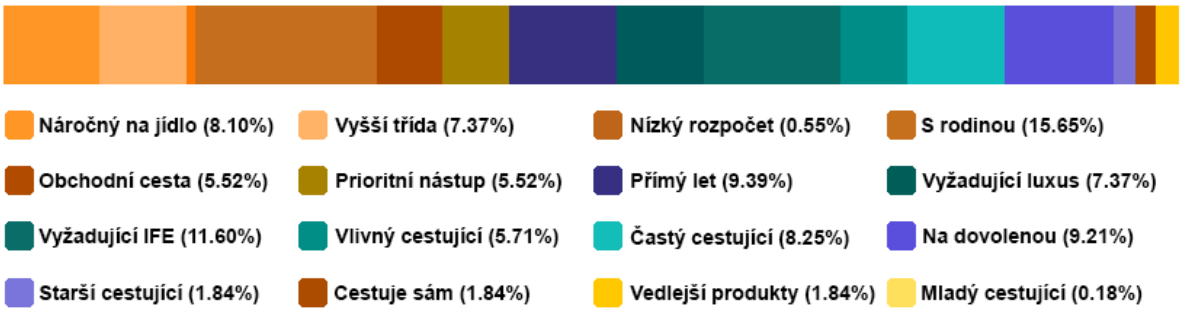

Obr. 13 *Proces vzniku personalizované nabídky z nákupního požadavku* [43]

Nejjednodušší případ, který může nastat, je obdržení nákupního požadavku od identifikovaného zákazníka. Takový proces je pro obchodníka letecké společnosti přímočarý a pro zákazníka lehce pozorovatelný. Identifikovaným zákazníkem může být například člen věrnostního programu nebo cestující zaregistrovaný na webové stránce letecké společnosti. Po přihlášení do uživatelského účtu dojde ke spárování se zákaznickým profilem a při zadání vyhledávacího požadavku je možné vidět, že výsledná nabídka koreluje především s historickými nákupy. Uživatel přitom může být vybízen k tomu, aby na sebe prozradil více detailů za účelem poskytnutí individuálnější nabídky.

Cílení na anonymního zákazníka naopak představuje poměrně velkou výzvu. Pro ilustraci si představme zákazníka, který na internetové stránce letecké společností hledá let do Paříže. Jeho aktivita naznačuje, že se chystá letět do Paříže, a do jeho webového prohlížeče je poslán tzv. soubor typu „cookie“. Tento soubor je schopný zmapovat aktivitu návštěvníka webové stránky, ale nemusí nutně hledat osobní informace a jeho životnost může být limitovaná do zavření okna prohlížeče. Základním účelem souboru cookie je následovat uživatele na jiné webové stránky (např. sociální sítě) a zobrazovat na nich personalizovanou reklamu. Tímto způsobem je nasbíráno obrovské množství dat, která jsou analyzována za účelem vytvoření tržních segmentů. Hypotetický uživatel může být např. identifikován jako častý cestující nebo jako cestující navštěvující rodinu v zahraničí, což jsou relativně obecké segmenty, ale z hlediska marketingu je i taková informace užitečná. Analýza však může být provedena opravdu zevrubně, čímž je možné identifikovat mikrosegmenty, jako např. cestující letící do New Yorku na konferenci o onemocněních kardiovaskulárního systému. To naznačuje, že taková aktivita vyžaduje umísťování trvalých souborů typu

cookie. Uživatel takto může být sledovat po delší dobu a setkávat se např. s reklamou na produkt, který vložil do nákupního košíku, ale poté internetovou stánku opustil. I přesto, že uživatel nemá založený žádný zákaznický účet, mu tedy může být nabízet vysoce personalizovaný produkt [44].

Přestože jsou sobory cookies vhodným nástrojem v oblasti marketingu, v globálních distribučních systémech se k personalizaci využívá už několikrát zmiňované prediktivní modelování. Po zadání nákupního dotazu anonymního zákazníka proběhne korelace parametrů vyhledávání s dříve vysledovanými trendy a náhledy z Big Data analýzy, a na konci vznikne zákaznický profil uživatele, na jehož základě jsou mu skrz distribuční systém prezentovány relevantní výsledky vyhledávání [45].




Obr. 14 *Příklad zákaznického profilu vzniklého z Big Data analýzy* [45]

Dalšího zvýšení personalizace produktu může být docíleno:

- Personalizací založenou na chování zákazníka nebo jeho absence („Moc nás mrzí, že jsme vás neviděli na naší lince do Paříže poté, co jsme se 12 týdnů v řadě těšili z vaší společnosti!“)
- Personalizací založenou na sociálních médiích („Několik z vašich přátel v poslední době navštívilo Řím. Nabízíme Vám 20 % slevu na naše lety, abyste se tam sám mohl podívat.“)
- Personalizace s ohledem na prodej vedlejších produktů („ Posíláme Vám kupón na bezplatnou návštěvu letištního salonku v Praze při vaší příští cestě!“)
- Personalizace zahrnující celou cestu místo jen jejích částí („Doufáme, že s námi příští týden zažijete příjemný let do Istanbulu. K tomu vám můžeme nabídnout ubytování v akční ceně 199EUR a limuzínu zdarma“)
- Personalizace založená na lokalitě („Vidíme, že jste právě přistál v Mnichově, ale vaše cílová destinace je Augsburg. Víte, že se tam můžete dostat vlakem, který odjíždí za 30 minut?“)
- Personalizace založená na narušení provozu („Mrzí nás, že nejspíše nestihnete následující let. Přijmete nabídku místa v business třídě zítra ráno v 8 hodin?“)

Důraz by měl být kladen také na mobilní platformu, která se stále více v oblasti distribuce dostává do popředí. V případě současné generace představují mobilní a sociální kanály dokonce dominantní komunikační prostředky. Ukazuje se, že nejmladší generace cestujících vykazuje následující atributy:

- Před samotným nákupem prověří průměrně 10,4 zdrojů online informací
- 75 % má založený profil na v sociálních médiích
- 83 % spí s telefonem u postele
- 84 % říká, že názory ostatních uživatelů internetu mají reálný dopad na jejich cestovní rozhodnutí
- 57 % aktualizuje svůj profil v sociálních médiích každý den během cestování

Tento trend naznačuje, že do budoucna jakákoliv distribuční strategie, která neklade silný důraz na mobilní platformu a z ní vyvozenou Big Data analýzu, povede k neúspěchu [37].

Jako další příklad možnosti praktického využití personalizace můžeme uvést její efektivní aplikaci na palubě letadla, která předpokládá vybavení palubních průvodčí tablety s nahranými detailními informacemi o věrnostním statusu cestujících, nákupních návycích z předchozích letů a dřívějších, dobrých či špatných zkušenostech z cestování s danou leteckou společností. Tato technologie umožní palubním průvodčím přístup k preferencím jednotlivých pasažérů, předchozím zážitkům a celosvětový přístup do databáze letecké společnosti, čímž může být zajištěno poskytnutí cestovního zážitku ušitého na míru, který cestující v budoucnu dost možná začnou očekávat [46]. V budoucnu tak palubní průvodčí mohou vědět, kdy máme narozeniny, kolik cukru si dáváme do kávy, znát naše alergie, preferovaná místa v letadle a historii ztrát našich zavazadel. Jedním pohledem do tabletu je také možné zjistit, kterých 5 cestujících v kabině představuje pro leteckou společnost největší hodnotu [47]. Kromě zlepšení vztahu s cestujícím tak s sebou tento přístup samozřejmě nese i potenciál vyšších výnosů z palubního prodeje. Poskytováním relevantních vedlejších produktů – od duty free zboží, přes palubní zábavu až po zboží spojené s konkrétní destinací (např. vstupenky na atrakce v destinaci) – prostřednictvím tradičního nákupu od palubních průvodčích nebo nově vznikajících kanálů (např. prostřednictvím IFE), mohou letecké společnosti výrazně zvýšit výkonnost palubního prodeje [46].

Se snahou o poskytování personalizovaného produktu se ale letecké společnosti také pomalu učí chodit po tenké čáře předělující poskytování excelentního servisu a vyvolávání znepokojujících pocitů [47]. Personalizace se totiž velmi lehce dokáže zvrhnout v dotěrnost přesažením „faktoru znepokojení“, pokud při ní není postupováno se souhlasem zákazníka, a zároveň i jemným a transparentním způsobem [37][46]. Například British Airways nedávno čelily skandálu spojenému se zjištěním, že narušily onu tenkou hranici, když daly za úkol svým palubním průvodčím aktivně

vyhledávat na Facebooku členy svého věrnostního programu, zapamatovávat si jejich tváře a na základě toho být schopní osobního přivítání na palubě letadla. Poté, co tato kauza vyšla na světlo, British Airways změnily svůj přístup a umožnily cestujícím na základě jejich osobní žádosti dostávat nepersonalizovaný servis. To samo o sobě ale neznamená, že o nich letecká společnost přestane sbírat data. American Airlines naopak přímo zakázaly svým palubním průvodčím ukládat většinu dat o svých cestujících, jako např. informaci o tom, co si dali k snídani, aby tak nemohlo dojít k případnému narušení jejich soukromí. Hledání správné polohy hranice mezi personalizovaným servisem a narušováním soukromí nejspíše bude nějakou dobu trvat a bude nutné odpovědět na celou řadu sociálních otázek. Cestující vyžadují, aby o nich letecké společností shromažďovaly právě takové množství informací, které jim umožní být pro konkrétního cestujícího užitečné, ale rozhodně si nepřejí, aby jim letecké společnosti „viděly až do ložnice“. Je tedy vhodné vědět, že cestující rád pije cappuccino, ale ne to, že jeho pes se jmenuje Punťa [47].

## 4.2 Prediktivní údržba

Prediktivní údržba představuje takový program řízení údržby, který se zakládá na monitorování jednotlivých komponentů letadel v reálném provozu, předjímá požadavky na údržbu a umožňuje její provedení těsně před tím, než dojde ke kritické poruše. Základem pro posun od preventivní údržby, která vyžadovala výměnu komponent v pevných časových intervalech, je dostupnost informací, a to jak historických dat, tak i současného stavu. Pokrok informačních technologií přispěl k větší možnosti shromažďovat informace o aktuálním mechanickém stavu zařízení a lepší schopnosti shromažďování a analýzy detailních historických dat o údržbě. To umožňuje přejít k této nové filozofii údržby. Zahrnutím většího množství informací do údržby chceme řešit především problém ohromných nákladů plynoucích z údržby letadel a situací, kdy jsou letadla z různých důvodů zcela mimo provoz. Další motivace k prediktivní údržbě představují náklady ušlé příležitosti neprovozuschopného letadla, snaha o maximalizaci využití letadel a právě i dostupnost nových technologií. Současnou výzvou, které čelí organizace údržby letadel, je schopnost vidět dopad jejich přístupu k údržbě na celkové náklady, vliv na využití letadla, personál a další provozní zdroje, letadla samotná a především inventář náhradních dílů.

Klíčem k správnému přístupu je uvědomění, že údržba znamená řízení souboru událostí vyžadujících údržbu. Veškerá aktivita v údržbě se nakonec soustřeďuje na řešení těchto událostí, které jsou ve své podstatě plánovaného nebo neplánovaného charakteru. Řízení údržby vyžaduje získávání a používání znalostí o jednotlivých událostech v údržbě.

Plánovaná údržba spotřebovává zdroje předvídatelným způsobem, ale možnost předvídání s sebou nese i náklady – mnohá údržba je prováděna v zbytečně velkém předstihu. Plánovaná údržba je prováděna v čase, kdy je letadlo nejméně potřebné

v provozu a prostředky potřebné k údržbě (jako náhradní díly a pracovní síla) mohou být doručené na požadované místo s předstihem a být tak optimálně dostupné.

Naproti tomu neplánovaná údržba znamená, že údržba je prováděna pouze tehdy, když je nezbytně potřebná. Bohužel se nedá plánovat, v zásadě má negativní dopad na provoz letecké společnosti a objevuje se v nejméně vhodnou dobu, tedy když je letadlo nejvíc potřeba v provozu. Neplánovaná údržba vyžaduje plánování a zajištění potřebných náhradních dílů a pracovní síly na místech, kde by tato neplánovaná údržba mohla být potřebná. Proces předjímání neplánované údržby závisí na kvalitě dostupných informací a flexibilitě v přístupu k jejím konkrétním požadavkům. Ať už k ní přistupujeme jakkoliv, dá se předpokládat, že bude vyžadovat zajištění mnohem většího množství různých prostředků ve velkém předstihu. S čím větším množstvím neplánované údržby se potýkáme, tím detailnější informace o historii údržby, konfiguraci zařízení a vzorcích opotřebení nebo provozních okolnostech budeme potřebovat. Čím více se posuneme od preventivní k prediktivní údržbě, tím více se budeme oddalovat od souhrnného využití informací, protože preventivní údržba stanovuje pevné intervaly údržby, zatím co preventivní údržba uvažuje různá specifika dané situace a navrhuje zásah těsně před selháním. Jelikož údržba spotřebovává zdroje, je nutné hledat informace o plýtvání, identifikovat nevyužité zdroje a hledat možnost lepšího časového využití letadel v jednotlivých zásazích údržby [48].

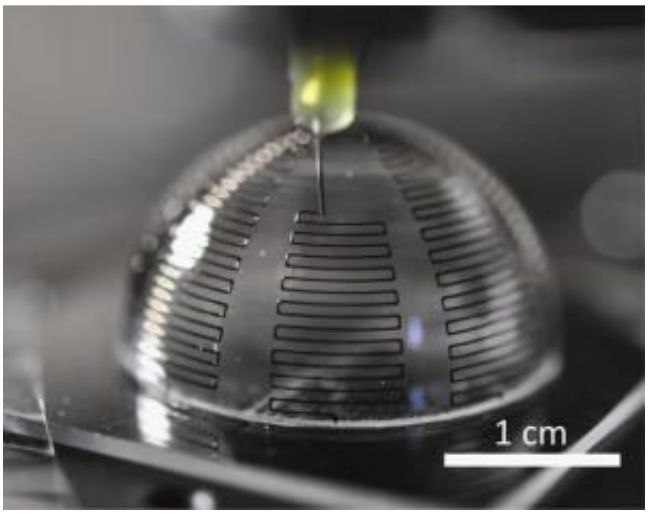
Letecké společnosti využívají prediktivní údržbu k snížení nákladů, přeměně neplánované údržby v plánovanou údržbu a minimalizaci prostojů [49]. Potenciál k úsporám je obrovský. Údržba má na svědomí zhruba 10 % provozních nákladů letecké společnosti a téměř polovinu všech zpoždění. Uzemnění dopravního letounu na zemi v průběhu údržby může způsobit náklady až 10 tisíc USD na hodinu. Snížení prostojů prostřednictvím efektivněji plánované údržby může dopravci přinést velké finanční výhody [50]. Tento přístup už se nějakou dobu uplatňuje v údržbě motorů, ale dnes letecké společnosti mohou provádět prediktivní analýzu mnohem efektivněji, protože nová letadla produkují mnohem více dat a nástroje k jejich zpracování dosáhly velmi vyspělé úrovně [49]. Vybavením kritických součástí letadel senzory je umožněn přenos informací inženýrům na zemi, kteří po přistání letadla mohou čekat připravení s náhradním dílem a začít s okamžitou údržbou [50].

V oblasti prediktivní údržby můžeme sledovat opravdový boom internetu věcí, protože nová letadla jsou lako létající servery [49]. Nejnovější letouny typu Boeing 787 Dreamliner a Airbus A350 jsou připojené k síti ještě víc než cestující na jejich palubách [50]. Boeing 787 vyprodukuje v průměru asi 500 GB systémových dat během jednoho letu, zatím co Airbus A350 je vybaven téměř 6000 senzory v různých částech letadla, které v průběhu jednoho dne dokáží vytvořit zhruba 2,5 Tb dat, a očekává se, že jeho novější verze, která by měla na trh přijít v roce 2020, bude produkovat třikrát tolik dat [51]. Největší dopravní letoun světa, Airbus A380, jehož první let byl proveden již před

deseti lety, je vybaven 25 tisíci senzory monitorujícími zhruba 200 tisíc parametrů [50][52]. Jeho nová verze Airbus A380-1000, schopná přepravy až 1000 cestujících, počítá s vybavením každého křídla až 10 tisíci senzory [51].

Letecké motory jsou dnes vybaveny senzory zachycujícími detaily každého aspektu jejich provozu, což znamená, že dopad vlhkosti, tlaku vzduchu a teploty mohou být určeny mnohem přesněji [51]. Např. Pratt & Whitney významně změní způsob správy jejich motorů nové generace za účelem zvýšení efektivity, snížení mechanického opotřebení a umožnění rychlejšího servisování. Jejich řada motorů PW1000G je charakterizována snížením hluku a spotřeby paliva, kterých bylo dosaženo užitím „umělé inteligence“ předpovídající nároky motoru a příslušným přestavením škrticí palivové klapky. Klíčem k zprovoznění takového motoru byla instalace 5000 senzorů. Motory z předchozí řady PW6000 a PW8000 jich přitom obsahovaly nanejvýš 100. Těchto 5000 senzorů produkuje ohromující množství dat – 10 GB/s a motor. To znamená 1,02 Tbps nebo 2,04 Tbps pro typický dvoumotorový letoun typu Airbus A320NEO nebo Boeing 737MAX. P&W předpovídá, že jejich potřeba zpracování dat dosáhne 12 PB ročně, což donutí tuto společnost k významným investicím do datových center a super-výpočetního výkonu [53].

Vytváření „chytrých součástí“ letadel pomocí velkého množství senzorů dalo vzniknout novým technologiím, jako např. technologii 3-D inking společnosti General Electric Co. Ta využívá počítačem ovládaný stylus podobný jehle ne širší než list papíru k vytváření malinkatých senzorů na součástkách, ze kterých jsou postavené proudové motory a jiné komplexní struktury. Tyto senzory jsou schopné měřit napětí součástí pracujících v nepříznivých podmínkách (např. místa s vysokými teplotami) a informace o nich přenášet. Takové senzory vyžadují nové materiály, jako např. keramiku, která je schopná snést vysoké teploty, nebo kombinace materiálů kompatibilních s lopatkami turbíny. Potřebné jsou i nové výrobní technologie. Senzory byly tradičně vyrobeny a nalisovány na požadovanou součástku. Tato technologie ale nefunguje dobře v případě válcových ploch, nebo když je potřeba senzor, který sahá za roh. 3-D inking se používá také v nanotechnologii k nanášení materiálů na velmi malé předměty. Materiál schopný snímání požadovaných parametrů ve formě prášku se zkombinuje s různými polymery a rozpouštědly, čímž vznikne gumovitá viskózní hmota, která má konzistenci podobnou zubní pastě. Tato hmota je pomocí stylusu nanesena na povrch předmětu v tenkých vrstvách, dokud nevznikne požadovaný senzor. Poté je předmět vystaven teplotě, při které se z hmoty vypaří nadbytečný materiál a keramické nebo kovové částečky tvořící senzor se spojí a připojí k součástce. Tento proces umožňuje umístění senzorů i na těžko dostupná místa a kladení různých typů senzorů (např. jeden pro teplotu a druhý pro tlak) vedle sebe do omezeného prostoru. Tato technologie může mít přínos i v jiných ohledech – např. přidání hrany k součástce, která vyžaduje lepší aerodynamické vlastnosti [54].




Obr. 15 *Vznik senzoru pomocí technologie 3-D inking* [54]

Dalším důležitým prvkem prediktivní údržby je technologie pro získávání dat, která zajišťuje monitorování dat z letu v reálném čase. Předními firmami v oblasti získávání dat jsou Honeywell a Rockwell Collins. Ty mohou těžit především z neutuchajícího zájmu různých subjektů podnikajících v odvětví letecké dopravě o monitorování dat v reálném čase. Výrobci letadel, motorů a avioniky, letecké společnosti, výrobci displejů do kokpitu a další výrobci letadlových částí, chtějí mít data o tom, jak si jejich produkty vedou za letu a během jeho různých fází, aby mohli určit chování jejich výrobků za různých podmínek a kam by měli směřovat jejich vylepšení.

Podle firmy Teledyne existují 3 hlavní důvody, proč letecké společnosti vylepšují systémy pro akvizici dat. V první řadě se letecké společnosti snaží o implementaci hardwaru pro získávání dat, který jim zprostředkuje víc dat a víc kontroly nad analýzou a přenosem dat pro neomezené aplikace ve vzduchu a na zemi. Druhý důvod vychází z regulatorních požadavků, které vyžadují nový software pro získávání dat nebo příslušný hardware pro jeho podporu. Zatřetí může být důvodem dosavadní nízká spolehlivost současného systému pro zisk dat. Teledyne nabízí řešení těchto potřeb pomocí rozšíření, které dokáže zdokonalit současná letadla jako A330, A340, B737 nebo B747 a další. Jedná se o jakési koncentrátory dat, které jsou připojené k 40 až 50 sběrnicím přenosového standardu ARINC 429, dalším datovým sběrnicím a analogovým kanálům. Sbírána jsou data ze všech palubních systémů. V závislosti na infrastrukturních schopnostech spojených s konkrétním letadlem může letecká společnost volit cestu automatického přenosu nebo manuálního stahování provozních dat podle potřeby [55].

Prediktivní údržba je dnes spojená především s nástroji vyvinutými samotnými výrobci letadel a letadlových částí. Většina zákazníků společnosti Pratt & Whitney například využívá její *Advanced Diagnostics and Engine Management* (ADEM) k pomoci při plánování údržby motorů. Společnost Honeywell vyvinula nástroj *Predictive Trend Monitoring and Diagnostics* (PTMD), poskytující prediktivní technologie pro pomocné

zdrojové jednotky (APU), takže dokáže předpovědět, kolik hodin zbývá do větších oprav. K iniciativě monitorování výkonů letadel se připojili také výrobci draků. Software společnosti Airbus *AIRcraft Maintenance Analysis* (Airman), využívaný 106 zákazníky, neustále monitoruje stav letadel a přenáší informace o poruchách nebo varovné zprávy pozemnímu personálu. Tento nástroj nabízí rychlý přístup k potřebným dokumentům, které popisují kroky k řešení problému, seřazeným podle pravděpodobnosti na úspěch. Podobný systém, *Airplane Health Management* (AHS), používá v 2000 letadlech i 53 zákazníků společnosti Boeing. Sám Boeing ho popisuje jako cestu k prediktivní údržbě, na které je k němu potřeba přidat zkušenosti a znalosti o konstrukci letadla a data včetně provozních dat a dat o údržbě. Data přispívající k možnostem bohatší predikce pochází z různých zdrojů a velké datové soubory jsou přenášeny za letu nebo stahovány po přistání.

Pokud výše zmíněné systémy indikují, že daná součást má lepší výkonnostní charakteristiky než se čekalo, může být její výměna nebo oprava odsunuta. Naopak pokud výkonnost není tak dobrá, jak by měla být, může k opravě dojít o to dřív. Tyto systémy ale bohužel ještě nejsou vůbec dokonalé. Opravdovou výzvou v prediktivní údržbě bývá často nacházení opravdových problémů bez nadměrného množství falešných alarmů. Problémem prediktivních systémů výrobců v současné době je, že dávají příliš mnoho alarmů a zaměstnanci řešící problém jim pak nevěří. A řešení, které nikdo nepoužívá, je vlastně k ničemu. K potlačení tohoto problému začaly některé konzultantské firmy v oblasti prediktivní údržby, jako např. Taleris nebo FCE, využívat techniku detekce anomálií, jejímž cílem je minimalizace jak falešných alarmů, tak i promeškání kritických událostí [49].

Detekce anomálií je snahou o nalezení toho, o čem nevíme, že máme hledat. Hledáme anomálie, ale přitom netušíme, jaké budou mít charakteristiky. Kdybychom je znali, mohli bychom použít jinou formu strojového učení, zvanou klasifikace, nebo napsat specifická pravidla pro vyhledávání anomálií. Obecně se však nejedná o správný začátek. Klasifikace je formou supervizovaného učení, kdy máme k dispozici příklad každého druhu věci, kterou hledáme. Pomocí učícího algoritmu se tak vytvoří model, který využije vlastností nových dat k jejich zařazení do příslušné kategorie. Když máme příklad normálních a určitého množství anomálních situací, model nám dokáže pomoci zařadit nové situace jako normální nebo anomální. Ale i v případě, že některé druhy anomálií známe, je vždy dobré přistupovat k tomuto problému s otevřenou myslí, protože se můžeme setkat s anomálií, kterou jsme dosud neznali.

K tomuto účelu se využívá tzv. nesupervizovaná detekce anomálií, kterou použijeme v případě, kdy přesně nevíme, co hledáme. Detekce anomálií je objevovací proces, který nám pomůže určit, co se děje, a na co bychom se při hledání měli zaměřit. Program pro detekci anomálií musí odhalit zajímavé vzorce a spojitosti v datech samotných, což detektor dělá tím, že nejprve identifikuje nejpodstatnější aspekt

detekce anomálií – to, co je normální. Jakmile se to modelu podaří, je náš program schopný v datech vyhledávat hodnoty ležící vně normálního souboru a označit je jako anomálie, které v drtivé většině případu indikují chování vedoucí k poruše [56].

# 5 Big Data jako podpora rozhodování provozního dispečinku

Provozní dispečink letecké společnosti Travel Service, a.s., sídlící na letišti Václava Havla v Praze, spadá pod provozní úsek této letecké společnosti a jeho hlavním úkolem je její operativní řízení a řešení provozních odchylek. Stručně řečeno žádný letový řád žádné letecké společnosti není možné realizovat ideálně, to znamená přesně tak, jak byl na začátku navržen. V jádru provozního dispečinku leží snaha o co možná nejpřesnější dodržování navrženého letového řádu, což se ale vhledem k množství faktorů ovlivňujících letecký provoz nemůže nikdy stoprocentně podařit.

Úkolem dispečera je, aby dokázal správně zanalyzovat nově nastalou, zpravidla nestandartní situaci, vyhodnotit, jaké nové podmínky a omezení z ní vyplývají, a učinit rychlé a správné rozhodnutí vedoucí k udržitelnosti stávajícího provozu a nejlépe také minimalizaci negativních dopadů, včetně zpoždění a dodatečných nákladů. Z povahy tohoto povolání plyne, že dispečer letecké společnosti musí být nejen zkušeným zaměstnancem se zevrubnou znalostí leteckého provozu na různých úrovních, ale zároveň musí umět myslet kontextuálně v rámci vysoce regulovaného prostředí, kterým letecká přeprava osob rozhodně je. Kromě toho, že dispečer dokonale zná požadavky svojí vlastní pracovní pozice, musí zároveň na základní úrovni rozumět i činnostem, pravomocem a oblastem zodpovědnosti ostatních oddělení a konkrétních pracovních pozic v rámci letecké společnosti, aby si v případě provozních změn dokázal do konečných důsledků představit jejich dopad na různé funkce v rámci letecké společnosti, popř. aby tyto funkce dokázal do jisté míry sám zastoupit.

Ekonomie leteckého provozu velí, že letadlo generuje zisk pouze během letu. Vzhledem k nízkým maržím v odvětví letecké dopravy je proto snahou každé letecké společnosti maximalizovat využití letadla. V průběhu letní sezóny dosahuje využití každého letadla ve flotile mnohdy až 20 hodin čistého letového času denně, což nechává pouze minimální manévrovací prostor pro případ, kdy některé z plně využitých letadel ztratí letovou způsobilost např. z důvodu technické závady. S rostoucím množstvím současně neprovozuschopných letadel mnohdy letecká společnost už není schopná pokrýt všechny lety vlastní kapacitou, z čehož plyne další činnost provozního dispečinku, a to poptávání letadel (tj. volné kapacity) od jiných leteckých společností prostřednictvím ACMI leasingu.

Mezi další činnosti, spadající do kompetence provozního dispečinku, patří především Flight Watch, základní povinnost každého dispečera, která v sobě skrývá potřebu neustálého monitorování provozu, aktuálního statusu všech letů a čtení provozních zpráv od handlingových společností, za účelem vytvoření detailního povědomí o současné situaci, jejího porovnání s letovým plánem a včasného řešení nesrovnalostí mezi nimi. S tím souvisí také ošetřování letových plánků včetně jejich zpožďování a předsouvání. Samostatnou pozici, kterou by měl zastávat nejzkušenější dispečer, si vyžaduje ošetřování ATC slotů. Tato činnost sestává z monitorování predikcí ATC

slotů pro lety společnosti Travel Service v internetové aplikaci oddělení Network Manager organizace Eurocontrol a komunikace s jejími operátory za účelem vylepšení slotů nebo naopak prodloužení bez nutnosti zpožďování plánku. Nejspíše nejdůležitějším úkolem letového dispečera je však optimalizace letového řádu v závislosti na vyhlášení tzv. AOG (Aircraft On Ground) statusu některého z letadel, který představuje ztrátu jeho letové způsobilosti. V tom případě je důležité, aby lety, které měly tímto letadlem být původně operovány, byly optimálně usazeny na zbývající letadla ve flotile, a to tak, aby tím vzniklé zpoždění bylo co nejnižší. Dispečeři mají na starost také správu aktuálně volné kapacity a její nabízení ostatním leteckým společnostem za úplatu. V neposlední řadě se starají o organizaci činností na pražské bázi společnosti Travel Service, včetně komunikace s handlingovou společností Menzies Aviation, s.r.o, objednávání paliva, radiové komunikace s přistávajícími a odlétajícími letadly, řešení problémů s cateringem, koordinace technickým úsekem společnosti, provozní spolupráce s provozním dispečinkem ČSA, atd.

Nařízení Evropského parlamentu a Rady (ES) č. 261/2004, které se zabývá ochranou cestujících v případě odepření nástupu na palubu letadla a významného zpoždění nebo zrušení letu, mění pro letového dispečera jakoukoliv událost s negativním dopadem na plynulost letového provozu na závod s časem. Mimo jiné uvádí, že je-li cestujícímu odepřen nástup na palubu, let je zrušen, nebo je jeho zpoždění na příletu do cílové destinace větší než 3 hodiny, má takový cestující nárok na odškodnění [57][58]:

**V rámci EU**

- do 1500 km 250 EUR
- nad 1500 km 400 EUR

**Mezi letištěm v EU a letištěm mimo EU**

- do 1500 km 250 EUR
- mezi 1500 a 3500 km 400 EUR
- nad 3500 km 600 EUR

Společnost Travel Service provozuje letadla typu Boeing 737-800 s kapacitou 189 míst a Airbus A320 s kapacitou 180 míst. V krajním případě tak pro tuto leteckou společnost kompenzace pro cestující na typickém letu z Prahy na Rhodos s ortodromickou vzdáleností 1884 km a zpožděného více jak 3 hodiny, představují náklad ve výši až 75600 EUR, resp. 72000 EUR, což jsou zhruba 2 miliony českých korun [59]. Zřejmě není potřeba zdůrazňovat, že profitabilita takového letu vyplacením kompenzací v takové výši značně klesne. Proto je naprosto nezbytné, aby byl dispečer letecké společnosti schopný velmi rychlého a efektivního rozhodování, které v ideálním případě povede k stlačení zpoždění pod 3 hodiny. Nutno podotknout, že i při nejlepší

vůli dispečera mnohdy různé neovlivnitelné faktory, jako např. podprůměrná práce handlingové společnosti nebo nespolupracující cestující, snížení zpoždění nedovolí. I na to však dispečer musí myslet, mít rozpracovaných několik variant a podle aktuální situace se přiklánět k té, která je dané situaci nejpřiměřenější. Aby toho však byl schopen, potřebuje k podpoře rozhodovacího procesu vždy včasné, přesné a úplné informace!

V následujících podkapitolách se zaměříme na možnost využití Big Data analýzy k podpoře rozhodovacího procesu na provozním dispečinku společnosti Travel Service. Pro tento účel budeme uvažovat jednoduché modelové situace, na kterých se pokusíme reprezentovat možnost integrace Big Dat do provozu letecké společnosti. Tyto příklady poslouží především k pochopení základní využitelnosti Big Dat v letecké společnosti.

## 5.1 Zdroje informací využívané dispečery letecké společnosti Travel Service a.s.

Základem jakékoliv Big Data strategie jsou zcela nepřekvapivě data. V této podkapitole se pokusím identifikovat všechny relevantní zdroje dat, které jsou dostupné dispečerům letecké společnosti Travel Service, a.s., a které mohou ležet v základu Big Data strategie provozního dispečinku spočívající v podpoře rozhodovacího procesu.

1. **EFA** (Extranet Flight Application)
   Internetová aplikace vyvinutá pro specifické potřeby společnosti Travel Service, poskytující dispečerům přístup ke komplexním provozním informacím včetně funkce Flight Graph, grafické interpretaci aktuální a minulé provozní situace a budoucího letového řádu. Dále se v ní nachází údaje z letových plánků a údaje o posádkách, počty cestujících stažené z rezervačního systému, údaje o případném nákladu či přepravě nebezpečného zboží, poznámky různých oddělení společnosti ke konkrétnímu letu, ATC Flight Plan, údaje z Journey Logu, údaje o kontraktech k plnění paliva na letišti Václava Havla, zprávy METAR a TAF pro aktuální destinace a jejich záložní letiště, provozní příručky a nařízení, atd.
2. **AIMS** (Airline Information Management System)
   Software AIMS slouží k vytváření letového řádu, usazování letů na konkrétní letadla a na úrovni provozního řízení k provádění provozních změn včetně přesouvání letů mezi letadly, úpravám časů odletu a příletu, přidávání nových letů, vyhlašování letů za prázdné, rozdělování, slučování letů, atd. V rámci funkce Flight Watch jsou do něho zapisovány údaje z provozních zpráv. V případě provozních změn slouží k vygenerování aktuálních informací o letovém řádu, které jsou zaslány na příslušná pracoviště, včetně handlingových společností, Ohlašovny letových provozních služeb, Letištní dispečink Letiště Václava Havla, atd.

3. **SkyWatcher**

   Jedná se o program vyvinutý speciálně pro Travel Service, který reprezentuje aktuálně spočítané a podané letové plánky. Kromě toho, že podobně jako funkce Flight Graph aplikace EFA zobrazuje aktuální provozní situaci, umožňuje zpožďování a předsouvání letových plánků včetně některých funkcí ošetřování ATC slotů, např. zasíláním zpráv RFI (Ready for Improvement), SWM (Slot Improvement Proposal Wanted Message), atd. SkyWatcher navíc umožňuje přístup k letovému plánku a informačnímu bulletinu (PIB). Jeho prostřednictvím je také možné komunikovat s letadly vybavenými systémem ACARS.

4. **SITATEX**

   Prostřednictvím služby SITATEX jsou vyměňovány zprávy provozního významu především s handlingovými společnosti a dalšími partnery. Zprávy jsou přijímány především ve formátu definovaném v IATA Airport Handling Manuálu, a jedná se hlavně o zprávy typu MVT (Aircraft Movement Messages), obsahující informace o času vytlačení letadla ze stojánky a času odletu, předpokládaném času příletu do destinace, počtu cestujících, případné délce a důvodu zpoždění, času přistání v destinaci a příjezdu na letištní stojánku, a další doplňující informace ve volné řeči. Speciální zprávou MVT je zpráva typu Delay (ED), která se posílá do cílové destinace v případě předpokládaného zpoždění většího než 20 minut a kromě velikosti zpoždění udává i jeho důvod. Dalšími významnými provozními zprávami jsou LDM (Load Message), obsahující informace o naložení letadla a PSM (Passenger Service Message) informující o cestujících vyžadujících po příletu do cílové destinace asistenci. Kromě zpráv v standardizovaném formátu umožňuje SITATEX výměnu zpráv ve volné řeči.

5. **Chráněná aplikace oddělení Network Manager**

   Registrovaný přístup do aplikace oddělení Network Manager (NM) umožňuje dispečerům společnosti Travel Service prostřednictvím funkce Flight List kontrolovat, že konkrétní letový plánek je v „bruselském“ systému podaný a aktuální. Dále dovoluje sledovat predikce ATC slotů a seznámit se s omezeními, které mají na svědomí snížení kapacity a rozdělování slotů. Po přidělení ATC slotu konkrétnímu letu je možné prostřednictvím aplikace E-Helpdesk zažádat o jeho zlepšení či prodloužení přímou komunikací s pracovníky NM. Funkce Flight List navíc nabízí porovnání velikosti přiděleného slotu se sloty přidělenými ostatním letům spadajícím pod stejnou slotovou regulaci, a další funkce.

6. **Automatizovaný informační systém AMIS**

   AMIS je neveřejný webový portál provozovaný Českým hydrometeorologickým ústavem, na kterém lze získat meteorologické informace pro civilní letectví, včetně zpráv METAR/SPECI, TAF, SIGMET/AIRMET, meteorologických map, atd.

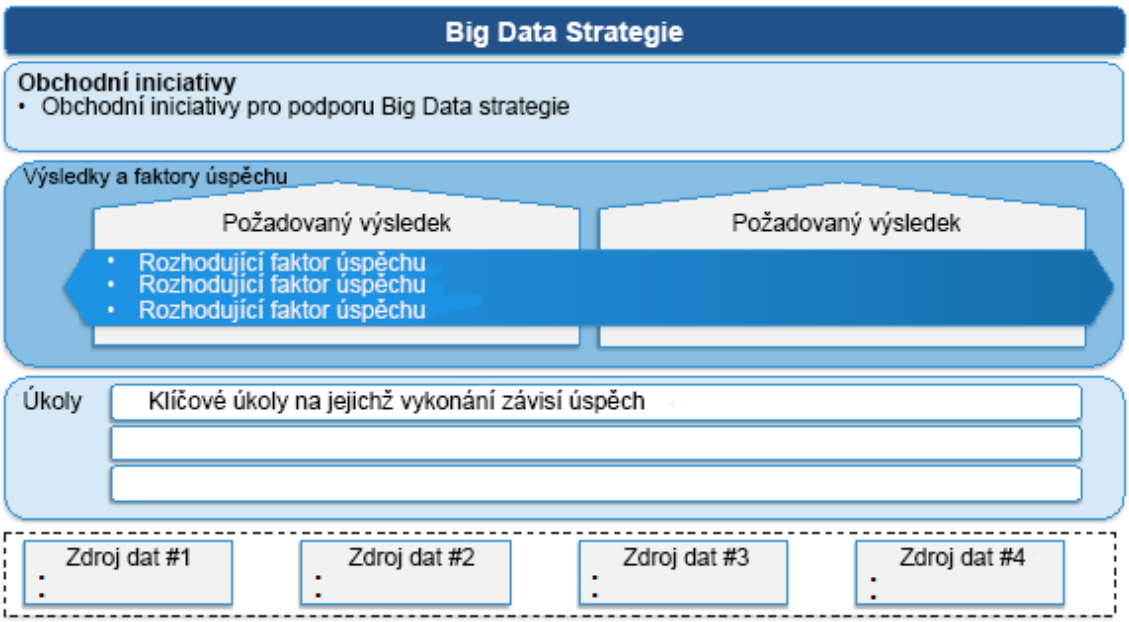
7. **FlightRadar24.com**

Webová aplikace založená na technologii ADS-B, která slouží k monitorování letadel a umožňuje dispečerům určovat přibližnou polohu letadel v daném čase. Tato webová stránka obsahuje také databázi dat o letech, avšak vzhledem k tomu, že jejich přesnost není zaručena, by časy vzletů a přistání letadel z ní neměly nahrazovat provozní zprávy přijaté prostřednictvím SITATEXu [60].

8. **Great Circle Mapper**
   Internetová aplikace umožňující měření ortodromické vzdálenosti mezi dvěma letišti, která je využívána v případě nutnosti rozhodnutí o poskytnutí či neposkytnutí služby cestujícím nebo jejího rozsahu v případě, když je závislá na ortodromické vzdálenosti dvou letišť (může se jednat o podávání občerstvení cestujícím na zpožděném letu) [59].
9. **E-mailový klient**
   Velké množství provozně důležitých informací, včetně provozních zpráv a informací o důležitých událostech v provozu a ACMI nabídek a požadavků, se na provozní dispečink dostane prostřednictvím e-mailového klienta. Jedná se o nástupce SITATEXu a pomalu ale jistě ho začíná plně nahrazovat.

## 5.2 Big Data strategie pro podporu rozhodovacího procesu provozního dispečinku

Abychom dokázali při tvorbě naší Big Data strategie pro podporu rozhodování provozních dispečerů udržet její relevanci vůči firemnímu procesu, budeme se držet strukturovaného dokumentu navrženého v [61] k zavádění malých pilotních Big Data projektů a graficky rozvrženého na následujícím brázku.




Obr. 16 *Struktura pomocného dokumentu pro implementaci pilotního Big Data projektu* [61]

Dokument pro stanovení Big Data strategie se skládá z následujících částí:

- **Big Data strategie** definuje rozsah, na který se v rámci Big Data iniciativy budeme soustředit. Jejím účelem je jasné definování obchodního cíle
- **Obchodní iniciativa** představuje projekt v trvání 9-12 měsíců s pevně stanovenými finančními a obchodními cíli, představujícími metriku, vůči které bude měřen úspěch obchodní iniciativy
- **Výsledky a faktory úspěchu.** Výsledky definují požadovaný nebo ideální konečný stav. Rozhodující faktory úspěchu říkají, co je potřeba udělat, aby byla obchodní iniciativa úspěšná
- **Úkoly.** Sekce dokumentu definující jaké konkrétní úkoly je potřeba splnit.
- **Zdroje dat.** Zdůraznění datových zdrojů potřebných k podpoře Big Data strategie. Sesbírána data jsou integrována v prediktivním nástroji, který na jejich základě dokáže vydat soubor doporučení pro optimalizaci současného stavu.

V následujících podkapitolách se zaměříme na dva základní pilotní Big Data projekty, zaměřené na rozhodovací procesy na provozním dispečinku, které mohou do budoucna sloužit jako reference pro budoucí holistické zavedení Big Data analýzy ve společnosti Travel Service. Jedná se o podporu rozhodování v případě zpoždění letu s velkým množstvím transferových cestujících a podporu rozhodování při overbookingu.

## 5.3 Podpora přestupu transferových cestujících

Společnost Travel Service obvykle zaznamenává největší počet transferových cestujících v letní sezóně především na linkách ze Splitu a Dubrovníku. Těchto cestujících někdy může v Praze přestupovat až 100 na více než desítku letů společnosti Travel Service, ale i ČSA a ve výjimečných případech i na lety ostatních společností. Zpoždění letu, z něhož více než polovina cestujících přestupuje v Praze na návazné lety, může mít negativní dopad na celkovou profitabilitu. V momentě, kdy se dispečer dozví o možném zpoždění takového letu, musí vzít do úvahy mnoho proměnných, na základě kterých rozhodne, jak situaci vyřešit. K tomu, aby ji dokázal vyřešit optimálně, se však nejprve musí dostat k velkému souboru potřebných informací. Problematiku optimálního rozhodování v případě zpoždění letu s velkým množstvím transferových cestujících, se pokusíme zjednodušit díky Big Data strategii.

*Big Data strategie*

1. Definování Big Data strategie
   **Podpora a optimalizace rozhodovacího procesu v případě zpoždění letu s transferovými cestujícími**
2. Obchodní iniciativy, které přímo podporují Big Data strategii

- **Vytvoření integrovaného informačního zdroje zpracovávajícího všechny informace relevantní k optimálnímu rozhodování a schopného reakce na měnící se podmínky**
- **Zvýšení situačního povědomí dispečerů o dění v provozu**

3. Výsledky a faktory úspěchu
   - **Dosažení optimálního zpoždění v síti letů společnosti Travel Service**
   - **Zvýšení úspěšnosti doručení zavazadel transferových cestujících**
   - **Zlepšení image obchodní značky a spokojenosti cestujících**
   - **Zvýšení celkové úspěšnosti přestupu transferových cestujících na 90 %**
   - **Snížení nákladů spojených se zmeškáním návazných letů na 50 % současného stavu**
4. Úkoly
   - **Sběr potřebných informací ze všech relevantních zdrojů**
   - **Integrace a analýza všech sesbíraných informací a vyhodnocování vlivu každého letu na základě jeho aktuálního statusu na celkovou situaci v síti**
   - **Navrhování opatření pro každý jednotlivý let za účelem optimalizace zpoždění v síti**
   - **Zhodnocení ekonomických faktorů pro minimalizaci nákladů**
   - **Predikce vlivu současného stavu na budoucí vývoj v síti**
5. Zdroje dat

   **Provozní zprávy** (především zprávy typu ED upřesňující velikost zpoždění; jména cestujících přítomných na palubě; naložení letadla a umístění zavazadel transferových cestujících; zpráva MVT s časy odletu letadla), **Informace o transferových cestujících** (počet transferových cestujících, jejich jména a návazné lety), **Informace o návazných letech** (poletí návazný let na čas?; ze kterého terminálu a stojánky poletí?), **Informace o ATC slotech** (Hrozí přidělení ATC slotu při zpoždění návazného letu pro čekání na transferové cestující? Jak to ovlivní provoz?), **Normy posádek** (Nedojde zpožděním letů k narušení schopnosti posádek operovat následné lety? Pokud ano, máme v záloze jinou posádku? Za jak dlouho může být schopná operovat daný let a jak to ovlivní další provoz?), **Informace o dalších letech do požadovaných destinací** (Operuje dnes naše nebo jiná společnost další let do požadované destinace? Kolik je volných míst?), **Historická obchodní data** (Jaké náklady s sebou zpožďování vzhledem k minulým zkušenostem ponese?), a další.

Představme si následující modelovou situaci: Na lince ze Splitu do Prahy letí 60 transferových cestujících, kteří v Praze přestupují na lety společnosti Travel Service do Paříže, Říma a Valencie a let ČSA do Soulu. Po SITATEXu provozní dispečink přijal ED zprávu, ze které vyplývá, že na příletu do Prahy bude mít tento let zpoždění 1 hodinu. Tato zpráva automaticky putuje do prediktivního Big Data modulu, který určí,

že cestující do Paříže let bez problému stihnou, přestup cestujících do Říma také není ohrožen, protože tento let je také zpožděn, odlet do Valencie doporučí zpozdit o 20 minut, protože další let do Valencie společnost operuje až za 2 dny a navíc to neohrožuje budoucí provoz, protože dle aktuálně spočteného letového plánku má být letadlo ve Valencii o 30 minut dřív, než bylo plánováno. Cestující do Soulu prediktivní nástroj doporučí ubytovat na hotelu a koupit jim letenky na druhý den se společností Korean Air, která má v současné chvíli volnou kapacitu, protože zpoždění tohoto dálkového letu by s sebou neslo neúnosně velké náklady (např. kvůli cestujícím, kteří přestupují na návazné lety v Soulu).

Výše popsaná úvaha by dispečerovi nejspíše trvala několik desítek minut a velmi pravděpodobně by i tak nedokázal postihnout veškeré dopady zpoždění linky ze Splitu. Nejspíše by našel nějaké obecné obecné řešení, které by však nemohlo odpovídat specifickým požadavkům jednotlivých přestupů. Navíc se může jednat o jeden z mnohých problémů, které dispečer v daný moment řeší, a nemůže mu tak věnovat dostatečnou pozornost. Big Data analýza je naproti tomu schopná okamžitě vyhodnotit nastalou situaci a prezentovat dispečerovi návrhy optimálního řešení, v důsledku čehož významně sníží jeho pracovní zátěž.

## 5.4 Podpora rozhodování při overbookingu

Overbooking je nástrojem Revenue managementu, který slouží k maximalizaci zisku ze sedačkové kapacity daného letadla. V jeho základu stojí statistiky konkrétní linky, které mohou např. ukázat, že se na let zpravidla dostaví pouze 90 % cestujících, kteří si zakoupili letenku. Letecká společnost se proto může rozhodnou nabízet 110 % sedačkové kapacity letadla a spoléhat na to, že v konečném důsledku bude letadlo na dané lince 100 % obsazeno. Mohou však nastat případy, kdy se všichni cestující na let dostaví, a potom je potřeba rozhodnout, kterým cestujícím bude odepřen nástup na palubu a nabídnut alternativní způsob přepravy a kompenzace. Ve většině případů jsou cestující vybíráni na základě toho, kdy se dostaví k odbavení. Těm posledním je zpravidla odepřen nástup.

Tento postup však nemusí vždy představovat optimální řešení. To může poskytnout Big Data analýza, která dokáže zvážit veškeré aspekty a důsledky odepření nástupu konkrétnímu cestujícímu.

*Big Data strategie*

1. Definování Big Data strategie
   **Podpora ideálního výběru cestujících, kterým bude odepřen nástup při overbookingu**
2. Obchodní iniciativy, které přímo podporují Big Data strategii
   - **Vytvoření integrovaného zdroje dat pro podporu řešení overbookingu**

- **Minimalizace negativních dopadů způsobených nevhodným výběrem cestujících, kterým byl odepřen nástup**

3. Výsledky a faktory úspěchu
   - **Dosažení 90 % úspěšnosti při oslovení předem vytipovaných cestujících, kteří by mohli být ochotní přijmout kompenzace a alternativní způsob přepravy**
   - **Maximalizace výnosů ze sedačkové kapacity letadel**
   - **Minimalizace nákladů spojených se zajištěním náhradní přepravy a s tím spojených služeb**
   - **Zlepšení image obchodní značky a spokojenosti zákazníků**
4. Úkoly
   - **Zajištění potřebných dat, jejich integrace a analýza**
   - **Včasné zjištění overbookingu a indentifikace nejvhodnějších cestujících k odepření nástupu na palubu**
   - **Určení nejvhodnějšího alternativního způsobu přepravy**
5. Zdroje dat

   **Itineráře cestujících** (Transferoví cestující? Cestující s dětmi nebo ve skupině? Cestující potřebující asistenci?), **Letové řády ostatních leteckých společností** (Operuje do této destinace v brzké době let jiná společnost? Mají na palubě místo?), **Historická obchodní data** (Zajištění minimalizace nákladů na základě předchozí zkušenosti), **Data sociálních sítí** (Odhad sentimentu cestujících), a další.

Díky Big Data analýze existuje velká šance, že se racionálním způsobem podaří vybrat cestující, jimž odepření nástupu do letadla z důvodu overbookingu způsobí pouze minimální, nebo žádné komplikace, a budou ochotní souhlasit s alternativním způsobem přepravy. Především se nejspíše bude jednat o mladé cestující letící za zábavou a bez přestupu v cílové destinaci, kterým přílet s několikahodinovým zpožděním nebude příliš vadit, zvlášť v případě vyplacení příslušné finanční kompenzace.

## Závěr

Letecké společnosti se v deregulovaném prostředí velmi často pohybují na tenké hranici předělující finanční zisk a ztrátu. Big Data pravděpodobně v dnešní době představují jeden z nejlepších nástrojů pro podporu obchodních strategií leteckých společností, které v sobě nesou potenciál toto finanční balancování překlopit na stranu zisku. Jelikož je letecká doprava odvětvím disponujícím masivním množstvím historických a nově generovaných dat, jeví se tento přístup založený na datech jako na míru ušitý tomuto odvětví. Letecké společnosti již data vlastní. Nyní je čas přijít s inovativním způsobem jejich využití, který je omezen pouze představivostí a kreativitou při kombinaci různých datových zdrojů, mezi kterými chceme najít korelace.

Letecké společnosti tradičně sbírají velké množství zákaznických dat, která představují základní úroveň využití Big Data analýzy. Již dnes je v silách leteckých společností poskytovat personalizovaný produkt na základě zákaznických profilů vytvořených ze zákaznických dat, které umožní nabízení vysoce specifických produktových balíčků jednotlivým cestujícím. Zákazník dostane přesně takový servis, jaký mu vyhovuje, a který může zároveň posílit loajalitu zákazníka k značce.

Snahou každé letecké společnosti je maximální využití její flotily. Tento plán je však často narušen nutností neplánovaných zásahů údržby, které s sebou často nesou šíření zpoždění sítí leteckého dopravce a dodatečné náklady. Big Data v tomto směru pomáhají pomocí prediktivní analýzy, která dokáže na základě analýzy signálů ze senzorů umístěných na různých součástech letadel s dostatečným předstihem rozpoznat jejich anomální chování a převést neplánovanou údržbu na plánovanou údržbou těsně před selháním součásti. Kromě toho, že bude možné lépe predikovat nežádoucí provozní události, maximalizovat využití flotily a snížit zpoždění z technických důvodů na minimum, dojde navíc k lépe předpověditelnému spotřebovávání omezených zdrojů při údržbě a umožnění racionálnější správy inventáře náhradních dílů.

V případě, kdy je potřeba řešit provozní události z pozice provozního dispečinku, je potřebné pracovat s včasnými, přesnými a úplnými informacemi, které jsou relevantní k optimálnímu vyřešení nastalé situace. Vzhledem k tomu, že na situaci může současně působit mnoho různých faktorů, není vždy v silách dispečera vyhodnotit situaci optimálně. Máme-li k dispozici dostatečné množství provozních dat, můžeme pomocí Big Data analýzy okamžitě vyhodnotit nejlepší řešení s ohledem na současný i budoucí stav sítě leteckého dopravce.

Letecká společnost je schopná udržet si konkurenční výhodu, pokud poskytuje produkt, který je unikátní, cenný pro zákazníka, a zároveň je velmi těžko imitovatelný konkurencí. Je zřejmé, že Big Data představují prostředek, který může být využíván

mnohými leteckými společnostmi a nejedená se tedy o nic unikátního či neimitovatelného. Dá se předpokládat, že Big Data přinesou největší konkurenční výhodu především společnostem, které je dokáží využít v rané fázi jejich vývoje. Do budoucna se však stanou obecným standardem, který přinese okamžité vzhledy do současné situace a návrhy řešení pro optimalizaci budoucího vývoje.

## Seznam použité literatury


[1] MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. Big Data. 1. vyd. Brno: Computer Press, 2014, 256 s. ISBN 978-80-251-4119-9.

[2] The Information Age and the Printing Press: Looking Backward to See Ahead. RAND Corporation [online]. [cit. 2015-11-08]. Dostupné z: http://www.rand.org/pubs/papers/P8014/index2.html

[3] What Impact Did the Invention of the Printing Press Have on the Spread of Religion? Synonym [online]. [cit. 2015-11-08]. Dostupné z: http://classroom.synonym.com/impact-did-invention-printing-press-spread-religion-6617.html

[4] „Internet Security“. Bletchley Park. The Mansion, Bletchley Park, Sherwood Dr, Bletchley, Milton Keynes MK3 6EB, Velká Británie. 30.8.2015

[5] The History and Evolution of the Internet, Media, and News in 5 Infographics. ACI [online]. [cit. 2015-11-08]. Dostupné z: http://aci.info/2013/10/24/the-history-and-evolution-of-the-internet-media-and-news-in-5-infographics/

[6] Data. TechTerms [online]. [cit. 2015-11-09]. Dostupné z: http://techterms.com/definition/data

[7] SHIFTING THE STORAGE PARADIGM (PART ONE): THE EVOLUTION OF DATA. DDN Storage [online]. [cit. 2015-11-09]. Dostupné z: http://www.ddn.com/blog/shifting-storage-paradigm-part-one-evolution-data/

[8] The Evolution of Data. VISUAL CAPITALIST [online]. [cit. 2015-11-09]. Dostupné z: http://www.visualcapitalist.com/evolution-of-data/

[9] A Relational Database Overview. Oracle [online]. [cit. 2015-11-09]. Dostupné z: https://docs.oracle.com/javase/tutorial/jdbc/overview/database.html

[10] HAN, Hu, Wen YONGGANG, Chua TAT-SENG a Li XUELONG. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. Access, IEEE [online]. 2014, (2): 652-687 [cit. 2015-11-09]. DOI: 10.1109/ACCESS.2014.2332453. Dostupné z: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=

[11] Bits &Bytes: The History Of Data Storage. Backupify [online]. [cit. 2015-11-09]. Dostupné z: https://www.backupify.com/history-of-data-storage/

[12] The birth of the web. CERN [online]. [cit. 2015-11-09]. Dostupné z: http://home.cern/topics/birth-web

[13] Total number of Websites. Internet Live Stats [online]. [cit. 2015-11-09]. Dostupné z: http://www.internetlivestats.com/total-number-of-websites/

[14] What Is Cloud Computing. Real Simple [online]. [cit. 2015-11-09]. Dostupné z: http://www.realsimple.com/work-life/technology/organizing-time-savers/what-is-cloud-computing

[15] Cloud Computing - How it all works. Youtube [online]. [cit. 2015-11-09]. Dostupné z: https://www.youtube.com/watch?v=TTNgV0O_oTg

[16] The Evolution of Mobile Technologies: 1G to 2G to 3G to 4G LTE. Qualcomm [online]. [cit. 2015-11-09]. Dostupné z: https://www.qualcomm.com/videos/evolution-mobile-technologies-1g-2g-3g-4g-lte

[17] he History Of Smartphones: Timeline. The Guardian [online]. [cit. 2015-11-09]. Dostupné z: http://www.theguardian.com/technology/2012/jan/24/smartphones-timeline

[18] The Internet of Things: How the Next Evolution of the Internet Is Changing Everything. Cisco [online]. [cit. 2015-11-09]. Dostupné z: http://www.iotsworldcongress.com/documents/4643185/3e968a44-2d12-4b73-9691-17ec508ff67b

[19] Gartner Says 4.9 Billion Connected "Things" Will Be in Use in 2015. Gartner [online]. [cit. 2015-11-09]. Dostupné z: http://www.gartner.com/newsroom/id/2905717

[20] Morgan Stanley: 75 Billion Devices Will Be Connected To The Internet Of Things By 2020. Business Insider [online]. [cit. 2015-11-09]. Dostupné z: http://www.businessinsider.com/75-billion-devices-will-be-connected-to-the-internet-by-2020-2013-10

[21] Internet Users Send 204 Million Emails Per Minute. Mashable [online]. [cit. 2015-11-09]. Dostupné z: http://mashable.com/2014/04/23/data-online-every-minute/#ChtHf6qPWSqu

[22] The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature. Research Trends [online]. [cit. 2015-11-09]. Dostupné z: http://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/

[23] 'Big Data' Is One Of The Biggest Buzzwords In Tech That No One Has Figured Out Yet. Business Insider [online]. [cit. 2015-11-09]. Dostupné z: http://www.businessinsider.com/companies-not-embracing-big-data-2014-8

[24] The Big Data Conundrum: How to Define It? MIT Technology Review [online]. [cit. 2015-11-09]. Dostupné z: http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/

[25] NOVOVIČOVÁ, Jana. Pravděpodobnost a matematická statistika. Vyd. 1. Praha: České vysoké učení technické, 1999, 154, iii s. ISBN 80-010-1980-2.

[26] Nate Silver's Election Predictions a Win for Big Data, The New York Times. Advertising Age [online]. [cit. 2015-11-09]. Dostupné z: http://adage.com/article/campaign-trail/nate-silver-s-election-predictions-a-win-big-data-york-times/238182/

[27] Correlation and Causation. Dr. Wheeler's Website [online]. [cit. 2015-11-09]. Dostupné z: https://web.cn.edu/kwheeler/logic_causation.html

[28] Pojem post hoc ergo propter hoc. Slovník cizích slov [online]. [cit. 2015-11-09]. Dostupné z: http://slovnik-cizich-slov.abz.cz/web.php/slovo/post-hoc-ergo-propter-hoc

[29] How do hardware costs compare for NoSQL versus RDBMS databases? Quora [online]. [cit. 2015-11-09]. Dostupné z: https://www.quora.com/How-do-hardware-costs-compare-for-NoSQL-versus-RDBMS-databases

[30] File Systems (FAT, HPFS, NTFS). Yale [online]. [cit. 2015-11-09]. Dostupné z: http://www.yale.edu/pclt/BOOT/IFS.HTM

[31] Understanding file systems. UFS Explorer [online]. [cit. 2015-11-09]. Dostupné z: http://www.ufsexplorer.com/und_fs.php

[32] How the Google File System Works. How Stuff Works [online]. [cit. 2015-11-09]. Dostupné z: http://computer.howstuffworks.com/internet/basics/google-file-system.htm

[33] What is NoSQL? MongoDB [online]. [cit. 2015-11-09]. Dostupné z: https://www.mongodb.com/nosql-explained

[34] MapReduce Tutorial. Hadoop [online]. [cit. 2015-11-09]. Dostupné z: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

[35] MapReduce. Steve Krenzel [online]. [cit. 2015-11-09]. Dostupné z: http://stevekrenzel.com/articles/finding-friends

[36] HDFS, MapReduce, and YARN (Core Hadoop). Cloudera [online]. [cit. 2015-11-09]. Dostupné z: http://www.cloudera.com/content/www/en-us/products/apache-hadoop/hdfs-mapreduce-yarn.html#hdfs

[37] DAVENPORT, Thomas H. At the Big Data Crossroads: turning towards a smarter travel experience [online]. [cit. 2015-11-09]. Dostupné z: http://www.bigdata.amadeus.com/assets/pdf/Amadeus_Big_Data.pdf

[38] BIG DATA TAKES OFF: Crunching the numbers to give passengers more comfortable and cost-efficient flights. Future of Transportation [online]. [cit. 2015-11-09]. Dostupné z: http://sciamfot.com/big-data-takes-off/

[39] The Evolution Of Customer Experience: Why Customer-Centric Airlines Will Lead the Market. Sabre [online]. [cit. 2015-11-09]. Dostupné z: http://sabre-2.hs-sites.com/experience

[40] BIG DATA REVOLUTION: TRAVEL INDUSTRY LEVERAGING BIG DATA FOR COMPETITIVE ADVANTAGE. Ascend [online]. [cit. 2015-11-09]. Dostupné z: http://www.ascendforairlines.com/2013-issue-no-4/big-data-revolution

[41] Boeing becoming more data driven after realizing big data benefits. Hortonworks [online]. [cit. 2015-11-09]. Dostupné z: http://hortonworks.com/big-data-insights/boeing-becoming-more-data-driven-after-realizing-big-data-benefits/

[42] Commercial Aviation and Aerospace: Big Data Analytics for Advantage, Differentiation and Dollars. IBM [online]. [cit. 2015-11-09]. Dostupné z: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA

[43] Smarter Airlines: Creating smarter airlines with Big Data. IATA [online]. [cit. 2015-11-09]. Dostupné z: http://www.iata.org/events/Pages/airlinesintl-webinar.aspx

[44] ROBERTS, Mary Lou a Debra L ZAHAY. Internet marketing: integrating online and offline strategies. 3rd ed. Mason, OH: South-Western Cengage Learning, c2013, xxii, 484 p. ISBN 978-113-3625-902.

[45] What do your passengers want? IATA [online]. [cit. 2015-11-09]. Dostupné z: http://view6.workcast.net/?pak=6587249560375849

[46] How airlines can get the most from Big Data to improve the passenger experience and increase ancillary revenues. Future Travel Experience [online]. [cit. 2015-11-09]. Dostupné z: http://www.futuretravelexperience.com/2014/05/airlines-can-get-big-data-improve-passenger-experience-increase-ancillary-revenues/

[47] How Airlines Mine Personal Data In-Flight: Flight Attendants Are Likely to Know What Fliers Will Buy on Board. The Wall Street Journal [online]. [cit. 2015-11-09]. Dostupné z: http://www.wsj.com/articles/SB10001424052702304384104579139923818792360

[48] Putting Aviation Operations Data to Work. Teradata [online]. [cit. 2015-11-09]. Dostupné z: http://www.teradata.com/resources/white-papers/Putting-Aviation-Operations-Data-to-Work-eb5137/

[49] New Predictive MRO Tools Cut Costs. Aviation Week [online]. [cit. 2015-11-09]. Dostupné z: http://aviationweek.com/awin/new-predictive-mro-tools-cut-costs

[50] Smarter aircraft create a wealth of data but it remains underexploited. Financial Times [online]. [cit. 2015-11-09]. Dostupné z: http://www.ft.com/cms/s/2/3f956a92-0943-11e5-b643-00144feabdc0.html#axzz3r1GuocsA

[51] That's Data Science: Airbus Puts 10,000 Sensors in Every Single Wing!. Data Science Central [online]. [cit. 2015-11-09]. Dostupné z: http://www.datasciencecentral.com/profiles/blogs/that-s-data-science-airbus-puts-10-000-sensors-in-every-single

[52] Millions of data points flying in tight formation. Aerospace Manufacturing And Design [online]. [cit. 2015-11-09]. Dostupné z: https://www.onlineamd.com/article/millions-of-data-points-flying-part2-121914

[53] BIG DATA IN PLANES: NEW P&W GTF ENGINE TELEMETRY TO GENERATE 10GB/S. VR World [online]. [cit. 2015-11-09]. Dostupné z: http://vrworld.com/2015/05/08/big-data-in-planes-new-pw-gtf-engine-telemetry-to-generate-10gbs/

[54] GE Puts Sensors in Hard-to-Reach Places With 3-D Inking: Air Travel Could Be Made Safer With Tiny Monitoring Devices Inside Jet Engines. The Wall Street Journal [online]. [cit. 2015-11-09]. Dostupné z: http://www.wsj.com/articles/ge-puts-sensors-in-hard-to-reach-places-with-3-d-inking-1401374016

[55] Avionics Big Data: Impacting All Segments of the Aviation Industry. Avionics [online]. [cit. 2015-11-09]. Dostupné z: http://www.aviationtoday.com/av/issue/departments/products/Avionics-Big-Data-Impacting-All-Segments-of-the-Aviation-Industry_83744.html#.VkDkGfkvfIU

[56] DUNNING, Ted a B FRIEDMAN. Practical machine learning: a new look at anomaly detection [online]. First edition. Beijin: O'Reilly, 2014, iv, 58 pages [cit. 2015-11-09]. ISBN 14-919-1160-3.

[57] Air Passenger Rights. Your Europe [online]. [cit. 2015-11-09]. Dostupné z: http://europa.eu/youreurope/citizens/travel/passenger-rights/air/index_en.htm

[58] AIR PASSENGER RIGHTS EU COMPLAINT FORM. European Commission [online]. [cit. 2015-11-09]. Dostupné z:

http://ec.europa.eu/transport/themes/passengers/air/doc/complain_form/eu_complaint_form_en.pdf

[59] Great Circle Mapper [online]. [cit. 2015-11-09]. Dostupné z: http://www.gcmap.com/

[60] FlightRadar24 [online]. [cit. 2015-11-09]. Dostupné z: http://www.flightradar24.com/50.02,14.93/9

[61] SCHMARZO, Bill. Big data: understanding how data powers big business. Indianapolis, IN: John Wiley, 2013, 1 online zdroj (242 pages). ISBN 978-1-118-74003-3.

## Seznam obrázků a grafů

## Seznam tabulek