UPRAVLJANJE PODACIMA

1. ORGANIZACIJA I SKLADIŠTENJE PODATAKA U POSLOVNIM SISTEMIMA

2. ANALITIKA I PRETRAŽIVANJE PODATAKA

Prof. dr Jasna Soldić-Aleksić

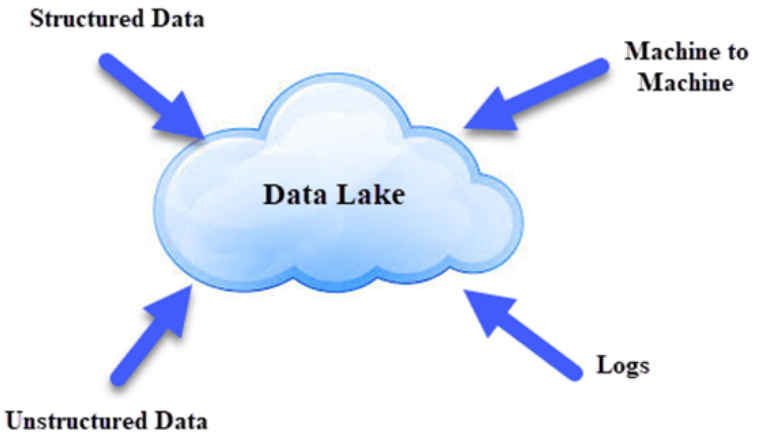
1.4. Jezero podataka – *Data lake*

Za potrebe skladištenja velike količine podataka bile su potrebne nove tehnologije. Naime, sa povećanjem količine i raznovrsnosti podataka u poslovnim sistemima, sve je prisutniji, već od ranije poznat, fenomen stvaranja **silosa podataka** (*data silos*) – postojanje podataka o istim entitetima na više mesta, fragmentisanost podataka, čime se umanjuje efikasnost njihovog korišćenja. Kao što smo videli sličan problem se javljao i kod prelaska sa tehnologije relacionih baza podataka na kreiranje skladišta podataka. Jedan od načina da se reše ovi problemi jeste kreiranje jezera podataka – *data lake.*

Jezero podata (eng. Data lake) predstavlja **centralizovani repozitorijum podataka**, gde je moguće skladištiti kako **struktuirane, tako i slabo struktuirane i nestruktuirane podatke**.

Za razliku od baza podataka i skladišta podataka u kojima se skladište struktuirani podaci, tehnologija jezera podataka je nastala kao odgovor na sve veću količinu posebno nestruktuiranih podataka.

Različiti tipovi podataka se skladište u jezeru podataka u sirovom, rudimentarnom obliku, dakle bez prethodne obrade. Prilikom skladištenja podataka u jezeru podataka kreiraju se *identifikatori* i *metadata tagovi*, koji služe za brže pronalaženje i učitavanje podataka. Posmatrano sa strane hardvera, **jezera podataka su obično konfigurisana kao klasteri relativno jeftinih i skalabilnih servera.** Ovi klasteri mogu biti konfigurisani bilo *on-premises* (na računarskoj opremu u poslovnom sistemu) ili u oblaku (*cloud based*), mada je prisutna tendencija migracije podataka na računare u oblaku.



Slika 20. Jezero podataka[1]

[1] https://www.guru99.com/data-lake-architecture.html

Jezero podataka pruža sledeće[2]:

- Centralizaciju i konsolidaciju podataka na jednom mestu
- Integraciju podataka u različitim formatima: struktuirani podaci, video, slike, binarni fajlovi, *batch* i *streaming* podaci;
- Podaci se u jezeru podataka mogu dugo čuvati relativno jeftino, sve do buduće upotrebe, kada se učitavaju za potrebe primene algoritama mašinskog učenja i napredne analitike;
- Mogućnost demokratizacije podataka i analitike, prevashodno primenom različitih alata za *self-service* pristup i analitiku podataka.

U sledećoj tabeli date su opšte karakteristike skladišta podataka i jezera podataka, čime se sumiraju sličnosti i razlike ovih tehnologija.

Tabela 4. Opšte karakteristike skladišta podataka i jezera podataka

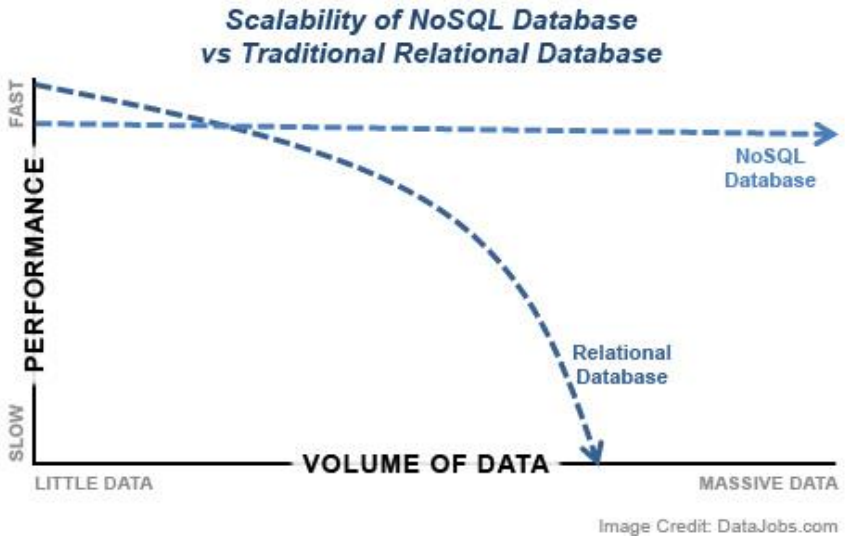
Karakteristike	***Data warehouse***	***Data lake***
Tip podataka	**Isključivo struktuirani podaci**	Svi tipovi podataka: **sirovi podaci**
Skalabilnost	**Ograničena i skupa**	Relativno **jeftina** za sve tipove podataka
Arhitektura	***Schema-on-write*** (šema za učitavanje podataka je poznata unapred)	***Schema-on-read*** (šema nije unapred poznata, već se podaci učitavaju u originalnom formatu)
Troškovi	Veoma **visoki**	**Niži**
Pristup korisnika	**Relativno jednostavan**, zahvaljujući dobro definisanoj i dokumentovanoj strukturi	**Komplikovan** i zahteva tehničko znanje o prirodi različitih podataka i njihovih veza
Prednosti	Korisnički interfejs je relativno poznat za korisnike tradicionalnih baza podataka	Fleksibilnost, niski troškovi, skalabilnost, skladištenje ogromne količine podataka za potrebe mašinskog učenja, napredne analitike, *data science* algoritama
Nedostaci	Skup sistem, vezanost za proizvođača, rigidnost pristupa i održavanja, ne može da skladišti nestruktuirane (sirove) podatke	Pristup i pretraživanje ogromne količine podataka je nemoguće bez odgovarajućih **alata za katalogizaciju podataka**
Mogućnost analitike	**Tradicionalne analitičke tehnike**	**Napredna analitika iz domena *data-science*** (npr. na nestruktuiranim podacima sentiment analiza, otkrivanje prevara, pranje novca i sl., text analitika).

[2] https://www.talend.com/resources/data-lake-vs-data-warehouse/

Tehnologija NoSQL ("Not Only SQL")

Tradicionalne tehniologije relacionih baza podataka i skladišta podataka nisu mogle da odgovore potrebama skladištenja i obrade ogromne količine podataka, koja se označava kao *Big data.*

Jedna od značajnih tehnologija koja može da odgovori izazovima velikog obima podataka jeste tehnologija NoSQL ("Not Only SQL") baza podataka. Naime, reč je o takvoj infrastrukturi baze podataka koja je dobro adaptirana za potrebe velikog obima podataka. Za razliku od klasičnih relacionih baza podataka koje su visoko struktuirane, NoSQL je baza podataka koja je nestruktuirana u svojoj osnovi. NoSQL se zasniva na **konceptu distribuiranih baza podataka**, gde se nestruktuirani podaci skladište i čuvaju na mnogobrojnim procesorskim jedinicama (*processing nodes*) na većem broju servera. Ovakva **distribuirana arhitektura** omogućava horizontalnu skalabilnost, što znači da kako se uvećava količina podataka dodaju se nove hardverske jedinice, čime se ne gubi na performansama baze. Za razliku od njih performanse klasičnih relacionih baza podataka drastično opadaju sa povećanjem obima podataka. Velike kompanije, kao što su Google, Amazon koriste NoSQL distribuiranu strukturu baze podatka.



Slika 21. Performanse tradicionalne relacione baze podataka i NoSQL baze[3]

Najpoznatija računarska okruženja za *Big data* analitiku, odnosno *Data lake* arhitekture su:

Hadoop kao open-source okruženje**, Amazon WS,** koji pruža niz proizvoda za Big data analitiku **i Microsoft Azure.**

[3] https://datajobs.com/what-is-hadoop-and-nosql

Hadoop predstavlja open source softverski ekosistem za Big data analitiku, gde je omogućeno da se velika količina podataka procesira simultano (paralelno procesiranje) na većem broju računara koji su povezani u klastere. Osnovne komponente Hadoop arhitekture su:

- **Hadoop Distributed File System (HDFS)** predstavlja prvenstveno sistem za čuvanje podataka (***storage system***) koji omogućava da se podaci čuvaju na mnogim različitim uređajima kao da se radi o jednom fajlu; radi se o **distribuiranom sistemu fajlova**.
- **Hadoop MapReduce** predstavlja centralnu komponentu Hadoop sistema, koja je zadužena **za procesiranje podataka**, odnosno za podelu složenog računarskog zadatka na manje računarske celine koje se mogu izvršavati paralelno na računarskim klasterima;
- **YARN (Yet Another Resource Negotiator)** je komponenta Hadoop ekosistema koja ima funkciju da obezbedi upravljanje računarskim resursima. Drugim rečima, ova komponenta se označava kao **operativni sistem Hadoop**-a, jer upravlja i nadzire kompletan radni tok i procesiranje na različitim mašinama.

Pored ove tri osnovne komponente Hadoop ekosistem čine i sledeće komponente: **Apache Hive** – open-source sistem za postavljanje upita, sumiranje podataka i analizu velike količine podataka. Ovaj sistem koristi jezik upita (HiveQL – HQL) koji je sličan SQL jeziku i prevodi SQL upite u MapReduce programske zadatke; **Apache Pig** je takođe Hadoop komponenta za analiziranje i pretraživanje ogromnih setova podataka koji su skladišteni u HDFS. Takođe je sličan SQL jeziku; **Apache HBase** komponenta se nalazi na vrhu HDFS sistema i zadužena je da obezbedi pristup u realnom vremenu čitanju i pisanju podataka u HDFS. Reč je o distribuiranoj NoSQL bazi podataka; **HCatalog** je bitna Hadoop komponenta za upravljanje skladištenjem podataka; **Apache Drill** je prva distribuirana SQL mašina za pretraživanje; **Apache Mahout** je komponenta koja sadrži mnoge *data science* algoritme za grupisanje, klasifikacije, kolaborativno filtriranje, automatsko pronalaženje značajnih obrazaca u velikoj količini podataka i sl; **Apache Sqoop** je Hadoop komponenta koja je zadužena za importovanje podataka iz različitih eksternih izvora, ali i za eksportovanje podatak iz Hadoop-a na eksterne izvore.

Na osnovu prethodnog može se sumirati da je pojava Hadoop –a otvorila prostor za sledeće[4]:

- Mogućnost da se skladište i obrađuju *velike količine raznovrsnih podatka* relativno brzo;
- *Računarska snaga* – proporcionalna je broju čvorova, tj. ukoliko ima više računarskih čvorova u Hadoop arhitekuri veća procesorska snaga je na raspolaganju;
- *Tolerantnost na greške* – ukoliko dođe do problema na nekim računarima, poslovi se automatski prenose na druge računare - čvorove i kopije svih podataka se automatski čuvaju;
- *Fleksibilnost* – za razliku od tradicionalnih relacionih baza podataka, podaci ne moraju da prođu postupak pretprocesiranja pre nego se skladište već se čuvaju u sirovom obliku;
- *Niski troškovi* – radi se o open-source računarskom okruženju;
- *Skalabilnost* – sistem se može jednostavno proširiti dodavanjem novih čvorova, uz relativno jednostavno administriranje.

[4] https://www.sas.com/en_lu/insights/big-data/hadoop.html)

2. ANALITIKA I PRETRAŽIVANJE PODATAKA

U prethodnom delu smo razmatrali skladištenje, način organizacije i čuvanje podataka u poslovnom sistemu. Na koji način se obrađuju i analiziraju ovi podaci? Podsetimo se da osnovni cilj obrade i analize podatka kreiranje nove poslovne vrednosti. Takođe, već smo naveli da je jedan od načina kreiranja nove poslovne vrednosti otkrivanje informacija i znanja iz podataka (engl. *Information and Knowledge Discovery*). Proces izdvajanja korisnog znanja iz velike količine podataka u literaturi je poznat još kao otkrivanje znanja u bazama podataka (engl. *Knowledge Discovery in Database – KDD*). Cilj *KDD* je da identifikuje valjane, nove, korisne, razumljive obrasce (šeme ili strukture) u podacima. **KDD podržavaju tri tehnologije:**

- **Prikupljanje i skladištenje velike količine podataka;**
- **Efikasni višeprocesorski računari;**
- **Algoritmi za otkrivanje modela podataka – *Data Mining* algoritmi.**

Evolucija koncepta i alata KDD-a mogla bi se predstaviti kroz sledeće etape:

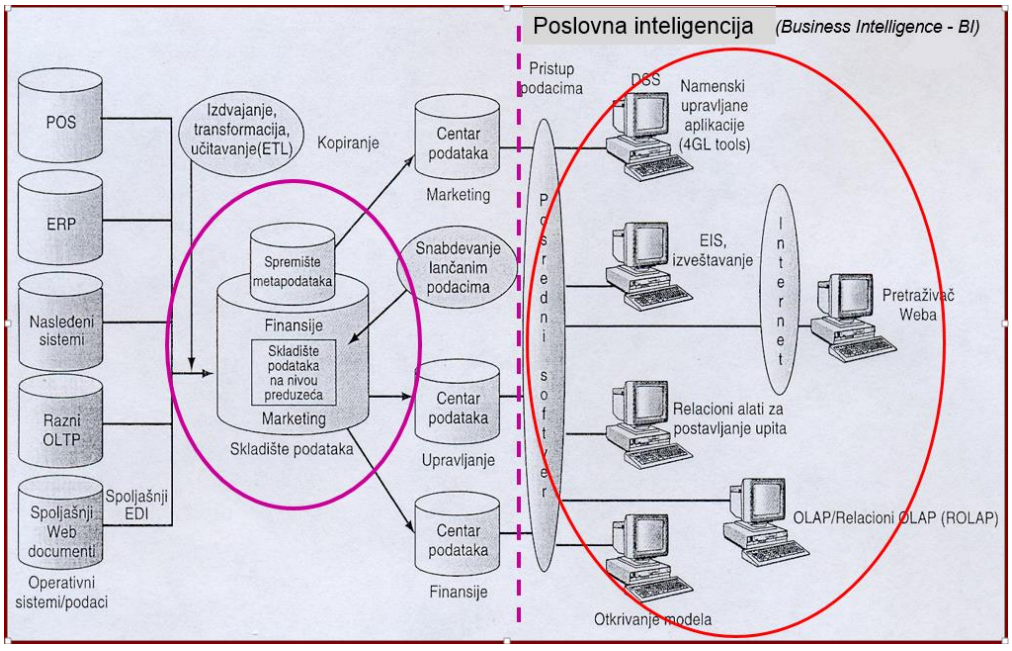
1. **Prikupljanje podataka** (1960-te): primena računara, nosioci podataka - trake, diskovi;
2. **Pristup podacima** (1980-te): korišćenje relacionih baza podataka, jezika upita - SQL;
3. **Skladištenje podataka i podrška pri odlučivanju** (početak 1990 ...): OLAP, multidimenzionalne baze podataka;
4. **Inteligentno otkrivanje modela podataka** (kraj 1990 ...): višeprocesorski računari, velike baze podataka, napredni algoritmi pretrage podataka – *data mining*;
5. **Koncept poslovne inteligencije** (***Business intelligence - BI***) (2000....) - složeni sistemi skladištenja podataka, postavljanja upita, primena DSS, EIS sistema, inteligentnih sistema, kompletna integracija različitih alata ...;
6. **Koncept Poslovne analitike** (***Business analytics - BA***) u periodu 2006 .... pojava fenomena *Big data*, pored tradicionalnih tenika analize podataka, široka primena algoritama mašinskog učenja, pojavljivanje nove disciopline – nauke o podacima (*data science*), razvoj svih oblasti veštačke inteligencija i njihova primena u poslovanju.

2.1.Poslovna inteligencija *(engl. Business Intelligence - BI)*

Koncept Poslovne inteligencije (engl. *Business Intelligence*) ima dugu istoriju. Njeni koreni datiraju od kraja 80-tih godina prošlog veka. Tradicionalno ovaj koncept se vezuje za korišćenje različitih alata za postavljanje upita i pripremanje izveštaja iz baza podataka. Međutim, ključna tehnologija za razvoj ovog koncepta jeste **tehnologija skladišta podataka – DW tehnologija.** Poslovna inteligencija obuhvata niz alata i tehnika za prikupljanje, skladištenje i transformisanje sirovih podataka u korisne i svrsishodne informacije za potrebe poslovne analize u formi izveštaja, ogovora na upite, dijagrama, grafikona, kontrolnih tabli (engl. *dashboards*) ili interaktivne vizuelizacije podataka. Reč je o procesu prikupljanja pravih podataka u pravo vreme u pogodnom formatu, a zatim sprovođenje njihove analize, koja ima za cilj da obezbedi pozitivan uticaj na poslovnu strategiju, taktiku i poslovne aktivnosti.

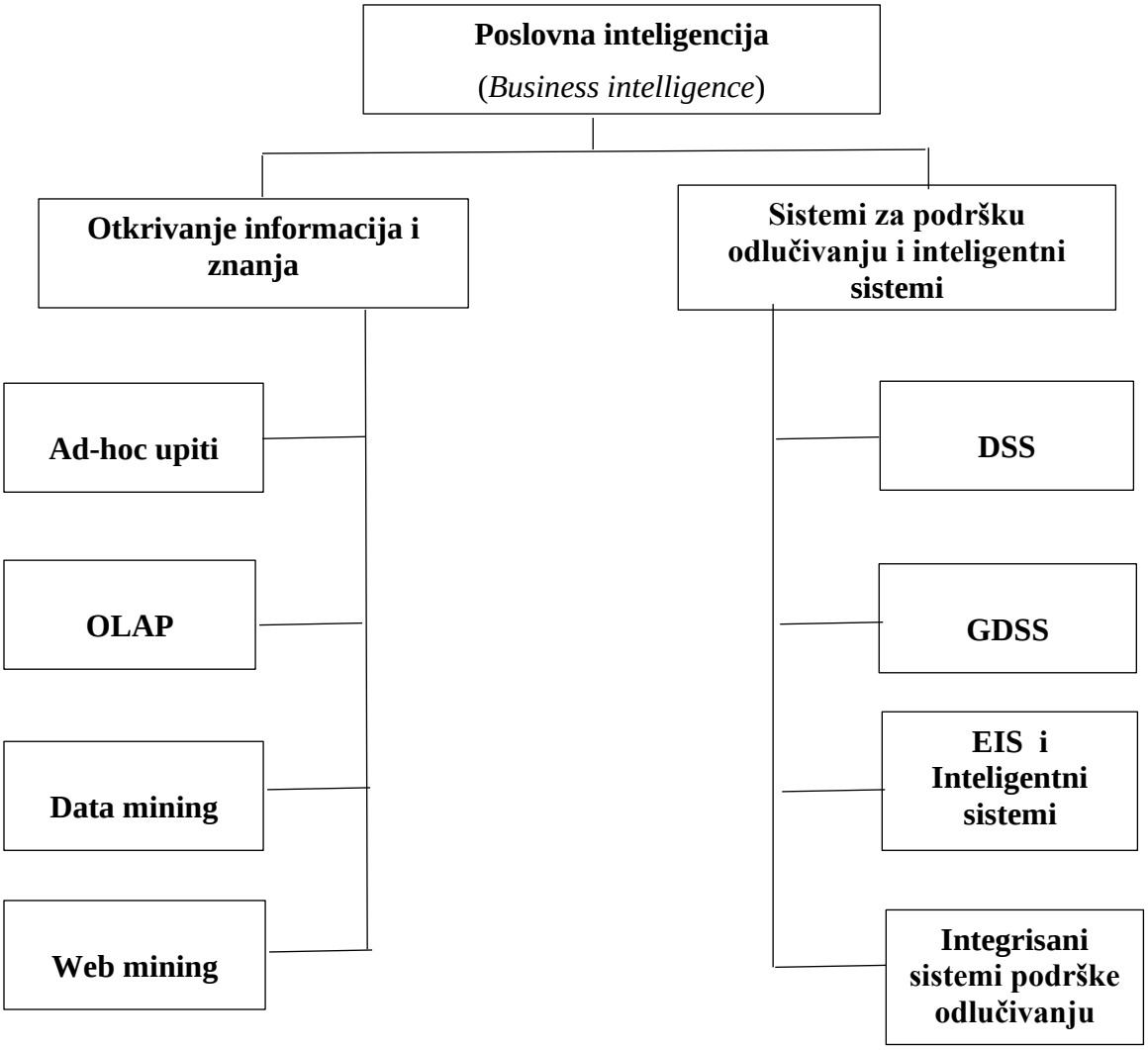
Poslovna inteligencija (*engl. Business Intelligence*) predstavlja niz procesa, tehnologija i alata koji imaju za cilj da se sirovi podaci transformišu u korisne informacije koje doprinose sprovođenju profitabilnih poslovnih aktivnosti.

Očigledno je pojam Poslovne inteligencije dovoljno širok da obuhvati mnoge softverske proizvode i usluge za prikupljanje, analizu, pristup podacima u cilju obezbeđenja donošenja dobrih poslovnih i strateških odluka. Na sledećoj slici je prikazana struktura i način funkcionisanja koncepta poslovne inteligencije.



Slika 22. Skladište podatka i koncept Poslovne inteligencije[5]

[5] Na osnovu Information Technology and Management, Turban, Leidner, McLean, Wetherbe, 5th ed. John Wiley & Sons, 2006, str.419



Slike 23. Alati poslovne inteligencije[6]

Jedna grupa alata za poslovnu inteligenciju obezbeđuje otkrivanje informacija i znanja iz podataka. Ovde je reč o sledećim alatima:

- ***Ad hoc* upiti,**
- Online analitička obrada (*Online Analytical Processing* - ***OLAP*),**
- Pretraživanje podataka (***Data mining***),
- Pretraživanje web-a (***Web mining***).

Koncept poslovne inteligencije pruža kompanijama niz mogućnosti, kao što su: veća produktivnost, bolja vidljivost poslovnih procesa, celovit pogled na poslovne procese, usmeravanje poslovnih aktivnosti, pojačana poslovna odgovornost, demokratizacija korišćenja

[6] Na osnovu Information Technology and Management, Turban, Leidner, McLean, Wetherbe, 5th ed. John Wiley & Sons, 2006, str.425

različitih alata za analitiku. Potencijalni problemi vezani za ovaj koncept su: visoki troškovi uvođenja BI tehnologija i alata, kompleksnost sistema, vremenski zahtevan proces uvođenja ovog koncepta.

Ovaj koncept nalazi primenu u mnogim industrijama čije se finkcionisanje zasniva na obradi velike količine podataka (*data-intensive industries*), kao što su finansijske institucije, telekomunikacione kompanije, maloprodaja, vazdušni saobraćaj, i sl.

Budućnost poslovne inteligencije karakterišu sledeći trendovi:

- ***Cloud-based BI i analitika*** – sve veći broj proizvođača nudi ove opcije u formi SaaS (*Software as a Service*), kao odgovor na sve veću količinu podataka;
- **Veštačka inteligencija** (*engl. Artificial Intelligence – AI*) i algoritmi mašinskog učenja u konceptu poslovne inteligencije;
- **Kolaborativna poslovna inteligencija** (*engl. Collaborative BI*) – povezivanje koncepta poslovne inteligencije sa alatima saradnje, kao što su društvene mreže;
- **Ugrađena BI** (*engl. Embedded BI*) – kombinovanje (ugradnja) softvera poslovne inteligencije sa drugim softverskim aplikacijama;
- **Napredna BI i analitika** (*engl. Advanced BI and Analytics*) – uključenje naprednih tehnika analitike podataka u koncept poslovne inteligencije.

2.2. Online analitička obrada
(On Line Analitical Processing - OLAP)

OLAP (***On Line Analitical Processing - OLAP)*** predstavlja tehnologiju koja omogućava korisnicima da nezavisno obavljaju analize podatke iz sistema ***višedimenzionalnih baza podataka.*** Radi se o alatima za interaktivnu analitičku obradu podataka. OLAP baze podataka se pojavljuju u formi jedne ili više kocki, što omogućava kreiranje upita relativno jednostavnije. Korisnici mogu da pristupaju i ekstrahuju poslovne podatke iz različitih perspektiva. Sa ovom tehnologijom **podaci se obično prethodno agregiraju** (i na drugi način obrađuju) i čuvaju u OLAP bazi podataka, što takođe doprinosi da se analiza obavi brže u poređenju sa korišćenjem relacionih baza podataka.

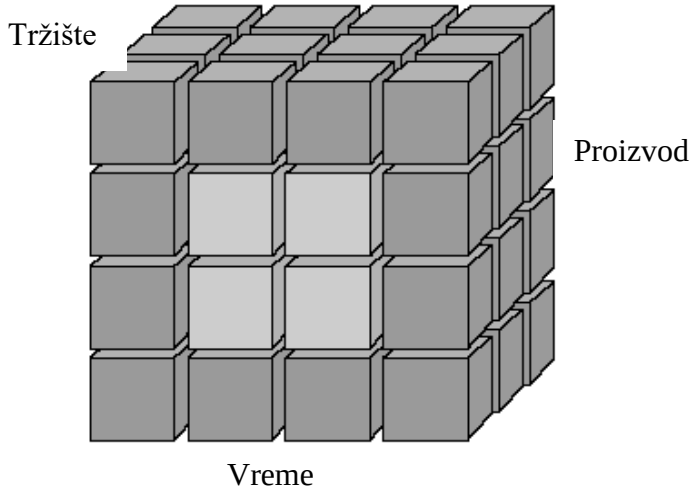
Osnovni ciljevi primene OLAP tehnologije su: da se odgovori na upite korisnika, kao i da se obavi analiza odnosa u podacima, pronađu „šabloni“ ponašanja, trendovi, izuzeci i sl. OLAP upiti imaju karakter online upita koji:

- pristupaju veoma velikim količinama podataka,
- analiziraju odnose između mnogih vrsta elemenata podataka,
- sadrže zbirne podatke (npr. obim prodaje, planirani iznosi i potrošeni iznosi),
- porede zbirne podatke u vremenu,
- predstavljaju podatke po različitim kriterijumima,
- obavljaju složena izračunavanja među elementima podataka,
- imaju mogućnost da brzo reaguju na korisničke zahteve.

Primena OLAP alata je zasnovana na sledećim principima:

1. višedimenzionalni pogled na podatke,
2. transparentnost podataka za korisnika,
3. jednostavan pristup podacima,
4. intuitivan pristup i operacije sa podacima,
5. klijent/server arhitektura,
6. jednostavnost i fleksibilnost dobijanja izveštaja,
7. podrška za višekorisnički pristup,
8. neograničen nivo dimenzija i agregiranja podataka.

U osnovi OLAP koncepta nalazi se OLAP kocka (engl. *OLAP cube*), koja predstavlja takvu strukturu podataka koja je pripremljena (izvršena optimizacija) za brzu analitičku obradu podataka. OLAP kocka sadrži višedimenzionalne podatke, koji su dobijeni iz različitih izvora. Na sledećoj slici je prikazana struktura OLAP kocke.



Slike 24. OLAP kocka

Na prethodnoj slici je prikazana trodimezionalna OLAP kocka, ali ona može imati i veći broj dimenzija. U okviru kocke podaci su uređeni u hijerahiji od logički nižeg nivoa ka višim nivoima: na primer za dimeziju Vreme mogu da postoje meseci, kvartali i godine.

Osnovne operacije koje se sprovode sa OLAP alatima su sledeće:

- **Agregiranje** (eng. *Roll-up*)
- **Ulazak u detalje** (eng. *Drill-down*)
- **Selekcija** (eng. *Slice and dice*)
- **Rotiranje** (eng. *Pivot*)

Operacija agregiranja podrazumeva kosolidaciju podataka, koja se sprovodi ili tako što se kreće u višem pravcu u hijerarhiji podataka, ili se vrši otklanjanje („oduzimanje“) jedne dimenzije. Na primer, kvartali se mogu agregirati u godinu. Operacija ulaženja u detalje je suprotna operaciji agregiranju. Ona se ispoljava u fragmentisanju podataka na manje delove. Ona se može obaviti ili tako što se ide u nižem pravcu (na dole) u hijerarhiji podataka, ili se „uvećava“ neka dimenzija. Na primer, dimenzija Vreme gde su podaci prikazani po kvartalima, može se prevesti u strukturu podataka gde je vremenska dimenzija prikazana po mesecima. Operacije selekcije (eng. *Slice and dice*) podrazumevaju da se izvrši selektovanje podataka prema zadatom filteru, koji može biti po jednoj dimenziji (engl. *slice*) ili po dve ili više dimenzija (engl. *dice*). Rezultat ove operacije je pod-kocka podataka (engl. *sub-cube*), koja zadovoljava tražene kriterijume. Na primer, prikazivanje podataka iz kocke samo za jedan proizvod i za

jedno tržište. Operacija rotiranja (*engl. Pivot*) podrazumeva da se rotiraju osnovne dimenzije podataka, tako da se menja način prikazivanja podataka po dimenzijama.

Vrste OLAP sistema

Osnovne tri vrste OLAP sistema su: ROLAP, MOLAP i HOLAP.

- **ROLAP (*Relational OLAP*)** čine alati za interaktivnu analitičku obradu koji koriste relacioni model kao osnovu svoje baze podataka. Naime ROLAP funkcioniše sa podacima koji se nalaze u relacionoj bazi podataka, a tabele sa činjenicama i dimenzijama se takođe čuvaju kao relacione tabele. Prednosti ovog modela su visoka efikasnost, kao i **skalabilnost** (efikasnost i kada se povećava obim podataka), a nedostaci su: angažovanje značajnih resursa, postojanje ograničenja u agregiranju podataka (ROLAP koristi SQL za sve kalkulacije kod agregiranja podataka), sporije performanse postavljanja upita u poređenju sa MOLAP alatima.
- **MOLAP (*Multidimensional OLAP*)** sistem čine alati za višedimenzionalnu interaktivnu analitičku obradu, koji koriste sopstvenu bazu podataka n-dimenzionalne matrične strukture. Ovi alati koriste OLAP kocku. Za razliku od ROLAP-a, u MOLAP-u se podaci prethodno agregiraju (i obavljaju druge kalkulacije) i čuvaju se pri kreiranju OLAP kocke, što čini **MOLAP bržim i efikasnijim sistemom** u poređenju sa ROLAP-om. U MOLAP-u se brzo obavljaju operacije selekcija (*slice and dice*) i ovaj model je jednostavniji za korisnike (posebno korisnike koji nemaju veliko iskustvo sa složenim upitima). Nedostatak MOLAP-a je da MOLAP nije pogodan za čuvanje detaljnih podataka, kao ni velike količine podataka.
- **HOLAP** (***Hybrid OLAP***) sistem predstavlja kombinaciju prethodne dve vrste OLAP tehnologija. Naime, ovaj sistem kombinuje prednosti MOLAP-a i ROLAP-a, a to su **brzo računanje MOLAP-a i skalabilnost ROLAP-a**. HOLAP sistem koristi dve baze podataka: agregirani podaci se čuvaju u multidimenzionalnim OLAP kockama, a detaljne informacije se čuvaju u relacionim bazama podataka. Međutim, kombinovanje ROLAP i MOLAP alata i aplikacija dovodi do izuzetne kompleksnosti HOLAP-a i potencijalnog preklapanja funkcionalnosti, što može da predstavlja značajan nedostatak ovih sistema.

  Pored ove tri osnovne vrste OLAP modela, postoje i sledeći modeli:

  DOLAP (*Desktop OLAP*) sistem podržava manje aplikacije pojedinačnih korisnika, koji učitavaju deo podataka iz baza podataka na svoj desktop računar i analiziraju ih.

Ovaj sistem pruža značajno manje funkcionalnosti u poređenju sa drugim modelima, ali je jeftiniji;

WOLAP (*Web OLAP*) je OLAP sistem kome se može pristupiti korišćenjem web pretraživača; ovaj sistem ima troslojnu arhitekturu – klijent, srednji sloj (*middleware*) i server baze podataka;

Mobile OLAP je sistem koji omogućava korisnicima da pristupe i analiziraju OLAP podatke koristeći mobilne uređaje;

SOLAP (*Spatial OLAP*) sistem je dizajniran sa ciljem da omogući upravljanje prostornim i drugim podacima u okviru GIS-a (engl. *Geographic Information System*).

OLTP vs OLAP

Na kraju izlaganja o OLAP alatima, korisno je ukazati na bitne razlike između OLTP (engl. *Online Transaction Processing*) i OLAP sistema. OLTP je sistem koji podržava aplikacije koje su orijentisane na obradu transakcija, odnosno administriraju svakodnevne transakcije u poslovnom sistemu. Značajne karakteristike OLTP sistema su:

- OLTP sistem obrađuje transakcije koje sadrže relativno mali obim podataka;
- Indeksiranim podacima u bazi podataka može se lako pristupiti;
- Ovi sistemi imaju obično veliki broj korisnika;
- OLTP ima kratko vreme odgovora;
- Korisnici mogu direktno pristupiti bazama podataka;
- OLTP obično čuva podatke od prethodno nekoliko dana ili nedelje;
- Ovi sistemi su optimizovani za obradu transakcija, a ne za analitičke potrebe.

U sledećoj tabeli su sumirane razlike između OLTP i OLAP sistema:

Tabela 5: Razlike između OLTP i OLAP sistema

OLTP	**OLAP**
OLTP je sistem za online **obradu transakcija**	OLAP je sistem za online **analitičku obradu**
OLTP je podrška za **obradu transakcija**	OLAP je podrška za proces **donošenja odluka**

OLTP je dizajniran za poslovne operacije **u realnom vremenu**	OLAP je dizajniran za analizu poslovnih rezultata po različitim atributima i kategorijama
OLTP i njihove transakcije su izvori podataka	OLTP baze podataka su izvori podataka za OLAP
OLTP karakteriše **veliki broj kratkih online transakcija**	OLAP karakteriše **veliki obim podataka (individualnih i agregiranih)**
OLTP koristi **tradicionalne DBMS**	OLAP koristi tehnologiju skladišta podataka – **DW**
Uglavnom se koriste operacije: *insert, update* i *delete* podataka u bazi podataka	Koriste se uglavnom operacije **filtriranja**
OLTP baze podataka se **brzo menjaju** - velika volatilnost	OLAP baza podataka se **ne menja često, povremeno**
Veoma kratko vreme odgovora sistema (meri se u milisekundama)	**Relativno kratko vreme odgovora** sistema (meri se u sekundama, minutima)
Dozvoljene su ***read/write* operacije**	Uglavnom ***read,*** a veoma retko *write* operacije
Upiti su **standardizovani i relativno jednostavni**	**Kompleksni upiti**
Kompletan back-up	**Back-up se povremeno** pravi
Moguće su hiljade korisnika OLTP baze podataka	Moguće je do stotine korsnika OLAP baze podataka

Primeri OLTP sistema su: rad bankomata – ATM, online bankarstvo, online rezervacija avionskih karata, primanje porudžbina, dodavanje artikla u kolica za kupovinu u e-trgovini i sl.

Svaki DW sistem je istovremeno i OLAP sistem. Jedan primer primene OLAP alata može biti sledeći: jedna kompanija može poželeti da uporedi prodaju svojih proizvoda u Decembru sa prodajom u Avgustu na tržištu A, a zatim da uporedi ove rezultate sa prodajom na drugom tržištu B, za koje se podaci čuvaju u drugoj bazi podataka.

2.3. Napredne tehnike pretraživanja podataka

– Data mining

Savremeni koncept pretraživanja (i istraživanja) podataka, poznat pod nazivom Data mining (DM), je naučna disciplina koja se pojavila početkom 2000-tih godina. Poreklo ove naučne discipline je vezano za više različitih naučnih oblasti: sa jedne strane, za razvoj većeg broja akademskih disciplina, gde se posebno izdvajaju statistika i mašinsko učenje, a sa druge strane

za praksu obrade velike količine podataka, koja se ostvaruje uz primenu tehnologija upravljanja bazama podataka i sistema za podršku odlučivanja. Osnovna ideja ovog koncepta je efikasno otkrivanje značajnih prikrivenih struktura (modela i lokalnih obrazaca – „*patterns*“) u ogromnim količinama podataka smeštenim u bazama i skladištima podataka. Naime, čitav koncept počiva na primeni različitih tehnika (algoritama) pomoću kojih se mogu pronaći i izdvojiti nepoznate i neočekivane informacije, bilo da su to veze podataka, obrasci ponašanja, korelacije, pravila, ili trendovi, iz poslovnih podataka kao sirovog materijala. Na osnovu prethodnog može se zaključiti da je Data mining naučna disciplina koja se bavi pretraživanjem velike količine podataka u cilju otkrivanja nekih prikrivenih (do tada nepoznatih) obrazaca ponašanja i/ili nepoznatih veza u podacima.

> ***Data mining*** je „otkrivanje značajnih novih korelacija, pravila i trendova u procesu ispitivanja ogromne količine podataka smeštenih u skladištima podataka, koristeći tehnologije prepoznavanja obrazaca ponašanja, kao i statističke i matematičke tehnike”. (*Gartner Group*)

Data mining možemo shvatiti kao metaforu, koja treba da ukaže na ogroman obim podataka (rudnike podataka) koje treba „prekopati“ u potrazi za izuzetno vrednim otkrićima u poslovnom smislu – informacijama i znanjem. Često se u literaturi kada se govori o *data mining*-u koriste i izrazi *Knowledge Discovery in Databases* (KDD) ili *Knowledge extraction.*

Generalno posmatrano *data mining* koncept i tehnike primenjuju se za rešavanje sledećih tipova problema:

- klasifikacija,
- ocenjivanje
- predviđanje
- grupisanje,
- otkrivanje pravila udruživanja – asocijacije („*association rules*“),
- eksplorativna analiza podataka – posebno vizuelizacija podataka i sl.

Prva tri zadatka predstavljaju primere direktnog *data mining*-a, dok su sledeća tri zadatka primeri primene indirektnog *data minig*-a. Direktni *data minin*g je pristup od vrha ka dnu (engl. *top-down approach*) koji se koristi u situaciji kada apsolutno znamo šta je naš cilj: da se kreira model koji će objasniti jednu posebnu promenljivu pomoću ostalih promenljivih. Indirektni *data mining* je pristup od dna ka vrhu (engl. *bottom-up approach*), koji “ostavlja podacima da sami govore za sebe”, a ima za cilj da otkrije vezu između promenljivih, bez izdvajanja neke

promenljive posebno. Konkretno za rešavanje navedenih zadataka koriste se mnoge statističke tehnike i tehnike iz oblasti mašinskog učenja. Sve DM tehnike mogu se grubo podeliti na dve velike grupe:

- deskriptivne tehnike i
- prediktivne tehnike.

Postoji mnoštvo takvih metoda i algoritama, kao i niz njihovih modifikacija, od kojih se kao najznačajnije izdvajaju sledeće: grupisanje - klasterisanje, drvo odlučivanja, *Bayesove* mreže, neuronske mreže, *fuzzy* logika, genetski algoritmi, metoda analize potrošačke korpe, kao i mnoge tradicionalne statističke tehnike (razni regresioni modeli i modeli predviđanja, diskriminaciona analiza, korelacione tehnike i dr.).

DM tehnike se koriste u mnogim poslovnim oblastima. Prema izveštajima *Gartner* grupe (*www.gartner.com*), najveći broj uspešnih kompanija koje se nalaze na listi *Fortune* magazina koriste data mining u svom poslovanju. Oblasti primene data mining-a su:

- Maloprodaja i prodaja,
- Bankarstvo,
- Berzansko poslovanje i trgovina hartijama od vrednosti,
- Osiguranje,
- Marketing,
- Industrijska prerada i proizvodnja,
- Telekomunikacije,
- Računarski hardver i softver,
- Državna uprava, odbrana, rad policije,
- Vazduhoplovne kompanije,
- Zdravstvena zaštita,
- Radio i TV difuzija.

Osnovne prednosti koje donosi primena DM tehnika i koje govore u prilog njihovog korišćenja su: obezbeđuju kompanijama korisne informacije i znanje koji im pružaju kompetitivnu prednost, pružaju pomoć kompanijama u procesu donošenja odluka, podržava se automatsko generisanje trendova, predviđanja, otkrivanja „neobičnih“ ponašanja (na primer, *outlier analysis*), doprinosi se otkrivanju uobičajenih ili, pak, neobičnih „obrazaca“ kupovine, otkrivaju se posebne grupe kupaca za marketinške svrhe, predviđa se mogućnost da kupci napuste korišćenje usluga kompanija, doprinosi se diferencijaciji između profitabilnih i

neprofitabilnih kupaca, doprinosi se optimizaciji web stranica time što se obezbeđuju prilagođene ponude svakom posetiocu itd.

Glavni nedostatak primene ovih tehnika jeste činjenica da je za korišćenje DM softvera potrebno iskustvo ili značajna obuka korisnika, a takođe, izbor data mining algoritma je izazov za korisnike, jer zahteva posedovanje usko domenskog znanja (ekspertize) o pojedinim DM tehnikama.

2.4. Pretraživanje teksta (*Text mining)*

Pretraživanje teksta (*Text mining*), ili kako se još naziva analitika teksta (*Text analytics*), se može definisati kao proces istraživanja i analize velike količine nestruktuiranih tekstualnih podataka primenom softvera koji može da otkrije specifične koncepte, razne „obrasce“ (*patterns*), posebne teme, ključne reči i druge karakteristike u tekstualnim podacima. Dakle, za razliku od *data mining*-a, gde se podrazumeva pretraživanje velike količine uglavnom struktuiranih podataka, *text mining* se bavi pretraživanjem slabo struktuiranih i nestruktuiranih podataka. Podsetimo se procena da je preko 70% podataka u kompanijama nestruktuiranog ili slabo struktuiranog tipa.

> ***Text mining*** je proces pretraživanja i analize slabo struktuiranih i nestruktuiranih tekstualnih podataka, primenom softverskih alata za otkrivanje posebnih tematskih celina, ključnih reči, drugih atributa teksta, kao i karakteričnih veza između tekstualnih elemenata.

Cilj pretraživanja teksta jeste da pomogne kompanijama da pronađu potencijalno značajne informacije u velikoj količini tekstualnih dokumenata, planova, web dokumenata, elektronskih poruka, log datoteka, komunikacija na društvenim mrežama i drugim tekstualnim izvorima.

Proces pretraživanja teksta se sprovodi kroz nekoliko faza: identifikovanje teksta koji je interesantan za pretraživanje, zatim grupisanje, kategorizacija i dodavanje oznaka (engl. *tags*) na različite komponente teksta, sumiranje tekstualnih nizova, kreiranje taksonomije, kao i izdvajanje korisnih informacija kao što su, na primer, učestalost pojavljivanja pojedinih reči, ili veze između pojaviljivanja pojedinih tekstualnih elemenata. Zatim se primenjuju različiti analitički modeli kako bi se otkrili interesantni obrasci u podacima koji mogu da generišu poslovnu vrednost. U ovom procesu koriste se algoritmi obrade prirodnog jezika – NPL

algoritmi (engl. *Natural Language Processing*), kao i algoritmi „dubokog učenja“ (engl. *deep learning*), koji se zasnivaju na proširenju algoritama veštačkih neuronskih mreža.

Generalno se može reći da pretraživanje teksta pomaže organizacijama u sledećem:

- da pronađu skriveni sadržaj dokumenata, uključujući i dodatne korisne relacije u tekstu,
- da klasifikuju web sadržaj,
- da grupišu dokumenta prema zajedničkim temama (na primer, da identifikuje sve klijente osiguravajuće kompanije čije su žalbe slične),
- da ukažu na potencijalno lažne reklamacije,
- da sagledaju i ocene kandidate za posao na osnovu, na primer, prisustva pojedinih reči u njihovim biografijama,
- da blokiraju *spam u* elektronskim porukama.

Jedna od najpoznatijih primena tehnika pretraživanja teksta jeste analiza osećanja (eng. *sentiment analysis*) ili kako se još označava istraživanje mišljenja (eng. *opinion mining*). Ova analiza se koristi za otkrivanje osećanja, mišljenja i stavova potrošača o kompaniji, o proizvodima i uslugama, o dobrim i lošim stranama poslovanja kompanije. Primenom ove analize pretražuju se različiti tekstualni sadržaji – *online* pregledi, elektronske poruke, mišljenja na društvenim mrežama, dokumenta o komunikaciji sa potrošačima preko *call center*-a i sl. Jasno je da se na ovaj način mogu otkriti i pozitivni i negativni stavovi i mišljenja potrošača, koji mogu značajno doprineti unapređenju usluga potrošačima, kreiranju novih marketinških kampanja, sagledavanju tržišnih pogodnosti i/ili pretnji, ponašanju konkurencije i sl.

2.5. Pretraživanje Web-a (*Web mining)*

Pretraživanje web-a (engl. Web mining) predstavlja jednu granu oblasti pretraživanja podataka (*data mining*-a) koja je usmerena na otkrivanje „obrazaca“ ponašanja na web-u. Ovde se misli na pretraživanje kompletnog sadržaja na web-u, počev od sadržaja na web stranama, aktivnosti primenom web pretraživača (*browser logs*) i log dokumenata na serverima (*server logs*), dokumenti sa hipervezama i dr.

Web mining je proces primene tehnika i algoritama data mining-a da bi se izdvojile korisne informacije sa web-a: iz web dokumenata i usluga, web sadržaja, *hyperlink* dokumenata i *log* dokumenata sa web servera.

Razlikuju se tri kategorije *web mining*-a:

- *Web-content mining*: pretraživanje sadržaja Web strana i web dokumenata, koji čine tekst, slike i audio/video dokumenti, radi dobijanja korisnih informacija;
- *Web-usage mining*: pretraživanje i analiza pristupa Web stranama, odnosno *Web log* -ova, i drugih informacija povezanih za aktivnosti korisnika na Web-u, kao što su korisnikovo pretraživanje na više Web lokacija, poreklo korisnika, vreme zadržavanja korisnika na web strani, kretanje između strana, broj *click*-ova na pojedinim elementima *web* strane i dr.
- *Web-structure mining*: analiza čvorova i strukturnih veza web strane primenom teorije grafova. Ovde se analizira povezanost jedne web strane sa drugim web stranama, kao i struktura same web strane.

S obzirom na činjenicu da je savremeno poslovanje neraskidivo povezano sa *web* aktivnostima, može se reći da je primena *web minig-a* moguća u svim poslovnim oblastima. U nastavku je dato nekoliko karakterističnih **primena *web* mining-a:**

- filtriranje informacija (e-mails, novine, magazini i dr.),
- aktivnosti prismotre (konkurencije, tehnoloških inovacija, patenata i dr),
- pretraživanje *web log*-ova za potrebe marketinških istraživanja (*clickstream* analiza).

References:

Rainer R.K.Jr, Turban E., Introduction to Information Systems, prevod, Datastatus, 2009

Turban E., Leidner D., McLean E., Wetherbe J., (2006), Information Technology and Management, 5th ed. John Wiley & Sons, Inc, USA

Turban, E., Pollard, C., Wood, G. (2018), Information Technology for management, On-Demand Strategies for Performancem, Growth and Sustainability, 11th ed. John Wiley & Sons, Inc.

https://searchbusinessanalytics.techtarget.com/definition/text-mining

https://www.techopedia.com/definition/15634/web-mining

https://en.wikipedia.org/wiki/Web_mining

https://www.techopedia.com/definition/15634/web-mining

https://www.techopedia.com/definition/3801/garbage-in-garbage-out-gigo

https://searchdatamanagement.techtarget.com/definition/dirty-data

https://searchdatamanagement.techtarget.com/downloadOffer/fulfillment/252471304/a27271e448e12210VgnVCM1000000d01c80aRCRD?submit=1

https://www.r1soft.com/blog/dark-data-what-is-it-and-why-should-i-care

Digging up dark data: What puts IBM at the forefront of insight economy | #IBMinsight". SiliconANGLE. 2015-10-30. Retrieved 2015-11-03. https://siliconangle.com/2015/10/30/ibm-is-at-the-forefront-of-insight-economy-ibminsight/

https://www.gartner.com/en/information-technology/glossary/dark-data

https://www.ibmbigdatahub.com/blog/big-data-challenge-transformation-manufacturing-industry?

Garner Glosary, Information technology, Dark Data

https://www.gartner.com/en/information-technology/glossary/dark-data

https://www.guru99.com/data-warehousing-pdf.html

https://www.talend.com/resources/data-lake-vs-data-warehouse/

https://databricks.com/discover/data-lakes/introduction

https://databricks.com/discover/data-lakes/history