ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

DATA SCIENTIST

МОДУЛЬ 5. ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА ДЛЯ АНАЛИТИКИ ДАННЫХ И ВИЗУАЛИЗАЦИИ

Тема 1. Обзор современных BI систем (часть 1)

Мишин Александр Юрьевич
к.э.н., доцент департамента бизнес-информатики

Тема 1. Обзор современных BI систем

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Business Intelligence как процесс анализа информации, выработки интуиции и понимания для улучшенного и неформального принятия решений бизнес-пользователями, а также инструменты для извлечения из данных значимой для бизнеса информации

История термина Business Intelligence

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

- 1958 год, учёный Ханс Петер Лун в статье «A Business Intelligence System» в IBM System Journal: обеспечивающие бизнес системы - это системы, поддерживающие разумную деятельность (intelligence system).

- 1989 год, аналитик Gartner Ховард Дреснер: BI - это зонтичный термин для различных технологий, предназначенных для поддержки принятия решений

- Сейчас: BI – это совокупность технологий программного обеспечения и практик, направленных на достижение целей бизнеса путём наилучшего использования имеющихся данных

Определения Business Intelligence

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Business Intelligence - это:

- Процесс анализа информации, выработки интуиции и понимания для улучшенного и неформального принятия решений бизнес-пользователями (процесс получения знания)
- Инструменты, процессы, технологии, методы и средства для:
  - извлечения из данных значимой для бизнеса информации (превращения данных в информацию)
  - извлечения знаний (превращение информации в знания) и представления знаний (ML-модели, бизнес-визуализация)
- превращение знаний в действия бизнеса для получения ценности
- активности конечного пользователя в программных BI-продуктах

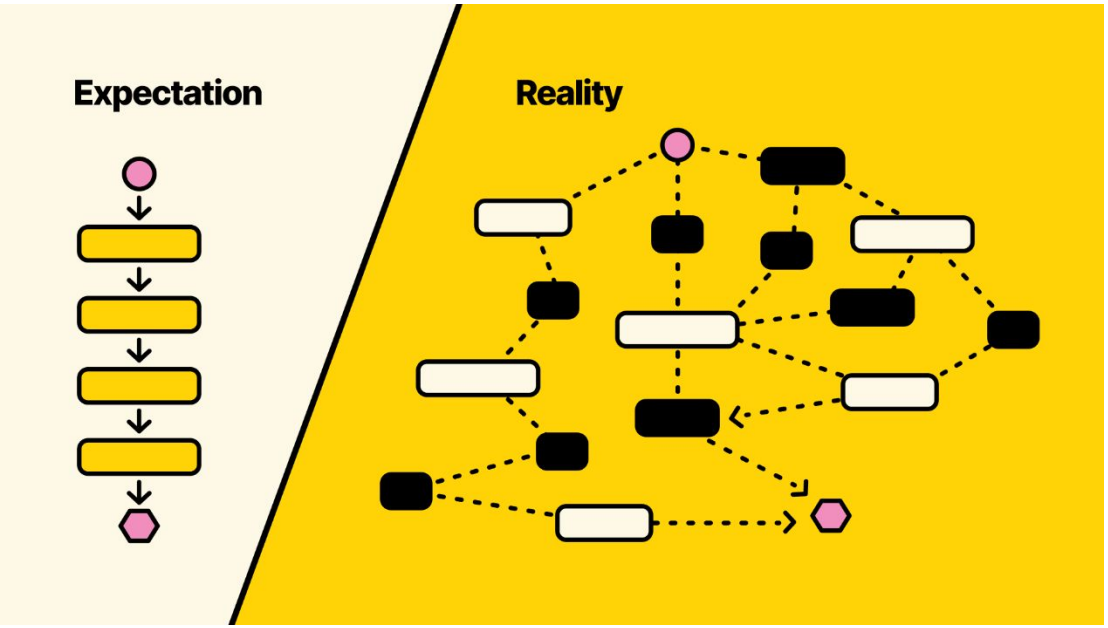
Извлечение знаний о бизнесе на примере технологии Process Mining

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Process Mining:

- группа методов, позволяющих проводить глубокий анализ бизнес-процессов на основе журналов событий
- автор концепции - Вил ван дер Аалст — профессор Эйндховенского технического университета (Голландия) и Квинслендского технического университета (Австралия)
- применяется для оценки многоэтапных процессов со сложной иерархией принятия решений, с большим количеством типичных, повторяющихся операций, которые логируются информационной системой
- позволяет восстановить фактическую, реальную модель массового бизнес-процесса, а не «экспертно-идеальную», регламентированную, игнорирующую многие варианты реализации событий

Влияние BI на бизнес на примере Process Mining



Минимально необходимая структура логов для Process Mining:

- событие
- идентификатор процесса;
- имя действия
- временная метка

Влияние BI на бизнес на примере Process Mining

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Задачи, решаемые в рамках Process Mining

- интеллектуальный анализ процессов в реальном времени
- анализ поведения клиента / сотрудника
- бенчмаркинг процессов
- анализ «что-если»
- расчет стоимости бизнес-процесса и входящих в него операций
- оценка временных и финансовых потерь
- анализ соблюдения требований и регламентов процессов
- выявление «бутылочных горлышек» процессов
- обнаружение избыточных звеньев процессов
- антифрод
- выявление зацикленности в моделях процессов
- моделирование и стресс-тестирование бизнес-процессов
- поиск аномалий в процессах
- оценка степени влияния каждого из факторов на процесс

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Процесс, технологии, методы и средства извлечения и представления знаний

Модель и моделирование

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Анализ данных:

- исследования, связанные с обсчетом многомерной системы данных, имеющей множество параметров;
- формирование представлений о характере явления, описываемого данными;
- средство проверки гипотез и решения задач исследователя
- использует различные математические методы

Термин «модель» (лат. modelium) означает «мера», «способ», «сходство с какой-то вещью».

Модель — объект или описание объекта, системы для замещения (при определенных условиях, предположениях, гипотезах) одной системы (то есть оригинала) другой системой для лучшего изучения оригинала или воспроизведения каких-либо его свойств.

Моделирование — универсальный метод получения, описания и использования знаний. Применяется в любой профессиональной деятельности.

Свойства моделей

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

- упрощенность - отображаются только существенные стороны объекта (модель проста для исследования или воспроизведения);
- конечность - оригинал отображается лишь в конечном числе его отношений, ресурсы моделирования конечны;
- приближенность - действительность отображается моделью грубо или приближенно;
- адекватность - моделируемая система успешно описана;
- целостность - реализует некая система (то есть целое);
- замкнутость - учитывается и отображается замкнутая система необходимых основных гипотез, связей и отношений;
- управляемость - имеется хотя бы один параметр, изменениями которого можно имитировать поведение моделируемой системы в различных условиях.

Виды данных

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

- неструктурированные:
  - произвольные по форме
  - могут включать включающие тексты и графику, мультимедиа (видео, речь, аудио)
- структурированные данные:
  - отражают отдельные факты предметной области
  - упорядоченные и организованные определенным образом с целью обеспечения возможности анализа
- слабоструктурированные данные:
  - для них определены некоторые правила и форматы в общем виде.
  - требуют меньших усилий для преобразования к структурированной форме
  - без процедуры преобразования непригодны для анализа

Способы сбора данных

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

- сбор данных из информационных систем
- получение данных на основе анализа косвенных источников информации
- сбор данных из мобильных устройств, устройств интернета вещей, веб-браузеров
- использование открытых датасетов
- OSINT / CSINT
- Data Sharing
- покупка данных у дата-брокеров или других специализированных компаний
- проведение собственных исследований и мероприятий по сбору данных
- ввод данных вручную на основе экспертных мнений
- другие источники

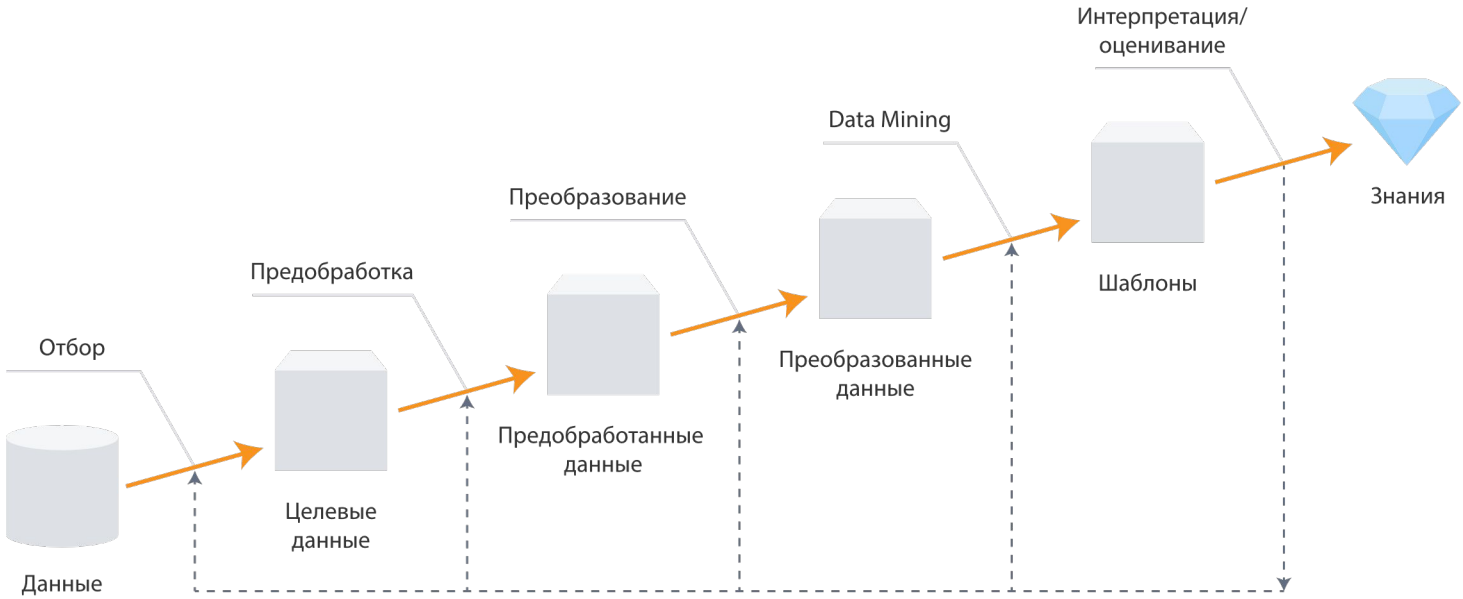
Knowledge Discovery in Databases

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Технология KDD (Knowledge Discovery in Databases):

- возникла в 1989 году
- основоположниками считаются Пятецкий-Шапиро и Усама Файад (Usama Fayyad)
- технология извлечение данных из баз данных
- не содержит описания конкретного алгоритма или математического аппарата
- описание последовательности действий, необходимых для извлечения знаний

Этапы KDD



Этапы KDD:

1. Выборка данных (используются методы фильтрации, запросы, экспертиза и экспертные данные)
2. Очистка
3. Трансформация – для того чтобы представить информацию в определенном виде. Например для прогнозирования временных рядов ряд преобразуется в скользящее окно. К трансформации относится квантование сортировка группировка и другие
4. Data Mining
5. Интерпретация

Современная концепция моделирования

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Базируется на информационном подходе:

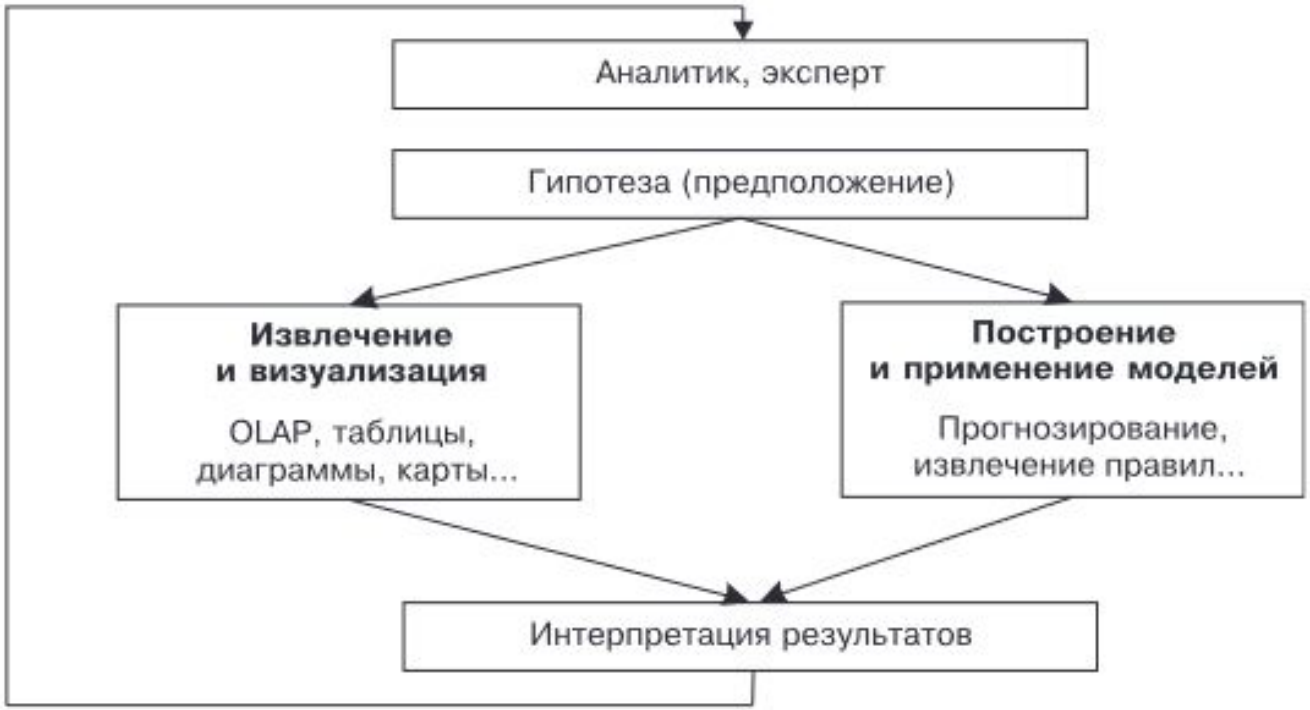
Модели:
- строятся от данных
- учитывают специфику моделируемого объекта
- требуют тщательного подхода к качеству исходных данных (консолидация данных, их очистка и обогащение)

Консолидация – добавление новых данных в датасет из других источников

Обогащение данных — добавление новых признаков в датасет, могущих повысить качество модели

Очистка – повышение качества данных (исправление ошибок, дополнение, стандартизация и устранение дубликатов данных)

Высокоуровневый алгоритм анализа данных



Н. Паклин, В. Орешков. Бизнес-аналитика: от данных к знаниям. Учебное пособие. 2-е издание, исправленное. Изд-во Питер, 2013

Высокоуровневый алгоритм моделирования

Любая модель (ML, Data Mining) теряет со временем свою эффективность

Н. Паклин, В. Орешков. Бизнес-аналитика: от данных к знаниям. Учебное пособие. 2-е издание, исправленное. Изд-во Питер, 2013

Основные роли в аналитической команде

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Роли в анализе данных:

- Эксперт в предметной области – ограничения для моделей, интерпретация и оценка результатов моделирования, формулировка гипотез
- Data Engineer (инженер данных) - проектирование, поддержка и оркестрация систем хранения данных (оркестрация - координирование работы сложных систем)
- Data Analyst (аналитик данных) – ETL-процессы, EDA, формулирование и проверка гипотез
- Data Scientist – углубленное понимание процесса моделирования, лучший подбор моделей и архитектуры нейронных сетей
- ML-разработчик, ML-инженер – создание промышленного ML-решения

Эффективность моделей

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Почему модель перестает работать:

- меняются взаимосвязи между факторами предметной области;
- меняется характер влияния факторов на модель
- появляются новые факторы (риски)
- модель узнается рынком и перестает работать
- проблемы с качеством данных
- недостаточно данных

Что делать:

- Индуктивное смещение
- оптимизация данных;
- управление качеством данных
- разработка и внедрение ML-платформ
- анализ предметной области
- больше сырых данных

Индуктивное смещение

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

- индуктивное смещение алгоритма машинного обучения – это набор предположений, определяющих критерии выбора модели алгоритмом машинного обучения
- есть два типа индуктивного смещения:
  - ограничивающее (restriction bias) – ограничивают набор моделей, которые алгоритмы будут использовать в процессе обучения
  - предпочтение (preference bias) – вынуждает алгоритмы обучения отдавать предпочтение определенным моделям в процессе обучения
- нет способа узнать, какое индуктивное смещение лучше всего подойдет для конкретной задачи
- пере/недообученные модели плохо обобщаются и не могут быть использованы для экземпляров, выходящих за пределы выборки

Недообучение и переобучение

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Две проблемы, ведущие к неправильному индуктивному смещению:

- Недообучение (underfitting) – модель прогнозирования слишком упрощена, чтобы представить связь между описательными и целевым признаком в обучающей выборке
- Переобучение (overfitting) – модель прогнозирования настолько сложна, что слишком точно приближает обучающую выборку и становится чувствительной к шуму в данных.

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

BI как совокупность технологий, программного обеспечения и практик, направленных на достижение целей бизнеса путём наилучшего использования имеющихся данных

BI-технологии

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

BI-технологии:

- *Ad hoc анализ*
- *ETL, технологии консолидации и трансформации данных*
- *Технологии управления качеством данных;*
- Технологии визуализации
- Технологии анализа данных, EDA, отчёты
- Технологии организации, хранения и доступа к данным (хранилища данных (Data Warehouse), витрины данных (DataMarts), технологии СУБД);
- OLAP (Online analytical processing)
- OLTP (Online transactional processing)
- HOLAP, ROLAP, MOLAP
- BPM-технологии (Business Performance Management)
- Data Mining
- Некоторые ML-технологии

Кроме того:

- Облачные технологии
- Технологии интеграции
- Мобильные технологии
- Технологии no-code и low-code разработки

BI-функционал есть во всех информационных системах.

BI-системы

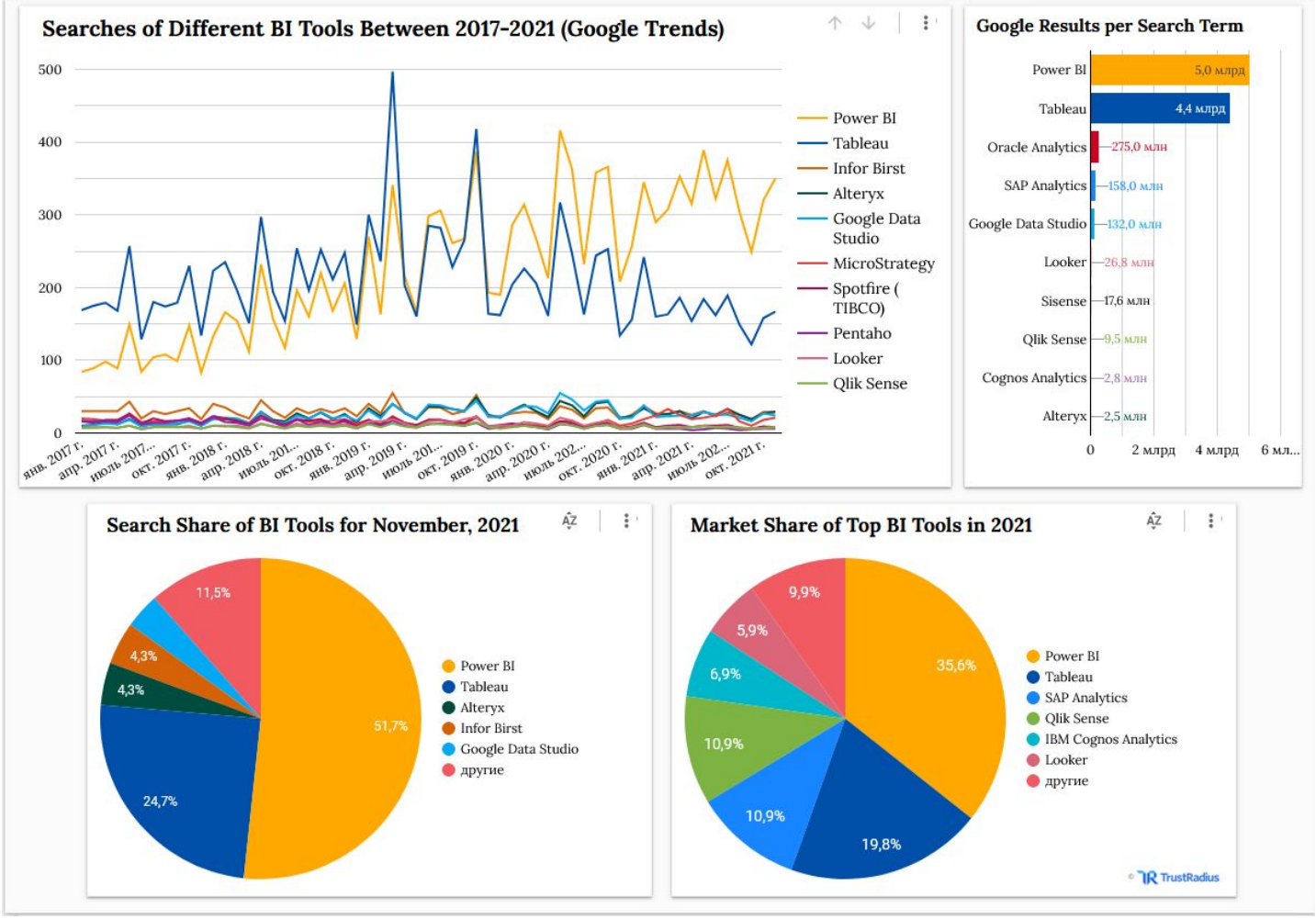
ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Континуум BI-решений:

- СУБД;
- BI-платформы (средства разработки BI-приложений для визуализации);
- корпоративные BI-наборы приложений;
- BI-модули ERP-систем;
- системы для анализа данных, DataMining и ML;
- отдельные BI-сервисы по BI-функциям;
- прочее.

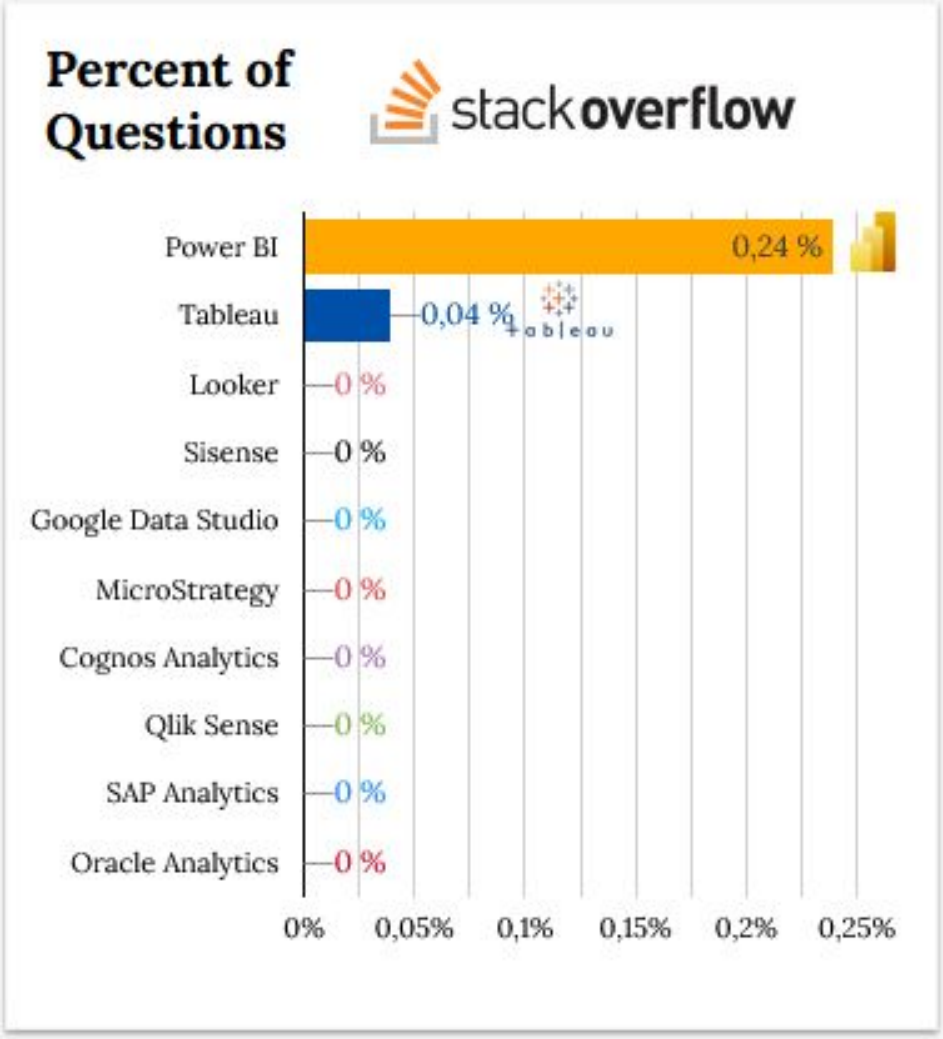
Квадрант Гартнера по BI-системам

Поисковые запросы по BI-системам



https://datastudio.google.com/reporting/03b3aed8-42e1-4423-bea9-e37b8e4e0f86/page/p_tuk4r2j3qc

Поисковые запросы по BI-системам



https://datastudio.google.com/reporting/03b3aed8-42e1-4423-bea9-e37b8e4e0f86/page/CV7iC

Сравнение функционала BI-систем

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Categories	Power BI	Sisense	Looker	Qlik Sense	Cognos Analytics	DOMO	Tableau	Amazon QuickSight	Oracle Analytics	MicroStrategy Analytics	Google Data Studio	Thought Spot	SAP Analytics	Spotfire
Company	Microsoft	Sisense	Google	Qlik	IBM	DOMO	Salesforce	Amazon	Oracle	MicroStrategy	Google	ThoughtSpot	SAP	TIBCO
Release year	2015	2004	2013	2013	2005	2011*	2003	2016	2012*	1989	2016	2012*	2015	1996
Price per creator (pm)	Pro: $9.99, Premium: $20	$160*	est. $5,000+ for 12 licenses	$30.00	$10-40	$83*	$70	$18.00	$40.00	$600/user or $300,000/CPU core*	$0	N/A*	$36	$125 + $9,000-50,000 for licensing*
Price per viewer (pm)	$9.99	$160*	~$15 per user after the $5,000pm fee*	$30.00	$10-40	$83*	$15	$5.00	$40.00	N/A*	$0	N/A*	$36	$25
On-prem cost (per month)	$4995*	?	$5,000+ for 12 licenses	annual $25,000 +10%*	$15-70 per user	?	same cost	?	?	?	n/a	$6,250*	?	?
Deployment Model	Cloud (poor on-prem)*	On-prem or Cloud	On-prem or Cloud*	On-prem or Cloud	On-prem or Cloud	Cloud	On-prem or Cloud	Cloud	On-prem or Cloud	On-prem or Cloud	Cloud	On-prem or Cloud	On-prem or Cloud	On-prem or Cloud
Development Environment	Desktop	Browser or Desktop	Browser	Browser	Browser or Desktop	Browser*	Desktop	Browser	Browser*	Desktop	Browser	Implementation	Browser*	Browser or Desktop
OS Agnostic	Windows only	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Windows: Full featured on Windows/ Mac: Limited, via web browser
Cloud Agnostic	Azure lock-in for hosting and some features	✓	✓	✓	✓*	✓	✓	✓*	✓*	✓	n/a	✓	✓*	✓
Database Agnostic	✓	Connector	Connector	✓	Connector	Connector	Connector	Connector	✓	Connector	✗	✓	Connector	✓
Preferred Data Model	Star-schema	Star-schema	Star-schema*	Snowflake / Star-schema	Star-schema/ Snowflake/ OLAP	Star-schema	Flat	Star-schema/ Snowflake*	Star-schema/ Snowflake*	Star-schema/ Snowflake*	Star-schema/ Snowflake	Star-schema*	Star-schema/ Snowflake*	Star-schema/ Snowflake
Full-Featured Free Version	✓	Free trial	✗	✗	Free trial	Free trial	Separate tool	Free trial	Free trial	✓	✓	✗	Free trial	Free trial
Share with anyone on the Internet	✓*	✓	✓	✗	✓	✗*	Separate tool	✗*	✓*	✗*	✓	✗*	✗*	✓

https://datastudio.google.com/reporting/03b3aed8-42e1-4423-bea9-e37b8e4e0f86/page/p_2qih72j3qc

Сравнение функционала BI-систем

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Categories	Power BI	Sisense	Looker	Qlik Sense	Cognos Analytics	DOMO	Tableau	Amazon QuickSight	Oracle Analytics	MicroStrategy Analytics	Google Data Studio	Thought Spot	SAP Analytics	Spotfire
R and Python Support	Only in the personal version	✓	✓	✓	✓	✓	✓	✗	✓	✓*	✗	R	R*	✓
Open-source Custom Visuals	✓	✓	✓	✓	✓	✓	✓*	✓	✓	✓	✓	✗	✗*	✓
Dynamic Cross-Filtering	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓*	✓
Search Analytics with NLP	✓	✓*	✗	✓	✓	✓*	✓	✓	✓	✓	✗	✓	✓	✓
Data Preparation Tools	✓	✓	✓*	✓	✓	✓	Separate tool	✓	✓	✓*	Separate tool	✓	✓	✓
Data Modelling Tools	✓	✓	✓	✓	✓	✓*	Separate tool	✓	✓	✓*	✗	✓	✓	✓
Mobile App	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Embed Analytics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Built-in Row Level Security	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Simultaneous Development by Multiple Parties	✗	?	✗*	?	✗	?	✗*	?	?	?	✓	?	?	✗
Security Protocols/Certs Support	?	?	?	?	?	?	?	?	?	?	?	?	?	?
User-friendliness (developer)	High	?	High	?	High*	High	High; after slightly higher learning curve	?	?	?	High	?	?	High; after slightly higher learning curve
Dashboards appeal	High	?	High	?	?	High	High	?	?	?	High	?	?	High

https://datastudio.google.com/reporting/03b3aed8-42e1-4423-bea9-e37b8e4e0f86/page/p_2qih72j3qc

Ad hoc анализ

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Ad hoc отчёты:

- отчёты Ad hoc не являются стандартными для организации
- генерируются с помощью нерегламентированных запросов (ad hoc query) к базе, хранилищу или витрине данных
- архитектура данных организации не оптимизирована для их быстрого выполнения

ETL

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Extract, Transform, Load (ETL):

- представляет собой процесс переноса первичных данных из различных источников в аналитическое приложение или поддерживающее его ХД
- является составной частью этапа консолидации в анализе данных
- ETL-операции происходят во временных таблицах (промежуточной области)
- должен учитывать все особенности используемой в хранилище модели
- содержит три укрупненных этапа:
  - извлекает данные из источников
  - преобразуют их в формат, поддерживаемый системой хранения и обработки
  - загружает в нее преобразованную информацию

Проблемы качества данных

- Качество данных — совокупность свойств и характеристик данных, определяющих степень их пригодности для анализа.
- Оценка качества анализируемых данных вместе с их очисткой может занимать до 80 % времени всего процесса анализа

https://habr.com/ru/post/548220/

Схема данных

Схема базы данных включает данные обо всех объектах в базе данных:

- поля;
- таблицы;
- отношения;

а также:

- триггеры;
- представления;
- индексы.

Проблемы качества данных

1. Проблемы с признаками (значениями переменных, столбцами в табличном представлении датасета)

- недопустимые значения, которые лежат вне нужного диапазона
- отсутствующие значения, которые не введены, бессмысленны или не определены
- орфографические ошибки
- многозначность (например, «БД» может быть сокращением для словосочетания «большие данные» или «база данных»)
- перестановка слов, обычно встречается в текстовых полях свободного формата
- вложенные значения – несколько значений в одном признаке, например, в поле свободного формата

2. Проблемы с записями – объектами, которые являются строками датасета и описываются значениями признаков

- нарушение уникальности
- дублирование записей
- противоречивость записей (один и тот же объект описан различными значениями признаков)
- неверные ссылки (нарушение логических связей между признаками)

Методы очистки данных для проблемных случаев

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Проблема	Решение
Несколько противоречивых записей	Удалить все
	Удалить все, кроме последнего
	Вычислить вероятность появления каждого из противоречивых значений и выбрать наиболее вероятное
Пропуск в данных	Аппроксимация (для временного ряда – в окрестности аппроксимируемой точки)
	Определение наиболее правдоподобного значения (задействованы все данные)
	Метод индикатора
Аномальное значение	Аномальные значения удаляется
	Аномальные данные заменяются на ближайшие граничные значения
Шум	Спектральный анализ
	Авторегрессионные методы
Орфографические ошибки в данных	Замена, изменение формата, применение тезаурусов

Инструменты очистки данных

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Написание собственными силами кода, исправляющего ошибки в данных на одном из следующих языков:

- Python
- R
- VBA

Использование инструментов автоматизированной очистки данных, встроенных в БД:
Microsoft SQL Server data Quality Services;

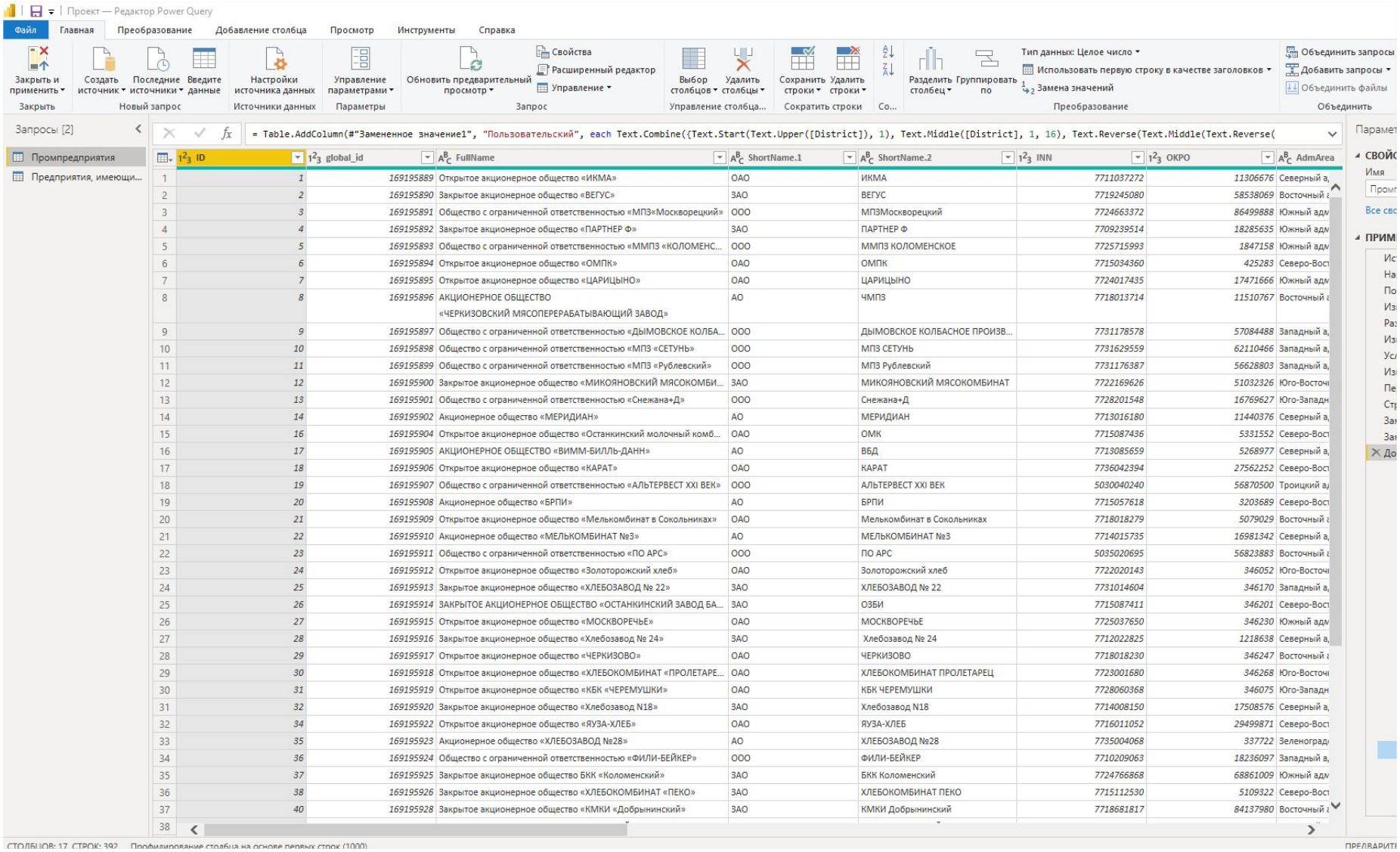
- Hive;
- Azure;
- IBM InfoSphere Information Server for Data Quality
- SAP Data Quality Management
- AWS Glue
- и т.д.

Использование пакетов анализа данных:

- Microsoft Power BI
- IBM SPSS
- SAS® Data Quality
- Loginom;
- и др.

Технологии очистки данных на примере Microsoft Power BI

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ



	ID	global_id	FullName	ShortName.1	ShortName.2	INN	OKPO	AdmArea
1	1	169195889	Открытое акционерное общество «ИКМА»	ОАО	ИКМА	7711037272	11306676	Северный а,
2	2	169195890	Закрытое акционерное общество «ВЕГУС»	ЗАО	ВЕГУС	7719245080	58538069	Восточный а
3	3	169195891	Общество с ограниченной ответственностью «МПЗ«Москворецкий»	ООО	МПЗМосворецкий	7724663372	86499888	Южный адм
4	4	169195892	Закрытое акционерное общество «ПАРТНЕР Ф»	ЗАО	ПАРТНЕР Ф	7709239514	18285635	Южный адм
5	5	169195893	Общество с ограниченной ответственностью «ММПЗ «КОЛОМЕНС...	ООО	ММПЗ КОЛОМЕНСКОЕ	7725715993	1847158	Южный адм
6	6	169195894	Открытое акционерное общество «ОМПК»	ОАО	ОМПК	7715034360	425283	Северо-Вост
7	7	169195895	Открытое акционерное общество «ЦАРИЦЫНО»	ОАО	ЦАРИЦЫНО	7724017435	17471666	Южный адм
8	8	169195896	АКЦИОНЕРНОЕ ОБЩЕСТВО «ЧЕРКИЗОВСКИЙ МЯСОПЕРЕРАБАТЫВАЮЩИЙ ЗАВОД»	АО	ЧМПЗ	7718013714	11510767	Восточный а
9	9	169195897	Общество с ограниченной ответственностью «ДЫМОВСКОЕ КОЛБА...	ООО	ДЫМОВСКОЕ КОЛБАСНОЕ ПРОИЗВ...	7731178578	57084488	Западный а,
10	10	169195898	Общество с ограниченной ответственностью «МПЗ «СЕТУНЬ»	ООО	МПЗ СЕТУНЬ	7731629559	62110466	Западный а,
11	11	169195899	Общество с ограниченной ответственностью «МПЗ «Рублевский»	ООО	МПЗ Рублевский	7731176387	56628803	Западный а,
12	12	169195900	Закрытое акционерное общество «МИКОЯНОВСКИЙ МЯСОКОМБИ...	ЗАО	МИКОЯНОВСКИЙ МЯСОКОМБИНАТ	7722169626	51032326	Юго-Восточ
13	13	169195901	Общество с ограниченной ответственностью «Снежана+Д»	ООО	Снежана+Д	7728201548	16769627	Юго-Западн
14	14	169195902	Акционерное общество «МЕРИДИАН»	АО	МЕРИДИАН	7713016180	11440376	Северный а,
15	16	169195904	Открытое акционерное общество «Останкинский молочный комб...	ОАО	ОМК	7715087436	5331552	Северо-Вост
16	17	169195905	АКЦИОНЕРНОЕ ОБЩЕСТВО «ВИММ-БИЛЛЬ-ДАНН»	АО	ВБД	7713085659	5268977	Северный а,
17	18	169195906	Открытое акционерное общество «КАРАТ»	ОАО	КАРАТ	7736042394	27562252	Северо-Вост
18	19	169195907	Общество с ограниченной ответственностью «АЛЬТЕРВЕСТ XXI ВЕК»	ООО	АЛЬТЕРВЕСТ XXI ВЕК	5030040240	56870500	Троицкий а/
19	20	169195908	Акционерное общество «БРПИ»	АО	БРПИ	7715057618	3203689	Северо-Вост
20	21	169195909	Открытое акционерное общество «Мелькомбинат в Сокольниках»	ОАО	Мелькомбинат в Сокольниках	7718018279	5079029	Восточный а
21	22	169195910	Акционерное общество «МЕЛЬКОМБИНАТ №3»	АО	МЕЛЬКОМБИНАТ №3	7714015735	16981342	Северный а,
22	23	169195911	Общество с ограниченной ответственностью «ПО АРС»	ООО	ПО АРС	5035020695	56823883	Восточный а
23	24	169195912	Открытое акционерное общество «Золоторожский хлеб»	ОАО	Золоторожский хлеб	7722020143	346052	Юго-Восточ
24	25	169195913	Закрытое акционерное общество «ХЛЕБОЗАВОД № 22»	ЗАО	ХЛЕБОЗАВОД № 22	7731014604	346170	Западный а,
25	26	169195914	ЗАКРЫТОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «ОСТАНКИНСКИЙ ЗАВОД БА...	ЗАО	ОЗБИ	7715087411	346201	Северо-Вост
26	27	169195915	Открытое акционерное общество «МОСКВОРЕЧЬЕ»	ОАО	МОСКВОРЕЧЬЕ	7725037650	346230	Южный адм
27	28	169195916	Закрытое акционерное общество «Хлебозавод № 24»	ЗАО	Хлебозавод № 24	7712022825	1218638	Северный а,
28	29	169195917	Открытое акционерное общество «ЧЕРКИЗОВО»	ОАО	ЧЕРКИЗОВО	7718018230	346247	Восточный а
29	30	169195918	Открытое акционерное общество «ХЛЕБОКОМБИНАТ «ПРОЛЕТАРЕ...	ОАО	ХЛЕБОКОМБИНАТ ПРОЛЕТАРЕЦ	7723001680	346268	Юго-Восточ
30	31	169195919	Открытое акционерное общество «КБК «ЧЕРЕМУШКИ»	ОАО	КБК ЧЕРЕМУШКИ	7728060368	346075	Юго-Западн
31	32	169195920	Закрытое акционерное общество «Хлебозавод N18»	ЗАО	Хлебозавод N18	7714008150	17508576	Северный а,
32	34	169195922	Открытое акционерное общество «ЯУЗА-ХЛЕБ»	ОАО	ЯУЗА-ХЛЕБ	7716011052	29499871	Северо-Вост
33	35	169195923	Акционерное общество «ХЛЕБОЗАВОД №28»	АО	ХЛЕБОЗАВОД №28	7735004068	337722	Зеленограде
34	36	169195924	Общество с ограниченной ответственностью «ФИЛИ-БЕЙКЕР»	ООО	ФИЛИ-БЕЙКЕР	7710209063	18236097	Западный а,
35	37	169195925	Закрытое акционерное общество ББК «Коломенский»	ЗАО	ББК Коломенский	7724766868	68861009	Южный адм
36	38	169195926	Закрытое акционерное общество «ХЛЕБОКОМБИНАТ «ПЕКО»	ЗАО	ХЛЕБОКОМБИНАТ ПЕКО	7715112530	5109322	Северо-Вост
37	40	169195928	Закрытое акционерное общество «КМКИ «Добрынинский»	ЗАО	КМКИ Добрынинский	7718681817	84137980	Восточный а
38								

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Технологии OLAP и Data Mining

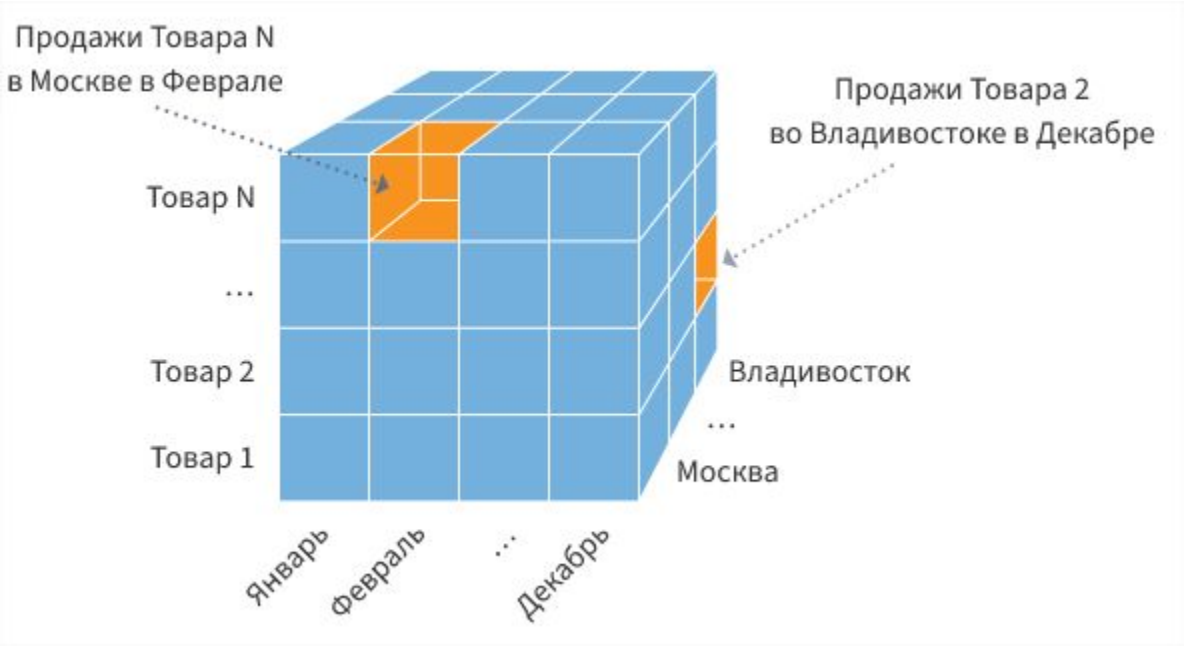
OLTP

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

OLTP - Online Transaction Processing

- технически это сервер реляционной БД и прилагаемые технологии;
- прилагается к любой комплексной информационной системе для бизнеса (ERP, CRM, АБС, SRM и т.д.);
- быстро выполняет простые операции (вставка, обновление или удаление элемента);
- очень медленно выполняет сложные запросы

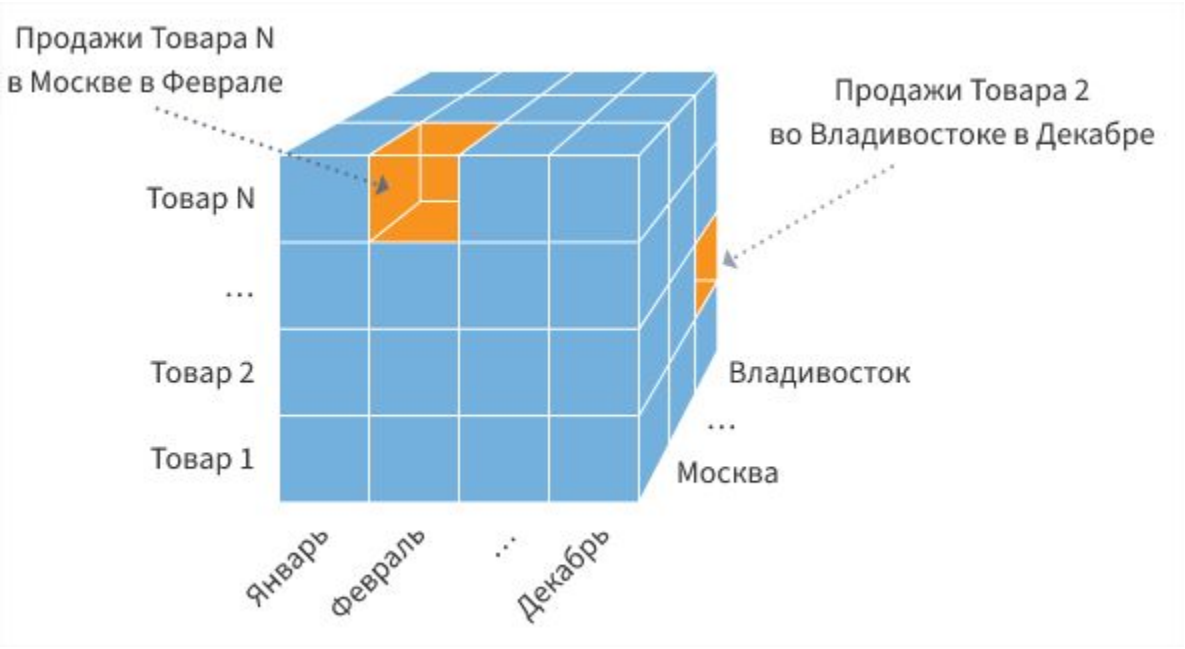
OLAP



OLAP - Online Analytical Processing:

- совокупность практик моделирования данных и баз данных
- технология хранения и обработки многомерных данных

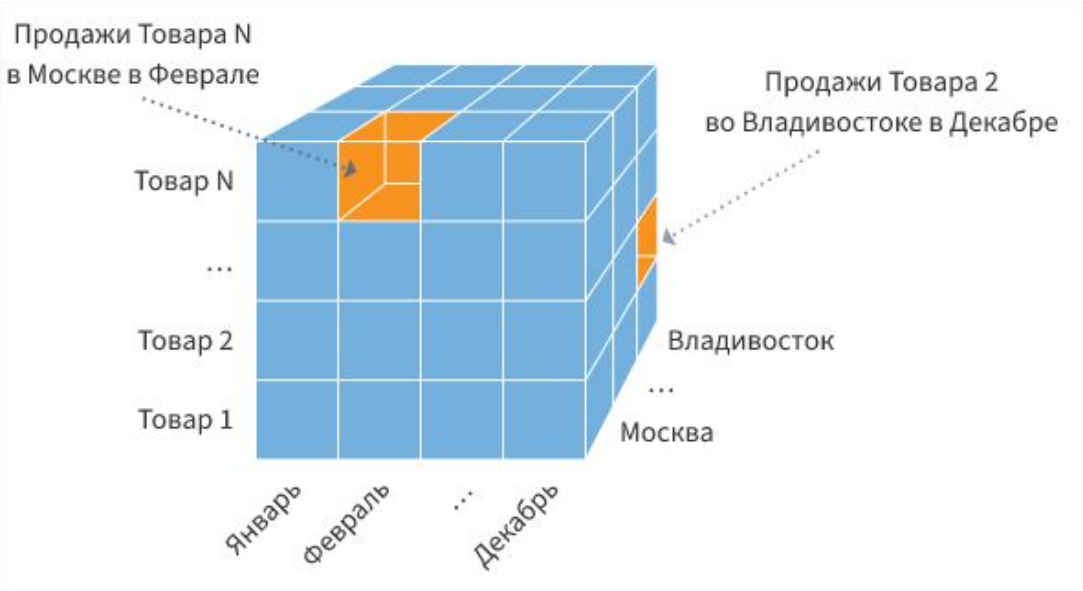
OLAP



OLAP - Online Analytical Processing:

- способен объединять классические таблицы в таблицы таблиц (OLAP-кубы)
- создание конкретных аналитических решений
- позволяет получать сложные аналитические отчёты в реальном времени

OLAP



Компоненты OLAP - Online Analytical Processing:

- база данных (БД)
- OLAP сервер (обработка многомерных структур данных и связь между БД и пользователями систем)
- приложения для работы пользователей (формирование запросов и визуализация полученных ответов)

Понятие Data Mining

Data Mining:

- методология и процесс обнаружения в больших массивах данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных областях
- автор концепции - Пятецкий-Шапиро
- данный термин впервые озвучен в 1989 году на одном из семинаров, посвященных технологиям поиска знаний в базах данных, проводимых в рамках Международной конференции по искусственному интеллекту (International Joint Conference on Artificial Intelligence) IJCAI-89

Понятие Data Mining

Data Mining:

- носит мультидисциплинарный характер, включая в себя элементы:
  - численных методов
  - математической статистики и теории вероятностей
  - теории информации и математической логики
  - искусственного интеллекта и машинного обучения

Задачи Data Mining

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Основные задачи Data Mining:

1. Классификация - определение категории для каждого объекта исследования (классификация заемщиков)
2. Прогнозирование - выявление новых возможных значений в определенной числовой последовательности (прогноз выручки компании)
3. Кластеризация (сегментации) - разбивка множества объектов на группы по каким-либо признакам (сегментация клиентов)
4. Определение взаимосвязей - выявление частоты встречающихся наборов объектов среди множества наборов (анализ покупок)
5. Анализ последовательностей - выявление закономерностей в последовательностях событий (анализ процессов)
6. Анализ отклонений - определение данных, значительно отличающихся от нормы (антифрод)