Интегриране на Amazon Web Services(AWS) и Hadoop за обработка на Големи данни

Integrating Amazon Web Services (AWS) and Hadoop for Big Data processing

Гено Стефанов[1]

Абстракт

В днешния дигитален свят обемът на генерираните данни е огромен, което прави тяхното ефективното съхранение, обработка и анализ от съществено значение за организациите. Правилната интеграция на AWS и Hadoop може да предостави многобройни възможности на бизнес организациите, като ще им даде възможност да се възползват от предимствата на двете технологии. AWS предлага мащабируема и гъвкава среда за облачни услуги(cloud services), докато Hadoop отворена, разпределена система за обработка на Големи данни. В настоящия доклад ще бъдат разгледани и анализирани предимствата и възможностите за интеграция на двете технологии.

Abstract

In our digital world the volume of data generated is enormous, making its efficient storage, processing, and analysis essential for organizations. Proper integration of AWS and Hadoop can provide numerous opportunities for business organizations, allowing them to take advantage of the advantages of both technologies. AWS offers a scalable and flexible environment for cloud services, while Hadoop is an open, distributed system for processing Big Data. This report will examine and analyze the advantages and opportunities for integration of both technologies.

Ключови думи: Големи Данни, Hadoop, AWS, Big Data.

JEL: O

1. Въведение

С нарастването на обема на данните идващи от източници като социални мрежи, Интернет на нещата (IoT) и електронна търговия, организациите се изправят пред нуждата от мощни платформи за управление на данните. Hadoop, като популярна система за съхранение и обработка на Големи данни, предлага разпределено съхранение и паралелна обработка, която значително подобрява производителността. Въпреки това, управлението на Hadoop в локална инфраструктура изисква сериозни инвестиции и ресурси. AWS предлага облачна платформа с динамично мащабиране и висока надеждност, която значително намалява разходите за инфраструктура и улеснява внедряването на Hadoop в облачна среда.

Интеграцията между AWS и Hadoop позволява на организациите да мащабират динамично обработката на данни, да използват управлявани услуги като Amazon EMR (Elastic MapReduce) и да намалят сложността на управлението на инфраструктурата. Този доклад представя основните аспекти и ползи от интеграцията на Hadoop с AWS.

[1] Гл.ас.д-р. катедра Информационни Технологии и Комуникации, УНСС, e-mail: genostefanov@unwe.bg

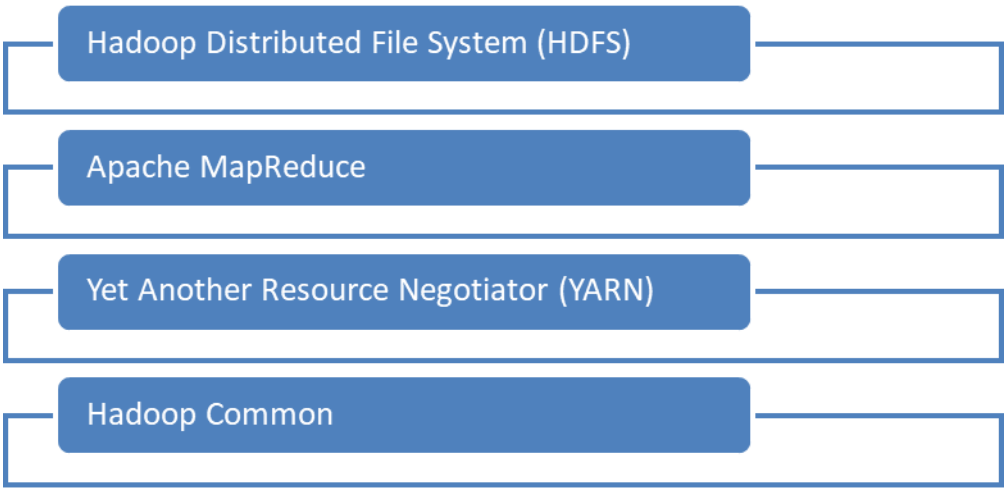
2. AWS и Hadoop

Hadoop

В съвременния свят, дигитализацията на икономиката и живота генерира огромно количество данни. Управлението на тези данни, както и тяхното ефективно използване става все по трудно. Hadoop от своя страна е един от мощните инструментите за обработка на тези големи данни.

Hadoop се състои от няколко основни елемента [1] (фиг. 1):

- Hadoop Distributed File System (HDFS) - HDFS осигурява разпределено съхранение на данни
- Apache MapReduce - предоставя модел за обработка на данни в разпределена среда.
- Yet Another Resource Negotiator (YARN) – платформа, която управлява и изчислителните ресурси в клъстерите.
- Hadoop Common – набор от библиотеки, които подпомагат други аспекти на Hadoop, като мрежови компоненти, сигурност и конфигурация



Фигура 1: Основни компоненти на Hadoop

AWS

AWS (Amazon Web Services) е водещата облачна платформа, която предлага широк спектър от услуги за облачно съхранение, обработка и управление на данни. AWS предоставя гъвкава и скалируема инфраструктура, която позволява на организациите да изграждат и разширяват своите IT ресурси в облака.

AWS предлага над 200 различни услуги, които обхващат различни области, включително изчисления, съхранение на данни, бази данни, мрежови услуги, аналитични инструменти, машинно обучение, изкуствен интелект, интернет на обектите (IoT) и много други. Някои от най-известните услуги на AWS включват Amazon EC2 (Elastic Compute Cloud) за виртуални сървъри, Amazon S3 (Simple

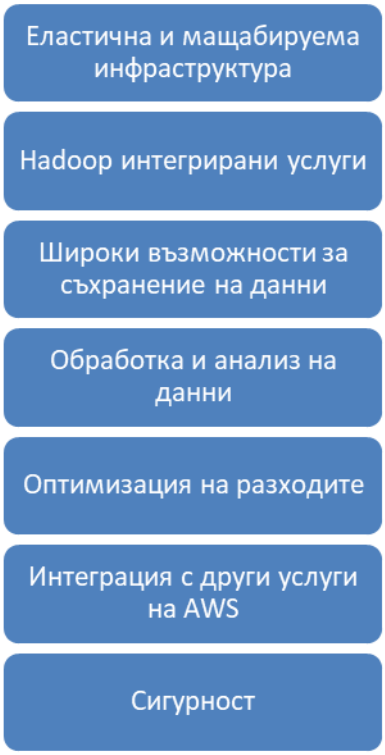
Storage Service) за облачно съхранение на данни, Amazon RDS (Relational Database Service) за управление на релационни бази данни и Amazon Lambda за безсървърно изчисление.

3. Възможности за интеграция

Някои от ключовите възможности и ползи от интегрирането на AWS и Hadoop са (фиг. 2):

- Еластична и мащабируема инфраструктура: AWS осигурява мащабируема инфраструктура чрез услуги като Amazon Elastic Compute Cloud (EC2) и Amazon Elastic MapReduce (EMR), за които ще стане по-подробно в изложението. Използвайки тези услуги, организациите могат лесно да предоставят и мащабират клъстерите на Hadoop въз основа на техните нужди за обработка на данни. Тази еластичност позволява ефективно използване на изчислителните ресурси и оптимизиране на разходите.
- Hadoop услуги: AWS предлага Amazon EMR(Amazon Elastic MapReduce), която опростява внедряването и управлението на Hadoop клъстери. С EMR компаниите могат бързо и сравнително лесно да се подсигурят необходимата инфраструктура за Hadoop клъстерите. Освен това, ще могат да автоматизират управлението на Hadoop клъстерите, което ще им позволи да се съсредоточат повече върху обработката и анализа на данни, а не върху управлението на инфраструктурата.
- Широки възможности за съхранение на данни: AWS предоставя различни услуги за съхранение, като Amazon Simple Storage Service (S3), Amazon Elastic Block Store (EBS) и Amazon Elastic File System (EFS). Тези опции за съхранение могат да бъдат безпроблемно интегрирани с Hadoop, което позволява на ефективното съхранение и достъп до големи обеми от данни.
- Обработка и анализ на данни: Hadoop е предназначен за разпределена обработка на големи масиви от данни, което позволява на организациите да извършват пакетна обработка, анализи в реално време и сложни трансформации на данни. Чрез интегрирането на Hadoop с AWS услуги като EMR, организациите могат да използват силата на Hadoop за бързо обработка на големи масиви от данни, като същевременно се възползват от инфраструктурата и възможностите за управление на AWS.
- Оптимизация на разходите: AWS, като доставчик на облачни услуги предоставя модел на ценообразуване при който бизнеса плаща само за ресурсите, които използва. Това много добре съответства на разпределения характер на Hadoop, тъй като позволява скалируемост на инфраструктурата нагоре или надолу въз основа на нуждите, което води до оптимизиране на разходите.
- Интеграция с други услуги на AWS: AWS предлага широка гама от услуги, които могат да бъдат интегрирани в едно цялостно решение за големи данни. Например, AWS Glue може да се използва за каталогизиране на данни и като ETL (Extract, Transform, Load) инструмент, AWS Lambda за безсървърни (serverless) изчисления и Amazon Redshift за съхранение на данни. Тези услуги могат да подобрят общите възможности за обработка и анализ на данни, когато се комбинират с Hadoop.
- Сигурност: AWS предоставя функции за сигурност, което позволява лесното спазване на регулаторни изисквания и като цяло за информационната сигурност. Услуги като VPC (Virtual

Private Cloud), IAM (Identity Access Management) и Cognito могат да се използват за гарантиране на защитата и поверителността на данните в цялостното решение за големи данни.



Фигура 2: Възможности и ползи от интеграцията на AWS и Hadoop

Едно от основните предизвикателства при изпълнението на Hadoop върху AWS е как да се интегрира с други AWS услуги и източници на данни.

В този документ ще разгледаме някои от възможностите и най-добрите практики за интегриране на AWS с Hadoop, като се съсредоточим върху следните аспекти:

- Как да се използва Amazon Elastic MapReduce (EMR) като управлявана услуга за изпълнение на Hadoop клъстери в AWS.
- Как да се използва модула hadoop-aws, за да се позволи на Hadoop да получи достъп до данни, съхранявани в Amazon Simple Storage Service (S3).
- Как да се използва HiveQL скриптове за обработка и анализ на данни, съхранявани в S3 с помощта на Hadoop.

AWS Elastic MapReduce

Amazon EMR е управлявана услуга, която прави бързо, лесно и рентабилно стартирането на Apache Hadoop и Spark в среда на AWS. С Amazon EMR може да се стартира напълно функционален Hadoop клъстер за минути, без да се налага да се осигурява инфраструктура и конфигуриране. Освен това, в зависимост от нуждите и бюджета, клъстерите могат да се мащабират динамично нагоре и надолу.

Основните стъпки при стартиране на Hadoop клъстер с помощта на Amazon EMR са:

- Дефиниране и конфигуриране на броя и вид на EC2 инстанциите, които ще формират клъстерните възли.
- Софтуерната конфигурация на клъстера, включително версията на Hadoop и всички допълнителни приложения или библиотеки, които ще се инсталират.
- Настройките на сигурност на клъстера, като двойката ключове EC2, роли дефинирани в IAM и групите за сигурност(Security Groups).
- Местоположението на входни и изходни данни в S3 или други източници.

За изпълнението на горе описаните стъпки, AWS предоставя няколко възможности, а именно използването на AWS конзолата за управление, AWS CLI (Command Line Interface) и AWS SDK.

Hadoop и AWS S3

S3 е силно мащабируема и сигурна услуга за съхранение на обекти, която може да съхранява всякакво количество данни във всякакъв формат. S3 често се използва като езеро от данни (Data lake) или склад за данни (Data WareHouse) за съхранение на сурови или обработени данни. Hadoop използва модула hadoop-aws за да достъпи данни съхранявани в S3.

Модулът hadoop-aws включва в себе си S3A клиента, който е реализация на файлова система, съвместима с Hadoop, която позволява на приложенията на Hadoop да четат и записват данни от S3, използвайки стандартните API-та на файловата система на Hadoop. S3A клиентът поддържа функции като batch uploads, сървърно криптиране (server-side encryption) и оптимизация на производителността.

Включването на S3A клиента става чрез добавянето на hadoop-aws в списъка с допълнителни модули (HADOOP_OPTIONAL_TOOLS) в конфигурационния файл на Hadoop - hadoop-env.sh Също така трябва да се конфигурират някои параметри в core-site.xml файлът и да се добавят AWS идентификационни данни(като Access и Secret ключовете) и други настройки за достъп до S3. Примерен конфигурационен файл би изглеждал по следния начин:

<configuration>

<property>

<name>fs.s3a.access.key</name>

<value>YOUR_ACCESS_KEY</value>

</property>

<property>

<name>fs.s3a.secret.key</name>

<value>YOUR_SECRET_KEY</value>

</property>

<property>

<name>fs.s3a.endpoint</name>

<value>s3.amazonaws.com</value>

</property>

<property>

<name>fs.s3a.impl</name>

<value>org.apache.hadoop.fs.s3a.S3AFileSystem</value>

</property>

</configuration>

HiveQL за обработка и анализ на данни

вHive е Apache Hadoop базирана система за обработка на данни, която предоставя възможности за анализ и изпълнение на заявки (SQL queries) върху големи данни, съхранявани в HDFS или други системи. HiveQL (Hive Query Language) е част от Hive и представлява SQL-базиран език за запитвания, който позволява на обработката и анализа на структурни и полу структурирани данни в Hadoop.

Hive може да се интегрира с AWS чрез Amazon EMR. Освен това Hive има възможност да достъпи данни съхранявани в S3 с помощта на S3A клиента.

Използването на Hive за четене на данни от s3 изисква създаването на таблица. Примерен скрипт за създаване на Logs таблица в Hive:

```
CREATE EXTERNAL TABLE logs (

host STRING,

identity STRING,

user STRING,

time STRING,

request STRING,

status STRING,

size STRING,

referer STRING,

agent STRING

)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'

WITH SERDEPROPERTIES (

"input.regex" = "([^ ]*) ([^ ]*) ([^ ]*) (-|\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)(?: ([^ \"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\"))?"

)
```

LOCATION 's3a://your-bucket/logs/';

Така дефинираната таблица logs служи като указател към обектите в s3. Използвайки HiveQL езика данните от таблицата може да се прочетат и да се обработят. Примерен HiveQL скрипт:

SELECT status, count(*) AS count

FROM logs

GROUP BY status

ORDER BY count DESC;

Освен това данните от обработката (резултатът от горната заявка) могат да се запишат обратно в S3. Примерен HiveQL скрипт:

INSERT OVERWRITE DIRECTORY 's3a://your-bucket/results/'

SELECT status, count(*) AS count

FROM logs

GROUP BY status ORDER BY count DESC

4. Заключение

Интегрирането на AWS и Hadoop предлага възможности за използване на мащабируема инфраструктура, ефективно управление на клъстерите на Hadoop, обработка и анализ на големи данни, оптимизиране на разходите и подобряване на сигурността и съответствието на данните. Тя позволява на бизнес организациите да се съсредоточат върху извличането на знания от своите данни, като същевременно се възползват от силата и гъвкавостта на изчисленията в облака.

References

1. Stefanova, K., Kabakchieva, D., Big Data Approach and Dimensions for Educational Industry, Economic Alternatives, Issue 1, 2019, ISSN 1312-7462.
2. Apache Hadoop on Amazon EMR https://aws.amazon.com/emr/features/hadoop/
3. Hadoop-AWS module: Integration with Amazon Web Services https://hadoop.apache.org/docs/current2/hadoop-aws/tools/hadoop-aws/index.html
4. Boyanov L., The Digital World - The Change, The global digital transformation - enriching or impoverishing humanity, ISBN 978-619-239-637-4, Avangard Prima Publ., Sofia 2021, 188 p.
5. M. Tsaneva, "A Practical Approach For Integrating Heterogeneous Systems," Business management, no. 2, p. 11, 2019.
6. V. Mihova, Common Architecture Design of a Business Information System for Performance Management of the Business Applications, in 3rd International conference on application of information and communication technology and statistics in economy and education ICAICTSEE–2013, Sofia, Bulgaria, 2013.
7. P. Milev, Technological Issues of Storing Dynamic Data in a Relational Database on Research Projects, Trakia Journal of Sciences, vol. 13, pp. 22-25, 2015
8. E. Karkalikova, A. Murdjeva, Organization of Data in Data Lake – Real-Life Practice, 11th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education ICAICTSEE– 2021, Sofia, Bulgaria.

9. P. Milev, Approach for Analysis and Comparison of Search Query Results in Web Publications, 11th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education ICAICTSEE– 2021, Sofia, Bulgaria
10. Marzovanova M., Building Multi-Touch User Interface, 4TH International Conference on Application of Information and Communication Technology and Statistics in Economy And Education (ICAICTSEE-2014), 2014, (icaictsee.unwe.bg/past-conferences/ICAICTSEE-2014.pdf).
11. Mihova V., Murdjeva A. Metadata for generating a specific data warehouse. International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE-2012), Sofia, Bulgaria, 2012. (icaictsee.unwe.bg/past-conferences/ICAICTSEE-2012.pdf)