Талип А.К. Обзор экосистемы Hadoop в области больших данных // Международный журнал информационных технологий и энергоэффективности. – 2023. – Т. 9 № 7(45) с. 97–103

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.65

ОБЗОР ЭКОСИСТЕМЫ HADOOP В ОБЛАСТИ БОЛЬШИХ ДАННЫХ

Талип А.К.
КАЗАХСТАНСКО-БРИТАНСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ, Алматы, Казахстан, (50000, Казахстан, Алматы, ул. Толе би, 59), e-mail: talipaltynai00@gmail.com

Данные всегда были ключевым элементом общества, экспоненциально растущим на протяжении веков и представляющим вызовы для каждой системы, с которой они сталкиваются. Возможность быстро обрабатывать и манипулировать данными открывает множество возможностей для инноваций и прогресса. "Большие данные" - термин, который широко обсуждается, но что на самом деле означает этот термин? Как он переосмысливает перспективы в различных областях, от научных исследований до операций компаний, некоммерческих организаций, правительств и других учреждений? Откуда берутся эти данные, как они обрабатываются, и как результаты сохраняются и используются для будущих начинаний? И почему открытые технологии так важны для решения этих вопросов? В этой статье мы собираемся ответить на все эти вопросы, чтобы прояснить, что на самом деле означают "большие данные" и как они влияют на нашу повседневную жизнь. Экосистема Hadoop выступает ведущим решением для обработки и анализа огромных объемов данных. Она включает в себя набор инструментов с открытым исходным кодом, разработанных для решения основных проблем больших данных: объема, скорости и разнообразия. В основе Hadoop лежит распределенная система обработки данных, известная как Apache MapReduce, которая разбивает вычислительные задачи на фазы отображения и сведения, облегчая параллельную обработку на нескольких узлах кластера. Этот распределенный подход существенно повышает производительность анализа больших данных за счет использования мощности параллельных вычислений. Тем не менее, несмотря на свои преимущества, экосистема Hadoop также сталкивается с определенными проблемами.

Ключевые слова: большие данные, Hadoop, HDFS, MapReduce, Экосистема Hadoop, NameNode, DataNode, YARN.

OVERVIEW OF THE HADOOP ECOSYSTEM IN BIG DATA

Talip A.K.
KAZAKH-BRITISH TECHNICAL UNIVERSITY, Almaty, Kazakhstan (50000, Kazakhstan, Almaty, st. Tole bi 59), e-mail: talipaltynai00@gmail.com

Data has always been a crucial element of society, exponentially growing over centuries and presenting challenges to every system it encounters. The ability to rapidly process and manipulate data opens up numerous opportunities for innovation and progress. "Big data" is a term widely discussed, but what does it actually mean? How does it reshape perspectives in various fields, from scientific research to corporate operations, non-profit organizations, governments, and other institutions? Where do these data come from, how are they processed, and how are the results stored and utilized for future endeavors? And why are open technologies so important in addressing these questions? In this article, we aim to answer all these questions to clarify what "big data" really means and how they affect our everyday lives. The Hadoop ecosystem emerges as a leading solution for processing and analyzing vast volumes of data. It encompasses a set of open-source tools designed to address the fundamental challenges of big data: volume, velocity, and variety. At the core of Hadoop lies a distributed data processing system known as Apache MapReduce, which breaks down computational tasks into mapping and reducing phases, facilitating parallel processing across multiple nodes in a cluster. This distributed approach significantly enhances the performance of big data analytics by leveraging the power of parallel computing. However, despite its advantages, the Hadoop ecosystem also faces certain challenges.

Keywords: Big Data, Hadoop, HDFS (Hadoop Distributed File System), MapReduce, Hadoop Ecosystem, NameNode, DataNode, YARN.

Introduction

In the realm of modern technology, where data has be- come the lifeblood of innovation and progress, the Hadoop ecosystem stands as a titan among frameworks, offering a comprehensive solution to the challenges posed by Big Data. As organizations grapple with the ever-increasing volumes, varieties, and velocities of data, Hadoop emerges as a beacon of hope, providing a robust infrastructure for storage, processing, and analysis on an unprecedented scale. The Hadoop ecosystem, with its diverse array of tools and components, represents a paradigm shift in how we approach data management and analytics. From the foundational Hadoop Distributed File System (HDFS) to the versatile processing capabilities of MapReduce and beyond, each component plays a vital role in harnessing the power of Big Data.

Normally [1] the data size is like MB, GB for example considering a video which a few GB may 1GB, 2 GB or 5Gb or it can be some GB. An audio file which is 1000 x 1000 x 1000 terabyte, So in social media everyone shares picture, posts, audio file, video file etc. so it is certainly a large amount of data so this is what a big data is. Examples [2] of big data usage are almost as varied as the data itself. Some prominent examples you're proba- bly already familiar with include: social media networks analyzing their members' data to learn more about them and connect them with content and advertising relevant to their interests, or search engines looking at the relationship between queries and results to give better answers to users' questions.

The internet continues to expand as users find new ways to access information. With the advent of social media, individuals increasingly depend on the internet to fulfill their daily data requirements. In 2020, users generated a staggering 64.2 zettabytes (ZB) of data, surpassing the total number of observable stars in the universe. Pro- jections indicate that data creation will escalate further, reaching an estimated 147 ZB by the conclusion of 2024. In this paper, we embark on a journey through the Hadoop ecosystem, exploring its key components, functionalities, and real-world applications. From understanding the core principles behind Hadoop's design to unraveling the in- tricacies of its various modules, our aim is to provide a comprehensive overview that equips readers with the knowledge needed to navigate the complexities of Big Data with confidence and proficiency. Join us as we delve into the heart of the Hadoop ecosystem and uncover the transformative potential it holds for the future of data- driven decision-making.

Hadoop Ecosystem

The Hadoop ecosystem is a cohesive assembly of open- source software tools, frameworks, and libraries metic- ulously crafted to synergize with Apache Hadoop, an adept framework for distributed storage and processing. This comprehensive ecosystem not only supplements the fundamental capabilities of Hadoop but also furnishes an array of solutions tailored for managing, processing, and analyzing data on a grand scale. Below will be several of the most 4 popular components.

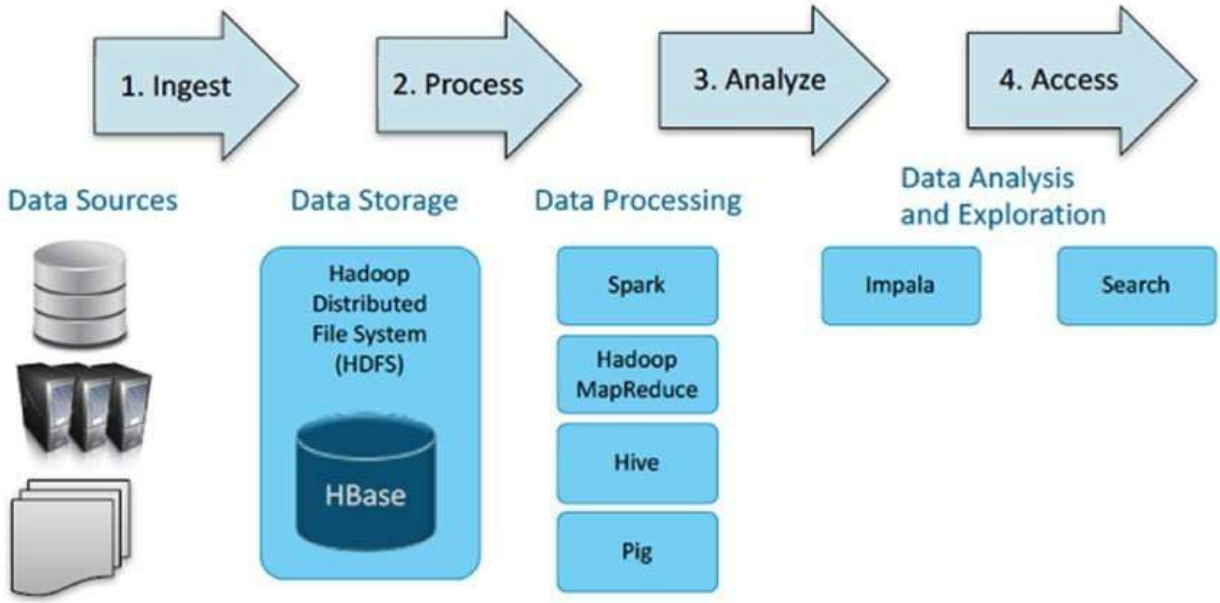

Figure 1 – Hadoop Ecosystem

Big Data

So, what exactly is “Big data”. Put in simple words, it is both structured and unstructured. Generally, it is so gigantic that is provides a challenge to process using conventional database and software techniques. [3]

In the realm of Big Data, there are typically two main lay- ers of data: Data Storage Layer: The Data Storage Layer is where enormous amounts of data are kept. It’s like a giant warehouse where data is stored across multiple locations, such as Hadoop Distributed File System (HDFS), Amazon S3, Google Cloud Storage, and similar platforms. This layer ensures that the data is stored securely and can be easily accessed for further analysis and processing.

Normally the data size is like MB, GB for example considering a video which a few GB may 1GB, 2 GB or 5Gb or it can be some GB. An audio file which is 1000 x 1000 x 1000 terabyte, So in social media everyone shares picture, posts, audio file, video file etc. so it is certainly a large amount of data so this is what a big data is. [4] Data Processing Layer: At this stage, data undergoes analysis, processing, and transformation using different tools like Apache Spark, Apache Hadoop MapReduce, Apache Flink, and similar technologies. This is where valuable insights are extracted from vast amounts of data, and various tasks like combining data, selecting specific data, changing its form, and even teaching computers to learn patterns (machine learning) are carried out.

Hadoop Distributed File System

Hadoop DistributedFile System (HDFS) splits the large data files into parts which are managed by different ma- chines in the cluster. Each part is replicated across many machines in a cluster, so that if there is a single machine failure it does not result in data being unavailable. [5] Every 3 seconds, a Datanode sends a signal called a Heartbeat to the Namenode to signal its status as alive. If no Heartbeat is received for a duration of 10 minutes, a ’Heartbeat Lost’ condition occurs, and the respective DataNode is considered to be inactive or unavailable.

MapReduce

One of the best-known methods for turning raw data into useful information is what is known as MapReduce. MapReduce is a method for taking a large data set and

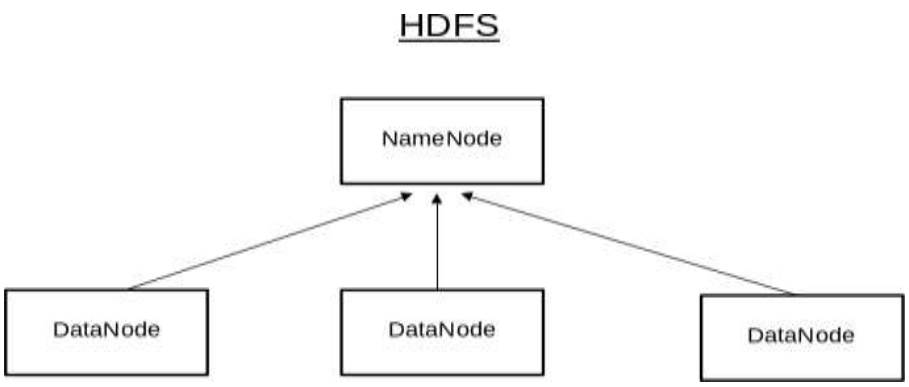

Figure 2 – Hadoop in System Design

performing computations on it across multiple computers, in parallel. It serves as a model for how to program and is often used to refer to the actual implementation of this model. [2] MapReduce is programming model or a software framework used in Apache Hadoop. Hadoop MapReduce is provided for writing applications which process and analyze large data sets in parallel on large multinode clusters of commodity hardware in a scalable, reliable and fault tolerant manner. Data analysis and processing uses two different steps namely, Map phase and Reduce phas [6]

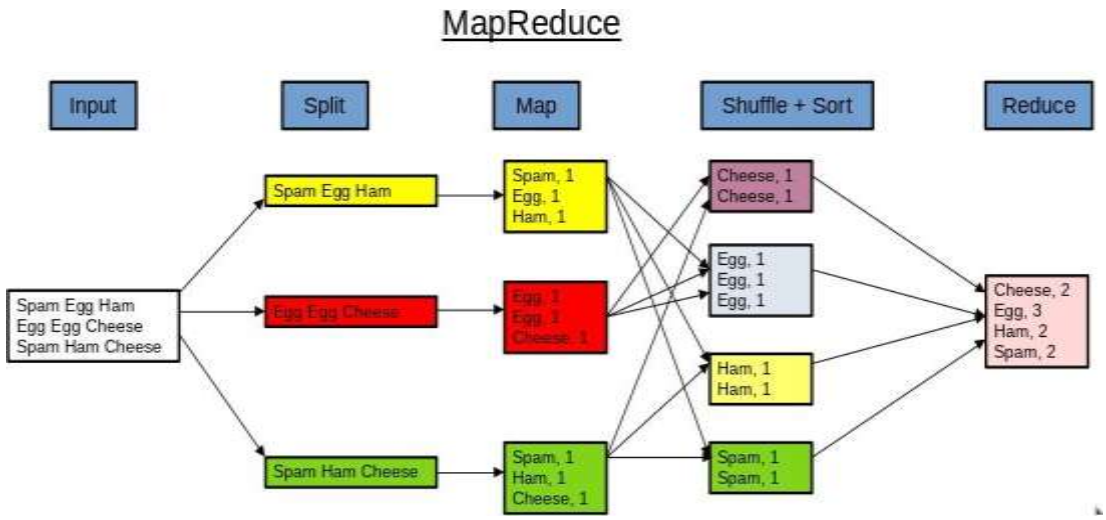

Figure 3 – MapReduce scheme

The input data is initially divided into smaller chunks. These chunks are then processed concurrently by map tasks. After processing, the data chunks are sorted and labeled with occurrence numbers. During the reduce task, aggregation occurs, and the final output is generated. [6]

Yet Another Resource Negotiator (YARN)

YARN's basic idea is to split up the two major func- tionalities of the JobTracker, resource management and job scheduling into separate daemons. The idea is to have a global ResourceManager and per- application Appli- cationMaster. The ResourceManager arbitrates resources among all the applications in the system and it has two components: Scheduler and Applications Manager. [7] The sequence of steps is as follows:

After clients submit MapReduce jobs, they are for- warded to the Resource Manager.

The Resource Manager handles resource allocation and management. It assigns a Container, containing physical resources like CPU and RAM, to initiate the Application Manager.

The Application Manager registers with the Resource

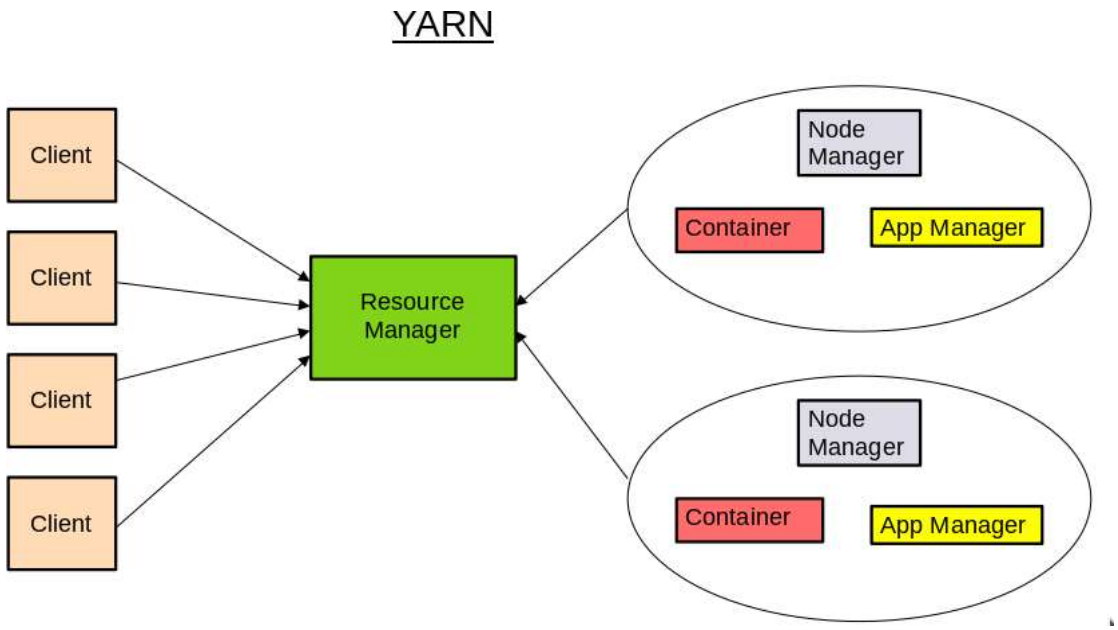

Figure 4 – The workflow of YARN

Manager and requests containers from the Node Manager, which oversees nodes and monitors resource usage.

The Container executes the application code.

Upon completion of processing, the Application Man- ager deregisters from the Resource Manager.

YARN is a core Hadoop service that supports two major services: –Global resource management (ResourceMan- ager) –Per-application management (ApplicationMaster). It offers advantages such as optimized resource utiliza- tion, high scalability, support for various programming languages beyond Java, innovative programming models and services, and flexibility. [8]

Future works and Discussion

Future works and ongoing discussions surrounding the target architecture of the Hadoop ecosystem for solving problems in the big data paradigm are centered on addressing emerging challenges and exploring new oppor- tunities. Here are some areas of focus for future work and discussion:

Scaling and Performance Optimization: As big data continues to grow exponentially, efforts are being made to enhance the scalability and performance of the Hadoop ecosystem. This includes optimizing resource manage- ment, reducing data transfer overhead, and exploring new parallel processing techniques.

Real-time and Stream Processing: The demand for real-time analytics and stream processing has increased significantly. Future work is geared towards integrating real-time processing capabilities into the Hadoop ecosys- tem. Technologies like Apache Kafka and Apache Storm are being explored to enable efficient and low-latency stream processing.

Data Privacy and Security: With the increasing con- cerns surrounding data privacy and security, there is ongoing research and discussion on enhancing the security features of the Hadoop ecosystem. This includes strength- ening authentication mechanisms, encryption techniques, access control, and auditing capabilities. Additionally, efforts are being made to integrate privacy-preserving techniques into data processing frameworks to protect sensitive data while enabling meaningful analysis.

Advanced Analytics and Machine Learning: The inte- gration of advanced analytics and machine learning capa- bilities within the Hadoop ecosystem is an area of active exploration. Researchers and practitioners are working on integrating popular machine learning frameworks (e.g., TensorFlow, PyTorch) and developing scalable algorithms that can take advantage of the distributed computing capabilities of Hadoop.

Data Governance and Compliance: Future work is focusing on improving data governance capabilities within the Hadoop ecosystem. This involves establishing better metadata management, data lineage, and data quality mechanisms. Additionally, efforts are being made to en- sure compliance with data protection regulations (e.g., GDPR, CCPA) and developing tools that facilitate regu- latory compliance in big data environments.

Integration with Cloud and Hybrid Environments: In- tegration of the Hadoop ecosystem with cloud and hybrid environments is gaining prominence. Discussions revolve around leveraging cloud-native services, containerization (e.g., Docker, Kubernetes), and serverless computing (e.g., AWS Lambda, Azure Functions) to enhance the deploy- ment, scalability, and flexibility of Hadoop in cloud and hybrid infrastructures.

Simplification of Development and Management: Fu- ture work aims to simplify the development and manage- ment of Hadoop applications. This includes developing higher-level abstractions, improving tooling support, and streamlining the development lifecycle. Initiatives such as Apache Ambari and Cloudera Manager focus on providing user-friendly interfaces for managing Hadoop clusters, monitoring performance, and optimizing resource utiliza- tion.

Through ongoing research, experimentation, and dis- cussions in these areas, the target architecture of the Hadoop ecosystem will continue to evolve, enabling or- ganizations to effectively tackle the challenges of the big data paradigm and extract actionable insights from their data assets.

Conclusion

The Hadoop ecosystem’s architecture is designed to handle the challenges of big data. It provides a scalable and distributed framework for efficient processing, storage, and analysis of large amounts of data.

The key components of the Hadoop ecosystem are the Hadoop Distributed File System (HDFS) and MapReduce. HDFS allows data to be stored across multiple nodes, ensuring high availability and reliability. MapReduce en- ables parallel and distributed processing of data, making it suitable for large-scale workloads.

The ecosystem also includes tools like Apache Hive and Apache Spark, which extend its capabilities for querying, analysis, and machine learning tasks.

The target architecture of the Hadoop ecosystem is flexible and extensible, allowing integration with other tools and technologies. This flexibility enables organiza- tions to customize solutions and leverage a wide range of applications and libraries within the ecosystem.

In summary, the Hadoop ecosystem's architecture pro- vides scalability, fault tolerance, and distributed process- ing capabilities to handle big data challenges. It empowers organizations to extract insights and value from their data assets effectively.

Список литературы

1. М. Гупта, Ф. Патва и Р. Сандху, "Модель управления доступом на основе атрибутов для безопасной обработки больших данных в экосистеме hadoop", в материалах Третьего семинара ACM по управлению доступом на основе атрибутов, 2018, С. 13-24.
2. opensource.com, "Ресурсы больших данных о opensource.com," https://opensource.com/resources/big-data, текущий год, дата обращения: 27 апреля 2024 года.
3. П.П.Шарма и К.П.Навдети, "Защита больших данных в hadoop: обзор проблем безопасности, угроз и решений", Международный компьютерный журнал. Научно-технический журнал. Технология, том 5, № 2, С. 2126-2131, 2014.
4. С. А. Ханнан, "Обзор больших данных и hadoop", Международный журнал компьютерных приложений, том 154, № 10, 2016.
5. П. С. Хоннутаги, "Распределенная файловая система hadoop", Международный журнал компьютерных наук и информационных технологий (IJCSIT), том 5, № 5, С. 6238-6243, 2014.
6. М. Р. Гази и Д. Гангодкар, "Hadoop, mapreduce и hdfs: взгляд разработчиков", Procedia Computer Science, том 48, С. 45-50, 2015.
7. А. П. Кулкарни и М. Хандевал, "Обзор hadoop и введение в yarn", 2014.
8. С. Мехта и В. Мехта, "Экосистема Hadoop: введение", Международный журнал науки и исследований (IJSR), том 5, № 6, С. 557-562, 2016.

References

1. M. Gupta, F. Patwa, and R. Sandhu, "An attribute-based access control model for secure big data processing in hadoop ecosystem," in Proceedings of the Third ACM Workshop on Attribute-Based Access Control, 2018, pp. 13–24.
2. opensource.com, "Big data resources on opensource.com," https://opensource.com/resources/big-data, current year, ac- cessed: April 27, 2024.
3. P. P. Sharma and C. P. Navdeti, "Securing big data hadoop: a review of security issues, threats and solution," Int. J. Comput. Sci. Inf. Technol, vol. 5, no. 2, pp. 2126–2131, 2014.
4. S. A. Hannan, "An overview on big data and hadoop," Interna- tional Journal of Computer Applications, vol. 154, no. 10, 2016.
5. P. S. Honnutagi, "The hadoop distributed file system," Interna- tional Journal of Computer Science and Information Technologies (IJCSIT), vol. 5, no. 5, pp. 6238–6243, 2014.
6. M. R. Ghazi and D. Gangodkar, "Hadoop, mapreduce and hdfs: a developers perspective," Procedia Computer Science, vol. 48, pp. 45–50, 2015.
7. A. P. Kulkarni and M. Khandewal, "Survey on hadoop and introduction to yarn." 2014.
8. S. Mehta and V. Mehta, "Hadoop ecosystem: An introduction," International Journal of Science and Research (IJSR), vol. 5, no. 6, pp. 557–562, 2016.