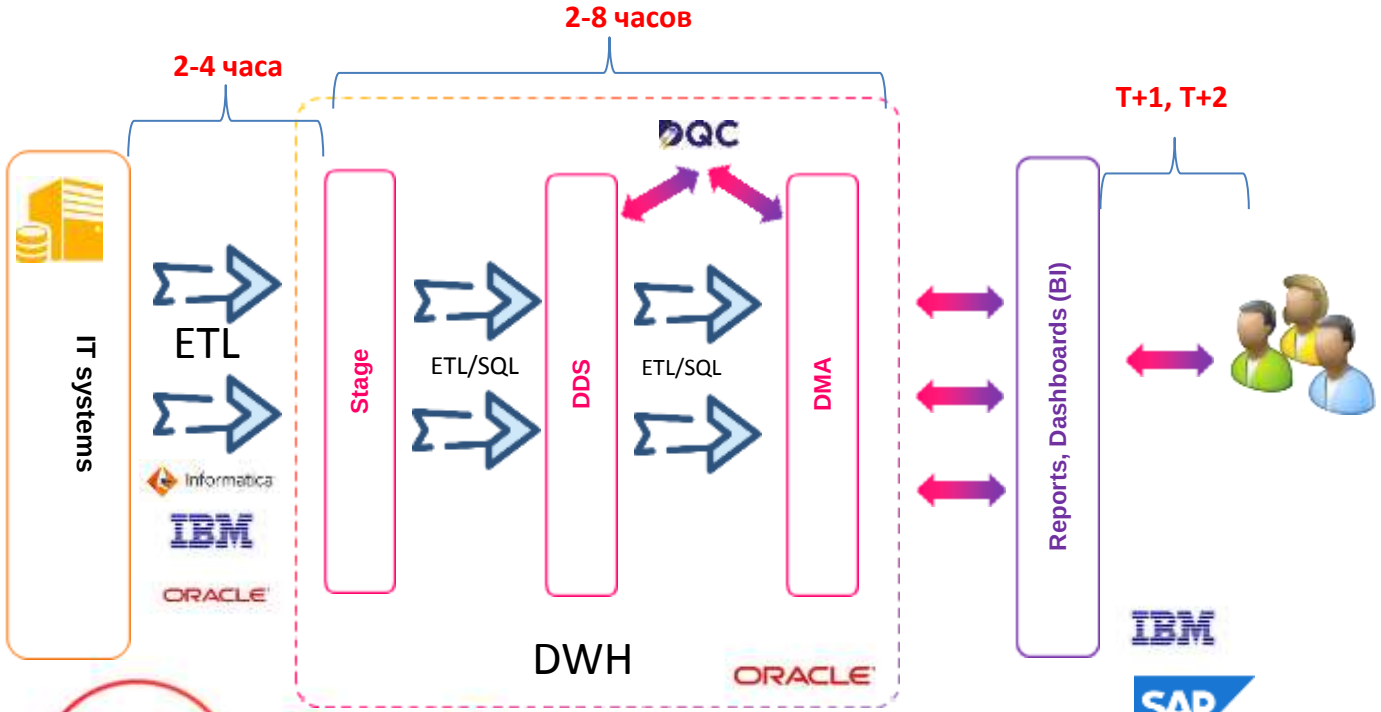
Neoflex
IT-solutions for business

Построение промышленной DWH без вендского ПО- Миф или реальность?

Орлов Олег- Главный Архитектор

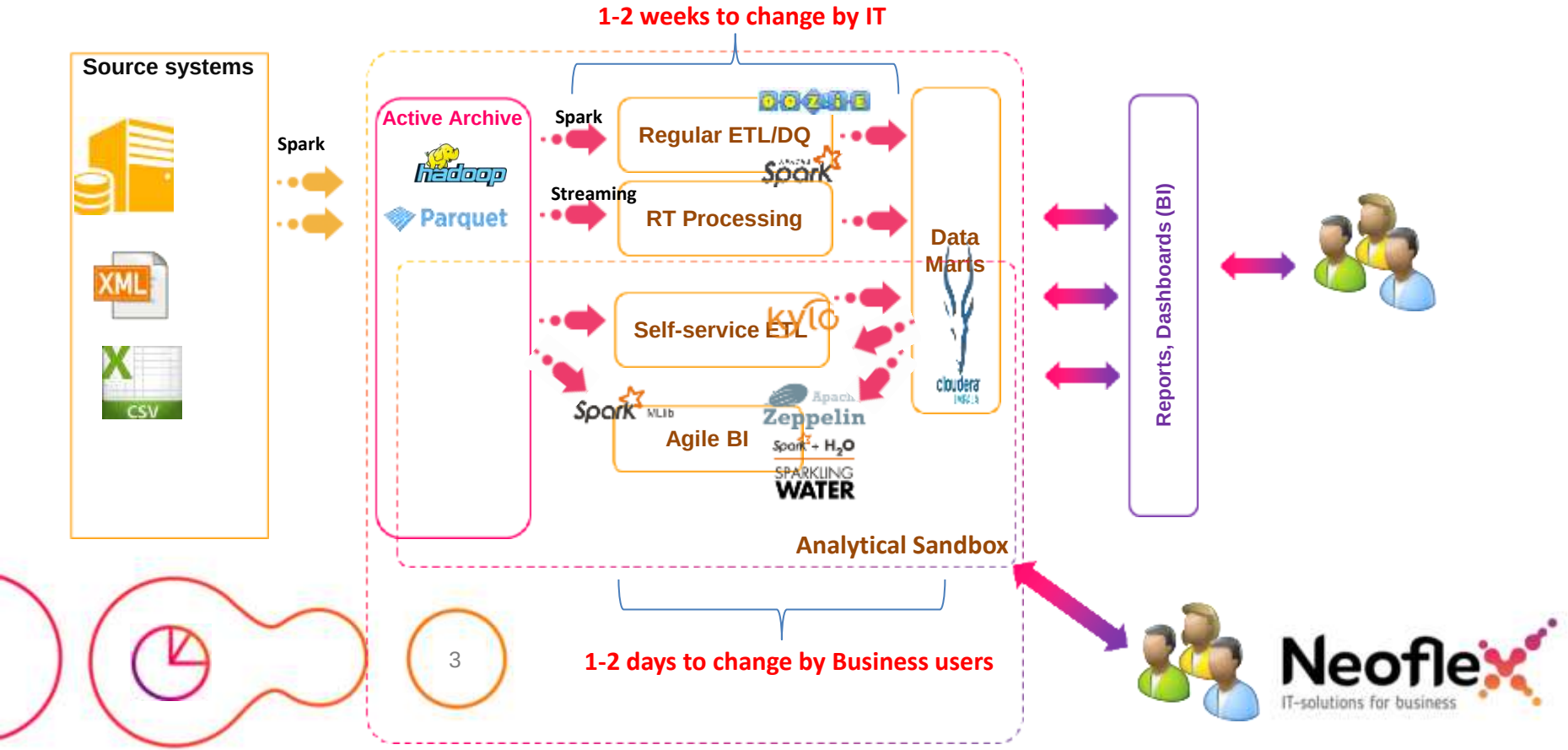
Компания Неофлекс

КЛАССИЧЕСКОЕ ХРАНИЛИЩЕ ДАННЫХ



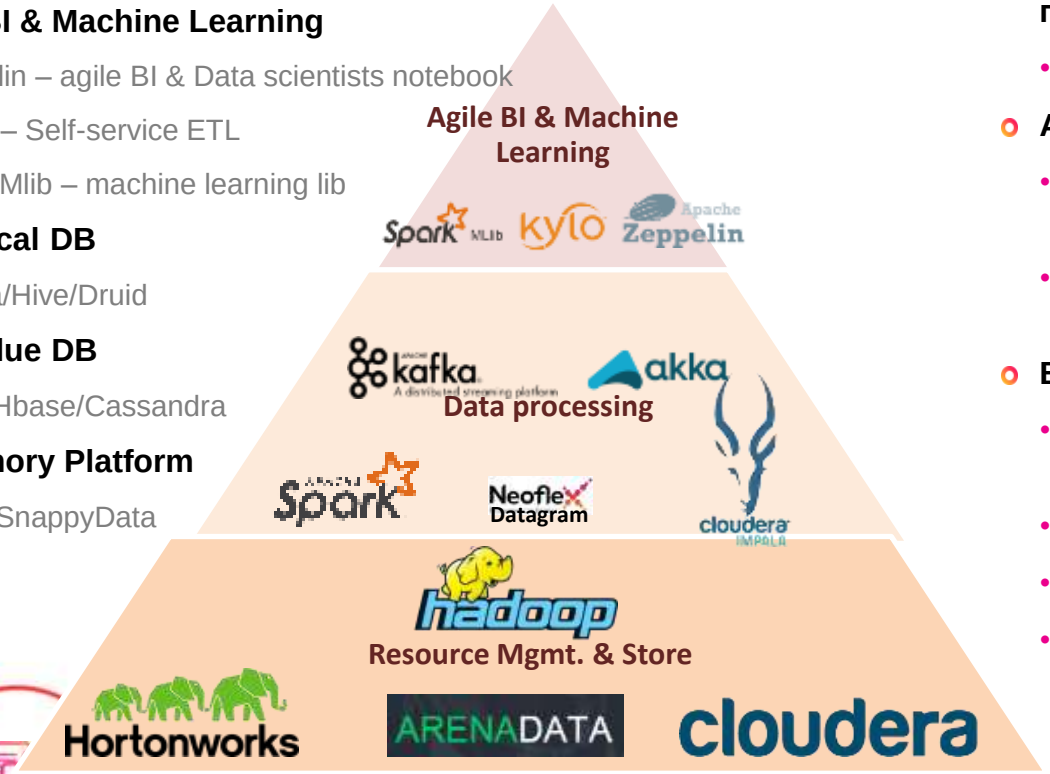
Neoflex
IT-solutions for business

СОВРЕМЕННАЯ АРХИТЕКТУРА



ТЕХНОЛОГИЧЕСКИЙ СТЕК

- **Agile BI & Machine Learning**
  - Zeppelin – agile BI & Data scientists notebook
  - KYLO – Self-service ETL
  - Spark Mlib – machine learning lib
- **Analytical DB**
  - Impala/Hive/Druid
- **Key-value DB**
  - Kudu/Hbase/Cassandra
- **In-memory Platform**
  - Ignite/SnappyData



- **Использование промышленных BigData платформ**
  - Hortonworks HDP, Cloudera CDH
- **Apache Hadoop**
  - YARN – управление распределенными вычислениями
  - HDFS – распределенное хранение большого объема разнородных данных
- **Batch & Stream Data Processing**
  - Neoflex Datagram – Rapid Application Development Platfrm
  - Apache Spark & Spark Streaming
  - Apache Kafka – message backbone
  - Akka Streams – stream processing

Neoflex
IT-решения для бизнеса

ПРОЕКТЫ

Клиент:

Банк TOP-10

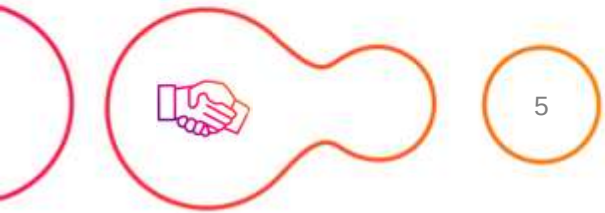
Цель проекта:

Построение витрин с информацией по клиентам и операциям банка и формирование обязательной отчетности

27M Операций ежедневно

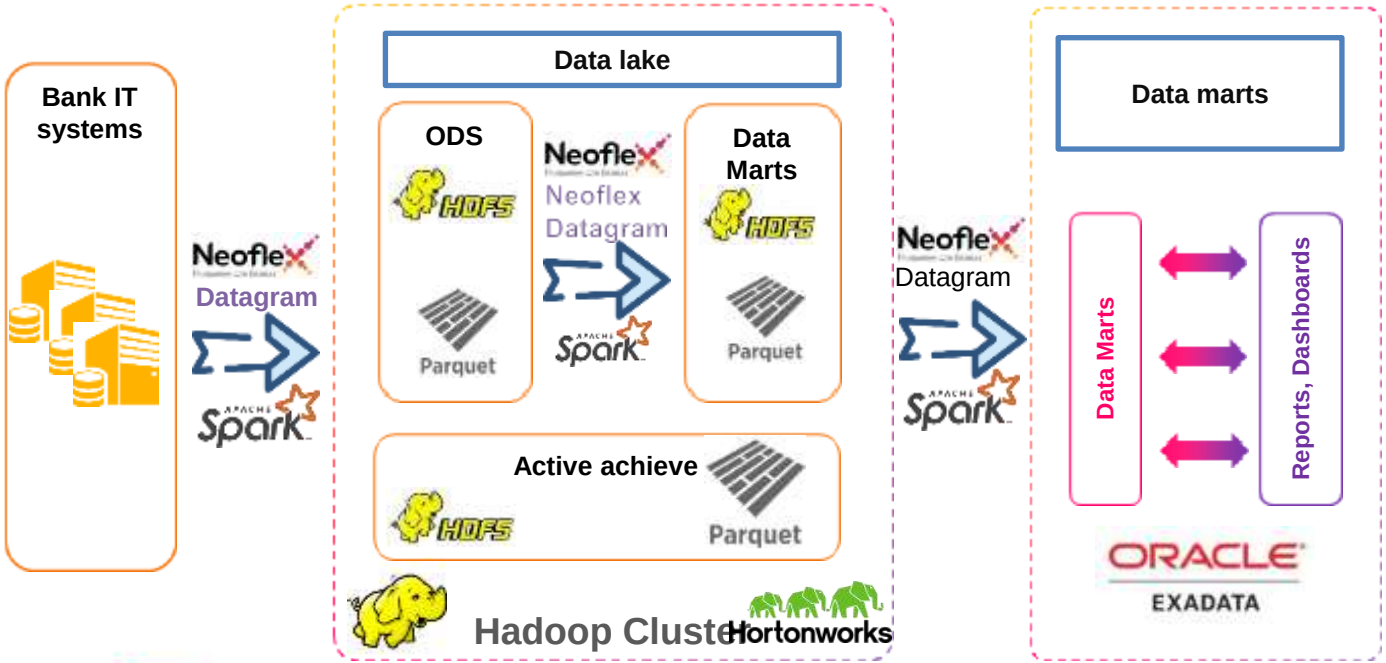
80+ Форм отчетности

1H Технологическое окно после закрытия дня



Neoflex
IT-solutions for business

АРХИТЕКТУРНЫЙ ОБЗОР



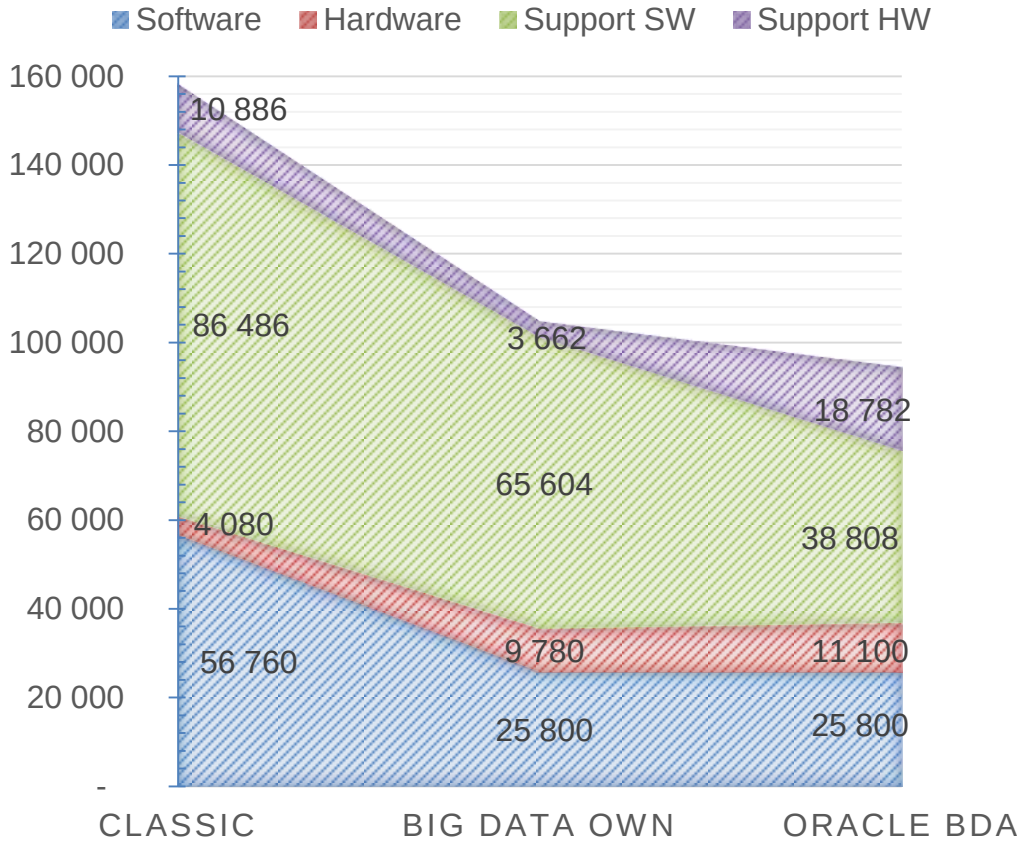
Особенности

- Комбинация подходов
- Высокая скорость обработки информации
- Экономия средств за счет использования BIG DATA

Neoflex
IT-solutions for business

СТОИМОСТЬ РЕШЕНИЯ

	Конфигурация	Итого TCO на 7 лет
Классическая архитектура	4 CPU Xeon (12 Cores) 256 Gb RAM HDD 12Tb Oracle Golden Gate Oracle DB 12 EE Oracle ODI 12 EE	185 752т.р.
Big Data индвидуальная конфигурация	6 узлов по: 2 CPU (16 Cores each) 256 Gb RAM HDD 8Tb Cloudera Enterprise 5 Oracle Golden Gate for BigData	104 846 т.р.
Oracle Big Data Appliance	6 узлов по: 2 CPU (22 Cores each) 256 Gb RAM HDD 8Tb Cloudera Enterprise 5 Oracle Golden Gate for BigData	94 490т.р.



Стоимость SW и HW Oracle – по данным Oracle
Иное – данные открытых источников
Расчет TCO по модели НКЦ
Все цены без НДС

ПРОЕКТЫ

BNP PARIBAS

Цель проекта:

Загрузка данных в DWH для формирования обязательной отчетности

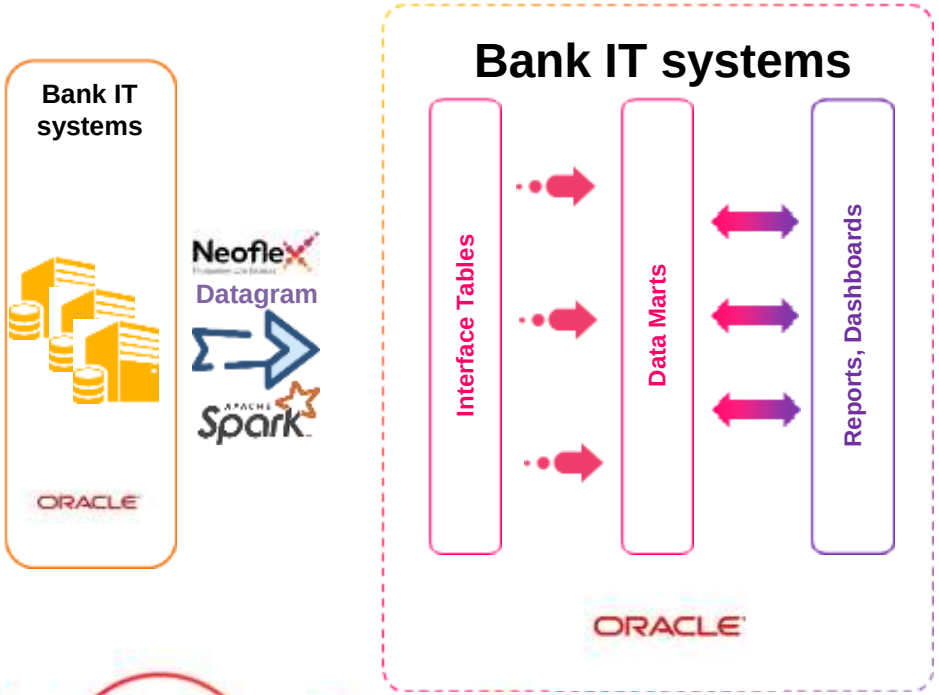
Результаты:

~80 Интерфейсных таблиц

27 Форм обязательной отчетности

Neoflex
IT-solutions for business

ПРИМЕНЕНИЕ КАК ETL



- **Особенности**
  - Сочетание классических технологий и BIG DATA
  - Высокая скорость обработки информации
  - Экономия за счет замены ETL зарубежного производства (IBM DataStage)

Neoflex
IT-solutions for business

ПРИМЕРЫ ПРОЕКТОВ: СИТУАЦИОННЫЙ ЦЕНТР

ЗАДАЧИ

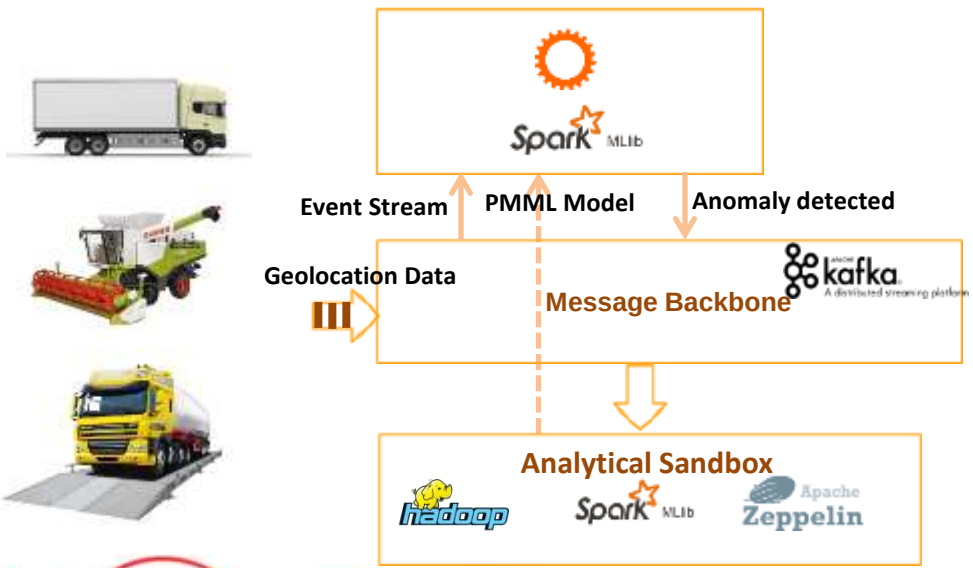
- On-line-мониторинг и детектирование подозрительных инцидентов
- Передача инцидентов в ситуационный центр
- «Машинное» выявление закономерностей
- Аналитическая отчетность

КЛИЕНТ

- Крупный агрохолдинг:
  - 65 000 га посевных площадей
  - ~ 600 единиц уборочной техники
  - ~ 1000 единиц машин для транспортировки
  - ~ 300 сообщений в секунду

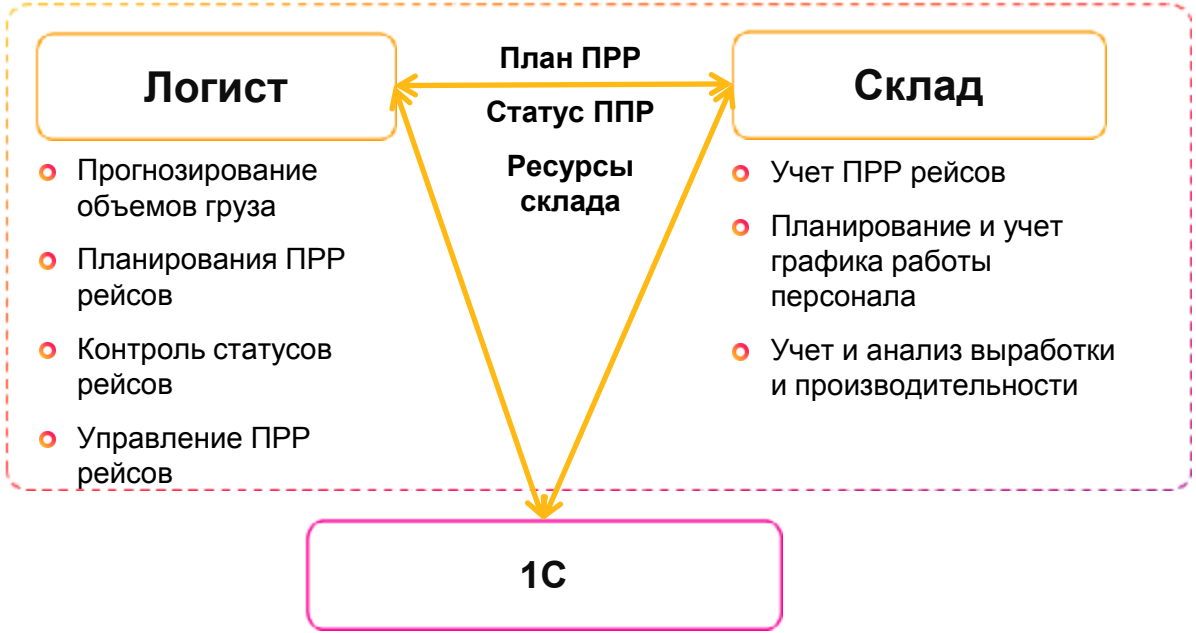
Neoflex
IT-solutions for business

СИТУАЦИОННЫЙ ЦЕНТР



- **Задача**: Выявление аномалий на потоке данных
  - Геолокация
  - Данные с пунктов взвешивания
- BigData & OnLine Analytics
  - Применение методов Machine Learning на потоке данных
    - K-Means, ARIMA…
  - Высокая масштабируемость
    - N x 100 000 транзакция в секунду
  - Немедленная реакция
    - Оповещение ответсвенных
    - Разбор ситуации ответственным сотрудником
- Analytical Sandbox
  - Разработка моделей выявления аномалий
  - Экспорт модели в PMML формате для онлайн обработки событий

ПРОЕКТ: ЦЕНТР УПРАВЛЕНИЯ ПЕРЕВОЗКАМИ



Крупная логистическая компания:

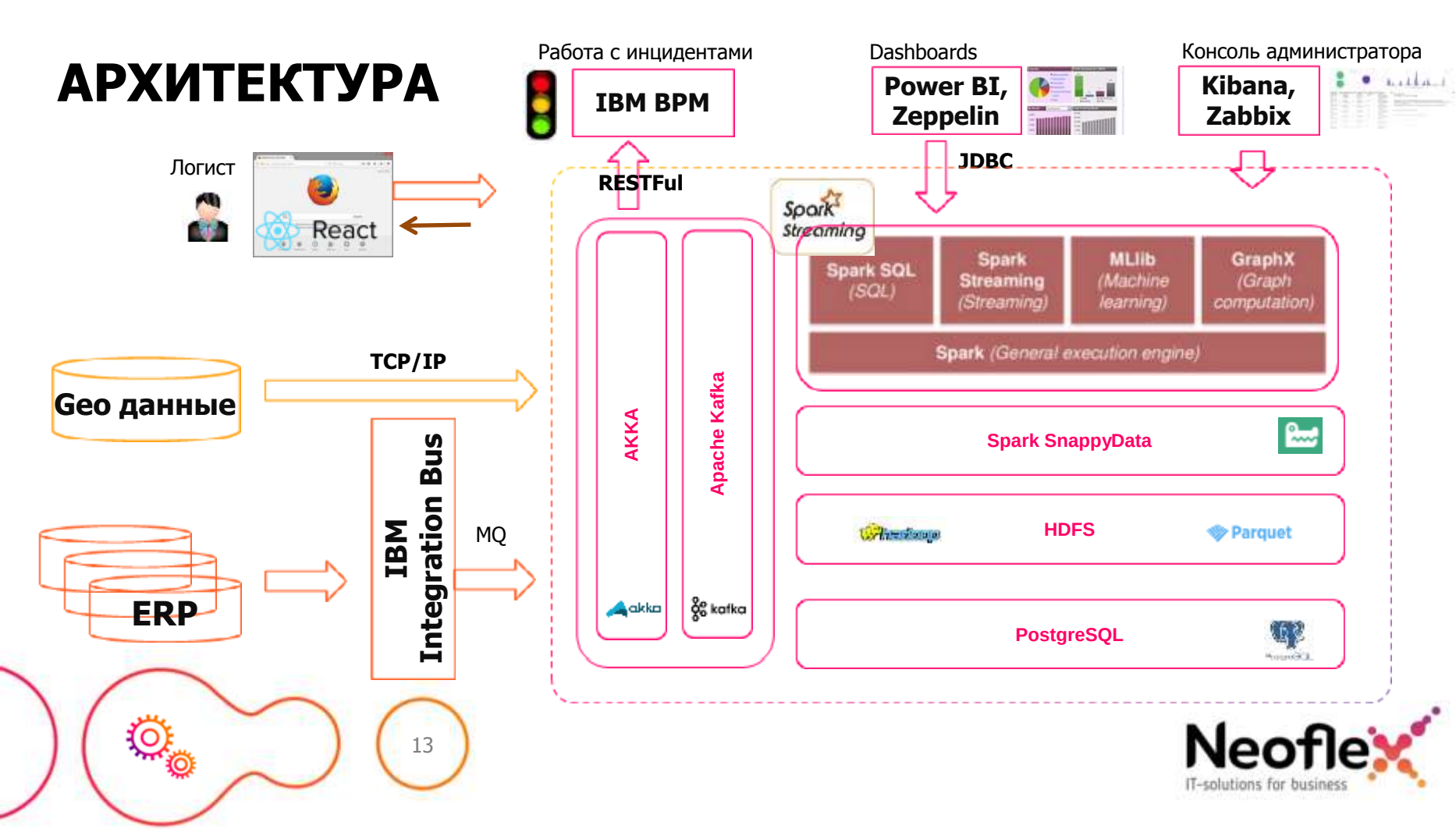
- Зона обслуживания - 100 000 городов
- 120 филиалов
- 1,5 млн клиентов ежегодно

Задачи:

- Снижение нагрузки складов
- Снижение времени доставки грузов

Neoflex
IT-solutions for business

АРХИТЕКТУРА



Neoflex
IT-solutions for business

ПРОЕКТЫ: ОЦЕНКА КРЕДИТНОГО РИСКА

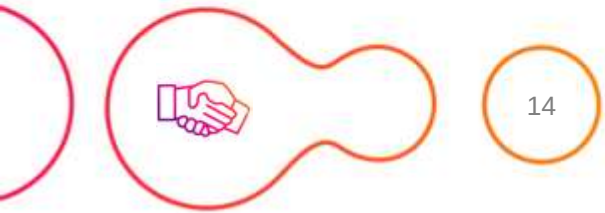
Клиент

Банк TOP-5

27МЛН Счетов

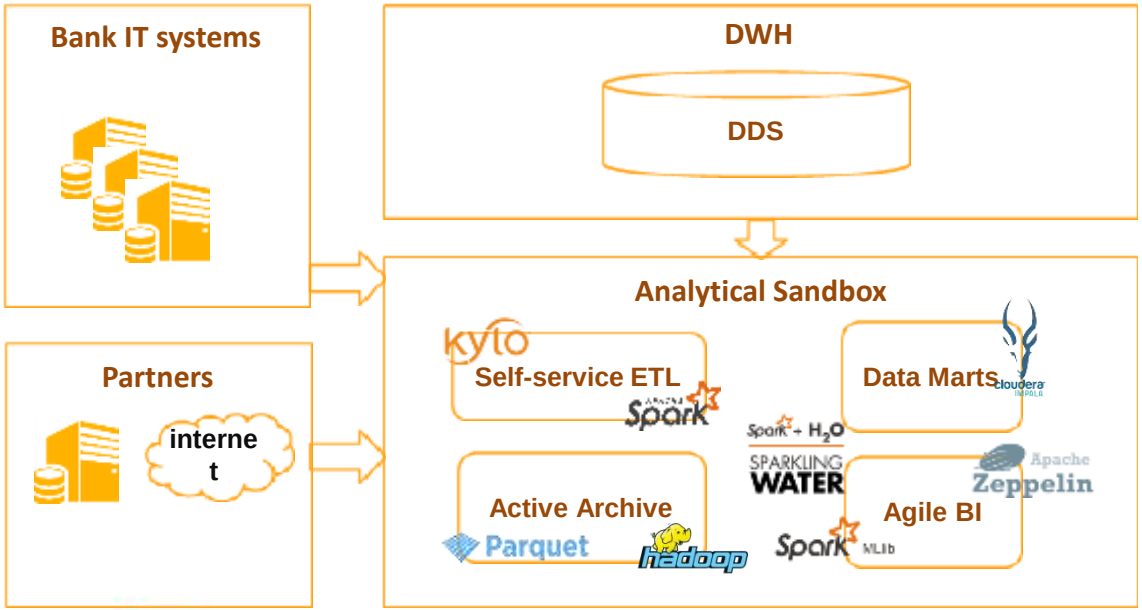
Цель проекта:

Построение аналитической “песочницы” для оценки заемщиков и отчета по факторам кредитного риска



Neoflex
IT-solutions for business

BIGDATA ANALYTICS & SELF-SERVICE BI

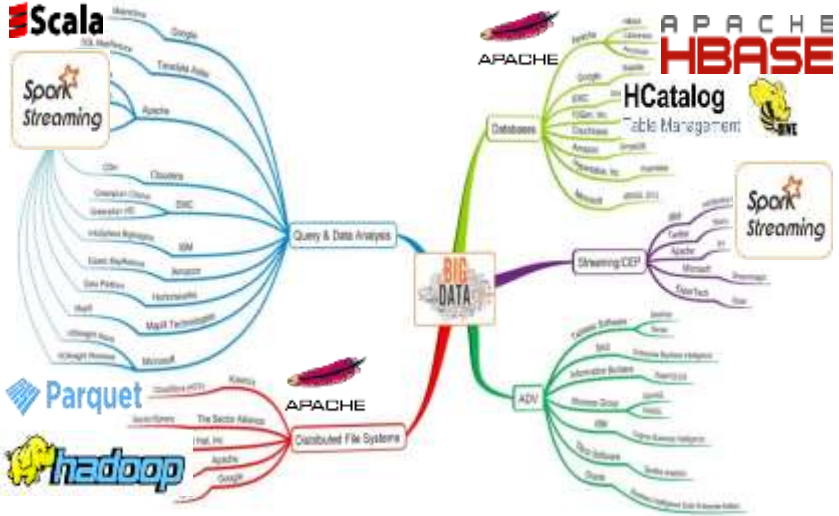


- **Особенности**
  - Требуются данные из систем Банка
  - И из внешних систем партнеров (БКИ, информация о банкротствах, налоговая отчетность клиентов)
  - Минимальная зависимость от IT department
- Self-service ETL для построения витрин
- Notebook Apache Zeppelin - Self-service BI
  - Machine Learning для построения моделей оценки заемщиков
- Построение Datamarts для отчетов
- Spago BI – средство построения отчётов

Neoflex
IT-solutions for business

ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ OPEN STACK

- Большое количество разнородных технологий
- Неоднородный , сложный в сопровождении код
- Низкая скорость разработки
- Ошибки стыковки технологий
- Высокая стоимость и дефицит квалифицированных специалистов на рынке



Neoflex
IT-solutions for business

NEOFLEX DATAGRAM

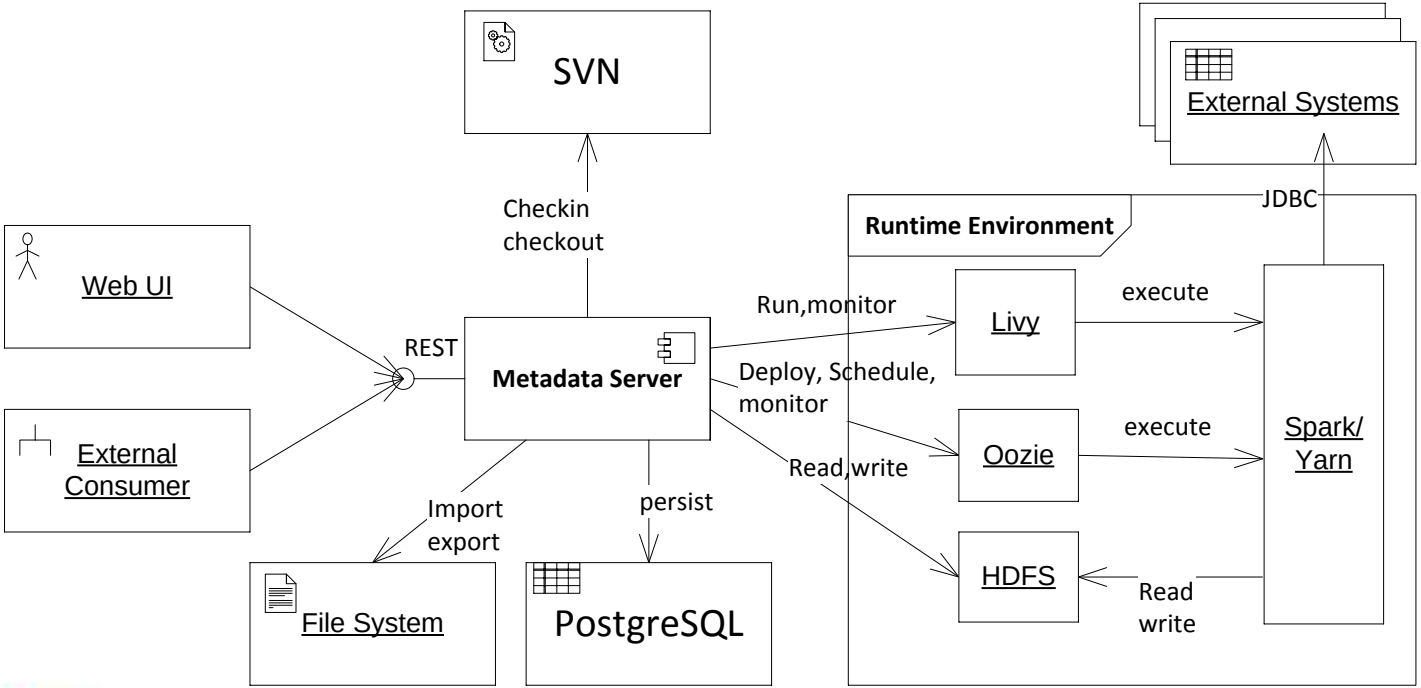
Neoflex Datagram

Фреймворк разработки приложений для (потоковой) обработки больших данных

- **Разработка**
  - Описание внешних источников
  - Дизайн трансформаций и последовательностей заданий
  - Генерация и компиляция Scala/Spark кода
- **Развертывание** приложений в среде исполнения
- **Планирование** запуска приложений
- **Мониторинг и управление**

Neoflex
IT-solutions for business

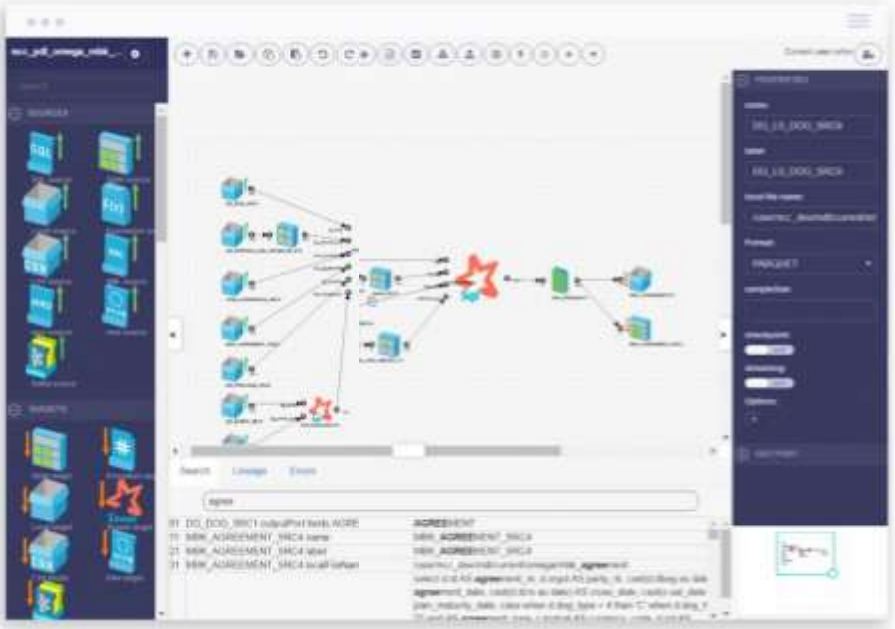
ОБЗОР АРХИТЕКТУРЫ



Neoflex
IT-solutions for business

ДИЗАЙНЕР ТРАНСФОРМАЦИЙ

- Визуальное проектирование маппингов и потоков данных
- Поиск элементов в сложных трансформациях
- Валидация отдельных элементов
- Трассировка преобразований отде полей

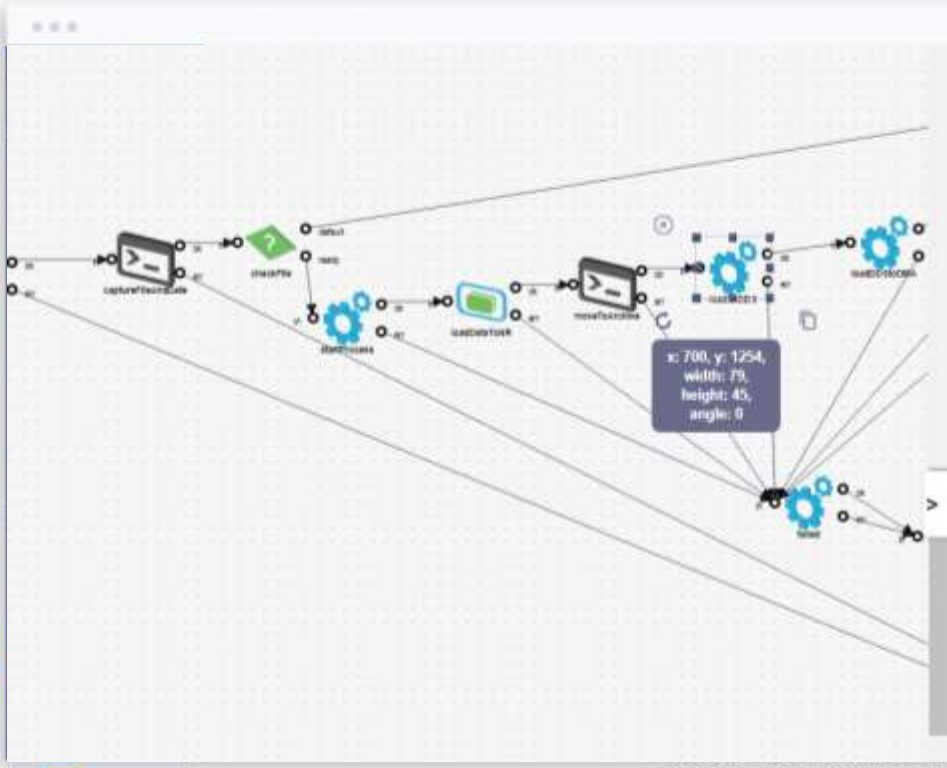


- Валидация трансформаций на предмет типичных ошибок
- Использование метаданных для описания источников данных
- Просмотр/преобразование/исполнение исходного кода
- Частичное исполнение трансформаций и просмотр результатов

Neoflex
IT-solutions for business

ДИЗАЙНЕР ПРОЦЕССОВ

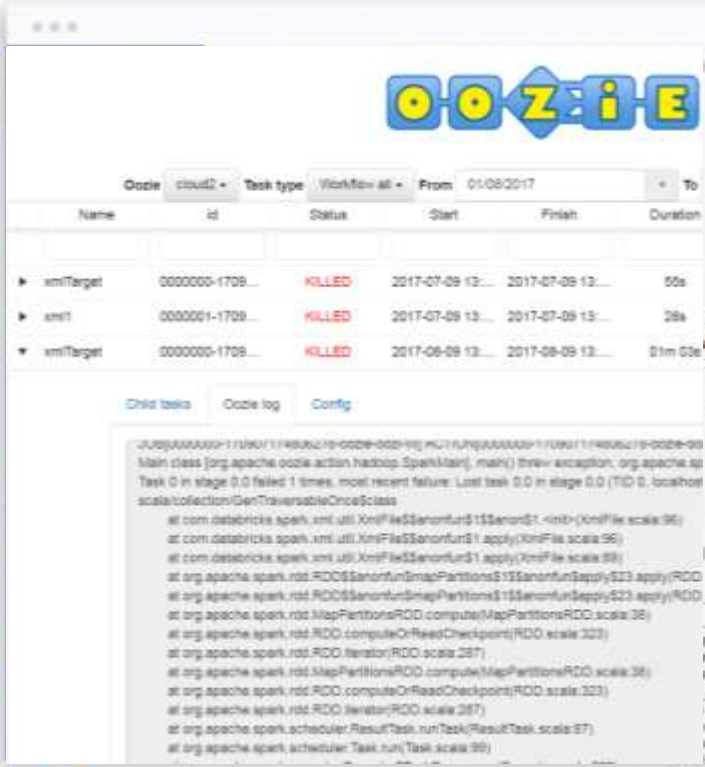
- Визуальное проектирование последовательностей задач
- Последовательное и параллельное выполнение задач
- Использование правил
- Обеспечение вложенности
- Запуск произвольных shell и java процедур
- Среда исполнения - Apache Oozie



Neoflex
IT-solutions for business

МОНИТОРИНГ

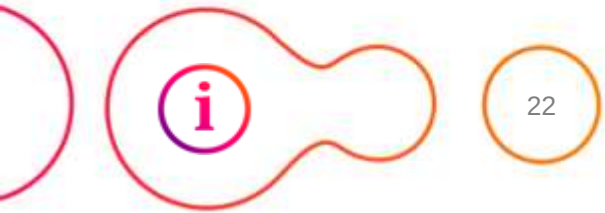
- **On-line мониторинг** сред исполнения - Apache Oozie/Livy
- **Просмотр** статусов задач и подзадач, параметров запуска, логов исполнения и ошибок баз данных
- **Старт, остановка и повторный запуск** выполнения заданий
- Просмотр **истории** запуска заданий
- Просмотр отклонённых записей (**rejects**)



Neoflex
IT-solutions for business

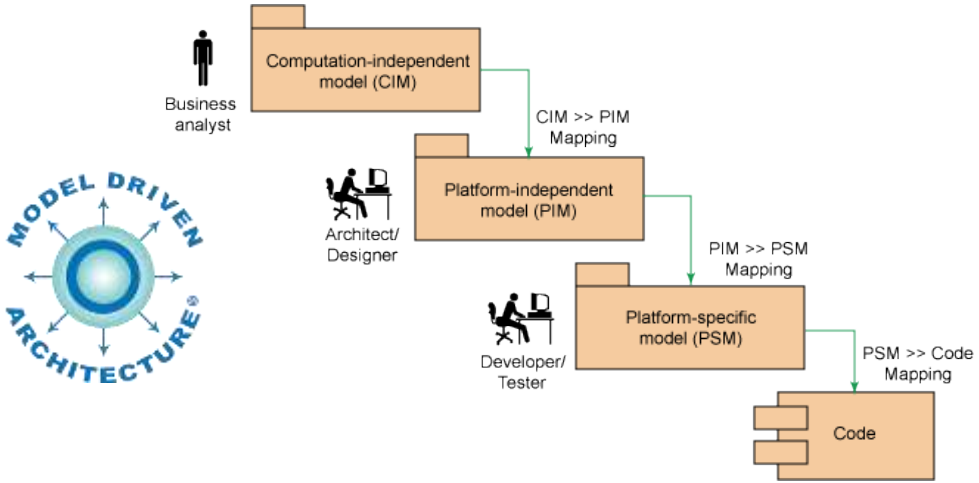
КЛЮЧЕВЫЕ ОСОБЕННОСТИ NEOFLEX DATAGRAM

- **Sources/Targets**
  - RDBMS sources/targets via JDBC (including stored procedures)
  - Text/structural/HDFS specific sources/targets: CSV, xml, avro, json, ORC, PARQUET, Spark Streaming support
  - Apache Hive, Apache Kafka
- **Типы трансформации**
  - join, sort, aggregation, union, selection, projections, pivot, explode arrays, sequence generation
  - Spark specific: Spark SQL – perform arbitrary SQL query over data streams (with Spark Catalyst support)
  - Model Based Analysis using Spark MLLib (decision trees, SVM, logistic regression etc.)
  - JBoss Rules (Drools) - Business Rule Management System
- **Versioning and teamwork (opt. locking, svn, projects)**
- **Multi-Environment support (dev->test->prod, Import/Export metadata, rewrite ursl)**
- **Security (ldap/Kerberos/passwords encryption/role model)**
- **Supplement tools (hdfs/livy/oozie consoles, Object Explorer)**



Neoflex
IT-solutions for business

КАК ЭТО РАБОТАЕТ



- Фреймворк разработан в архитектуре, управляемой моделями (**MDA**).
- Для управления моделями используется Eclipse Modelling Framework (**EMF**).
- Для хранения моделей используется **PostgreSQL/Hibernate/Teneo**.
- Для валидации моделей, преобразований модель-в-модель (**M2M**) и модель-в-текст (**M2T**) используется язык **Eclipse Epsilon**.
- Фреймворк использует следующие логические пакеты объектов: Authentication, Relational, ETL, Runtime, DWH, UI, Metadata.

Neoflex
IT-solutions for business

CONTACTS

Орлов Олег

Главный архитектор

www.en.neoflex.ru

Neoflex
IT-solutions for business