ОСОБЕННОСТИ ХРАНЕНИЯ И ОБРАБОТКИ БОЛЬШИХ ДАННЫХ ЧЕРЕЗ РАСПРЕДЕЛЕННЫЕ ФАЙЛОВЫЕ СИСТЕМЫ

Зайниддинов Олимжон Одил ўғли
магистр Самаркандского государственного
университета имени Шарофа Рашидова
Ахатов Акмал Рустамович
профессор Самаркандского государственного
университета имени Шарофа Рашидова, д.т.н.

Аннотация: В данной статье описаны процессы хранения данных Big Data через распределенные файловые системы и их особенности. Кроме того, широко раскрыты практические аспекты обработки данных через файловые системы.

Ключевые слова: Большие данные, система хранения и обработки данных, озеро данных, файловые данные.

Abstract: This article describes the processes of storing Big Data data through distributed file systems and its specific features. In addition, practical aspects of data processing through file systems are widely disclosed.

Key words: Big data, data storage and processing system, data lake, file data.

Сегодня сложно представить наше время без больших данных. Ведь сегодня информации так много, что возникает множество проблем ее систематизации, хранения и обработки. Основной проблемой хранения данных является систематизация данных через файловые системы. По этой причине практические аспекты возможности хранения и обработки больших данных раскрыты в данной статье в большом масштабе. Что

такое большие данные и их базы данных? Каковы его практические аспекты? этому уделяется особое внимание.

Большие данные (big data) — это очень большое количество неоднородных и быстро падающих цифровых данных, которые не могут быть обработаны обычными методами. В некоторых случаях в понятие больших данных входит и обработка этих данных. В основном объект анализа называется большими данными [1].

Способность совместно использовать диски, каталоги, и файлы по сети это одно из наиболее значительных достижений современных информационных технологий. Эта способность может существенно сократить требования к дисковому пространству компьютеров и облегчить совместную работу пользователей. Компьютеры с Microsoft Windows и MacOS Apple / MacOS X используют для этого механизм совместного использования дисков и директорий. В системах Linux / Unix для тех же самых задач традиционно используется NFS сетевая файловая система.

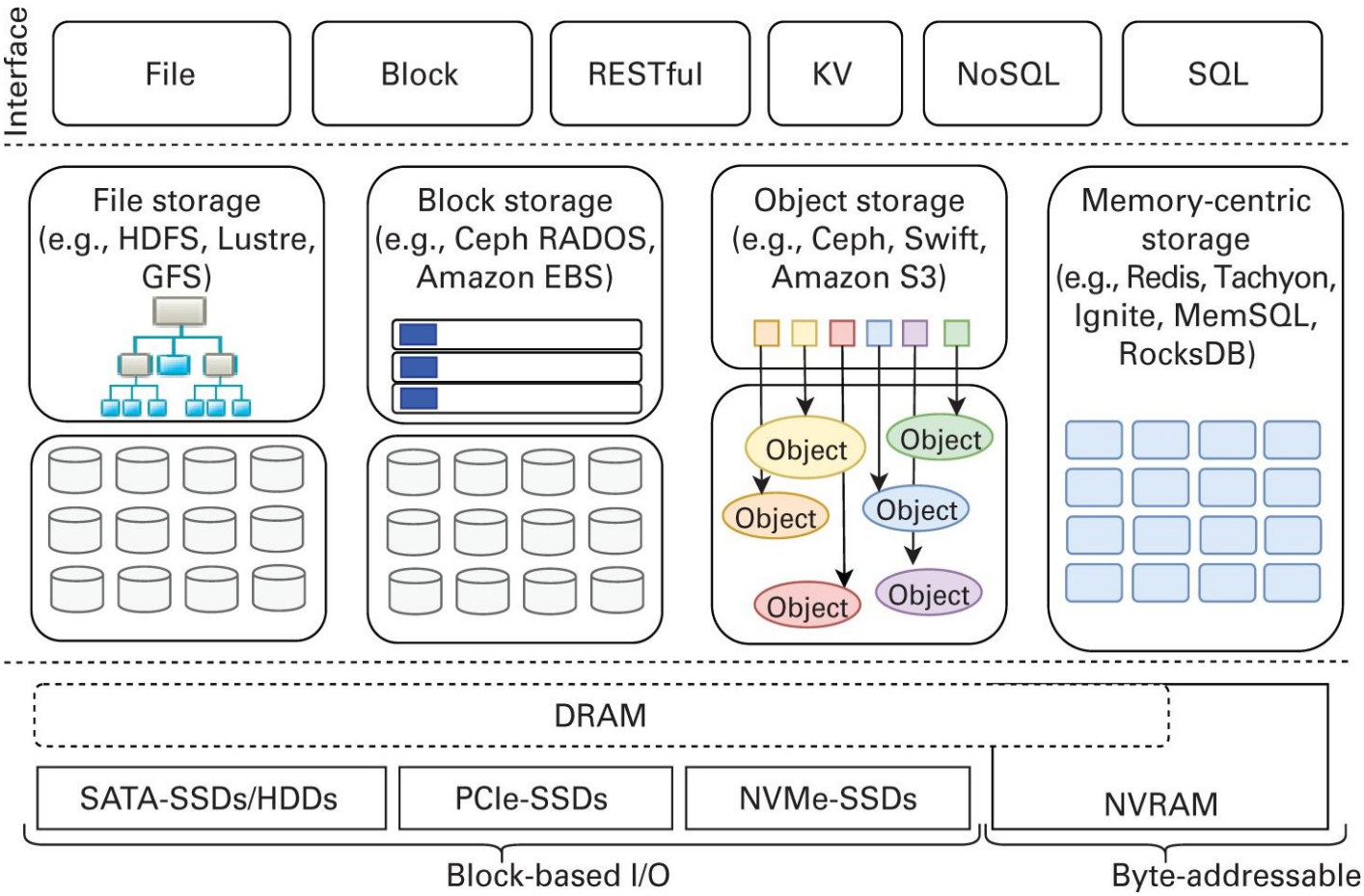
Сейчас сетевые файловые системы называют <<распределенными>>. Этот термин отражает тот факт, что многие из этих файловых систем имеют гораздо больше возможностей, чем простая передача данных по сети. Носители данных, связанные с этими файловыми системами, не обязательно могут быть расположены на одном компьютере они могут быть распределены между многими компьютерами.

Распределенные файловые системы OpenAFS и Coda имеют собственные механизмы управления разделами, которые упрощают возможности хранения общедоступной информации. Они так же поддерживают дублирование способность делать копии разделов и сохранять их на других файловых серверах. Если один файловый сервер становится недоступным, то все равно к данным, хранящимся на его разделах, можно получить доступ с помощью имеющихся резервных копий этих разделов.

Самое главное различие между подходом Windows / MacOS (совместное использование каталогов и дисков) и подходом Linux, MacOS X и других Unix-подобных многопользовательских операционных систем в том, как эти операционные системы используют и организовывают разделы. Windows / MacOS экспортируют разделы как отдельные каталоги или диски, и удаленные системы, которые хотят обратиться к общедоступным устройствам, должны обязательно подключить их к себе.

Мы классифицируем различное ПО промежуточного уровня для хранения применяемого в облачных решениях и кластерных средах Больших данных по четырём основным категориям: файловой хранилище, блочное хранилище, объектное хранилище и ориентированное на оперативную память хранилище (Рисунок 1). Файл и блок - это методы доступа к хранилищу на основе файловой системы. Оба эти метода традиционно применялись для для предоставления приложениям доступа к данным на блочных устройствах хранения, таких как твердотельные накопители и жёсткие диски. Основное различие между ними состоит в том как само приложение способно взаимодействовать с блоками данных на своём устройстве хранения; файловая система обеспечивает доступ через API интерфейсы на основе файлов, таких как POSIX, то есть Portable Operating System Interface - интерфейс переносимой операционной системы (The Open Group, 2011), а блочное хранилище предоставляет прямой доступ к блокам необработанных (сырых, raw) данных. С наступлением бума Больших данных было сделано наблюдение, что бóльшая часть производимых данных является "неструктурированной", неизменной и масштабируется на многие Петабайты в географически распределённых кластерах. Это потребовало разработки новых подходов, которые лучше бы подходили к работе с хранилищем как с обособленными элементами, а также с настраиваемыми метаданными для каждого из таких обособленных элементов вместо файлов или блоков. С другой стороны, по

причине снижения с каждым годом стоимости микросхем оперативной памяти большое внимание уделяется "обработке в оперативной памяти" (in-memory), которая сосредотачивается на максимально возможном применении "активной памяти". Для дополнения этим, как научные исследования, так и аналитика центров обработки данных (ЦОД) обратились к сосредоточенным на оперативной памяти системам хранения, включая хранилищ ключ- значение в оперативной памяти (KV, key- value), а также базы данных и файловые системы в оперативной памяти. На Рисунке 3.1 также продемонстрированы различные интерфейсы приложений (такие как файловый, блочный, службы REST/ RESTfull (Fielding, 2000), KV, SQL (ISO/IEC, 2016), NoSQL (NoSQL Database.org, 2021)), которые обрабатывают рабочие нагрузки Больших данных для доступа к таким различным решениям хранения.



Рысонок 1. Различные типы параллельных и распределённых систем

хранения. ***(PCIe, Peripheral Component Interconnect Express; RADOS, Reliable, Autonomic Distributed Object Store; SATA, Serial AT Attachment).***

В данной главе мы изучим необходимые основы для этих четырёх систем хранения промежуточного уровня и обсудим образцы из реального мира с соответствующими интерфейсами, применяемыми для разработки приложений Больших данных крупного масштаба.

Термин «большие данные» появился в 2008 году. Клиффорд Линч, редактор журнала Nature, использовал термин «большие данные» в специальном выпуске, посвященном быстрому росту объема данных в мире. Однако большие данные существовали и раньше. По мнению экспертов, потоки с более чем 100 ГБ данных в день называются большими данными.

Исходя из этого, возникает следующий важный вопрос, связанный с безопасностью хранения данных и их использования. Например, является ли эта или другая аналитическая платформа, где потребители автоматически отправляют свои данные, безопасными? Кроме того, многие представители бизнеса подчеркивают отсутствие высококвалифицированных аналитиков и маркетологов, способных эффективно управлять большими объемами данных и решать с их помощью конкретные бизнес-задачи.

Несмотря на все сложности внедрения Big Data, бизнес планирует увеличить инвестиции в этом направлении. Согласно исследованию Gartner, медиа, розничные, телекоммуникационные, банковские и сервисные компании являются лидерами в области инвестиций в большие данные.

Список использованной литературы

1. Big Data and Big Data Analytics: Concepts, Types and Technologies November 2018 DOI:10.21276/ijre.2018.5.9.5 Authors: Youssra Riahi

2. Bernard Marr."Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance". John Wiley& Sons Ltd, 2015

3. Efficient development of high performance data analytics

4. Andrea De Mauro, Marco Greco and Michele Grimaldi."What is Big Data? A Consensual Definition and a Review of Key Research Topics". In"AIP Proceedings"2014,"4th International Conference on Integrated Information".

5. Sofia Berto Villas-Boas."Big Data in Firms and Economic Research". Applied economics and Finance, Vol. 1, No. 1; May 2014.

6. Тезисы докладов конференции «Большие данные в национальной экономике», Москва, 21 октября 2014 г.