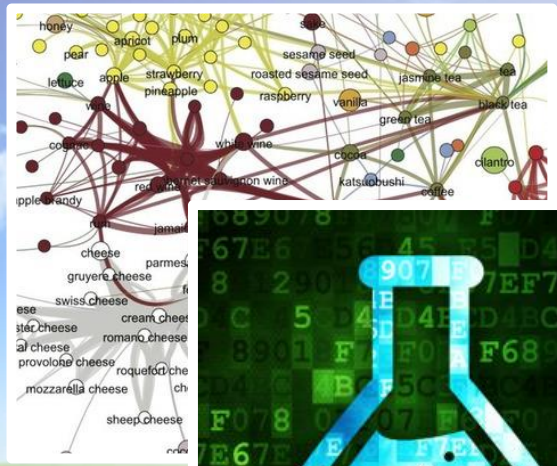
informatica
Put potential to work.™

DIS

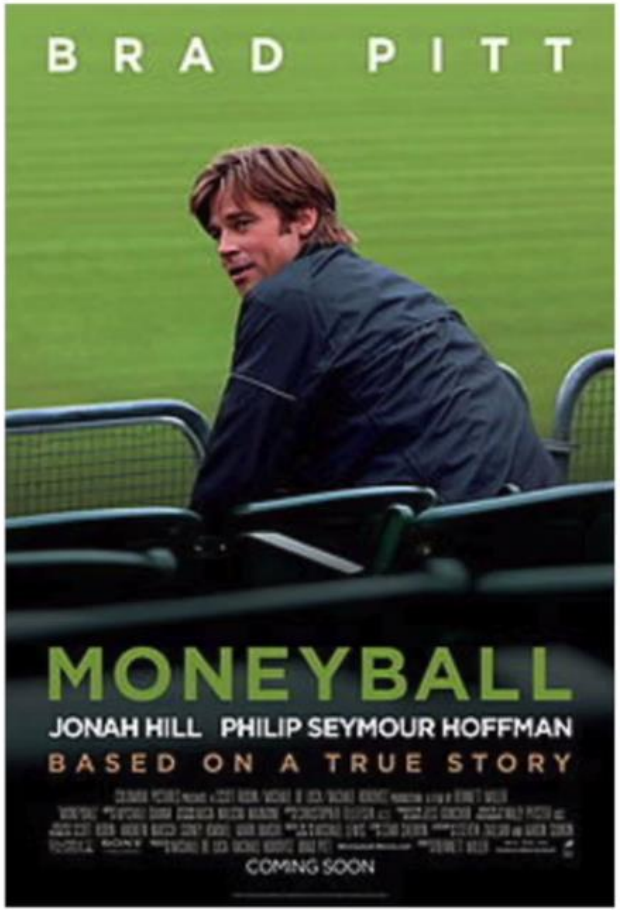
Как Informatica поможет в проектах Data Science

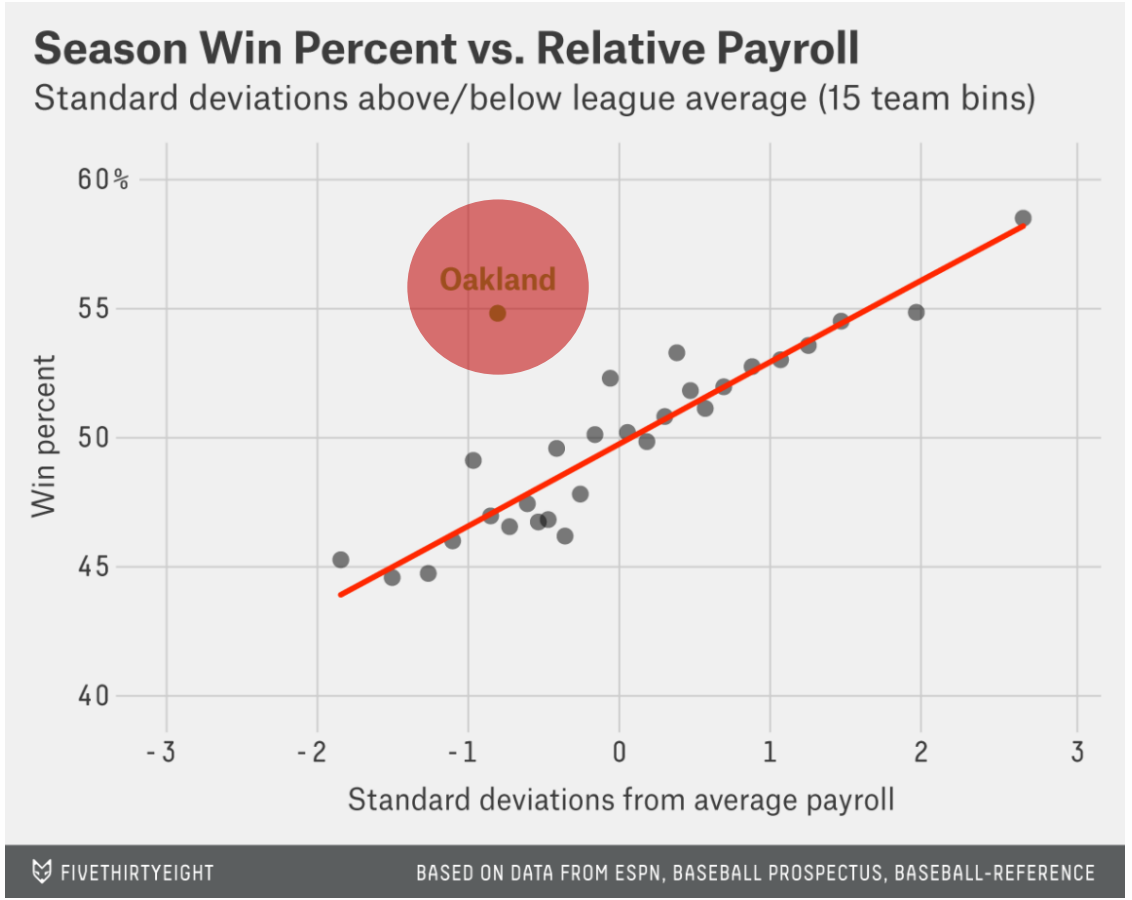


DIS Group

Петр Борисов,
Руководитель направления Big Data

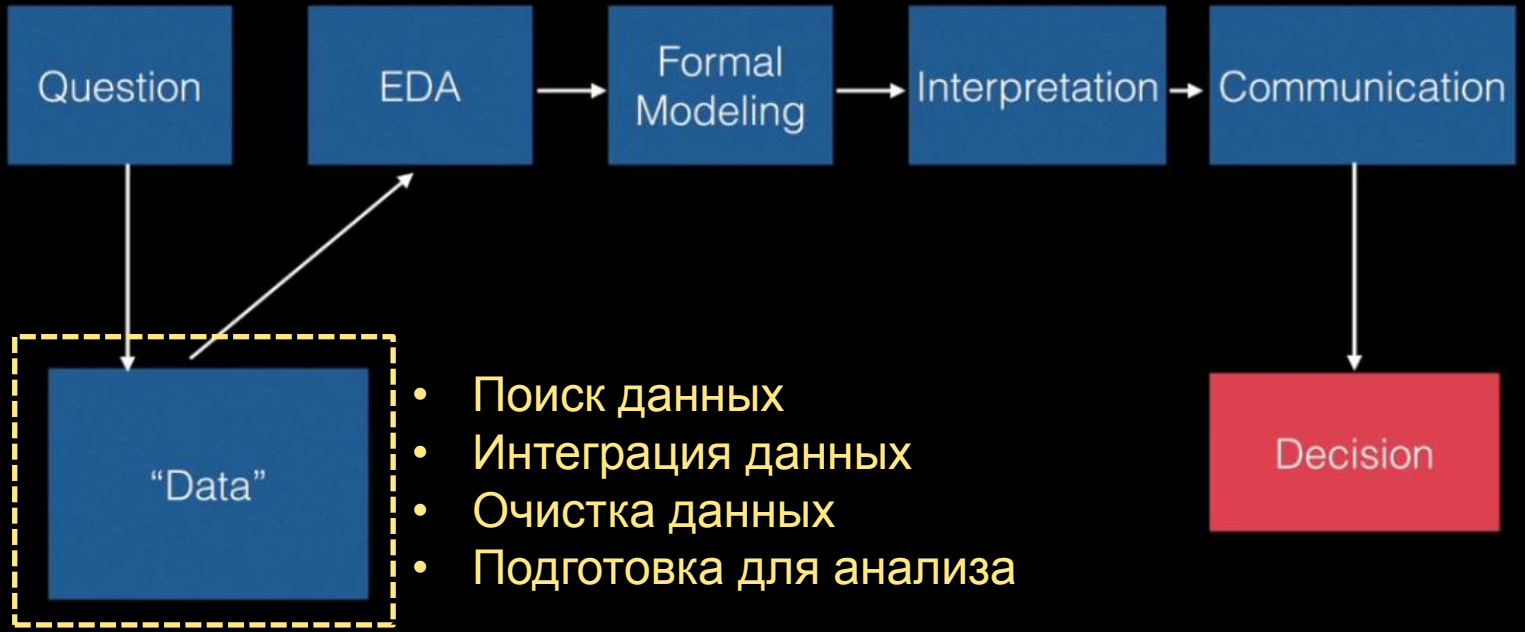
https://en.wikipedia.org/wiki/Moneyball_(film)





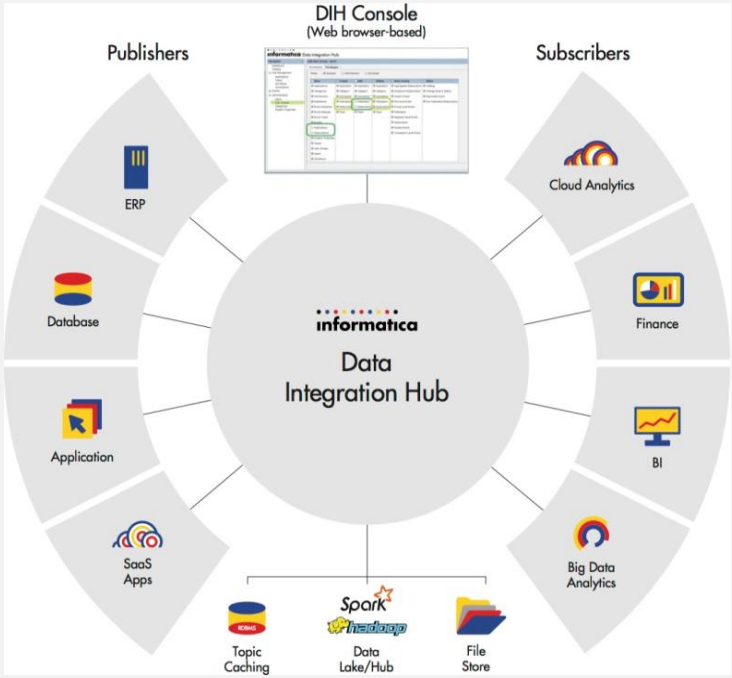
BRIAN CAFFO
ROGER D. PENG
JEFFREY T. LEEK

Structure of a Data Science Project



- Поиск данных
- Интеграция данных
- Очистка данных
- Подготовка для анализа

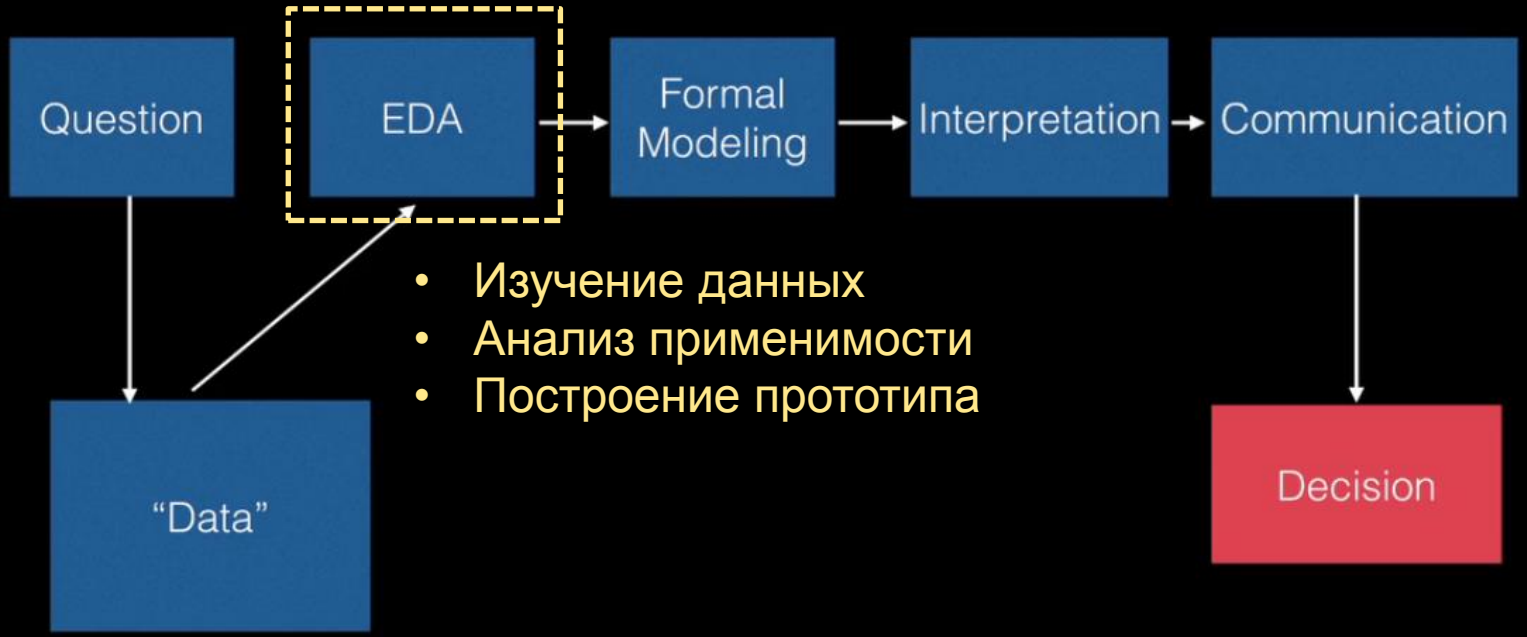
Интеграция и подготовка данных – Informatica DIH



- Wizard для настройки новых потоков данных
- 200+ подключение к различным системам
- Правила обеспечения качества данных
- Сокращение точек интеграции за счет использования publish-subscribe модели
- Прозрачное управление всеми workflow по интеграции и обработке данных
- Автоматическая генерация потоков обработки данных или использование custom-процессов
- Единая среда управления для классических потоков интеграции данных, потоков Big Data, потоков Cloud и Hybrid интеграции

BRIAN CAFFO
ROGER D. PENG
JEFFREY T. LEEK

Structure of a Data Science Project

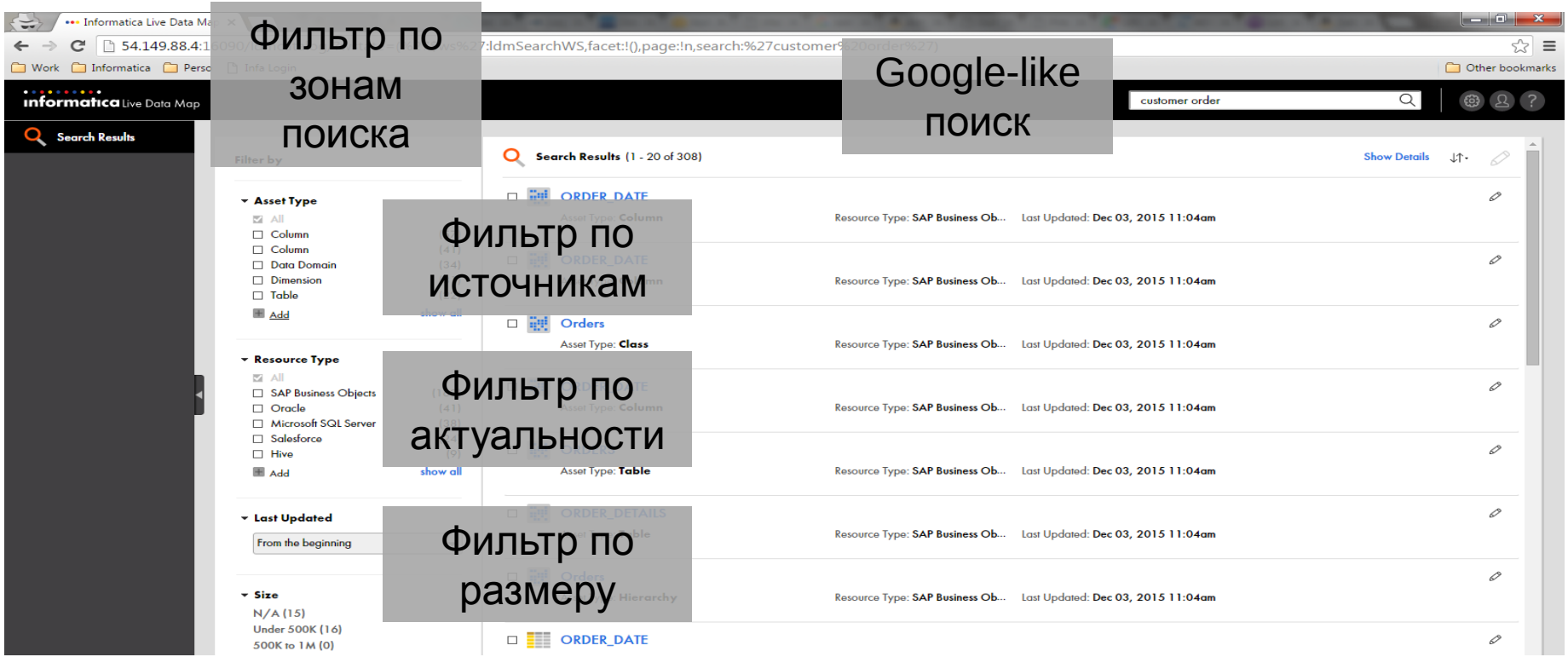


- Изучение данных
- Анализ применимости
- Построение прототипа

EDA – Informatica Intelligent Data Lake

DIS

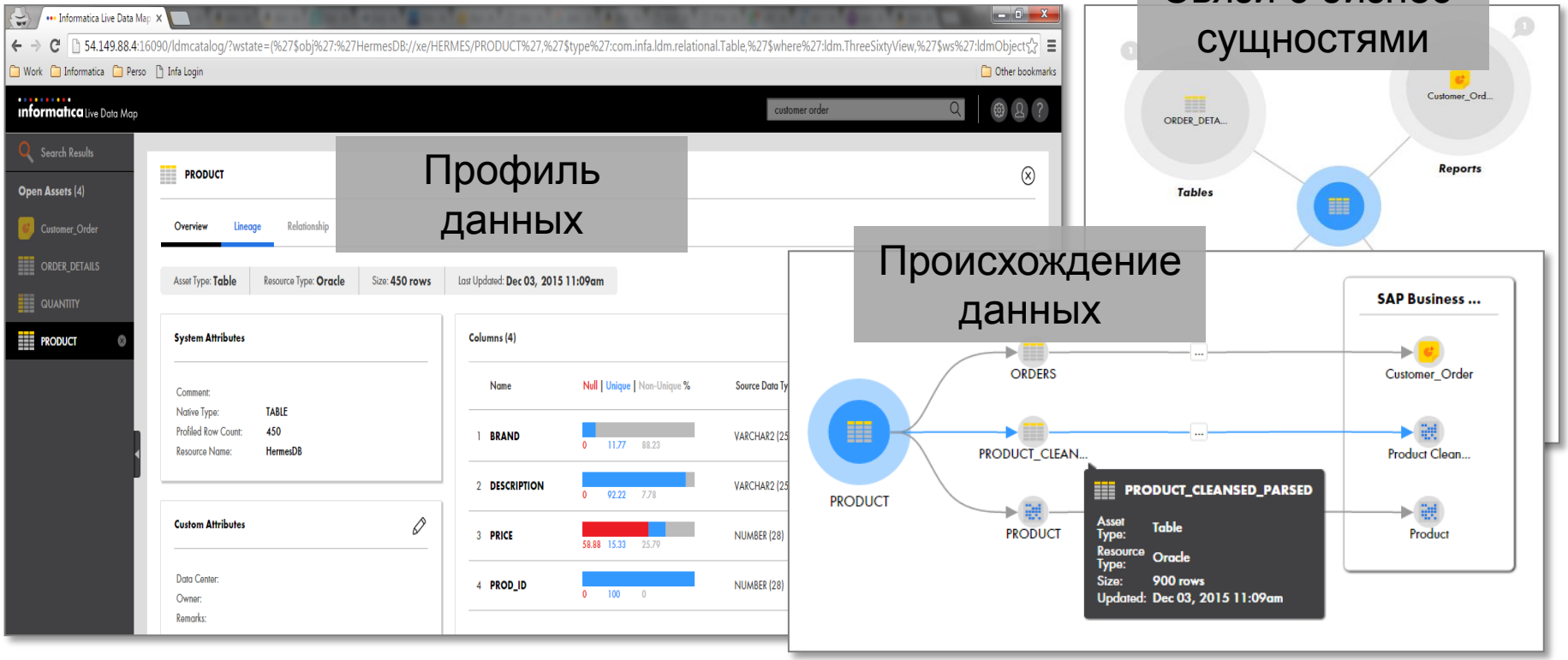
Поиск необходимых данных среди всех корпоративных источников информации



EDA – Informatica Intelligent Data Lake

DIS

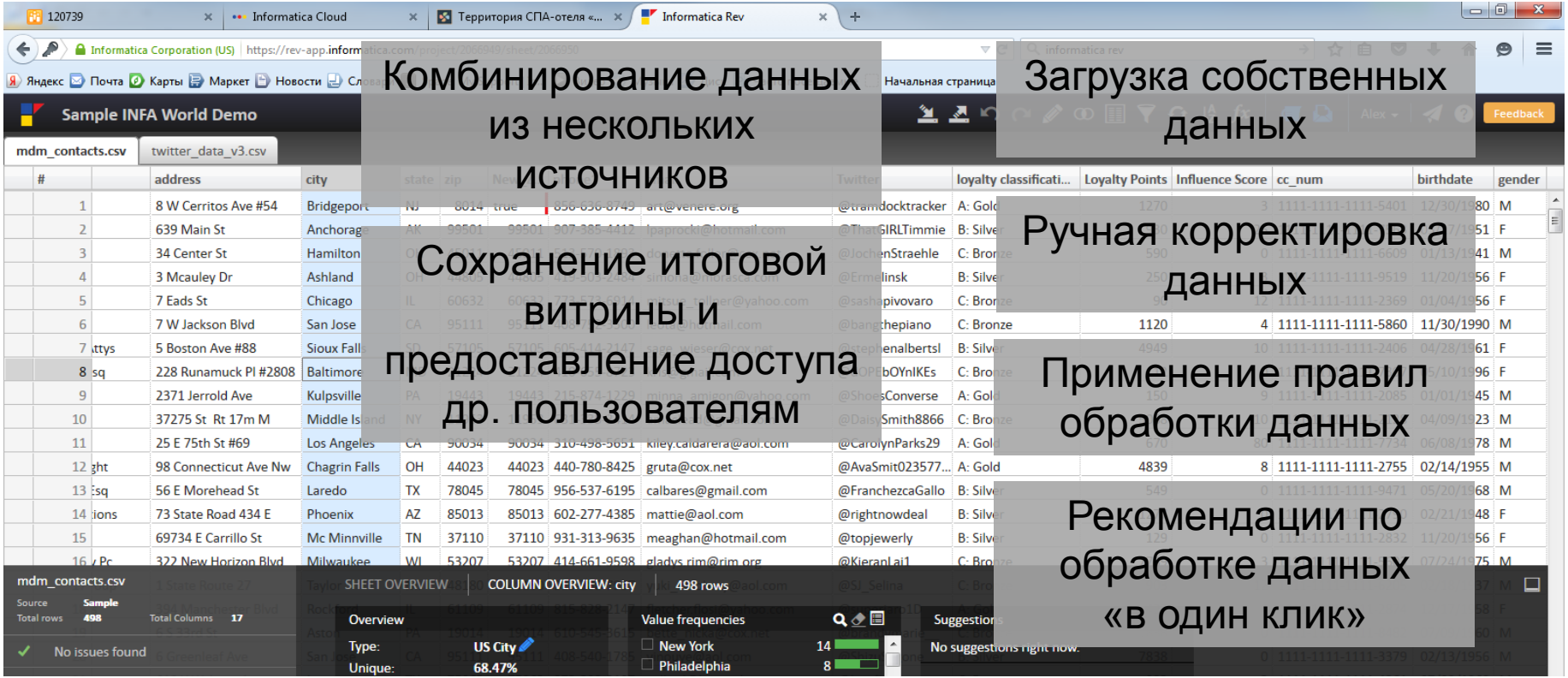
Оценка бизнес-значимости и применимости данных для анализа



EDA – Informatica Intelligent Data Lake

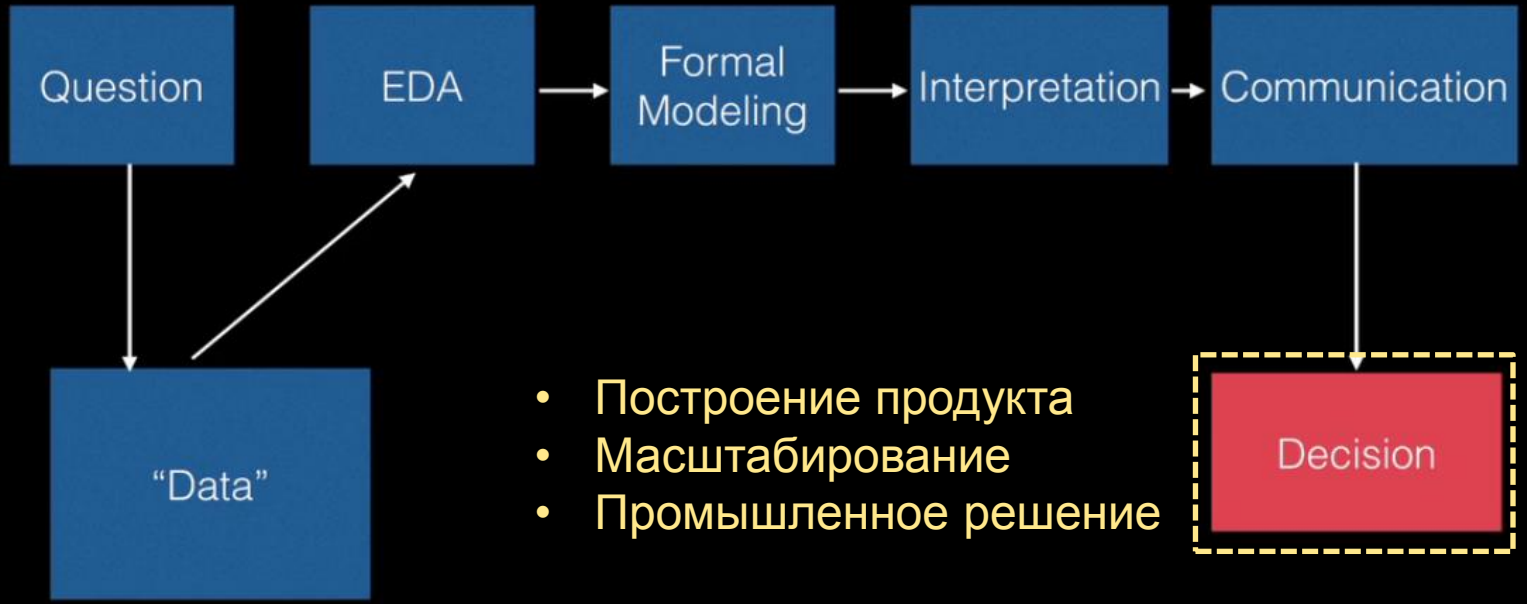
DIS

Поиск необходимых данных среди всех корпоративных источников информации



BRIAN CAFFO
ROGER D. PENG
JEFFREY T. LEEK

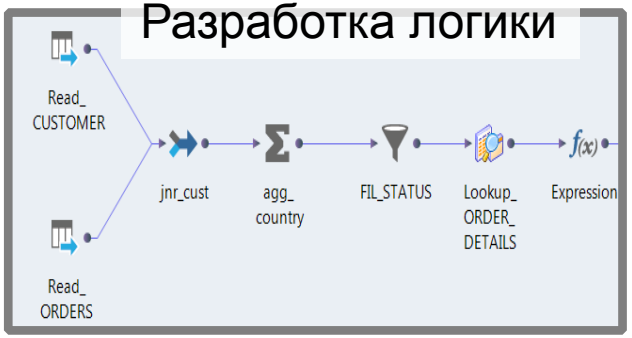
Structure of a Data Science Project



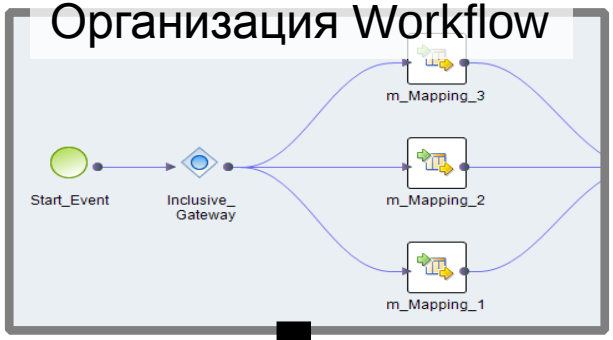
- Построение продукта
- Масштабирование
- Промышленное решение

Informatica Big Data Management

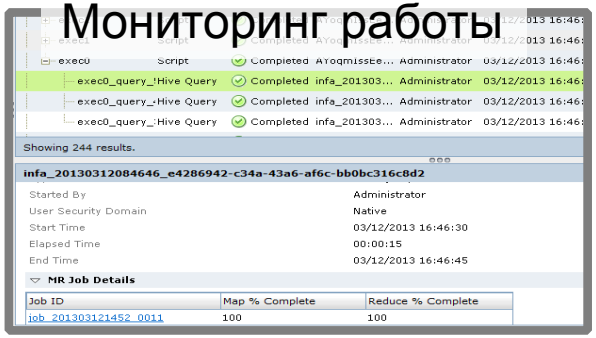
Разработка логики



Организация Workflow



Мониторинг работы

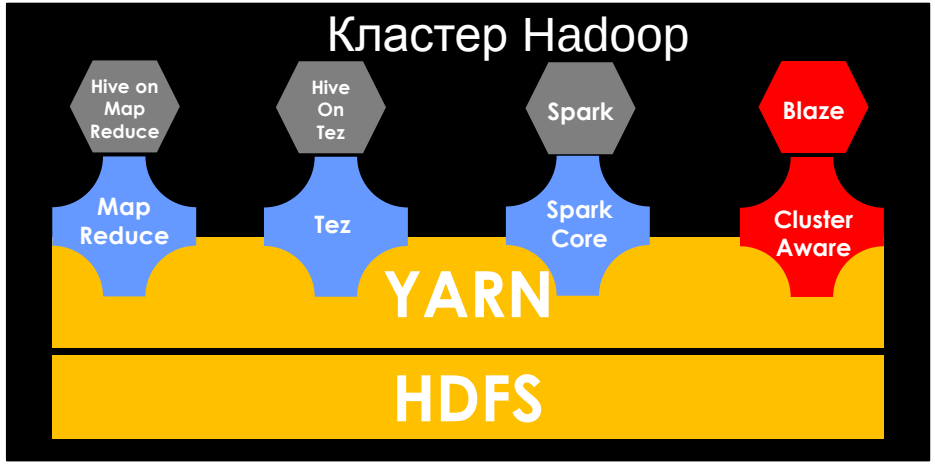


Smart Executor

Native

Native Informatica Data Transformation Engine

Кластер Hadoop



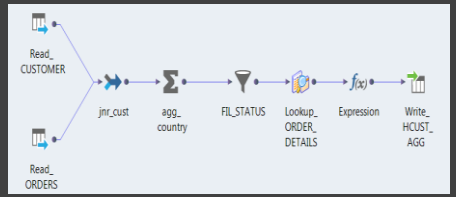
SQL

Database Pushdown

Informatica Intelligent Streaming

DIS

Internet of Things (IoT)



Kafka

Alerting

Transformations

Rules Engine

Spark Streaming

Kafka

Applications

Alerts

Data

Geo-Spatial

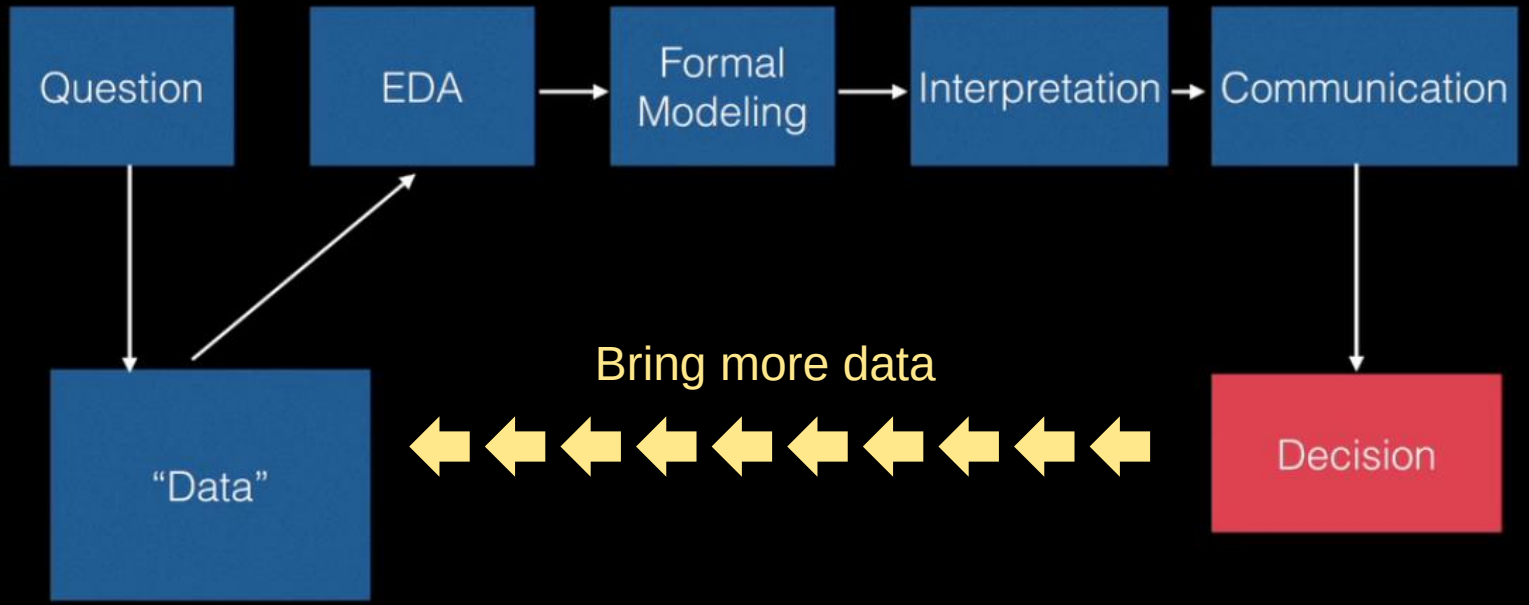
MDM

Campaign

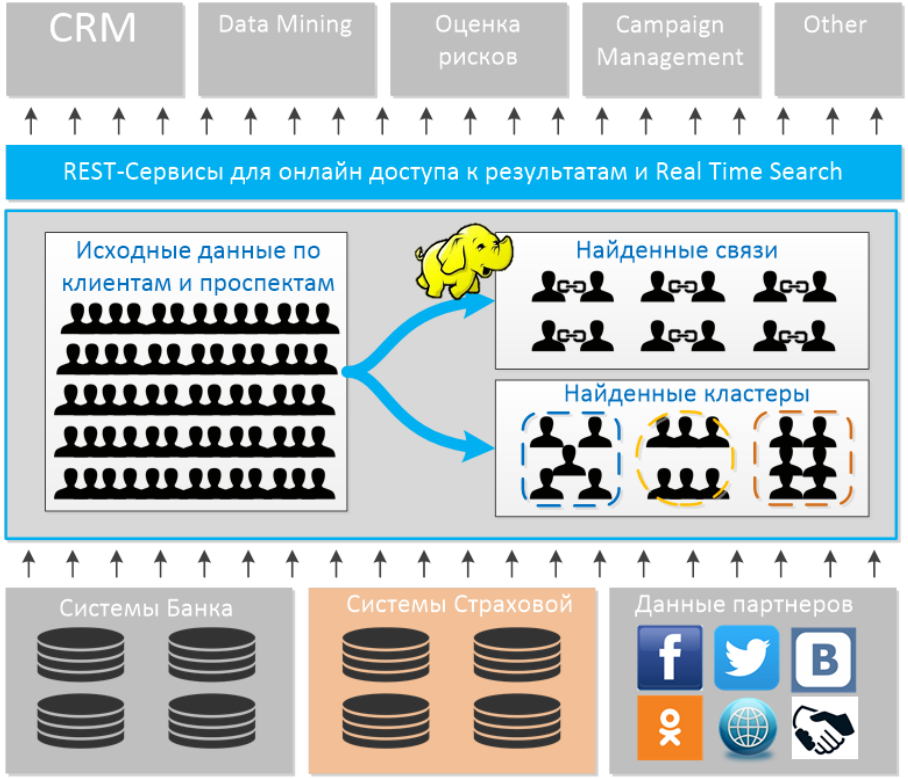
Hadoop

BRIAN CAFFO
ROGER D. PENG
JEFFREY T. LEEK

Structure of a Data Science Project



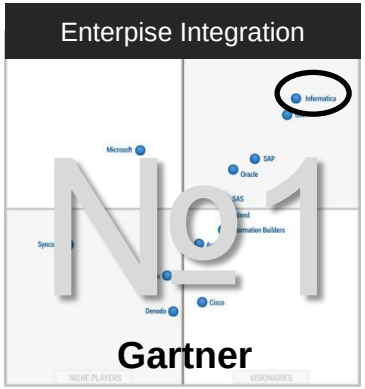
Big Data Relationship Management

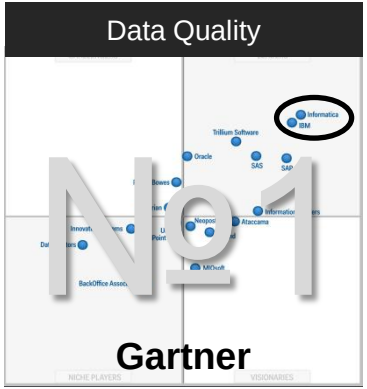


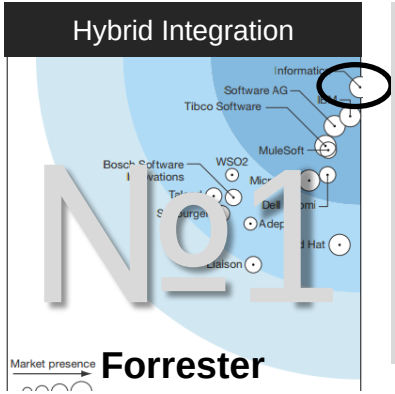
- Сопоставление данных по нечеткой логике
- Матчинг в условиях низкого качества исходных данных
- Multi search (нахождение связей на разных языках)
- Нахождение связей и объединение в кластеры
- Pushdown в Hadoop
- Онлайн API-сервисы

О компании Informatica

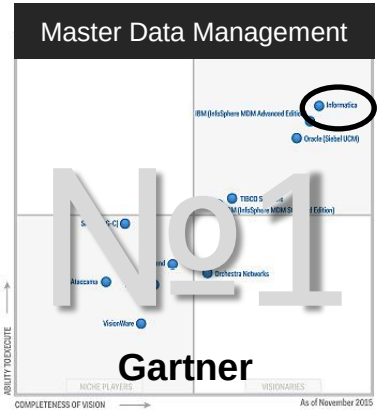
Независимый лидер во всех аспектах по управлению данными!

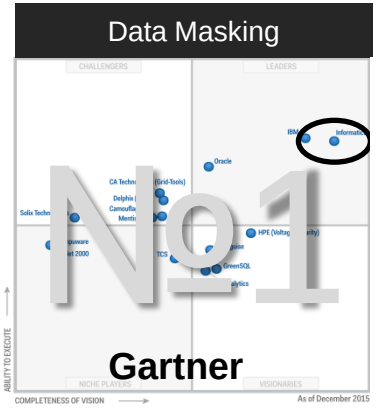


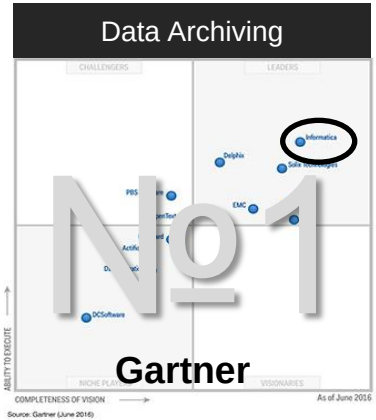




- Более 7 тыс. заказчиков
- 70% из списка Global 500
- Более 450 партнеров в различных областях
- Разработка ведется в США, Европе, Индии и Австралии





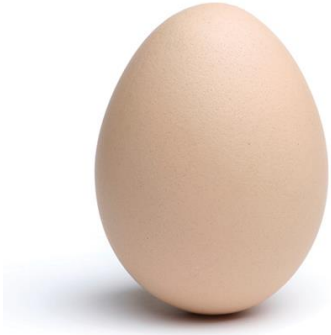


- 37 из ТОП-50 банков США
- 20 из 27 крупнейших банков Европы
- 9 из ТОП-10 глобальных инвестиционных банков
- 31 из ТОП-35 мировых банков
- 15 из ТОП-20 российских банков

Преимущества универсальной платформы

DIS

Простота

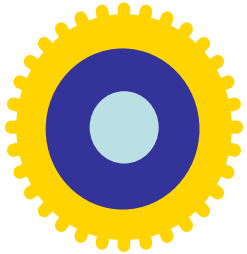


- Быстрый старт
- Доступность экспертизы
- Снижение hand coding

Скорость

- Time to Market
- Скорость разработки
- Скорость развития

Эффективность



- Простота поддержки
- Работа с ключевыми вендорами Hadoop

informatica