Ukládání a příprava dat – Projekt
Štruktúry vytvorených CSV súborov

Matej Berezný, Ondrej Valo, Švenk Adam
xberez03, xvaloo00, xsvenk00

1 1. dotaz skupiny A

V rámci prvej úlohy sa načítavali kolekcie 'hospitalizovany' a 'statistika_celkovo' v ktorý sa dáta následne združovali podla dátumov po 30 dňoch a využili ako unikátne kľúče. Zanechali sa atribúty v hospitalizovaných: 'pacient_prvni_zaznam'. A v štatistika celkovo: 'prirustkovy_pocet_nakazenych', prirustkovy_pocet_vylecenych', 'prirustkovy_pocet_provedenych_testu'
Následne boli premenované:

'datum' -> 'date'
'prirustkovy_pocet_nakazenych' -> 'Positive'
'prirustkovy_pocet_vylecenych' -> 'Recovered'
'prirustkovy_pocet_provedenych_testu' -> 'Tests'
'pacient_prvni_zaznam' -> 'Hospitalized'

Nakoniec boli tabuľky spojené pomocou dátumov.

Date	Category	Count

Tabulka 1: Štruktúra CSV súboru k 1. dotazu A

2 2. dotaz skupiny A

V druhej úlohe sa využila kolekcia z databáze 'nakazeny_kraj' z ktorej boli atribúty '_id', 'vek' a 'kraj_nuts_kod' uložené do požadovaného CSV súboru, nepotrebné atribútu sa neukladali, atribút 'vek' sa premenoval na 'Age' a atribút 'kraj_nuts_kod' bol premenovaný na 'Region'

_id	Age	Region	Count

Tabulka 2: Štruktúra CSV súboru k 2. dotazu A

3 Dotaz skupiny B

V dotaze skupiny B boli využité kolekcie: 'obyvatelia' a 'nakazeny_kraj'.

Kde v kolekcií obyvatelia sa dátum z dátového typu string pretypoval na date a zachovali sa iba atribúty a niektoré ich hodnoty:
Atribút 'vuzemi_cis' kde hodnoty sa rovnajú '100',
Atribút 'vek_cis' kde hodnoty sa rovnajú '0.0',
Atribút 'pohlavi_cis' kde hodnoty sa rovnajú '0.0',
Atribút 'casref_do' kde hodnota je najaktuálnejší rok',
'hodnota' -> 'total_people'

A v kolekcií nakazený kraj boli dáta zoskupené podľa dátumu po troch mesiacoch a podla kódu kraja. A brali sa záznami kde kód kraja nebol 0. Premenované atribúty boli:
'kraj_nuts_kod' -> 'Region'
'datum' -> 'Date'

Nakoniec bol pridaný stĺpec 'metric' ktorý sa rovná nakazaným za mesiac/celkový počet ľudí.

Nakoniec obe kolekcie boli pripnuté za seba a uložené do jedného CSV súbora.

Region	**index**	Date	infected_per_month	total_people	metric

Tabulka 3: Štruktúra CSV súborov k dotazu B

4 Vlastný dotaz 1

V prvom vlastnom dotaze boli využité kolekcie: 'hospitalizovani_ockovanie', 'zemreli_ockovanie', 'jip_ockovanie' z ktorých boli extrahované údaje o očkovaných/neočkovaných pacientoch, ďalej 'ockovanie_kraj', z ktorej sa získali údaje o očkovaniach v jednotlivých krajoch a tabuľky 'obyvatelia', z ktorej sa získal celkový počet obyvateľov pre jednotlivé kraje za posledný rok. Vo všetkých kolekciách sa dáta zoskupujú podľa dátumov po jednom mesiaci a zároveň sa dátum využíva ako unikátny kľúč. Následne sa stĺpec 'dátum' premenuje na 'date'.
V kolekcii obyvatelia sa berú dáta kde kód územia je hodnota 100 (iba kraje), a hodnota stĺpcov vek_cis a pohlavi_cis je 0.0 (nezáleží na veku či pohlaví) a kde casref_do je posledný rok. V kolekcií ockovany_kraj sa premenuje stĺpec druhych_davek na Vaccinated % a prepočíta sa jej hodnota na percentá k pomeru počtu obyvateľov. Nakoniec sa kolekcie spájajú podľa dátumu do dvoch tabuliek, kde jedna je pre nevakcinovaných a druhá pre vakcinovaných. Kde obe berú dáta z viacerých tabuliek. Kde vo finále vo vakcinovanej tabuľke date nám predstavuje dátum Category predstavuje kategóriu ako úmrtie alebo hospitalizácia a percentá počet percent ku počtu obyvateľov. nakoniec je ešte jedna kolekcia predstavujúca preočkovanosť obyvateľstva.

Date	Category	Percent

Tabulka 4: Štruktúra CSV_0 a CSV_1 súborov k vlastnému dotazu 1 - Očkovaní/Neočkovaní

Date	Vaccinated %

Tabulka 5: Štruktúra CSV_2 súboru k vlastnému dotazu 1

5 Vlastný dotaz 2

V druhom vlastnom dotaze sa využili kolekcie 'hospitalizovany' a 'statistika_celkovo'
V kolekcii hospitalizovaný boli dáta zoskupené podľa dátumu po jednom mesiaci a takýto dátum bol využitý ako unikátny kľúč.
V kolekcii štatistika celkovo bol dátum pretypovaný z string na date a dáta boli zoskupené podla dátumu po jednom mesiaci. A kolekcie boli prepojené pomocov dátumov

Následne boli niektoré stĺpce a premenné premenované:
'datum' -> 'date'
'prirustkovy_pocet_nakazenych' -> 'New cases'
'jip' -> 'Intensive care unit'
'kyslik' -> 'Oxygen'
'upv' -> 'Artificial lung ventilation'
'ecmo' -> 'ECMO'
'tezky_upv_ecmo' -> 'ALV + ECMO'

Date	Category	Count

Tabulka 6: Štruktúra CSV súboru k dotazu B

6 Dotaz skupiny C

V dotaze skupiny C boly použité kolekcie 'obyvatelia', 'obce', 'ockovanie_zariadenia' a 'ockovaci_mista. Pomocou kolekcie 'obyvatelia' bolo možné získať dáta týkajúce sa počtu obyvateľov jednotlivých okresov a taktiež aj ich prislúchajúcemu vekovému rozloženiu. Následne boli z kolekcie 'obce' získané údaje o počte infikovaných a tie boli vyfiltrované za časové obdobie posledných 4 štvrťrokov (čiže obdobie posledného roka). Kolekcia 'ockovaci_mista' slúžila ako zdroj dát týkajúcich sa počtu vykonaných očkovaní spoločne s kolekciou 'ockovanie_zariadenia', ktorá slúžila na mapovanie vykonaného očkovania v očkovacom zariadení na konkrétny okresný celok. Samotná štruktúra výsledného CSV súboru je nasledovná, pričom zvýraznené atribúty predstavujú atribúty, na ktorých bola vykonaná normalizácia a diskretizácia:

Region	Age [0-15]	Age [15-55]	Age [55+]	Infected	**Vaccination_percentage**	**Vaccination**

Tabulka 7: Štruktúra CSV súboru k dotazu C