



ДМИТРИЙ МУН

Руководитель проекта Smart Data
Ukimet, АО НИТ

Опыт построения Data Lake из 60
государственных баз данных и
обзор Data Driven Decisions на основе
межведомственных данных

**Опыт построения Data Lake
из 60 государственных баз
данных, и обзор Data Driven
Decision cases на основе
межведомственных данных**

Дмитрий Мун

Руководитель проекта Smart Data Ukimet в АО «Национальные Информационные Технологии» (Оператор «электронного правительства» Республики Казахстан).

8-летний опыт работы в области анализа данных в секторе B2G. В настоящее время внедряю концепцию Data Driven Government.

Проблема – ведомственная принадлежность данных

Согласно законодательству, каждое государственное учреждение может анализировать только свои собственные данные

Министерство труда и социальной
защиты населения



Место работы лица, согласно
данным о пенсионных взносах

Министерство
здравоохранения



Данные о вакцинации,
здоровье и болезни
человека

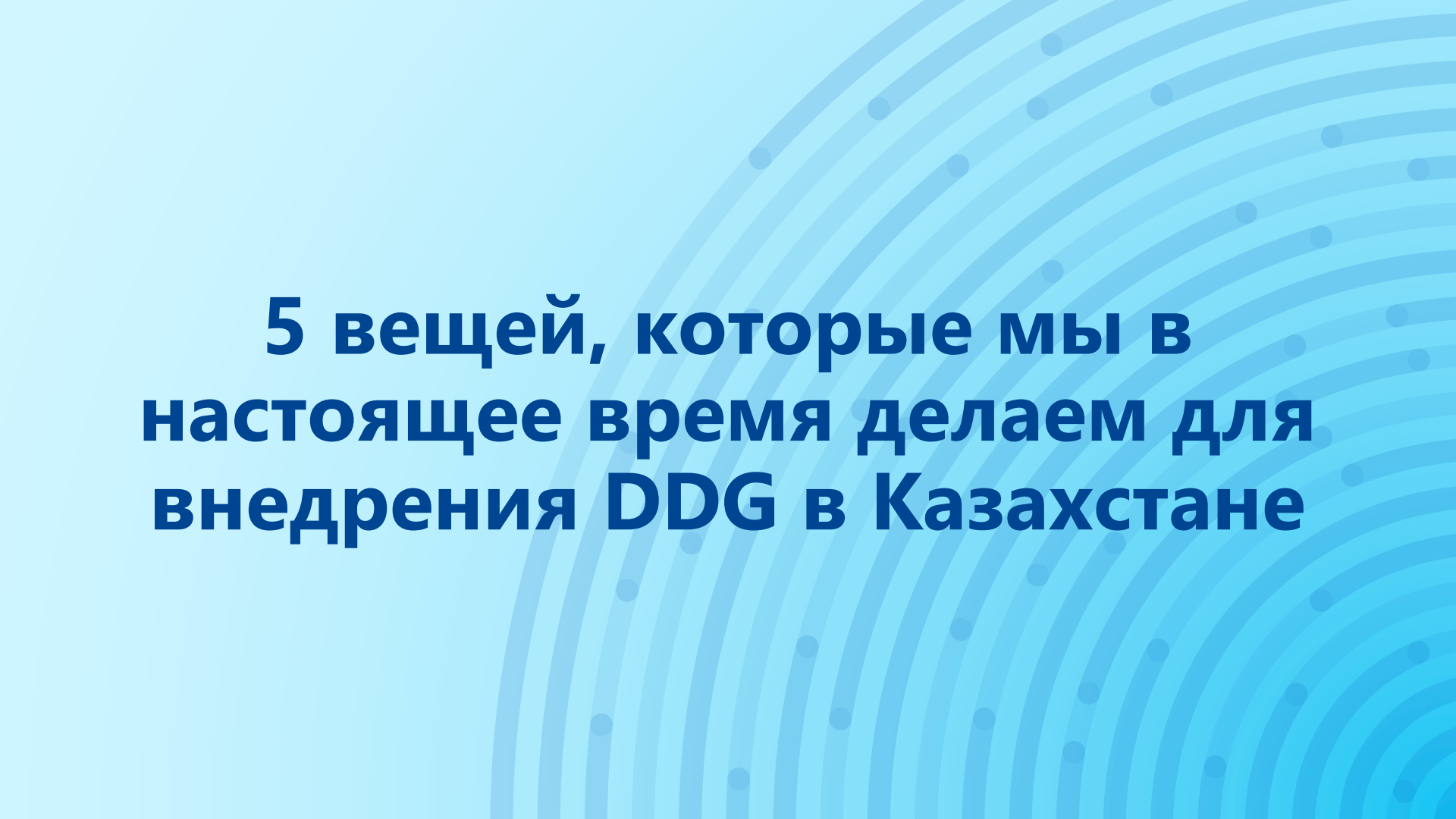
Министерство
юстиции



Адрес проживания
лица

Мы хотим проанализировать все эти данные **в одном месте** для принятия решений, основанных на данных.

Какой процент людей, которые еще не были вакцинированы, в каждом районе и в каждой компании?

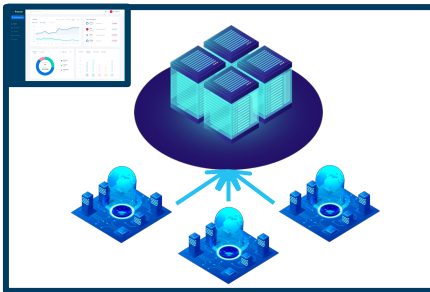
The background features a series of concentric, semi-circular lines in various shades of blue, ranging from light to dark. Small, solid blue dots are scattered across the lines, creating a pattern reminiscent of a fingerprint or a stylized globe. The overall aesthetic is clean, modern, and tech-oriented.

5 вещей, которые мы в настоящее время делаем для внедрения DDG в Казахстане

1. Оператор данных



АО «НИТ» ПРЕДОСТАВЛЯЕТ ГОС ОРГАНАМ АНАЛИТИЧЕСКУЮ ПЛАТФОРМУ (SDU) для ПРИНЯТИЯ ЭФФЕКТИВНЫХ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ ОСНОВАННЫХ НА ДАННЫХ



1. Формирование Озера данных страны «Data Lake»

1. Сбор, обогащение и актуализация данных из всех БД ГО и Квази ГО;
2. Обезличивание персональных данных;
3. Обеспечение удобства работы с данными в Государстве

2. Конструктор Анализа данных

KNIME Analysis Platform

Workflow Coach 10

- Recommended Nodes
- Joiner
- Column Filter
- Row Filter
- Concatenate
- Partitioning
- Number To String
- Concatenate
- Statistics
- Column Rename
- Missing Value
- String Manipulation
- Multi-based Row Filter
- Node Repository
- Manipulation
 - Column
 - Split & Combine
 - Join
 - Concat
 - Database
 - Manipulation
 - Database Joiner

CSV Reader Node 1

CSV Reader Node 2

KNIME Console

```
=====  
*** Welcome to KNIME Analytics Platform v3.7.2 (20200829) ***  
=====  
Copyright by KNIME AG, Zurich, Switzerland  
=====  
Log file is located at: C:\Users\user\knime-workspace\metadata\knime.log  
WARN: MetadataCacheInjector Could not read the intro page: Read timed out  
WARN: CSV Reader 211 No settings available  
WARN: CSV Reader 212 No settings available
```

3. Конструктор визуализации данных

SDU

Upload statement

Type in to search

Activity

Income Outcome Net profit

1 week 1 month 3 months 1 year 5 years

Last transactions

Vendor	License fee	Amount
GitHub	Vendor	-\$1 600.00
Nike	Customer	\$1 600.00
Slack	Vendor	-\$1 600.00
Slack	Slack license 756	-\$1 600.00
GitHub	Vendor	-\$1 600.00

Income Outcome

June

Software 65%

Salaries 25%

Operating 10%

Transactions 23

Amount -\$13,000

TG AR AS NEW SE

2. Правила шифрования



2. Правила шифрования

Шифрование персональных данных в различных государственных органах должно быть одинаковым. Шифрование скрывает личность, но предоставляет всю информацию, собранную из 60 баз данных по этому вопросу



ИИН:
#####1
Адрес:
Нур-Султан



ИИН:
#####1
Место работы:
АО «НИТ»



ИИН:
#####1
Статус вакцинации:
QazVac



ИИН	#####1
Адрес	Ул. Абая
Место работы	АО «НИТ»
Статус вакцинации	QazVac

Этот алгоритм дает нам возможность проводить межведомственную аналитику

3. Упрощенная и простая интеграция

3. Исключение из правил для упрощения интеграции баз данных



```
01 // First, log in
02 loginResult loginResult=null;
03 SoapBindingStub sfdc=null;
04 sfdc = (SoapBindingStub) new SfdcServiceLocator().getSoap();
05 // login
06 loginResult = sfdc.login("username","password");
07
08 // The set up some security related items
09 // Reset the SOAP endpoint to the returned server URL
10 sfdc.setProperty(SoapBindingStub.ENDPOINT_ADDRESS_PROPERTY,loginResult.getServerUrl());
11 // Create a new session header object
12 // Add the session ID returned from the login
13 sessionHeader.showNew SessionHeader();
14 sh.setSessionId(loginResult.getSessionId());
15 sfdc.setHeaders(new SfdcServiceLocator().getServiceName().getNamespaceURI(),
16 "sessionHeader",sh);
17
18 // now that we're logged in, make some calls - retrieve information about the user
19 getUserInfoResult userInfo = sfdc.getUserInfo();
20
21 // create a new account object locally
22 Account account = new Account();
23 account.setAccountNumber("123456789");
24 account.setName("My New Account");
25 account.setBillingCycle("Monthly");
26
27
28 SOBJect[] objects = new SOBJect[2];
29 objects[0] = account;
30
31 // persist the object
32 saveResults() saveResults = sfdc.createSOBJects();
```

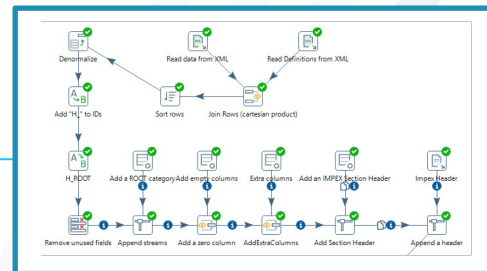


внешние и внутренние шлюзы
электронного правительства

Дорогостоящая и слишком
длительная интеграция
(более 2 месяцев)



ETL (Extract Transform Load)



Прямой доступ к БД ГО по ЕТС ГО



Дешевая и быстрая
интеграция (1 неделя)

4. Инструменты для работы с данными



4. Open Source технологии для работы с данными



Pentaho – Конструктор для сбора, интеграции, преобразования данных



Knime - Конструктор без кода для построения аналитических алгоритмов и машинного обучения



Apache Superset - Конструктор без кода для визуализации



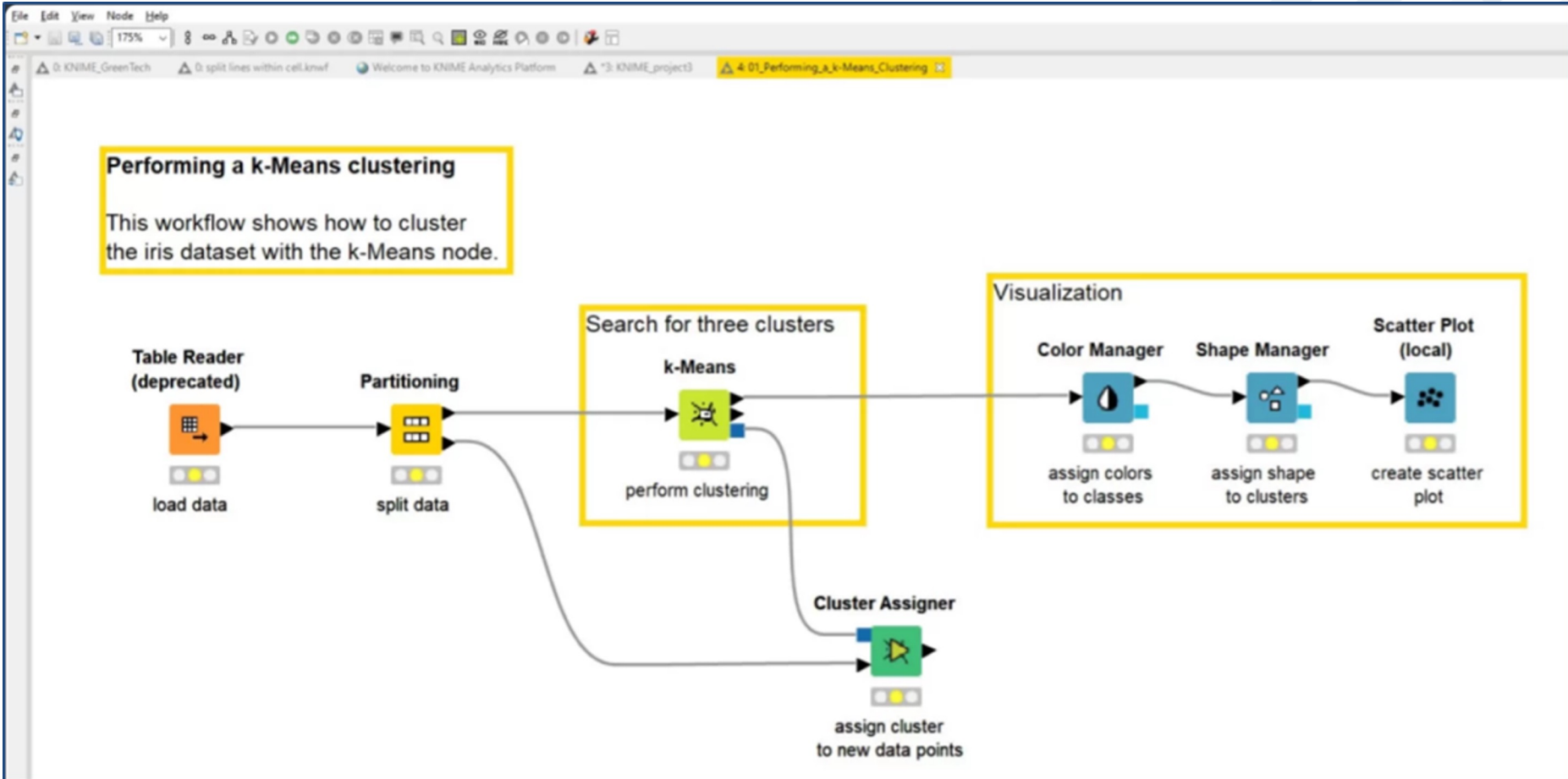
Amundsen – Инструмент для управления метаданными (каталог данных)
(гlossарий, справочник с описанием всех полей и таблиц)

Современный набор технологий, которые уже используются:



Преимущества:

- Никаких лицензий не требуется
- Нет необходимости в дорогостоящей технической поддержке, разработка бесплатных инструментов
- Простые в использовании инструменты для понимания данных
- Гибкая настройка



Performing a k-Means clustering

This workflow shows how to cluster the iris dataset with the k-Means node.

Search for three clusters

k-Means
perform clustering

Visualization

Color Manager
assign colors to classes

Shape Manager
assign shape to clusters

Scatter Plot (local)
create scatter plot

Cluster Assigner
assign cluster to new data points

5. Команда



5. Компетентная команда в АО "НИТ"

Бизнес-аналитика



- Бизнес-аналитика
- Подготовка ТЗ

Data Engineering



- Сбор данных (интеграция)
- Очистка и подготовка данных
- Автоматизация отчетов

Data Analysis



- Прогнозирование и планирование
- Машинное обучение
- Управление рисками

Data Visualization



- Понятные дашборды
- Мониторинг ключевых показателей эффективности

Как лежат базы данных



x60



Переход на подходы Data Mesh



Unified Star Schema





Data Driven Decision кейсы



Поручение Президента РК от 4 марта 2020 года:

*“Завершить подключение всех государственных органов к информационно-аналитической системе **«Smart Data Ukimet»** и обеспечить разработку соответствующих аналитических моделей (кейсов)”.*

Из Закона об информатизации:

16) Оператор ЭП осуществляет сбор, обработку, хранение, передачу данных для осуществления аналитики в целях реализации функций государственными органами в порядке, определенном уполномоченным органом.

Smart Data Ukimet сейчас

Бюро Нац. статистики

Данные комитета статистики

Мин. Труда

Зарботные платы, пенсионные отчисления, иностранные работники, трудоспособное население

Мин. Цифровизации

Все государственные услуги

Мин. Индустрии

Все данные по транспорту

Мин. Образования

Данные по дет.садам, школам, ВУЗам

Мин. Здрав.

Данные по состоянию здоровья населения и лекарственному об-ю

Мин. Фин

Данные по гос.закупкам, гос.имуществу, фин и бюджет. отчетности, таможенной декларации

Мин. Юст

Данные по Физ. Юр. Лицам, имуществу, эталонный банк НПА

Мин. Энергетики

Данные по добыче, переработки, и потребления минерального сырья

Акимат г.Алматы

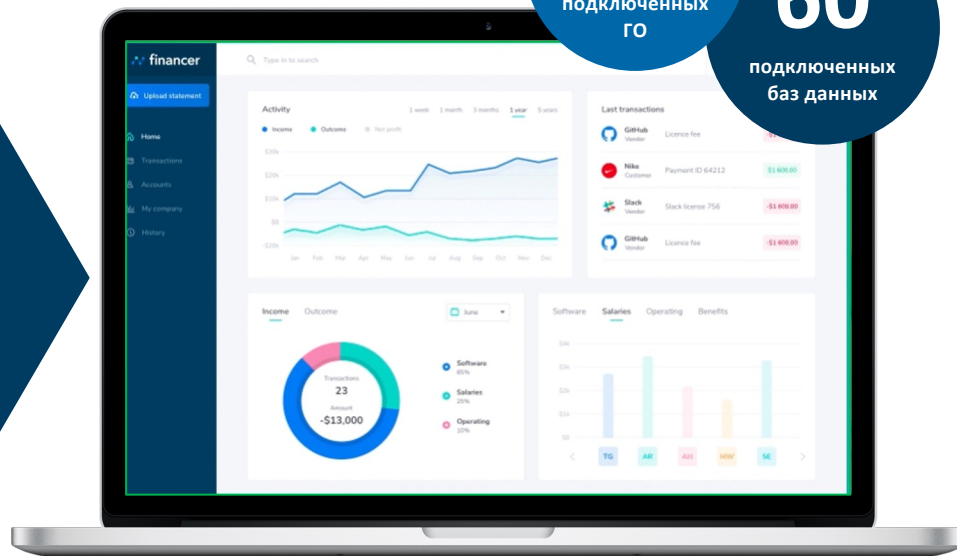
все данные комитета статистики

10

подключенных
ГО

60

подключенных
баз данных



ЦГО

Визуализировано 150
показателей СЭР

Обучено 50
человек

Реализован Сит
центр МСХ

DDD кейс: Соц пособия умершим на 1,5 млрд.

Обновление Ежедневно / Ежемесячно в одной БД

МИО

Визуализация 170
показателей Регионов
в разработке

~ экономия затрат
2 млрд. тг.

57 БД ГО, которые обогащают Data Lake SDU

Гос. орган	Название ИС
МЗ	ИС «Электронный регистр онкологических больных»
	ИС «Электронный регистр диспансерных больных»
	ИС «Система управления ресурсами»
	ИС «Дополнительный компонент подушевого норматива»
	ИСЛО - информационная система «Лекарственное обеспечение»
	Электронный регистр стационарных больных
	Бюро госпитализации
	АПП Банк - Амбулаторно поликлиническая помощь Банк Направлений
	РПН – Регистр прикрепленного населения
РБЖФВ - Регистр беременных и женщин фертильного возраста	
МИИР	Информационно-аналитическая система транспортной базы данных и мониторинга динамики безопасности перевозок
МНЭ	ИС «Метаданные»
	ИС «База данных агрегированных показателей» (БДАП)
	ИС «КЛАСС»
	ИС «Монолист»
МТСЗН	СПП «Регистр субъектов предпринимательства»
	Центральная база данных «Оралман»
	Иностранная рабочая сила
	Организация обработки платежей
	Рынок труда
	Социальная помощь ГДП, АСП, ЖП (Схемы)
	ЦБД (включая схемы MAIN и SOLIDARY)
	Автоматизированная информационная система «Электронное назначение пенсионных выплат и пособий» (Е-макет)
ИС "Централизованный банк данных лиц, имеющих инвалидность"	
Автоматизированная информационная система «Охрана труда и безопасность»	
МЭ РК	Интегрированная информационная система «Единая государственная система управления недропользованием Республики Казахстан
МОН	НОБД - Национальная образовательная база данных
Акимат г. Алматы	АИС "Население"

МФ	Подсистема «Финансовый контроль. Система управления рисками» (emof dm kz)
	Подсистема «Формирование и мониторинг исполнения стратегического плана МФ РК» (spr oper plan)
	Подсистема «Исполнение ГБ в части государственного заимствования и кредитования» (rpt)
	Реестр государственного имущества
	Депозитарий финансовой отчетности
	Интегрированный таможенный тариф
	ЭГЗ – Электронные государственные закупки
	Подсистема «Сбор и консолидация финансовой и бюджетной отчетности» (CONS_DATA, CONS_DICT, CONS_META)
	ЕХД - ИС Единое хранилище данных
	ИХД - ИС «Интегрированное хранилище данных»
МЦРИАП	Статистика
	«Информационная система «Система управления рисками» КГД МФ РК
	Автоматизированная система таможенного и налогового администрирования (ИС АСТАНА-1)
	Адресный регистр
	Интегрированная информационная система центров обслуживания населения
	«Государственная база данных «Е-лицензирование»
	Интернет портал Государственных Органов
	Портал Электронного Правительства
	Мобильное правительство МГОВ
	Открытое правительство
МЮ	Единый контакт-центр
	РН - Регистр недвижимости
	Автоматизированная информационная система органов исполнительного производства
	АИС «Национальный институт интеллектуальной собственности»
	Юридические лица
	Физические лица
	Реестр движимого имущества, 2012 год
	Записи актов гражданского состояния
Эталонный контрольный банк нормативных правовых актов Республики Казахстан в электронном виде	

“Оценка семейного благополучия”

Мы объединили информацию из различных баз данных о каждой семье: семейные связи, различные виды доходов, рабочая сила, недвижимость, автомобиль, уровень образования, здоровье, домашний скот, бизнес, банковские кредиты и т. д.

ID Семьи	Населенный Пункт	Регион	Район	Уровень благосостояния	Уровень среднедушевого дохода (СДД) в семье	Информация о занятости членов семьи	Наличие в собственности семьи недвижимого имущества	Наличие в собственности семьи движимого имущества	Наличие в собственности семьи коммерческого имущества	Наличие сельскохозяйственной техники	Наличие в собственности земельного участка	Наличие в собственности семьи коммерческого недвижимого имущества	Наличие у семьи личного подсобного хозяйства	Учредитель в ЮЛ и ИП	Присутствие в семье четырех и более несовершеннолетних детей и учащейся молодежи до 23 лет
1001602	Без прописки	Атырауская область	Атырау Г.А.	E	Семья без доходов	Семьи с трудоспособными и не работающими членами семьи	Семьи без жилья	В семье нет автотранспорта	В семье нет коммерческого транспорта	В семье отсутствует сельхозтехника	Семьи без земельного участка	В семье нет коммерческой недвижимости	В семье нет личного подсобного хозяйства	В семье нет учредителей ТОО или ИП	В семье нет 4-х и более несовершеннолетних детей и учащейся молодежи до 23 лет
10015860	Алмалинский район	г.Алматы	Алмалинский	E	Уровень среднедушевого дохода в семье – от черты бедности до 1 прожиточного минимума	Семьи без трудоспособных членов семьи	В семье только одно жилье, менее 18 кв. метров на 1 члена семьи	В семье 1 (один) автотранспорт	В семье нет коммерческого транспорта	В семье отсутствует сельхозтехника	В семье есть земельный участок под индивидуальное жилищное строительство	В семье нет коммерческой недвижимости	В семье нет личного подсобного хозяйства	В семье есть ИП	В семье нет 4-х и более несовершеннолетних детей и учащейся молодежи до 23 лет
						Информация о наличии у членов семьи хронических заболеваний	Информация о прикреплении детей в семье к дошкольным и школьным образовательным учреждениям	Информация о прикреплении членов семьи к медицинским учреждениям	Информация об среднем образовании взрослых членов семьи	Информация об профессиональном образовании взрослых членов семьи	Информация об высшем образовании взрослых членов семьи	Информация о участии в ОСМС	Закредитованность семьи	Задолженность семьи по кредитам	
						В семье нет лиц с хроническими заболеваниями	В семье нет детей до 18 лет	Семьи с прикрепленными членами семьи к медицинскому учреждению	Семьи, в которых у взрослых членов семьи нет среднего образования	Семьи, в которых у взрослых членов семьи нет профессионального и среднеспециального образования	Семьи, в которых у взрослых членов семьи нет высшего образования	Семьи в которых нет участников ОСМС	Семья без кредитов	Семья не имеет задолженности по кредитам больше 90 дней и свыше 1000 тенге	
						В семье нет лиц с хроническими заболеваниями	В семье нет членов состоящие на Д-учете	Семьи с прикрепленными членами семьи к медицинскому учреждению	Семьи, в которых у взрослых членов семьи нет среднего образования	Семьи, в которых у взрослых членов семьи нет профессионального и среднеспециального образования	Семьи, в которых у взрослых членов семьи нет высшего образования	Семьи в которых есть участники ОСМС	Семья есть кредиты	Семья не имеет задолженности по кредитам больше 90 дней и свыше 1000 тенге	

“Цифровая карта семьи”



Количество всех семей в РК



6,553,188

СЕМЕЙ



Количество семей по фильтру



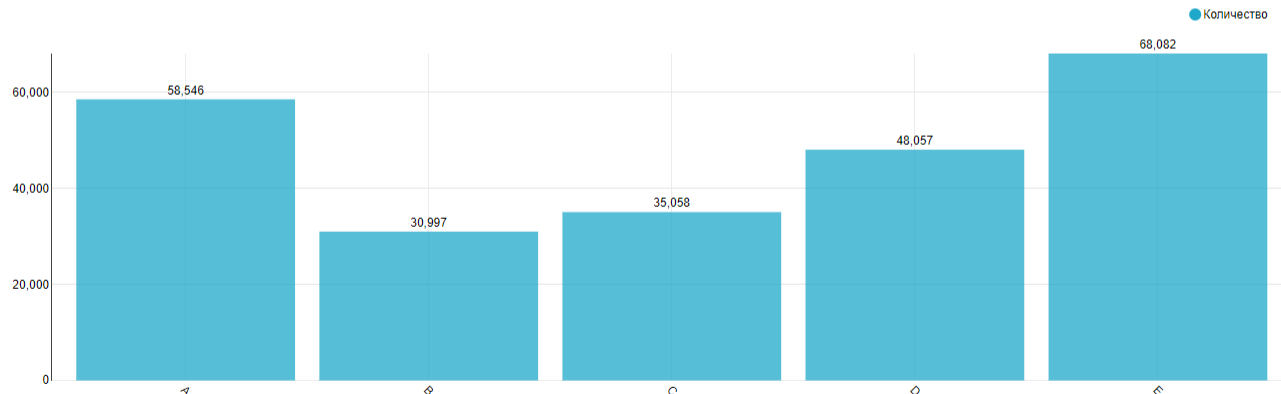
68,082

СЕМЕЙ

Количество семей по уровню благополучия (по сегментам)

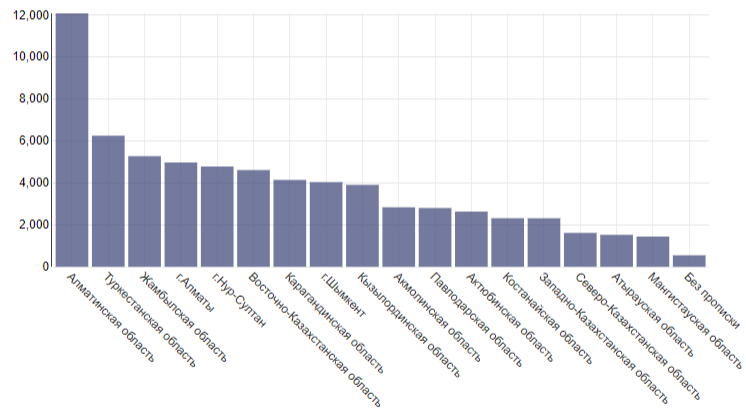


- A** Благополучный (>16)
- B** Удовлетворительный (12-15)
- C** Неблагополучный (9-11)
- D** Кризисный (5-8)
- E** Экстренный (<5)



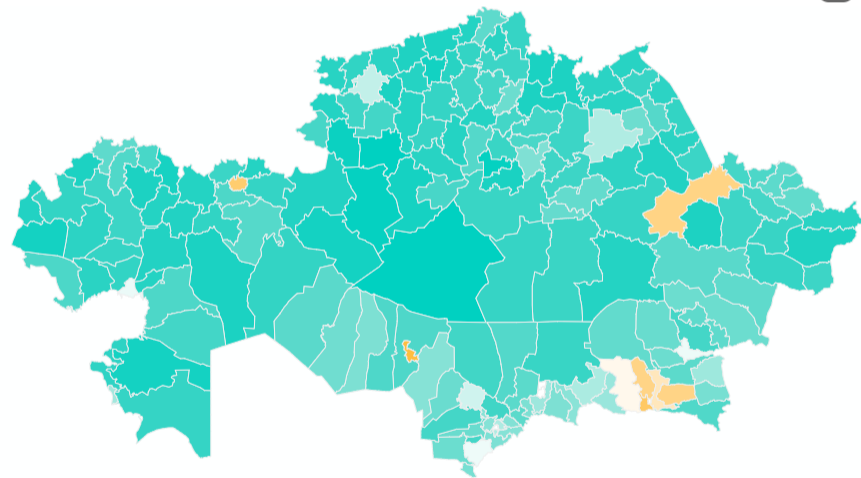
“Цифровая карта семьи”

Количество по регионам



2

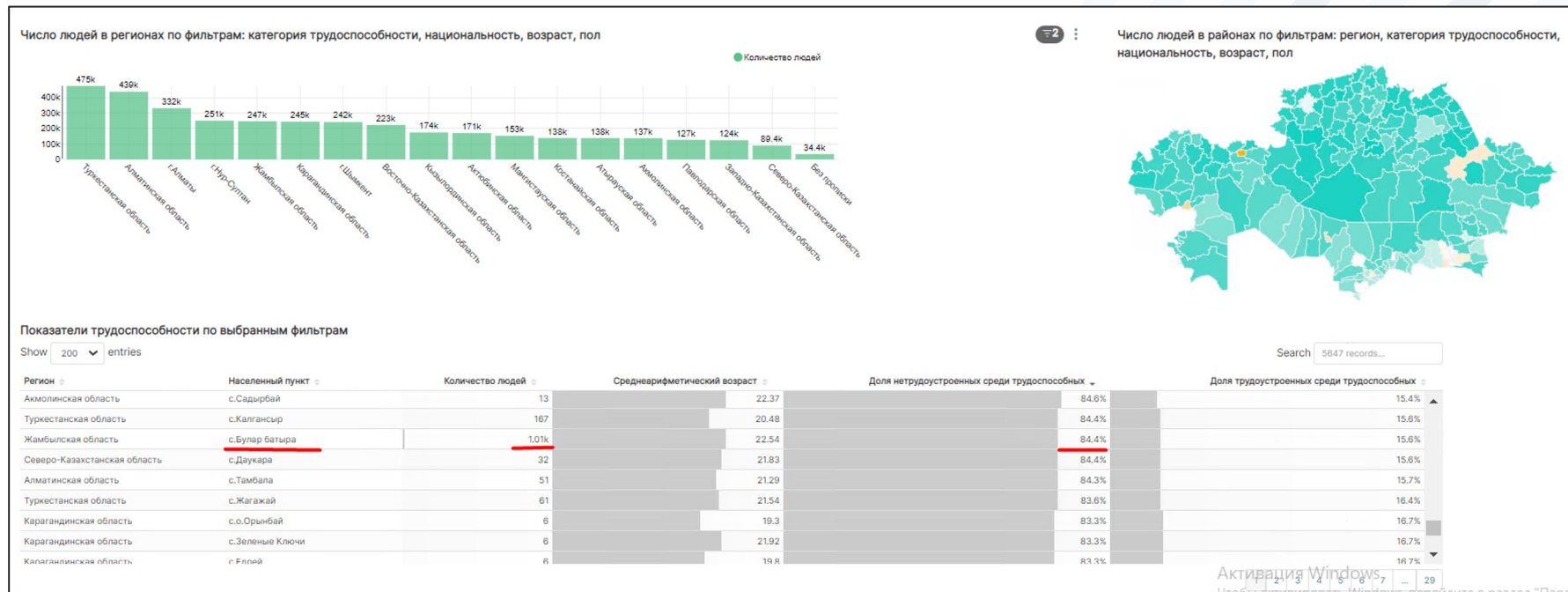
Количество по районам



2

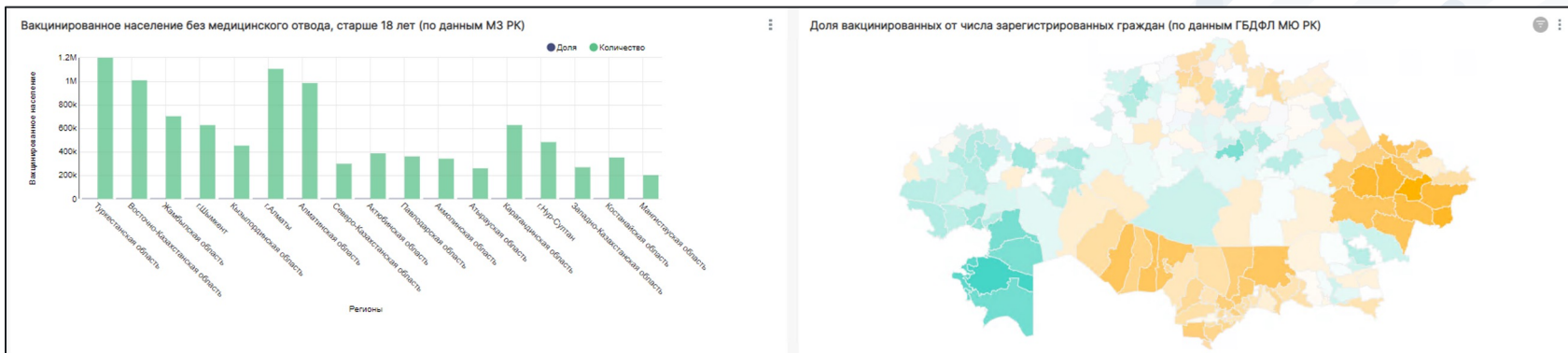
Мониторинг социальных рисков

Здесь мы видим, где необходимо планировать меры государственной поддержки трудоустройства молодежи (до 29 лет).



Где необходимо повысить уровень вакцинации?

данные о месте жительства – МЮ
данные о вакцинации – МЗ
пенсионный взнос -МТСЗН



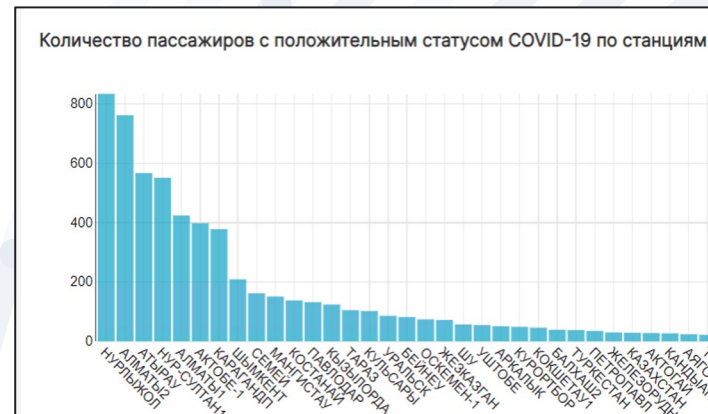
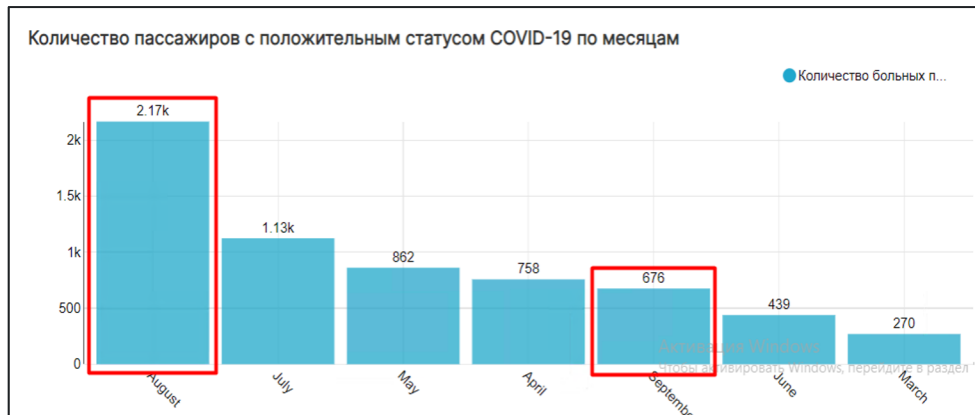
Пассажиры АО “КТЖ”, зараженные Covid-19

Мы использовали:

Базу данных билетов - КТЖ

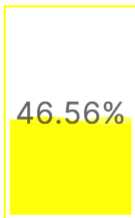
базу данных Covid-тестов – МЗ

Более 4000 железнодорожных пассажиров отправляются в пункт назначения с положительным результатом теста на COVID

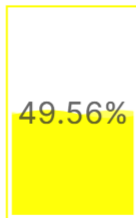


Здесь мы можем увидеть, на каких станциях необходимо установить термометры на входе

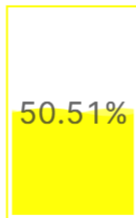
город Шымкент



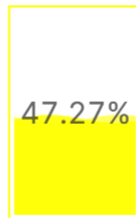
г. Алматы



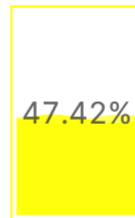
г. Нур-Султан



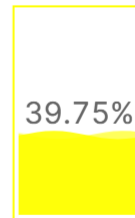
Жамбылская обл.



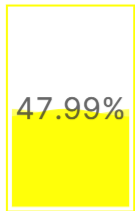
Акмолинская обл.



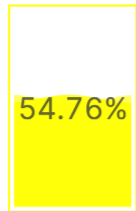
ВКО



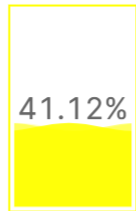
Карагандинская обл.



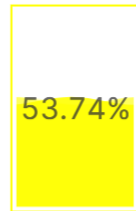
Павлодарская обл.



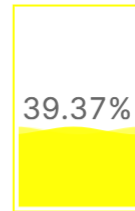
Атырауская обл.



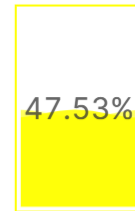
Кызылорд. область



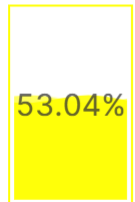
Актюбинская обл.



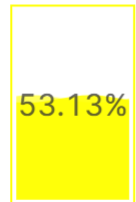
СКО



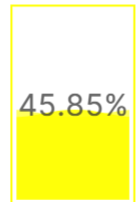
Костанайск. обл.



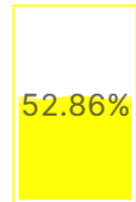
Алматинская обл.



Туркестанская обл.



Мангистауская обл.



Прогноз остатков топлива на АЗС

Фильтр по организациям

SNA.ORGANIZATION

Type or Select [sna.ORGANIZATION]

APPLY

Фильтр по АЗС

SNA.AZS

D-262 x

APPLY

Мониторинг НП по АЗС

Show 50 entries

sna.REGION	sna.AZS	sna.NEFTEPRODUKT_NAME	PROC
Туркестанская область	АЗС "Жаушыкым" ИП "Утегенов Толеген Нурбекович"	Дизельное топливо	25.4%
Туркестанская область	АЗС "Жаушыкым" ИП "Утегенов Толеген Нурбекович"	Бензин	33.4%
город Нур-Султан	АЗС №1 "Бизнес Петролеум"	Не подакцизный товар	99.8%
город Нур-Султан	АЗС №1 "Бизнес Петролеум"	Дизельное топливо	99.7%
город Нур-Султан	АЗС №1 "Бизнес Петролеум"	Бензин	99.1%
Актюбинская область	АЗС №8 ТОО "АЗС Лидер"	Дизельное топливо	41.7%
Актюбинская область	АЗС №8 ТОО "АЗС Лидер"	Дизельное топливо	99.8%
Актюбинская область	АЗС №8 ТОО "АЗС Лидер"	Бензин	94.1%
Актюбинская область	АЗС №8 ТОО "АЗС Лидер"	Бензин	95.7%
Актюбинская область	АЗС №8 ТОО "АЗС Лидер"	Бензин	98.9%
Кызылординская область	АЗС №09 ТОО "TS Онтустик Жанар Май"	Бензин	71.5%
Кызылординская область	АЗС №09 ТОО "TS Онтустик Жанар Май"	Бензин	77.8%
Кызылординская область	АЗС №09 ТОО "TS Онтустик Жанар Май"	Бензин	66.5%
Кызылординская область	АЗС №09 ТОО "TS Онтустик Жанар Май"	Дизельное топливо	71.4%
Карагандинская область	Газпром АЗС 402	Бензин	87.9%
Карагандинская область	Газпром АЗС 402	Бензин	93.6%
город Нур-Султан	АЗС АстанаНефтеСнаб №1	Бензин	98.8%
город Нур-Султан	АЗС АстанаНефтеСнаб №1	Бензин	100.0%
город Нур-Султан	АЗС АстанаНефтеСнаб №1	Бензин	100.0%



Вопросы по Open Data

Айбек +7 775 653 56 60

The background features a series of concentric, semi-circular arcs in various shades of blue, ranging from light to dark. Small, semi-transparent blue dots are scattered across these arcs, creating a sense of depth and movement. The overall composition is clean and modern.

Спасибо!