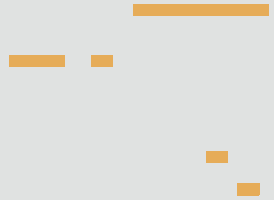
ORACLE

ORACLE

Обзор платформы управления данными

Андрей Пивоваров



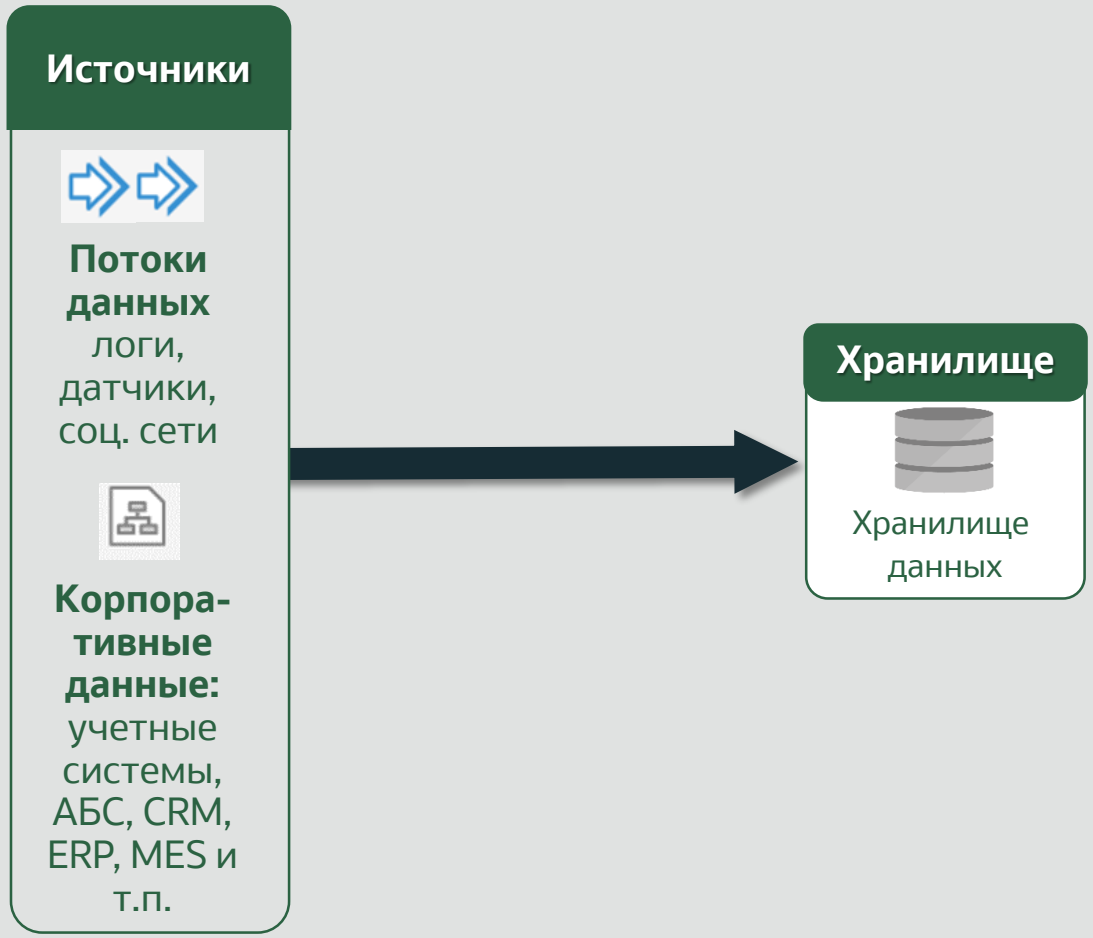
Платформа управления данными

Источники

Потоки данных
логи, датчики, соц. сети

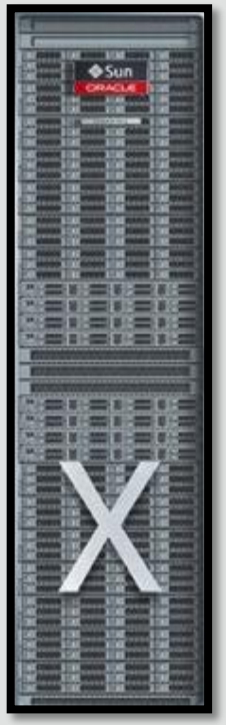
Корпора-тивные данные:
учетные системы, АБС, CRM, ERP, MES и т.п.

Платформа управления данными



Принципы Exadata

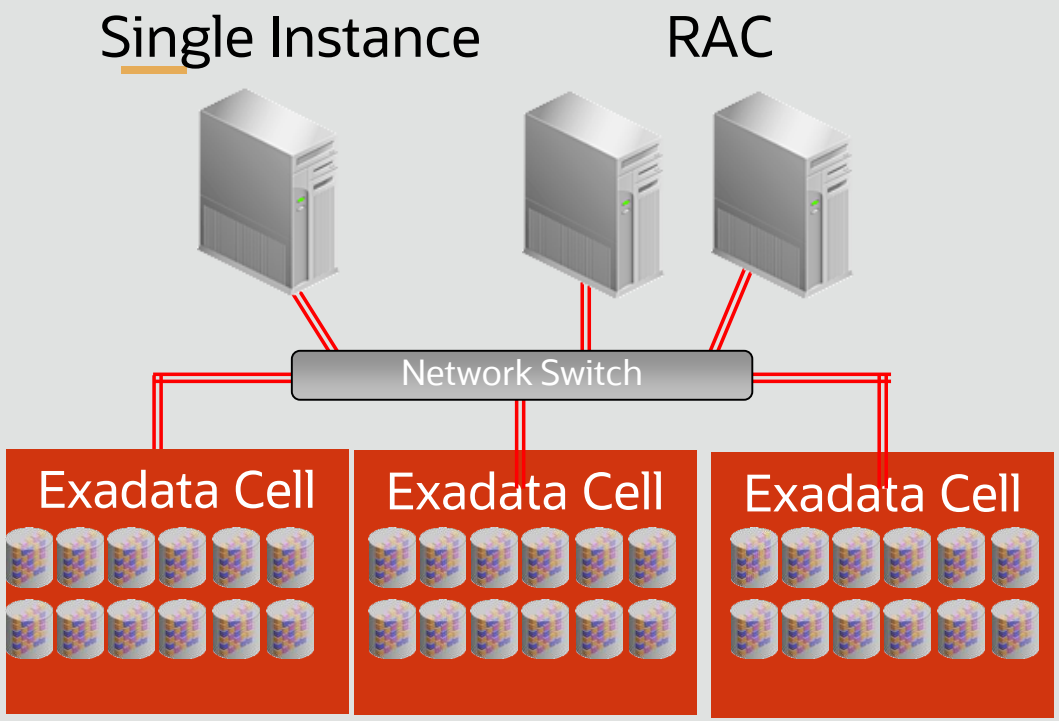
Цель: Создать лучшую платформу для всех типов нагрузки баз данных



- Идеальное HW для баз данных - Масштабируемость, оптимизированные вычисления, сеть и СХД для высочайшей производительности и снижения затрат
- Умное ПО – специальные алгоритмы вносят улучшения и ускорения для всех типов обработки данных: **OLTP, Аналитики, Консолидации**
- Интеграция всего стека – Оптимизации всего стека, плюс тестирование, патчирование, упрощение поддержки для уменьшения затрат

Идентичные системы в облаке и в ЦОД заказчиков

Архитектура Exadata



- Каждая ячейка Exadata – самостоятельный сервер с установленными дисками и ПО Exadata
- Данные «размазаны» между многими ячейками Exadata
- Нет ограничения на количество ячеек в системе
- Ячейки выполняют множество операций, которые в традиционной архитектуре делает Oracle
- Ячейки работают в режиме MPP

Уникальные возможности Exadata для аналитики

Smart Scan (SQL Offload)

Обработка данных на ячейках Exadata уменьшает сетевой трафик и освобождает ресурсы сервера БД

Flash Cache (PCI NVMe Flash)

Горячие данные автоматически попадают в PCI Flash, неактивные хранятся на дешевых дисках

Storage Indexes

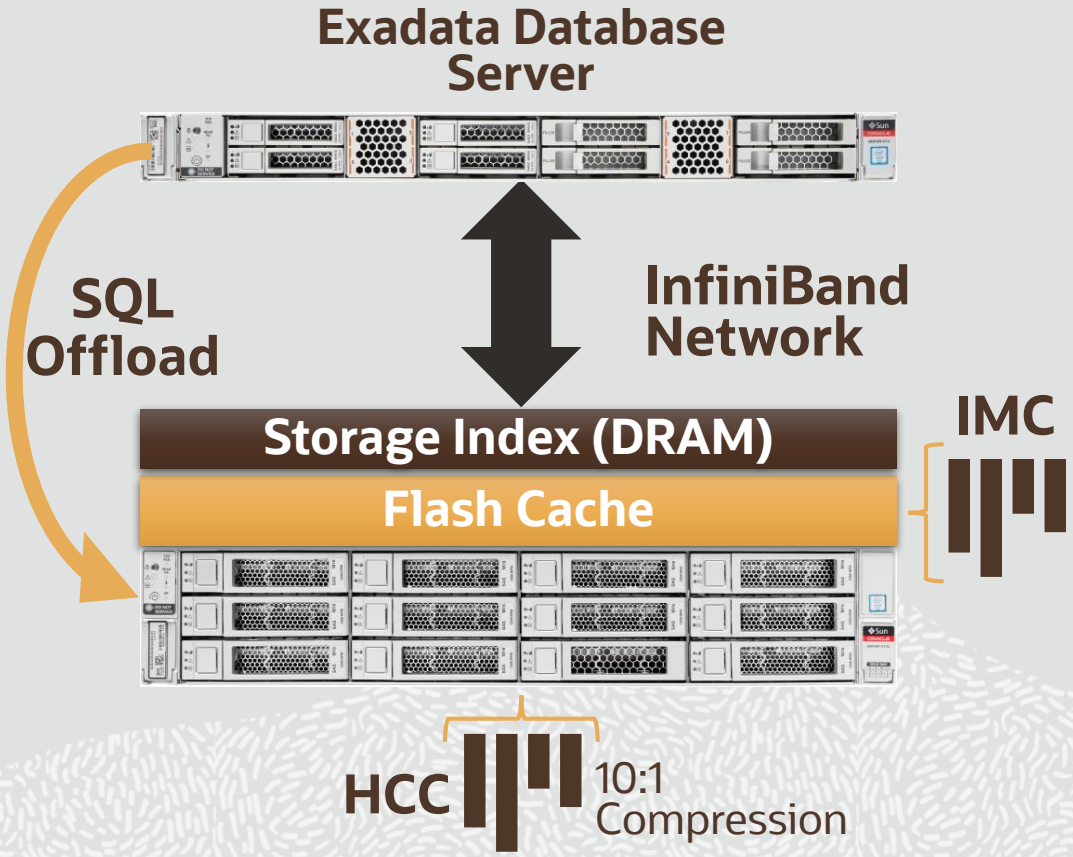
Позволяют сокращать чтения заведомо ненужных данных

Hybrid Columnar Compression (HCC)

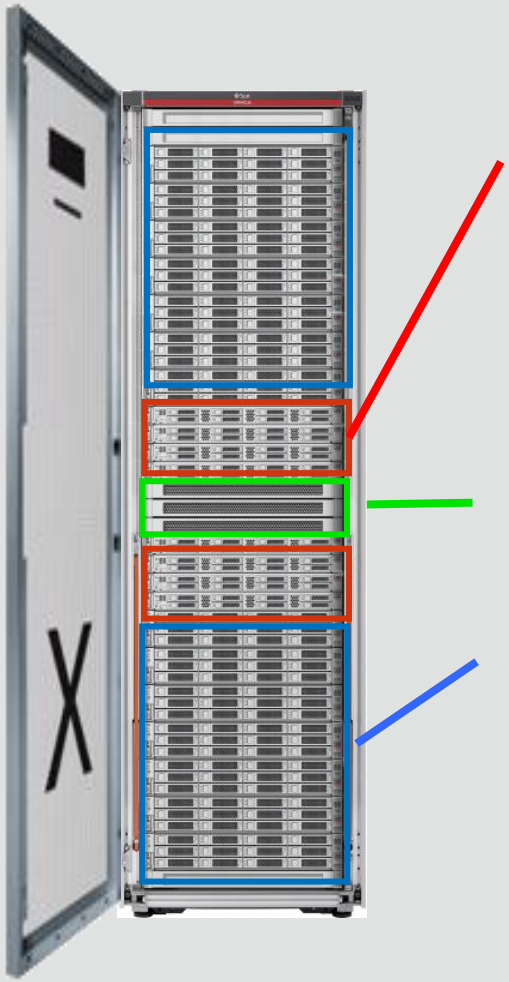
Сжатие данных в колоночном формате позволяет уменьшить объем данных на дисках, сократить ввод-вывод и ускоряет запросы.

In-Memory Columnar (IMC)

Еще больше ускоряет выполнение запросов



Exadata X8M (changes from X8 in red)



Same Scale-Out 2-Socket Database Servers

- Latest 24 core Intel Cascade Lake
- Spectre & Meltdown mitigated in silicon – no software overhead, more secure

100Gb RDMA over Converged Ethernet (RoCE) Internal Fabric

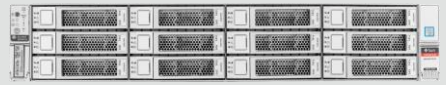
Scale-Out Intelligent 2-Socket Storage Servers

- 1536 GB Persistent Memory (PMEM) per storage server accelerates I/O, 21.5TB of PMEM per rack
- Three tiers of storage: PMEM, NVMe, HDD

Enhanced Consolidation with new KVM virtualization

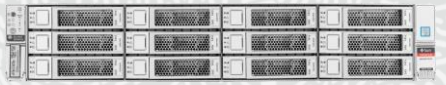
Database Server

High-Capacity (HC) Storage



Extreme Flash (EF) Storage

Extended (XT) Storage



Oracle Autonomous Database

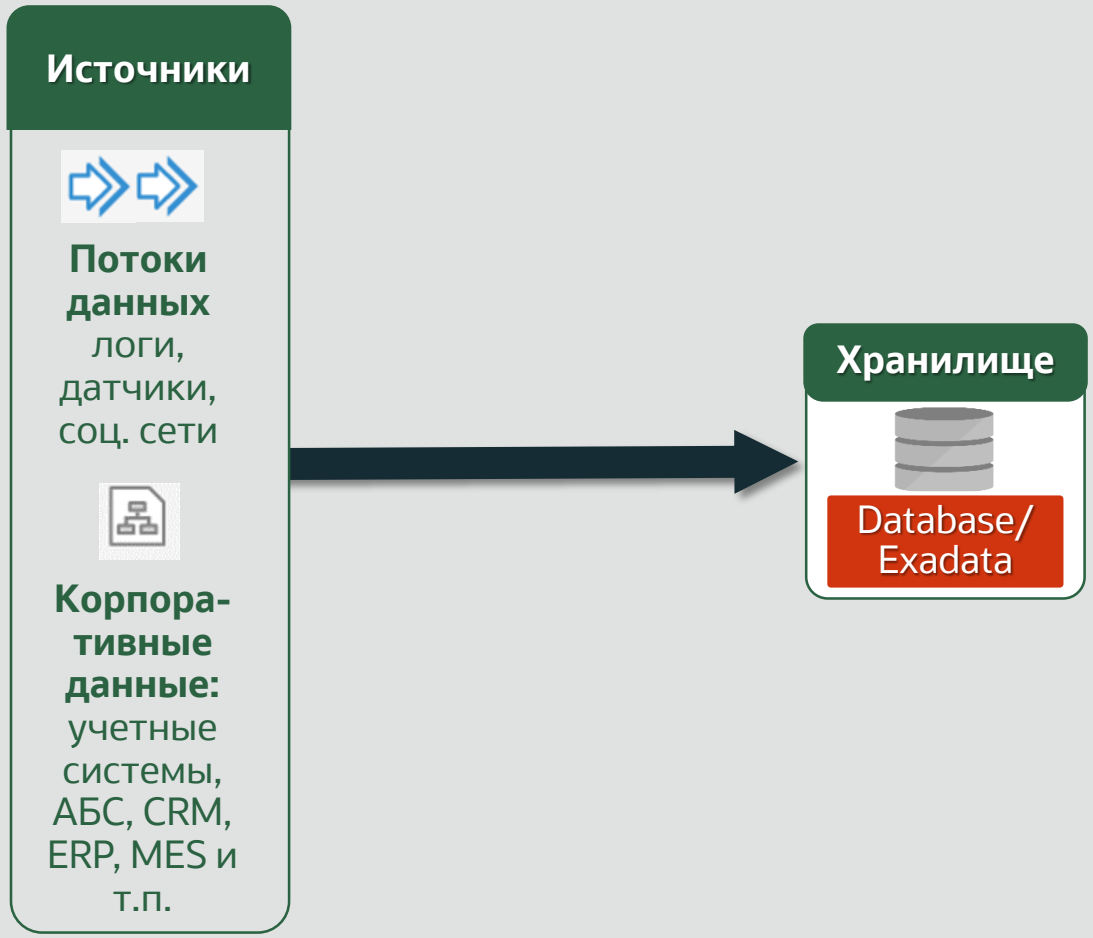
ORACLE AUTONOMOUS DATABASE =  +  + ORACLE CLOUD

Автоматизация на уровне инфраструктуры

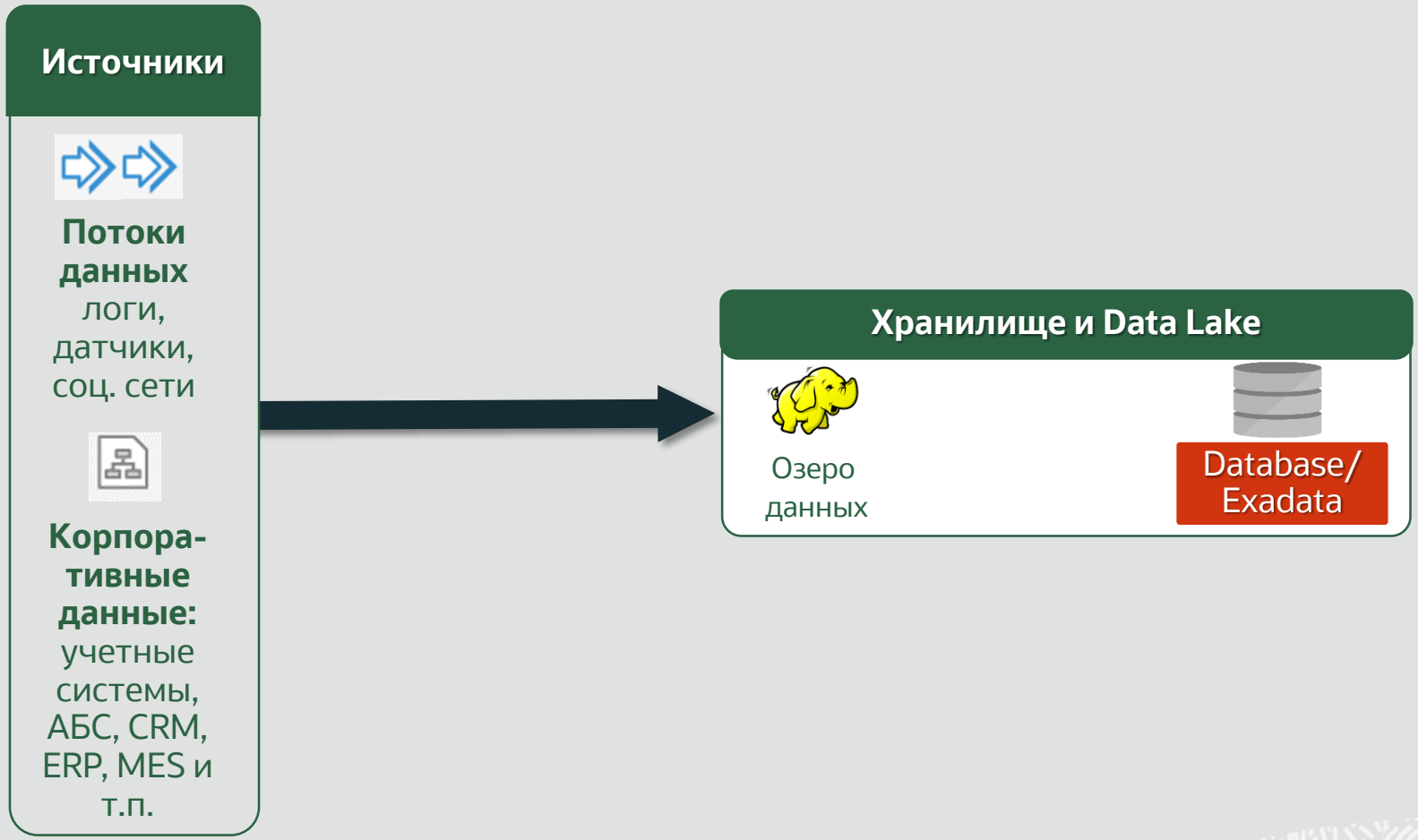
Автоматизация на уровне СУБД

Автоматизация администрирования + машинное обучение

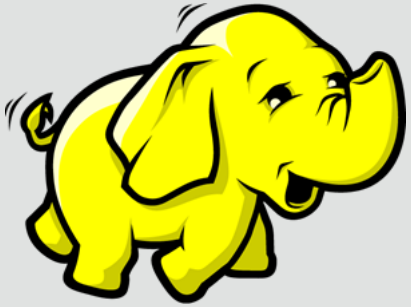
Платформа управления данными



Платформа управления данными



Hadoop

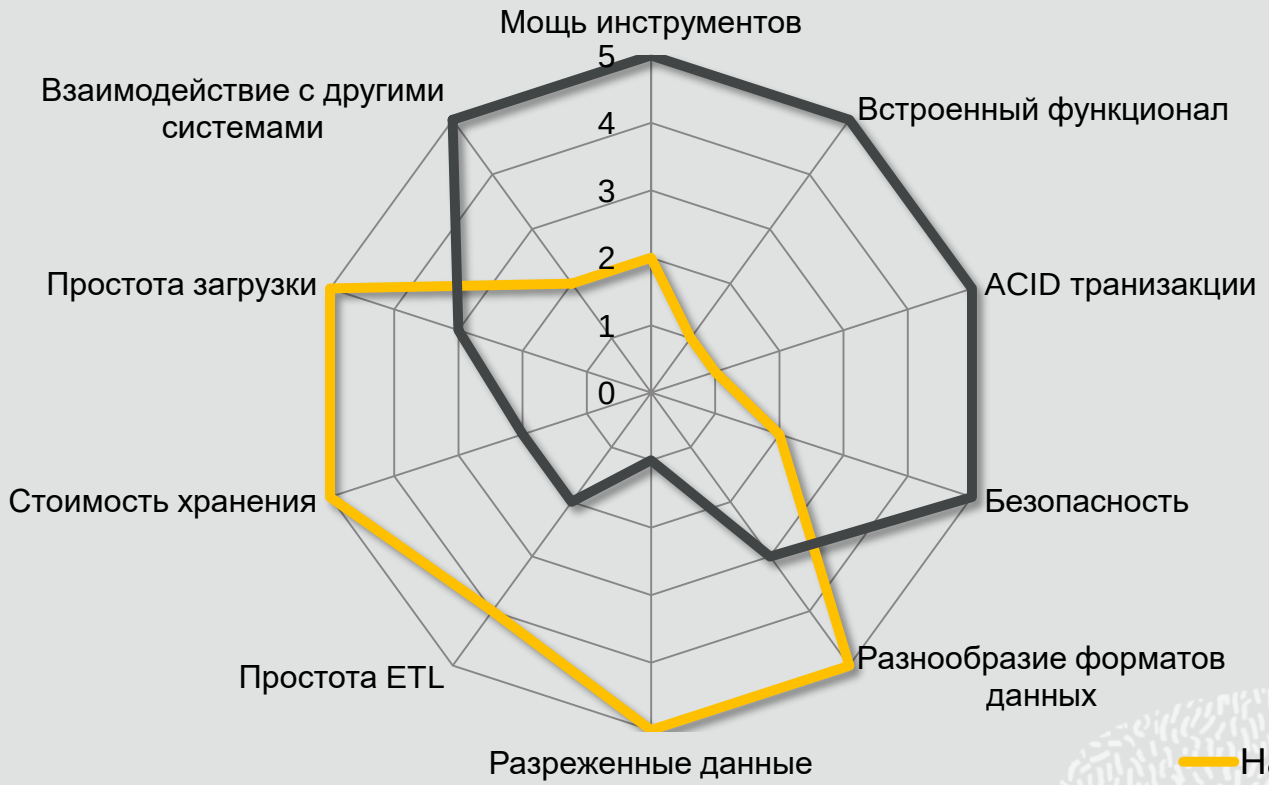


- Apache Hadoop - распределенная масштабируемая вычислительная архитектура. Не СУБД!
- Данные хранятся в распределенной файловой системе HDFS
- Данные обрабатываются при помощи разных движков (MapReduce, Spark и др.)
- Одна из самых популярных платформ для хранения и обработки больших объемов данных
- Подходит для аналитических задач
- Очень быстро развивается
- Oracle совместно с Cloudera производит программно-аппаратный комплекс для Hadoop (и Oracle NoSQL DB)

Для чего используются технологии Больших данных?

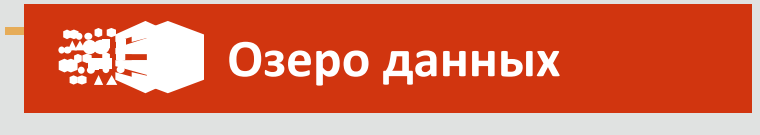
- Удешевление хранения традиционных данных из СУБД
  - Возможность дешево хранить и иметь к ним доступ, если понадобится
- Хранение и обработка полуструктурированных и неструктурированных данных
  - Аудио, Видео
  - Тексты
  - Логи
  - И т.п.
- Не нужно заранее структурировать и перекладывать данные в СУБД
  - Возможность работы непосредственно с исходными данными, например с бинарными

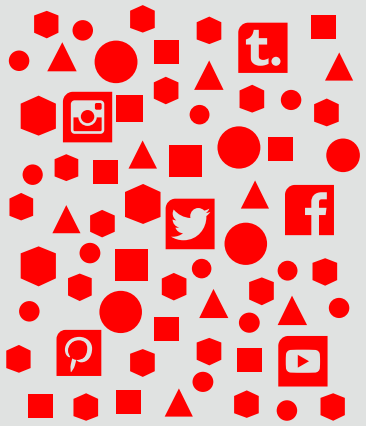
Разные подходы – разные преимущества и недостатки



- У Hadoop свои плюсы
- У СУБД свои

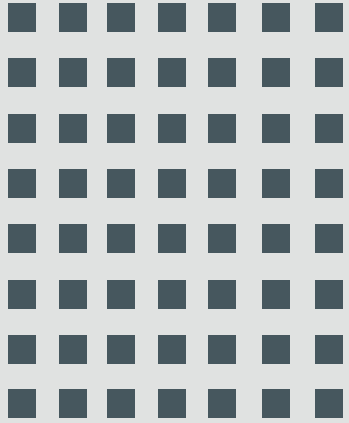
Преимущества построения систем с озером данных





Новые источники

+



Традиционные источники данных

Дешевое хранение

Только значимые и актуальные данные живут в реляционном ХД

Гибкость

В озере хранятся любые данные, не нужна предопредленная структура и модель хранения

Предварительная обработка данных на распределенном кластере

Big Data Appliance X7-2

Sun Oracle X7-2L Servers with **per server**:

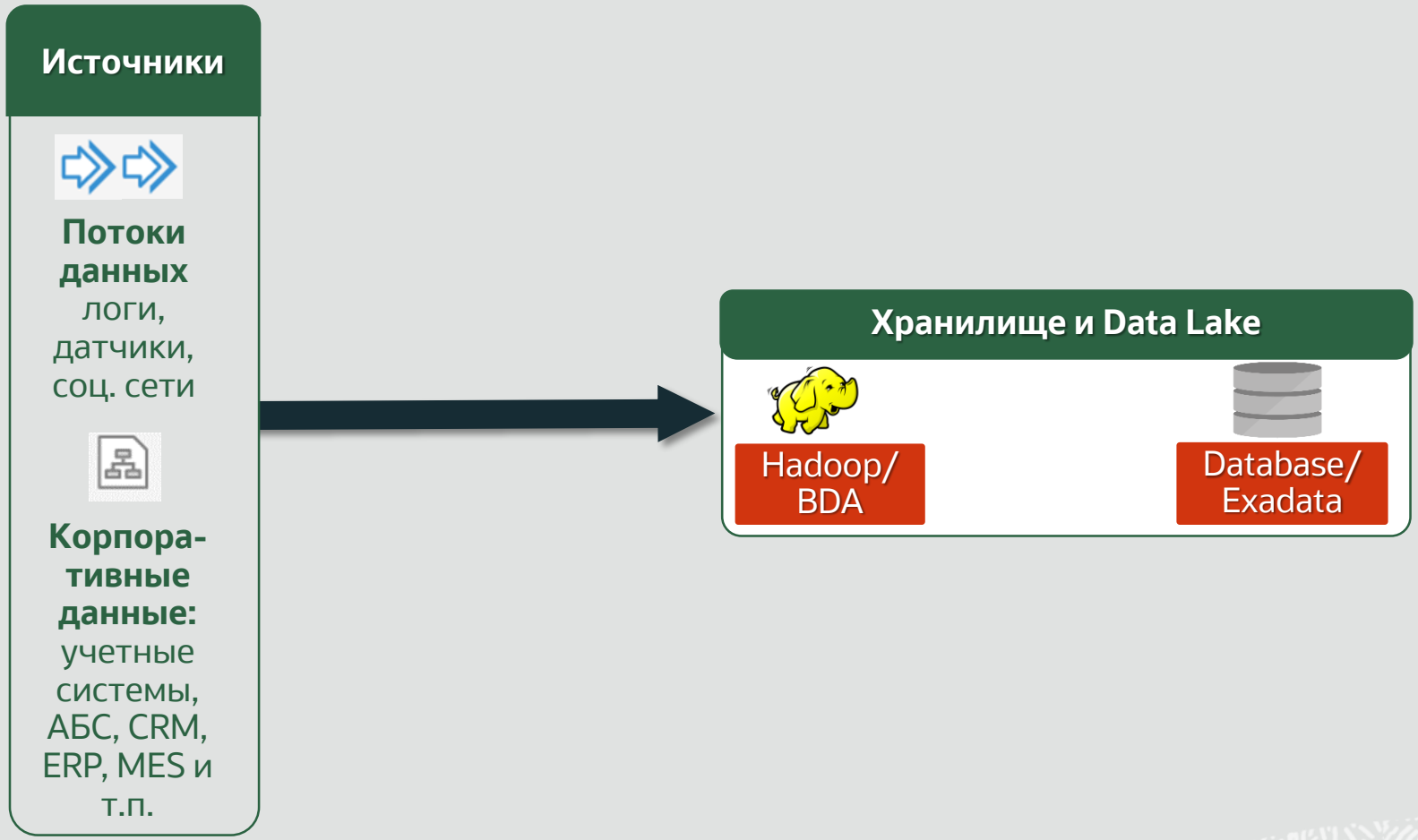
- 2 * 24 Core (2.1GHz) Intel Xeon Platinum 8160 (“Skylake”) Processors
- 256 GB DDR4-2666 Memory, expandable to 1.5TB
- 12 * 10TB, 7,200 RPM Disks for a total of 120TB storage
- 40Gb/sec InfiniBand Internal Network, 10Gb Ethernet External

Included Software (4.10):

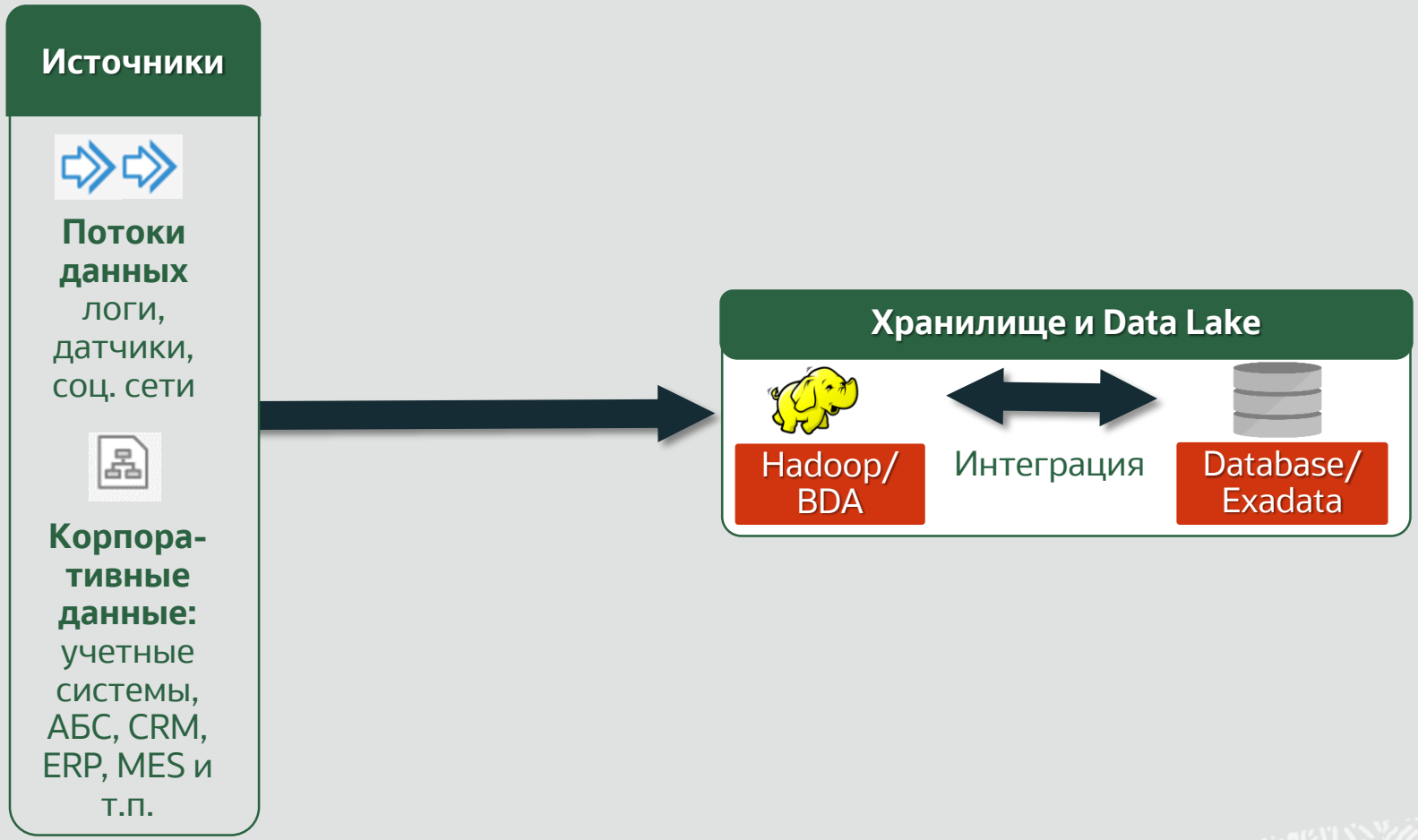
- Oracle Linux 6.x & Oracle Linux 7.x for Edge Nodes
- Oracle Big Data SQL*
- Cloudera Distribution of Apache Hadoop 5.12.1 – EDH Edition
- Cloudera Manager 5.12.1
- Oracle R Distribution
- Oracle NoSQL Database CE

* Oracle Big Data SQL is separately licensed

Платформа управления данными



Платформа управления данными



Недостатки существующих систем Больших данных

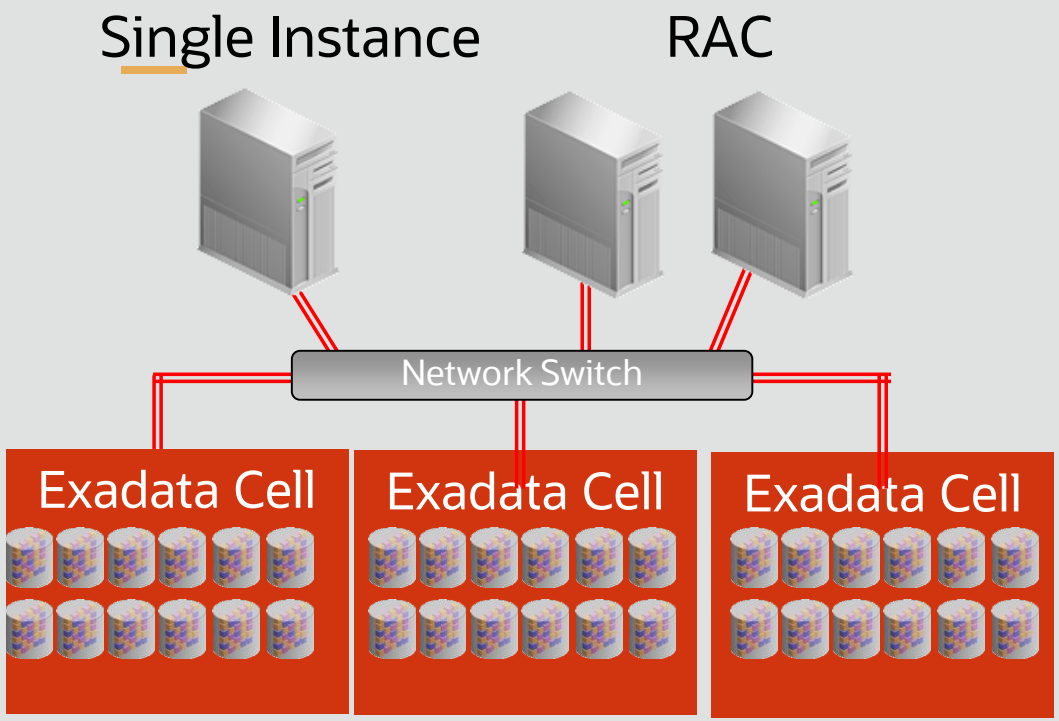
- Для работы с Hadoop и реляционными базами данных требуются разные навыки
- Существующие механизмы доступа к данным в Hadoop функционально ограничены или работают медленно
- Конечные пользователи используют разные инструменты для работы с Hadoop и реляционными базами

Apache **Hive**

Apache Hive

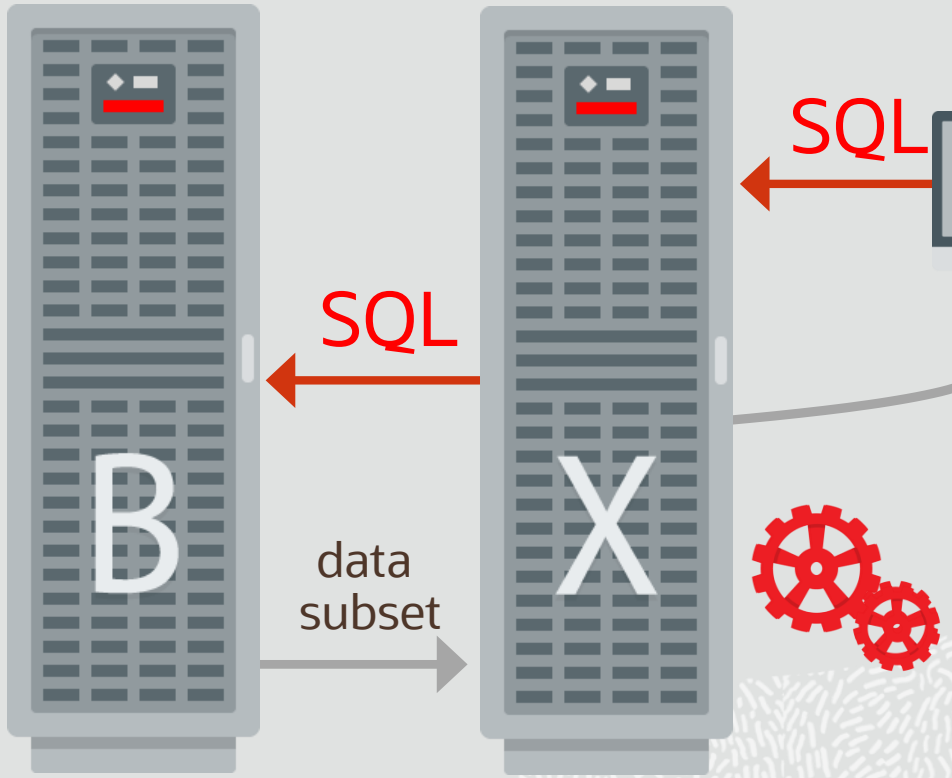
- Один из самых популярных проектов для обработки данных над Hadoop
- Инфраструктура, эмулирующая реляционную СУБД над Hadoop
- Есть SQL-подобный язык HiveQL
- Позволяет строить аналог сверхбольших хранилищ данных в Hadoop

Архитектура Exadata



- Каждая ячейка Exadata – самостоятельный сервер с установленными дисками и ПО Exadata
- Данные «размазаны» между многими ячейками Exadata
- Нет ограничения на количество ячеек в системе
- Ячейки выполняют множество операций, которые в традиционной архитектуре делает Oracle
- Ячейки работают в режиме MPP

Как работает Oracle Big Data SQL



Оффлоадинг запросов на Data Nodes

Hadoop & NoSQL

Oracle Database

На сервер базы данных быстро загружается небольшое подмножество данных

Оффлоадинг запросов на Exadata Storage Servers

Задачи, которые решает Big Data SQL

1. Использование единых метаданных и инструментов для работы с разными источниками

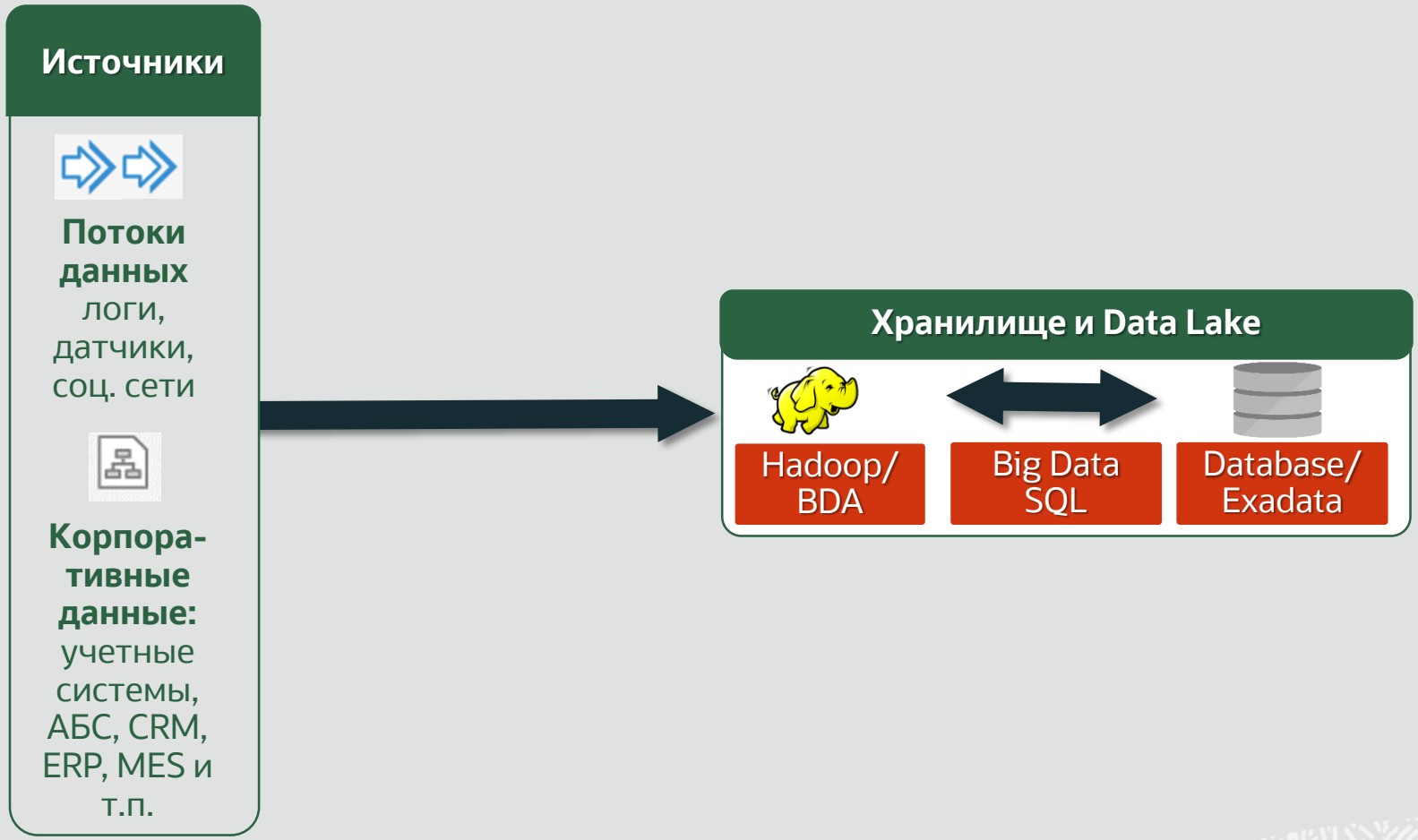
   Не нужны разные квалификации в зависимости от источника данных

2. Увеличение производительности запросов при работе с большими объемами

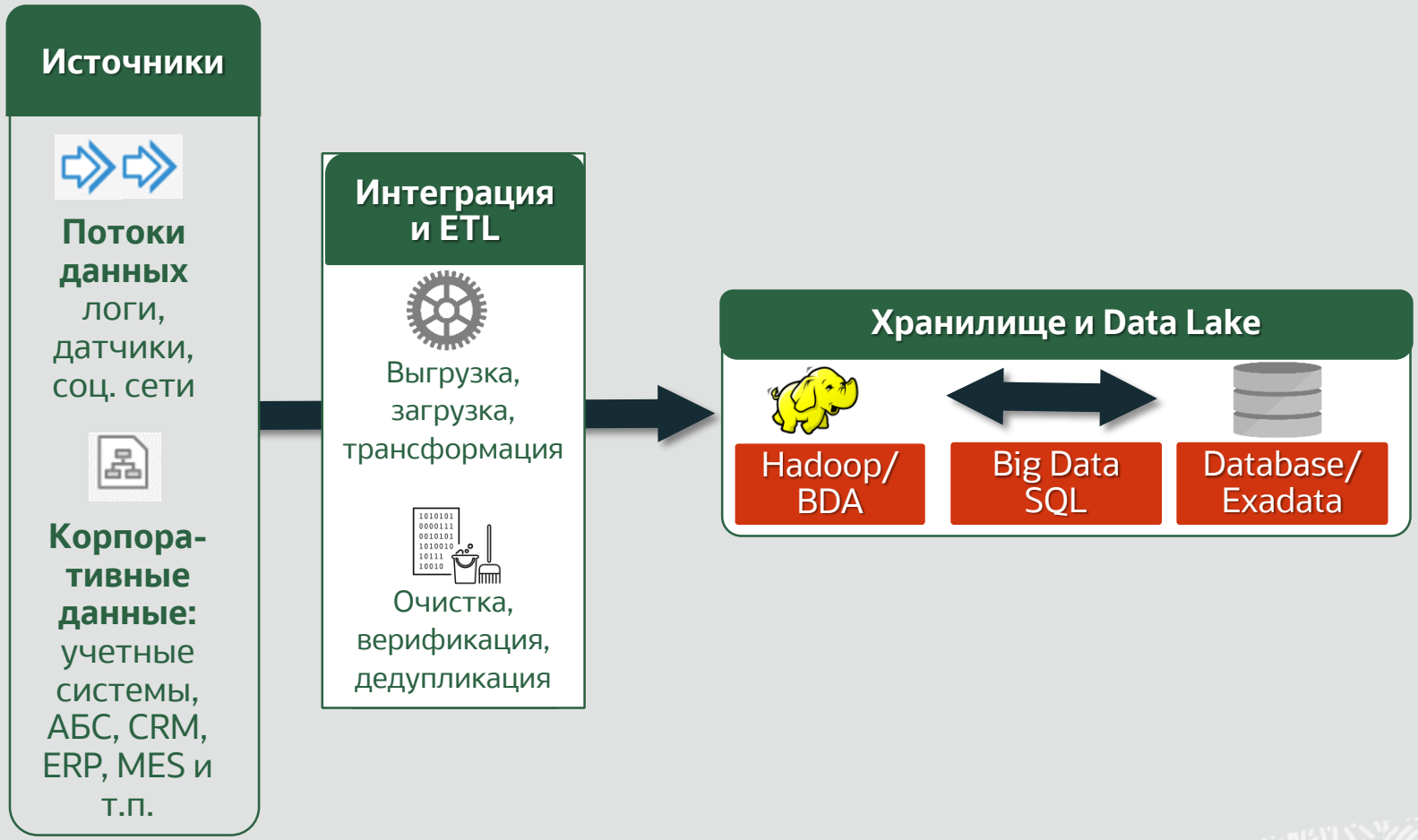
   Используя наработки Exadata

3. Использование возможностей Oracle SQL для работы с данными
4. Возможность использования данных Oracle, Hadoop, NoSQL, Kafka в одном SQL запросе
5. Безопасность

Платформа управления данными



Платформа управления данными

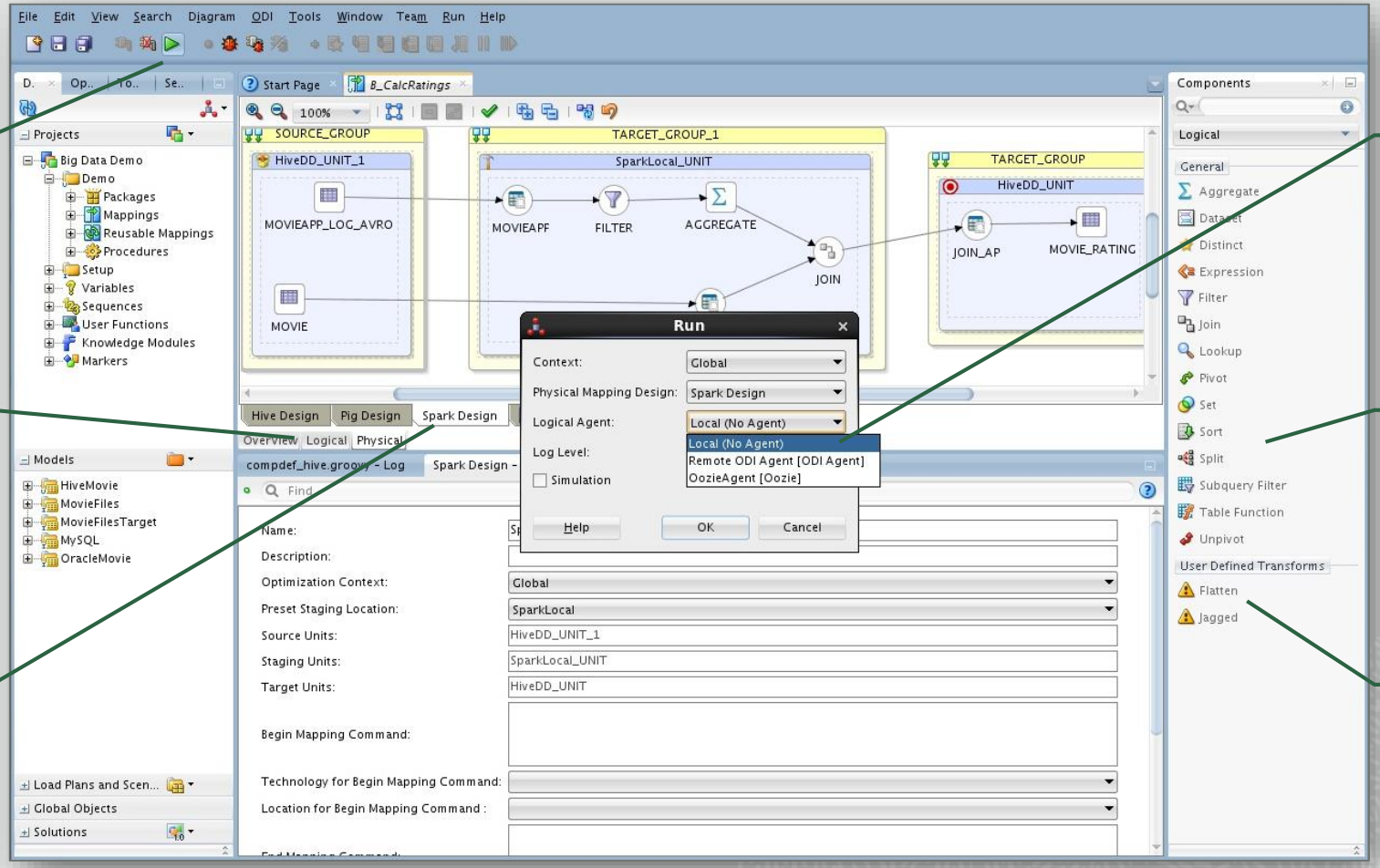


Oracle Data Integrator

Не требуется отдельный ETL сервер

Логический и физический дизайн разделены

Физическое выполнение кода SQL, Hive, Pig, Spark



Использование Oozie или ODI Java Agent

Библиотека операторов

Возмоно определять свои функции

Модули знаний

Проще физическое проектирование и короче время реализации

Изменяемая архитектура модулей знаний

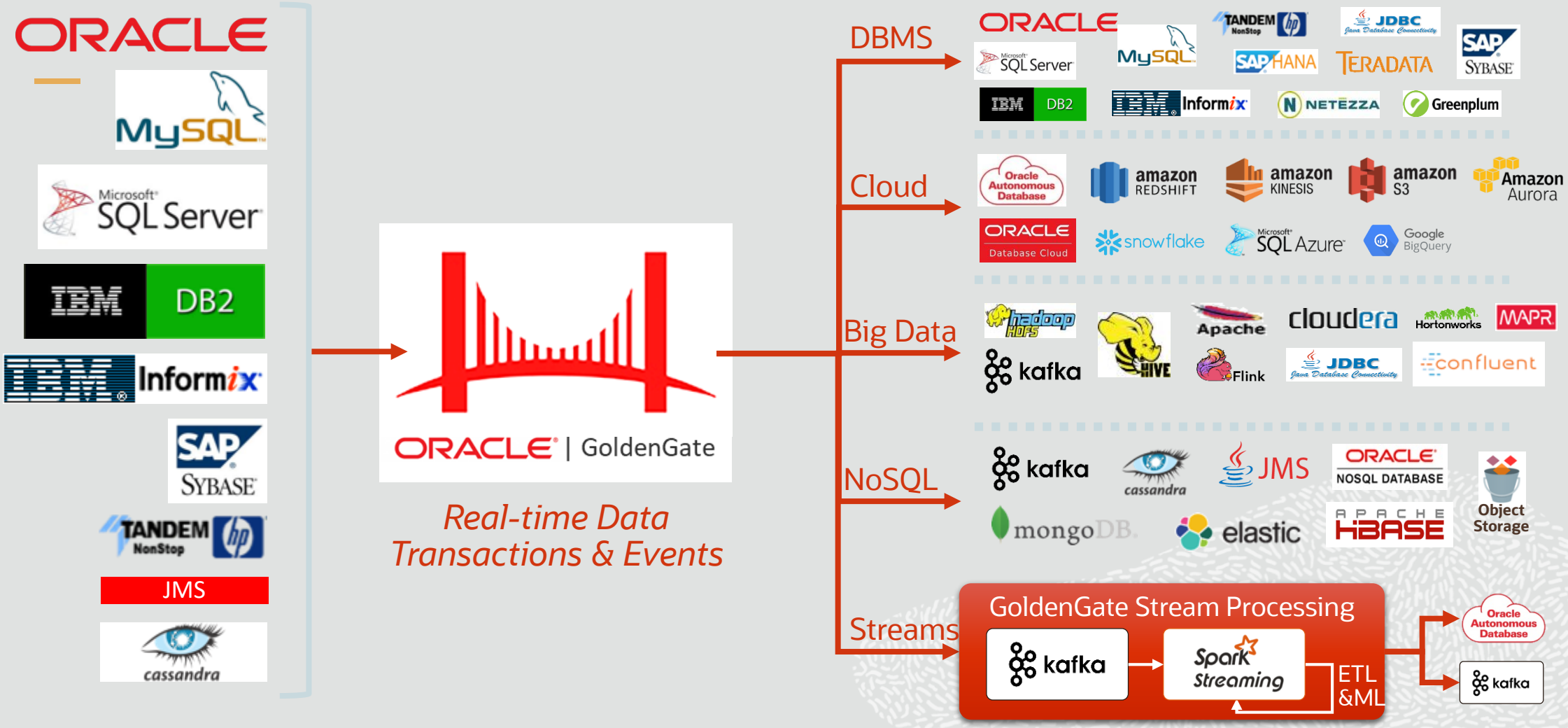
Примеры встроенных модулей знаний:

Oracle	Sqoop	Hive	HBase	Oracle Merge	SAP ERP
SAP BW	Oracle Datapump	Oracle DBLink	JMS	External Tables	Teradata
Oracle Spatial	Siebel	eBusiness Suite	IBM DB2	Netezza	DBaaS

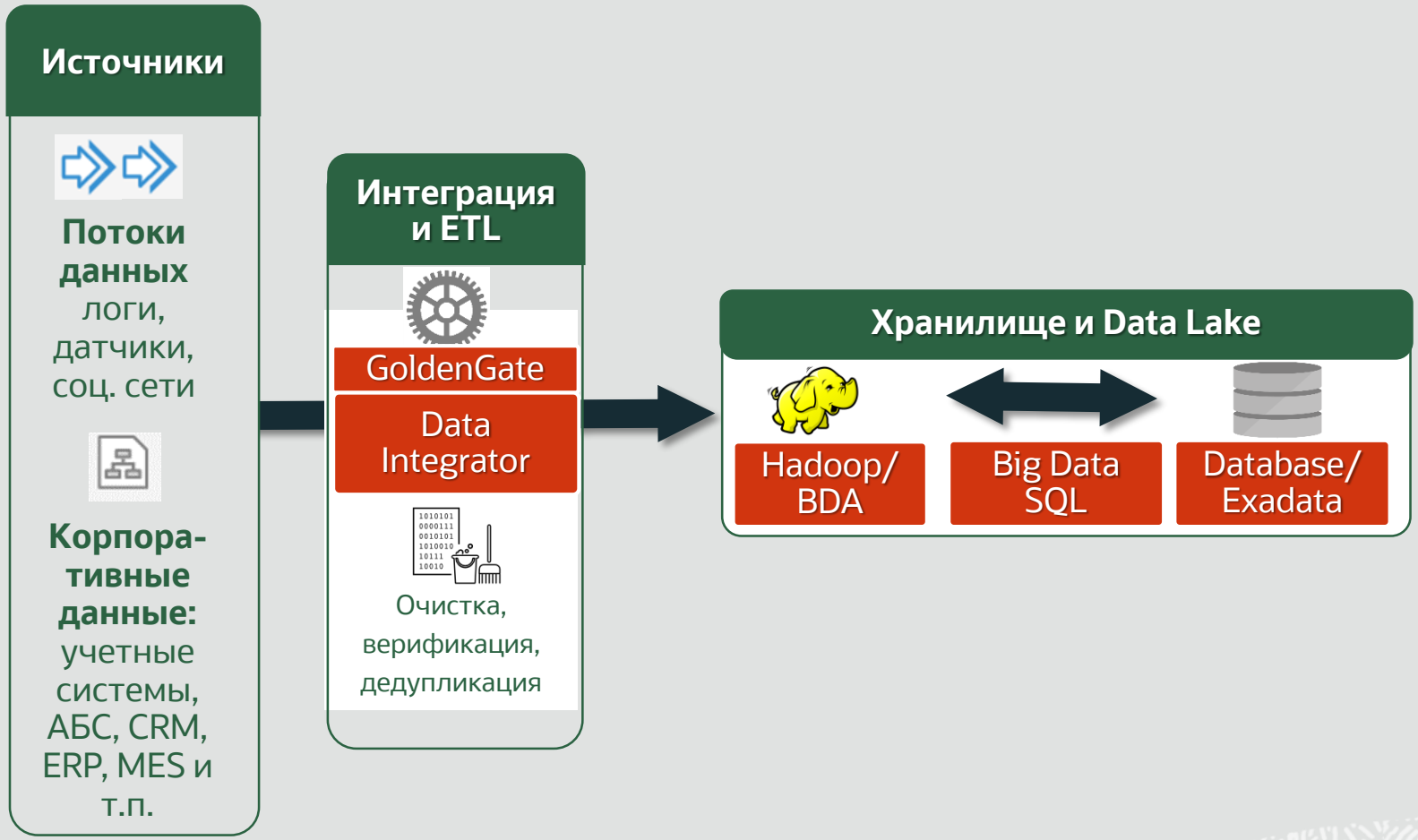
Ключевые преимущества:

- Ускорение разработки и упрощение обслуживания с использованием шаблонов
- Легко расширить и добавить новые лучшие практики
- Обеспечивает предсказуемость и снижает стоимость владения

Oracle GoldenGate | Платформа для репликации данных



Платформа управления данными



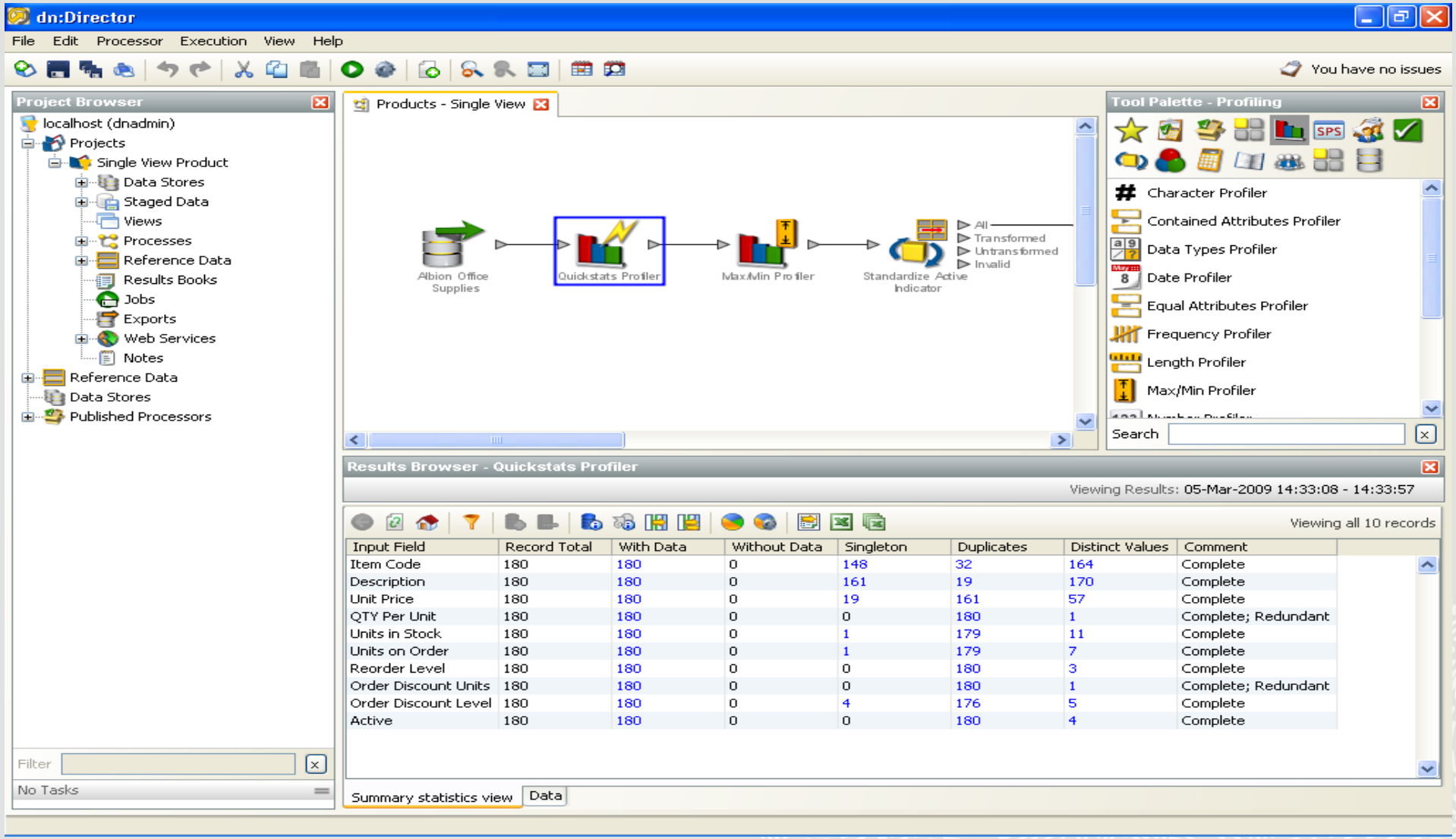
Oracle Enterprise Data Quality

- Простой GUI, без кодирования
- Коллективная работа
- Разработано для бизнес-пользователей

- Полностью настраиваемые правила
- Никаких «черных ящиков»
- Высокая производительность

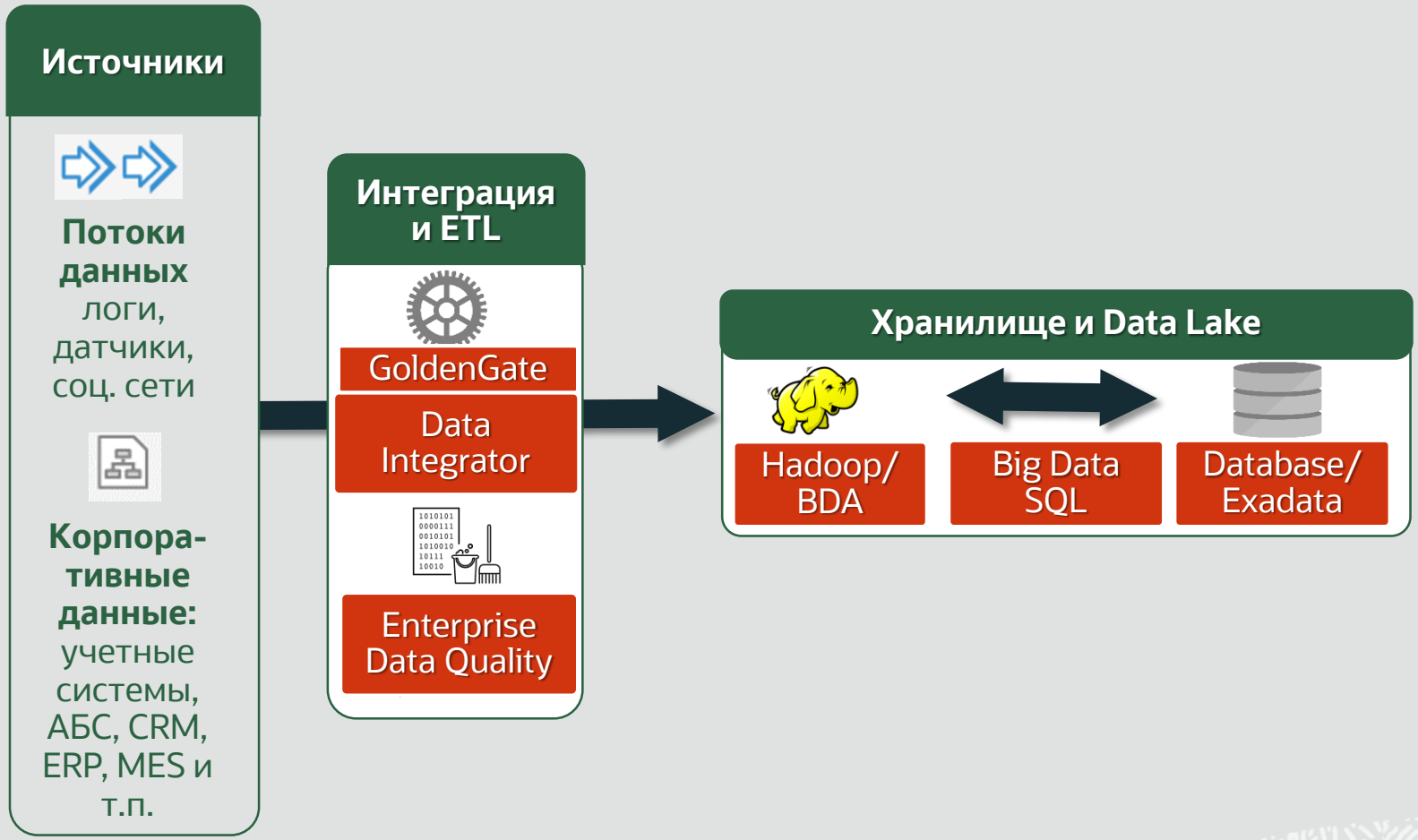
- Интегрированное решение
- Быстрое внедрение и интеграция
- Низкие затраты на обслуживание

Основной интерфейс EDQ

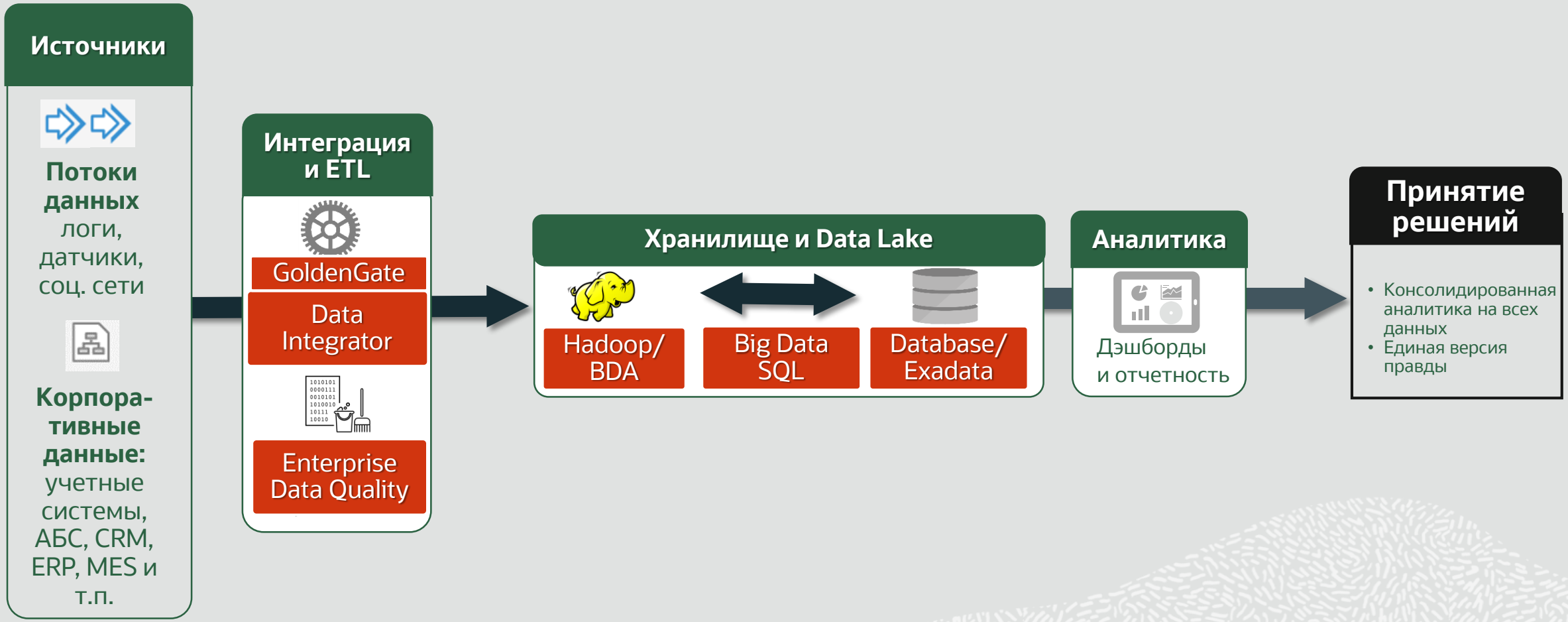


Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values	Comment
Item Code	180	180	0	148	32	164	Complete
Description	180	180	0	161	19	170	Complete
Unit Price	180	180	0	19	161	57	Complete
QTY Per Unit	180	180	0	0	180	1	Complete; Redundant
Units in Stock	180	180	0	1	179	11	Complete
Units on Order	180	180	0	1	179	7	Complete
Reorder Level	180	180	0	0	180	3	Complete
Order Discount Units	180	180	0	0	180	1	Complete; Redundant
Order Discount Level	180	180	0	4	176	5	Complete
Active	180	180	0	0	180	4	Complete

Платформа управления данными

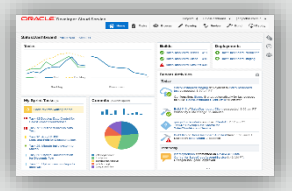


Платформа управления данными



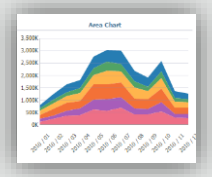
Основные возможности Oracle BI и DV

Classic BI



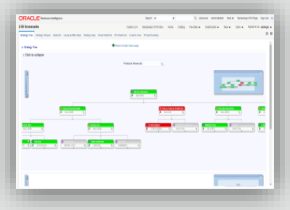
Информационные панели

Пространственная аналитика



Интерактивные отчеты

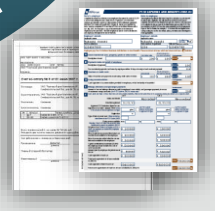
Рассылки, мониторинг



Ключевые Показатели Эффективности

Мобильная аналитика

Интеграция с MS Office

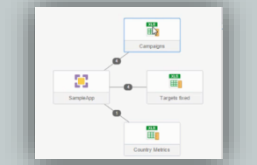


Регламентированные формы

Data Visualization

Data Visualization

История анализа

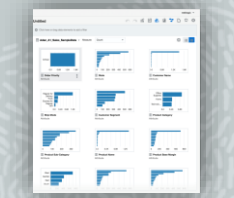


Внешние источники

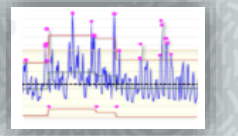
Подготовка данных

DV Mobile

Аналитика по поисковому запросу

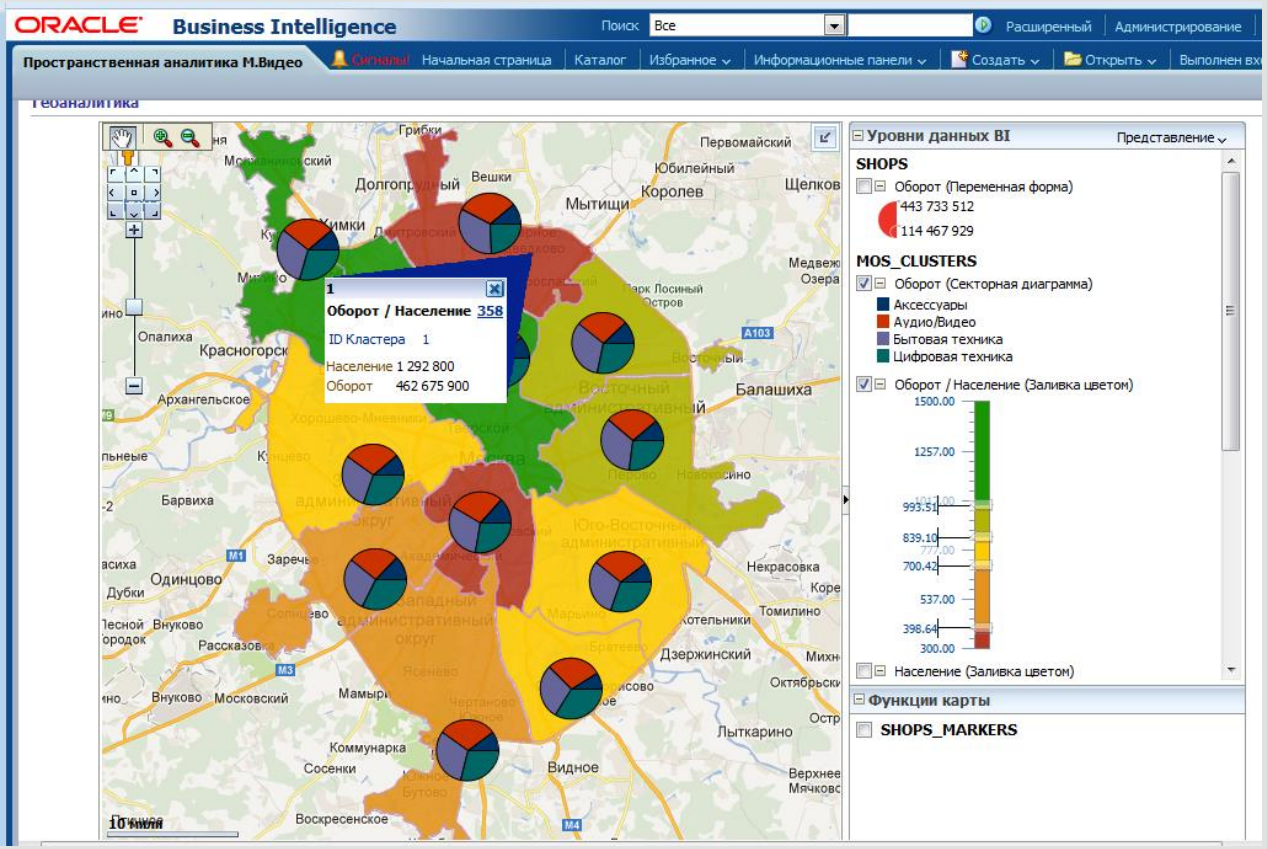


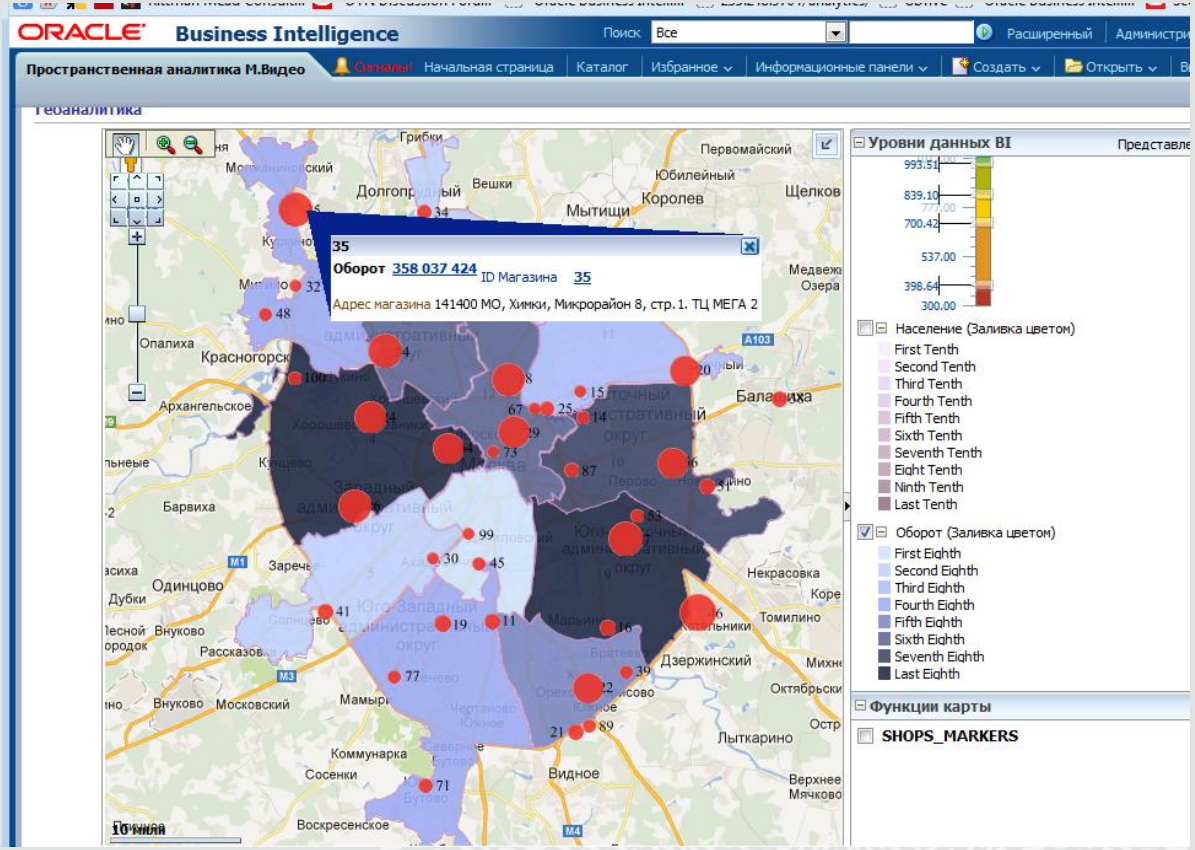
Профилирование



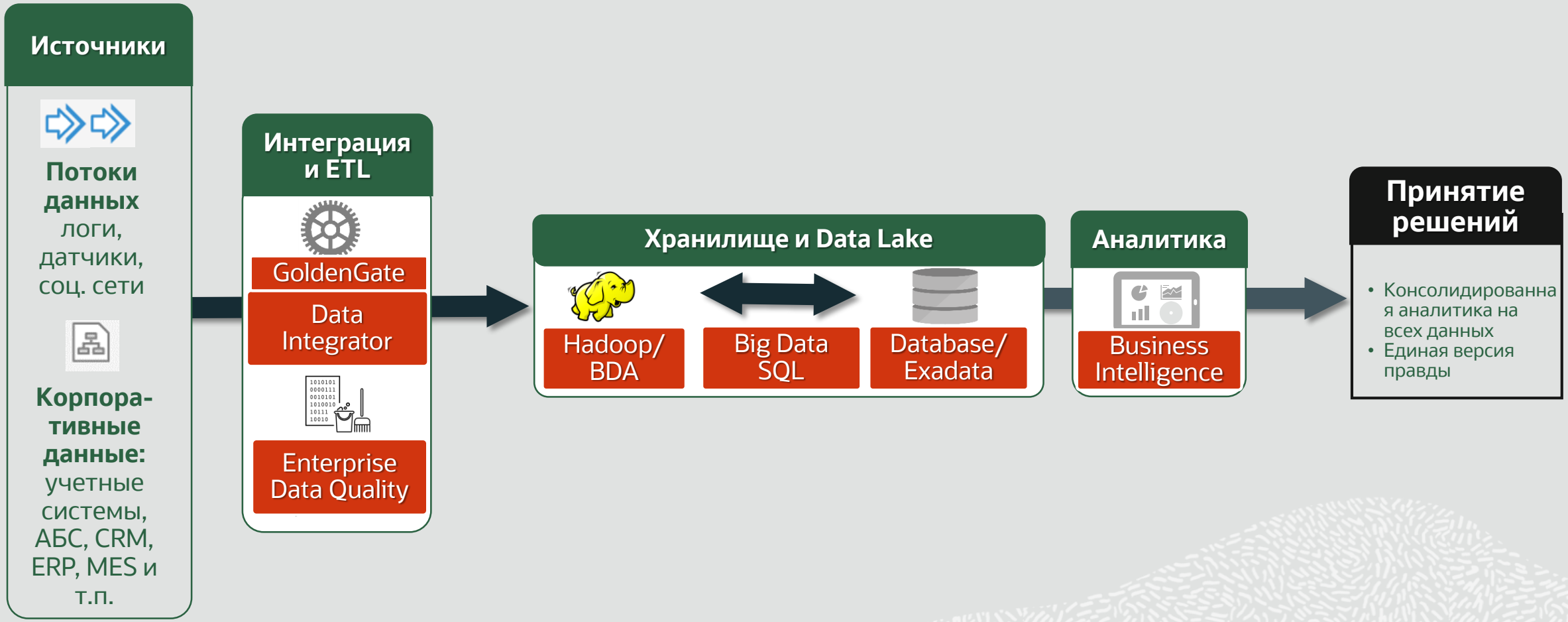
Продвинутая аналитика

Геоаналитика в BI

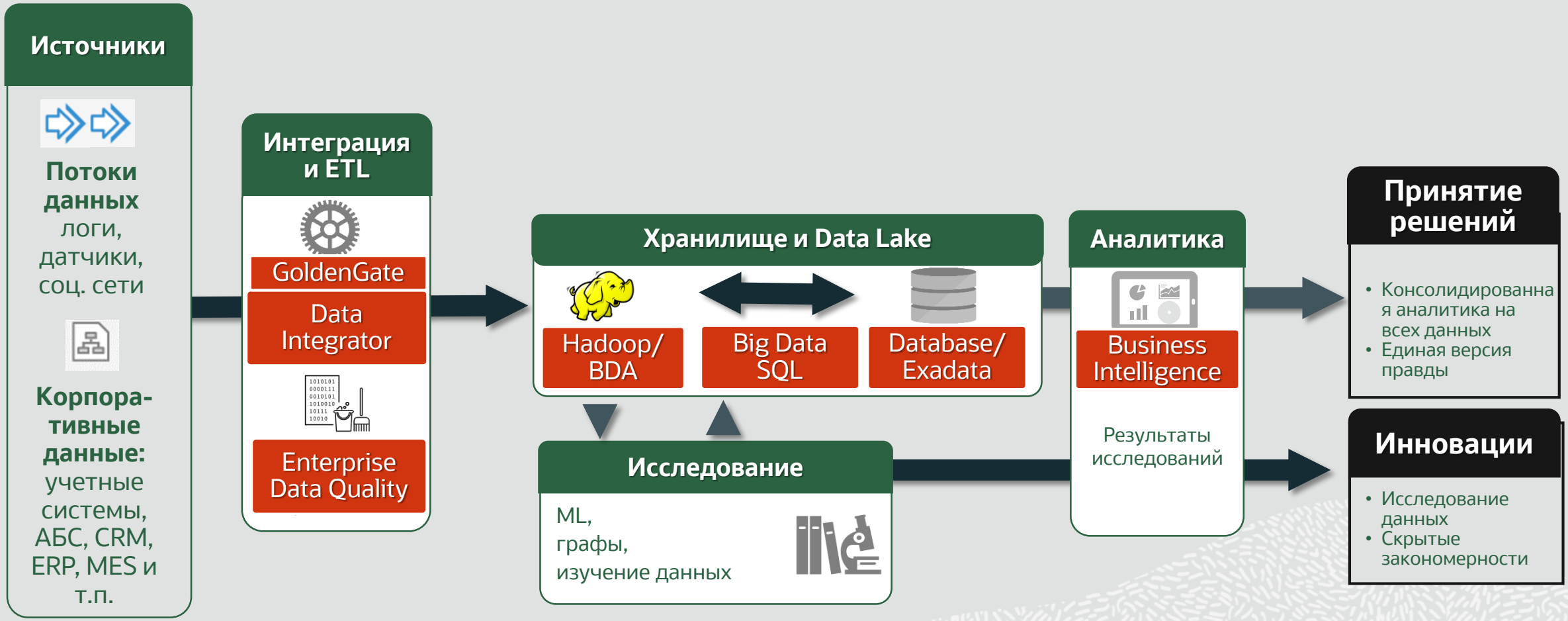




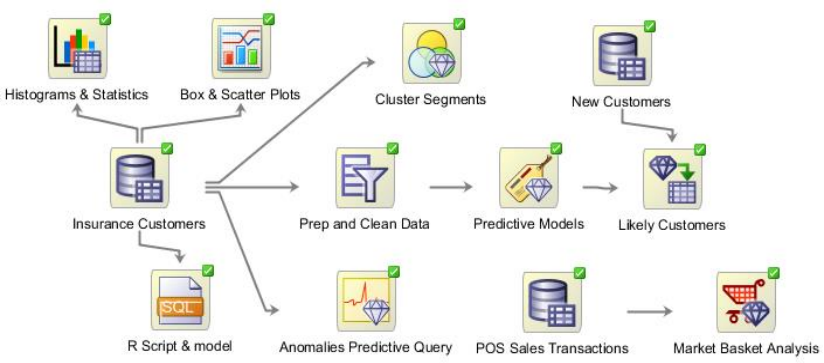
Платформа управления данными



Платформа управления данными



Advanced Analytics option for Oracle Database
Возможности для ML внутри базы данных

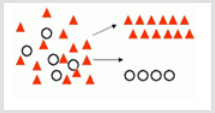


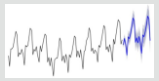
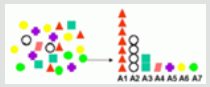
Data Mining

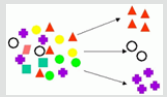
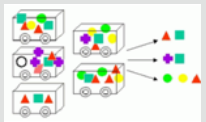
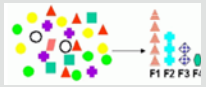
- Графический интерфейс на платформе SQL Developer
- Готовые инструменты для подготовки данных, построения и применения моделей ML
- Расчет полностью внутри БД Oracle

Advanced Analytics: Oracle Data Mining

Основные алгоритмы «из коробки»



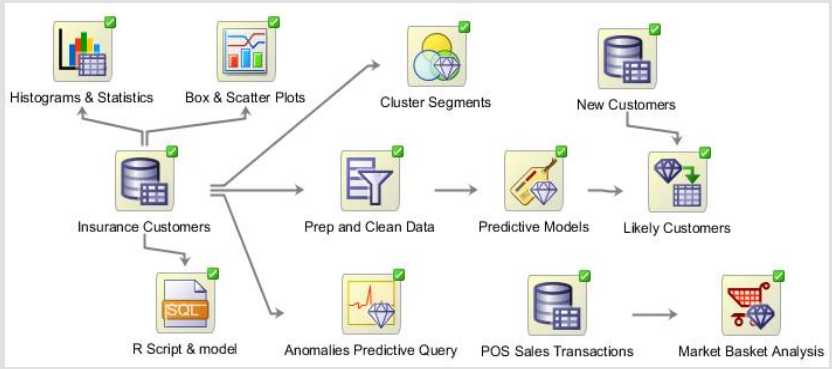



Техника	Применимость	Алгоритмы
Классификация	Прогнозирование принадлежности к классу (дискретной величины). Например, прогноз оттока клиентов, факта поломки по показаниям датчиков.	Naïve Bayes; Logistic Regression (GLM); Decision Tree; Random Forest Neural Network ; Support Vector Machine; Explicit Semantic Analysis
Регрессия	Прогнозирование численной величины. Например, lifetime value, house value.	Linear Model; Generalized Linear Model; Support Vector Machine (SVM); Stepwise Linear regression Neural Network
Временные ряды	Прогнозирование численных значений на временной оси, учитывая тренды и сезонность. Например, прогноз выручки, продаж в течении дня, месяца, года.	Holt-Winters, Regular & Irregular, with and w/o trends & seasonal Single, Double Exp Smoothing
Важность атрибута	Ранжирует атрибуты по влиянию на целевой атрибут. Например, поиск фактора, который влияет на положительный отклик на предложение	Minimum Description Length Principal Comp Analysis (PCA) Unsupervised Pair-wise KL Div
Обнаружение аномалий	Выявляет необычные и подозрительные случаи на основе их отклонения от нормы. Например, обнаружение мошенничества в страховании, уплате налогов	One-Class Support Vector Machine
Кластеризация	Для исследования данных и обнаружения естественных групп. Например, построение клиентских сегментов.	Hierarchical K-Means Hierarchical O-Cluster Expectation Maximization (EM)
Ассоциативные правила	Правила, отражающие часто совместно встречающиеся события. Анализ покупательской корзины, кросс-продажи, размещение товаров в магазинах.	A priori/ market basket
Выделение признаков	Создание новых атрибутов как линейных комбинаций существующих. В т.ч. для анализа текстов, семантического анализа, распознавания образов	Principal Comp Analysis (PCA) Non-negative Matrix Factorization Singular Value Decomposition (SVD) Explicit Semantic Analysis (ESA)

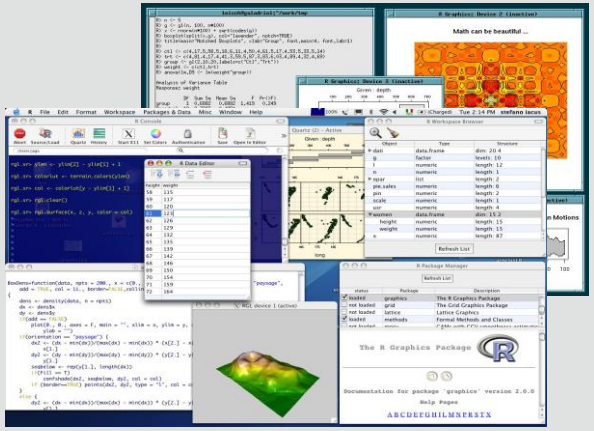
Advanced Analytics option for Oracle Database

Возможности для ML внутри базы данных



Data Mining

- Графический интерфейс на платформе SQL Developer
- Готовые инструменты для подготовки данных, построения и применения моделей ML
- Расчет полностью внутри БД Oracle



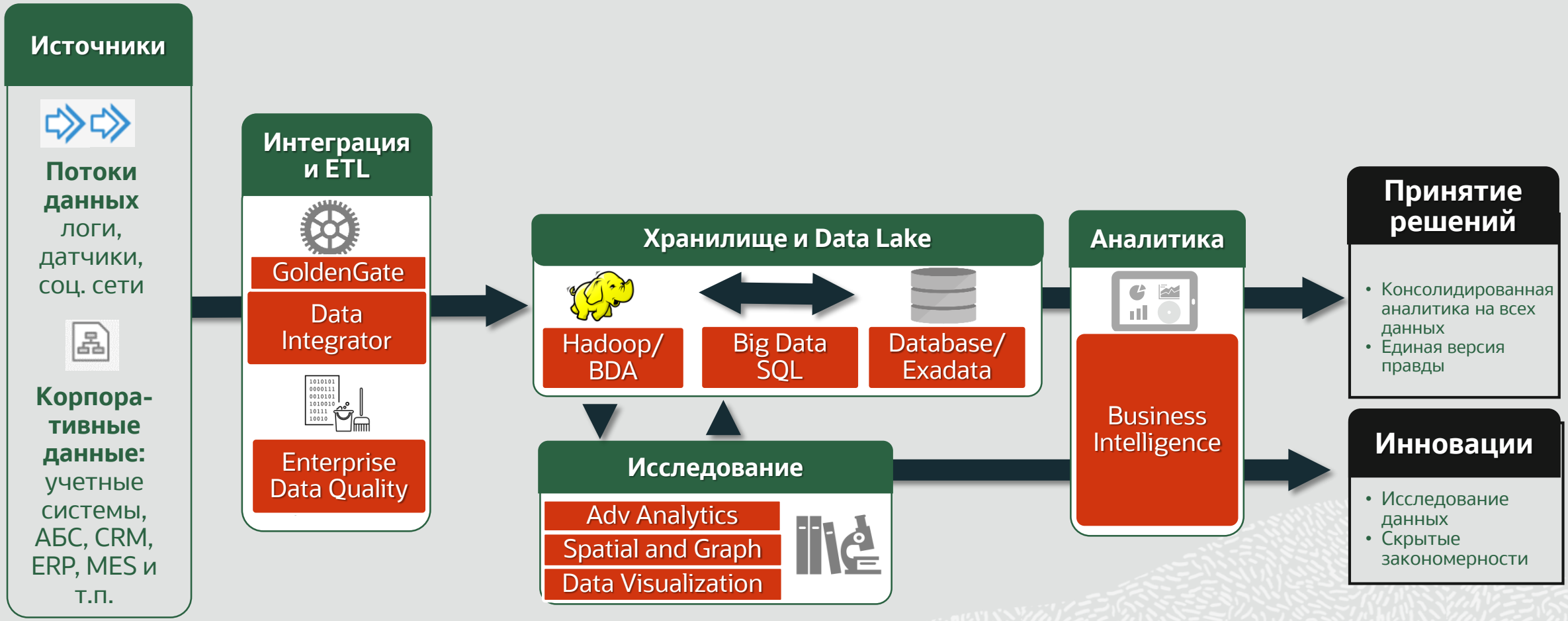
R Enterprise

- R – популярная open-source среда и язык статистической обработки данных
- >3000 готовых пакетов для аналитики
- Продвинутые возможности визуализации и работы с данными
- R интегрирован с БД Oracle

OML4Py

- Python для задач машинного обучения
- Интеграция Python с БД Oracle, аналогично R
- Скоро будет доступен для скачивания на OTN

Платформа управления данными

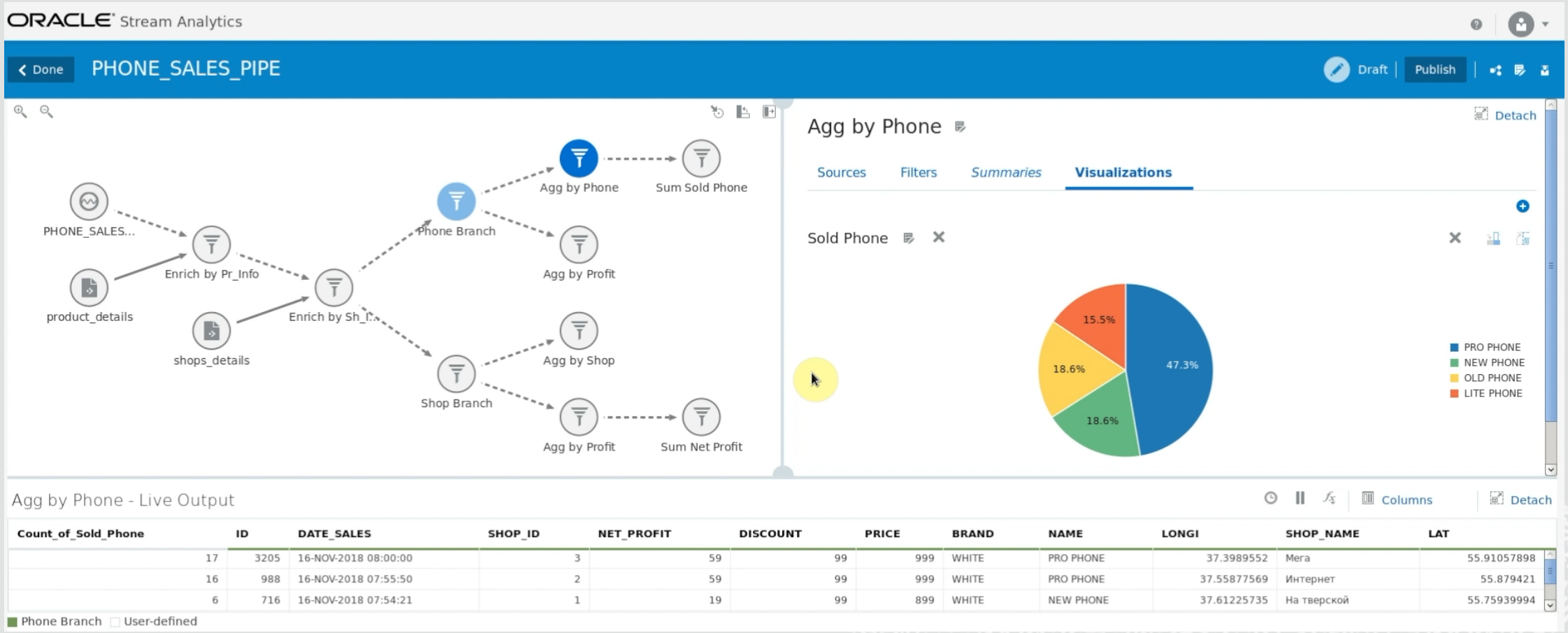


Платформа управления данными

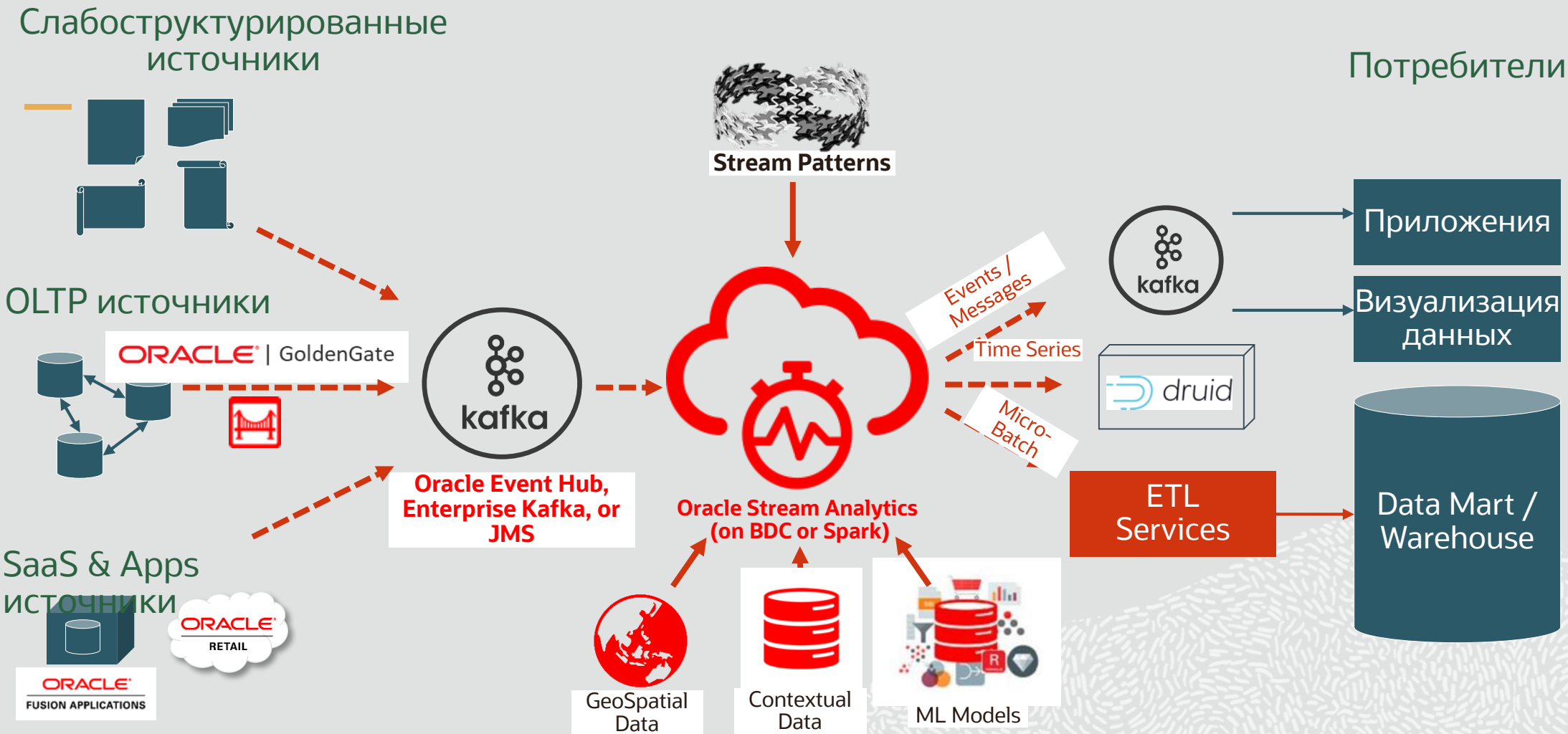
Oracle Stream Analytics

Обработка и анализ событий в реальном времени

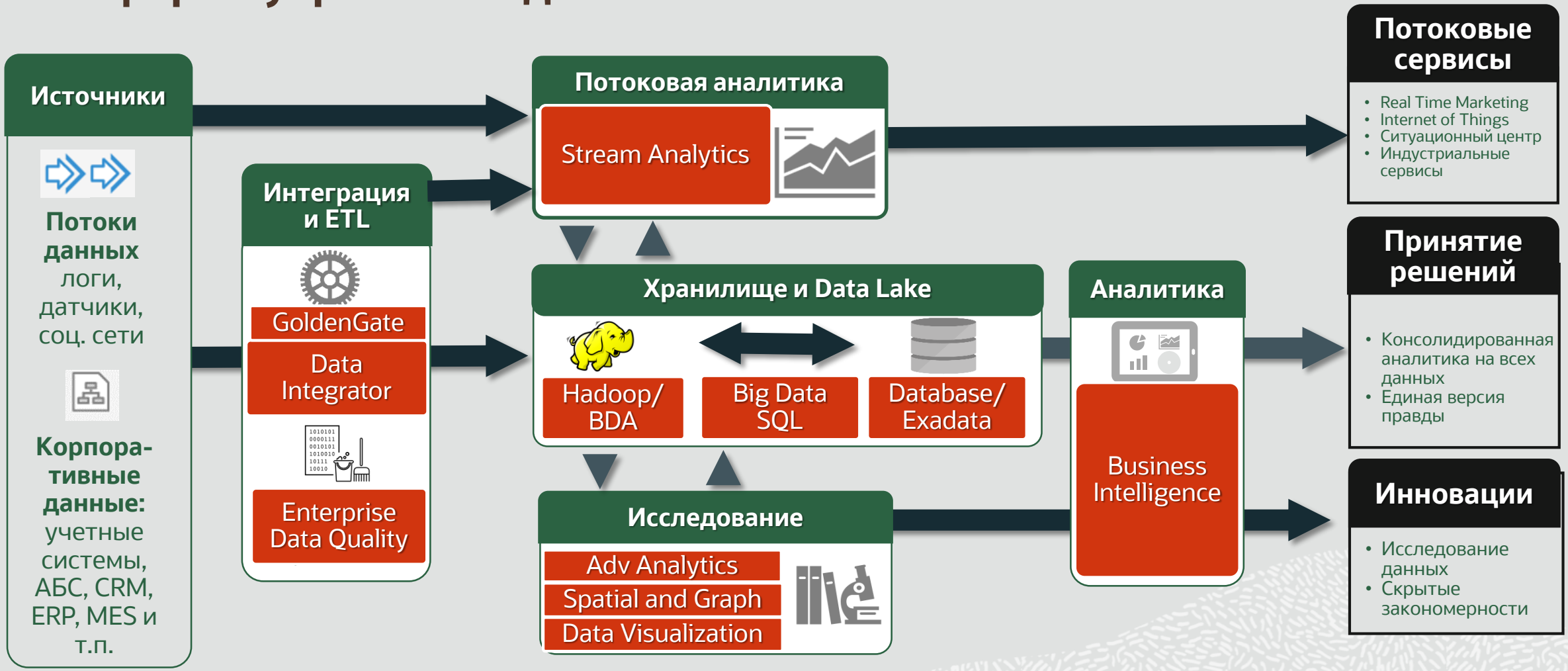
Oracle Stream Analytics



Принцип работы Oracle Steam Analytics



Платформа управления данными



Always Free – What's Included

Autonomous Database

2 x Databases
20 GB each

Compute

2 x VMs
1 GB Memory each

Storage

100 GB Block
10 GB Object
10 GB Archive

Networking/ Load Balancing

10 Mbps LB
10 TB Outbound Data Transfer

Monitoring / Notifications

500M Metrics Ingestion
1B Metrics Retrieval
1M Notifications
1K Emails

Available to All New and Existing Cloud Accounts

ДНИ ORACLE В МОСКВЕ

Oracle Modern **Cloud Day**

30 октября 2019

#OracleMCD

ПАРАЛЛЕЛЬНЫЕ СЕССИИ

Сессия 2. Аналитика, Большие Данные и Машинное Обучение	
13:30	**Вступление, обзор Oracle Data Management Platform** Андрей Пивоваров, Руководитель группы перспективных технологий предпроектного консалтинга Oracle СНГ
14:00	**Новинки бизнес аналитики и визуализации данных** Павел Дубинин, Ведущий консультант Oracle
14:30	**Создаём платформу Data Science - as -Service на технологиях Oracle** Олег Сиротюк, Ведущий консультант Oracle
14:55	**Возможности Oracle Exadata для машинного обучения** Алексей Пылкин, Ведущий консультант Oracle
15:10	**High Performance Computing от Oracle: сценарии использования и метрики производительности** Иван Веткасов, Ведущий консультант, Oracle

ДНИ ORACLE В МОСКВЕ

Oracle Modern **Cloud Day**

30 октября 2019

#OracleMCD

ПАРАЛЛЕЛЬНЫЕ СЕССИИ

Сессия 2. Аналитика, Большие Данные и Машинное Обучение	
15:40	Перерыв
16:00	**Потоковая аналитика данных с помощью Oracle Stream Analytics 19** Александр Моисеев, Ведущий консультант Oracle
16:30	**Аналитические платформы Oracle для построения решений промышленного интернета вещей** Иван Диканев, Эксперт-консультант ФОРС
16:45	**Интеграция больших данных от локальных систем до облака с Oracle Data Integrator** Наталья Кусова, Ведущий консультант Oracle
17:15	**Новые возможности Oracle Database для хранилищ данных** Сергей Томин, Ведущий консультант Oracle

Спасибо!

Андрей Пивоваров

Andrey.Pivovarov@oracle.com