Univerzita Pardubice

Fakulta ekonomicko-správní

Big data v organizaci

Bakalářská práce

2020 Milan Zeman

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2019/2020

ZADÁNÍ BAKALÁŘSKÉ PRÁCE
(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Milan Zeman**
Osobní číslo:	**E17624**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Téma práce:	**Big data v organizaci**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování

Cílem práce je charakterizovat základní vlastnosti strukturovaných a nestrukturovaných dat v organizaci. Zaměření bude na životní cyklus dat, modelování a nakládání s daty.

Osnova:
- Základní pojmy související se zpracovávanou problematikou.
- Životní cyklus dat.
- Tvorba modelů.

Rozsah pracovní zprávy: **cca 35 stran**
Rozsah grafických prací:
Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam doporučené literatury:

GRONWALD Klaus-Dieter. *Integrated business information systems: a holistic view of the linked business process chain ERP-SCM-CRM-BI-big data*. New York: Springer Berlin Heidelberg, 2017. ISBN 978-3-662-53290-4.
HOLUBOVÁ, Irena, KOSEK, Jiří, MINAŘÍK Karel a David NOVÁK. *Big Data a NoSQL databáze*. Praha: Grada, 2015. Profesionál. ISBN 978-80-247-5466-6.
LEE, James, WEI, Tao a Suresh Kumar MUKHIYA. *Hands-On Big Data Modeling*. Birmingham, UK: Packt Publishing, 2018. ISBN 978-1788620901.
MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. *Big Data*. Brno: Computer Press, 2014. ISBN 978-80-251-4119-9.

Vedoucí bakalářské práce: **doc. Ing. Stanislava Šimonová, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **2. září 2019**
Termín odevzdání bakalářské práce: **30. dubna 2020**

L.S.

doc. Ing. Romana Provazníková, Ph.D.
děkanka

doc. Ing. Pavel Petr, Ph.D.
vedoucí ústavu

V Pardubicích dne 2. září 2019

PROHLÁŠENÍ

Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávání, zveřejňování a formální úpravu závěrečných prací, ve znění pozdějších dodatků, bude práce zveřejněna prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 30. 6. 2020

Milan Zeman

PODĚKOVÁNÍ

Chci poděkovat doc. Ing. Stanislavě Šimonové, Ph.D. za její odbornou pomoc a cenné rady během zpracování bakalářské práce. Dále chci poděkovat své rodině a přátelům za podporu během studia.

ANOTACE

Bakalářská práce se zabývá charakteristikou Big data. Úvodní část práce se věnuje základním pojmům týkajících se dat. Druhá kapitola popisuje základní vlastnosti Big data. Třetí kapitola charakterizuje hlediska, podle kterých jsou zpracovány další části bakalářské práce. Hlediska se týkají životního cyklu dat, architektury Big data, technologií pro práci s Big data a datových modelů.

KLÍČOVÁ SLOVA

Big data, nestrukturovaná data, strukturovaná data, životní cyklus dat, datové modely

TITLE

Big data in an organization

ANNOTATION

The bachelor thesis deals with the characteristics of Big data. The introductory part of the thesis deals with basic concepts related to data. The second chapter describes the basic properties of Big data. The third chapter characterizes the aspects according to which other parts of the bachelor's thesis are processed. Aspects relate to the data lifecycle, Big data architecture, technologies for working with Big data and data models.

KEYWORDS

Big data, unstructured data, structured data, data life cycle, data models

OBSAH

SEZNAM ILUSTRACÍ A TABULEK

SEZNAM ZKRATEK A ZNAČEK

ACID	atomicity, consistency, isolation, durability
BI	Business Intelligence
DAS	Direct Attached Storage
DIKW	Data, Information, Knowledge, Wisdom
GFS	Google File System
HDFS	Hadoop Distributed File System
IOPS	input / output operations per second
IoT	Internet of things
IS	Information system
IT	Information technology
JVM	Java Virtual Machine
NAS	Network Attached Storage
NBDRA	NIST Big Data Reference Architecture
NIST	National Institute of Standards and Technology
NoSQL	not only SQL
SAN	Storage Area Network
SQL	Structured Query Language

ÚVOD

V organizacích proudí hromady nevyužitých dat. Pokud se data začnou nekontrolovatelně hromadit vysokou rychlostí, lze hovořit o termínu Big data. Převážná část těchto dat je ve nestrukturované formě a organizace si jich příliš nevšímají. Ovšem tato data mohou být pro organizace potencionálním zlatým dolem. Jen je třeba najít správné postupy a řešení, jak s těmito daty pracovat.

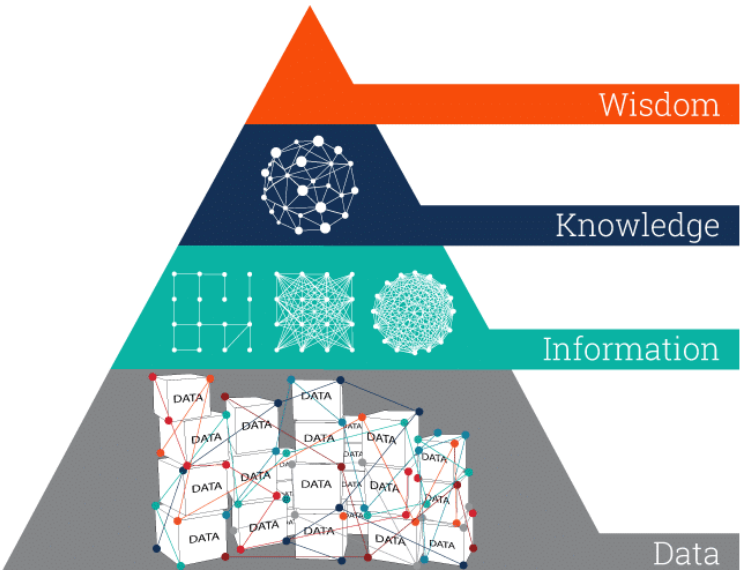
První část bakalářské práce popisuje základní pojmy související s daty. Charakterizuje kvalitu a typy dat podle jejich struktury a podle místa vzniku. Ve druhé kapitole se čtenář dozví základní charakteristiky Big data, pochopí samotný pojem Big data, seznámí se s historií, využitím v současnosti, a také s možnou budoucností a riziky, která od Big data hrozí. Třetí kapitola charakterizuje další směřování bakalářské práce podle zvolených hledisek. Další kapitoly se zabývají jednotlivými hledisky. Nejprve je popsán životní cyklus dat, v další kapitole je charakterizována architektura Big data, dále technologie pro práci s Big data a na závěr jsou popsány datové modely u SQL a NoSQL databází

Na téma bylo čerpáno z mnoha zdrojů, převážně zahraničních. Českých zdrojů vhodných pro vypracování práce bylo nalezeno minimum a většinou se týkaly jen obecných vlastností Big data. Tato bakalářská práce může sloužit jako výukový materiál pro ty, kdo se o Big data zajímají.

1 DATA

Data jsou fakta, signály nebo symboly reprezentující nějakou událost z reálného světa, ovšem bez kontextu nemají žádný smysl. Jedná se o čísla, znaky a další symboly.

Znalostní pyramida



Obrázek 1: Znalostní pyramida

Zdroj: [9]

Pyramida DIKW (Data, Information, Knowledge, Wisdom) nebo také znalostní pyramida reprezentuje vztahy mezi daty, informacemi, znalostmi a moudrostí (Obrázek 1). Je zde pevně stanovená forma – data se nachází na prvním místě, za nimi následují informace, dále znalosti a nakonec moudrost. Každým krokem směrem nahoru pyramida odpovídá na otázky o počátečních datech a přidává jim hodnotu. Čím více jsou data obohacena smyslem a kontextem, tím více znalostí a poznatků z nich lze získat. Na vrcholu pyramidy se získané znalosti a postřehy mění do zkušeností, pomocí kterých lze činit důležitá rozhodnutí. Vyšší stoupání v pyramidě znamená vyšší zapojení lidské činnosti a zároveň nižší nebo žádné možnosti algoritmizace [9].

Pokud je datům přidělený význam, stávají se z nich informace. Jedná se tedy o soubor údajů, které jsou jednotně uspořádané a umožňují jednodušší ukládání a vyhledávání dat. Jestliže se informace neberou pouze jako popis shromážděných skutečností, ale jako možnost k dosažení cílů, mění se informace ve znalosti. Znalosti ukrývají výhody, které dávají výhody podnikům proti jejich konkurenci. Moudrost se nachází na vrcholu pyramidy a aby se tam podnik dostal,

musí si odpovědět na otázky „proč něco dělat“ nebo „co je nejlepší“. Jedná se o znalosti aplikované do praxe. Moudrost je schopnost vybrat nejlepší způsob, jak dosáhnout požadovaného výsledku na základě znalostí a je výsledkem zkušeností nebo znalostí dřívějších pokusů o dosažení úspěšného výsledku [9].

Metadata

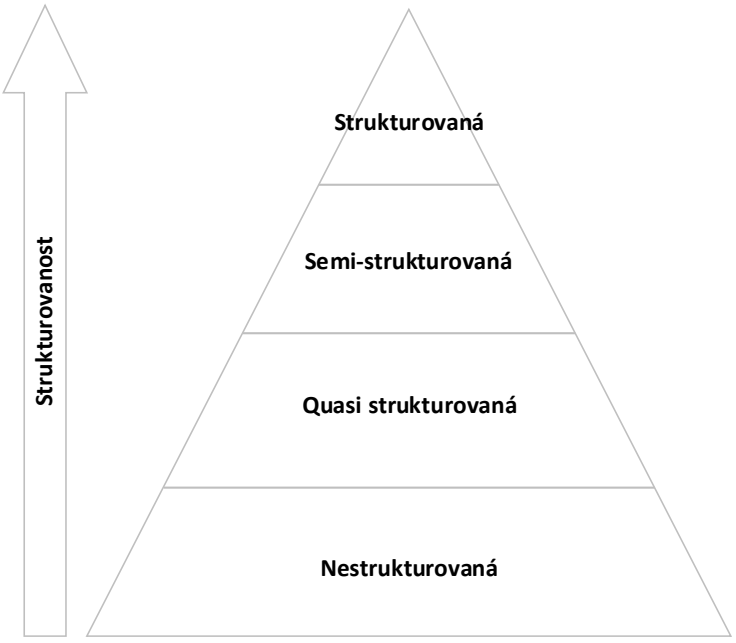
Metadata jsou data poskytující informace o dalších datech. Shrnují informace o datech, což usnadňuje vyhledávání a práci s konkrétními daty. Metadata lze vytvářet ručně nebo automaticky. Ruční vytvoření zajistí větší přesnost metadat, automatické vytvoření bude obsahovat více základních informací. Metadata lze chápat jako odkazy na data [14].

1.1 Typy dat

Data se mohou dělit podle mnoha různých faktorů. V následujících případech budou data rozdělena podle své struktury a podle toho, kde vznikla.

1.1.1 Podle struktury

Data mohou v organizaci existovat v různých formátech. Hlavní dva formáty dat jsou data strukturovaná a nestrukturovaná. Mezi nimi se nachází další typ dat, semi-strukturovaná. V některých případech jsou uváděna i Quasi strukturovaná data (Obrázek 2).



Obrázek 2: Strukturovanost dat

Zdroj: Zpracováno dle [1]

Strukturovaná data

Strukturovaná data jsou data přehledně uspořádaná pro snadný přístup a jejich další zpracování. Jedná se obvykle o kvantitativní data. Výhoda strukturovaných dat spočívá v jejich jednoduchosti. Data lze snadno zadávat, ukládat, prohledávat nebo s nimi manipulovat. Data jsou obvykle ukládána v relačních databázích a tabulkách [11].

U strukturovaných dat je pevně stanovený datový model, podle kterého budou data zaznamenávána. Model určuje jaké typy dat budou ukládány, zda číselné datové typy, text, datum atd. nebo jaká omezení platí při zadávání dat (např. stanovení maximální délky textu).

Nestrukturovaná data

Nestrukturovaná data můžeme označit jako všechno ostatní. Mají sice vnitřní strukturu, ale nejsou předdefinovaná podle datových modelů nebo schémat a nemohou být uspořádané v relačních databázích. Neobsahují žádná metadata popisující obsah dokumentu. Pro správu tohoto typu dat se používají nerelační databáze, NoSQL databáze nebo datová jezera [11].

Nestrukturovaná data tvoří 80 % všech generovaných dat a neustále svůj objem zvětšují. Nalezení informací uložených v nestrukturovaných datech může být velmi složité a zdlouhavé. Organizace, které jsou schopny pracovat s nestrukturovanými daty mají konkurenční výhodu a mohou odhalit hlubší záměry nebo chování zákazníka [11].

Semi-strukturovaná data

Semi-strukturovaná data jsou zvláštním typem dat. Neodpovídají sice datovému modelu, ale mají určitou strukturu. Semi-strukturovaná data nemohou být uložena v řádcích ani sloupcích databází. Obsahují metadata, která se používají k seskupování dat a k jejich hierarchickému uspořádávání [13].

Quasi strukturovaná data

Jedná se o textová data s nepravidelnými datovými formáty, která lze s dostupnými nástroji, úsilím a časem zformátovat. Příkladem mohou být data z kliknutí na webu, která nebudou konzistentní v hodnotách ani ve formátech [1].

1.1.2 Podle místa vzniku

Podle místa vzniku se data rozlišují na interní a externí. Jako interní data se označují data vzniklá uvnitř firmy, například z databáze zákazníků nebo z jejich nákupů. Externí data vznikají

primárně mimo podnik. Mezi externí data se řadí otevřená data, data z výzkumů a nakoupené databáze [6].

Interní data tvoří většinovou část informací o tom, jak se chovají zákazníci daného podniku. Důležité je využít potenciál vlastních firemních dat před tím, než podnik nakoupí externí data. Z interních dat lze obvykle velmi dobře vytvářet predikce, které odhalují chování zákazníka. Teprve až podnik vytěží hlavní potenciál interních dat, může se porozhlédnout po externích zdrojích [6].

Externí data prohlubují informace o zákaznících. Pokud podnik vstupuje na nový trh a chce přilákat zákazníky nebo se chce vyrovnat konkurenci, stávají se externí data nezbytným pomocníkem. Propojení interních a externích dat pomáhá optimalizovat výběr potenciálních zákazníků [6].

1.2 Kvalita dat

Kvalita dat ovlivňuje správné rozhodování a fungování organizace. Kvalitní data znamenají úspěšné fungování podniku, za nekvalitní data mohou firmy zbytečně přicházet o peníze. Kvalita dat se týká pěti hlavních znaků – přesnost, úplnost, spolehlivost, relevance a aktuálnost [16]:

- Přesnost – Jsou informace správné v každém detailu?
- Úplnost – Jak komplexní jsou informace?
- Spolehlivost – Jsou informace v rozporu s jinými důvěryhodnými zdroji?
- Relevance – Opravdu tyto informace potřebujete?
- Aktuálnost – Jak aktuální jsou informace? Lze je použít pro reporting v reálném čase?

Přesnost dat určuje, zda jsou informace správné. Pro přesnost se stačí zeptat na otázku, zda data odrážejí skutečnou situaci. Jedná se o klíčovou charakteristika kvality dat, protože nepřesné informace mohou způsobit závažné problémy. Úplnost určuje, jak komplexní informace jsou. Při pohledu na úplnost údajů stačí přemýšlet o tom, zda jsou všechna potřebná data k dispozici. Pokud jsou informace neúplné, mohou být nepoužitelné. Spolehlivost v oblasti charakteristik kvality dat znamená, že část informací není v rozporu s jinou informací v jiném zdroji nebo systému. Spolehlivost je zásadní vlastností kvality dat. Pokud si informace protiřečí, nemůže se těmto datům důvěřovat. Mohlo by dojít k chybě, která by stála peníze a poškodila podnik. Musí existovat dobrý důvod, proč informace shromažďovat. Pokud jsou shromažďovány irelevantní informace, ztrácí se tím čas i peníze a analýzy nebudou

tak přínosné. Aktuálnost odkazuje na to, jak aktuální informace jsou. Pokud byla data shromážděna v minulé hodině, pak jsou aktuální, pokud tedy nepřijdou nové informace, kvůli kterým budou předchozí informace zbytečné. Neaktuální informace mohou vést ke špatným rozhodnutím organizací. Mezi další znaky kromě výše jmenovaných může patřit integrita dat, dostupnost nebo efektivita [16].

Data v organizaci musí být kvalitní, důraz je přitom kladen hlavně na obchodní a finanční procesy. Tyto procesy ovlivňují existenci podniku. Mezi typická data organizace patří data z prodeje nebo fakturace. Kvalita se týká jak transakčních dat, například faktury nebo objednávky, tak i kmenových dat, mezi který patří data o zákaznících nebo dodavatelích [15].

Příčiny nekvalitních dat se mohou nacházet ve špatně nastavených procesech nebo pravidlech organizace, nedůslednost pracovníků při vytváření a ukládání dat, špatná implementace podnikových aplikací nebo jejich nesprávná funkcionalita. Organizace používají například nástroje pro čištění nekvalitních dat v datových skladech [15].

Mezi důsledky nekvalitních dat patří opakování procesů kvůli nekvalitním vstupům a neefektivnosti, chybné reporty kvůli duplicitním nebo nekonzistentním datům, ztráta znalostí kvůli odchodu zaměstnanců, ztráta zákazníků nebo zvýšená rizika. Důležitým krokem ke zkvalitnění dat je odstraňování příčin a ne jenom důsledků [15].

1.3 Dílčí souhrn

Význam dat ve společnosti stále narůstá a jejich správné pochopení lze využít ke zlepšení fungování podniku. Čím kvalitnější data, tím lépe mohou pomoci podniku při rozhodování a určování cílů. Podle toho, kde data vznikají, se dělí na interní a externí. Podle své struktury se data rozlišují na strukturovaná, nestrukturovaná, quasi strukturovaná a semi-strukturovaná.

2 BIG DATA

Slovní spojení Big data naznačuje, že se bude jednat o velký objem dat. V českém prostředí se ustálil tento anglický termín, nicméně se lze setkat i s překlady do češtiny, jako jsou velká data nebo veledata. V této bakalářské práci bude používán anglický termín „Big data".

Jak velká tato data jsou? Pro samotný pojem neexistuje přesně vymezená definice, kterou by každý uznával. Definice Big data se mohou lišit z pohledu na vlastnosti dat, technologie pro zpracování dat nebo z možných dopadů na společnost [7].

Jedna z nejpoužívanějších definic, známá také jako „3V" (z anglických slov Volume, Variety a Velocity), pochází od americké společnosti Gartner [5]:

„Big data jsou data, jejichž objem, rychlost a různorodost vyžadují efektivní a inovativní formy pro zpracování informací k lepšímu pochopení a rozhodování."

Definice od Microsoftu, která poukazuje na důležitost technologií, zní [7]:

„Big data popisují proces aplikování seriózního výpočetního výkonu – nejnovějšího ve strojovém učení a umělé inteligenci – na masivní a vysoce komplexní soubory informací."

Big data se chápou jako data, jejichž rostoucí objem, rychlost a různorodost jsou příliš náročné pro zpracování tradičními nástroji a proto je třeba hledat nástroje nové. Jako tradiční nástroj lze chápat například firemní databáze, které nedokáží s tak velkým množstvím dat pracovat. Proto musí firmy hledat nové možnosti, díky kterým lze data zpracovat a porozumět jim.

2.1 Historie Big data

Poprvé byla myšlenka Big data naznačena vědcem Peterem Denningem v roce 1990. Ve svém článku „Saving All the Bits" vyslovil větu [44]:

„Je možné stavět stroje, které umí rozpoznávat nebo předpovídat vzorce v datech, aniž by pochopily význam těchto vzorců. Takové stroje mohou být nakonec dostatečně rychlé, aby zvládly velké datové proudy v reálném čase."

Roku 1997 Michael Cox a David Ellsworth publikovali článek „Application-controlled demand paging for out-of-core visualization". Je to tedy poprvé, co byl použit termín Big data ve významu, jak je chápán i v dnešní době [50]:

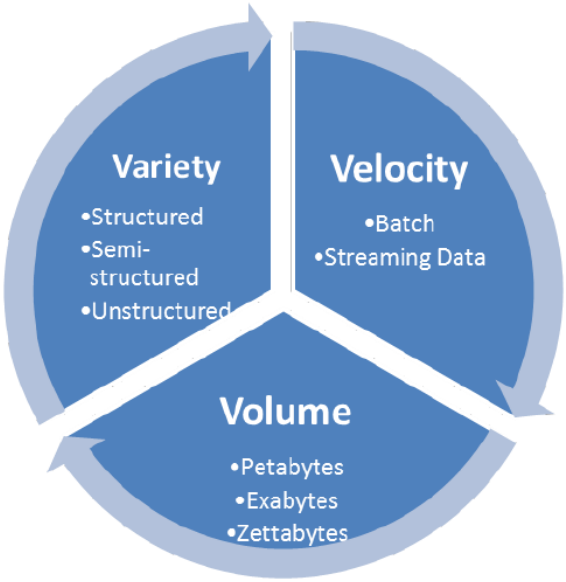
„Vizualizace představuje zajímavou výzvu pro počítačové systémy: datové sady jsou obecně poměrně velké a zdaňují kapacity hlavní paměti, místního disku a dokonce i vzdáleného disku.

Tomu říkáme problém velkých dat. Pokud se datové sady nevejdou do hlavní paměti (v jádru) nebo když se nevejdou ani na lokální disk, nejběžnějším řešením je získat více prostředků.“

V roce 2001 představil Doug Laney, analytik skupiny Meta Group (dnes Gartner), výzkumný dokument s názvem „3D Data Management: Controlling Data Volume, Velocity, and Variety.” V dokumentu jsou charakterizovány tři hlavní charakteristiky, které se obecně používají i v dnešní době – objem, rychlost a různorodost [51].

2.2 Charakteristiky Big data – 3V, 5V, 10V

Objem (Volume), rychlost (Velocity) a různorodost (Variety) jsou tedy základní vlastnosti Big data a označují se jako 3V (Obrázek 3). S rostoucím množstvím dat se postupně zaváděly i nové charakteristiky. Stejně jako u definic, ani u charakteristik není dána jednoznačnost, kolik „V“ se má používat. Mezi další z často užívaných charakteristik jsou důvěryhodnost (Veracity) a hodnota (Value) a dohromady tvoří 5V [3]. Existuje spousta dalších charakteristik, např. variabilita (Variability), zranitelnost (Vulnerability), platnost (Validity), volatilita (Volatility) nebo vizualizace (Visualization), což je označováno jako 10V [8]. V této práci je dále charakterizováno 5V.



Obrázek 3: Charakteristiky 3V

Zdroj: [17]

Objem (Volume)

Množství dostupných dat, která je potřeba analyzovat a zpracovávat, se stále rozrůstá. Více než 90 % dat vzniklo v posledních letech. Data dosahují velikostí v rámci terabytů až zettabytů. Takto velký objem dat nemůže být zpracován běžnými nástroji a vyžaduje tedy odlišné technologie ke zpracování. Bez vhodných technologií přicházejí organizace o možnost získat potencionální výhody z neprozkoumaných dat [3].

Různorodost (Variety)

Data mohou být v různých datových formátech. Mezi hlavní tři typy dat patří strukturovaná, nestrukturovaná a semi-strukturovaná. Různorodost vyžaduje odlišné možnosti zpracování dat. Uvádí se, že více jak 80 % ze všech dat proudících organizacemi, se nachází ve nestrukturované formě. [10].

Rychlost (Velocity)

Rychlost představuje problém zrychlení generování nových dat a současně se zvyšující požadavky na zrychlení jejich zpracování tak, aby byly výsledky dodány v reálném čase [10]. Data mohou být uložena a zpracována různými způsoby. Mezi tyto způsoby patří dávková zpracování, zpracování near-time (téměř v reálném čase), real-time (v reálném čase) nebo data streaming (datové proudy) [18]:

- Dávkové zpracování – Data jsou zpracována až když jsou k dispozici výpočetní prostředky. Dochází ke zpracování bloků dat, které jsou po určitou dobu uloženy na nějakém místě. Příkladem je zpracování všech transakcí ve finanční společnosti za poslední týden. Data obsahují miliony denních záznamů, které mohou být uloženy jako soubor. S konkrétním souborem budou prováděné požadované analýzy.
- Zpracování near-time – Data se zpracovávají při výskytu události, ale nedokážou zareagovat okamžitě. Zákazník odešle žádost o kreditní kartu, ale musí počkat hodinu než bude žádost vyhodnocena.
- Zpracování real-time – Data jsou zpracována téměř v době, kdy jsou zadána nebo kdy jsou potřebná. Lze tak reagovat na události v okamžiku jejich vzniku. Zákazník například odešle žádost o kreditní kartu a během několik vteřin obdrží odpověď.

- Data streaming – Umožňuje zpracovávat data v reálném čas hned jakmile dorazí a detekovat podmínky v krátkém období od přijetí. Data lze po zpracování vkládat do nástrojů pro analýzu a získat tak okamžité výsledky.

Důvěryhodnost (Veracity)

Důvěryhodnost nebo také pravdivost se týká kvality analyzovaných dat. Určuje jak přesná nebo pravdivá data mohou být [10]. Správná interpretace Big data zajišťuje, že jsou výsledky relevantní a proveditelné. Důvěryhodnost pomáhá vyfiltrovat to, co je důležité a vytváří hlubší porozumění datům a jejich kontextu [19].

Hodnota (Value)

Hodnota patří mezi nejdůležitější vlastnosti Big dat. Organizace jsou schopny lépe porozumět svým zákazníkům, zacílit na ně své produkty, optimalizovat vlastní procesy nebo zlepšit výkonnost podniku [3].

2.3 Uplatnění Big data

Důležitost dat se netýká toho, kolik se jich v podniku nachází, ale toho, jak se s daty pracuje a jak jsou využívána. Každá organizace používá data svým vlastním způsobem. Čím účinněji, tím více roste i potenciál dat.

2.3.1 Úspory a podpora

Správně provedená analýza dat může přinést následující dopady [20]:

- úspora nákladů;
- rychlejší a lepší rozhodování;
- porozumění tržním podmínkám;
- kontrola online reputace;
- posílení akvizice a retence zákazníků;
- řešení problému inzerentů;
- síly inovací a vývoje produktů.

Úspora nákladů

Některé nástroje Big data mohou podnikům přinést výhody týkajících se jejich nákladů. Pomáhají při vyhledávání efektivnějších způsobů podnikání [23].

Rychlejší a lepší rozhodování

Vysoká rychlost nástrojů může snadno identifikovat nové zdroje dat, které pomáhají podnikům okamžitě analyzovat data a rychle se rozhodovat na základě získaných poznatků [23].

Porozumění tržním podmínkám

Analýzou Big data můžete lépe porozumět současným tržním podmínkám. Například analýzou nákupního chování zákazníků může společnost najít produkty, které se prodávají nejvíce a podle tohoto trendu vyrábět produkty. Tím se může společnost dostat před své konkurenty [23].

Kontrola online reputace

Nástroje Big data dokážou analyzovat myšlení zákazníků, proto může podnik získat zpětnou vazbu o tom, kdo co říká o daném podniku. Nástroje mohou pomoci se sledováním a zlepšováním online přítomnosti podniku [23].

Posílení akvizice a retence zákazníků

Zákazník je nejdůležitějším aktivem, na kterém závisí podnikání. Neexistuje jediný podnik, který by si mohl nárokovat úspěch, aniž by musel nejprve vytvořit solidní zákaznickou základnu. Avšak i se zákaznickou základnou si podnik nemůže dovolit přehlížet vysokou konkurenci, před kterou stojí. Pokud se firma pomalu dozví, co zákazníci hledají, pak je velmi snadné začít nabízet výrobky nízké kvality. Nakonec dojde ke ztrátě klientely, což celkově nepříznivě ovlivní obchodní úspěch. Použití Big data umožňuje podnikům sledovat různé vzory a trendy související se zákazníky. Pozorování chování zákazníků je důležité pro vyvolání loajality [20].

Řešení problému inzerentů

Analytika Big data může pomoci změnit všechny obchodní operace. To zahrnuje schopnosti jako přizpůsobovat se očekávání zákazníků, změnit produktovou řadu společnosti nebo zajistit, aby marketingové kampaně byly silnější [20].

Síly inovací a vývoje produktů

Další obrovskou výhodou Big data je schopnost pomoci podnikům se inovovat a rozvíjet své produkty [20].

2.3.2 Příklady využití v současnosti

Maloobchod

Na základě Big data lze získat informace o chování zákazníků nebo jejich nákupních preferencích. To mohou obchody využít ke zkvalitnění poskytovaných služeb a vytváření služeb na míru. Mezi příklady dat v maloobchodu patří přizpůsobování výrobků a služeb zákazníkům na míru, nabídka zboží zákazníkům přes mobilní notifikace nebo evidence všech informací o zákazníkovi a následné nabídce individuálních služeb [21].

Průmysl

Využití dat v průmyslu zasahuje od výroby až po servis strojů. Příkladem může být monitorování a optimalizace výroby, snižování spotřeby materiálu a energie, vzdálené řízení výroby nebo prediktivní servis, kdy se minimalizují nečekané odstávky výroby [22].

Logistika a doprava

Logistikou proudí obrovské hromady zboží a tím se vytváří i hromady dat. V logistice jsou tři základní možnosti využití Big data. První možností je zlepšení provozu a optimalizace tras. Druhou možností je předvídání času při dodání zboží. Další z možností je odhadování množství zasílaných balíků podle počasí nebo chřipkového období [22].

Zdravotnictví

Big data mají ve zdravotnictví obrovský potenciál. Existují dvě možnosti, jak data o pacientech využívat. První možností je telemedicína, kde je zjišťován stav pacienta na dálku a dochází k případnému poskytování služeb. Druhá možnost je využívání dat při diagnostice onemocnění, k čemuž pomáhá umělá inteligence [21].

Zemědělství

Big data se využívá i v zemědělství k usnadnění a zefektivnění práce nebo ke zvýšení produkce. Příkladem ze zemědělství jsou senzory pro dobytek, které kontrolují stav a chování zvířat. Na trhu se nacházejí moderní traktory, drony nebo další stroje, které jsou naváděny pomocí GPS [22].

Energetika

V energetickém sektoru má analýza získaných dat potenciál pro obchodní i strategické účely. Využití analýzy dat se nabízí v oblastech skladování energie, řízení elektrizačních soustav

a stability nebo řízení spotřeby energie. Také lze pomocí získaných dat nabídnout zákazníkovi lepší služby [22].

Bankovnictví

V Bankovnictví se získaná data týkají např. převodů, plateb nebo výběrů z bankomatů. Podle transakcí je možné určit pro koho klient pracuje, jaká je jeho pozici, kdo jsou jeho kolegové v práci a rodinní příslušníci. Big data se využívají pro lepší komunikace se zákazníky a nabízení lepších služeb na míru [21].

Pojišťovnictví

Pojišťovny mají k dispozici data o klientech. Pomocí dat lze detekovat pojistné podvody. Analýza umožňuje zlepšit poskytované služby, jako je např. určení, které klienty na pobočku pozvat, vyřizování požadavků online nebo určení rizikovosti konkrétní osoby podle jízdního chování [21].

2.4 Budoucnost Big data

Jakmile globální data začala exponenciálně růst před deseti lety, nevykazovala žádné známky zpomalení. Agregují se hlavně přes internet, včetně sociálních sítí, požadavků na vyhledávání na webu, textových zpráv a mediálních souborů. Další obrovský podíl dat vytvářejí zařízení a senzory IoT. Jsou to klíčové faktory růstu globálního trhu Big data. Mezi předpověďmi o budoucnosti Big data v příštích pěti letech jsou tyto příklady [24]:

- Objemy dat se budou dále zvyšovat a migrovat do cloudu.
- Strojové učení se bude dále vyvíjet.
- Poroste zájem o datové vědce.
- Zvýší se nejistota ochrany osobních údajů.
- Dojde k nárustu rychlých a akčních dat.

Budoucnost Big data z pohledu organizací

Analýza Big data slibuje změnu způsobu, jakým podniky fungují ve financích, zdravotnictví, výrobě a dalších odvětvích. Drtivá velikost Big data může v budoucnu způsobit další výzvy, se kterými se organizace budou muset vypořádat. Hrozí rizika v oblasti ochrany osobních údajů a bezpečnosti, nedostatku odborníků na data a problémů při ukládání a zpracování dat [24].

Většina odborníků se však shoduje, že Big data budou znamenat velkou hodnotu. Vzniknou nové kategorie pracovních míst a dokonce celá oddělení odpovědná za správu dat ve velkých organizacích. Objeví se nové regulační struktury a standardy chování, protože společnosti nadále používají osobní údaje spotřebitelů. Většina společností také přejde z generování dat na data založená na aktivních datech a obchodních poznatcích [24].

2.5 Rizika Big data

Big data jsou pro většinu podniků nesmírně ohromující koncept. Přináší řadu výzev a rizik. Jasně stanovený plán a správná technologie pomohou úspěšně bojovat proti rizikům Big data. Mezi největší rizika podniků patří [26]:

- neorganizovanost dat;
- ukládání a uchovávání dat;
- náklady;
- nekompetentní analytika;
- ochrana osobních údajů.

Neorganizovanost dat

Big data jsou vysoce univerzální. Pochází z mnoha zdrojů a v mnoha podobách. Data jsou buď strukturovaná nebo nestrukturovaná a pocházejí z online a offline zdrojů. Všechna tato data se hromadí každý den, každou minutu. Je ohromující, že podniky účinně řeší takové neorganizované a zazděné datové soubory. Dobře naplánovaná strategie správy vás může zbavit temných dat a pomoci vám to pochopit [25].

Ukládání a uchovávání dat

Toto je jedno z nejviditelnějších rizik spojených s Big data. Když se data hromadí tak rychlým tempem a v tak obrovských objemech, prvním problémem je jeho ukládání. Tradiční metody a technologie ukládání dat nestačí k uložení velkých dat a jejich zachování. Podniky dnes potřebují přechod k cloudovým řešením pro ukládání dat, aby mohly účinně ukládat, archivovat a přistupovat k velkým datům [26].

Náklady

Proces ukládání, archivace, analýzy, vykazování a správy velkých dat zahrnuje náklady. Mnoho malých a středních podniků si myslí, že Big data jsou pouze pro velké podniky a nemohou si to

dovolit. Při pečlivém rozpočtování a plánování zdrojů však lze náklady na Big data zmírnit [25].

Nekompetentní analytika

Bez řádné analýzy jsou Big data v organizaci k ničemu. Analytika je to, co dělá data smysluplnými, což managementu poskytuje cenné informace pro přijímání obchodních rozhodnutí a plánování strategií růstu. Vzhledem k tomu, že data rostou tak alarmujícím tempem, zjevně chybí kvalifikovaní odborníci a technologie pro efektivní analýzu Big data. Podniku hrozí nesprávné pochopení dat a nesprávné rozhodování. Používání správných nástrojů je zásadní pro přijímání příslušných rozhodnutí z velkého datového projektu [25].

Ochrana osobních údajů

Big data představují riziko soukromí dat. Podniky na celém světě používají citlivé údaje, osobní informace o zákaznících a strategické dokumenty. Bezpečnostní incident může vést k soudním sporům a přísným trestům. Přijetí opatření na ochranu osobních údajů již není jen dobrá iniciativa, je to nutnost dodržování předpisů [25].

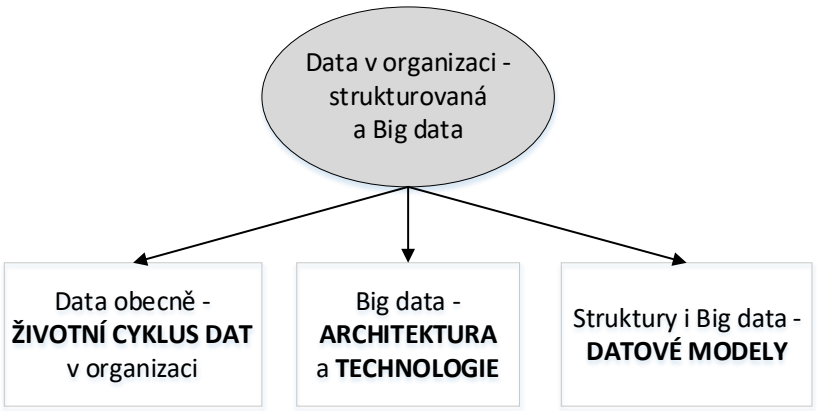
2.6 Dílčí souhrn

Základní vlastnosti Big data jsou určeny jako – objem, rychlost a různorodost dat. Pokud se data v podniku týkají alespoň jedné z těchto vlastností, lze je označit jako Big data. Určování základních charakteristik se postupně rozšiřovalo, a proto existují skupiny vlastností označované kromě 3V také 5V nebo 10V. Existuje mnoho uplatnění, jak mohou podniky z Big data těžit. Mezi hlavní uplatnění patří zisk, úspora nákladů nebo větší automatizace podniku. Na druhou stranu díky svým vlastnostem hrozí u Big data i některá rizika. Ta se mohou týkat nákladů nebo ztrátou kontroly nad organizovaností dat.

3 URČENÍ HLEDISEK: ŽIVOTNÍ CYKLUS, ARCHITEKTURA, TECHNOLOGIE, MODELOVÁNÍ

Cílem bakalářské práce je charakteristika strukturovaných a nestrukturovaných dat v organizaci se zaměřením na Big data. Pro další zpracování bakalářské práce jsem se rozhodl se zaměřit na čtyři vybraná hlediska, které schematicky vyjadřuje Obrázek 4.

První z popsaných hledisek je životní cyklus dat. Hledisko popisuje tok dat organizací z obecného pohledu, od požadavků na data až po jejich využívání. Další dvě z hledisek – architektura a technologie se týkají Big data. Architektura popisuje prostředí Big data a jeho funkčnost. K charakterizování typických technologií pro práci s Big data slouží hledisko technologie. Poslední hledisko se zaměřuje na datové modely u SQL a NoSQL databází, týká se tedy jak strukturovaných, tak i nestrukturovaných dat.

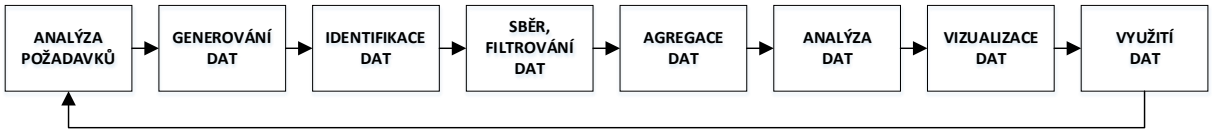


Obrázek 4: Určení hledisek pro zpracování bakalářské práce

Zdroj: Vlastní zpracování

4 HLEDISKO: ŽIVOTNÍ CYKLUS DAT

Strukturovaná a nestrukturovaná data jsou data, která procházejí podnikem. Proto je do bakalářské práce zařazena obecná kapitola o životním cyklu dat. Na začátku procesu je analýza požadavků na data od uživatelů. U koncových fází životního cyklu se hledá odlišný přístup. Jedná se hlavně o způsoby jak data využít a jak je vizualizovat. Následující kapitola nerozlišuje strukturovaná a nestrukturovaná dat, charakterizuje data z obecného pohledu. V kapitole budou popsány jednotlivé činnosti životního cyklu (Obrázek 5).



Obrázek 5: Životní cyklus dat v organizaci

Zdroj: Vlastní zpracování

4.1 Analýza požadavků

Prvním krokem životního cyklu dat je zjištění, jaká data budou potřebná pro dosažení cílů organizace. K tomuto účelu slouží analýza požadavků. Analýza požadavků zahrnuje úkoly, které vstupují do rozhodování o cílech podniku [41].

Analýzu požadavků lze rozdělit na sběr požadavků, analýzu požadavků a zaznamenání požadavků. Při sběru požadavků dochází k přímé komunikaci s uživatelem a následné dotazování na požadavky systému. Po sběru následuje samotná analýza požadavků, která identifikuje a řeší nepřesné a nejasné požadavky. Při zaznamenávání požadavků dochází k jejich dokumentování [47].

Analýza požadavků znamená správně analyzovat, dokumentovat, validovat a spravovat softwarové nebo systémové požadavky. Kvalitní požadavky jsou zdokumentovány, lze je měřit, testovat, sledovat a pomáhají k identifikaci obchodních příležitostí. Usnadňují také návrh systému [41].

Mezi cíle požadavků patří [41]:

- Vymezení funkčnosti systému.
- Lepší pochopení vývojářů na systém.

- Definování omezení.
- Odhad nákladů a času na vytvoření systému.

4.2 Generování dat

Převážná část generovaných dat pochází ze tří primárních zdrojů: sociální data, strojová data a transakční data. Kromě toho musí společnosti rozlišovat mezi interně generovanými daty a externě generovanými daty. Důležitým faktorem je také to, zda jsou data nestrukturovaná nebo strukturovaná. Nestrukturovaná data nemají předdefinovaný datový model, proto vyžadují více zdrojů, aby byla pochopena [32].

Zdroje strukturovaných dat

Strukturovaná data mohou být generovaná buď lidskou činností nebo stroji. Mezi příklady dat generovaných strojem patří senzorová data, například GPS. Dále se může jednat o zachycování dat o činnostech serveru nebo sítě během jejich fungování. Taková data mohou předvídat narušování bezpečnosti. Spousta finančních systémů je předdefinováno podle pravidel, které automatizují procesy. Příkladem jsou data o obchodování s akciemi. Posledním z uvedených příkladů dat generovaných stroji jsou data generovaná při nákupu [3].

Mezi strukturovaná data generovaná člověkem patří jakákoliv vstupní data, které člověk vloží do počítače, např. svoje osobní údaje. Taková data umožňují podnikům chápat chování zákazníka. Dalším příkladem jsou click-stream data, u kterých dochází ke generování dat při každém kliknutí člověka na webové stránce. Tento způsob opět odhaluje chování zákazníka. Mezi další příklad dat generovaných člověkem patří data z počítačových her [3].

Zdroje nestrukturovaných dat

Stejně jako u strukturovaných dat, i nestrukturovaná data mohou být generována buď z lidské činnosti nebo pomocí strojů. Mezi nestrukturovaná data generovaná člověkem patří textové soubory (zpracování textu, tabulek, prezentací), e-maily, data ze sociálních médií jako je Facebook nebo Twitter, z webů jako jsou YouTube nebo Instagram. Data lze získávat také z textových zpráv nebo podle polohy na mobilním telefonu. Dále se data získávají z komunikace (chatování nebo telefonní hovory). Data mohou být generována i z různých médií, jako jsou MP3, digitální fotografie, videa nebo zvukové záznamy [12].

Mezi nestrukturovaná data generovaná strojem patří satelitní snímky, které dokáží získat údaje o počasí nebo např. údaje o pohybech vojáků. Vědecká data mohou být získávána

z průzkumu vesmíru nebo zkoumání ropy a zemního plynu. Senzorová data dokáží získat data například o dopravě nebo o počasí [12].

Zdroje semi-strukturovaných dat

Mezi zdroje semi-strukturovaných dat patří např. značkovací jazyk XML, datový formát JSON nebo NoSQL databáze. Zdrojem semi-strukturovaných dat může být stejně jako u nestrukturovaných dat i e-mail [13].

4.3 Příprava a sběr dat

Výzkumy odhadují, že proces přípravy dat zabere až 80 % celkového času analýzy. Pro podniky je to i nadále hlavní překážkou pro rychlou a přesnou analýzu. Proces přípravy dat umožňuje komukoli rychle přeměnit veškerá nezpracovaná data z více zdrojů na rafinovaná informační aktiva, takže je lze použít pro přesnou analýzu a cenné obchodní poznatky. Samoobslužný proces přípravy dat se rychle stává dovedností, která je vyžadována pro rostoucí počet analytiků dat, vědců údajů a obchodních uživatelů. Tito jednotlivci se učili a osvojovali si tuto novou dovednost na podporu svých každodenních obchodních zpravodajských aktivit a analytických iniciativ [33].

Identifikace dat

Prvním krokem přípravy dat je identifikace datových souborů a jejich zdrojů. Důkladněji provedená identifikace může odhalit skryté vzorce a korelace. Mnoho organizací si je vědomo, že mnoho interně generovaných dat nebylo v minulosti plně využito k plnému potenciálu. Díky využití nových nástrojů získávají organizace nový přehled o dříve nevyužitých zdrojích nestrukturovaných dat v e-mailech, záznamech služeb zákazníkům, datech senzorů a bezpečnostních protokolech. Kromě toho existuje velký zájem o hledání nového úhlu pohledu založeného na analýze dat, která jsou primárně vně organizace, jako jsou sociální média, umístění mobilního telefonu, provoz a počasí. Nové a neočekávané vztahy a korelace mezi elementy se mohou projevit pouze zkoumáním velmi velkých objemů dat. Tyto vzorce mohou například poskytnout nahlédnutí do preferencí zákazníka pro nový produkt [31].

Sběr a filtrování dat

Další částí přípravné fáze je sběr a filtrování dat. Během této fáze dochází ke shromažďování dat ze všech zdrojů identifikovaných v předchozím kroku. Data jsou podrobena automatizovanému filtrování pro odstranění poškozených dat nebo dat nepotřebných.

U externích nestrukturovaných dat se může často jednat o irelevantní data a budou tak vyřazena z filtrování. Data, která jsou označena jako poškozená, mohou obsahovat záznamy s chybějícími nebo neplatnými hodnotami a datovými typy. Data filtrovaná pro jednu analýzu, mohou být pro jinou analýzu cenná. Proto je vhodné před pokračováním uložit kopii původního soubor dat. Některá z dat identifikovaných pro analýzu mohou dorazit v nekompatibilním formátu. Rozdílné typy dat pocházejí častěji z externích zdrojů. Extrakce nesourodých dat slouží k transformaci do formátu, který lze použít pro analýzu. Rozsah požadované extrakce a transformace závisí na typech analýz a možnostech řešení. Neplatná data mohou zkreslovat výsledky analýzy. Fáze čištění tedy slouží k vytvoření pravidel validace a odstranění všech známých neplatných dat. Řešení Big data často přijímají nadbytečná data různých typů. Toho lze využít k prozkoumání vzájemných propojení mezi daty. Výsledkem pak je sestavení validačních parametrů a doplnění chybějících platných dat [31].

Agregace dat

Data mohou být rozložena do více datových sad, nebo se mohou vícekrát nacházet ve více souborech. Z tohoto důvodu je vyžadováno, aby byla data propojena pomocí polí nebo určit soubor, který představuje správnou hodnotu. Provedení agregace může být komplikované, buď kvůli rozdílům ve struktuře dat, nebo z pohledu sémantiky. Časově se může jednat o velice náročnou operaci. Propojení těchto dat může vyžadovat složitou logiku, avšak provádí se automaticky bez potřeby zásahu člověka. V této fázi je třeba zvážit budoucí požadavky na analýzu dat, kvůli podpoře opětovné použití dat. Stejná data mohou být ukládána ve více formách, ovšem jedna forma může být vhodnější pro určitý typ analýzy než jiná [31].

4.4 Analýza dat

Fáze analýzy se může provádět opakovaně, dokud nebude objeven příslušný vzorec nebo korelace. Jedná se hlavně o případ průzkumných analýz dat. Náročnost této fáze závisí na požadovaném výsledku analýzy. Analýzu dat lze klasifikovat jako konfirmační nebo průzkumnou spojenou s dolováním dat. U potvrzující analýzy dat je předem navržena příčina vyšetřovaného jevu. Navrhovaná příčina nebo předpoklad se nazývá hypotéza. Data se poté analyzují, aby se prokázala nebo vyvrátila hypotéza a poskytly se odpovědi na konečné otázky. Obvykle se používají techniky vzorkování dat. Neočekávané nálezy nebo anomálie se mohou ignorovat [31].

4.5 Vizualizace dat

Fáze vizualizace se věnuje používání technik nástrojů k vytvoření grafického výsledku. Lidé v podniku musí být schopni porozumět výsledkům, aby získali hodnotu z provedené analýzy. Vizualizace umožňuje získávat odpovědi na otázky, které uživatelé předtím neznali. Stejné výsledky mohou být prezentovány řadou různých způsobů, které mohou ovlivnit interpretaci výsledků [36].

4.6 Využití dat

Po zpřístupnění výsledků mohou být objeveny další možnosti, jak využít výsledky řešené analýzy. Podle povahy řešených problémů existuje šance na vytvoření modelu, který zapouzdří nové poznatky o povaze vzorů a vztahů, které existují v datech. Model může být buď matematická rovnice, nebo soubor pravidel. Modely lze použít ke zlepšení logiky podnikových procesů a logiky aplikačních systémů a mohou tvořit základ nového systému nebo softwarového programu [31].

Výsledky analýzy mohou být ručně nebo automaticky vkládány přímo do podnikových systémů, aby se zlepšilo jejich chování a optimalizoval se výkon. Příkladem může být zákazník na internetovém obchodu, o kterém analýza zpracovala výsledky. Tomuto zákazníkovy se budou podle výsledků generovat doporučené produkty ke koupi. Modely mohou být použity ke zlepšení logiky programování v podnikových systémech nebo mohou být základem nového systému. Díky identifikovaným vzorcům, korelacím a anomáliím může dojít k optimalizaci podnikových procesů. Příkladem je konsolidace přepravních tras. Modely mohou vést k příležitostem ke zlepšení logiky obchodních procesů. Výsledky analýzy mohou také varovat před možnými výstrahami [31].

4.7 Dílčí shrnutí

Životní cyklus dat začíná sběrem požadavků. Účelem sběru požadavků je zajištění, že vytvářená a spotřebovávaná data budou splňovat obchodní cíle podniku a budou srozumitelná pro všechny zúčastněné strany. Data mohou být generována z mnoha různých zdrojů, převážně v nestrukturované formě. Příprava dat zahrnuje kroky, jako jsou identifikace dat, filtrování dat, sběr dat a agregace. V analýze dat dochází k vyhledávání vzorců a korelací mezi daty. Vizualizace usnadňuje uživatelům pochopení vztahů mezi daty. Po pochopení vztahů lze získat nové poznatky o datech a tyto poznatky využít v prospěch podniku.

5 HLEDISKO: ARCHITEKTURA BIG DATA

Organizace zajímající se o Big data by měly mít základní znalosti o tom, jak jsou prostředí Big data navržena, jak jsou provozována v podnikových prostředích a jak data proudí různými vrstvami organizace. Pochopení základů architektury Big Data pomůže s rozhodováním a porozuměním, jak se k sobě hodí jednotlivé komponenty Big Data.

Aby bylo možné těžit z potenciálu Big data, je nutné mít k dispozici potřebné technologie pro analýzu obrovského množství dat. Protože je Big Data vývojem z tradiční analýzy dat, měly by se technologie Big Data přizpůsobit stávajícímu IT prostředí podniku. Z tohoto důvodu je užitečné mít společnou strukturu, která vysvětluje, jak Big Data doplňují a jak se liší od stávajících analytik, BI, databází a systémů. Společná struktura se nazývá referenční architektura [27].

5.1 Referenční architektura

Referenční architektura je dokument nebo soubor dokumentů, která poskytuje doporučené struktury a integrace produktů a služeb IT k vytvoření řešení. Referenční architektura ztělesňuje uznávané osvědčené postupy v oboru a obvykle navrhuje optimální způsob doručení pro konkrétní technologie, jako jsou hardware, software, procesy, specifikace a konfigurace, jakož i logické komponenty a vzájemné vztahy. Architektury pomáhají projektovým manažerům, vývojářům softwaru, podnikovým architektům a IT manažerům spolupracovat a efektivně komunikovat o implementačním projektu. Referenční architektura předpokládá a odpovídá na nejčastější otázky. V důsledku toho pomáhají týmům vyhnout se chybám a zpožděním, ke kterým může dojít bez použití testované sady osvědčených postupů a přístupů k řešení [34].

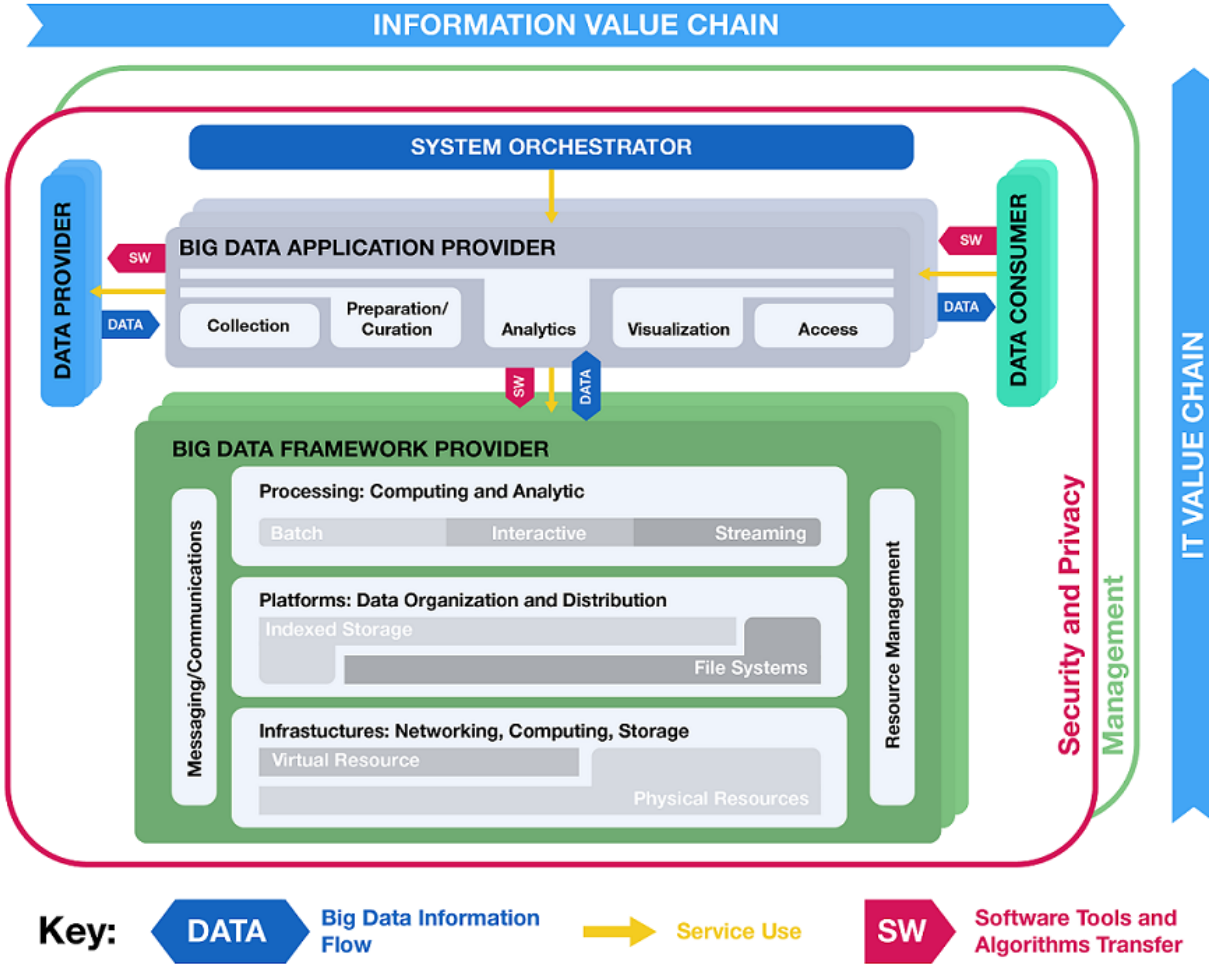
Cílem referenční architektury je vytvořit otevřený standard, který může každá organizace využít ve svůj prospěch. Národní institut pro standardy a technologie (NIST), jedna z předních organizací ve vývoji norem, vyvinula takovou referenční architekturu – NIST Big Data Reference Architecture (NBDRA – Obrázek 6) [27].

Důvody pro použití referenční architektury [27]:

- Společný jazyk pro různé zúčastněné strany;
- Dodržování společných norem, specifikací a vzorů;
- Konzistentní metody implementace technologie k řešení podobných sad problémů;

- Porozumění různým komponentám, procesům a systémům Big Data v kontextu konceptuálního modelu Big Data zaměřeného na dodavatele a technologie;
- Analýza kandidátských standardů pro interoperabilitu, přenositelnost, opakovatelnost a rozšiřitelnost.

NBDRA je přístup nezávislý na prodejci a může jej použít jakákoli organizace, která si klade za cíl vyvinout architekturu Big Data. Architektura Big Data představuje systém Big Data složený z pěti logických funkčních komponent spojených prostřednictvím rozhraní služeb [27].



Obrázek 6: NBDRA

Zdroj: [27]

NBDRA je organizována kolem pěti hlavních rolí a více pod-rolí zarovnaných podél dvou os představujících dva hodnotové řetězce Big Data: Informační hodnota na vodorovné ose a Informační technologie na vertikální ose. V ose Informační hodnota je hodnota vytvářena prostřednictvím sběru dat, integrace, analýzy a aplikace výsledků sledujících hodnotový

řetězec. V rámci osy IT je hodnota vytvářena poskytováním sítí, infrastruktury, platforem, aplikačních nástrojů a dalších IT služeb pro hostování a provozování Big data na podporu požadovaných datových aplikací. Na průsečíku obou os je role poskytovatele aplikací velkých dat, což naznačuje, že analytika dat a její implementace poskytují hodnotu zúčastněným stranám velkých dat v obou hodnotových řetězcích [27].

5.2 Komponenty Big data architektury

Pět hlavních rolí NBDRA představuje logické komponenty každého prostředí Big Data a jsou přítomny v každém podniku [27]:

- Systémový orchestrátor (System orchestrator);
- Poskytovatel dat (Data provider);
- Poskytovatel Big data aplikací (Big data application provider);
- Poskytovatel Big data frameworků (Big data framework provider);
- Spotřebitel dat (Data consumer).

Dvě dimenze, kterých se role týkají, jsou [27]:

- Řízení;
- Zabezpečení a soukromí.

Tyto dimenze poskytují služby a funkčnost pěti hlavním rolím v oblastech specifických pro Big Data a jsou zásadní pro jakékoli řešení Big Data [27].

Systémový orchestrátor

Systémová orchestrace je automatizované uspořádání, koordinace a správa počítačových systémů, middlewaru a služeb. Zajišťuje, aby různé aplikace, data a komponenty infrastruktury prostředí Big Data spolupracovaly. Aby toho bylo dosaženo, jsou využity pracovní postupy, procesy automatizace a řízení změn. Prostředí IT pro velké údaje sestává z kolekce mnoha různých komponent aplikací, dat a infrastruktury. Orchestrátor zajišťuje, že všechny tyto komponenty spolupracují synchronně [28].

Poskytovatel dat

Poskytoval dat zavádí do systému Big Data nová data nebo informační zdroje pro vyhledávání, přístup a transformaci. Jednou z klíčových vlastností Big Data je její rozmanitost, což znamená, že data mohou přicházet v různých formátech z různých zdrojů. Vstupní data mohou mít

podobu textových souborů, obrázků, zvuku, webových blogů atd. Zdroje mohou zahrnovat interní podnikové systémy (ERP, CRM, Finance) nebo externí systém (zakoupená data, sociální zdroje). V důsledku toho mohou mít data z různých zdrojů různé aspekty zabezpečení a ochrany soukromí [52].

K přenosu dat dochází mezi poskytovatelem dat a poskytovatelem Big data aplikací. K tomuto přenosu dat obvykle dochází ve třech fázích: zahájení, přenos a ukončení. Zahajovací fáze zahajuje jedna ze dvou stran a často zahrnuje určitou úroveň autentizace. Fáze přenosu dat tlačí data směrem k poskytovateli Big data aplikací. Ukončovací fáze kontroluje, zda byl přenos dat úspěšný a zaznamenává výměnu dat [27].

Poskytovatel Big data aplikací

Poskytovatel Big data aplikací je komponenta architektury, která obsahuje obchodní logiku a funkčnost, která je nezbytná pro transformaci dat do požadovaných výsledků. Společným cílem této komponenty je extrahovat hodnotu ze vstupních dat a zahrnuje následující činnost [27]:

- sběr;
- příprava;
- analytika;
- vizualizace;
- přístup k datům.

Rozsah a typy aplikací, které se používají v této komponentě referenční architektury, se velmi liší a jsou založeny na povaze a podnikání podniku. Pro finanční podniky mohou aplikace zahrnovat software pro detekci podvodů, aplikace pro kreditní skóre nebo ověřovací software. Ve výrobních společnostech může být komponentou poskytovatele velkých datových aplikací správa zásob, optimalizace dodavatelského řetězce nebo software pro optimalizaci trasy [27].

Poskytovatel Big data frameworků

Poskytovatel frameworků nabízí zdroje a služby, které mohou být využity poskytovatelem aplikací. Dále také poskytuje základní infrastrukturu architektury Big data. V této komponentě jsou data ukládána a zpracovávána na základě návrhů optimalizovaných pro prostředí Big data [27].

Role poskytovatele Big data frameworků lze rozdělit na dílčí role [27]:

- Infrastruktura: vytváření sítí, výpočetní technika a ukládání
- Platformy: organizace a distribuce dat
- Zpracování: výpočetní a analytické

Vrstva infrastruktury se zabývá potřebami vytváření sítí, výpočetní techniky a ukládání, aby bylo zajištěno, že velké a rozmanité formáty dat mohou být ukládány a přenášeny nákladově efektivním, bezpečným a škálovatelným způsobem. Ve svém jádru je klíčovým požadavkem úložiště velkých dat to, že je schopen zpracovat velmi velké množství dat a neustále se přizpůsobovat růstu organizace a že může poskytovat operace vstupu nebo výstupu za sekundu (IOPS) nezbytné k dodání dat do aplikací. IOPS je měřítkem výkonu úložiště, které se dívá na rychlost přenosu dat [27].

Vrstva platformy je kolekce funkcí, které umožňují vysoce výkonné zpracování dat. Platforma zahrnuje možnosti integrace, správy a použití úloh zpracování dat. V prostředích velkých dat to efektivně znamená, že platforma musí usnadňovat a organizovat distribuované zpracování na řešeních distribuovaného úložiště. Jednou z nejrozšířenějších platformových infrastruktur pro řešení Big Data je open source framework Hadoop. Důvod, proč společnost Hadoop poskytuje tak úspěšnou platformovou infrastrukturu, je kvůli prostředí sjednoceného úložiště (distribuované úložiště) a zpracování (distribuované zpracování) [27].

Vrstva zpracování poskytuje funkce pro dotazování dat. Prostřednictvím této vrstvy jsou prováděny příkazy, které jsou s daty prováděny za běhu. Často je to prostřednictvím provádění algoritmu, který spouští úlohu zpracování. V této vrstvě probíhá vlastní analýza. Usnadňuje dosažení požadovaných výsledků a hodnoty [27].

Spotřebitel dat

Spotřebitelem dat může být skutečný koncový uživatel nebo jiný systém. Používá rozhraní nebo služby nabízené poskytovatelem Big data aplikací. Umožňuje získat přístup k zájmovým informacím. Rozhraní mohou zahrnovat hlášení o datech, získávání dat nebo jejich vykreslování Mezi činnosti spotřebitele dat patří [52]:

- hledání a načítání;
- stahování;
- analýza;
- reporting;
- vizualizace;

- práce s daty pro vlastní procesy.

5.3 Big data v organizaci – výhody a rizika

Výhody zavedení Big data architektury

Objem dat, která jsou k dispozici pro analýzu, každým dnem roste. A existuje více zdrojů datových proudů než kdy jindy, včetně dat dostupných ze senzorů provozu, senzorů stavu, protokolů transakcí a protokolů aktivit. Organizace musí být schopna dát datům smysl a použít je včas, aby ovlivnila kritická rozhodnutí. Použití architektury Big data může pomoci firmě ušetřit peníze a provádět kritická rozhodnutí, včetně [28]:

- Snížení nákladů. Velké datové technologie, jako je Hadoop a cloudová analytika, mohou výrazně snížit náklady, pokud jde o ukládání velkého množství dat.
- Rychlejší a lepší rozhodování. Pomocí streamingové komponenty architektury Big data se lze rozhodovat v reálném čase.
- Předpovídání budoucích potřeb a vytváření nových produktů. Big data pomohou s analýzou potřeb zákazníků a předpovídáním budoucích trendů.

Rizika Big data architektury

Po správném provedení může architektura ušetřit firmě peníze a pomůže předpovídat důležité trendy, ale není to bez problémů. Při práci s Big daty je důležité myslet na následující problémy [35]:

1) kvalita dat;
2) měřítko;
3) bezpečnost;
4) vyspělost technologií.

Ad 1) Kvalita dat

Kdykoli se pracuje s různými zdroji dat, tak je kvalita dat výzvou. Data mohou být duplicitní nebo mohou úplně chybět, proto by analýza dat nemusela být spolehlivá. Data se musí analyzovat a připravit, než se budou moci spojit s dalšími daty pro analýzu [35].

Ad 2) Měřítko

Hodnota velkých dat je v jeho objemu. To se však může stát také významným problémem. Pokud není architektura navržená pro zvětšení, může rychle narazit na problémy. Náklady

na podporu infrastruktury se mohou zvýšit nebo může dojít ke snížení výkonu. Oba problémy by se měly řešit ve fázi plánování budování architektury Big data [35].

Ad 3) Bezpečnost

Big data mohou poskytnout skvělý přehled o datech, ale je také náročné je chránit. O data se mohou zajímat podvodníci a hackeři – Mohou se pokusit přidat vlastní falešná data nebo přesouvat data za účelem získání citlivých informací. Kybernetický zločin dokáže zpracovat data a přemístit je do datového jezera podniku [35].

Ad 4) Vyspělost technologií

Mnoho technologií používaných v oblasti Big data se vyvíjí. Zatímco základní technologie Hadoop jako např. Hive a Pig se stabilizovaly, tak nové objevující se technologie, jako je např. Spark, přinášejí s každou novou verzí rozsáhlé změny a vylepšení [28].

5.4 Dílčí souhrn

Architektura Big data popisuje prostředí Big data, dále zobrazuje, jak toto prostředí je provozováno v podniku, jaké komponenty obsahuje a jaká data proudí mezi jednotlivými komponentami architektury. Referenční architektura je spojení technologií současného prostředí podniku a technologií Big data. Použití architektury Big data může pomoci firmě ušetřit peníze, provádět lepší rozhodování a předpovídat budoucí trendy. Použití architektury ale není bez problémů, mohou nastat problémy s kvalitou dat, bezpečností nebo problémy s technologiemi.

Organizace NIST vyvinula referenční architekturu NBDRA. NBDRA obsahuje pět hlavních komponent, které se týkají každého podniku. První z komponent je systémový orchestrátor, který dohlíží na ostatní komponenty a zajišťuje, aby komponenty správně spolupracovaly. Poskytovatel dat zavádí data do systému. Poskytovatel Big data aplikací obsahuje obchodní logiku a funkčnost architektury. Poskytovatel Big data frameworků nabízí zdroje a služby, které využívá poskytovatel Big data aplikací. Spotřebitelem dat může být skutečný koncový uživatel nebo jiný systém a používá rozhraní nebo služby k získávání informací.

6 HLEDISKO: TECHNOLOGIE PRO PRÁCI S BIG DATA

Technologie Big Data se vztahují na softwarové nástroje navržené za účelem analýzy, zpracování a extrahování informací z velkého množství nestrukturovaných nebo polostrukturovaných dat, které nelze zpracovat pomocí relačních databází nebo tradičních systémů zpracování [37].

Technologie Big data lze rozdělit do dvou hlavních typů – operativní a analytické [45]:

- Operativní technologie Big data produkují data, které vytvářejí lidé při svých běžných denních činnostech. Jedná se o sociální média, dále o online transakce, nebo data od konkrétní organizace (např. online obchodování, online nákupy, online rezervace aj.). Tato data lze považovat jako prvotní data, která se používají ve druhé technologie, analytické.
- Analytické technologie pak slouží k prozkoumávání a analýze těchto dat, tedy pracují s ohromným množstvím dat, která jsou zásadní pro podnikatelské rozhodnutí. Tyto technologie jsou využívány v oblastech jako např. akciový marketing, předpovědi počasí nebo nalézání souvislostí v záznamech o zdravotním stavu.

Obrázek 7: Kategorie technologií Big data

Zdroj: [37]

Datové technologie se dělí do čtyř hlavních kategorií podle jejich činnosti (Obrázek 7) [37]:

- Datové skladování
- Data mining
- Analýza dat

- Vizualizace dat

6.1 Technologie pro datové skladování

Technologií pro datové skladování se chápe infrastruktura úložiště, která je navržena speciálně pro ukládání, správu a načítání obrovského množství dat. Úložiště umožňuje ukládání a třídění dat takovým způsobem, že lze k datům snadno přistupovat, používat je a zpracovávat. Úložiště je také možné flexibilně škálovat podle potřeby [38].

Technologie mohou být klasifikovány buď jako úložiště přímo připojené k počítači (DAS) nebo síťové úložiště. Síťové uložiště lze dále rozdělit na NAS a SAN. NAS lze popsat jako servery sloužící k připojení technologií k ukládání dat k síti. SAN slouží jako síť k propojení úložných zařízení [38].

Distribuovaný úložný systém používá počítačové sítě k ukládání informací na více než jednom uzlu. Komponenty distribuovaného úložiště pro Big data lze klasifikovat do tří úrovní [27]:

- Systémové soubory – Bez systémových souborů by informace umístěné na paměťovém médiu představovaly jeden velký soubor dat. Rozdělením dat na části se mnohem snadněji identifikují. V oblasti Big data jsou důležitými systémovými soubory GFS a HDFS.
- Databáze – NoSQL
- Programovací modely – Big data jsou obvykle uložena na stovkách až tisících serverů. Pro přístup k těmto datům byly vyvinuty modely paralelního programování, které zvyšují výkon NoSQL databází. Příkladem je MapReduce.

Hadoop

Hadoop poskytuje řešení všech problémů v oblasti Big data. 90 % světových dat je nyní přesunuto do Hadoopu. Hadoop byl navržen pro ukládání a zpracování dat v distribuovaném prostředí pro zpracování dat s komoditním hardwarem s jednoduchým programovacím modelem. Může ukládat a analyzovat data přítomná v různých strojích s vysokými rychlostmi a nízkými náklady. Komponentami Hadoop jsou HDFS, MapReduce a YARN. Hadoop je využíván například společnostmi Microsoft nebo Intel [37].

Hadoop si lze představit jako výpočetní prostředí, které umožňuje distribuovat zpracování velkých datových souborů napříč klastry počítačů pomocí jednoduchých programovacích modelů. Je navržen tak, aby se rozšířil od jednotlivých serverů po tisíce strojů, z nichž každý

nabízí místní výpočet a úložiště. Spíše než spoléhat se na hardware, který poskytuje vysokou dostupnost, je knihovna sama navržena tak, aby detekovala a řešila selhání v aplikační vrstvě, takže poskytuje vysoce dostupnou službu na vrcholu clusteru počítačů, z nichž každý může být náchylný k selhání [42].

NoSQL

Databáze NoSQL nejsou tabulkové a ukládají data jinak než relační tabulky. NoSQL nevyžaduje předem danou datovou strukturu a dokáže přijímat data různých typů a velikostí. Tyto databáze sice mají volnější strukturu, ale na oplátku kladou vyšší nároky na procesor a úložiště [4].

Mezi výhody NoSQL patří [2]:

- Flexibilní škálovatelnost – NoSQL škálují horizontálně, tzn. zpracování každé úlohy mohou tyto systémy distribuovat v rámci clusterů (množiny uzlů).
- Flexibilní datový model – Nevyžadují databázové schéma.
- Orientace na efektivní čtení – Rychlé zpracování dotazů v obrovském množství a krátkém čase.
- Ekonomický aspekt – Na uzly se nekladou žádné speciální požadavky. Může se jednat o běžně dostupné a levné počítače.

Mezi výzvy NoSQL patří [2]:

- Zralost – Databáze jsou sice efektivní, ale ještě nedosahují prověřené robustnosti jako u relačních databází.
- Uživatelská podpora – Vývoj databází je překotný, ale k dispozici ještě není tolik kvalitních poskytovatelů.
- Administrace – Složitější instalace a údržba.
- Standardizace přístupu k datům – Každá databáze má svůj dotazovací jazyk a programátorské rozhraní (API). Složitější dotazy vyžadují lepší programátorské znalosti.
- Experti – Na trhu není dostatek expertů v oblasti NoSQL databází.

NewSQL

NewSQL databáze jsou moderní formou relačních databází, jejichž cílem je srovnatelná škálovatelnost s databázemi NoSQL při zachování transakčních záruk tradičních databázových

systémů. Příkladem NewSQL databáze je VoltDB. NewSQL lze charakterizovat těmito znaky [53]:

- SQL je primární mechanismus pro interakci aplikací.
- Podpora ACID pro transakce.
- Neuzamykatelný mechanismus řízení souběžnosti.
- Architektura poskytující mnohem vyšší výkon.

6.2 Technologie pro data mining

Data mining se týká technik hromadné těžby nebo extrakce dat, které se provádějí na velkých sadách nebo objemech dat. Data mining se primárně provádí za účelem získání a načtení požadovaných informací nebo vzorů z velkého množství dat. To se obvykle provádí na velkém množství nestrukturovaných dat, která jsou v průběhu času uložena. Dolování dat obvykle pracuje na algoritmech vyhledávání, upřesňování, extrakce a porovnávání dat. Velké dolování dat také vyžaduje podporu ze základních výpočetních zařízení, konkrétně z jejich procesorů a paměti, pro provádění operací a dotazů na data. Techniky a procesy těžby velkých dat jsou také používány v analytice Big data a BI k poskytování souhrnných, cílených a relevantních informací, vzorců nebo vztahů mezi daty, systémy, procesy a dalšími [39].

Presto

Presto je distribuovaný SQL Query Engine s otevřeným zdrojovým kódem pro spouštění interaktivních analytických dotazů proti zdrojům dat všech velikostí od gigabajtů po petabajty. Presto umožňuje dotazování dat v Hive, Cassandra, Relational Databases a Proprietary Data Stores. Presto využívají například společnosti Netflix nebo Facebook [37].

Rapid Miner

RapidMiner je centralizované řešení, které obsahuje velmi silné a robustní grafické uživatelské rozhraní, které uživatelům umožňuje vytvářet, poskytovat a udržovat prediktivní analýzu. Umožňuje vytvářet velmi pokročilé pracovní postupy, podporu skriptování v několika jazycích [37].

Elasticsearch

Elasticsearch je vyhledávač založený na knihovně Lucene. Poskytuje distribuovaný, plně kompatibilní fulltextový vyhledávač s webovým rozhraním HTTP a dokumenty JSON bez schémat [37].

6.3 Technologie pro analýzu dat

Cílem analýzy Big data je odhalit vzory a propojení, které by jinak byly neviditelné a které by mohly poskytnout cenné informace o uživatelích, kteří je vytvořili. Jako výsledek mohou podniky získat výhodu nad svou konkurencí a dělat lepší obchodní rozhodnutí. Sofistikované softwarové programy se používají pro analýzu Big data, ale nestrukturovaná data použitá v analýze Big data nemusí být vhodná pro běžné datové sklady. Vysoké požadavky na zpracování Big data mohou také způsobit špatné uložení tradičních datových skladů. V důsledku toho se objevily novější, větší prostředí a technologie pro analýzu dat [40].

Apache Kafka

Apache Kafka je platforma distribuovaného streamování. Streamovací platforma má tři klíčové schopnosti – vydavatel, odběratel a spotřebitel. Je to podobné frontě zpráv nebo systému podnikových zpráv [37].

Splunk

Splunk zachycuje, indexuje a koreluje data v reálném čase ve vyhledávacím úložišti, ze kterého může generovat grafy, sestavy, výstrahy, řídicí panely a vizualizace dat. Používá se také pro správu aplikací, zabezpečení a dodržování předpisů a také pro obchodní a webovou analýzu [37].

KNIME

KNIME umožňuje uživatelům vizuálně vytvářet datové toky, selektivně provádět některé nebo všechny kroky analýzy a kontrolovat výsledky, modely a interaktivní pohledy. KNIME je psán v Javě a založen na Eclipse a využívá svého mechanismu rozšíření k přidání pluginů poskytujících další funkčnost [37].

Apache Spark

Spark poskytuje funkce In-Memory Computing pro poskytování Speed – zobecněného prováděcího modelu, který podporuje širokou škálu aplikací, a API Java, Scala a Python pro snadný vývoj [37].

6.4 Technologie pro vizualizaci dat

Vizualizace je grafické znázornění informací a dat. Pomocí vizuálních prvků, jako jsou grafy, grafy a mapy, poskytují nástroje vizualizace dat přístupný způsob, jak vidět a porozumět

trendům, odlehlým hodnotám a vzorům v datech. V oblasti Big data jsou nástroje a technologie vizualizace dat nezbytné pro analýzu obrovského množství informací a pro rozhodování na základě údajů [43].

Tableau

Tableau je výkonný a nejrychleji rostoucí nástroj vizualizace dat používaný v Business Intelligence Industry. Analýza dat je u Tableau velmi rychlá a vytvořené vizualizace jsou ve formě dashboardů a pracovních listů [43].

6.5 Programovací jazyky pro práci s Big data

Existuje spousta programovacích jazyků, ale jen některé dokážou pracovat s Big data. Mezi nejpoužívanější z nich patří Python, jazyk R, Java a Scala. Některé z těchto jazyků jsou lepší pro rozsáhlé analytické úkoly, zatímco jiné vynikají v provozování Big data a IoT [49]. Tabulka 1 porovnává základní vlastnosti vybraných programovacích jazyků.

Tabulka 1: Porovnání vybraných programovacích jazyků

	Python	**R**	**Java**	**Scala**
Rychlost			✓	✓
Snadné používání	✓	✓		
Rychlé učení	✓			
Analýza dat	✓	✓		✓
Univerzálnost	✓		✓	✓
Podpora Big data	✓	✓	✓	✓
Rozhraní s jinými jazyky	✓	✓		

Zdroj: Zpracováno dle [49]

Python

Univerzální povaha jazyka Python znamená, že může být použit v širokém spektru případů použití. Programování Big data je jednou z hlavních oblastí aplikace. Python je univerzální a open-source programovací jazyk. Programování v Pythonu znamená méně kódování ve srovnání s jinými programovacími jazyky. Kromě toho Python automaticky nabízí pomoc

při identifikaci a přiřazení datových typů. Jako open-source podporuje Python více platforem. Může být tak spuštěn v různých prostředích, jako jsou Windows nebo Linux. Knihovny pro analýzu a manipulaci s Big data, jako jsou např. Pandas, NumPy nebo SciPy jsou založené na Pythonu. Jednou z nevýhod Pythonu v oblasti programování Big data je jeho menší rychlost [30].

Jazyk R

Jazyk R je statistický jazyk a používá se k vytváření datových modelů, které lze použít pro efektivní a přesnou analýzu dat. R nabízí velké úložiště balíčků a uživatelé mají téměř každý typ nástroje k provedení jakéhokoli úkolu ve zpracování Big data. R lze bez problémů integrovat do aplikací Apache Hadoop a Apache Spark pro zpracování a analýzu Big data. Jeden problém s použitím R jako programovacího jazyka pro Big data je, že nemá příliš obecný účel. Kód napsaný v R není implementovatelný do výroby a obecně musí být přeložen do jiného programovacího jazyka, jako je Python nebo Java [49].

Java

Některé z tradičních velkých datových frameworků, jako je Apache Hadoop a všechny nástroje v jeho ekosystému, jsou založeny na Java a dnes se v mnoha podnicích stále používají. Jednou z hlavních nevýhod Javy je její složitost při kódování. Java nepodporuje iterativní vývoj na rozdíl od novějších jazyků, jako je Python, a to je oblast, na kterou se budou zaměřovat budoucí verze Java. Navzdory nedostatkům zůstává Java silným uchazečem, pokud jde o preferovaný jazyk pro programování Big data, kvůli jeho historii a neustálému spoléhání se na tradiční nástroje a frameworky Big data [49].

Scala

Scala je přechod objektově orientovaných a funkčních programovacích paradigmat, rychlý a robustní a oblíbený výběr jazyka pro mnoho profesionálů v oblasti Big data. Scala běží na JVM, což znamená, že kódy napsané v Scale lze snadno použít v ekosystému Big Data založeném na Java. Jedním z významných faktorů, které odlišují Scalu od Javy, je to, že Scala je ve srovnání s mnohem méně podrobnou. V Scale můžete psát stovky řádků matoucího kódu Java na méně než 15 řádků. Jedním z negativních aspektů Scaly je však její strmá křivka učení ve srovnání s jazyky jako Go a Python, což může odradit začátečníky, kteří ji chtějí používat [49].

6.6 Dílčí shrnutí

Technologie pro práci s Big data lze rozdělit do čtyř kategorií podle jejich účelu. Mezi tyto kategorie patří datové skladování, data mining, analýza dat a vizualizace data. Každá z kategorií má typické technologie, které jsou vhodné k jejich používání. Mezi technologie datového skladování patří např. Hadoop nebo NoSQL databáze. Pro data mining lze použít technologie Presto, Rapid Miner nebo Elasticsearch. Analýzu lze provádět např. pomocí technologií Apache Kafka, Splunk nebo KNIME. K vizualizaci dat lze využít technologii Tableau.

Který z programovacích jazyků vybrat záleží na případu, který chce organizace rozvinout. R je vhodný pro spoustu statistických výpočtů. Pro vyvíjení streamingových aplikací pro Big data je vhodnější jazyk Scala. Pro strojové učení k využití Big data a vytváření prediktivních modelů je vhodný jazyk Python. Pro vytváření řešení Big data pomocí tradičně dostupných nástrojů je jazyk Java. Jazyky lze kombinovat a tím dosáhnout účinnějšího a výkonnějšího řešení [49].

7 HLEDISKO: DATOVÉ MODELY

Datové modelování je proces, který definuje a analyzuje datové požadavky potřebné pro podporu procesů v informačních systémech organizací. Při datovém modelování spolupracuje vývojář s potencionálními uživateli IS. Cílem datového modelování je vytvoření datového modelu pro data, která mají být uložena v databázi. Model představuje konceptuální reprezentaci datových objektů, asociaci mezi objekty a pravidly [46].

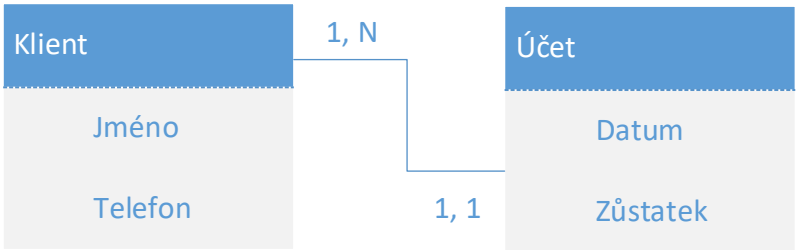
7.1 Datové modely v SQL databázích

Primárním cílem použití datového modelu u SQL databází je [46]:

- Zajištění přesného zobrazení všech datových objektů požadovaných v databázi.
- Datový model pomáhá navrhovat databázi na koncepční, fyzické a logické úrovni.
- Struktura datového modelu pomáhá definovat relační tabulky, primární a cizí klíče a uložené procedury.
- Poskytuje jasný obrázek o základních datech a může být použit vývojáři databází k vytvoření fyzické databáze.
- Je také užitečné identifikovat chybějící a nadbytečná data.
- Přestože počáteční vytvoření datového modelu je náročné na práci a čas, z dlouhodobého hlediska je jeho údržba levnější a rychlejší.

U SQL databází existují tři různé typy datových modelů – konceptuální, logický a fyzický.

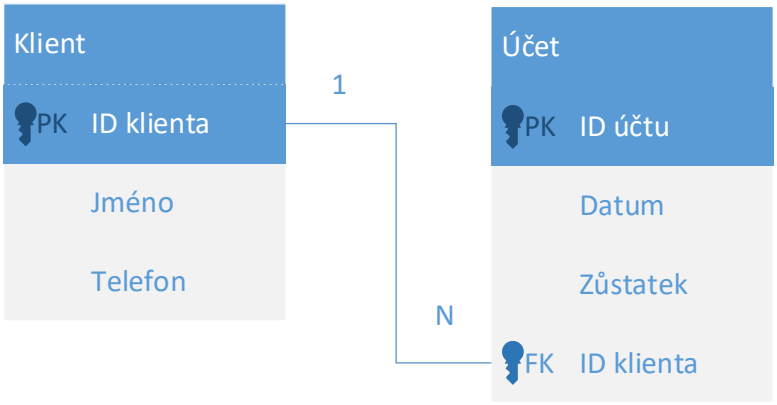
Konceptuální model (Obrázek 8) definuje, co systém obsahuje. Entity a vztahy jsou definovány podle potřeb podniku. Cílem je organizace, rozsah a definice obchodních konceptů a pravidel [46].



Obrázek 8: Konceptuální model

Zdroj: Vlastní zpracování

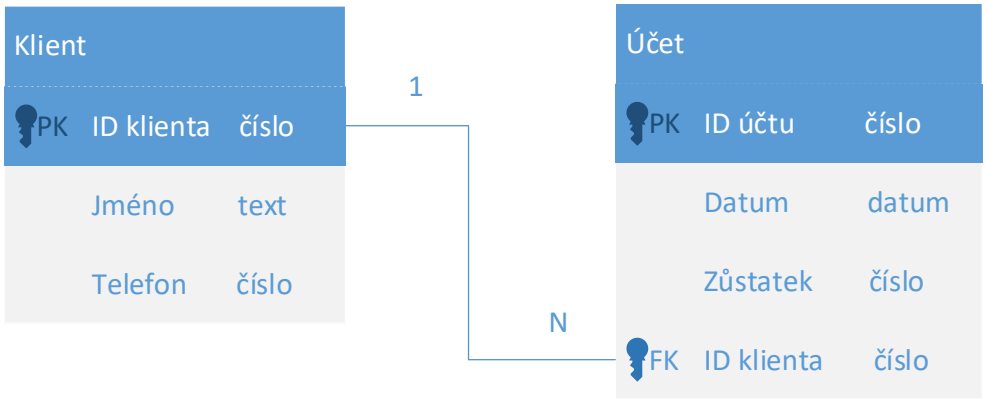
Logický model (Obrázek 9) podrobněji definuje, jak by měl být systém implementován bez ohledu na implementaci. Přidává další informace k prvkům konceptuálního modelu – zahrnuje entity a jejich vztahy, atributy, primární a cizí klíče. Na této úrovni nastává normalizace. Účelem je vypracování technické mapy pravidel a datových struktur. Bude tak vytvořen základ pro fyzický model [54].



Obrázek 9: Logický model

Zdroj: Vlastní zpracování

Fyzický model (Obrázek 10) popisuje, jak bude systém implementován do databáze. Tento typ pomáhá vizualizovat strukturu databáze a modelovat klíče sloupců databáze, omezení, indexy a další funkce [54].



Obrázek 10: Fyzický model

Zdroj: Vlastní zpracování

7.2 Datové modely v NoSQL databázích

Vhodné modely a prostředí úložiště nabízejí následující výhody pro Big data [48]:

- Výkon – Rychlé dotazování na požadovaná data a snížení propustnosti vstupů a výstupů.
- Náklady – Snížení nadbytečné redundance dat, znovupoužití výpočetních výsledků a snížení nákladů na skladování a výpočetní prostředky pro Big data systém.
- Účinnost – Zlepšení uživatelských zkušeností a zvýšení efektivity využívání dat.
- Kvalita – Zvýšení konzistentnosti statistik dat a snížení možnost výskytu chyb.

Proto je nepochybné, že Big data vyžadují vysoce kvalitní metody modelování dat pro organizaci a ukládání dat, což umožňuje dosáhnout optimální rovnováhy mezi výkonem, náklady, efektivitou a kvalitou [48].

Technologie ukládání Big data jsou rozděleny do kategorií podle jejich datového modelu. Mezi čtyři typy datového modelu patří klíč-hodnota (Obrázek 11), model sloupcově orientovaný (Obrázek 12), dokumentově-orientovaný model (Obrázek 13) a grafový model (Obrázek 14) [29].

7.2.1 Databáze typu klíč-hodnota

Databázové systémy typu klíč-hodnota mohou ukládat prakticky jakékoli objekty na základě unikátních klíčů. Operace nad úložišti typu klíč-hodnota jsou pak poměrně jednoduché a primárně neposkytují žádný způsob, jak s daty manipulovat nebo je vyhledávat na základě uloženého obsahu – jen podle určeného klíče. Tato úložiště bývají extrémně rychlá a dají se velice efektivně distribuovat. Datový model u těchto systémů je absolutně svobodný. Redis a DynanoDB jsou populární databáze typu klíč-hodnota [2], [56].

Namespace *User*

Klíč:	*userID*
Hodnota:	*userProfile* *sessionData* *shoppingCart* • *item 1* • *item 2*

Obrázek 11: Databáze typu klíč-hodnota

Zdroj: [2]

7.2.2 Sloupcové databáze

Datový model sloupcových systémů si lze představit jako tabulku, u které je možné do každého řádku volně přidávat sloupce bez nutnosti přidat je i do řádků ostatních. Toto volné nakládání se sloupci nesnižuje výkonnost sloupcových databází, které bývají masivně distribuované. Základním pojmem popisovaného datového modelu je řádek. K identifikaci řádku slouží klíč řádku. Klíč řádku může obsahovat velké množství sloupců. Každý sloupec v daném řádku má název, hodnotu a časové razítko. Jednotlivé sloupce jsou sdruženy do rodin sloupců. Při návrhu schématu databáze se definují pouze rodiny sloupců. Jednotlivé řádky mohou v rámci dané rodiny sloupců volně vytvářet libovolné množství sloupců s libovolnými názvy. Příklady oblíbených sloupcových databází jsou Cassandra a HBase [2], [57].

user_id (klíč řádku)	název sloupce	název sloupce	…
	hodnota sloupce	hodnota sloupce	…
1	login	first_name	…
	honza	Jan	…
4	login	age	
	david	35	…
5	first_name	last_name	…
	Jana	Novotná	…
…			…

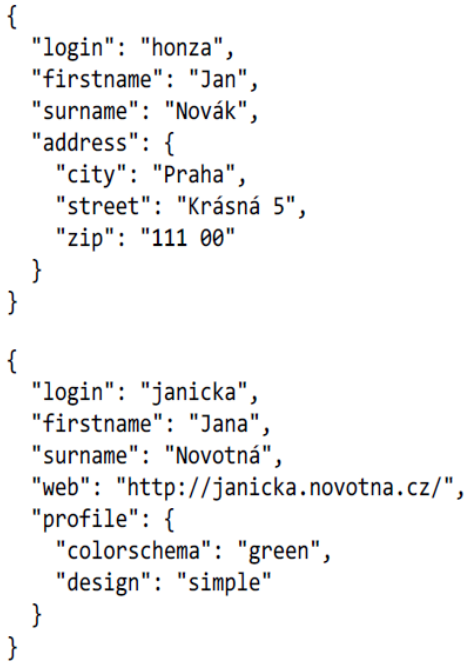
Obrázek 12: Sloupcová databáze

Zdroj: [2]

7.2.3 Dokumentové databáze

Dokumentové databáze ukládají a spravují různé druhy strukturovaných dokumentů (datových struktur), o kterých se předpokládá, že mají samopopisný charakter. Obsahuje tedy kromě dat i metadata. Příkladem je formát JSON nebo XML. Tyto formáty si lze představit jako stromové datové struktury, které obsahují asociativní pole (název a hodnota), seznamy a základní datové typy. Dokumentové databáze umožňují přistupovat k dokumentům a vyhledávat v nich podle jejich obsahu. Dokumenty slouží nejen pro ukládání dat, ale také pro komunikaci s klienty a aplikacemi. Výhodou použití datového modelu je možnost manipulace se všemi daty v rámci jedné operace (čtení, zápis a aktualizace). Tento model je vhodný pouze při modelování vztahu 1:1 nebo 1:N. Nevýhodou je, že pokud se do nadřazeného dokumentu vloží další vnořené

záznamy, velikost dokumentu může výrazně vzrůst a tím ovlivňovat rychlost čtení, zápisu a přenosu dat. Příkladem dokumentové databáze je MongoDB [2], [55].

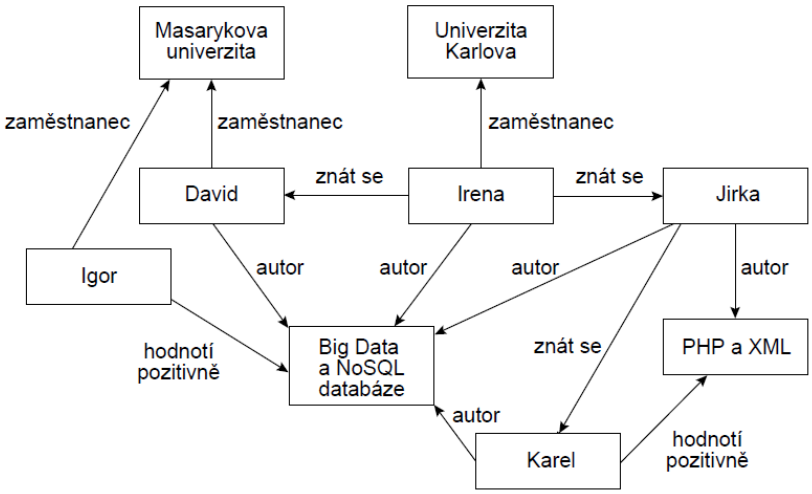


Obrázek 13: Dokumentová databáze

Zdroj: [2]

7.2.4 Grafové databáze

Grafové databáze se výrazněji liší od předchozích. Systém je určen pro data, která je vhodné modelovat a dotazovat jako grafy (objekty a vztahy mezi nimi). Data jsou vnitřně strukturována tak, aby bylo možné efektivně vyhodnocovat různé grafové úlohy, jako je hledání sousedů nebo cest v grafu. Uzel v grafu odpovídá jednomu objektu, který může mít množinu atributů jako např. jméno nebo věk. Hrany, typicky orientované, představují vztahy mezi objekty. Hrany také mohou mít atributy, jako je typ, doba platnosti vztahu, podmínky platnosti vztahu apod. Příkladem grafových databází jsou Neo4j nebo JanusGraph [2].



Obrázek 14: Grafová databáze

Zdroj: [2]

7.3 Dílčí shrnutí

Datové modelování je proces používaný k definování datových požadavků. Datové modelování zahrnuje spolupráci vývojáře s potenciálními uživateli dat. Cílem datového modelu je zajistit, aby všechny datové objekty vyžadované databází byly zcela a přesně reprezentovány.

U SQL databází se datové modelování dělí na tři úrovně – konceptuální, logickou a fyzickou. Konceptuální úroveň popisuje základní požadavky systémy. Logická úroveň popisuje pravidla bez ohledu na relační databázi. Fyzická úroveň se týká skutečné implementace databáze.

Při použití NoSQL databází se nedodržují postupy jako u SQL a zvažují se jiná schémata. Typy NoSQL databází se dělí podle jednotlivých datových modelů – databáze typu klíč-hodnota, dokumentové databáze, sloupcové databáze a grafové databáze.

ZÁVĚR

Bakalářská práce byla zaměřena na strukturovaná a nestrukturovaná data v organizaci, přičemž zaměření bylo zejména na Big data, protože jejich význam pro firmy stále vzrůstá.

Nejdříve byly charakterizovány základní vlastnosti Big data, označované jako 3V, kterými jsou objem, rychlost a různorodost dat. Postupné rozšiřování charakteristik vedlo ke skupinám vlastností označovaných také 5V nebo 10V. Existuje více uplatnění, jak mohou podniky z Big data těžit. Mezi hlavní uplatnění patří zisk, úspora nákladů nebo větší automatizace podniku. Na druhou stranu díky svým vlastnostem hrozí u Big data i některá rizika, která se mohou týkat nákladů nebo ztráty kontroly nad organizovaností dat.

Pro další zpracování bakalářské práce byla vybrána čtyři hlediska – životní cyklus dat, architektura Big data, technologie pro práci s Big data, a datové modely u SQL a NoSQL databází.

Každé hledisko bylo popsáno v samostatné kapitole včetně konkrétních příkladů a schémat. Životní cyklus dat byl charakterizován obecně bez rozlišení na data strukturovaná a nestrukturovaná. Kapitola Architektura a kapitola Technologie jsou zaměřeny výhradně na Big data. Poslední hledisko, charakteristika datových modelů a modelování, se zaměřilo opět na obě skupiny dat, tzn. na strukturovaná i nestrukturovaná.

Práce měla přinést přehled o základních aspektech této problematiky. Byla proto určena čtyři hlediska a ta byla postupně charakterizována na základě dostupných zdrojů, které byly převážně zahraniční. Práce představuje shrnutí základních témat v dané problematice a může být využita jako výukový materiál pro studenty, což byl také její účel. Lze proto konstatovat, že cíl práce byl naplněn.

POUŽITÁ LITERATURA

[1] GRONWALD Klaus-Dieter. *Integrated Business Information Systems: A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data*. Berlin: Springer Berlin Heidelberg, 2017. ISBN 978-3-662-53290-4.

[2] HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. *Big Data a NoSQL databáze*. Praha: Grada Publishing , 2015. Profesionál. ISBN 978-80-247-5466-6.

[3] LEE, James, Tao WEI a Suresh Kumar MUKHIYA. *Hands-On Big Data Modeling*. Birmingham (United Kingdom): Packt Publishing, 2018. ISBN 978-1788620901.

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

[5] BEYER, Mark a Douglas LANEY. The Importance of 'Big Data': A Definition. *Gartner* [online]. Stamford (Connecticut): Gartner, 2012 [cit. 2020-04-17]. Dostupné z: https://www.gartner.com/en/documents/2057415/the-importance-of-big-data-a-definition

[6] MATOUŠEK, Jan. BIG DATA V DŮCHODU - JSOU DALŠÍ A DALŠÍ DATA OPRAVDU SPÁSOU MARKETINGU A OBCHODU? *Data Mind* [online]. Praha: Data Mind, 2014 [cit. 2020-04-15]. Dostupné z: https://www.datamind.cz/cz/blog/big-data-v-duchodu-jsou-dalsi-a-dalsi-data-opravdu-spasou-marketingu-a-obchodu

[7] YAMU, Claudia, Alenka POPLIN, Oswald DEVISCH a Gert DE ROO. *The Virtual and the Real in Planning and Urban Design: Perspectives, Practices and Application*. New York: Routledge, 2017. ISBN 978-1-138-28348-0.

[8] FIRICAN, George. The 10 Vs of Big Data. *TDWI* [online]. Renton (Washington): TDWI, 2017 [cit. 2020-07-20]. Dostupné z: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

[9] What is the DIKW Pyramid? *Ontotext* [online]. Sofia (Bulgaria): Ontotext, ©2020 [cit. 2020-06-26]. Dostupné z: https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/

[10] The Four V's of Big Data. *Big Data Framework* [online]. Big Data Framework©, 2019 [cit. 2020-04-18]. Dostupné z: https://www.bigdataframework.org/four-vs-of-big-data/

[11] PICKELL, Devin. Structured vs Unstructured Data – What's the Difference? *Learning Hub* [online]. Chicago (Illinois): G2.com, 2018 [cit. 2020-04-18]. Dostupné z: https://learn.g2.com/structured-vs-unstructured-data

[12] TAYLOR, Christine. Structured vs. Unstructured Data. *Datamation* [online]. Nashville (Tennessee): TechnologyAdvice, 2018 [cit. 2020-04-18]. Dostupné z: https://www.datamation.com/big-data/structured-vs-unstructured-data.html

[13] IHRITIK. What is Semi-structured data? *GeeksforGeeks* [online]. Noida (Uttar Pradesh): GeeksforGeeks [cit. 2020-05-18]. Dostupné z: https://www.geeksforgeeks.org/what-is-semi-structured-data/

[14] Metadata. *Sociální síť pro business - ManagementMania.com* [online]. Wilmington (Delaware): ManagementMania.com, 2016 [cit. 2020-04-19]. Dostupné z: https://managementmania.com/cs/metadata

[15] Kvalita dat (Data quality). *ManagementMania.com* [online]. Wilmington (Delaware): ManagementMania.com, 2018 [cit. 26.06.2020]. Dostupné z: https://managementmania.com/cs/kvalita-dat-data-quality

[16] LEVY SARFIN, Rachel. 5 Characteristics of Data Quality. Syncsort Blog [online]. Pearl River (New York): Syncsort, 2019 [cit. 2020-06-26]. Dostupné z: https://blog.syncsort.com/2019/07/data-quality/5-characteristics-of-data-quality/

[17] IVANOV, Todor, Nikos KORFIATIS a Roberto ZICARI. On the inequality of the 3V's of Big Data Architectural Paradigms: A case for heterogeneity. *ResearchGate* [online]. Frankfurt Am Mein : 2013. https://www.researchgate.net/figure/BM-Big-Data-characteristics-3V-Adopted-from-Zikopoulos-and-Eaton-2011_fig1_258247680

[18] SHIFF, Laura. Real Time vs Batch Processing vs Stream Processing. *BMC Blogs* [online]. Houston (Texas): BMC Software, 2020 [cit. 2020-06-26]. Dostupné z: https://www.bmc.com/blogs/batch-processing-stream-processing-real-time/

[19] VERACITY: THE MOST IMPORTANT "V" OF BIG DATA. *GutCheck* [online]. GutCheck: Denver (Colorado), 2019 [cit. 2020-06-26]. Dostupné z: https://www.gutcheckit.com/blog/veracity-big-data-v/

[20] BAUMANN, Peter a Morris RIEDEL. Big Data - Definition, Importance, Examples & Tools. *RDA* [online]. RDA, 2019 [cit. 2020-06-26]. Dostupné z: https://www.rd-

alliance.org/group/big-data-ig-data-development-ig/wiki/big-data-definition-importance-examples-tools

[21] Big Data Use Cases. *Oracle* [online]. Redwood City (California): Oracle, ©2020 [cit. 2020-06-26]. Dostupné z: https://www.oracle.com/big-data/guide/big-data-use-cases.html

[22] HRTÚSOVÁ, Tereza a Radek NOVÁK. BIG DATA V ČR: POJEM VS. REALITA. *Česká spořitelna* [online]. Praha: Česká spořitelna, 2018 [cit. 2020-06-26]. Dostupné z: https://www.csas.cz/content/dam/cz/csas/www_csas_cz/Dokumenty-korporat/Dokumenty/Analytici/Big_Data_v_CR.pdf

[23] Why is big data analytics important? *SAS* [online]. Cary (North Carolina): SAS Institute, ©2020 [cit. 2020-06-26]. Dostupné z: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

[24] KHVOYNITSKAYA, Sandra. The future of big data: 5 predictions from experts for 2020-2025. *Itransition* [online]. Austin (Texas): Itransition, 2020 [cit. 2020-06-26]. Dostupné z: https://www.itransition.com/blog/the-future-of-big-data

[25] MARR, Bernard. The Biggest Risks of Big Data. *LinkedIn* [online]. Sunnyvale (California): LinkedIn, 2015 [cit. 2020-06-26]. Dostupné z: https://www.linkedin.com/pulse/5-biggest-risks-big-data-bernard-marr

[26] ESTUATE. Are you fighting the 5 biggest risks of big data? *Estuate* [online]. Milpitas (California): Estuate, 2017 [cit. 2020-06-27]. Dostupné z: https://www.estuate.com/company/blog/content/are-you-fighting-5-biggest-risks-big-data

[27] Big Data Framework: *Enterprise Big Data Professional*. Bonn: Big Data Framework, 2018. ISBN 978-90-828958-0-3.

[28] Big data architecture style. *Microsoft* [online]. Redmond (Washington): Microsoft, 2019 [cit. 2020-06-26]. Dostupné z: https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data

[29] SIDDIQA, Aisha, Ahmad KARIM a Abdullah GANI. Big data storage technologies: a survey. *Frontiers of Information Technology & Electronic Engineering* [online]. 2017, **18**(8), 1040-1070 [cit. 2020-06-27]. DOI: 10.1631/FITEE.1500441. ISSN 2095-9184. Dostupné z: http://link.springer.com/10.1631/FITEE.1500441

[30] ROSE, Scarlett. Why is Python Programming a perfect fit for Big Data? *Medium – Get smarter about what matters to you.* [online]. San Francisco (California): Medium, 2019 [cit. 2020-07-13]. Dostupné z: https://towardsdatascience.com/why-is-python-programming-a-perfect-fit-for-big-data-5ac54ee8f95e

[31] ERL Thomas, Wajid KHATTAK, and Paul BUHLER. Big Data Fundamentals: Concepts, Drivers & Techniques. Cham: Springer, 2016. ISBN 978-0-13-429107-9.

[32] Sources of big data: Where does it come from? *CloudMoyo* [online]. Bellevue (Washington) CloudMoyo, ©2020 [cit. 2020-06-27]. Dostupné z: https://www.cloudmoyo.com/blog/data-architecture/what-is-big-data-and-where-it-comes-from/

[33] PRESS, Gill. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. *Forbes* [online]. Jersey City: Forbes Media, 2016 [cit. 2020-06-27]. Dostupné z: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#64c065696f63

[34] WHAT IS A REFERENCE ARCHITECTURE? *Hewlett Packard Enterprise* [online]. San Jose (California): Hewlett Packard Enterprise Development LP, ©2020 [cit. 2020-06-27]. Dostupné z: https://www.hpe.com/us/en/what-is/reference-architecture

[35] VEERESH, Kumar. Challenges of Big Data Architecture. *Medium* [online]. San Francisco (California): Medium, 2020 [cit. 2020-06-27]. Dostupné z: https://medium.com/@veeeresh423kumar/challenges-of-big-data-architecture-ef931f968c95

[36] TONIDANDEL, Scott, Eden KING a Jose M. CORTINA. *Big data at work: the data science revolution and organizational psychology*. New York: Routledge, Taylor & Francis Group, 2016. ISBN 978-1-84872-581-2.

[37] KIRAN, Ravi. Top Big Data Technologies that you Need to know. *Edureka* [online]. Bengaluru (Karnataka): Brain4ce Education Solutions, 2019 [cit. 2020-06-27]. Dostupné z: https://www.edureka.co/blog/top-big-data-technologies/

[38] Big Data Storage. *Techopedia* [online]. Edmonton (Alberta): Techopedia, ©2020 [cit. 2020-06-27]. Dostupné z: https://www.techopedia.com/definition/29473/big-data-storage

[39] Big Data Mining. *Techopedia* [online]. Edmonton (Alberta): Techopedia, ©2020 [cit. 2020-06-27]. Dostupné z: https://www.techopedia.com/definition/30215/big-data-mining

[40] Big Data Analytics. *Techopedia* [online]. Edmonton (Alberta): Techopedia, ©2020 [cit. 2020-06-27]. Dostupné z: https://www.techopedia.com/definition/28659/big-data-analytics

[41] REQTEST. Requirements Analysis – Understanding the Process & Techniques. *ReQtest* [online]. Stockholm: ReQtest, 2018 [cit. 2020-07-13]. Dostupné z: https://reqtest.com/requirements-blog/requirements-analysis/

[42] *Apache Hadoop* [online]. Houston (Texas): The Apache Software Foundation, ©2006-2020 [cit. 2020-06-27]. Dostupné z: https://hadoop.apache.org/

[43] Data visualization beginner's guide: a definition, examples, and learning resources. *Tableau* [online]. Seattle (Washington): Tableau Software, ©2003-2020 [cit. 2020-06-27]. Dostupné z: https://www.tableau.com/learn/articles/data-visualization

[44] DENNING, Peter J. Saving All the Bits. *American Scientist* [online]. **78**, 402-405. Dostupné z: http://denninginstitute.com/pjd/PUBS/AmSci-1990-5-savingbits.pdf

[45] TYAGI, Neelam. Top 10 Big Data Technologies in 2020. *Analytics Steps* [online]. Noida (Uttar Pradesh): Analytics Steps, 2020 [cit. 2020-06-27]. Dostupné z: https://www.analyticssteps.com/blogs/top-10-big-data-technologies-2020

[46] What is Data Modelling? Conceptual, Logical, & Physical Data Models. *Guru99* [online]. Ahemdabad (Gujrat): Guru99, ©2020 [cit. 2020-06-27]. Dostupné z: https://www.guru99.com/data-modelling-conceptual-logical.html

[47] ANALÝZA POŽADAVKU. *Internetová agentura SOVA NET* [online]. Brno: SOVA NET [cit. 2020-07-13]. Dostupné z: https://www.sovanet.cz/analyza-pozadavku/

[48] ZHANG, Leona. A Comparison of Data Modeling Methods for Big Data. *DZone* [online]. Morrisville (North Carolina): DZone, 2018 [cit. 2020-06-28]. Dostupné z: https://dzone.com/articles/a-comparison-of-data-modeling-methods-for-big-data

[49] VARANGAONKAR, Amey. Top 5 programming languages for crunching Big Data effectively. *Packt Hub* [online]. Birmingham (United Kingdom): Packt Publishing, 2018 [cit. 2020-07-13]. Dostupné z: https://hub.packtpub.com/top-5-programming-languages-big-data/

[50] COX, Michael a David ELLSWORTH. Application-Controlled Demand Paging for Out-of-Core Visualization. *Proceedings of the 8th conference on Visualization 97*. California: 1997. Dostupné z: https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf

[51] LANEY, Douglas. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies*. Stamford (Connecticut): META Group, 2001. Dostupné z: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[52] NIST Big Data Interoperability Framework: Volume 8, Reference Architecture Interfaces. *NIST Special Publication 1500-9* [online]. 2018, **8**, 5-12 [cit. 2020-06-28]. DOI: 10.6028/NIST.SP.1500-9. Dostupné z: https://doi.org/10.6028/NIST.SP.1500-9

[53] STROHBACH, Martin, Jörg DAUBERT, Herman RAVKIN a Mario LISCHKA. Big Data Storage. CAVANILLAS, José María, Edward CURRY a Wolfgang WAHLSTER, ed. *New Horizons for a Data-Driven Economy* [online]. Cham: Springer International Publishing, 2016, 2016, s. 119-141 [cit. 2020-06-28]. DOI: 10.1007/978-3-319-21569-3_7. ISBN 978-3-319-21568-6. Dostupné z: http://link.springer.com/10.1007/978-3-319-21569-3_7

[54] Conceptual, Logical and Physical Data Model. *Ideal Modeling & Diagramming Tool for Agile Team Collaboration* [online]. Hong Kong: Visual Paradigm [cit. 2020-07-23]. Dostupné z: https://www.visual-paradigm.com/support/documents/vpuserguide/3563/3564/85378_conceptual,l.html

[55] Comparing Document Databases and Relational Databases. *Couchbase* [online]. Santa Clara (California): Couchbase, ©2020 [cit. 2020-07-24]. Dostupné z: https://developer.couchbase.com/comparing-document-vs-relational/

[56] What Is a Key-Value Database? *Amazon Web Services (AWS)* [online]. Seattle (Washington): AWS [cit. 2020-07-24]. Dostupné z: https://aws.amazon.com/nosql/key-value/

[57] AKASH, Kumar. Difference between Row oriented and Column oriented data stores in DBMS. *GeeksforGeeks* [online]. Noida (Uttar Pradesh): GeeksforGeeks [cit. 2020-07-24]. Dostupné z: https://www.geeksforgeeks.org/difference-between-row-oriented-and-column-oriented-data-stores-in-dbms/