

ДАТА-ПЛАТФОРМА НОВОГО ПОКОЛЕНИЯ В DIGITAL-БАНКЕ



Александр
Заболоцкий
СТО по дата-технологиям

Альфа Банк



Петр
Пашков
Управляющий директор

▶ data sapience

ОСНОВНЫЕ ФАКТЫ ОБ АЛЬФА-БАНКЕ

Альфа Банк — универсальный
фиджитал-банк

Активы банка	→	11,3 трлн рублей
Чистая прибыль	→	210 млрд рублей
Клиенты ФЛ	→	37 млн / 3,2 трлн рублей
Клиенты ЮЛ	→	2 млн / 6 трлн рублей



ХРАНИЛИЩЕ В ЦИФРАХ Альфа Банк (2023)

170

систем-
источников
для DWH

300

пользовательских
витрин

4.5 Tb

ежедневный прирост
новых данных

6 Pb

общий объем хранения
данных на 1 кластере

120к

процедур загрузки
данных
для DWH

8.5к

отчетных форм
выпускается
в квартал

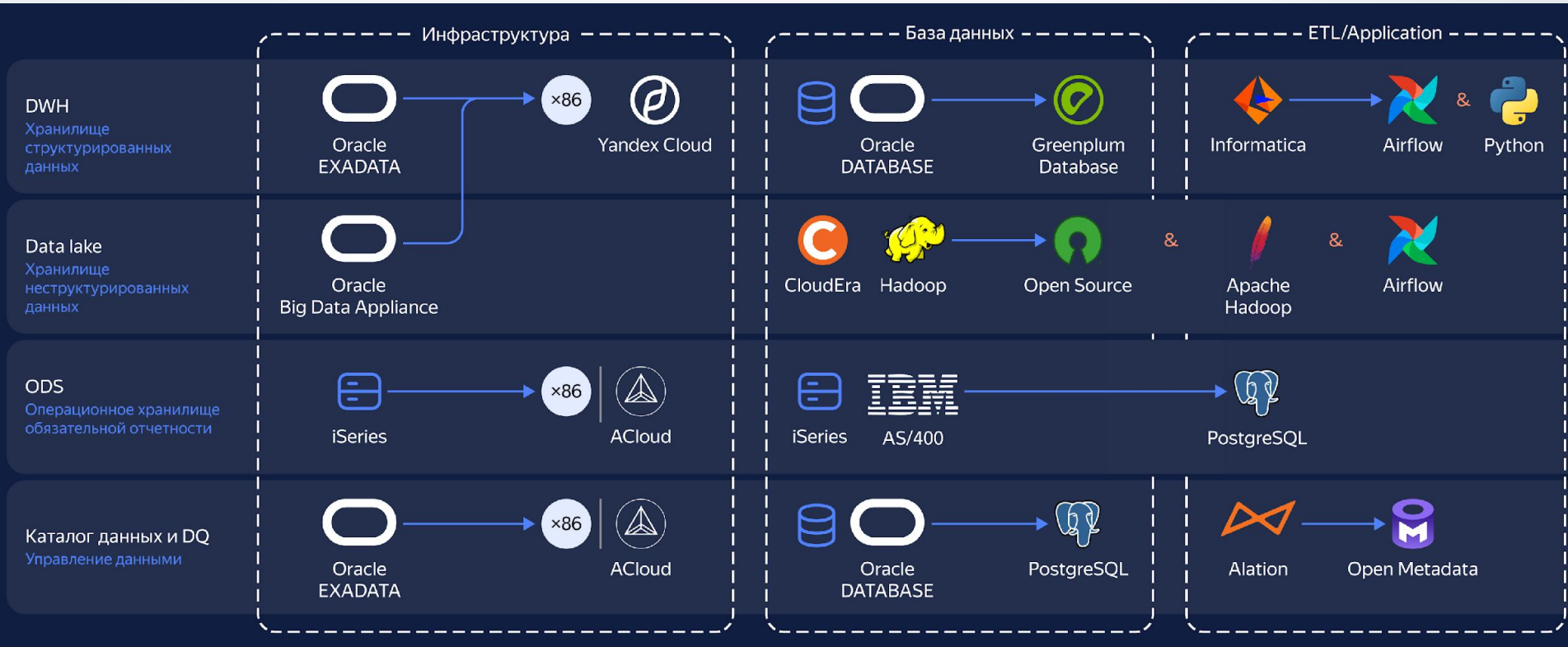
9к+

пользователей
DWH

ТЕХ.СТЕК ХРАНИЛИЩА Альфа Банк (2023)



РАЗВИЛКИ ПО ТЕХ.СТЕКУ



СТРАТЕГИЯ DATA-ФУНКЦИИ

Стратегия



NRT-доступ к данным
и NRT-отчетность



Lakehouse: DWH
и Data Lake в одном



MAU ↑
Эффективность ↑



Data-ready подход
для бизнеса



TCO ↓
эффективнее, чем у MPP



Решение проблем vendor-lock
и ИЗ

И ТРЕБОВАНИЯ К ПРОДУКТУ

Требования к продукту

Batch- и real-time обработка
из коробки



Высокопроизводительные
SQL compute



Возможность гетерогенной
работы из коробки



Независимое масштабирование
compute и storage



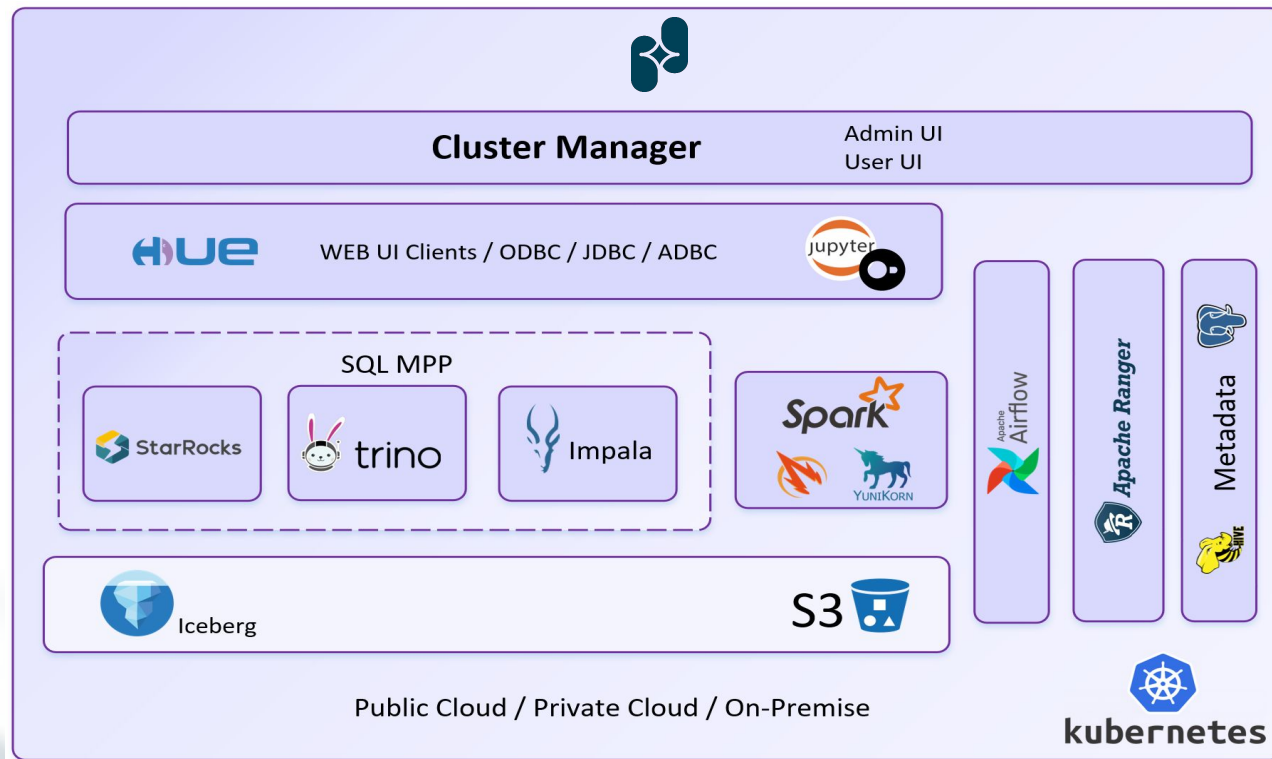
Мультитенантность в единой
инфраструктуре



Открытый интероперабельный
стек технологий



АРХИТЕКТУРА LAKEHOUSE Nova в Альфа Банк



Все в Kubernetes. Storage и Compute аппаратно разделены

ВНЕДРЕНИЕ BIG DATA PLATFORM

Бизнес-продукты

Построение отчетных форм

Power BI

Superset

Qlicksense

Tableau

Fine Bi

Per-ая отчетность

Web Apps

UI для доступа к аналитике

HUE

Dbeaver

Jupyter Hub

Управление и мониторинг

Сбор метрик, мониторинг, логирование

Prometheus

Grafana

Админ-консоль и ресурс-менеджер платформы

Kubernetes

#разделение compute и storage
#гармонизация загрузки в Big data и DWH

Платформа управления метаданными, каталог данных

OpenMetadata

DQMON

ПУМА

Слой передачи

Интеграционный слой

#новые возможности near real time обработка

Spark

Kafka

Flink

Обработка данных

Работа аналитическими запросами

Spark SQL

Trino

StarRocks

#единая точка работы со таблицами и файлами на sql like сервисе

Преобразование и обработка в структурированном виде

#sql like работа с данными

Hive

Impala

Iceberg

#замена недоступных элементов стека

Хранение данных

Хранение данных

#миграция на x86 стандарт инфраструктуры

Oracle

x86

Oracle БД

Greenplum

HDFS

S3

PostgreSQL

Tarantool

ClickHouse

Интеграционный слой

Обработка и оркестрации загрузки данных

AirFlow

Python

Spark

PXF

IPC

Debezium

новые возможности near real time загрузка

Flink

Kafka

Vision replicator

Oracle GoldenGate

☐ Существующий

☐ Потенциальный

☐ Альтернативы друг другу

● Долгосрочно выбывающие компоненты

ВНЕДРЕНИЕ BIG DATA PLATFORM

<i>ПОКАЗАТЕЛЬ</i>	<i>ТЕКУЩЕЕ СОСТОЯНИЕ</i>	<i>ПЛАНЫ</i>
Количество источников	40	600+
Количество NRT-отчетов и приложений	55	5к+
Объем сжатых данных (NRT + batch)	750 Tb	5 Pb
Количество таблиц	100	200к+
Количество пользователей	200	4к+
Объем кластера (кол-во ядер, объем данных)	1к ядер 2,4 Pb	2к ядер 5 Pb
Производительность загрузки (RPS upsert)	Avg – 4к Max – 10к	Avg – 8к Max – 20к



СПАСИБО ЗА ВНИМАНИЕ

Александр Заболоцкий
СТО по дата-технологиям, Альфа-Банк
AZabolotskiy@alfabank.ru

Петр Пашков
Управляющий директор, Data Sapience
Petr.Pashkov@glowbyteconsulting.com