**ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК**

**RT.DATALAKE**

2025

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

| Термин | Определение |
| --- | --- |
| Разработчик | Организация, которая выполняет разработку задач (в том числе анализ требований, проектирование, приёмочные испытания) в процессе жизненного цикла. |
| Пользователь | Лицо или группа лиц, извлекающих пользу из Hadoop в процессе его применения. |

ОГЛАВЛЕНИЕ

[ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ 2](#_Toc53136463)

[1 ФУНКЦИОНАЛЬНЫЕ ХАРАКТЕРИСТИКИ 4](#_Toc53136464)

[2 ПРИНЦИПЫ ФУНКЦИОНИРОВАНИЯ RT.DATALAKE 5](#_Toc53136465)

[2.1 Структура и функционирование RT.DataLake 5](#_Toc53136466)

[2.2 Реализация информационного потока 6](#_Toc53136467)

[3 БИЗНЕС-ПРОЦЕССЫ 7](#_Toc53136468)

ФУНКЦИОНАЛЬНЫЕ ХАРАКТЕРИСТИКИ

**RT.DataLake** обеспечивает реализацию следующих функций:

Быстрая обработка больших массивов неструктурированных данных.

Хранение данных.

Подготовка данных для выгрузки в другие хранилища и информационные системы.

Сбор статистики.

**RT.DataLake** обладает следующими особенностями:

1. Высокая мощность вычислений, достигаемая за счёт большого количества узлов.

Лёгкая масштабируемость за счёт увеличения количества узлов при росте объёма данных.

Высокая устойчивость к отказам.

Надёжность хранения данных, обеспечиваемая механизмами файловой системы.

Техническая поддержка, обеспечиваемая Разработчиком.

Протестированная совместимость компонентов.

Отсутствие проприетарных компонентов.

Возможность быстрого старта для оценки и тестирования.

ПРИНЦИПЫ ФУНКЦИОНИРОВАНИЯ RT.DATALAKE

Структура и функционирование RT.DataLake

**RT.DataLake** включает в себя следующие функциональные компоненты, предназначенные для решения соответствующих комплексов задач:

1. **HDFS** (Hadoop Distributed File System) — распределённая файловая система. Является технологией хранения файлов на узлах хранения данных (DataNodes), управление которым осуществляется узлом имён (NameNode).Надёжное хранение обеспечивается за счёт репликации информационных блоков между узлами данных.

**YARN** (Yet Another Resource Negotiator) — система планирования заданий и управления кластером. Представляет собой набор системных сервисов, обеспечивающих совместное использование, масштабирование и надёжность работы распределённых приложений. Является интерфейсом между аппаратными ресурсами кластера и приложениями, использующих его мощности для вычислений и обработки данных.

**MapReduce** — платформа программирования и выполнения распределённых вычислений с использованием отдельных стадий обработки больших объёмов данных: map (предварительная обработка данных), combine (частичная агрегация данных по ключу), shuffle (разделение данных на секции и сортировка промежуточных пар по ключу), reduce (свёртка обработанных данных).

**Hive** — СУБД, которая в качестве языка запросов использует HiveQL и которая позволяет выполнять сложные запросы над данными, хранимыми в **HDFS**.

**Knox** — REST API и шлюз приложений для компонентов Hadoop, обеспечивает единую точку доступа для всех HTTP-соединений с кластером Hadoop и систему единой аутентификации Single Sign On (SSO) для сервисов и пользовательского интерфейса компонентов **RT.DataLake**.

**Hue** (Hadoop User Experience) — веб-интерфейс для анализа данных на **RT.DataLake**.

Схема компонентов **RT.DataLake** представлена на схеме ниже.



Реализация информационного потока

Схема потоков данных **RT.DataLake** представлена на рисунке.



БИЗНЕС-ПРОЦЕССЫ

**RT.DataLake** может выполнять автоматизацию следующих бизнес-процессов:

1. Хранение большого объёма несистематизированных данных, которые собираются из большого количества источников и которые должны храниться продолжительное время для будущих целей компании или, например, для соблюдения законодательных мер.
2. Обогащение хранимых данных из иных источников, например, обогащение клиентских профилей данными из социальных сетей для маркетинговых целей.
3. Сбор информации о работе серверов и сети, а также действиях пользователей в системе для определения причин нарушения безопасности.
4. Структуризация различных типов данных и их использование для аналитики и дальнейшего построения модели и проверки предположений.