**СОДЕРЖАНИЕ**

[**ВВЕДЕНИЕ** 2](#_Toc57421258)

[**1 АНАЛИЗ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ В ЗАДАЧЕ ОПТИМИЗАЦИИ ХРАНЕНИЯ ДАННЫХ ДЛЯ СИСТЕМЫ ЭДО** 5](#_Toc57421259)

[1.1 Технологии хранения данных 5](#_Toc57421260)

[1.1.1 Понятие систем хранения данных 5](#_Toc57421261)

[1.1.2 Виды и характеристики запоминающих устройств 10](#_Toc57421262)

[1.1.3 Технология RAID 13](#_Toc57421263)

[1.1.4 Распределенные хранилища данных 18](#_Toc57421264)

[1.2 Обзор существующих хранилищ данных 21](#_Toc57421265)

[1.2.1 Amazon S3 21](#_Toc57421266)

[1.2.2 Mail Cloud Storage 26](#_Toc57421267)

[1.2.3 Yandex Object Storage 28](#_Toc57421268)

[1.3 Информационные процессы хранения данных в системах электронного документооборота 29](#_Toc57421269)

[1.4 Выбор методов оптимизации хранения данных для системы электронного документооборота 30](#_Toc57421270)

[1.5 Постановка задачи создания оптимизированного хранилища данных для системы электронного документооборота 30](#_Toc57421271)

[**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ** 32](#_Toc57421272)

# **ВВЕДЕНИЕ**

**Актуальность темы.** Первыесистемы электронного документооборота появились еще в 80-х годах прошлого века, но широкого распространения они не получили, из-за существующих в то время технических возможностей. Они были дороги в разработке, обслуживании и решали узкие задачи отдельно взятых компаний.

Сегодня же можно сказать, что данные системы захватывают все отрасли нашей с вами жизни. С их помощью повышают эффективность деятельности коммерческих компаний и промышленных предприятий, а в государственных учреждениях на базе технологий электронного документооборота решаются задачи внутреннего управления, межведомственного взаимодействия и взаимодействия с населением.

Система электронного документооборота (СЭД) – организационно-техническая система, обеспечивающая процесс создания, управления доступом и распространения электронных документов в компьютерных сетях, а также обеспечивающая контроль над потоками документов в организации.

Электронный документ – документ, созданный с помощью средств компьютерной обработки информации.

Неотъемлемой частью любой системы электронного документооборота является система хранения данных – это очевидно, так как электронные документы должны где-то храниться.

Система хранения данных (СХД) – комплекс аппаратных и программных средств, который предназначен для хранения и оперативной обработки информации.

С каждым годом, объем хранимой информации системами электронного документооборота неуклонно растет. Также в нынешнее время из-за пандемии COVID-19, системы ЭДО стали жизненно-необходимы для дистанционной работы предприятий. Отметим, что компании не торопятся увеличивать бюджет на увеличение объема хранилищ и их поддержку, разрыв между ростом объема данных и необходимыми расходами на их сопровождение продолжает увеличиваться. В следствии чего вопросы оптимизации хранилищ данных становиться все более и более остро.

**Цель и задачи работы.** Целью данной работы является программная реализация оптимизированного хранилища данных для системы электронного документооборота, которое реализует хранение, обработку и организацию доступа к данным.

Для достижения поставленной цели необходимо решить следующие задачи:

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

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

**Научная новизна.** Оптимизация хранения данных предназначенного для систем электронного документооборота. Наверно

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

**Структура работы.** Данная работа состоит из пояснительной записки, включающей в себя введение, три раздела, заключение, списка использованных источников и приложений.

В первом разделе осуществляется анализ методов оптимизации хранилищ данных, рассматриваются готовые решения распределенных хранилищ данных.

In progress…

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. **Технологии хранения данных**

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

### **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 запоминающие устройства [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 [2, 3]:

* 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.

Остановимся 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

### **1.1.2 Виды и характеристики запоминающих устройств**

Запоминающие устройство (ЗУ) – устройство, предназначенное для записи и хранения данных. В основе работы ЗУ может лежать любой физический процесс, обеспечивающий приведение системы к двум или более устойчивым состояниям [4].

По типу доступа к данным ЗУ делятся на:

* Устройства с последовательным доступом. Обращения к данным, происходят в заранее заданном порядке;
* Устройства с произвольным доступом. Единовременно можно получить доступ к любым данным по адресу.

По зависимости от электропитания:

* Энергозависимые. Данные после потери электропитания исчезают;
* Энергонезависимые. Данные после потери электропитания никуда не исчезают.

По возможности записи:

* Устройства с однократной записью, без возможности перезаписи;
* Перепрограммируемые устройства. ЗУ с возможностью многократной перезаписи, затрудненной долгим временем записи или ограниченным числом циклов записи;
* Устройства со свободной многократной записью.

По назначению:

* Внутренние устройства. Предназначены для хранения данных непосредственно необходимых во время выполнения программы. Обладают быстрым уровнем доступа к данным (так как подключены на прямую к процессору);
* Внешние устройства. Предназначены для хранения больших объёмов данных на сменных или фиксированных носителях. Обладают медленным уровнем доступа к данным.

Для решения задач электронного документооборота не подходят устройства с однократной записью и перепрограммируемые устройства, так как данные постоянно поступают новые данные, которые необходимо хранить. Также не подходят энергозависимые запоминающие устройства.

Рассмотрим самые распространенные запоминающие устройства на сегодня:

* Оперативные запоминающие устройства (ОЗУ). Является энергозависимым ЗУ, с произвольным доступом к данным, возможностью многократной записи данных и относиться к внутренним устройствам;
* HDD-диски (Hard Disk Drive). Энергонезависимое ЗУ, с произвольным доступом к данным, возможностью многократной записи данных и относиться ко внешним устройствам. Основан на принципе магнитной записи;
* SSD-диски (Solid State Drive). Энергонезависимое запоминающие устройство, с произвольным уровнем доступа, возможностью многократной записи данных и относиться ко внешним устройствам. Использует технологию флеш-памяти;
* Ленточные накопители (Стриммер). Энергонезависимое ЗУ, с последовательным уровнем доступа, возможностью многократной записи и относиться ко внешним устройствам. Основан на принципе магнитной записи.

Запоминающие устройства имеют следующие характеристики:

* Емкость. Максимально-возможный объем хранимых данных;
* 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.

ОЗУ обладает самой маленькой емкостью среди описанных ЗУ, но данное устройство обладает самой высоким средним временем доступа, скорости передачи данных, числом операций ввода-вывода, пропускной способностью. Все этим преимущества – благодаря прямому подключению к процессору. Из минусов отметим, что данные ЗУ, помимо малой емкости, еще и самые дорогие в соотношении цены и объема хранимой информации. Также ОЗУ является энергозависимой, и мы не можем гарантировать надежность данных.

HDD-диски – можно сказать, что они самые распространенные на сегодняшний день. Имеют следующие преимущества:

* 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.

SSD-диски по сравнению с HDD-дисками, имеют одно очень важное преимущество – это высокое среднее время доступа к данным, на порядок выше IOPS [5]. Для сравнения у HDD-дисков IOPS равняется 80-100, в то время как у SSD-дисков он более 8000 [5]. 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.

Ленточные накопители являются самыми дешевыми (в соотношении цены и объема хранимой информации), надежными и самыми медленными.

Таким образом ОЗУ хоть и самые быстрые, но не подходят для надежного хранения данных. 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 меньшего объема (по сравнению с HDD), 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. Ленточные носители – большие хранилища данных, но очень медленные, такого типа ЗУ подойдут для архивного хранения 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% [5]. 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 [6]. 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 [6, 7].

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 [7].



Рисунок 1.10 – RAID 10

Таблица 1.1 – Сравнение характеристик наиболее популярных уровней RAID.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип | Использование емкости, % | Производительность | | Надежность | Мин. количество дисков | Макс. 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 |

### **1.1.4 Распределенные хранилища данных**

Распределенное хранилище данных (Distributed Data Storage) – это инфраструктура, которая может быть разделена между множественными физическими серверами (узлами хранения данных), которые чаще всего, распределены между разными центрами обработки данных (ЦОД) [8].

Обычно представлены в виде программно-определяемой сети хранения (SDS) данных, с механизмами синхронизации и координирования данных между узлами, находящимся в кластере.

Software Defined Storage (SDS) – это архитектура объединения систем хранения данных, которое отделяет программного обеспечение от оборудования, на котором храниться информация. В отличии от NAS и SAN систем, организация сети данных происходит на программном уровне, для функционирования нет необходимости использования одного и того-же аппаратного оборудования. В одной сети могут находиться сервера с различной операционной системой, характеристиками [9].

Распределенные хранилища данных имеют следующие преимущества [8, 10]:

* Простота масштабируемости. Осуществляется очень просто, для этого необходимо добавить новый узел в сеть;
* Высокая надежность. Распределенные хранилища как правило хранят более одной копии данных на разных узлах кластера;
* Низкая стоимость. На сегодня представлено большое количество бесплатного программного обеспечения для организации распределенных хранилищ, можно использовать практические любые сервера;
* Высокая производительность. В виду того, что данные размещены на разных узлах одновременно, возможно получение данных с «ближайшего» узла к пользователю, параллельный доступ к данным.

Из минусов можно выделить сложность ввода в эксплантацию распределенного хранилища, также из-за возможности использования различных серверов, возможны трудности в обслуживании кластера.

Перечислим функциональные возможности распределенных систем хранения данных [11]:

* Функция архивного хранения. Так называемое «холодное хранение», однократная запись данных с целью долгосрочного хранения. Обычно к таким данным обращаются в исключительных случаях;
* Функция сжатия данных. Алгоритмическое преобразования данных, с целью уменьшения объема данных;
* Функция дедупликации данных. Технология обнаружения и исключения избыточных данных. В свою очередь бывает двух типов: на уровне файлов, на уровне блоков. В случае уровня файлов – дублирующийся файлы исключается из хранилища. В случае уровня блоков – исключаются повторяющиеся блоки данных;
* Функция шардинга (Sharding) данных. Технология распределения данных между различными узлами кластера;
* Функция репликации (Replication) данных. Способ синхронизации копий данных между узлами кластера, с целью увеличения надежности хранилища;
* Механизм георепликации. Практически тоже самое, что и репликация, в отличии от того, что целью является распространение копий данных между географически удаленных узлов;
* Механизм управления доступа к данным. Применяется для ограничения доступа данных посредством списков доступа;
* Функция шифрования. Используется для повышения уровня конфиденциальности хранимых данных;
* Функция самовосстановления. В случае если происходит исключительная ситуация, такая как отказ одного из узлов кластера, распределенное хранилище не выходит из строя, а перестраивается и продолжает свою работу;
* Функция кэширования;
* Функция балансировки нагрузки. Учет статистики работы сети и принятие решение о распределении передачи данных на другие узлы кластера.

Распределенные системы хранения данных как правило реализуют не все функции, которые перечислены выше. Для конкретного хранилища выбираются и реализуются только необходимые для решения поставленных задач, именно поэтому существует большое количество разнообразных хранилищ. В последнее большое развитие получили облачные хранилища данных.

Облачное хранилище данных (Cloud Storage) – это хранилище, данные, в котором, хранятся на многочисленных распределенных 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)) (узлах данных), 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 [12].

Преимущества пользования облачными хранилищами следующие [13]:

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

Далее рассмотрим недостатки облачных СХД. В самую первую очередь это цена использования данных хранилищ, они очень высоки. Второй момент – это обязательное подключение к интернету, если из-за сбоев он недоступен, недоступно и облачное хранилище. Самый большой минус – это возможная утечка данных. Известно множество случаев, к примеру, из-за некорректных прав доступа к облачным хранилищам Amazon S3, «утекло» 6 миллионов записей о клиентах сотовой компании связи «Verizon» [12].

Существует различные типы облачных хранилищ [13]: публичное, приватное, гибридное.

Публичное облачное хранилище – это инфраструктура, предоставляемая в пользование большому количеству пользователей. Данная инфраструктура располагается и находиться в собственности компании предоставляющей услуги.

Приватное (корпоративное) облачное хранилище – это хранилище, развернутое на собственной закрытой инфраструктуре. Значительно повышается уровень безопасности, но сильно увеличиваются затраты на поддержание инфраструктуры.

Гибридное облачное хранилище – это объединение преимуществ публичного и приватного облачного хранилища. Также как в случае приватного облачного хранилища существует собственная закрытая инфраструктура, но значительно меньшего размера. В тоже время существует доступ к общедоступному хранилищу. При таком подходе можно гарантировать высокий уровень безопасности для «важных» данных и хранить большие объемы данных в публичном облаке.

## **1.2 Обзор существующих хранилищ данных**

### **1.2.1 Amazon S3**

Amazon Simple Storage Service (S3) – это облачное хранилище данных, разработанная компанией «Amazon», впервые появился в марте 2006 года. Является системой хранения данных объектного типа.

Общий объем хранимых данных и количество объектов не ограничены. Размер отдельных объектов Amazon S3 может составлять от 0 байт до 5 ТБ. Можно хранить практически любые типы данных в любом формате. Amazon S3 предоставляет простой интерфейс веб-сервиса, который можно использовать для хранения и извлечения любых объемов данных в любое время из любого места в Интернете. Amazon S3 – это простое хранилище объектов на основе ключа. При хранении данных объектам назначается уникальный ключ, который может использоваться впоследствии для доступа к данным. Ключи могут иметь любые строковые значения; их можно создавать так, чтобы имитировать иерархические атрибуты.

Перечислим основные концепции для общего понимания работы [14]:

* Корзина (Bucket). Именованный контейнер для хранения объектов. Каждый объект обязательно должен находиться в корзине. Такое решение позволяет: логически разделять файлы по назначению, защищать файлы по уровням доступа, агрегировать информацию для отчетов;
* Объект (Object). Неделимая единица хранения, состоит из данных и метаданных (данных о данных). Метаданные представлены в виде пар ключ-значения, которые описывают объект, к примеру – дата последнего обращению к объекту. У каждого объекта обязательно должен быть определен ключ;
* Ключ (Key). Уникальный идентификатор объекта в корзине. Сочетание наименования корзины, ключа объекта и его версии – однозначно идентифицирует объект;
* Регион (Region). Центр обработки данных (ЦОД), в котором буду храниться объекты. При создании корзины, можно выбрать конкретный регион. Объекты никогда не покидают своего региона, только если это не было сделано намерено.

Версионность объектов – механизм, который позволяет сохранять состояние объектов до внесения изменений. В качестве примера рассмотрим обычный случай редактирования файла, после обновления файла, старое содержимое никуда не пропадает, доступ к нему доступен, обновления же получат просто новый идентификатор версии.

По всему миру создано большое количество центров обработки данных компании Amazon. На рисунке 1.11 изображена географическая карта нашего мира с отмеченными ЦОД компании Amazon. Как видно из рисунка 1.11 – на территории Российской Федерации нет ни одного центра обработки данных. Ближайшие доступные, это европейские центры, расположенные во Франкфурте, Ирландии, Лондоне, Париже, Стокгольме и Милане.



Рисунок 1.11 – Центры обработки данных компании Amazon

Amazon S3 предоставляет множество классов хранилищ, которые решают разные поставленные задачи и конечно же, различную стоимостью обслуживания. Существуют следующие классы хранилищ [14]:

* Amazon S3 Standard (S3 Standard). Предлагает высокую надежность, доступность и производительность объектного хранилища для хранения часто используемых данных. Обеспечивая низкую задержку и высокую пропускную способность;
* Amazon S3 Intelligent-Tiering (S3 Intelligent-Tiering). Создан для оптимизации расходов путем автоматического перемещения данных на наиболее экономичный уровень доступа без ущерба для производительности и роста операционных издержек;
* Amazon S3 Standard-Infrequent Access (S3 Standard-IA). Является идеальным выбором для хранения данных, доступ к которым осуществляется относительно редко, но при этом должен обеспечиваться быстро;
* Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA). Подходит для хранения данных, с редким доступом. В отличии от других классов, которые хранят данные как минимум в трех регионах, он хранит только в одной зоне доступности.
* Amazon S3 Glacier (S3 Glacier). Это безопасный, надежный и экономичный класс для архивации данных. Время извлечения данных может составлять от нескольких минут до нескольких часов;
* Amazon S3 Glacier Deep Archive (S3 Glacier Deep Archive). Это самый экономичный класс хранилища, с поддержкой долгосрочного хранения и цифровой архивации данных, доступ к которым запрашивается один‑два раза в год. Он создан для клиентов, которые хранят наборы данных 7–10 лет или дольше для выполнения нормативных требований.

Каждый из описанных классов гарантирует 99.99% надежность хранения данных, а также 99.99% доступности в течении года [14].

Особое внимание стоит уделить хранилищу Amazon S3 Intelligent-Tiering, который имеет возможность адаптивного управлению временем доступа к объектам. На рисунке 1.12 изображена схема принципа работы.



Рисунок 1.12 – Принцип работы хранилища Amazon S3 Intellegent-Tiering

Объекты, загруженные в данное хранилище, автоматически сохраняются на уровне для частого доступа. S3 Intelligent-Tiering работает путем отслеживания сценариев доступа и последующего перемещения объектов, доступ к которым не осуществлялся в течение 30 дней подряд, на уровень нечастого доступа. После активации одного или обоих уровней доступа к архивным данным S3 Intelligent-Tiering перемещает объекты, доступ к которым не осуществлялся в течение 90 последовательных дней, на уровень Archive Access, а затем по истечении 180 последовательных дней – на уровень Deep Archive Access. Если после этого запрашивается доступ к этим перемещенным объектам, S3 Intelligent-Tiering перемещает их обратно на уровень для частого доступа.

Поддерживает такие функции как [12]: архивное хранение, дедупликация данных на уровне файлов, георепликация данных, управления доступом к данным, шифрования данных, кэширования, самовосстановления, балансировки нагрузки.

Для расчета стоимости использования хранилища, компания Amazon создало онлайн-калькулятор, который расположен на их сайте. Для того, чтобы узнать конкретную стоимость, надо заполнить требования к хранилищу. Сперва необходим выбрать регион, для нас подходит Франкфурт (самый близкий к РФ), далее выбираем тип хранилища – S3 Intelligent-Tiering, определяем объем хранилища – 20 ТБ. Amazon требует выбрать планируемое количество запросов на добавления данных и получения, для нашего тестового примера выберем значение в 1 000 000 запросов каждого типа. Результаты расчета изображены на рисунке 1.13. На момент написания данной работы, стоимость составила 497,56 USD (37 894,90 рублей).



Рисунок 1.13 – Расчет стоимости месячного обслуживания хранилища для Amazon S3 Intelligent-Tiering

### **1.2.2 Mail Cloud Storage**

Mail Cloud Storage – является российским аналогом рассмотренного Amazon S3 [15]. Облачное хранилище (или S3) является объектного типа, SAN-архитектуры.

Данные хранятся в виде объектов в бакетах. Объект – это файл и любые дополнительные метаданные, описывающие файл. Чтобы сохранить файл его необходимо загрузить в бакет. Когда загружается файл как объект, можно установить разрешения доступа [16].

Бакеты – это контейнеры для объектов. Может создавать несколько. Доступ к каждому бакету можно контролировать, решая, кто может создавать, удалять и перечислять объекты в нем. Дополнительно можно просматривать журналы доступа для бакета и его объектов [16].

В отличии от Amazon S3, отсутствует понятие регионов, это обусловлено тем, что на сегодняшний день Mail Cloud Storage работает только на территории Российской Федерации. Заявлено о наличии 400 центров обработки данных (ЦОД) по всей территории РФ. Отсутствует версионность объектов.

Возможно хранить файлы размером от 0 байт до 32 ГБ. Общий размер хранилища не ограничен. Возможно хранить файлы любого формата. Для работы с хранилищем предоставлен простой веб-интерфейс, который имеет совместимость с веб-интерфейсом Amazon S3. Надежность хранилища составляет 99,5%.

Mail Cloud Storage отвечает всем требованиям закона «О хранении персональных данных» № 152-ФЗ [17].

Существуют следующие типы хранилищ [16]:

* Hotbox (горячее хранилище). Подходит для хранения данных для хранения часто используемых данных, обеспечивая низкую задержку и высокую пропускную способность;
* Icebox (холодное хранилище). Подходит для хранения файлов нечастого использования, таких как: архивы, резервные копии. Создано для хранения редко используемых данных, к которым при необходимости можно получить быстрый доступ;
* Backup. Используется для хранения файлов резервных копий виртуальных машин и баз данных.

Для расчета стоимости обслуживания, воспользуемся онлайн-калькулятором, расположенного на сайте владельца облачного хранилища. Для расчета воспользуемся тем-же примером, что и в случае Amazon S3. Результаты расчета представлены на рисунке 1.14. Под запросами 1 типа следует понимать запросы, на добавление и обновление данных. Запросы 2 типа – запросы на получение файлов из архива. Также необходимо указать размер исходящего трафика. Исходящий трафик – ограничивает объем данных, который может быть загружен с хранилища за месяц. На момент написания данной работы цена месячного обслуживания составляет 36 153,70 рублей.



Рисунок 1.14 – Расчет стоимости месячного обслуживания хранилища для Mail Cloud Storage

### **1.2.3 Yandex Object Storage**

Сервис Yandex Object Storage – это универсальное масштабируемое решение для хранения данных. Оно подходит как для высоконагруженных сервисов, которым требуется надежный и быстрый доступ к данным, так и для проектов с невысокими требованиями к инфраструктуре хранения [18].

Yandex Object Storage можно отнести также к объектному типу хранилища с SAN-архитектурой. Основные понятия следующие:

* Бакет. Логическая сущность, которая помогает организовать хранение объектов;
* Объект. Данные произвольного формата, загруженные пользователем.

Можно сказать, что данное хранилище также является аналогом Amazon S3. В отличии от Mail Cloud Storage присутствует возможность версионности объектов. Понятие регионов отсутствует, все данные хранятся на территории Российской Федерации. Ко всем объектам и бакетам можно настроить режимы доступа.

Возможно хранить файлы размером от 0 байт до 5 ТБ. Общий размер хранилища не ограничен. Возможно хранить файлы независимо от их формата. Для работы с хранилищем предоставлен простой веб-интерфейс, который имеет совместимость с веб-интерфейсом Amazon S3.

Yandex Object Storage также отвечает всем требованиям закона «О хранении персональных данных» № 152-ФЗ [17].

У Yandex Object Storage представлено два типа хранилищ [18]:

* Стандартное хранилище. Предназначено для активной работы с объектами;
* Холодное хранилище. Предназначено для длительного хранения объектов с редкими запросами на чтение.

Для определения стоимости месячного обслуживания хранилища, воспользуемся калькулятором представленном на сайте Yandex Cloud Storage. Возьмем знакомый пример. Расчет месячного обслуживания представлен на рисунке 1.15.



Рисунок 1.15 – Расчет стоимости месячного обслуживания для Yandex Cloud Storage

Таким образом рассмотренные облачные хранилища обладают схожим функционалом для хранения данных большого объема. Каждое из решений предоставляет различные типы хранилищ. Самым интересным хранилищем –является решение от компании Amazon с «умным» менеджментом расположения данных на разных хранилищах с различной скоростью доступа.

## **1.3 Информационные процессы хранения данных в системах электронного документооборота**

Тут расписываем предметную область для которой будем создавать решение.

* Кто такие операторы ЭДО?
* Какие они решают задачи?
* Какие файлы они должны хранить и сколько по продолжительности, требования со стороны закона? Расписываем про файлы отчетности в ведомства ФНС, ПФР, РОССТАТ, ФСС. Расписываем про файлы обмена между контрагентами. У каждого документа есть история документооборота. Для каждого система должна создавать печатные формы (которые тоже надо хранить). Форматы документов (xsd-валидация и т.д);
* IDEF0 процесса системы ЭДО?
* Какие есть задачи для оптимизации? Тип из-за того, что данных много к ним нужен быть быстрый доступ, объемы данных растут без остановки. Есть % архивных данных, есть % горячих данных. Нормы операторов ЭДО. Может быть там могут быть требования по хранению данных.

## **1.4 Выбор методов оптимизации хранения данных для системы электронного документооборота**

Тут расписываем про то, что облачные решение — это конечно очень хорошо, но есть проблемы с тем, что обмен данными происходит по сети Интернет и это уже по умолчанию не может происходить быстро, тип предлагаем использовать облачные только для архивных данных и не более. Вопросы безопасности.

Выбираем методы – есть такое метод такой и +-. Версионность. Репликация, Шардинг.

Говорим о том, что берем решение от Amazon с уровнями данных. Прикручиваем дедупликацию данных на уровне файлов.

Встает необходимость разработки ПО, которое и будет «перекидывать» файлы с уровня на уровень, объединять хранилища в одно. Абстрагируем систему ЭДО от знаний устройства хранилища в целом (скрываем все за api (итерфейс прикладного программирования)).

## **1.5 Постановка задачи создания оптимизированного хранилища данных для системы электронного документооборота**

Вводим понятия Cold Storage, Warm Storage, Hot Storage. закупаем RAID массив из SSD в качестве Warm Storage, закупаем оперативку для кэша (куда помещаются прям очень горячие данные).

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Понятие системы хранения данных [Электронный ресурс]. URL: <https://itglobal.com/ru-ru/company/glossary/shd-sistema-hraneniya-dannyh/> (дата обращения 26.11.2020);
2. 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.;
3. И.В. Савин. Анализ систем хранения данных // Журнал «Известия Тульского государственного университета. Технические Науки». – 2019. – С. 193-196;
4. Запоминающие устройство [Электронный ресурс]. URL: <https://ru.wikipedia.org/wiki/Запоминающее_устройство> (дата обращения 26.11.2020)
5. Джиблазде З.Г., Гасюк К.В., Садыков Д.К. SSD и HDD // Журнал «Вестник современных исследований». – 2020. – С. 4-6;
6. Исалёв А.С. Анализ надежности технологии хранения данных RAID // Журнал «Актуальные проблемы авиации и космонавтики». – 2016. – С. 959-961;
7. Шувалов Н.В. Технология RAID // Журнал «Аллея науки». – 2017. – С. 825-828;
8. Distributed Storage: What’s Inside Amazon S3? [Электронный ресурс]. URL: <https://cloudian.com/guides/data-backup/distributed-storage/> (дата обращения 27.11.2020);
9. What is software-defined storage? [Электронный ресурс]. URL: <https://www.redhat.com/en/topics/data-storage/software-defined-storage> (дата обращения 27.11.2020)
10. Understanding distributed data storage [Электронный ресурс]. URL: <https://www.googlinux.com/understanding-distributed-data-storage/> (дата обращения 27.11.2020);
11. Мазур Э.М. Распределенные системы хранения данных: анализ, классификация и выбор // Журнал «Перспективы развития информационных технологий». – 2015. – С. 33-60;
12. Облачные хранилища данных [Электронный ресурс]. URL: <https://ru.wikipedia.org/wiki/Облачное_хранилище_данных> (дата обращения 26.11.2020);
13. Margaret Rouse. Cloud Storage [Электронный ресурс]. URL: <https://searchstorage.techtarget.com/definition/cloud-storage> (дата обращения 27.11.2020);
14. Документация облачного хранилища Amazon S3 [Электронный ресурс]. URL: <https://aws.amazon.com/ru/s3/> (дата обращения 26.11.2020);
15. Архитектура S3: 3 года эволюции Mail.ru Cloud Storage [Электронный ресурс]. URL: <https://habr.com/ru/company/mailru/blog/513356/> (дата обращения 26.11.2020);
16. Документация облачного хранилища Mail Cloud Storage [Электронный ресурс]. URL: <https://mcs.mail.ru/storage/> (дата обращения 26.11.2020);
17. Федеральный закон «О персональных данных» от 27.07.2005 № 152-ФЗ [Электронный ресурс]. URL: <http://www.consultant.ru/document/cons_doc_LAW_61801/> (дата обращения 26.11.2020);
18. Документация облачного хранилища Yandex Object Storage [Электронный ресурс]. URL: <https://cloud.yandex.ru/services/storage> (дата обращения 26.11.2020);