Описание по разработке модели данных и построении корпоративного хранилища данных ДО АО ВТБ Казахстан

**Информация по имеющимся в Банке инструментам и системам-источникам:**

|  |  |
| --- | --- |
| СУБД | * Vertica Analytics Platform
 |
| ETL-инструмент | * Informatica PowerCenter
* Informatica PowerExchange
* Oracle Data Integrator
 |
| BI-инструменты | * Qlik Sense
* SAP BO BI
 |
| Системы-источники | * СУБД: Oracle, MS SQL Server, PostgreSQL
* Данные из файлов (Excel, csv, txt)
* Количество источников: 4
 |

**Цель:** построение корпоративного хранилища данных – системы, из которой возможно получение исторических, аналитических и консолидированных бизнес-показателей Банка, оперативных данных в режиме реального или около-реального времени (near real-time); обеспечение доступности финансовых данных с задержкой не более 0,5 - 1 часа.

**Задачи**:

1. Разработка концепции КХД;
2. Разработка архитектуры КХД;
3. Разработка модели данных;
4. Развертывание и разработка КХД (в том числе по части загрузки данных);
5. Автоматизация отчетности;
6. Проведение тестирования и приемо-сдаточные работы;
7. Гарантийное сопровождение не менее 6 месяцев (в том числе по настройке процессов загрузки данных).

**Концепция КХД**

1) создание и ведение справочников, используемых для построения финансовых и аналитических отчетов;

2) описание принципов владения информацией, разграничения предметных областей, прав доступа для пользователей;

3) обеспечение доступности и качества данных;

4) обеспечение хранения версионности отчетов (после отправки отчетов ставить отметку, отправленная версия отчета должна быть закрыта для изменений);

5) обеспечение управления данными;

6) простота разработки и сопровождения системы;

7) описание тайминга основных процессов, для достижения целей проекта.

**Архитектура КХД**

1) описание процедур интеграции с системами-источниками для извлечения информации и порядка выгрузки данных из систем-источников;

2) описание механизмов контроля целостности и консистентности при загрузке информации из источников в КХД и управления данными;

3) реализация процедур фиксации процессов, выполненных с ошибками при загрузке данных из источников, описание типов ошибок и возможные варианты решения критичных ошибок;

4) описание требований к характеристикам программного и аппаратного обеспечения в конфигурации КХД;

5) описание архитектуры кластеров в массово-параллельной аналитической СУБД (Vertica), архитектуры хранения данных, слоев в КХД, ETL-процедур;

6) описание архитектуры слоя высокопроизводительных витрин данных на базе MPP-решения Vertica для использования BI-инструментами и в отчетах (для работы с оперативными данными и неоперативными);

7) обеспечение возможности работы со всем объемом накопленных данных с использованием различных фреймворков и языков обработки данных (SQL, BI, ML-инструменты);

8) обеспечение возможности линейной масштабируемости системы (как путем подключения новых источников данных, так и за счет расширения аналитических показателей);

9) описание и внедрение системы по контролю качества данных на каждом этапе загрузки данных (в том числе межформенные контроли); описание логики проверок качества данных;

10) описание и внедрение инструментов по управлению данными для описания, каталогизирования бизнес-показателей и увеличения прозрачности структуры и загрузки данных;

11) фиксация изменений, влияние изменений на витрины данных и отчеты (маркировка и актуализация изменений до оперативных витрин и отчетов).

Архитектура КХД также должна обеспечивать:

* + возможность хранения и извлечения данных как измененных, так и помеченных как удаленные;
	+ возможность импорта данных различных форматов (XML, CSV, XLS, TXT и др.) через программные интерфейсы для сбора информации в КХД, отсутствующей в системах-источниках, и иметь подробную документацию функциям экспорта/импорта данных;
	+ разграничение по ролевым доступам;
	+ возможность параллельных расчетов
	+ информационную безопасность и защиты от несанкционированного доступа;
	+ Поддержку работы с национальными символами на всех уровнях (в базе данных, приложениях, отчетах);
	+ возможность нотификации определенных пользователей при наступлении тех или иных событий (в т.ч. в ETL-процессах);

Архитектура также должна предусматривать загрузку, как структурированных исторических данных, так и неструктурированных и полуструктурированных данных в кластер Озера данных для дальнейшей обработки аналитиками и учеными по данным.

Загрузка данных в хранилище из систем-источников должна быть проработана с учетом того, что системы-источники являются высоконагруженными продуктивными системами.

**Модель данных**

1) формулировка бизнес-глоссария показателей;

2) определение взаимосвязей между показателями, описание логики верификации показателей;

3) формирование перечня справочников, реестра сущностей и показателей;

4) определение каждой сущности, ее компонентов и взаимосвязей между ними в структуре КХД;

5) построение ER-диаграмм для сущностей и витрин КХД;

6) описание глубины истории и периодичности обновления информации в хранилище на уровне каждого слоя КХД;

7) описание слоев КХД, их предназначение, методы хранения данных в каждой области, способы использования этих данных;

8) описание формирования и преобразования данных, методов загрузки данных (ETL-процессов), способов обновления данных (полное, инкрементальное обновление), проектирование витрин данных;

Модель данных КХД должна обеспечивать непротиворечивость данных, ретроспективность данных, доступность данных на любом уровне детализации.

Модель данных должна обладать модульной структурой и позволять последующее расширение/дополнение сущностями и показателями, как из вышеперечисленных систем-источников в этом ТЗ, так и из других новых систем.

**Анализ, подготовка и разработка системы КХД**

1) сбор и анализ (актуализация) требований по отчетам и бизнес-показателям, анализ периодичности формирования и детализации этих показателей;

2) подготовка информации по требованиям к программно-аппаратному комплексу;

3) разворачивание MPP-платформы, создание структур данных и слоев хранилища;

4) разработка ТЗ;

5) разработка ETL-процедур и витрин данных;

6) разработка отчетов;

7) разработка процедур по контролю качества на соответствие описанным правилам;

8) тестирование выгруженных из систем-источников сведений на соответствие требованиям. Корректировка модели данных, описания контролей и проверок данных в случае выявления ошибок и расхождений; корректировка ETL-процедур, алгоритмов расчета показателей в отчетах и витрин данных;

9) формирование предложений по процедурам приема-передачи, участие в тестировании результатов и нагрузочных испытаниях хранилища.

Список отчетов к автоматизации будет предоставлен в Приложении 1 согласно установленным приоритетам. Шаблоны отчетов и описание формирования отчетности будут предоставлены сотрудниками Банка.

**Дополнительные требования при построении КХД**

*Параметры производительности:*

При стандартной эксплуатации система должна быть в рабочем состоянии в режиме «24/7» 365 дней в году.

С точки зрения производительности система должна отвечать следующим требованиям:

• определенные администратором пользователи системы должны быть уведомлены, когда возможно ухудшение производительности в результате каких-либо действий по сравнению с обычной производительностью в графическом интерфейсе или путем отправки уведомлений (e-mail, SMS, мессенджеры, Push);

• должна поддерживаться обработка не менее 200 одновременных пользовательских сессий в любое заданное время без ухудшения производительности;

• должна быть обеспечена возможность хранения детальных данных согласно срокам, установленным регулятором, но не менее 5 лет;

• хранение обработанных и рассчитанных данных (витрин) должно быть не менее 5 лет.

• должна быть обеспечена возможность передачи как обработанных, так и «сырых» данных в Озеро данных, и хранение не менее 5 лет;

*Параметры администрирования:*

• должна быть возможность выполнять задачи сопровождения системы (обновление ПО, АО) с минимальным простоем решения. Должен быть сформулирован подход к снижению времени простоя;

• должны быть обеспечены функции оповещения по системе и приложению (потеря данных, сбой процессов).

• должна быть произведена оценка, сколько сотрудников, какие роли и навыки необходимы для администрирования и эксплуатации КХД в продуктивной среде при стандартных условиях;

• должно поддерживаться оповещение администраторов об отказах по электронной почте и SMS;

• должен поддерживаться мониторинг по порогу хранения данных (например, по нагрузке, на свободное место) и нотификации по достижении данных порогов.

**Тестирование, прием-передача результатов**

Методика тестирования (интеграционного, функционального) должна быть разработана в соответствии с Техническими заданиями и предоставляется на согласование Заказчику. Тестирование будет проводиться совместно Исполнителем (внешней компанией по аутсорсингу) и заказчиками от Банка путем проверки результатов реализации модели данных и на корректность финансовых показателей в отчетности.

Документация проекта:

1. Концепция КХД (описание поставленных задач, описание бизнес процессов создания, развития и поддержки КХД, принципы разграничения прав доступа, дорожная карта развития КХД);
2. Архитектура КХД (описание по аппаратной и программной части, описание слоев хранения данных, логика формирования слоев данных, описание витрин данных и отчетов, описание объемов хранимой информации с учетом принципа масштабируемости, архитектура Data Lake);
3. Глоссарий показателей и сущностей;
4. «Дерево» показателей (графическое описание зависимости сущностей и показателей);
5. База справочников/классификаторов;
6. Описание модели данных, ER-диаграммы;
7. Описание ETL-процедур, логики преобразования данных, верификации данных (процедуры контроля качества данных);
8. Описание требований к программной и аппаратной части;