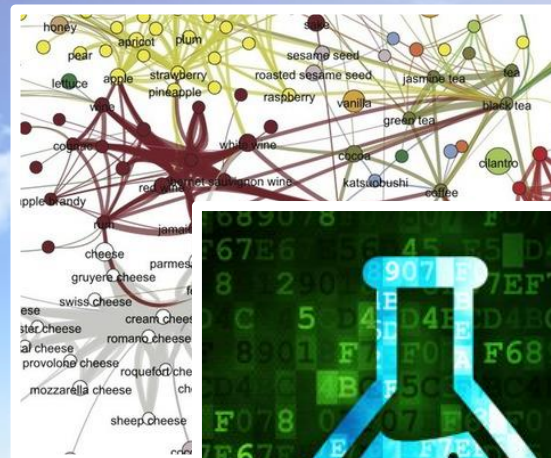
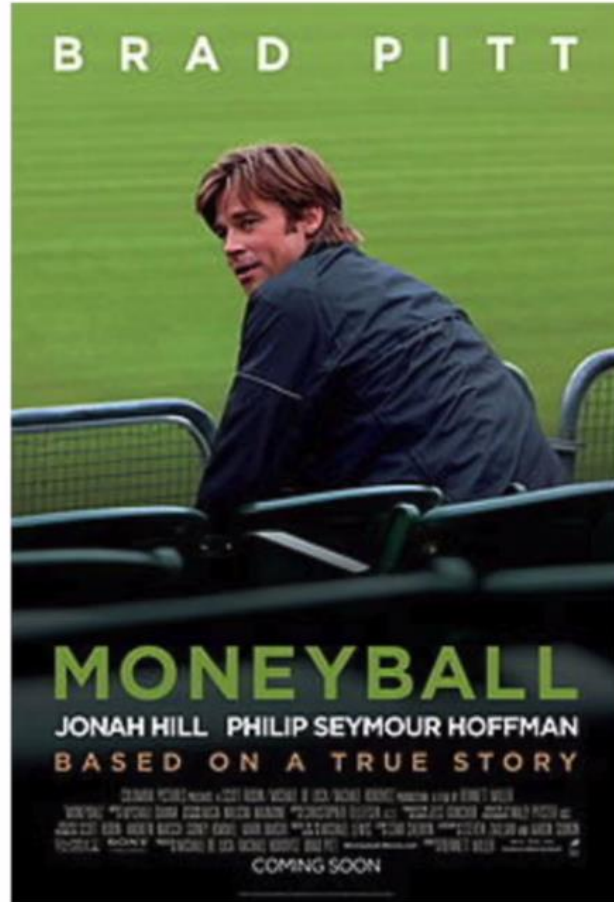


Как Informatica поможет в проектах Data Science



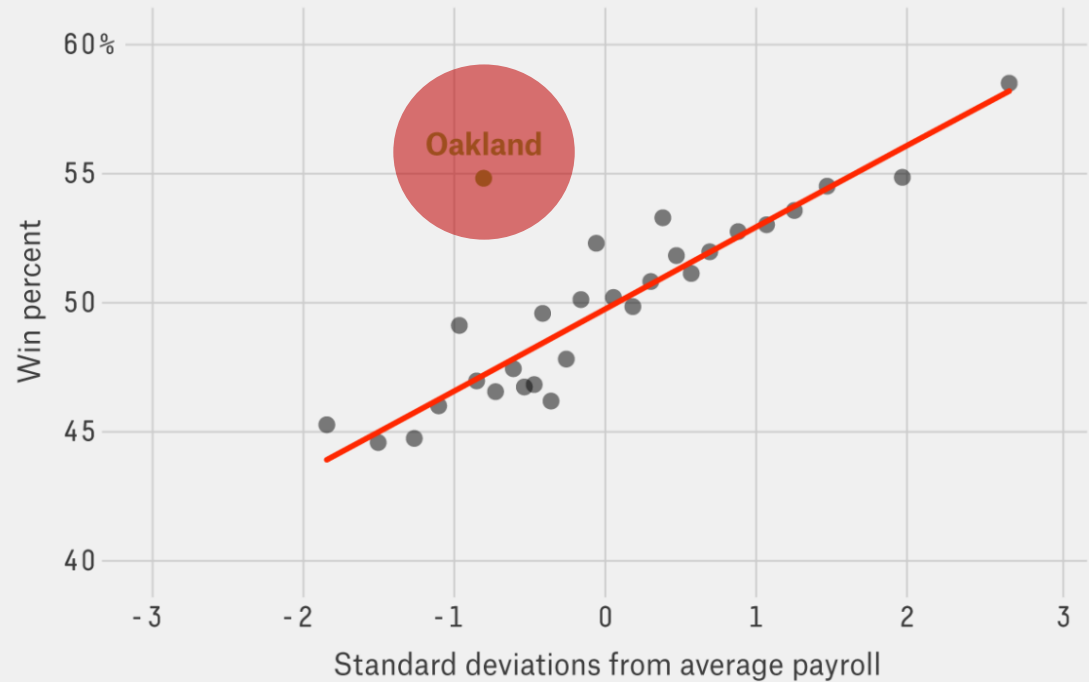
DIS Group

Петр Борисов,
Руководитель направления Big Data

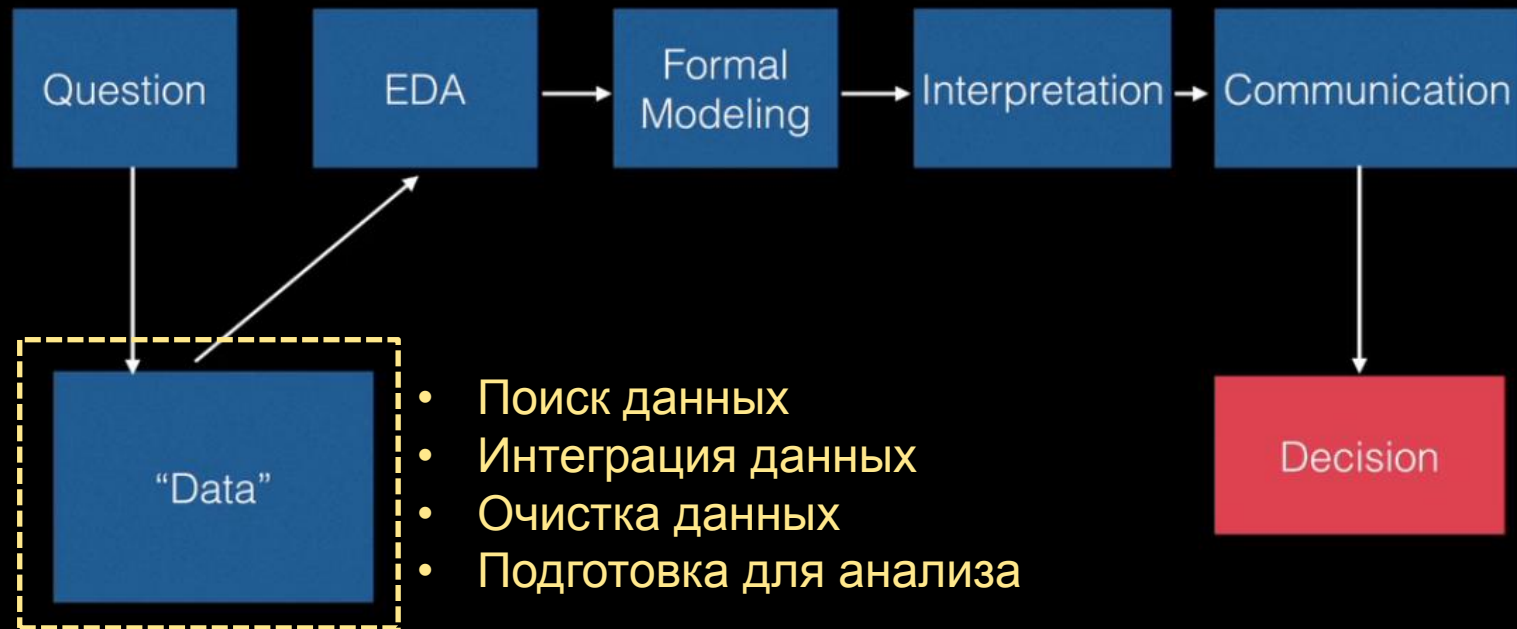


Season Win Percent vs. Relative Payroll

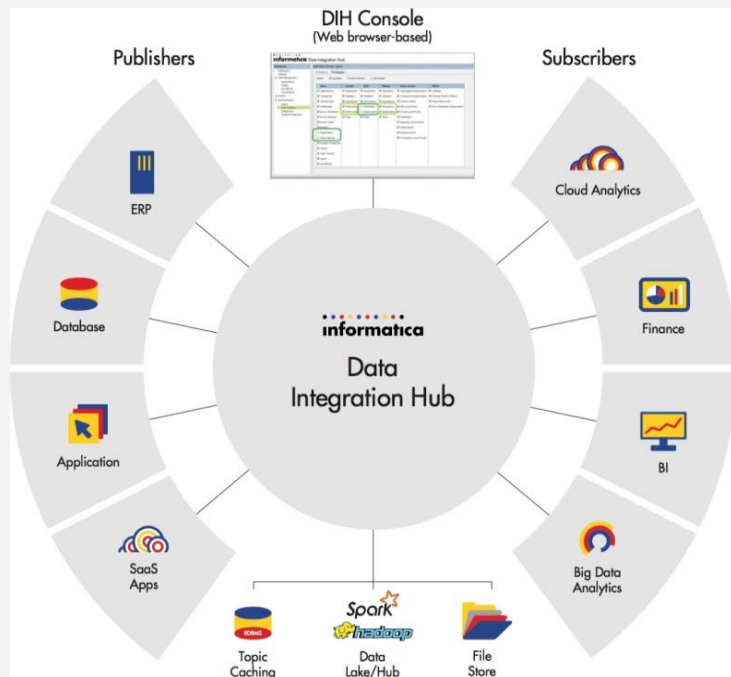
Standard deviations above/below league average (15 team bins)



Structure of a Data Science Project

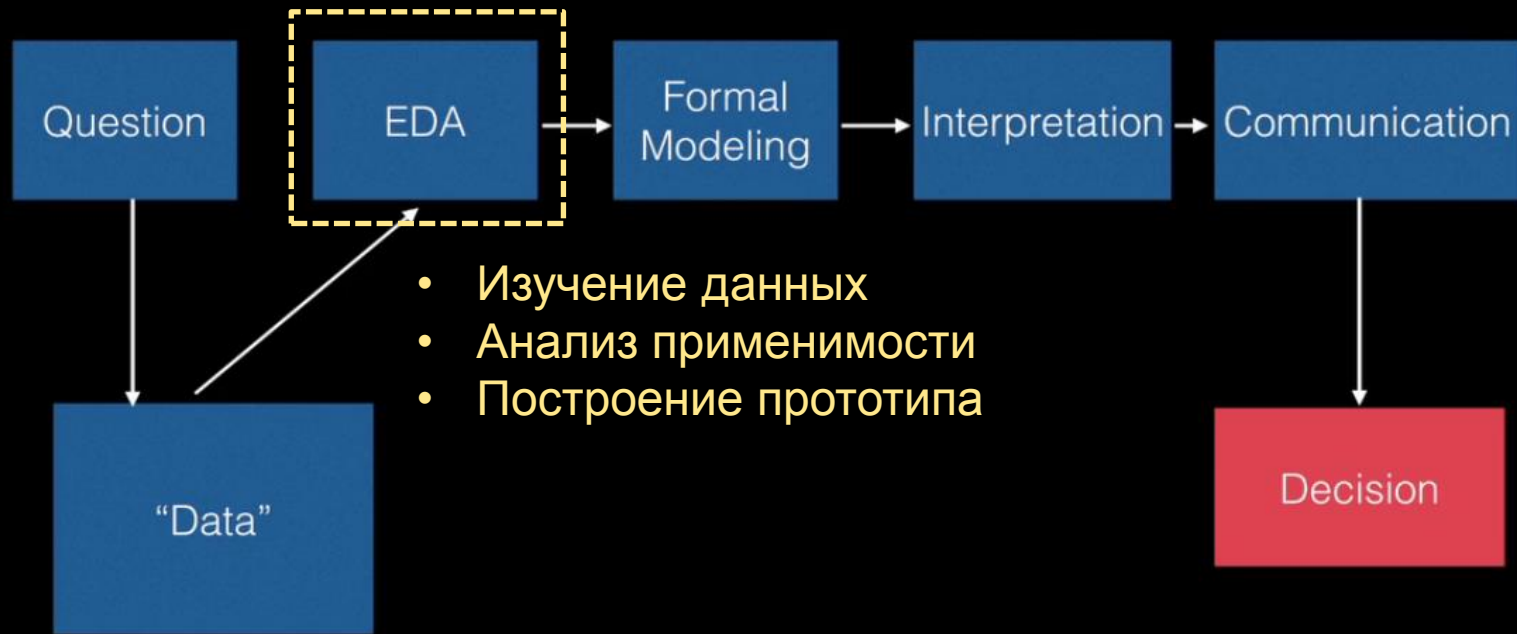


Data Integration Hub



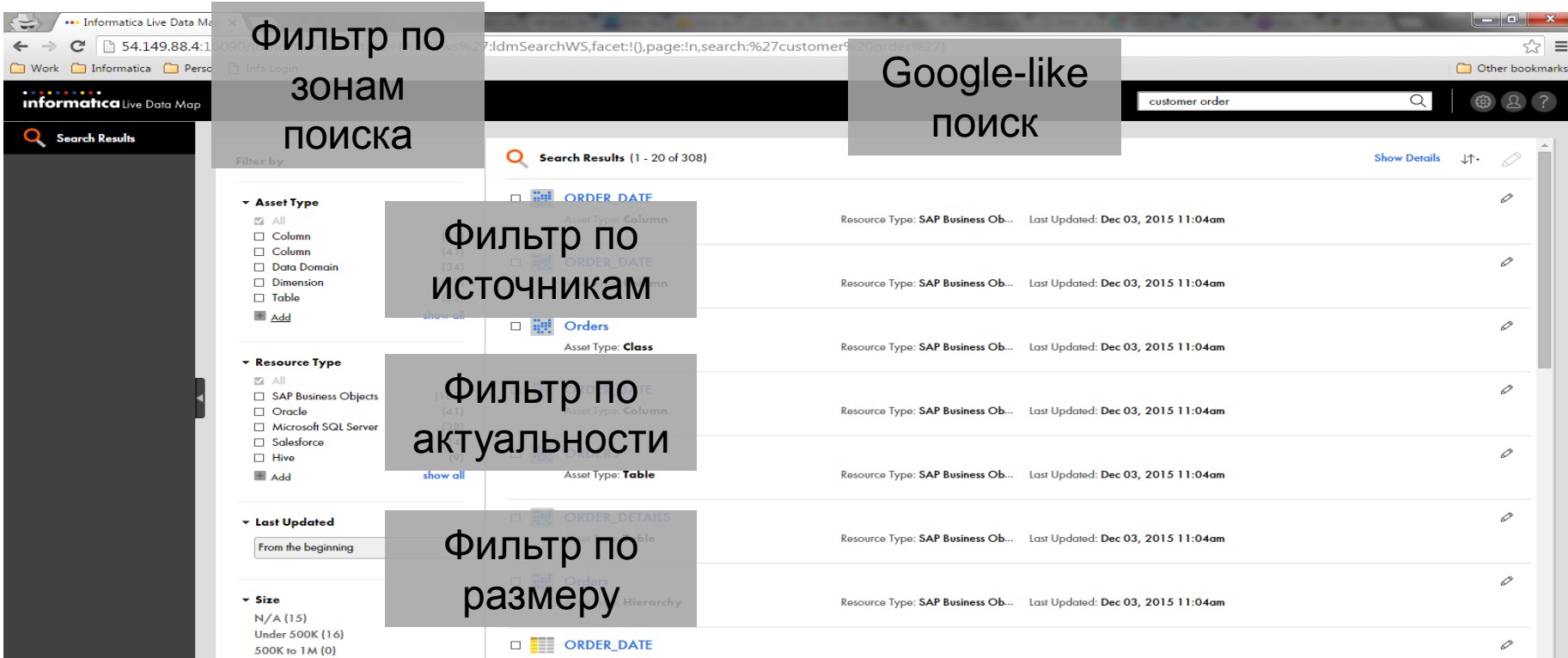
- Wizard для настройки новых потоков данных
- 200+ подключение к различным системам
- Правила обеспечения качества данных
- Сокращение точек интеграции за счет использования publish-subscribe модели
- Прозрачное управление всеми workflow по интеграции и обработке данных
- Автоматическая генерация потоков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

Structure of a Data Science Project



EDA – Informatica Intelligent Data Lake

Поиск необходимых данных среди всех корпоративных источников информации



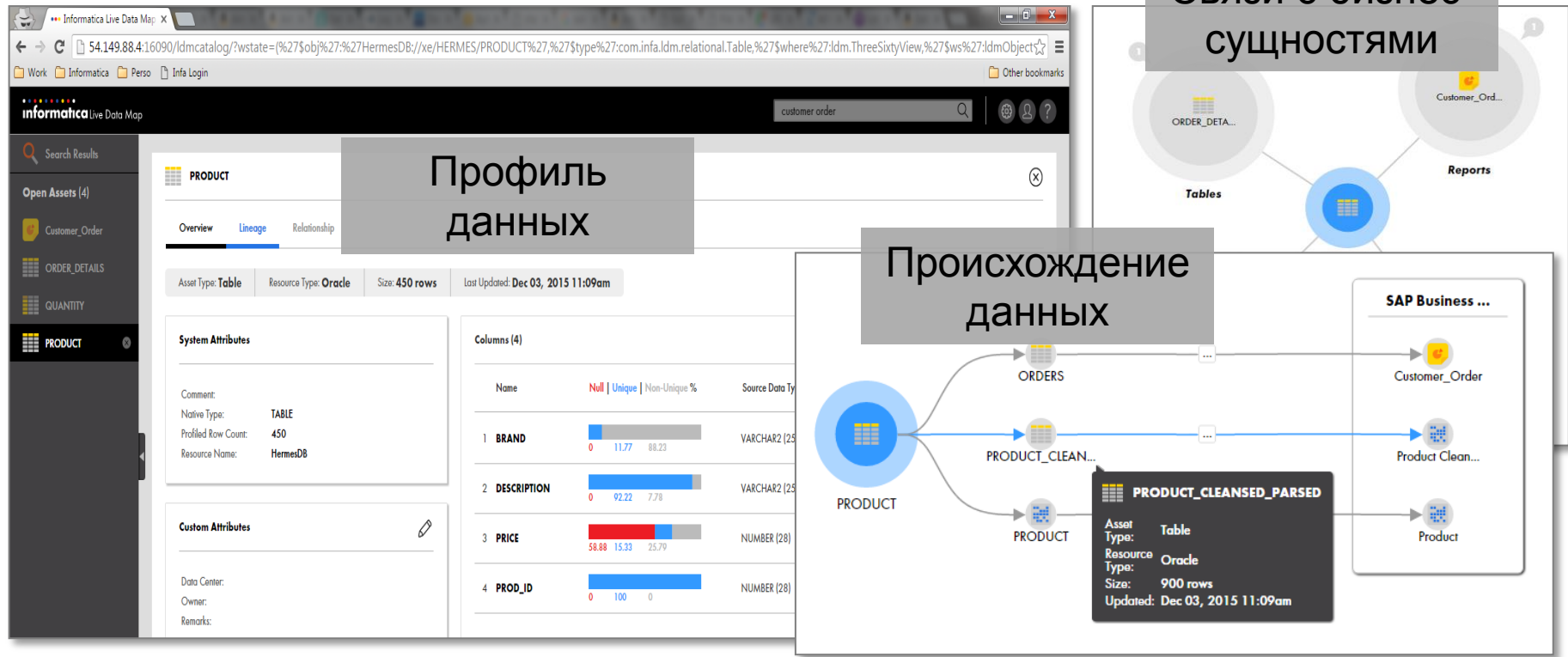
The screenshot displays the Informatica Live Data Map search results page. The interface includes a left sidebar with filter categories, a top search bar, and a main results area. Five callout boxes highlight specific features:

- Фильтр по зонам поиска** (Filter by search zones): Points to the left sidebar containing filters for Asset Type, Resource Type, Last Updated, and Size.
- Google-like ПОИСК** (Google-like search): Points to the search bar at the top right of the interface.
- Фильтр по источникам** (Filter by source): Points to the 'Asset Type' filter section in the sidebar.
- Фильтр по актуальности** (Filter by freshness): Points to the 'Last Updated' filter section in the sidebar.
- Фильтр по размеру** (Filter by size): Points to the 'Size' filter section in the sidebar.

The search results table shows columns for Asset Name, Asset Type, Resource Type, and Last Updated. The results are filtered to show only 'ORDERS' and 'ORDER_DATE' assets, all of which are 'SAP Business Objects' and were last updated on 'Dec 03, 2015 11:04am'.

EDA – Informatica Intelligent Data Lake

Оценка бизнес-значимости и применимости данных для анализа



EDA – Informatica Intelligent Data Lake

Поиск необходимых данных среди всех корпоративных источников информации

The screenshot displays the Informatica Cloud web interface. At the top, there are browser tabs for 'Informatica Cloud', 'Территория СПА-отеля...', and 'Informatica Rev'. The main content area shows a 'Sample INFA World Demo' with two CSV files: 'mdm_contacts.csv' and 'twitter_data_v3.csv'. A large table of data is visible, with columns including '#', 'address', 'city', 'state', 'zip', 'New York', 'Twitter', 'loyalty classificati...', 'Loyalty Points', 'Influence Score', 'cc_num', 'birthdate', and 'gender'. The table contains 498 rows of data. Overlaid on the interface are several semi-transparent boxes with Russian text describing key features: 'Комбинирование данных из нескольких источников' (Combining data from multiple sources), 'Загрузка собственных данных' (Loading own data), 'Ручная корректировка данных' (Manual data correction), 'Сохранение итоговой витрины и предоставление доступа др. пользователям' (Saving the final data warehouse and providing access to other users), 'Применение правил обработки данных' (Applying data processing rules), and 'Рекомендации по обработке данных «в один клик»' (Data processing recommendations 'with one click'). At the bottom, there are sections for 'Overview' (showing 'US City' with a 68.47% unique value frequency), 'Value frequencies' (showing 'New York' and 'Philadelphia' with 14 and 8 occurrences respectively), and 'Suggestions' (showing 'No suggestions right now.').

Комбинирование данных из нескольких источников

Загрузка собственных данных

Ручная корректировка данных

Сохранение итоговой витрины и предоставление доступа др. пользователям

Применение правил обработки данных

Рекомендации по обработке данных «в один клик»

#	address	city	state	zip	New York	Twitter	loyalty classificati...	Loyalty Points	Influence Score	cc_num	birthdate	gender
1	8 W Cerritos Ave #54	Bridgeport	NY	8014	true	@tramdocktracker	A: Gold	1270	3	1111-1111-1111-5401	12/30/1980	M
2	639 Main St	Anchorage	AK	99501	99501	@TheSIRLTimmie	B: Silver	590	0	1111-1111-1111-6609	01/13/1951	F
3	34 Center St	Hamilton	OH	44605	44605	@JochenStraehle	C: Bronze	250	0	1111-1111-1111-9519	11/20/1956	F
4	3 McAuley Dr	Ashland	OH	44805	44805	@Ermilinsk	B: Silver	1120	4	1111-1111-1111-5860	11/30/1990	M
5	7 Eads St	Chicago	IL	60632	60632	@sashapivovaro	C: Bronze	4949	10	1111-1111-1111-2406	04/28/1961	F
6	7 W Jackson Blvd	San Jose	CA	95111	95111	@banchepliano	C: Bronze	150	9	1111-1111-1111-2385	01/07/1945	M
7	5 Boston Ave #88	Sioux Falls	SD	57105	57105	@Ghebenalberts1	B: Silver	670	8	1111-1111-1111-7734	04/09/1923	M
8	228 Runamuck Pl #2808	Baltimore	MD	21208	21208	@ShoresConverse	A: Gold	549	0	1111-1111-1111-9471	05/20/1968	M
9	2371 Jerrold Ave	Kulpsville	PA	19443	19443	@DaisSmith8866	C: Bronze	129	0	1111-1111-1111-2832	11/20/1956	F
10	37275 St Rt 17m M	Middle Island	NY	11967	11967	@CarolynParks29	A: Gold	4839	8	1111-1111-1111-2755	02/14/1955	M
11	25 E 75th St #69	Los Angeles	CA	90034	90034	@AvaSmit023577...	A: Gold	549	0	1111-1111-1111-9471	05/20/1968	M
12	98 Connecticut Ave Nw	Chagrin Falls	OH	44023	44023	@FranchezzaGallo	B: Silver	129	0	1111-1111-1111-2832	11/20/1956	F
13	56 E Morehead St	Laredo	TX	78045	78045	@rightnowdeal	B: Silver	129	0	1111-1111-1111-2832	11/20/1956	F
14	73 State Road 434 E	Phoenix	AZ	85013	85013	@topjewelry	B: Silver	129	0	1111-1111-1111-2832	11/20/1956	F
15	69734 E Carrillo St	Mc Minnville	TN	37110	37110	@KieranLai1	C: Bronze	129	0	1111-1111-1111-2832	11/20/1956	F
16	322 New Horizon Blvd	Milwaukee	WI	53207	53207							

mdm_contacts.csv

Source: Sample

Total rows: 498

Total Columns: 17

✓ No issues found

Overview

Type: US City

Unique: 68.47%

Value frequencies

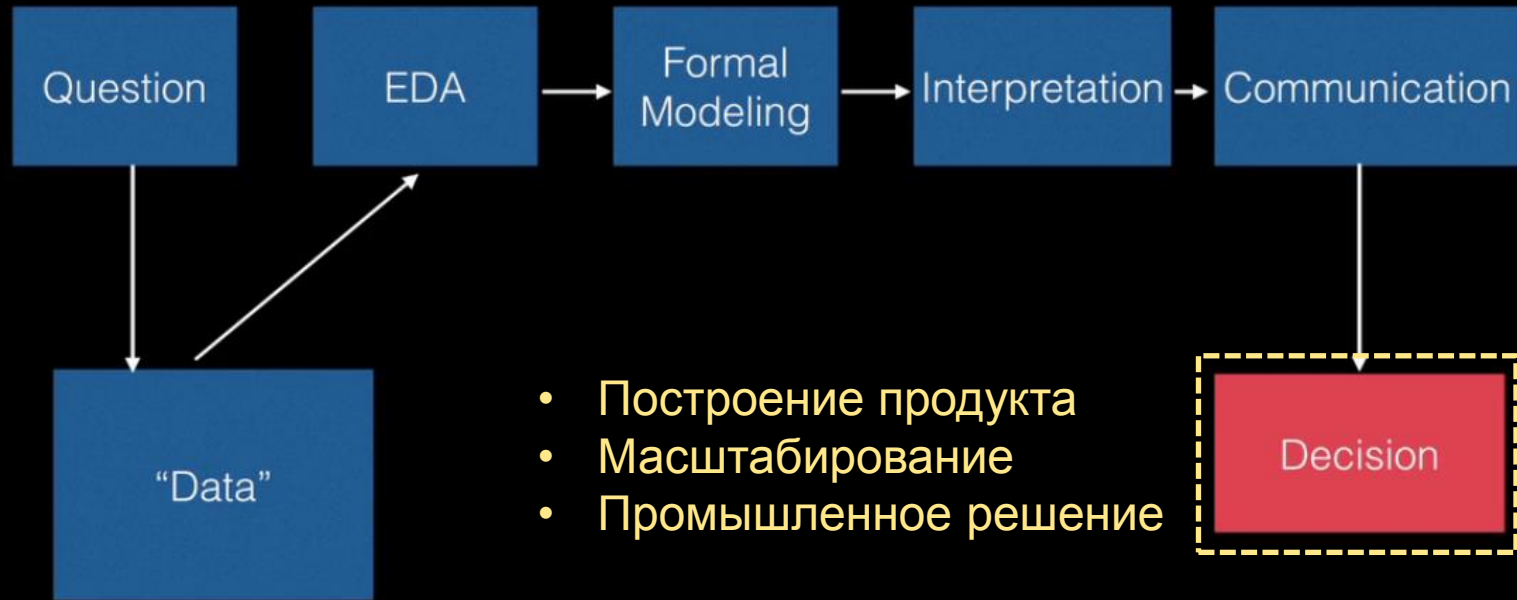
New York: 14

Philadelphia: 8

Suggestions

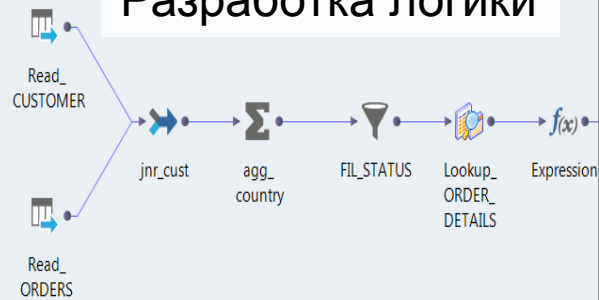
No suggestions right now.

Structure of a Data Science Project

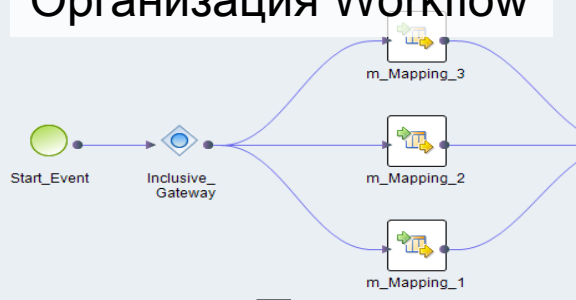


Informatica Big Data Management

Разработка логики



Организация Workflow



Мониторинг работы

The screenshot displays the Informatica job monitoring interface. It shows a list of jobs with columns for job name, status, user, and time. Below the list, there is a section for 'Showing 244 results.' and a table for 'MR Job Details'.

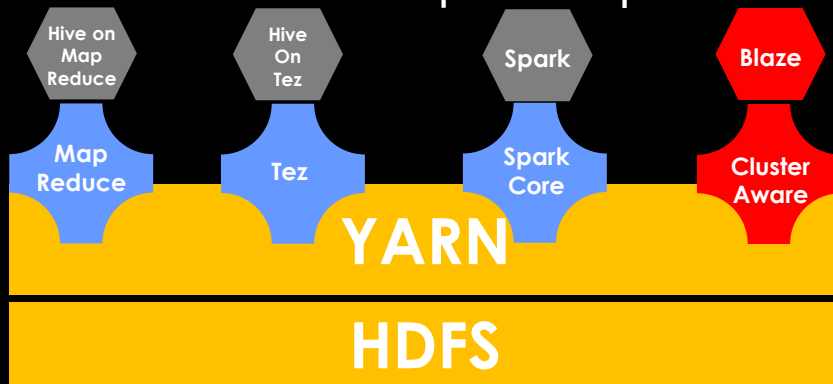
Job ID	Map % Complete	Reduce % Complete
job_201303121452_0011	100	100

Smart Executor

Native

Native
Informatica
Data
Transformation
Engine

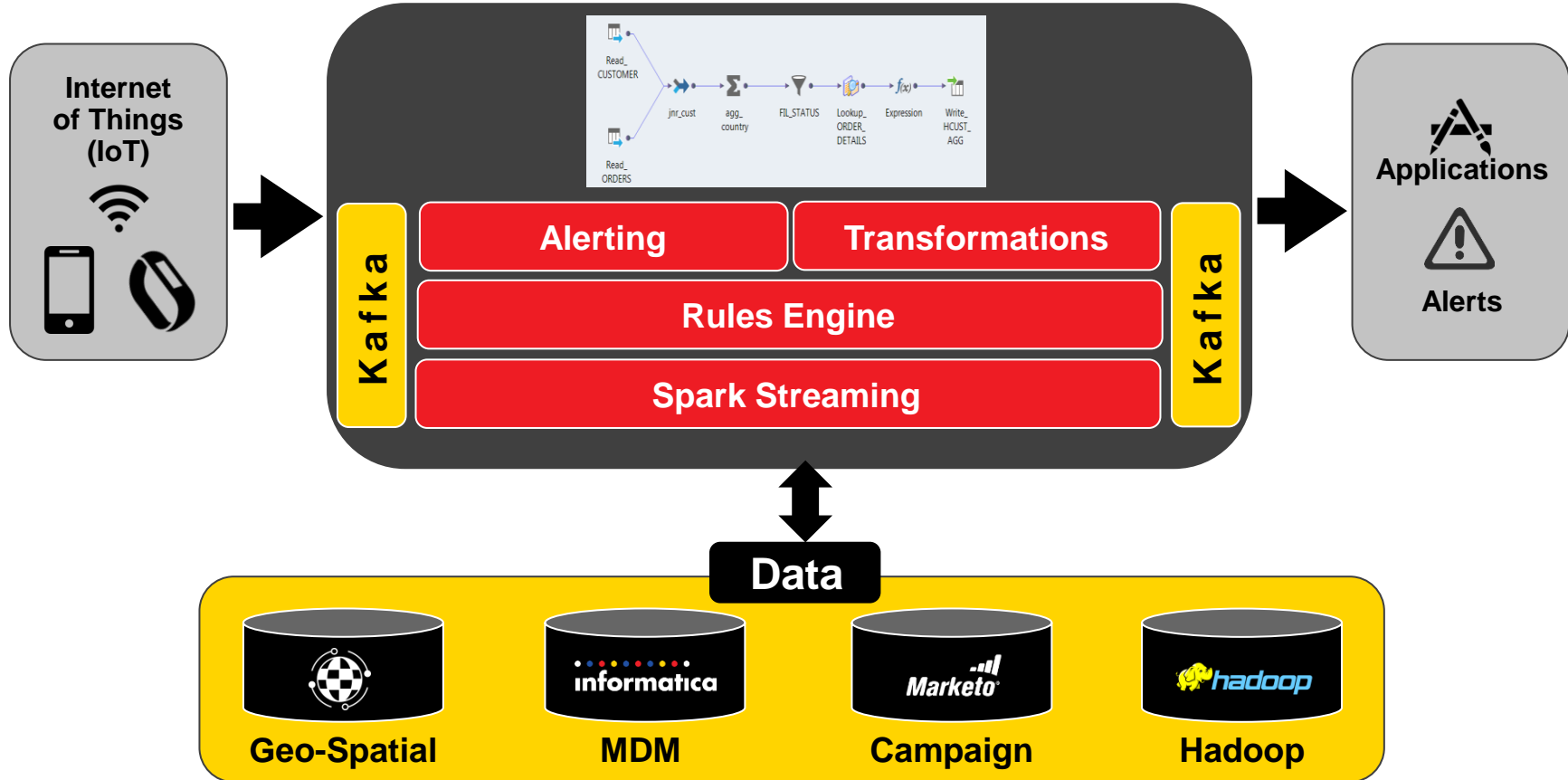
Кластер Hadoop



SQL

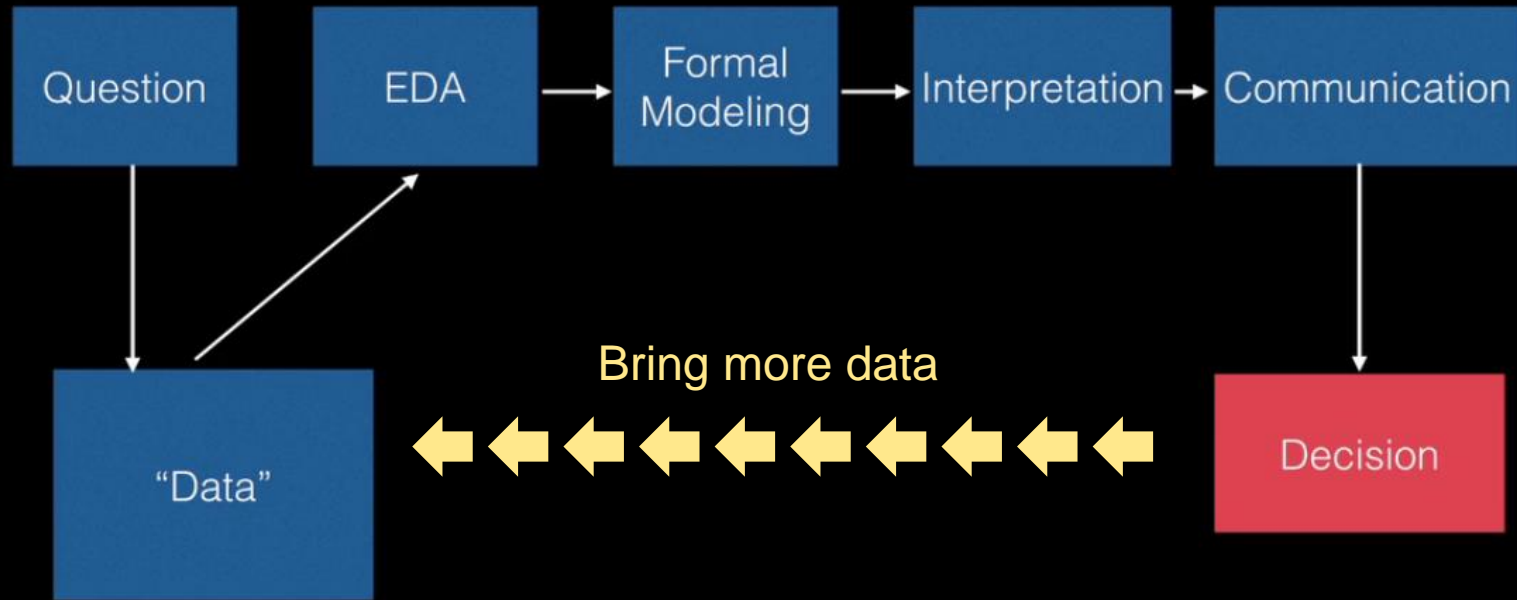
Database
Pushdown

Informatica Intelligent Streaming

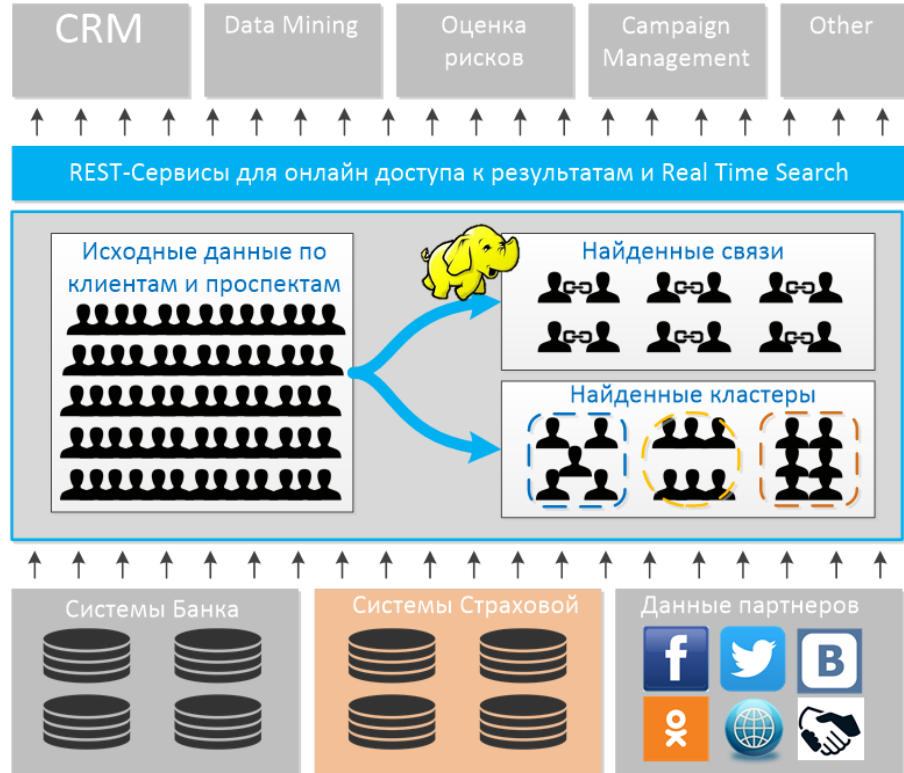


BRIAN CAFFO
ROGER D. PENG
JEFFREY T. LEEK

Structure of a Data Science Project



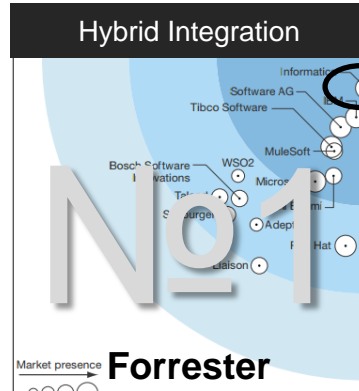
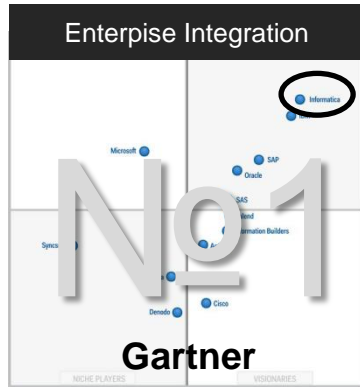
Big Data Relationship Management



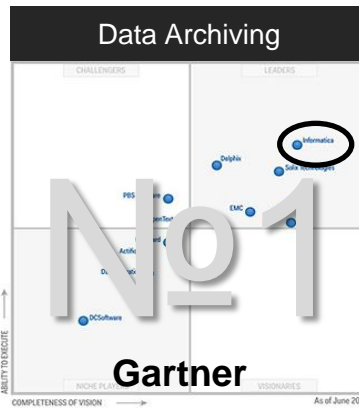
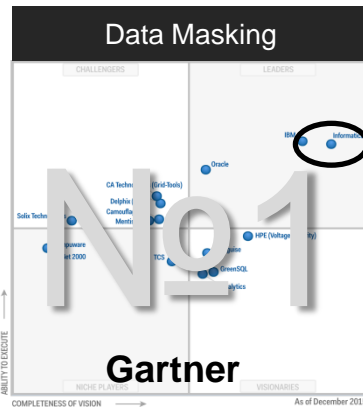
- Сопоставление данных по нечеткой логике
- Матчинг в условиях низкого качества исходных данных
- Multi search (нахождение связей на разных языках)
- Нахождение связей и объединение в кластеры
- Pushdown в Hadoop
- Онлайн API-сервисы

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

Простота



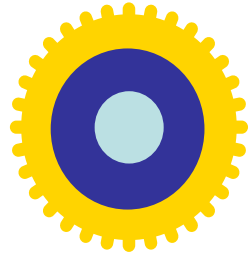
- Быстрый старт
- Доступность экспертизы
- Снижение hand coding

Скорость



- Time to Market
- Скорость разработки
- Скорость развития

Эффективность



- Простота поддержки
- Работа с ключевыми вендорами Hadoop