

ORACLE®

Новости OOW по интеграции данных

Дарья Алешкова
Менеджер по продуктам
интеграции СНГ

22 ноября, 2018

Oracle Day

ORACLE®

Управления
информацией как
основа бизнеса

UBER



NETFLIX



amazon.com



Spotify

airbnb



Мы должны
двигаться вперед
имея данные и
учиться управлять
на основе данных

Все уровни интеграции

Объединенное решение для всех интеграционных нужд

Интеграция для

Приложений

Интеграция для

Инфраструктуры

Интеграция для

Аналитики

Application
Integration

API
Management

Data
Replication

Bulk Data
ETL & E-LT

Process
Integration

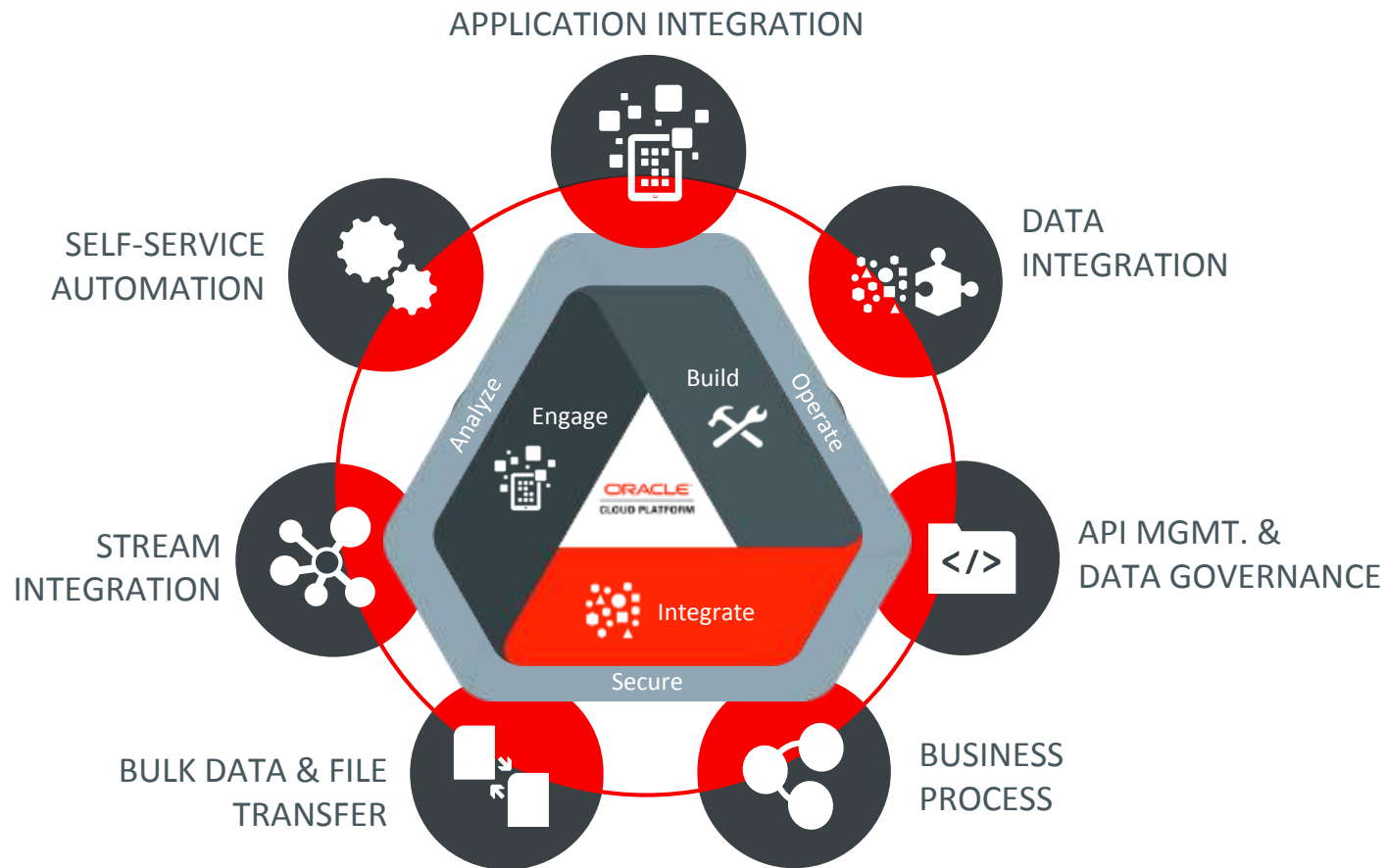
Stream
Processing

Metadata
Management

Data
Quality

Единая технологическая платформа (PaaS)

Интеграционные решения для бизнеса



- **Единая платформа**
 - Для любых задач интеграции
- **0-Кода**
 - Разработка мышкой
- **Полная**
 - Поддержка гибридных платформ
- **Единое окно**
 - Для управления и мониторинга

Инновационные Технологии

#1 Realtime / Streaming
Data Integration Tool

#1 Pushdown / E-LT
Data Integration Tool

1st to certify replication with
Streaming Big Data

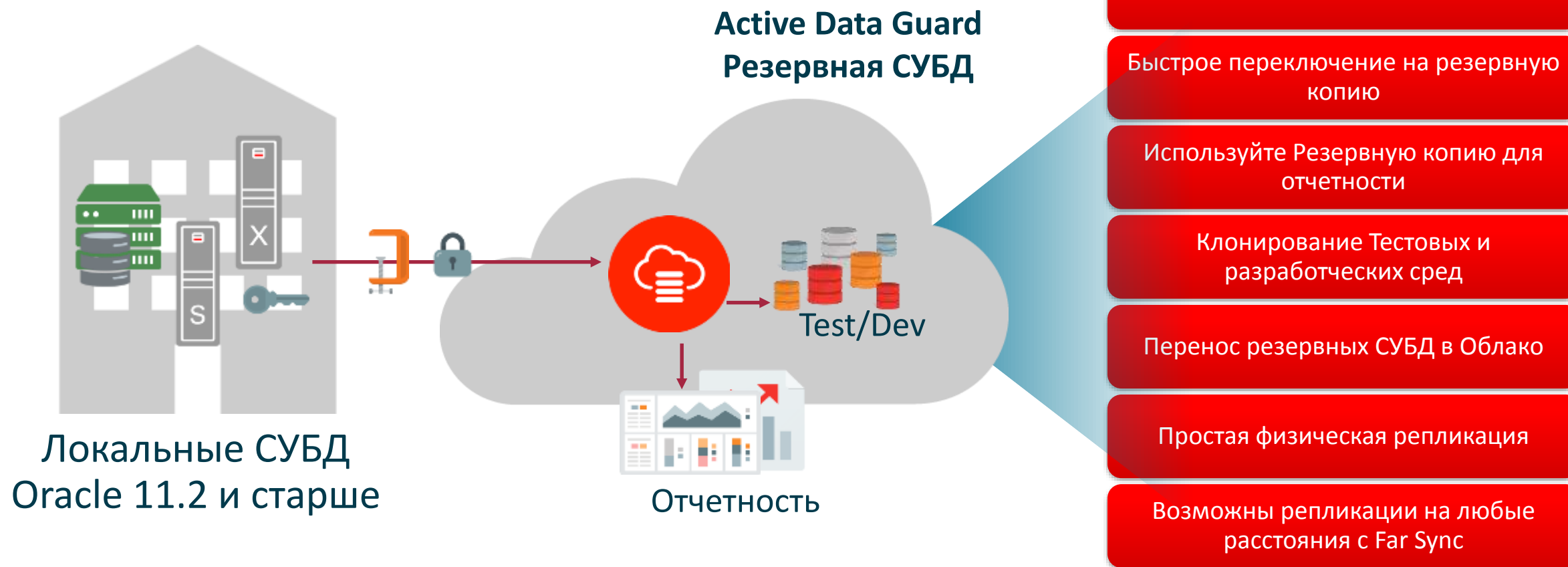
1st to certify E-LT tool with
Apache Spark/Python

1st to power Data Preparation
w/ML + NLP + Graph Data

1st to offer Self-Service &
Hybrid Cloud solution

Создание копии данных для анализа данных одной системы Oracle Active Data Guard

Oracle Active DataGuard основа для построения **высоко доступных решений** на Oracle Database, предоставляя мгновенное **аварийное восстановление**, **физическая репликация**, и **доступный на чтение резервный сервер** для целей отчетности.



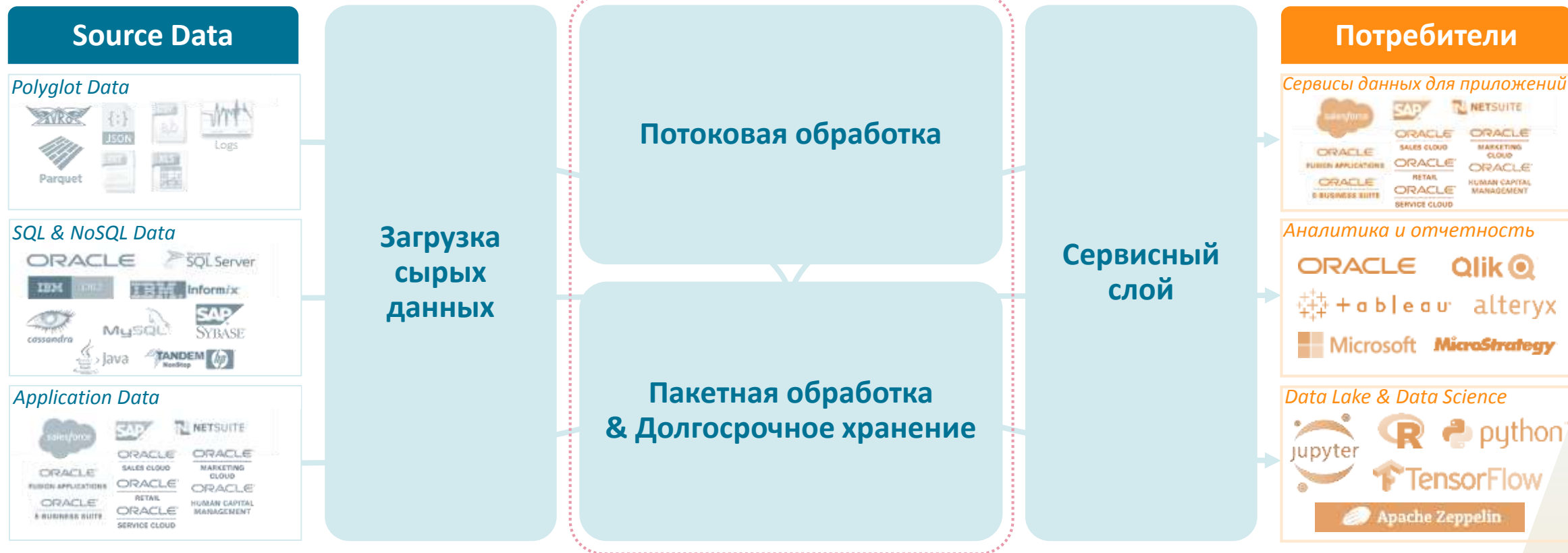
www.oracle.com/technetwork/database/availability/dr-to-oracle-cloud-2615770.pdf

Каковы современные тренды?

#1 – Загрузка в любой момент доступности данных

#2 – Обработка с любой задержкой

#3 – Данные доступные в формате, удобном для потребителя



#4 – Serverless Инфраструктура
#5 – Data at rest на Object Storage

Что такое современная загрузка данных?

#1 – Загрузка в любой момент доступности данных

Исходные данные

Polyglot Data



SQL & NoSQL Data



Application Data



Загрузка сырых данных



Bulk Copy Utilities

Копирование данных это плохо, но перегрузка источника хуже

- Физическое совмещение данных – это обычно первый шаг в Data Lake
- Оптимизация для минимального влияния на источники
- Репликация измененных данных – это лучшая опция для баз данных

Event driven – обязательно в новом мировом порядке

- Переносим данные тогда, когда возможно и не ждем job scheduler
- Некоторые данные должны обрабатываться пакетно

Поддержка ключевых стилей виртуализации и переноса данных

- Параллельное копирование для файловых данных
- Репликация для реляционных данных
- Пакетное извлечение или storage репликация для полной копии
- Федерация данных/Виртуализация «хорошо иметь», но не все источники могут хорошо работать

Первоначальная загрузка подчиняется законам физики

- Загрузка терабайтов занимает время
- Оптимизация для пакетных программ загрузки или big data программ
- Магии не существует ☹️

GoldenGate – современный инструмент

НЕРЕЛЯЦИОННЫЕ ДАННЫЕ

ИНТЕГРАЦИЯ С ЯДРОМ

УДАЛЕННЫЙ ЗАХВАТ

УПРОЩЕНИЕ

МИКРОСЕРВИСЫ

КОНТЕЙНЕРЫ

МОНИТОРИНГ

ПОТОКОВАЯ АНАЛИТИКА

ОБЛАКА

ПОДПИСКА



**GOLDENGATE ДЛЯ ПОТОКОВОЙ ОБРАБОТКИ
BIG DATA:**



Что такое современная обработка данных?

Потоквые Use Cases

- Clickstream аналитика
- Рекомендательные движки
- Обнаружение мошенничества

Pipeline “Logic” Layer

- Создаем data pipelines, правила и встраиваем Machine Learning
- Архитектурная изоляция: независимость от движка
- Повышение usability для аналитиков

Потоковые движки

- Предпочтение Oracle: Flink для реальной потоковой обработки, Spark Streaming для микробатчей и ML
- Другие: Storm, Kafka Streams, Samza или проприетарный

#2 – Обработка с любой задержкой



Пакетные Use Cases

- ETL Offloading
- Загрузка Data Lake
- Масштабируемая аналитика

Пакетная (MPP) обработка

- Движки: Spark для большинства, Hive или Flink в некоторых случаях
- Хранение: Object Storage

Интерактивный доступ

- SQL для прямых запросов к большим объемам данных:
 - Hive SQL (простые)
 - Spark SQL (простые)
 - Sparkline OLAP (расширенные)
- Machine Learning & Graph
 - MLlib
 - PGX, GraphX

** In your own data center you can have mixed workloads run from same physical clusters (ie; Kappa-style) but from the Cloud you should only pay for what you use and not care how the infrastructure is managed...*

Что такое современный сервисный слой?

Streaming API для Real-Time данных

- Publish and Subscribe (Kafka) с дружественными REST APIs
- Основанный на REST означает, что API Gateways может быть использован для настройки ACLs
 - HTTPS передача и отсутствие файлового доступа
 - Data redaction на уровне API / contract level

SQL-based доступ для интерактивной отчетности

- Пакетный перенос данных (ETL) во внешний dw/data marts
- Прямой SQL доступ к данным в Data Lake (Spark OLAP)
 - Наиболее распространен
- Интеграция с LDAP для безопасности

Прямой доступ к сырым данными для специалистов

- Прямой доступ к object storage или HDFS
- Особенно полезно для ML / AI программ
 - Прямой доступ без необходимости перемещать данные
- Локальный доступ выдается на объектном уровне (например, файл данных)

#3 – Данные доступные в формате, удобном для потребителя

Сервисный слой

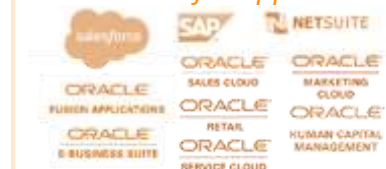


SQL



Потребители

Data Services for Applications



EDW & Analytics for Reporting



Data Lake & Data Science



Что такое современная инфраструктура?

Public Cloud Infrastructure устанавливаем стандарты низкой стоимости и высокой производительности

- Платим только за использование, serverless стиль работы снижает нагрузку с клиентов
- Быстрые compute сервисы и «плоская» сеть дает 5x прирост производительности *
- Очень низкая стоимость хранения с практически бесконечной надежностью 99.999999%

Источники данных

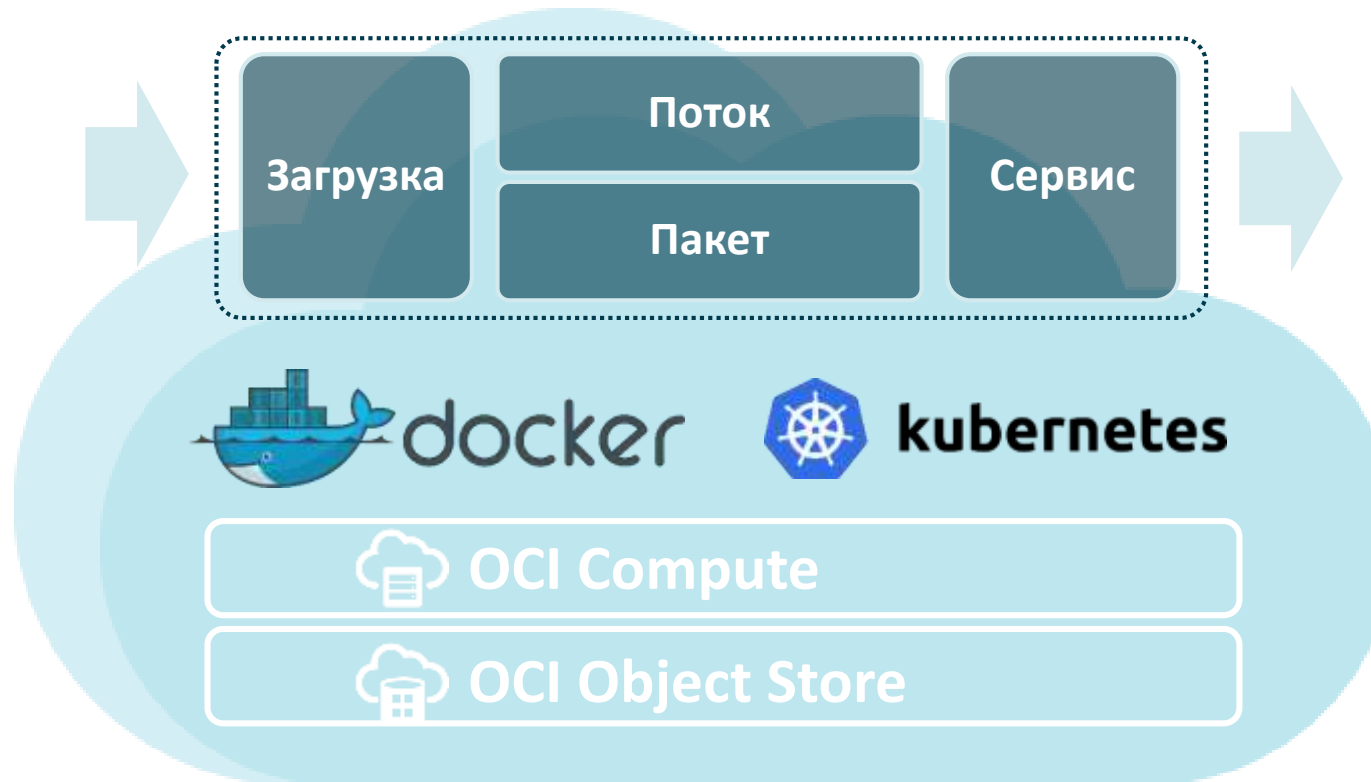
Polyglot Data



SQL & NoSQL Data

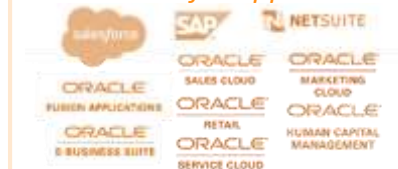


Application Data



Потребители

Data Services for Applications



EDW & Analytics for Reporting



Data Lake & Data Science



#4 – Serverless Инфраструктура
#5 – Data at rest на Object Storage

Современный подход



ЛЮБЫЕ
ДАННЫЕ



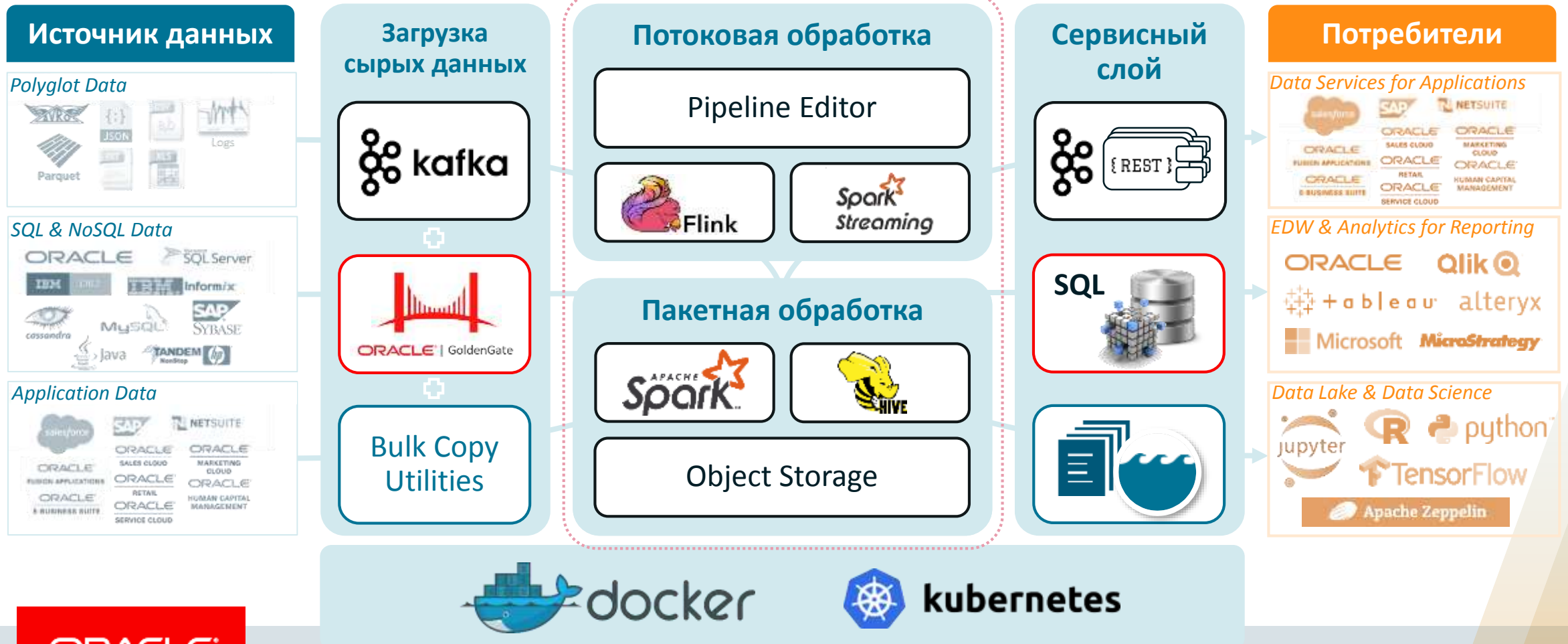
ЛЮБАЯ
ЗАДЕРЖКА



ЛЮБОЙ
ФОРМАТ



В ЛЮБОМ
МЕСТЕ



Современный подход? Oracle Cloud



ЛЮБЫЕ
ДАННЫЕ



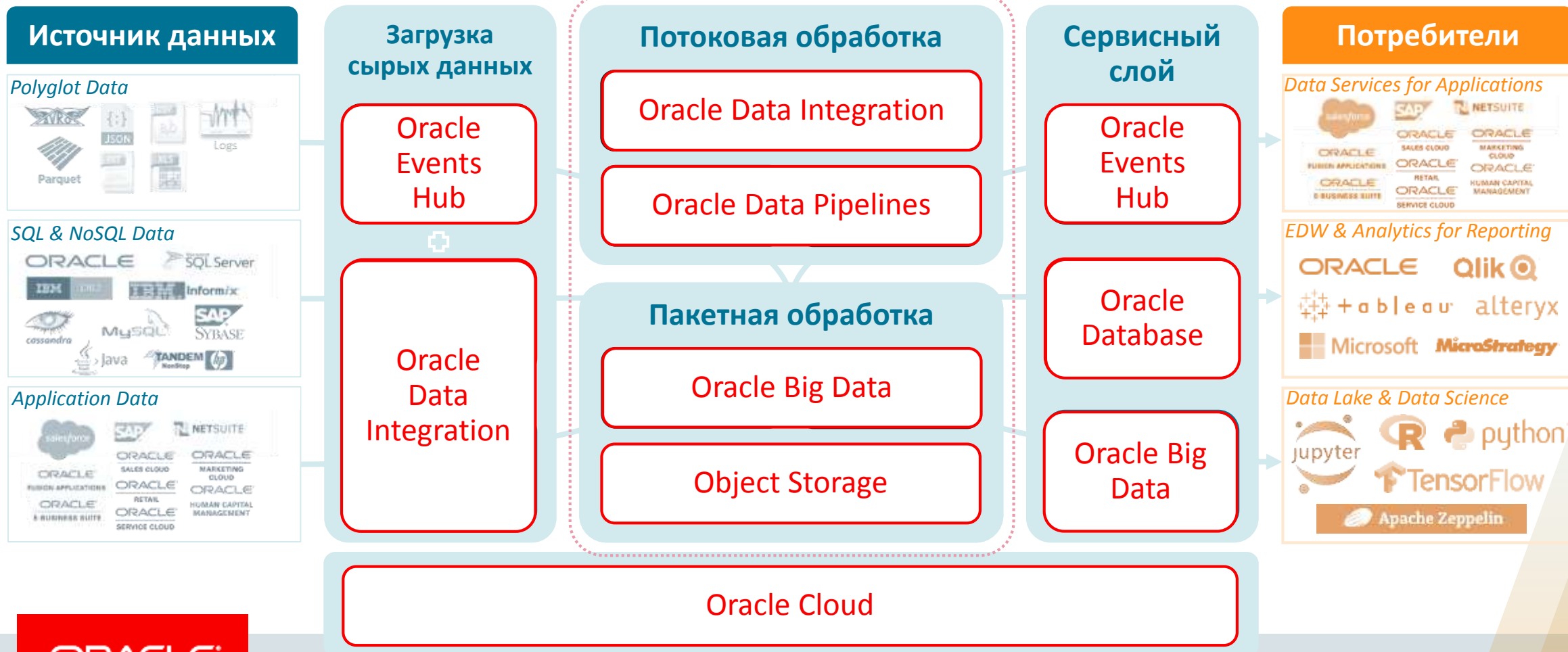
ЛЮБАЯ
ЗАДЕРЖКА



ЛЮБОЙ
ФОРМАТ



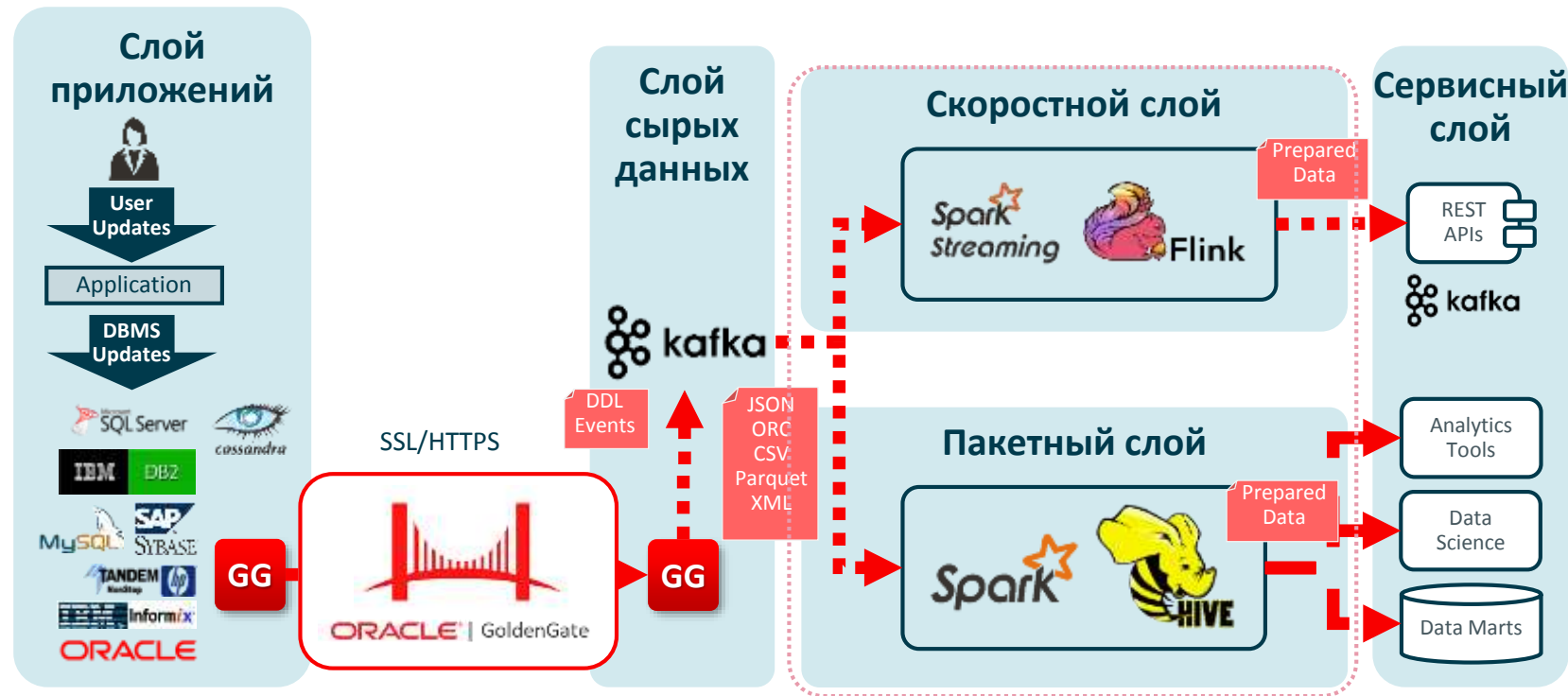
В ЛЮБОМ
МЕСТЕ



GoldenGate для Kappa / Streaming

Проверено передовыми клиентами: Kappa ETL архитектура

- Мы видим множество клиентов GG использующих различные потоковые движки: Spark Streaming, Flink, Kafka Streams, Apex/DataTorrent, Samza, Kinesis Firehose, Storm, etc.



EBay обрабатывает 200 млрд tx в день;
более 25 TB изменений в день через
GoldenGate и задержка составляет
менее 2 seconds (Flink)



LinkedIn использует GG на >200 базах в 5
ЦОД по всему миру
(Samza для обработки)



Quickbooks.com использует GG для
Oracle, SQL Server и DB2, работающих в
AWS (GG+Kafka как фабрика данных,
питающая data science/ML платформу)



Apple iTunes использует GG+Kafka для
загрузки транзакций в 5,000+ узловой
Data Lake (для data science)



General Motors использует GG+Kafka и
GG+S3 для переноса транзакций из 600+
баз в их Data Lake

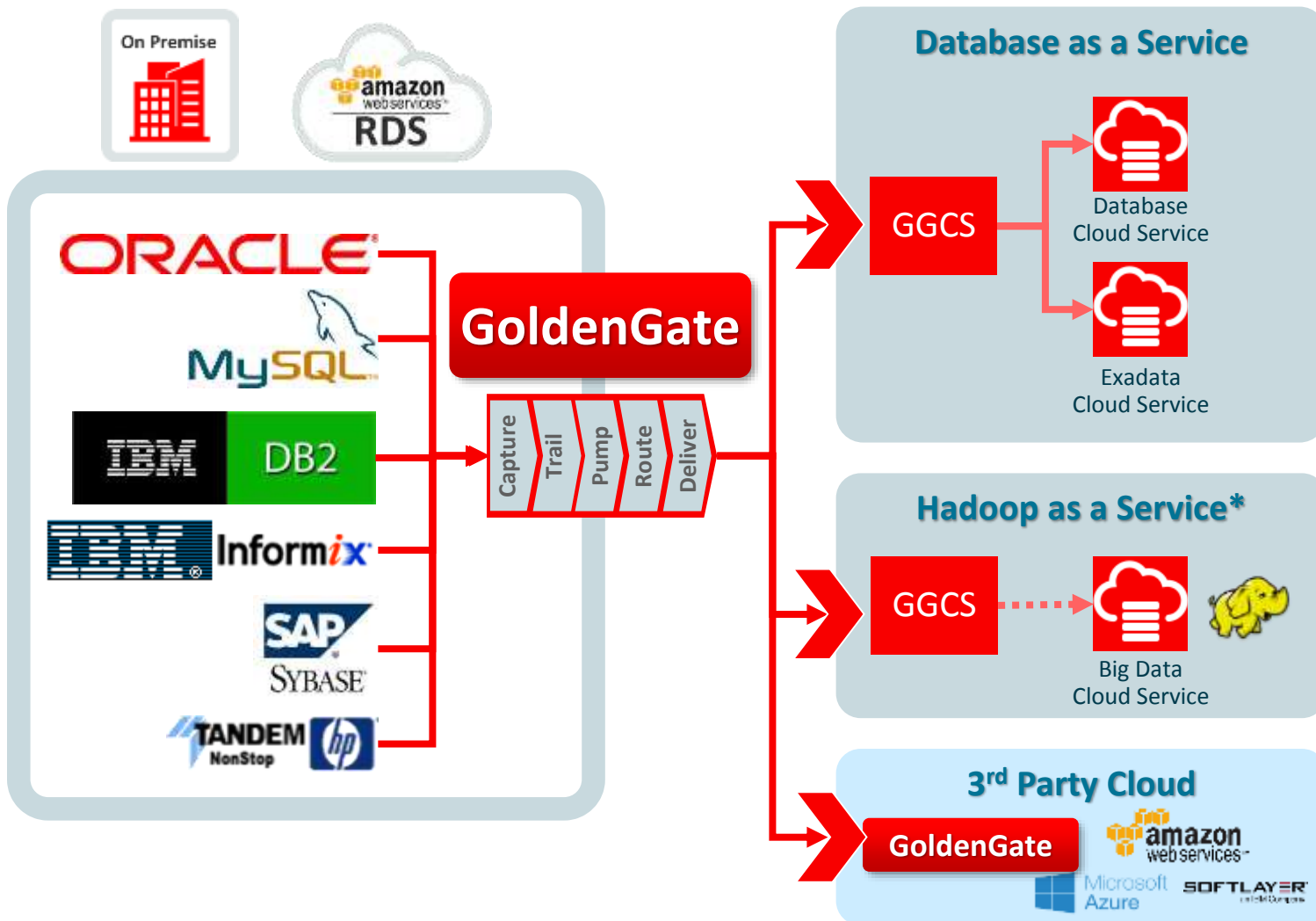


Maersk использует GG+Kafka для
realtime IoT отслеживаний посылок
(таможня в том числе)



MGM переносит всю IT data architecture
на GG+Kafka для streaming

Консолидация данных из гетерогенных источников. Oracle GoldenGate



GoldenGate Cloud Service

- ✓ Репликация в реальном времени в Oracle Cloud по подписке или по часам

Преимущества

- ✓ **Измененные данные** доставка только измененных данных более эффективна чем пакеты
- ✓ **Без потерь** надежный и проверенный инструмент
- ✓ **В реальном времени** всегда актуальные данные

Гетерогенность

- ✓ **Миграция и доставка с СУБД и Облаков 3х поставщиков**

Бизнес цели GG 12.3



Простота



Управляемость



Скорость

Микросервисы

Oracle GoldenGate 12.3



Новое поколение GoldenGate Service Architecture

Масштабируемое, основанное на REST управление, микросервисная архитектура

Oracle Database Sharding Support

Active-active репликация между Shard Groups

Автоматическое обнаружение конфликтов и их разрешение

Встроенное CDR для Oracle без необходимости менять приложение

Гетерогенное развитие

Поддержка DDL в MySQL

Удаленный захват и применение для z/OS

Generic JDBC

Дает возможность реплицировать в любую базу с JDBC драйвером

Cloud! Cloud! Cloud!

Cloud@Customer and ExaCS

Расширенная поддержка Big Data и NoSQL

Адаптеры для Cassandra, MongoDB, Kinesis, Kafka

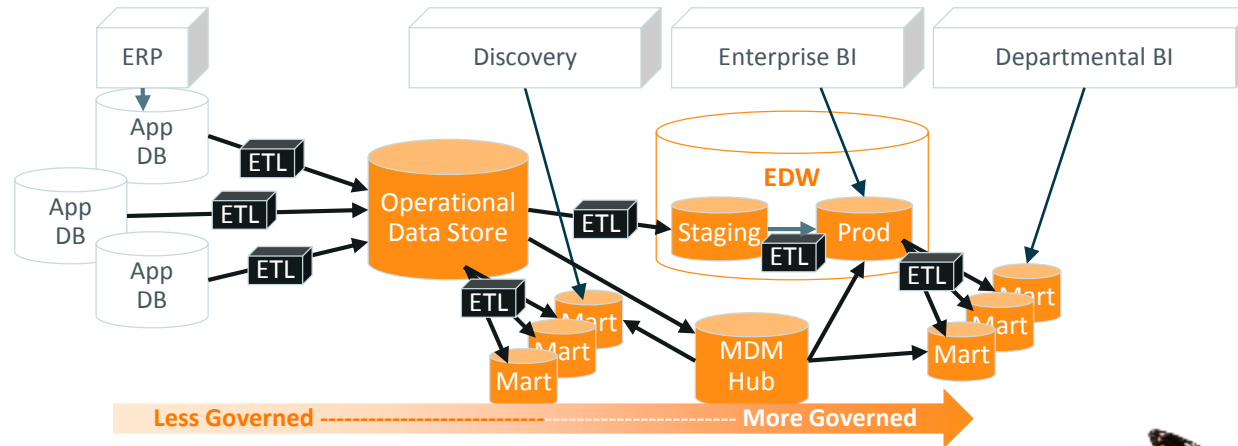
Parallel Apply

Enables support for multiple source files and offload the parallelism to separate server



Новые тенденции в построении хранилищ и озер

ТРАДИЦИОННАЯ:



Classical Data Management: Hub and Spoke

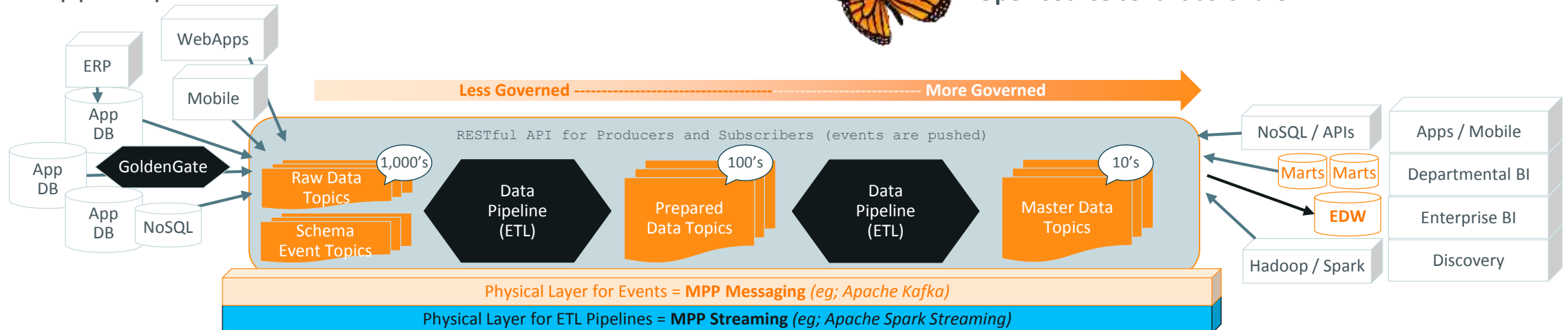
- **Invasive** on Sources
- **High Latency** / SLA
- **Mainly Relational Views**
- **Heavy IT process overhead**
- **Vendor-centric software**
- ODS & ETL Hubs
- EDW/Mart Hubs
- MDM/RDM Hubs
- Static Data Lake Hubs

Next-Gen: Streaming Databus/Kappa

- **Low impact** on Sources
- **Low Latency** (< 1 second)
- **Variety of Data Formats**
- **More Agile DevOps processes**
- **Open source centric software**
- Pub/Sub for Staging
- ETL in Pipelines
- Analytics/CEP in Stream
- Data is in Motion



СЛЕДУЮЩЕЕ ПОКОЛЕНИЕ:



Почему Oracle Data Integrator?

- Чтобы дать реальную гибкость и точность
- Чтобы легко переходить между платформами и стилями обработки данных
- ODI's EL-T архитектура важна при работе с большими объемами
- ODI умеет выполнять pushdown любых трансформаций данных в Data Lake

Oracle Data Integrator 12.2.1 – PS2+

- **Big Data**

- Spark Streaming Support
- Hive Complex Type
- Hive Storage Format Support
- Hadoop Usability Enhancements (Phase 2)
- Mongo DB

- **Cloud**

- RESTful Services
- BICS
- Amazon Redshift

- **Lifecycle Management**

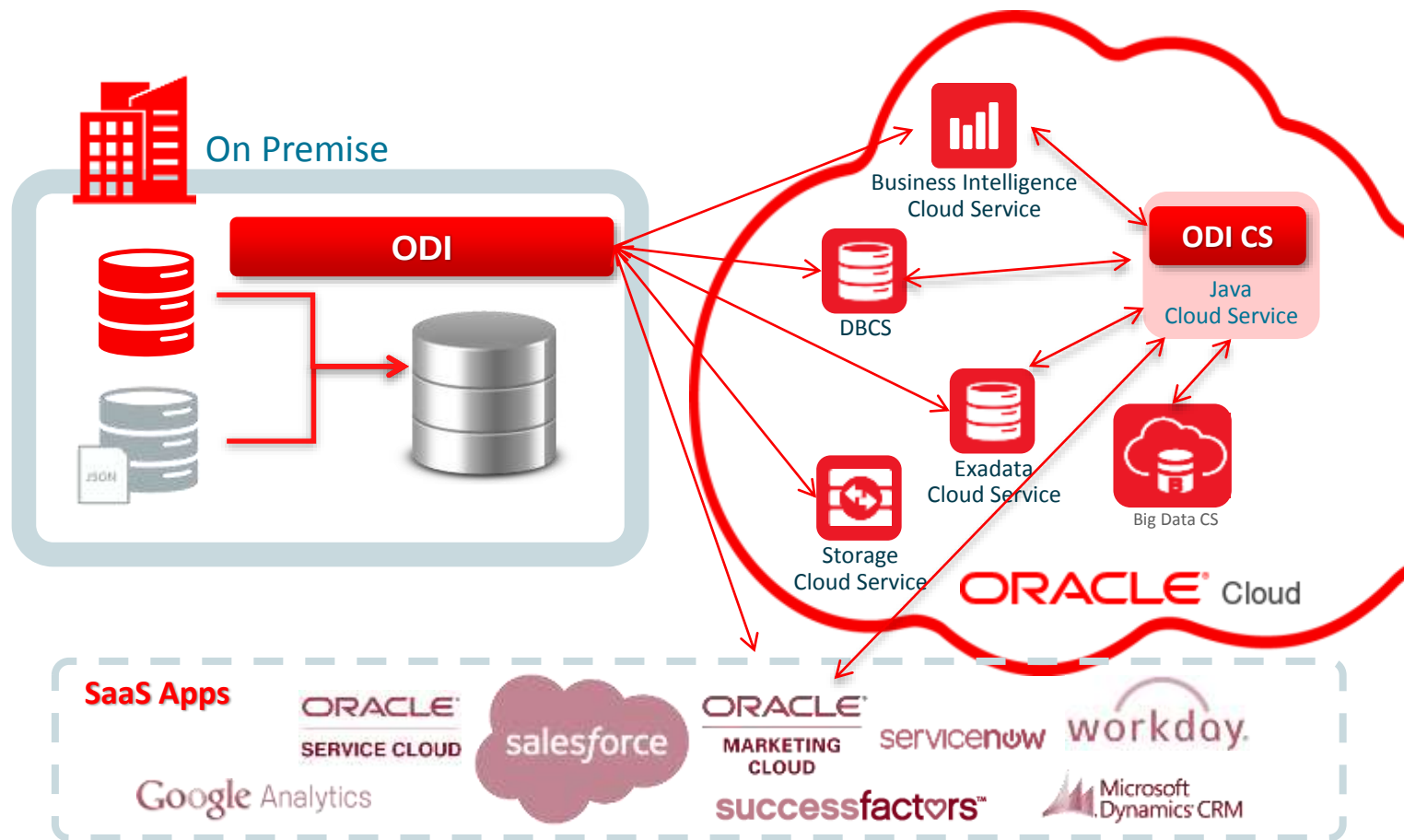
- Git
- Smart Merge

- **ETL Enhancements**

- Unified KM Editor

Построение Озера данных. Oracle Data Integrator

Oracle Data Integrator



Data Integrator в Облаке

- ✓ Внедряй ODI где удобнее и запускай ETL там где хранятся данные

Ключевые преимущества

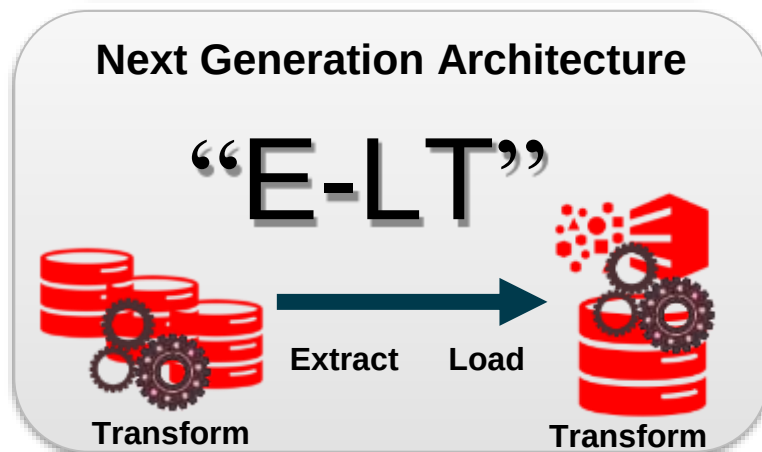
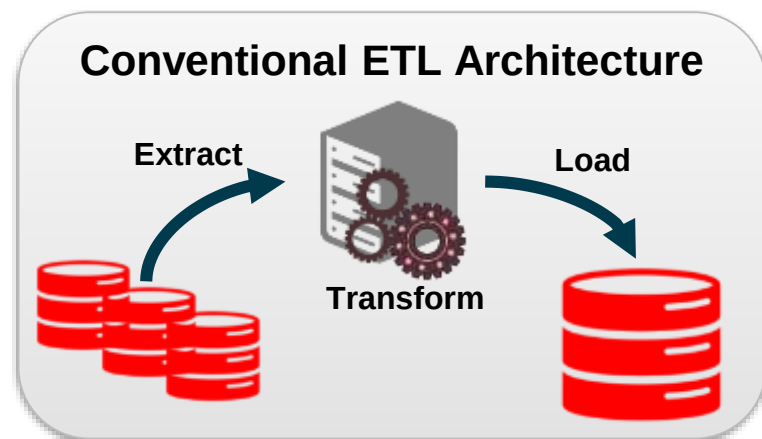
- ✓ **Интегрирован** с OPC
- ✓ **Гибкое внедрение** архитектура позволяет подготавливать данные локально, в Облаке или на Big Data
- ✓ **Широкий список коннекторов** любые источники приложения или СУБД

Независимость

- ✓ **Гибкий** логика трансформации не зависит от технологии, вычисляй там где выгодно

Лидер по производительности

Экстремально быстрая обработка данных при низких затратах



Е-LT представляет собой гибкую архитектуру для оптимальной производительности для любой платформы

Преимущества

- Использует преимущество трансформаций на основе внутренних функций Leverages set-based transformations
- Повышение эффективности загрузки, снижение нагрузки на канал
- Использует преимущество существующей инфраструктуры: Как Серверов, так и ПО

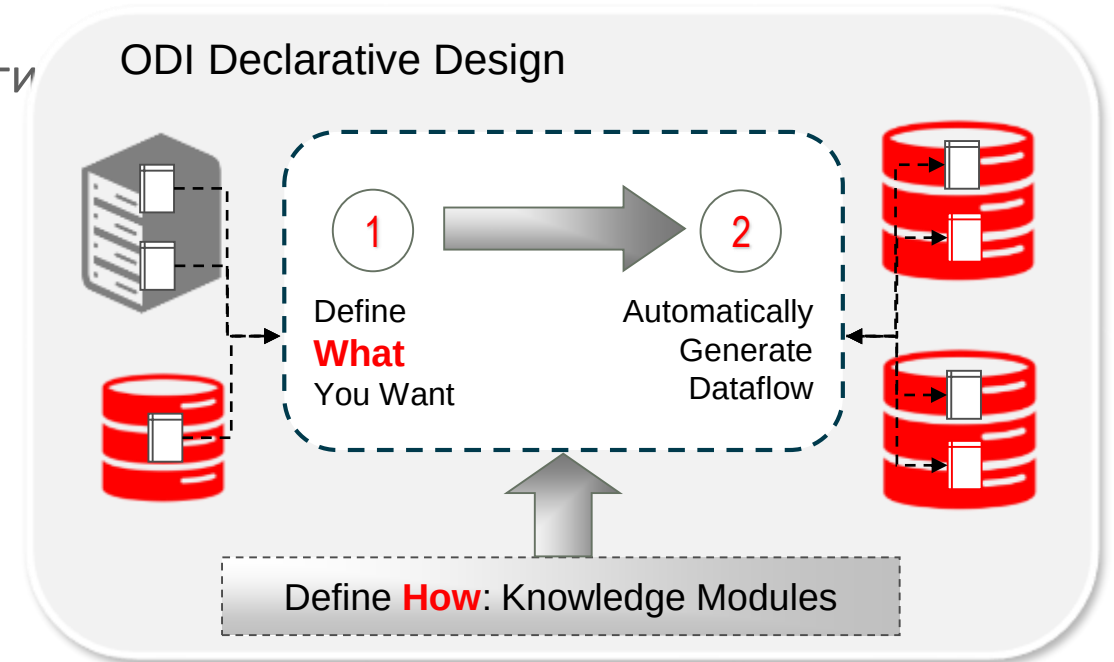
Декларативный дизайн

Повышение производительности разработки для быстрого получения результата

- **Легкий и универсальный** для разработки как простых так и сложных преобразований
- **Надежный и Переиспользуемый:** Создавайте Маппинги один раз для любой физической имплементации
- **Быстро** определяем и перерабатываем маппинги, с высокой степени автоматизации
- **Расширяемая** модель для любых механизмов интеграции

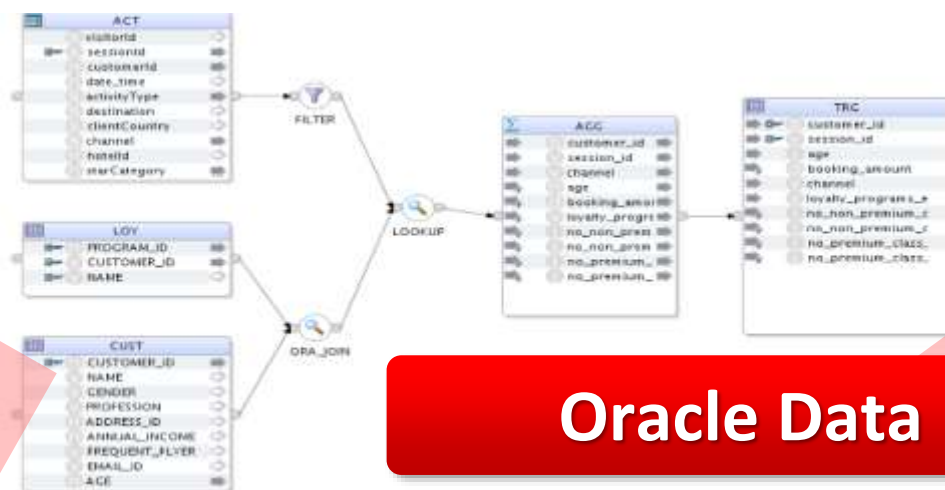
Преимущества

- Существенно сокращает время обучения
- Быстро и легко разрабатывать, быстро внедрять
- Легко поддерживать



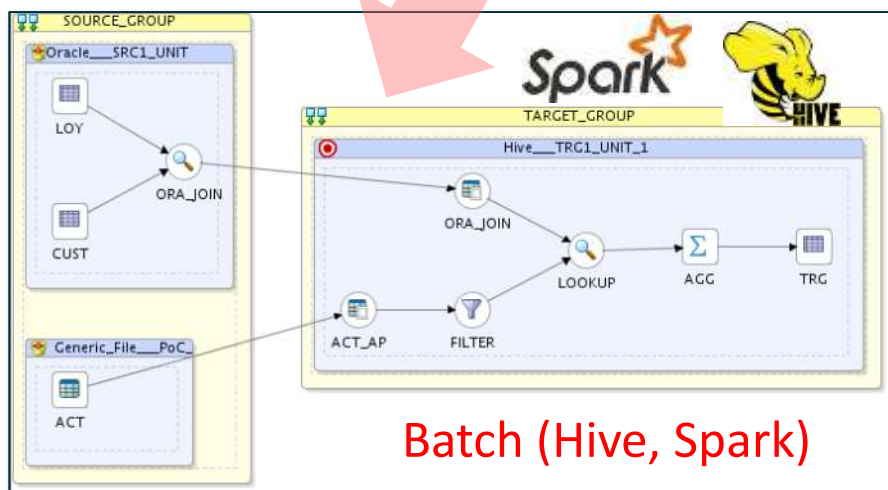
Генерация **Пакетов и Поточков** из единого Мappings

Единый логический дизайн

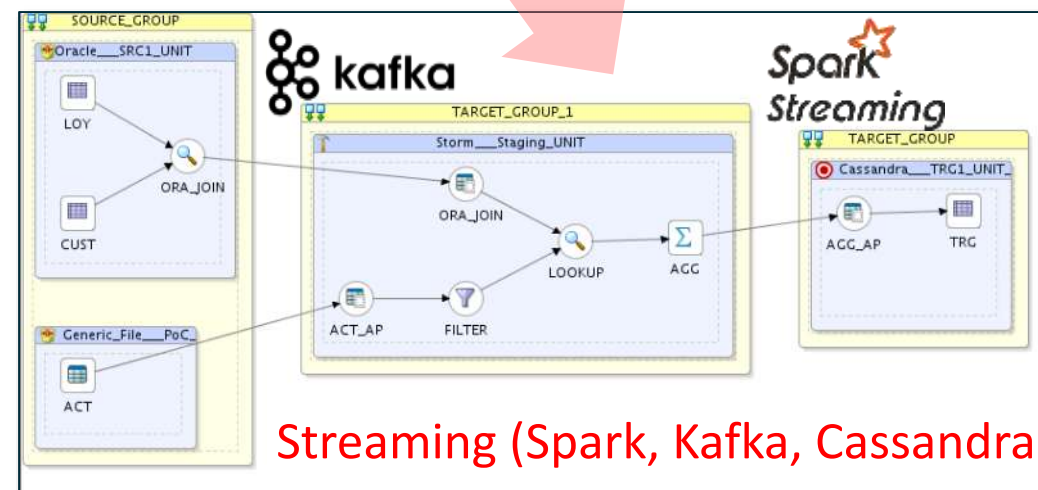


Oracle Data Integrator

Гибкая реализация

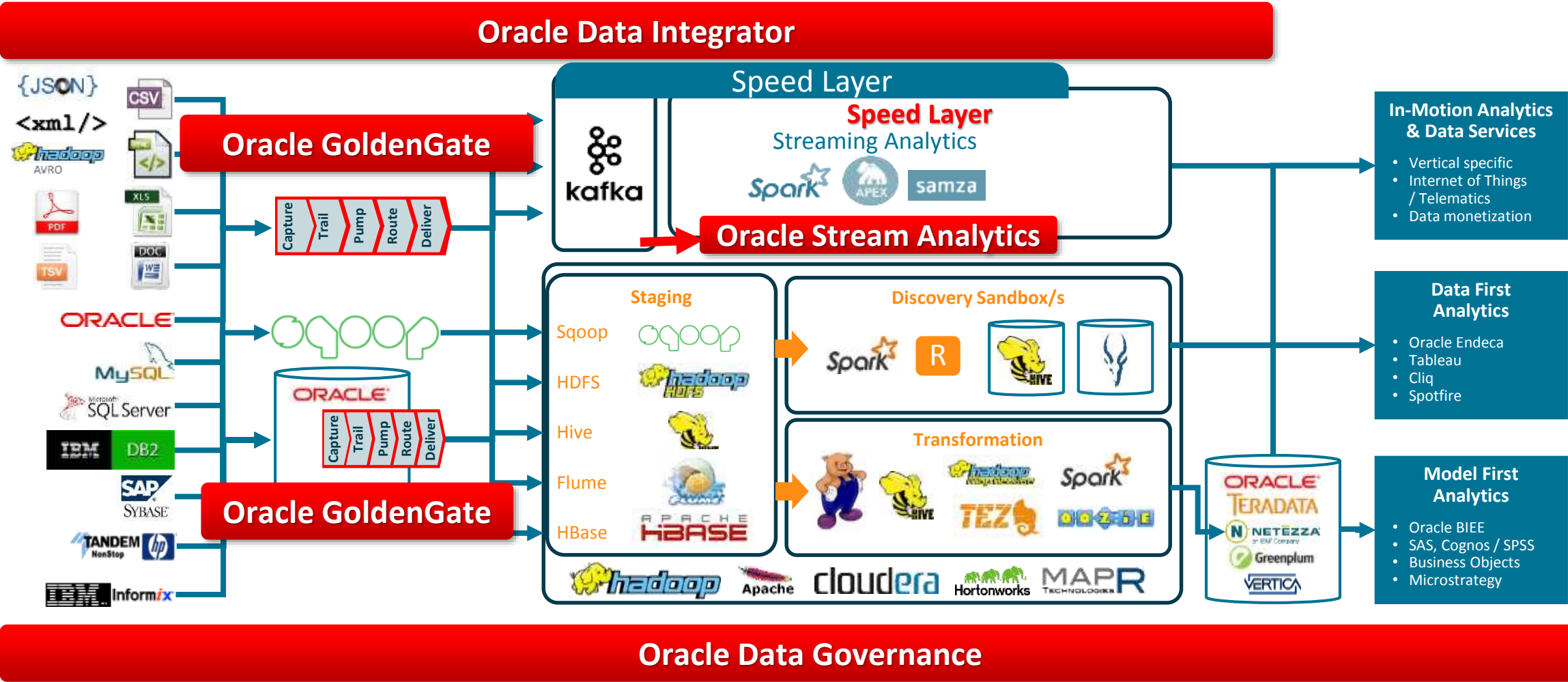


Batch (Hive, Spark)

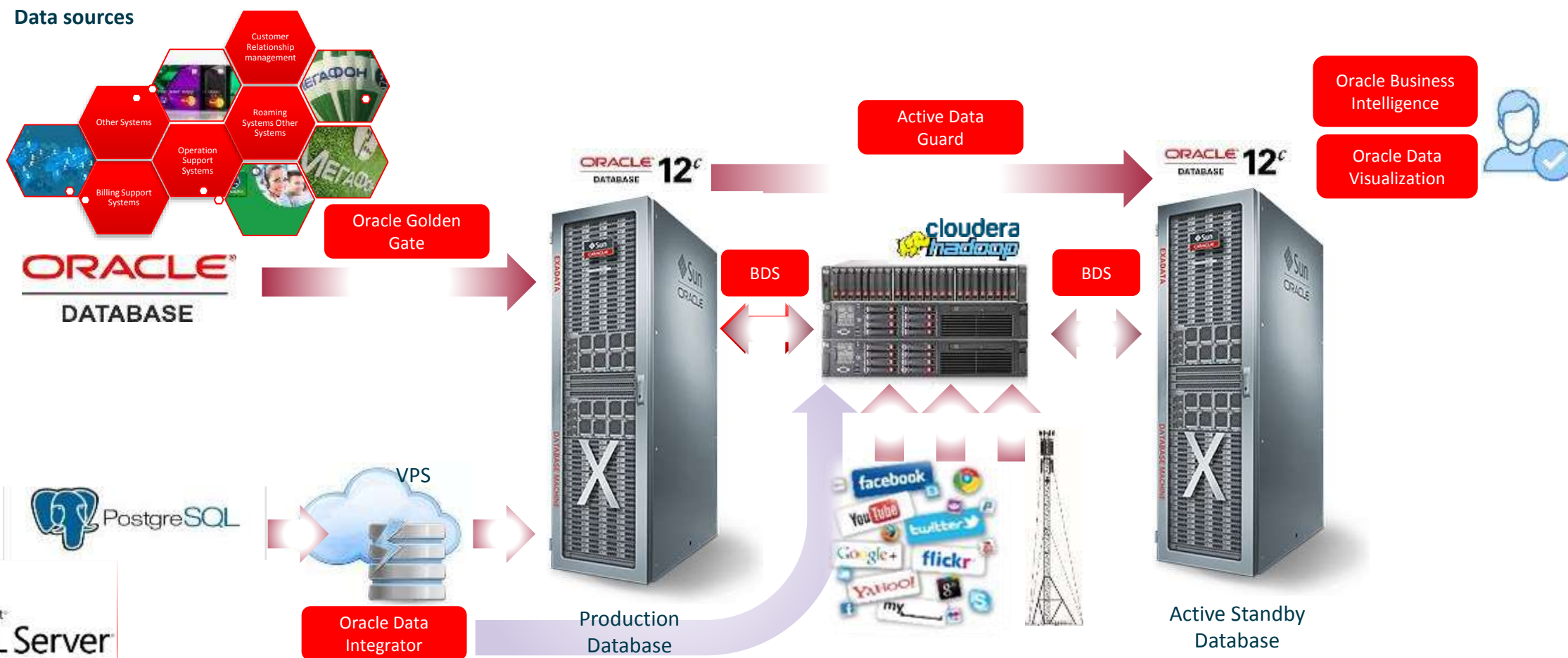


Streaming (Spark, Kafka, Cassandra)

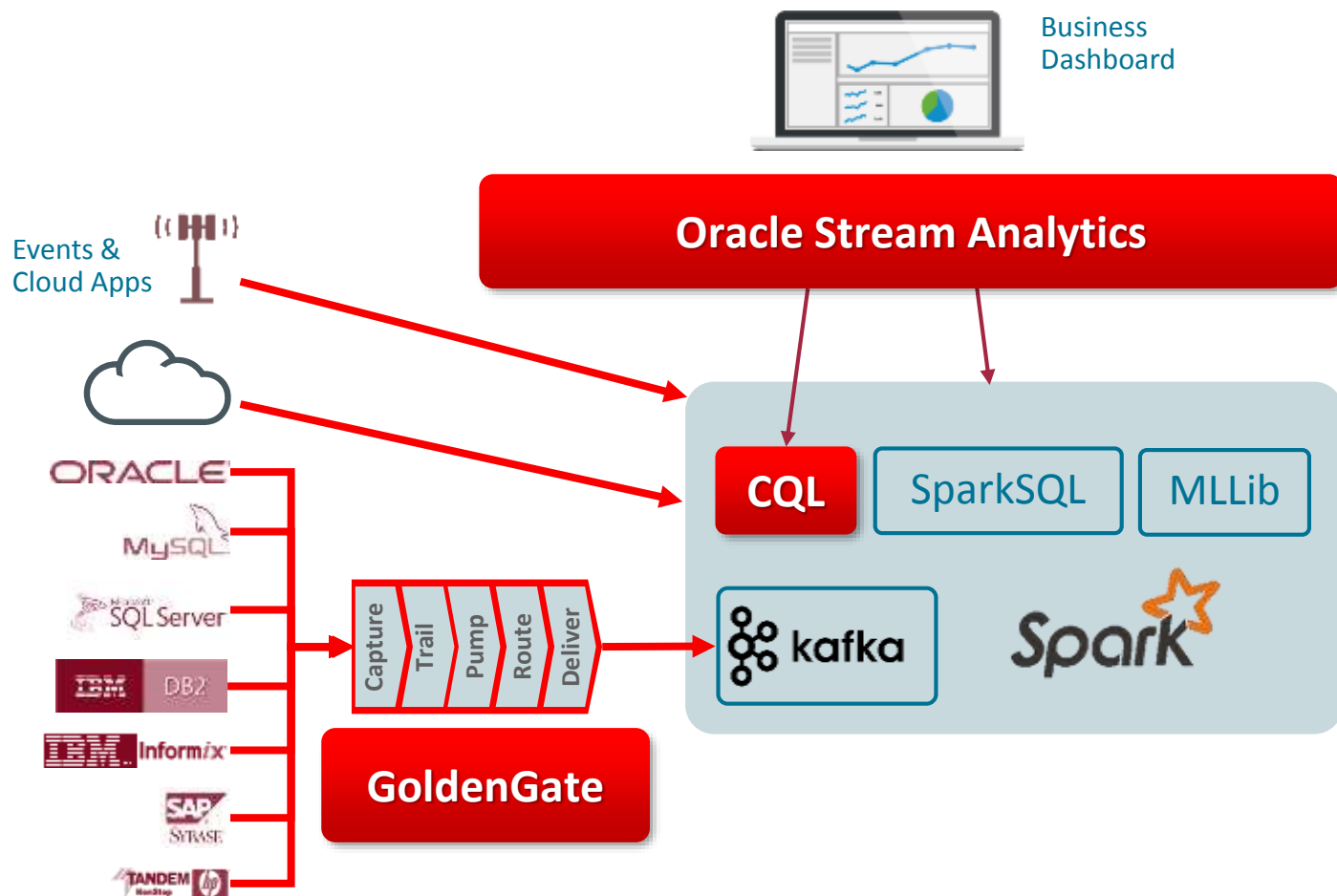
Референсная архитектура Oracle для построения Озера



Озеро Данных на примере клиента



Oracle Stream Analytics для Больших Объемов Big Data



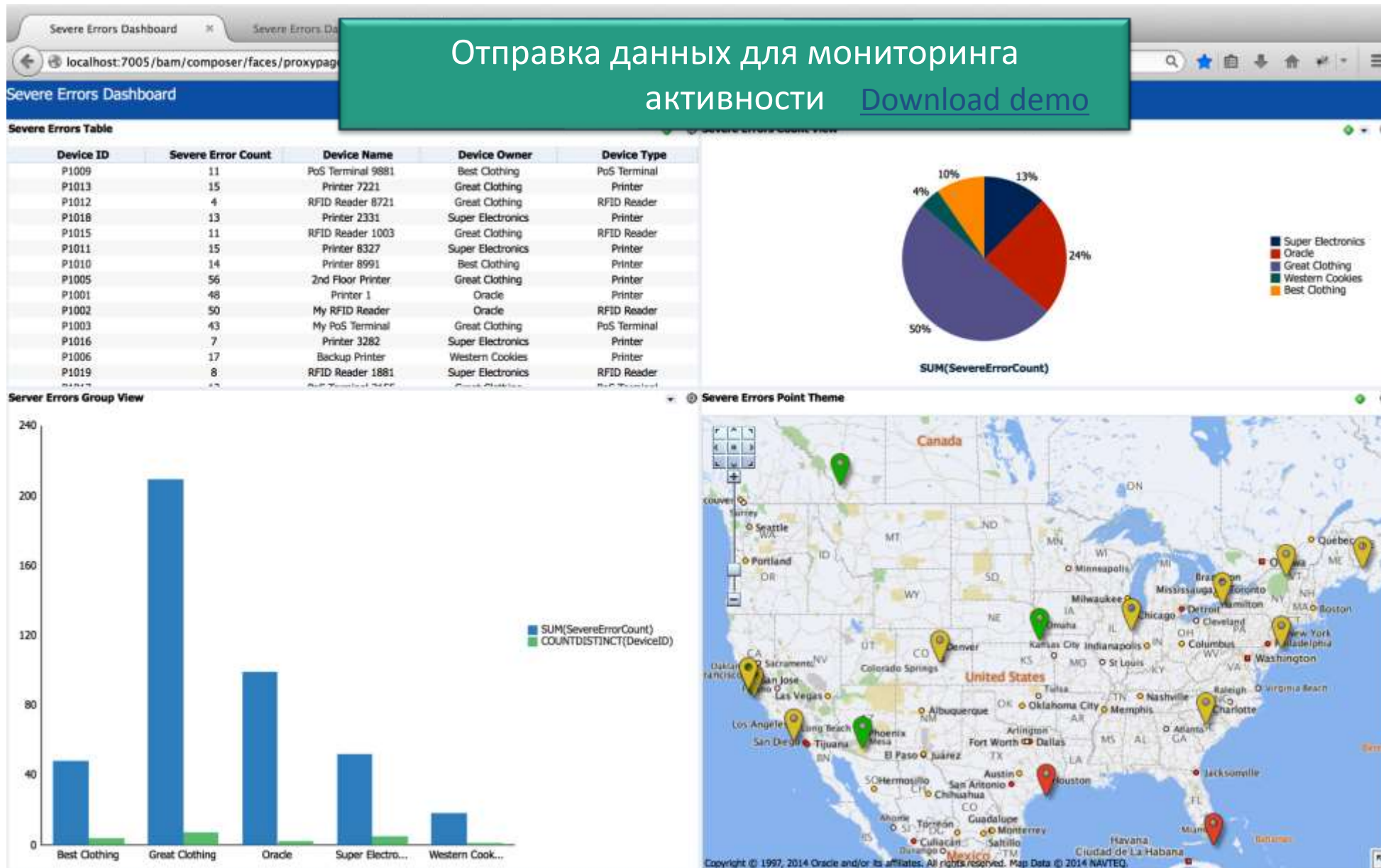
Потоковая аналитика

- ✓ Анализ потоков данных в памяти
- ✓ Простой интерфейс для построения приложений
- ✓ Пред интегрирован с Oracle Spatial, Coherence, Apache Kafka, Apache Spark и Oracle GoldenGate

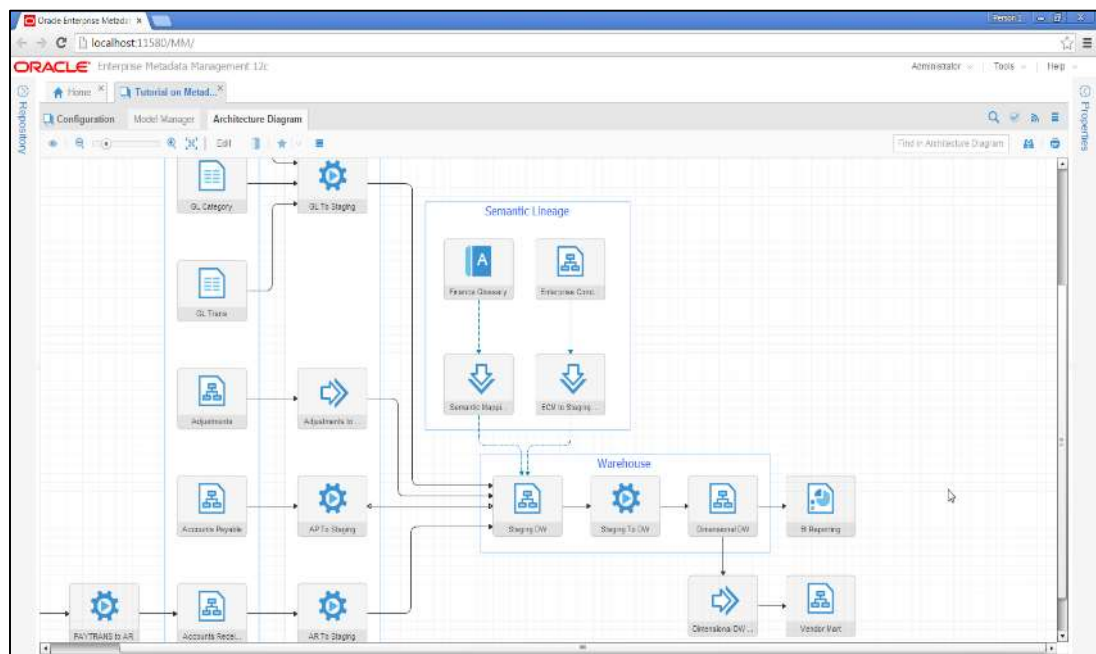
Ключевые преимущества

- ✓ **Легко начать** интуитивный веб-интерфейс
- ✓ **Гибкость** движок может работать самостоятельно или на Spark
- ✓ **Интеграция** с GoldenGate для загрузки потоков изменений СУБД

Отправка данных для мониторинга активности [Download demo](#)



Oracle Metadata Management для управления данными



Business/Metadata Catalog

- ✓ Связь отчета с источником
- ✓ Анализ Влияния
- ✓ Аудит, версионность, Сопоставление
- ✓ Работа команд, чаты
- ✓ Аннотации и метки
- ✓ Автоматический захват

Расширенный сбор данных

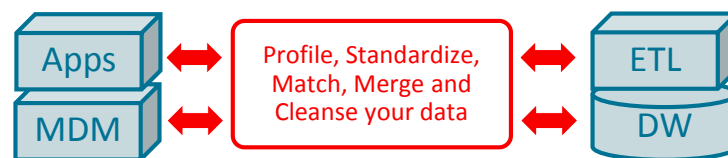
- ✓ 3rd Party BI Metadata
- ✓ 3rd Party ETL Metadata
- ✓ 3rd Party DB Metadata
- ✓ 3rd Party Modeling Tools
- ✓ Hadoop & NoSQL Metadata
- ✓ Metadata Standards

1. **Business Glossary** – простой инструмент ведения каталогов и иерархий

2. **Metadata Management** – горизонтальный и семантический анализ взаимосвязей любых данных

Oracle Enterprise Data Quality

Oracle Enterprise Data Quality обеспечивает **надежные бизнес данные** предоставляя фундамент для **профилирования и стандартизации данных, сопоставление и объединения возможностей по очистке данных.**



Ключевые преимущества:

Просто использовать

Инструмент доступный для бизнеса, подходит для больших и малых проектов

Высокая производительность

Подходит для сравнения записей в реальном времени данных приложений и мастерданных

Разнообразные правила

Масштабируемые решения для решения различных задач по очистке данных

Лидер в Gartner Magic Quadrant и Forrester Wave:



Simple to start:

- Health check for your data – quick and easy setup for data profiling
- Intuitive business user friendly toolkit for building enterprise data quality rules



Представляем Data Integration Platform Cloud

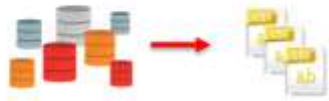
Универсальные пакеты

		Data Integration Platform Cloud		
		Standard Edition	Enterprise Edition	Governance Edition
Solution Use Cases:	Features:			
- Database Migrations (offline) - Bespoke Data Marts - DW Automation (batch) - Database E-LT & ETL - Profiling & Lineage (basic)	Intuitive Web Dashboard & Service Home	✓	✓	✓
	Bulk Data Movement & Transformations	✓	✓	✓
	Basic Profiling with SaaS & PaaS Connectivity	✓	✓	✓
	Admin, Monitoring, Basic Governance	✓	✓	✓
- Database Continuity - Database Migrations (online) - DW Automation (stream/batch) - Real-time Reporting/Analytics - Big Data ETL (stream/batch)	Streaming Database Transactions		✓	✓
	Streaming ETL and Complex Transformations		✓	✓
	Native Big Data (stream/batch) Support		✓	✓
	Multi-Directional Database Replication		✓	✓
- Data Profiling (advanced) - Data Cleansing (real-time/batch) - Data Lineage (comprehensive) - Audit and Repair, Workflow	Advanced Business Glossary and Business Rules			✓
	Comprehensive Data Profiling & Cleansing			✓
	Compliance, Audit and Case Management			✓
	Data Lineage – Business, Vertical, Horizontal			✓

Примеры задач

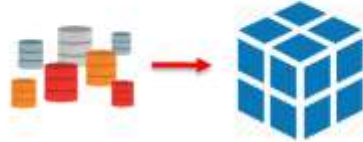
Export Data Into Files

export data from any source into flat files



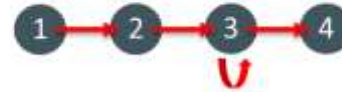
Load Cube

ingest, transform and load star schema



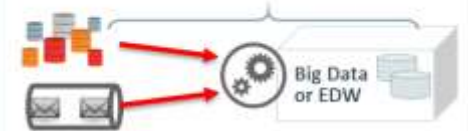
Sequence Tasks

orchestrate Tasks execution, error handling



Transform Data

ingest, perform simple to medium data transformations



Import Data From Files

load files from any location into any target



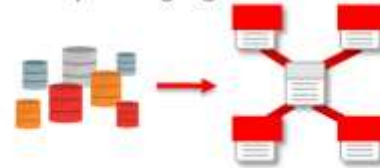
Create Snapshot Table

capture a snapshot of data as of point in time



Load Dimension

ingest, transform and load slowly changing dimensions



Invoke ODI Job

execute ODI Load Plan or Scenario



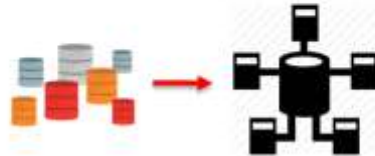
Profile Data

scan and statistically profile data sets (from DB or Big Data)



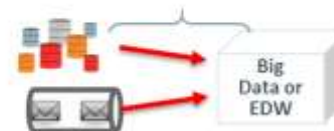
Build Data Warehouse

create, ingest, transform and load data warehouse



Copy Data

data movement into staging area – real-time or batch



Map Data

ingest, stage, transform and load DW/Big Data



DIPС Издания – Ключевые кейсы применения и функционал

Data Integration Platform Cloud Standard

Core Use Cases

- Data Warehouse and Data Mart Integration
- Data Migrations

Core Capabilities

- Bulk Data Movement
- ETL Transformations
- Push-down Data Processing
- Basic Data Profiling

- ODI-Enterprise Edition + Data Profiling

Data Integration Platform Cloud Enterprise

Core Use Cases

Standard Edition plus:

- Big Data Integration
- Data Synchronization
- ZDT Migrations
- Active-Active
- Real-time Data Warehouse

Core Capabilities

Standard Edition plus:

- Big Data & Streaming ETL
- Real-time Replication/Streams

Standard Edition Plus:

- GoldenGate (all)
- ODI for Big Data

Data Integration Platform Cloud Governance

Core Use Cases

Enterprise Edition plus:

- Data Health Check
- Enterprise Data Cleansing
- Data Lineage (Phase 2)

Core Capabilities

Enterprise Edition plus:

- Data Profiling & Validation
- Match & Merge
- Glossary & Lineage (Phase 2)

Enterprise Edition Plus:

- Enterprise Data Quality
- Metadata Mgmt (Phase 2)

Sold as an option of Enterprise

DIPС Решаемые задачи



Рядом с нами

Communications	Finance & Banking	Media
Energy & Industrial	Health & Insurance	Retail
Services	High Tech	Other

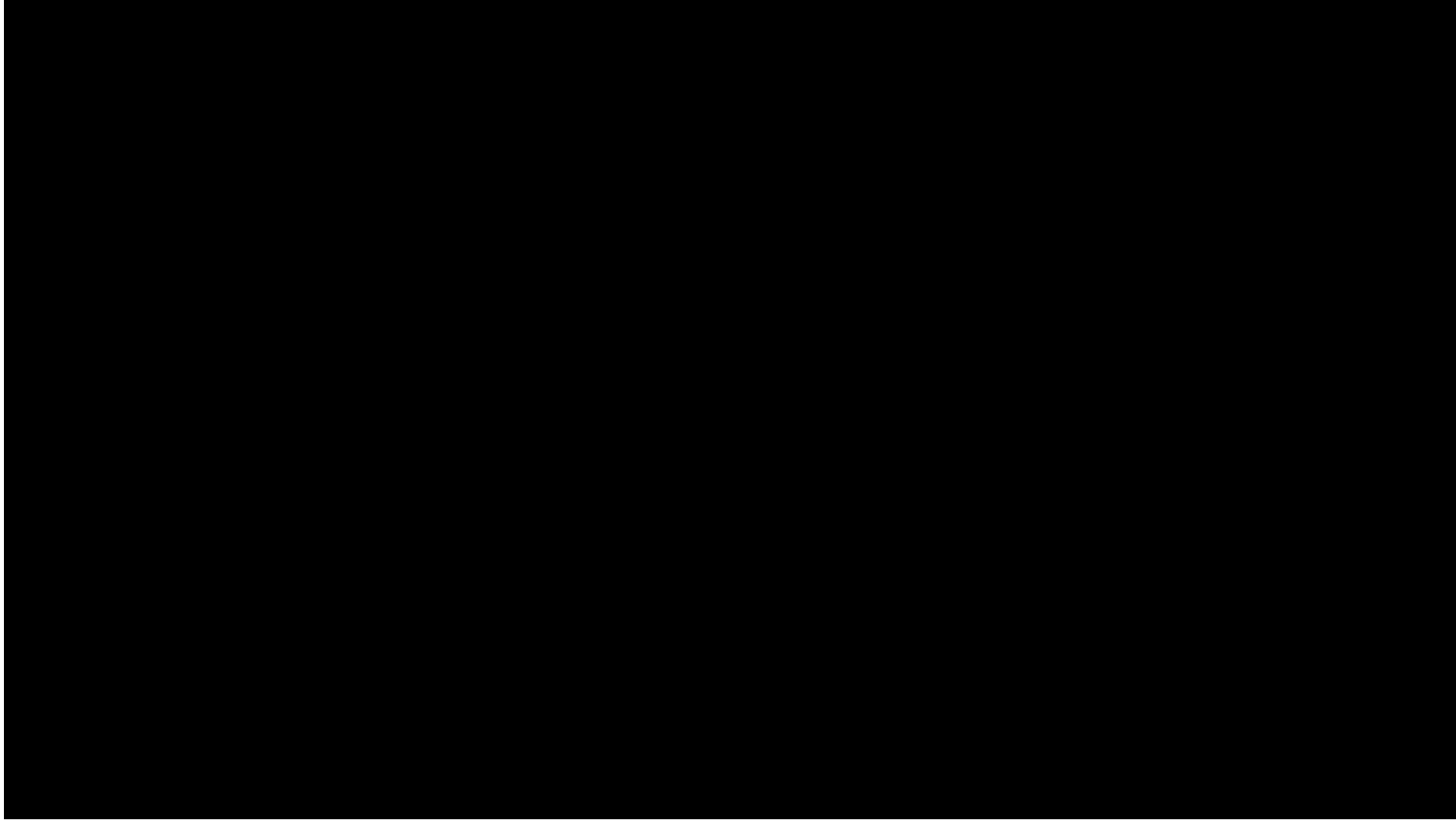
- Съездите за покупками в **Ашан**
- Оплатите покупки картой **Сбербанк** или **ВТБ24**
- Проверьте задолженность по налогам в **ФНС**
- Позвоните родителям с помощью **Tele2** или **Мегафон**
- Получите cashback в банке **Открытие** или **Тинькофф**
- Посмотрите курс акций на **ММВБ**
- Измените свой профиль или добавьте новый skill в **LinkedIn.com**
- Купите в **Starbucks** латте с помощью карты лояльности
- Заплатите за покупку с помощью **PayPal**
- Кликните на **Google AdWords** ссылку
- Выберите рекомендованный продукт в **EBay.com**
- Купите песню в **Apple iTunes**
- Забегите за хлебом в **Пятерочку**
- Получите ОСАГО в **Росгосстрах**

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

ORACLE®

Демонстрация



[Download demo](#)