УДК 004.78

ДОСТОИНСТВА И НЕДОСТАТКИ АППАРАТНЫХ РЕШЕНИЙ СИСТЕМ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Черепанов С.П.

Национальный исследовательский ядерный университет «МИФИ», Москва, e-mail: semen.cherepanov@phystech.edu

«Большие данные» – это термин, который определяет не только обширные по характеристикам объема, разнообразия, скорости и/или вариабельности данные, превосходящие возможности обычных баз данных, но и неструктурированную информацию, перед обработкой и анализом которых бессильны традиционные алгоритмы и аппаратные архитектуры хранения данных. Настоящая статья посвящена обзору и сравнению наиболее популярных аппаратных архитектур для хранения и обработки больших данных. В наши дни большие данные используются как в отдельных организациях, так и в отдельных отраслях, обеспечивая поддержку их деятельности. Главным образом рассмотрены быстродействие, защищенность, удобство в настройке и использовании, а также плюсы и минусы рассматриваемых аппаратных архитектур. Для эффективных обработки и хранения больших данных необходимо не только выбрать правильное программное обеспечение и выбрать средства разработки, обеспечить необходимый бюджет и найти квалифицированных разработчиков, но и правильно подобрать аппаратную архитектуру. Выбор такой архитектуры определяется множеством факторов и различными бизнес-требованиями, уникальными в каждом отдельно взятом случае, такими как безопасность, конфиденциальность, доступный бюджет и минимально необходимое быстродействие.

Ключевые слова: большие данные, аппаратная архитектура, управление данными, типовая архитектура, компьютерная безопасность

ADVANTAGES AND DISADVANTAGES OF HARDWARE BIG DATA SYSTEMS

Cherepanov S.P.

National Research Nuclear University MEPhI, Moscow, e-mail: semen.cherepanov@phystech.edu

«Big data» is not only about volume of data, it is also about unstructured information that can not be handled by traditional algorithms and hardware architectures of data storages. The present paper is devoted to review and comparison of the most popular hardware architecture for storage and processing of big data. Nowadays, big data is used or created by separate companies or for particular sectors and is intended to provide the company's activity. Mainly, considered the performance and security of the examined hardware architectures. If you want to properly process and store big data, it is necessary not only to choose effective software, choose development tools, not only to provide the necessary budget and not only to find the qualified developers, but also to choose a correct hardware architecture. Such choice is determined by a lot of factors and different business requests, unique in each separate case, such as security, privacy, affordable budget, and minimal affordable performance.

Keywords: big data, hardware, data management, reference architecture, computer security

В настоящее время вопрос создания востребованных систем обработки больших данных очень актуален. Под большими данными подразумеваются наборы данных, обширные по характеристикам объема, разнообразия, скорости и/или вариабельности, которые требуют масштабируемой архитектуры для эффективности хранения, манипулирования и анализа. Как правило, каждая система обработки больших данных уникальна, так как предназначена для выполнения функций в рамках конкретной организации [1].

Главной целью создания аппаратного решения является уменьшение алгоритмической сложности обработки данных, что позволяет обеспечивать бесперебойную работу системы, простоту ее контроля, а также достигается многократное ускорение работы системы. Разработка систем обработки больших данных сейчас актуальна и востребована, так как используется во множестве областей, таких как банковское дело, коммуникации, муниципальное управление, медицина, производство и добыча натуральных ресурсов и т.д. [2].

Для создания системы обработки больших данных необходимо осуществить выбор соответствующей аппаратной архитектуры. Аппаратной архитектурой систем обработки больших данных является набор физических компонентов и их взаимосвязи [3].

На данный момент существует множество различных аппаратных архитектур со своими особенностями. Целью настоящей статьи является сравнение наиболее часто используемых аппаратных архитектур и выбор архитектуры, обеспечивающей рациональный процесс создания систем обработки больших данных.

Материалы и методы исследования. В ходе изучения литературы о различных аппаратных архитектурах систем обработки больших данных было выявлено отсутствие формализованного способа их сравнения.

На основании этого было принято решение о создании шаблона паспорта аппаратной архитектуры.

Шаблон паспорта аппаратной архитектуры должен включать следующие разделы:

– Название системы;
– Объем данных;
– Пропускная способность;
– Задержка доступа;
– Избыточность данных;
– Энергопотребление;
– Криптоустойчивость;
– Достоинства;
– Недостатки.

Шаблон паспорта аппаратной архитектуры представлен в табл. 1.

Для сравнения различных аппаратных архитектур систем обработки больших данных было принято решение создать паспорта трех наиболее распространенных и популярных в настоящее время аппаратных архитектур, таких как Google Cloud Platform, Amazon Web Services и NetApp. Данные сервисы рассматриваются в статье как аппаратные архитектуры, поскольку предоставляют возможность для создания кластера хранения больших данных.

Паспорт аппаратной архитектуры Google Cloud Platform представлен в табл. 2.

Паспорт аппаратной архитектуры Amazon Web Services представлен в табл. 3.

Паспорт аппаратной архитектуры NetApp представлен в табл. 4.

Результаты исследования и их обсуждение

1. Сравнение объемов данных.

Архитектуры Google Cloud Platform и Amazon Web Services не ограничены объемом. NetApp SG6060 имеет максимальный объем 2 Пб. Таким образом, максимальный объем данных позволяет использовать все три архитектуры для хранения больших данных.

2. Сравнение пропускной способности.

Пропускная способность архитектур Google Cloud Platform и Amazon Web Services ограничены пропускной способностью внешней сети Интернет. NetApp SG6060 имеет 4 выхода Ethernet с максимальной пропускной способностью 25 Гб/с каждый. При достаточной скорости сети Интернет все три архитектуры можно использовать для работы с большими данными.

Таблица 1

Шаблон паспорта аппаратной архитектуры

Название архитектуры	Полное название, сокращенное
Объем данных	Объем хранимых данных в терабайтах
Пропускная способность	Максимально возможная скорость приема/передачи данных
Задержка доступа	Минимальное время доступа к данным, МС
Избыточность данных	Дупликация данных, расходы на ее поддержку
Энергопотребление	Расход потребления электроэнергии
Криптоустойчивость	Защищенность данных от взлома
Достоинства	Основные достоинства архитектуры
Недостатки	Основные недостатки архитектуры

Таблица 2

Паспорт архитектуры Google Cloud Platform

Название архитектуры	Google Cloud Platform
Объем данных	Неограниченный
Пропускная способность	Зависит от региона, в котором расположены сервер и клиент. Для сервера EU – 489 MB/s [4]
Задержка доступа	Зависит от региона, в котором расположены сервер и клиент. Для сервера EU – 23 мс [4]
Избыточность данных	Так как данные хранятся на удаленном устройстве, избыточность гарантируется компанией Google
Энергопотребление	Отсутствует
Криптоустойчивость	Присутствует множество возможностей для кибератак злоумышленников [5]. Данные могут оказаться утеряны, заблокированы или украдены
Достоинства	Отсутствие необходимости детальной настройки архитектуры
Недостатки	Низкая криптоустойчивость, скорость ограничена скоростью интернет-соединения

Таблица 3

Паспорт архитектуры Amazon Web Services

Название архитектуры	Amazon Web Services
Объем данных	Неограниченный
Пропускная способность	Зависит от региона, в котором расположены сервер и клиент. Для сервера EU North 1 – 295 MB/s [6]
Задержка доступа	Зависит от региона, в котором расположены сервер и клиент. Для сервера EU North 1 – 30 мс [6]
Избыточность данных	Так как данные хранятся на удаленном устройстве, избыточность гарантируется компанией Google
Энергопотребление	Отсутствует
Криптоустойчивость	Присутствует множество возможностей для кибератак злоумышленников [7]. Данные могут оказаться утеряны, заблокированы или украдены
Достоинства	Отсутствие необходимости детальной настройки архитектуры.
Недостатки	Низкая криптоустойчивость, скорость ограничена скоростью интернет-соединения

Таблица 4

Паспорт архитектуры NetApp SG6060

Название архитектуры	NetApp
Объем данных	Представлен в нескольких вариантах от 232 до 696 Тб, с возможностью расширить объем более чем до двух Пб с помощью дополнительных серверных полок
Пропускная способность	4 x 25 Гб/с
Задержка доступа	Менее 0.5 мс
Избыточность данных	Доступна возможность создавать RAID-массивы. Также есть возможность создания резервного хранилища в облаке
Энергопотребление	В среднем 4975 КВт*ч и до 7639 КВт*ч в моменты пиковой нагрузки
Криптоустойчивость	Есть встроенная система защиты данных Commvault
Достоинства	Масштабируемость и удобство настройки
Недостатки	Необходима серверная стойка и охлаждение

3. Сравнение задержки доступа.

Высокие задержки доступа архитектур Google Cloud Platform и Amazon Web Services связаны с удаленностью серверов от конечного пользователя. У архитектуры NetApp SG6060 задержка доступа значительно ниже, достигает значений менее чем в 0.5 мс. Архитектура NetApp значительно превосходит архитектуры Google Cloud Platform и Amazon Web Services по времени доступа, что может быть критично в областях работы с большими данными, чувствительных к скорости данных.

4. Сравнение избыточности данных.

Архитектуры Google Cloud Platform и Amazon Web Services генерируют избыточность данных в полном объеме, так как данные дублируются на различных удаленных серверах. Архитектура NetApp позволяет использовать RAID-массивы и создавать дублирующее хранилище на удаленном сервере. Таким образом, все три архитектуры обладают достаточными возможностями для гарантии избыточности данных.

5. Сравнение энергопотребления.

Стоимость и обслуживание энергопотребления архитектур Google Cloud Platform и Amazon Web Services включены в стоимость их использования. Энергопотребление архитектуры NetApp требует отдельных систем питания и охлаждения.

6. Сравнение криптоустойчивости.

Архитектура NetApp обладает встроенной системой защиты данных. Помимо этого, есть возможность создания защищенной сети, что в совокупности значительно повышает криптоустойчивость архитектуры. Архитектуры Google Cloud Platform и Amazon Web Services неоднократно подвергались кибератакам. Данные могут быть потеряны или украдены. Таким образом, архитектуры Google Cloud Platform и Amazon Web Services подходят только для хранения обезличенных данных либо данных, не представляющих коммерческой ценности без дополнительной обработки, что сильно сужает их область применения. В свою очередь, криптоустойчивость архитектуры

NetApp позволяет использовать ее в любых областях.

7. Сравнение достоинств.

Архитектуры Google Cloud Platform и Amazon Web Services не требуют дополнительной настройки для использования, не имеют физического ограничения в объеме данных. Архитектура NetApp обладает высокой пропускной способностью, низкой задержкой доступа и хорошей защищенностью данных. Легко масштабируется на несколько физических устройств.

8. Сравнение недостатков.

Низкая криптоустойчивость архитектур Google Cloud Platform и Amazon Web Services не позволяет использовать их для хранения персональных данных клиентов либо коммерчески важных данных. Необходимость настройки, проведения системы охлаждения и хорошей системы электроэнергии архитектуры NetApp может стать существенным недостатком для небольшой компании.

Выводы

Проведенный сравнительный анализ архитектур систем обработки больших данных показал следующее.

1. Наиболее рациональной архитектурой обработки больших данных является NetApp вследствие ее полной самодостаточности. Архитектура NetApp обладает высокой пропускной способностью и низкой задержкой доступа. Данная архитектура обладает достаточной защищенностью данных от взлома и имеет возможность дублирования данных с помощью RAID-массива. Архитектура легко масштабируется на объемы данных в несколько Петабайт.

2. Поскольку для функционирования архитектуры NetApp требуется предварительная настройка, система охлаждения, а также достаточный уровень электроэнергии, то при создании обезличенных кластеров данных могут использоваться архитектуры Google Cloud Platform или Amazon Web Services.

3. Не рекомендуется использование архитектур Google Cloud Platform и Amazon Web Services для работы с персональными данными либо важной коммерческой информацией.

4. Несмотря на относительно низкую задержку доступа к данным архитектуры NetApp, она не подойдет для областей бизнеса, где требуются сверхвысокие скорости, таких как биржевые операции.

Список литературы

1. Проект ПНСТ Информационные технологии. Большие данные. Типовая архитектура. 78 с. [Электронный ресурс]. URL: https://sk.ru/documents/143/%D0%9F%D0%9D%D0%A1%D0%A2_%D0%91%D0%94_%D0%A2%D0%B8%D0%BF_%D0%B0%D1%80%D1%85_1%D0%B0%D1%8F.pdf (дата обращения: 16.08.2020).

2. Karlovits I. Technologies for using Big Data in the paper and printing industry. Journal of Print and Media Research Technology. 6. 75-83.10.14622/JPMTR-1706. 2017. P. 75–83.

3. Hardware architecture. [Electronic resource]. URL: https://en.wikipedia.org/wiki/Hardware_architecture (date of access: 16.08.2020).

4. Google Cloud Platform Network Test. [Electronic resource]. URL: https://cloudharmony.com/speedtest-for-google (date of access: 16.08.2020).

5. Why you should not use Google Cloud. [Electronic resource]. URL: https://medium.com/@serverpunch/why-you-should-not-use-google-cloud-75ea2aec00de (date of access: 16.08.2020).

6. Amazon Web Services Network Test. [Electronic resource]. URL: https://cloudharmony.com/speedtest-for-aws (date of access: 16.08.2020).

7. Rootkit in the Cloud: Hacker Group Breaches AWS Servers. [Electronic resource]. URL: https://www.cbronline.com/news/aws-servers-hacked-rootkit-in-the-cloud (date of access: 16.08.2020).