**ВВЕДЕНИЕ**

**Актуальность темы.** Главный посыл: разлить воды, что ЭДО это важно, дистанционные технология, упомнить по корону. Актуальность растет, нагрузка растет. Ссылка на статистику работы ЭДО. Более остро стоит вопрос хранения данный. На текущем этапе развития информационных технологий, аппаратных средств, программного обеспечения – стало возможным создание систем электронного документооборота. Под электронным документом принято считать – набор информации (текст, изображение, звукозапись, видеозапись и т.п.), который сохранен компьютере или сервере. Электронный документооборот (ЭДО) – это способ организации работы с электронным документами, хранящимся централизовано. Система электронного документооборота (СЭД) – это программное обеспечение, которое позволяет организовать работу с электронными документами.

На сегодняшний день существует множество разработанных и введенных в эксплуатацию систем электронного документооборота. Некоторые из них решают проблемы отдельно взятых организаций, которые упрощают и удешевляют работу внутри этих организаций. Как пример можно привести отдельно взятое конструкторское бюро, где чертежи изделий создаются с помощью специализированного программного обеспечения и хранятся в электронном виде. После создания чертежа, его можно отправить на утверждение, получить замечания, сделать доработки и все это можно сделать за компьютером, никуда не отходя. Другие, наоборот, решают проблемы обмена документами между организациями и подачи отчетности в государственные ведомства. Организации, которые, обеспечивают работу таких систем – называют операторами электронного документооборота.

Но чем больше электронных документов появляется в системе электронного документооборота, тем больше ресурсов необходимо для работы этой системы. В первую очередь – растет объем занимаемой памяти системы, во вторую очередь – падает эффективность работы системы. В среднем оператор электронного документооборота за год использует от 2 до 6 терабайтов памяти. Увеличить объемы хранилища не составляет труда, вопрос стоит в их стоимости этих работ. Если покупать дешевые накопители – будет существенно замедлен доступ к информации, купить дорогие – получим быстрый доступ к данным, но они стоят на несколько порядков дороже. К тому же многие данные могли стать не актуальными, и запросов на получение их будет намного меньше, так зачем переплачивать? Для решения данного вопроса нужен компромисс.

**Цель и задачи работы.** Целью данной работы является исследование методов оптимизации распределенного хранилища данных для системы электронного документооборота и создание программного модуля реализующий хранение, обработку и организацию доступа к данным. Исследования ли? Или реализация?

Для достижения поставленной цели необходимо решить следующие задачи:

* исследовать существующие методы и подходы к организации распределенного хранилища данных для системы электронного документооборота;
* разработать оптимизированный метод организации распределенного хранилища данных для решения задач электронного документооборота;
* разработать программный модуль хранения данных к системе ЭДО.

**Предмет и объект исследования.** Объектом исследования настоящей работы является система электронного документооборота. Предметом исследования является методы организации хранения данных.

**Научная новизна.** Разработка новых методов решений – организации хранилища данных для системы электронного документооборота. Ближе к оптимизации хранения данных по ряду характеристик

**Практическое значение работы.** Результаты данной работы могут представлять интерес для инженеров программного обеспечения, разрабатывающих системы электронного документооборота, а также организациям использующие системы электронного документооборота.

ПОМЕТКИ:

Когда идет оценки разных технологий надо делать упор на экономические факторы, импорт замещение.

Учесть вариант миграции данных с одного хранилища на другой!

Ссылка на законодательство персональных данных. Может быть требования по хранению на территории РФ.

**Исследование методов оптимизации распределенного хранилища данных для системы ЭДО**

**1 АНАЛИЗ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ В ЗАДАЧЕ ОПТИМИЗАЦИИ ХРАНЕНИЯ ДАННЫХ ДЛЯ СИСТЕМЫ ЭДО**

* 1. **Обзор существующих подходов к решению задачи**

В первую очередь, перед обзором существующих подходов к решению задачи хранения данных, необходимо дать определение систем хранению данных (СХД), их разновидности, как они устроены.

**1.1.1 Понятие систем хранения данных**

Система хранения данных – комплекс аппаратных и программных средств, который предназначен для хранения и оперативной обработки информации, как правило, большого объема. Информация – это файлы, в том числе медиа, структурированные (СУБД) и неструктурированные данные (big data), резервные копии, архивы. В качестве носителей информации используются жесткие диски [1]. СХД различаются по уровням хранения:

* Блочное хранилище, СХД используется как обычный диск, который можно форматировать, устанавливать на него ОС, создавать логические диски. Данные хранятся не файлами, а блоками, что ускоряет операции ввода-вывода. Подходит для высокопроизводительных вычислений, СУБД, хранения больших объемов данных;
* Файловое хранилище, данные хранятся в виде файлов, которые размещаются в каталогах. Такая СХД используется для хранения «холодной» информации, которая не требуется для операционных вычислений;
* Объектное хранилище, ориентировано на работу с большими неструктурированными данными объемом до петабайтов. Информация хранится не в виде файлов, а в виде «объектов» с уникальными идентификатором и метаданными. Используется в аналитике, big data, машинном обучении, для хранения «тяжелых» медиа-файлов и резервных копий, разработки и эксплуатации приложений в облаке, хостинга веб-сайтов.

Также системы хранения данных различаются по частоте использования хранимых данных:

* Системе краткосрочного хранения (online storage). Такого рода системы обязаны быть иметь высокие показатели скорости доступа к данным, содержат небольшой объем информации и как правило данные «живут» в них от двух недель до одного месяца;
* Системы средней продолжительности (near-line storage). Имеют средний показатели скорости доступа, средний объем информации, данные могут храниться на протяжении года;
* Системы долговременного хранения (offline storage). У таких систем низкой уровень скорости доступа, большие объемы информации, данные хранятся от года и больше.

Помимо этого, их также различают по типам подключения:

* DAS (Direct-attached storage). система хранения данных с прямым непосредственным подключением к серверу или рабочей станции, без помощи сети;
* NAS (Network-attached storage). Файловый сервер, который включен в локальную сеть;
* SAN (Storage Area Network). Сеть, которая объединяет разнотипные хранилища (диски, оптические приводы, ленточные массивы), но которые воспринимаются операционной системой как единое логическое хранилище данных, или как сетевой логический диск.

Далее хотелось бы остановиться на каждом типе системы хранения более подробно.

DAS – подразумевает прямое (непосредственное) подключение носителей информации к серверу либо рабочей станции. При этом накопители могут быть внутренними (установлены непосредственно в корпусе сервера) либо внешними. Самый простой пример DAS-системы – это жесткий диск, который расположен, внутри сервера или рабочей станции. Архитектура DAS системы представлена на рисунке 1.1.



Рисунок 1.1 – Архитектура DAS системы

Данная система имеет относительно низкую стоимость оборудования и очень хорошо подходит для хранения мультимедия данных большого объема благодаря высокой скорости обмена данными. Выделим преимущества данной системы:

* Низкая стоимость;
* Простота эксплуатации;
* Может быть быстро развернута;
* Высокая скорость обмена данными.

К большому сожалению DAS-системы плохо масштабируются. Устройства хранения имеют относительно небольшое количество портов, что ограничивает количество хостов, которые могут непосредственно подключаться к хранилищу. Ресурсы, которые не используются, не могут быть перераспределены. Именно по этим причинам были разработаны системы NAS и SAN, которые являются сетевыми.

NAS-системы – это сетевые системы хранения данных, непосредственно подключаемые к сети точно так же, как и сетевой принт-сервер или маршрутизатор. Данная система предоставляет доступ к файлам через IP сети.

Устройство NAS использует собственную операционную систему, которая оптимизирована для файловых операций ввода/вывода, а также освобождена от всех функций операционной системы, не связанных с обслуживанием файловой системы.

NAS используется для работы с данными файлового типа, к которым нужен коллективный одновременный доступ. NAS работает «поверх» существующей локальной сети, через общие коммутаторы/маршрутизаторы. Архитектура системы изображена на рисунке 1.2.



Рисунок 1.2 – Архитектура NAS системы

Преимущества данной системы являются:

* Эффективное использования всех ресурсов памяти;
* Высокая производительность серверов;
* Уменьшение нагрузки на сервер приложений;
* Прост в эксплуатации.

Из недостатков можно выделить сложность масштабируемости серверов, а также повышенную нагрузку на сеть.

SAN-системы – представляет собой специализированную сетевую инфраструктуру для хранения данных, которая связывает один или несколько разнотипных хранилищ в единую сеть. Данная сеть позволяет большому числу пользователей хранить свои данные в одном месте и совместно использовать их. В основном используется блочный тип хранения данных. Преимущества данной системы следующие:

* Централизованное управление данными;
* Высокий уровень быстродействия;
* Высокий уровень отказоустойчивости;
* Высокий уровень масштабируемости.

Но самым главным недостатком SAN является его цена, далеко не каждая фирма может позволить себе данную систему.

На рисунке 1.3 изображена архитектура SAN-системы.



Рисунок 1.3 – Архитектура SAN-системы

Мы дали определение понятию систем хранению данных, а также рассмотрели их возможные разновидности. При описании СХД, несколько раз упоминались термины: жесткий диск, носитель информации.

**1.1.2 Виды и характеристики жестких дисков**

Носитель информации – это любой материальный объект, используемый человеком, способный достаточно длительное время сохранять в своей структуре занесённую на него информацию, без использования источников энергии [2].

Жесткий диск – это носитель информации с произвольным доступом, основанное на принципе магнитной записи, другими словами всеми знакомый HDD-диск (Hard Disk Drive). Интересный факт, что жесткими дисками также принято называть твердотельный накопитель SSD (Solid State Drive), хотя они не основаны на принципе магнитной записи, а используют технологию флеш-памяти.

Жесткие диски имеют следующие характеристики:

* Емкость;
* Среднее время доступа. Высчитывается по простой формуле, среднее время поиска (mean seek time) суммируется с временем ожидания (mean wait time), т.е. временем извлечения информации с диска;
* Скорость передачи данных;
* Число операций ввода-вывода в секунду (Input/Output Per Second)
* Пропускная способность передачи данных (data throughput). Показывает, какой объем данных можно передать в единицу времени.

HDD-диски имеют следующие преимущества [3]:

* Низкая стоимость в перерасчете на объем хранимой информации;
* Неограниченное количество циклов записи;
* Возможность восстановления информации. Имеется в виду то, что при выходе из строя диска, данные находящиеся на нём можно восстановить.

Из недостатков выделим низкое среднее время доступа к информации, не большое количество операций ввода-вывода в секунду.

У SSD-дисков, по сравнению с жесткими дисками, имеется одно очень важное преимущество – это высокое среднее время доступа к данным, на порядок выше IOPS [3]. Для сравнения у HDD-дисков IOPS равняется 80-100, в то время как у SSD-дисков он более 8000 [3]. Но за такими важными преимуществами скрываются свои недостатки:

* Высокая стоимость в перерасчете на объем хранимой информации;
* Ограниченное количество циклов записи. По сути, после каждой записи информации на диск, физическая ячейка памяти «сжигает» ее.

Таким образом HDD-диски – это медленные и большие носители информации, которые могут успешно решать задачи хранения большого объема данных, не требующих быстрой скорости доступа к данным. В свою очередь SSD-диски – это быстрые хранилища малого объема, которые могут решать задачи быстрого доступа к данным. Также стоит отметить, что данные диски имеют ограниченный ресурс циклов перезаписи, что в свою очередь не позволяет использовать их в системах с частой перезаписью данных.

В данной научной работе – намерено отсутствует описание аппаратной реализации данных носителей информации, а также описание некоторых характеристик, потому что, это не является целью данной работы.

На сегодняшний день, в свободном доступе можно приобрести HDD-диски максимального объема в 16 терабайт (ТБ), в свою очередь SSD-диски максимального объема в 8 ТБ. Очевидно, что данных объемов будет недостаточно, для средних и больших систем электронного документооборота. Также HDD и SSD диски не обладают требуемыми уровнями надежности, прогноз выхода из строя диска является сложной. Согласно статистике компании «Backblaze», из 13 тысяч жестких дисков за 3 года отказывают примерно от 3.1% до 26.5% [4]. С целью увеличения общего объема памяти и для повышения уровня надежности систем хранения данных – была разработана технология RAID.

**1.1.3 Технология RAID**

Технология RAID (Redundant Array of Independent Disks) – массив независимых дисков с избыточностью хранения данных [4]. Избыточность означает то, что все байты данных при записи на один диск дублируются на другом диске, и могут быть использованы в том случае, если первый диск откажет. Кроме того, эта технология помогает увеличить IOPS. Существует множество различных уровней (конфигураций) RAID-массивов, которые каждый по-своему решает поставленные задачи надежности и производительности [4, 5].

RAID 0. Предполагает одновременное использование нескольких жестких дисков с целью существенного увеличения производительности рабочей станции. Информация разбивается на блоки данных фиксированной длины и записывается на несколько дисков поочередно. При отказе одного из дисков неработоспособной оказывается вся система, так как данные до этого были равномерно записаны по всем хранилищам из массива. Cхема данного уровня представлена на рисунке 1.4.



Рисунок 1.4 – RAID 0

RAID 1. Полностью решает проблему надежности, массив состоит из двух (или более) дисков, являющихся полными копиями друг друга. Только половина ёмкости массива отводиться под данные. Скорость считывания данных выше чем скорость записи данных. Схема уровня изображена на рисунке 1.5.



Рисунок 1.5 – RAID 1

RAID 2. Представляет собой улучшенную версию RAID 0. Данные все также записываются на несколько дисков, логически представляющих единое целое дисковое пространство, но было введено использования кода Хэмминга при записи данных. Данный код способен исправлять возникающие ошибки (в случае выхода из строя одного из дисков в массиве), используя проверочные последовательности. Коды коррекции ошибок требуют достаточно много дискового пространства, что повышает избыточность (для этих целей выделяются отдельные диски из массива). Схема представлена на рисунке 1.6.



Рисунок 1.6 – RAID 2

RAID 3. Схожа с RAID 2, с одной разницей, что вместо кода Хэмминга используется обычная для побитовых проверок контрольная сумма, построенная по принципу «исключающего ИЛИ». Для хранения контрольных сумм выделяется диск из массива, который отвечает высоким требования отказоустойчивости, так как обращения к этому диску происходят каждый раз, когда необходимо записать данные. За счет этого уменьшается избыточность данных, но одновременная обработка нескольких обращений к данным невозможна. Схема конфигурации представлена на рисунке 1.7.



Рисунок 1.7 – RAID 3

RAID 4. Практическая полная копия RAID 3, отличие заключается лишь в увеличенном блоке записываемых данных. Появляется возможность параллельного доступа к данным для считывания, для записи по-прежнему данная возможность отсутствует. Схема изображена на рисунке 1.8.



Рисунок 1.8 – RAID 4

RAID 5. Похож на RAID 3 и RAID 4, но контрольная сумма записывается не на отдельный диск, а по всему массиву, занимает примерно четвертую часть дискового пространства. Данное решение ускоряет операции считывания (так как не требуется обращаться к одному диску) и добавляет возможность параллельной записи данных. Схема данной конфигурации представлена на рисунке 1.9.



Рисунок 1.9 – RAID 5

Мы рассмотрели основные базовые уровни RAID-массивов, также у этой технологии есть возможность комбинирования уровней. Рассмотрим один из популярных, который представляет интерес для данной работы.

RAID 10 – зеркалированный массив, данные в котором записывают последовательно на несколько дисков, также как в RAID 0. Но у этих дисков существует полная копия как в RAID 1. Таким образом, данный массив объединяет в себе высокую отказоустойчивость и производительность. На рисунке 1.10, представлена схема уровня. Сравнительные характеристики уровней RAID-массивов представлена в таблице 1.1 [5].



Рисунок 1.5 – RAID 10

Таблица 1.1 – Сравнение характеристик наиболее популярных уровней RAID.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип | Использование емкости дисков, % | Производительность | | Надежность | Минимальное количество дисков | Максимальное количество дисков |
| чтения | записи |
| RAID 0 | 100 | Высокая | Высокая | Низкая | 2 | 16 |
| RAID 1 | 50 | Высокая | Средняя | Высокая | 2 | 2 |
| RAID 10 | 50 | Высокая | Высокая | Высокая | 4 | 16 |
| RAID 5 | 67-94 | Высокая | Средняя | Средняя | 3 | 16 |

Описанной теоретической части устройства систем хранения данных достаточно для того, чтобы перейти к рассмотрению существующих хранилищ данных. В последнее время, очень большую популярность набирают решения облачного хранения данных.

Облачное хранилище данных – это хранилище, в котором данные хранятся на многочисленных распределённых в сети [серверах](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D1%80%D0%B2%D0%B5%D1%80_(%D0%B0%D0%BF%D0%BF%D0%B0%D1%80%D0%B0%D1%82%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D0%B5%D1%81%D0%BF%D0%B5%D1%87%D0%B5%D0%BD%D0%B8%D0%B5)), предоставляемых в пользование клиентам, в основном, третьей стороной. Внутренняя структура серверов клиенту, в общем случае, не видна. Данные хранятся и обрабатываются в так называемом «[облаке](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D0%BB%D0%B0%D1%87%D0%BD%D1%8B%D0%B5_%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F)», которое представляет собой, с точки зрения клиента, один большой виртуальный сервер. Физически же такие серверы могут располагаться удалённо друг от друга географически [6].

Преимуществами данного типа хранилища, следующие:

* Доступ к данным возможен с любого сервера, рабочей станции, которая подключена к интернету;
* Возможен совместный доступ к данным;
* Оплачивается только тот объем хранилища, который фактически используется;
* Отсутствует необходимость заниматься приобретением, поддержкой и обслуживанием собственной инфраструктуры хранилища;
* Обеспечение целостности данных предоставляются провайдером «облачного» центра. Все возможные аппаратные сбои под ответственностью центра.

Преимущества очень хорошие, а какие есть недостатки у данных СХД? В самую первую очередь это цена использования данных хранилищ, они очень высоки (конкретные цифры будут приведены при рассмотрении конкретных облачных хранилищ). Второй момент – это обязательное подключение к интернету, если из-за сбоев он недоступен, недоступно и облачное хранилище. И наконец самый неприятный минус – это возможная утечка данных. Известно множество случаев, к примеру, из-за некорректных прав доступа к облачным хранилищам Amazon S3, «утекло» 6 миллионов данных о клиентах сотовой компании связи «Verizon» [6].

Отметим, что облачное хранилище может быть частным (корпоративным) [7] – это хранилище, развернутое на собственной инфраструктуре. При таком подходе мы теряем большую часть преимуществ, касающихся снижению расходов на оборудование, поддержку инфраструктуры, но тем самым получаем высокий уровень безопасности данных.

Обзор Amazon S3, Google File System, Microsoft Azure

Постановка задачи – из темы, что надо сделать «новое» «отличное» хранилище, которое соответствует требования оператора ЭДО по таким-то критериям.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Понятие СХД. https://itglobal.com/ru-ru/company/glossary/shd-sistema-hraneniya-dannyh/;
2. <https://www.elibrary.ru/item.asp?id=44151107>;
3. Богановский А.В. Анализ методов построения систем хранения данных // Перспективы развития информационных технологий. Сборник материалов XXXV Международной научно-практической конференции. – Новосибирск, 2017. – С. 13-17.;
4. Исалёв А.С. Анализ надежности технологии хранения данных RAID. <https://www.elibrary.ru/item.asp?id=27032132&>;
5. Шувалов Н.В. Технология RAID. <https://elibrary.ru/item.asp?id=29769452>;
6. <https://ru.wikipedia.org/wiki/Облачное_хранилище_данных>;
7. Архипова А.А., Высоцкая А.В. Современные облачные технологии. <https://elibrary.ru/item.asp?id=42330839>