## 1.slaids

Labdien! Mani sauc Lauma Svilpe un mana bakalaura darba temats ir neviendabīgu integrētu datu avotu evolūcijas apstrāde. Darba vadītāja: asociētā profesore datorzinātņu doktore Darja Solodovņikova.

## 2.slaids

Sākumam vēlos ieskicēt esošo situāciju saistībā ar datu glabāšanu.

- Kamēr 2000.gadā visā pasaulē glabāto datu apjoms bija mērāms vien simtos petabaitu,
- 2020.gadā tiek prognozēts, ka šis skaitlis sasniegs jau ap 35 zetabaitiem, kas ir ļoti ievērojams pieagums.
- Pēdējo gadu laikā ieviesti vairāki jauni rīki, tehnoloģijas un ietvari, kas atbalsta šāda liela apjoma datu analītiku, tomēr tie galvenokārt risina tikai problēmas, kas saistītas ar vaicājumu izpildes ātrumu, datu atlases ērtumu, vienkāršība u.c.

Neatrisināts paliek datu un to struktūru evolūcijas jautājums.

## 3.slaids

Balstoties uz pastāvošo problēmu, tiek izvirzīts bakalaura darba mērķis - atrast risinājumu, kā apstrādāt neviendabīgu integrētu datu avotu evolūcijas rezultātā radušās izmaiņas.

Lai sasniegtu izvirzīto mērķi,

- bija nepieciešams veikt literatūras analīze par datu noliktavām, lielajiem datiem un ETL procesiem
- izpētīt esošo Latvijas Universitātes Datorikas fakultātē izstrādāto datu avotu evolūcijas sistēmu
- izstrādāt neviendabīgu integrētu datu avotu evolūcijas apstrādes mehānismu, kas paredzēts iekļaušanai jau esošajā sistēmā.

## 4.slaids

Pašos pirmsākumos datu glabāšanai tika izmantoti ļoti dārgi un ierobežotas ietilpības mehānismi –

- perfokartes, magnētiskās lentas,
- pēc tam ar pavisam jauniem papildinājumiem nāca klajā diskatmiņa.
- Drīz vien diskatmiņas tika papildinātas ar datu bāzu pārvaldības sistēmām, kuru galvenais izmantošanas ieguvums bija ļoti ātra datu ievietošana sistēmā

Līdz ar datu glabāšanas veidu attīstību, tika ieviests jēdziens “lēmumu atbalsta sistēmas”, kas paredzētas stratēģisku lēmumu pieņemšanas atbalstam, iekļaujot tajā datu analīzi un apstrādi.

- Līdz ar šī jēdziena parādīšanos, klajā nāca dažādi OLAP jeb tiešsaistes analītiskās apstrādes rīki,
- kas vēlāk tika iekļauti lietojumsistēmās.
- Pēc tam tika ieviestas arī 4GL tehnoloģijas, kuru ideja bija padarīt programmatūras izstrādi tik vienkāršu, lai to varētu darīt jebkurš.

Kā jau noprotams, šeit nevar iezīmēt nekādu konkrētu struktūru –

radās nekārtība, ko mēdz saukt par “zirnekļa tīkla” vidi, kura mēdza izaugt neiedomājami sarežģīta, kā rezultātā to vairs nebija iespējams ne pienācīgi uzturēt, ne pilnveidot. Tai ar laiku pat radies nosaukums “dabiski attīstīta arhitektūra”, jo bija maz iespēju tās attīstību ietekmēt.

## 5.slaids

Datu noliktavas būtiski izmainīja IT speciālistu domāšanu – līdz šim pastāvēja uzskats, ka viena datu bāze ir paradzēta, lai glabātu jebkādam nolūkam paredzētus datus. Tomēr līdz ar datu noliktavas jēdziena parādīšanos kļuva acīmredzams, ka ir nepieciešamas dažādu veidu datu bāzes,

tādēļ “Zirnekļa tīkla” vide tika sadalīta divās atsevišķās daļās, lai glabātu datus pēc to nozīmes datu analītikas kontekstā.

## 6.slaids

Datu noliktavu kontekstā svarīgs jēdziens ir arī lielie dati, jo visbiežāk to glabāšanai un analīzei tiek izmantotas tieši datu noliktavas.

Lielos datus raksturo to apjoms – tie var būt terabaitus lieli datu faili, dažādi ieraksti, datu bāzes tabulas. Ātrums, kādā dati ienāk noliktavā – sērijveidā, reāllaikā vai kā nepārtraukta datu straume. Kā arī datu dažādība – tie sākotnēji ir strukturēti, nestrukturēti vai jaukti dati.

## 7.slaids

Lai neapstrādātus lielos datus ievietotu datu noliktavās, tiek izmantoti ETL procesi. Atšifrējums latviešu valodā saīsinājumam ETL nozīmē – iegūšana, transformācija, ielādēšana.

Datu iegūšanas procesā tie tiek savākti no dažādiem ārējiem datu avotiem,

kur tie tiek validēti, attīrīti un filtrēti.

Transformācijas procesā dažādu struktūru dati tiek pārveidoti vienā vienotā struktūrā.

Kad tas izdarīts, datus var ielādēt datu noliktavā.

## 8.slaids

Latvijas Universitātes Datorikas fakultātē piedāvāts datu noliktavas risinājums lielo datu analīzei, kurā iekļauti algoritmi dažādu evolūcijas rezultātā radušos izmaiņu atklāšanai. Praktiski risinājums pielietots publikāciju lielo datu sistēmā, kuras mērķis ir no vairākiem neviendabīgiem datu avotiem integrēt datus par Latvijas Universitātes darbinieku un studentu publikācijām un nodrošināt šo datu analīzi datu noliktavā.

Datu avotu evolūcijas sistēmas arhitektūras pamatkomponentes ir datu avoti, datu maģistrāle, metadatu glabātuve un adaptācijas komponente.

Datu avotu līmenī neviendabīgi jeb dažādu formātu dati tiek iegūti no dažādiem avotiem un ielādēti sistēmā tālākai to apstrādei.

Datu maģistrāle sastāv no vairākiem līmeņiem. Pirmajā līmenī tiek glabāti neapstrādāti dati, kas iegūti pa tiešo no datu avotiem. Katra nākamā maģistrāles līmeņa dati tiek iegūti no iepriekšējā līmeņa ar ETL procesu palīdzību.

🅞Sistēmas arhitektūras darbība pamatā balstīta uz metadatu glabātuvē esošajiem datiem. Izmantojot metadatu pārvaldības rīku un definējot dažādus metadatus, izstrādātājs nosaka, kā darbosies sistēma.

🅞Lielo datu noliktavas pamatelements, kas atbild par datu avotu un informācijas prasību izmaiņu apstrādi, ir adaptācijas komponente.

Bakalaura darba ietvaros adaptācijas komponente papildināta ar mehānismu, kas datu avotu evolūcijas rezultātā radušās izmaiņas adaptē sistēmas metadatos.

## 9.slaids

Datu avotu evolūcijas rezultātā iespējamas izmaiņas gan datu avotu struktūrā, gan īpašībās. Daži iespējamie izmaiņu piemēri no konkrētās publikāciju lielo datu sistēmas redzami ekrānā. Lai būtu iespējams turpināt datu ielasīšanu noliktavā no evolucionējušiem avotiem, nepieciešams veikt izmaiņu adaptāciju.

## 10.slaids

Lai apstrādātu datu avotu evolūcijas rezultātā radušās izmaiņas un veiksmīgi adaptētu tās sistēmas metadatos, bakalaura darba ietvaros izstrādāts mehānisms, kurš darbojas, balstoties uz dažādiem metadatiem.

Tāpat kā jau esošā sistēma, arī šis mehānisms izstrādāts, izmantojot Oracle SQL relāciju datu bāzi. Darba gaitā izstrādāti adaptācijas scenāriji tikai reāli publikāciju sistēmā notikušajiem izmaiņu veidiem un uz šiem datiem arī testēti.

🅞Izmaiņu adaptācijas operācijas ir darbības jeb soļi, kas jāveic, lai adaptētu izmaiņas sistēmā. Tās ir iespējami īsas un universālas, var būt gan aprakstošā tekstuālā formā, gan izpildāmas datu bāzes procedūras formā. Šīs darbības tiek atsevišķi glabātas ekrānā iezīmētajā tabulā.

🅞Izmaiņu adaptācijas scenārijs ir secīgu darbību virkne, kas tiek veikta, lai veiksmīgi adaptētu izmaiņu sistēmā. Šie scenāriji tiek glabāti ekrānā iezīmētajā tabulā. Gan šī, gan iepriekšminētā tabula tiek aizpildīta manuāli pirms izmaiņas rašanās.

🅞Lai sekotu līdzi izmaiņu adaptācijas procesam un glabātu informāciju par katras darbības izpildi, nepieciešama tabula, kurā glabājas katrai reāli notikušai izmaiņai atbilstošais adaptācijas scenārijs. Šajā tabulā ieraksti ģenerējas automātiski līdz ar izmaiņas rašanos.

🅞Neskatoties uz to, ka izmaiņas tips ir nosakāms jau pie tās rašanās, tas negarantē viennozīmīga izmaiņas adaptācijas scenārija esamību. Ir dažādi nosacījumi, kuru izpildes rezultātā scenārijs var zaroties. Tie var būt gan automātiski izpildāmi, gan manuāli apstrādājami un arī tiek glabāti atsevišķā tabulā un izveidoti manuāli pirms izmaiņas rašanās.

🅞Atsevišķi tiek glabāta informācija arī par katram scenārija solim atbilstošajiem nosacījumiem un to izpildi.

🅞Izmaiņu adaptācijas procesa laikā no izstrādātāja var tikt prasīti dažādi papildus dati, kas būs nepieciešami tālāko scenāriju darbību izpildei un nosacījumu pārbaudei.

🅞Izstrādātais risinājums ar esošo sistēmu saistīts pamatā izmantojot izmaiņu tabulu, kas glabā identifikatoru uz konkrēto reāli notikušo izmaiņu, kā arī **Tipu** un **Autoru** klasifikatorus.

## 11.slaids

Veicot pirmreizējo izmaiņas apstrādi, tiek apstrādātas tikai vēl neapstrādātās izmaiņas. Nosakot izmaiņas tipu, pareizajā secībā tiek atlasīti visi izmaiņas tipam atbilstošie scenārija soļi, pēc kā izveidoti konkrētajai izmaiņai atbilstošie procesa soļi. Pēc tam tiek atlasīti konkrētajam procesa solim atbilstošie scenārija soļu manuālie nosacījumi. Kad visa šī informācija saglabāta atbilstošajās tabulās, izmaiņas adaptācijas procesu var uzskatīt par sākušos.

## 12.slaids

Darba ietvaros netiek apskatīti visu iespējamo izmaiņu veidu apstrādes scenāriji – aprakstīti septiņi reāli publikāciju sistēmā notikušo izmaiņu adaptācijas scenāriji. Katra izstrādātā scenārija mērķis ir panākt, lai tas būtu pēc iespējas automātiski izpildāms, tomēr konceptuālu lēmumu pieņemšana par datu izmantošanu sistēmā nav iespējama bez cilvēka iejaukšanās.

## 13.slaids

Ekrānā redzams piemērs kā izskatās viens no izstrādātajiem izmaiņu adaptācijas scenārijiem. Piemēram, lai pievienotu jaunu datu maģistrāles līmeni, pirmkārt, nepieciešams, aprakstīt un izveidot jaunā līmeņa struktūru metadatos un definēt atbilstošos ETL procesus. Ja datu maģistrāles līmeņa aizpildīšanai nepieciešams pievienot jaunu datu avotu, tad tas tiek izdarīts, pievienojot avota datu kopas piemērus. Sistēma automātiski spēj iegūt pievienotās kopas struktūru un pievienot to jaunajam datu maģistrāles līmenim.

## 14.slaids

Tālāk ekrānā redzamajā shēmā peldceliņu diagrammā aprakstīta datu maģistrāles līmeņa dzēšana. Šīs izmaiņas adaptācija pamatā ir automātiska, jo sistēma spēj noteikt datu maģistrāles līmenim piesaistītās datu kopas un to avotus.

## 15.slaids

Darba gaitā izpildīti visi sākotnēji izvirzītie darba uzdevumi.

- Vispirms izanalizēta pieejamā teorētiskā literatūra par datu noliktavām, to arhitektūru un veidiem, kā arī tajās izmantotajiem ETL procesiem. Datu noliktavu pielietojuma kontekstā izpētīta un apkopota informācija par lielajiem datiem un to svarīgākajām raksturiezīmēm.
- Datu avotu evolūcijas kontekstā pētīta un analizēta esoša sistēma, īpašu uzmanību pievēršot tajā iekļautajai adaptācijas komponentei.
- Balstoties uz esošās sistēmas datu bāzes struktūras, kā arī pastāvošo metadatu glabāšanas mehānismu, izstrādāts risinājums, kas paredzēts datu avotu evolūcijas rezultātā radušos izmaiņu adaptācijai sistēmas metadatos.

Reāli notikušām datu avotu izmaiņām sastādīti pilni izmaiņu adaptācijas scenāriji, kas sastāv gan no automātiski izpildāmas funkcionalitātes (procedūru formā), gan manuāli veicamiem norādījumiem (veicamās darbības apraksta formā).

Pēc tam izstrādāta datu bāzes struktūra izmaiņu adaptācijas procesā radušos metadatu glabāšanai, kā arī datu atlases vaicājumi un procedūras pašas izmaiņu adaptācijas veikšanai.

## 16.slaids

Datu noliktavu jēdziens attīstījies līdz ar lielu datu plūsmu parādīšanos tehnoloģiju vidē. Šie dati tika uzkrāti, bet arvien vairāk tika novērotas dažādas "zirnekļa tīkla" vides, kuras bija grūti pārvaldīt, tāpēc drīz vien tika izdalīts atsevišķs datu noliktavas jēdziens.

Līdz ar datu noliktavu attīstību parādījies arī lielo datu jēdziens.

Taču, ņemot vērā lielo datu apjomu, dažādību un ātrumu, parādījušās datu avotu evolūcijas problēmas.

Latvijas Universitātes Datorikas fakultātē izstrādātā datu avotu evolūcijas sistēma iekļauj datu ieguvi no dažādiem avotiem, kā arī ETL procesus datu pārveidošanai vienotā struktūrā, lai tos būtu iespējams ievietot datu noliktavā.

Sistēmas darbība balstīta uz metadatiem, kas tiek glabāti par katru saņemto informācijas vienību un tās transformāciju integrācijai datu noliktavā.

Bakalaura darba ietvaros izstrādātais risinājums papildina esošo metadatu glabāšanas shēmu ar papildus datu bāzes struktūru un funkcionalitāti, lai veiktu datu avotu izmaiņu adaptāciju sistēmas metadatos.

Līdz ar to iespējams secināt, ka darba sākumā izvirzītais mērķis ir sasniegts.

Un šis rezultāts izmantojams gan kā esošās sistēmas papildinājums, gan kā universāls konceptuāls piemērs citu līdzīgu sistēmu papildināšanai ar šādu izmaiņu adaptācijas komponenti. Turpmāk plānots attīstīt sistēmu, iekļaujot visu iespējmo datu avotu izmaiņu veidu adaptācijas scenāriju realizāciju.