# Техническое задание на проведение консалтинга и внедрение платформы BigData

## Заказчик – Департамент Управления Хранилищем и Анализа Данных

## Предпосылки

Существующее хранилище построена на СУБД Postgre (строчная СУБД, колоночная распределенная citus), на нем построена система продуктовой, аналитической и управленческой отчетности на базе Apache SuperSet.

Также по этим данным строится пользовательская отчетность и проводится аналитика для решения локальных задач.

ELT/ETL батчевый, с использованием Apache Airflow. Идет разработка по внедрению CDC Kafka (Postgre источники, необходимо еще внедрить Oracle)

Продуктовая аналитика и фичи для ML-моделей не рассчитываются.

Начаты работы по внедрению DataCatalog – OpenMetadata. Внедряется Kafka, DBT.

Общий объем собранных данных порядка 40ТБ.

Отдельно хранение “холодных” данных в S3 не организовано.

Проектируемое решение должно поддерживать существующую функциональность (хранилище) + Data Lake + ML/DS и отчуждаемое от поставщика услуг по консалтингу.

##



## Область применения и границы

В рамках проекта предполагается проведение следующих работ:

* Проведение исследования текущего состояния системы и сценариев её использования;
* Разработка вариантов архитектуры нового хранилища и платформы данных, включающей корпоративную шину данных (на основе результатов проведенного исследования из пункта выше). Предлагаемая архитектура должна быть адоптивной к изменяющимся потребностям в получении и использовании данных, работы с данными;
* Разработка вариантов плана интеграции и способов организации работ по интеграции;
* Оценка ТСО для предложенных вариантов;
* Оценка стоимости и сроков интеграции/+миграции на новое решение;
* Оценка стоимости и сроков обучения пользователей.

В рамках проекта не предполагается перенос данных и процессов с текущей системы.

## Цели и задачи Департамента

Цели:

* Улучшить текущую скорость по доступности данных в landing слое;
* Дать/сохранить возможность ad-hoc аналитики для пользователей;
* Обеспечить горизонтальную масштабируемость решения;
* Снизить стоимость владения хранилищем и дата-платформой;
* Обеспечить Банк корпоративной шиной данных, позволяющей потребителям данных не строить индивидуальные соединения к источникам данных.
* Перенести нагрузку аналитических запросов от продуктов с мастер систем на хранилище данных (например, большие выписки за давние периоды, статическая информация и т.п.)

## Задачи

|  |  |  |
| --- | --- | --- |
| Блок | Вопросы | Ожидаемый результат |
| текущая архитектура | оценка производительности текущего стека и определения его максимума в загрузке данных, производства расчетов | оценить возможности каждого функционального сегмента на предмет аппаратных и программных пороговых значений:* Apache Airflow
* Kafka
* DWH old (postgres)
* DWH Cluster (citius)
* Apache Super Set
* Open Metadata Data Catalog
 |
| текущая архитектура | корректировка параметров текущей архитектуры для повышения производительности | рекомендации по каждому элементу |
| текущая архитектура | мониторинг элементов, важные аспекты мониторинга кроме стандартных CPU/RAM, up/down service | набор метрик для каждого элемента |
| текущая архитектура | метрики качества данных | необходимый набор метрик, их приоритезация, контроли, процесс создания/выбора |
| ETL | анализ/code review текущих процессов etl для осуществления целостной политики data management - сбор хранение обработка, контроль целостности и достоверности данных, ведение веток разработки в Git (корпоративный gitlab.cbk.kg) | Унификация процессов и процедур |
| будущая архитектура | Детализация векторов развития1. 1. Корпоративная шина данных

  2. Миграция продуктов с мастер систем в DWH за статичными данными  3. внедрение Big Data DS/ML, минимум для целей CVM(customer value management). решение должно быть таким, чтобы результат мог быть поставлен в другие платформы в том числе | сформировать аппаратную и программную архитектуру каждого вектора, ТСО, сроки внедрения в случае реализации Поставщиком* data lake, хранение данных

обработка и расчет данных с использование технологий DS/ML |
| будущая архитектура | горизонтальная клиентская витрина для продуктовых аналитиков | аналитики продуктов могут через web интерфейс (опционально) строить воронки клиентов |
| будущая архитектура/dataset | оценка оптимальности собранных датасетов, практики как лучше |  |
| будущая архитектура | Data Lineage | методологии, инструменты, практические результаты и примеры |
| Процессы/политики | определение методологий сбора, нормализации, хранения, доступа к данным и оценки качества данных | выбор методологии и необходимые артефакты для нее на всех этапах. Потери данных, работа с потерей данных от источника |
| Процессы/политики | data governance /методологии инструменты/ текущие в том числе) |  |
| Процессы/политики | риски - выявление имеющихся рисков и рекомендации по митигации | доступ к персональным/чувствительным данныминструменты для обеспечения безопасности |

Критерии отбора победителя

* цена отдельно на услуги консалтинга и услуги интеграций и внедрения элементов будущей архитектуры;
* сроки на основании детального плана работ на консалтинг и интеграционные работы (внедрение);
* артефакты закрытия этапов работ (что и в каком виде будет предоставлено, передано и т.п.);
* опыт проведения подобных с работ (оба блока: консалтинг и реализация архитектуры) с предоставлением рекомендаций прошлых Заказчиков, похожих по структуре и объему бизнеса, данных и развитию;
* Риски реализации проекта и меры митигации рисков;
* Выделенные инженеры на проект (не человеко-часы компетенций).