Computer Science and Application 计算机科学与应用, 2022, 12(1), 114-122
Published Online January 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.121013

Hans汉斯

大数据访问控制综述

薛　涛[1,2]，文　雨[2]

[1]中国科学院大学网络空间安全学院，北京
[2]中国科学院信息工程研究所，北京

收稿日期：2021年12月22日；录用日期：2022年1月19日；发布日期：2022年1月26日

摘　要

当今大数据时代，数据存储系统、大数据计算平台发展迅速，而访问控制作为保护数据的基础能力没有得到充分的考虑。首先，本文概括出大数据计算平台数据处理流程，并总结出其中的访问控制需求；然后按照访问控制需求综述并分析相应的访问控制技术。最后对未来访问控制技术的发展进行了展望。

关键词

大数据，访问控制，数据保护

The Overview of Big Data Access Control

Tao Xue[1,2], Yu Wen[2]

[1]School of Cyber Security, University of Chinese Academy of Sciences, Beijing
[2]Institue of Information Engineering, Chinese Academy of Sciences, Beijing

Received: Dec. 22nd, 2021; accepted: Jan. 19th, 2022; published: Jan. 26th, 2022

Abstract

Nowadays, in the era of big data, data storage system and big data computing platform are developing rapidly, but access control as the basic capability of data protection has not been fully considered. First, this paper summarizes the process of data processing in big data computing platform, and summarizes the corresponding access control requirements; then, according to those requirements, it summarizes and analyzes the corresponding access control technologies. Finally, the future development of access control technology is prospected.

Keywords

Big Data, Access Control, Data Protection

文章引用: 薛涛, 文雨. 大数据访问控制综述[J]. 计算机科学与应用, 2022, 12(1): 114-122.
DOI: 10.12677/csa.2022.121013

1. 引言

人类已经进入大数据时代，人们的生活习惯、工作方式、思维活动都受到了大数据的影响[1]。现今很多新颖的技术应用——例如，物联网技术，智能移动设备，社交网络平台——产生大量数据，计算机存储的数据量越来越大，数据类型也多种多样；大数据计算平台可以访问各种各样的数据源，处理数据的能力得到了大幅度提升。通过数据之间的联合分析产生非常有价值的信息：数据之间的联系、新颖的数据使用方法、高效的商业决策等，人们驾驭数据的能力也得到了提升。

访问控制是现代大数据计算平台(例如，Apache Spark, Apache Hadoop, Apache Flink [2])中基础的数据保护功能[3]。这些平台被设计之初主要考虑高效地处理大规模数据，而其中的访问控制功能设计不足。文献[4]指出大数据计算平台需要多种访问控制技术相互配合去保护数据。基于大数据计算平台的现状，本文概括出包括三个组件的数据处理流程：数据源、数据处理逻辑、数据处理测量，并根据这三个组件系统性地总结出对应的三方面访问控制需求；并根据访问控制需求，梳理现有的访问控制技术；最后对大数据访问控制领域的发展进行展望。

2. 访问控制需求

本节首先概括出大数据计算平台数据处理流程，然后根据该数据处理流程总结出相应的访问控制需求。如图 1 所示，现阶段大数据计算平台处理数据的流程概括如下：首先从各种各样的数据源中获取数据，然后获取的数据经过用户编写的数据处理逻辑，最终用户得到输出结果，此外数据处理逻辑执行过程中的中间数据也可以被用户测量。

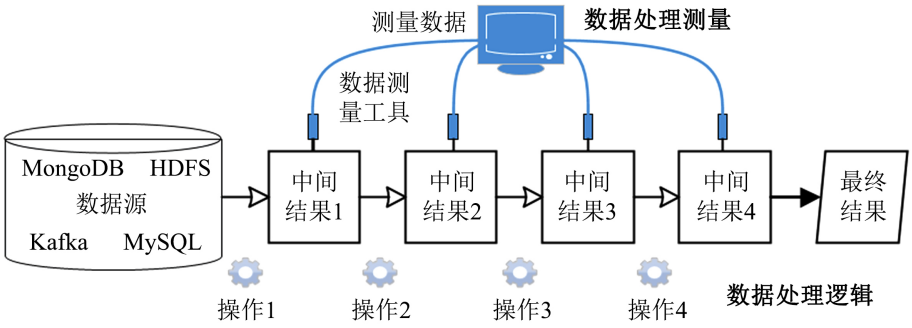

Figure 1. Data processing diagram of big data computing platform
图 1. 大数据计算平台数据处理示意图

数据源：在数据存储这一环节中，用户会根据应用场景的需求选择数据存储系统[5]，因此大数据计算平台需要有能力处理异构存储系统中的数据。大数据计算平台提供各种 API 用于接入多样的数据源，如此大数据计算平台上的用户可以方便地获取数据。

数据处理逻辑：在实际的应用场景中，用户会面临种类繁多的数据分析任务，比如，从数据源简单地加载数据，SQL 查询，流计算，机器学习等[6]。大数据计算平台底层设计有计算执行引擎，面向用户提供了可以灵活使用的算子、封装了 SQL 系统、设计了流计算引擎、封装了机器学习库等。用户可以方

便地实现各种数据处理逻辑，从而可以实现各种数据分析任务。用户设计的每一个数据分析任务可能会包含各种数据处理逻辑，一条数据处理逻辑由一系列的数据操作组成：首先从数据源读取数据，然后对数据进行一系列的转换[3]。

数据处理测量：在数据分析任务执行过程中，用户可能需要了解其中的细节数据[6]。大数据计算平台向用户提供可插拔的数据处理测量工具，比如 Apache Spark 中的累加器。用户可以使用这些数据处理测量工具任意抓取数据处理逻辑执行过程中的细节数据。通过对数据处理逻辑测量，用户可以优化数据处理逻辑、监控计算执行过程等等。

通过上文对大数据计算平台处理数据流程的概括，我们可以清晰地看出大数据计算平台需要多种访问控制技术相互配合去保护数据。我们需要考虑三个方面的访问控制技术：1) 数据源方面上的访问控制技术，2) 数据处理方面上的访问控制技术，3) 数据处理测量方面上的访问控制技术。数据源上的访问控制用于授权大数据计算平台用户可以使用的数据，其中的授权数据可能包含敏感数据；某些大数据计算平台用户可能可以看到敏感数据，而某些用户则不能，所以大数据计算平台也要具备保护敏感数据的能力，通过访问控制确保敏感数据对某些用户来说只能用于计算；但是，由于数据处理测量工具可以任意地抓取数据处理逻辑执行过程中的细节数据，因此大数据计算平台也需要通过访问控制技术确保某些用户不能通过数据处理测量工具抓取到敏感数据。下一章节，基于本段介绍的访问控制需求，我们分别展开相关访问控制技术的综述与分析。

3. 相关访问控制技术

本节基于第 2 节介绍的访问控制需求对访问控制技术展开综述与分析。

3.1. 访问控制技术——数据源

下面我们主要综述并分析以下几类数据源的访问控制技术：NoSQL 数据源，关系型数据源，文件系统数据源，消息队列。比如，Apache Spark 可以接入 Cassandra、HBase、MongoDB、Redis 等 NoSQL 类型数据库，可以接入 MySQL，SQL Server 等关系型数据库，可以接入 HDFS、LFS 等文件系统，可以接入 Kafka 等消息队列。数据源使用的数据模型可能不同[7] [8] [9]，相应的数据保护粒度也不同。下面我们从数据源上的数据模型出发去介绍访问控制技术。

在做数据分析时，数据分析人员会使用到关系型数据源[10]。关系型数据模型[11]中的核心组件包括：表、列、行、单元，因此对应的数据保护粒度可以包括：表级、列级、行级、单元级。有大量访问控制工作基于关系型数据模型[12]，现有的访问控制技术可以实现所有细粒度级别的数据保护；此外关系型的数据库/数据仓库(比如 MySQL，Hive)在工业界长期发展；因此关系型数据源上的访问控制技术已经很成熟。

虽然关系型数据库在访问控制方面已经很成熟，但是关系型数据库严格耦合于关系型数据模型。在大数据时代，关系型数据库很难满足实时数据分析的需求，在可用性、可扩展性上受到限制；NoSQL 数据库为了实现大数据场景下的实时数据分析、良好的可用性及可扩展性，它们采用了比关系数据模型更灵活的数据模型[13]。下面我们介绍三种数据模型：1) Bigtable [5]数据模型，其中的核心组件包括：表(与关系型数据模型中的表有很大不同，关系型数据模型中的表有预定义的 Schema，而这里的表一般有预定义列族，每一列族内可以包含若干列)，列族，列，行，单元，因此对应的数据保护粒度可以包括表级、列族级、列级、行级、单元级；2) 文档数据模型[5]，其中核心组件包括集合、文档、域，因此对应的数据保护粒度可以包括集合级、文档级、域级；3) 键－值数据模型[5]核心组件是键，因此对应的数据保护粒度为键级。基于以上介绍的数据模型，学术界/工业界为 NoSQL 数据库设计了各种各样的访问控制技术。

NoSQL 数据库发展之初，普遍采用粗粒度的访问控制机制，甚至没有访问控制机制；这为 NoSQL

中隐私数据的保护带来挑战，也就产生了各种访问控制技术[7]。K-VAC [14]可以为采用类似于 Bigtable 数据模型的数据库(比如，Cassandra 与 HBase)提供各种数据保护粒度，最细粒度为单元级，其中访问控制策略的制定可以基于数据内容、用户角色、上下文信息。K-VAC 提升了系统的访问控制能力。K-VAC 可用于需要对存储数据实施严格访问控制的场景(比如，医疗卫生)。为了进一步提升 NoSQL 数据库的安全性，Shalabi 等[15]提出了一个基于角色的访问控制模型及协议，用于加密执行访问控制，攻击者即使绕过访问控制机制也不能读到敏感数据。HBase 原生地支持基于角色的访问控制，但是数据访问控制粒度为粗粒度，用于授权的元素不够丰富。而基于属性的访问控制模型可以提供非常灵活的授权，Huang L 等[16]为 HBase 定制了基于属性的访问控制模型，该模型可以支持上下文，最细粒度为列级。MongDB 采用文档数据模型[13]。MongoDB 原生地支持基于角色的访问控制，但是数据访问控制粒度为集合级。为了满足 MongoDB 实施隐私策略的需求以及细粒度的访问控制需求，Mem [17]提出一种基于目的的策略实施方案，最细粒度为文档级。为了进一步加强 Mem 的访问控制能力，ConfinedMem [18]在 Mem 的基础之上进行扩展，最细粒度升级为域级，并且支持基于内容的访问控制，基于上下文的访问控制。Redis 采用键－值模型。Redis 最初采用的安全模型是：用户一旦获取数据库访问权限，即可任意访问数据[9]；为了进一步限制用户对键的访问以提高数据安全性，Redis 采用访问控制列表授权用户对某些键的访问权限。表 1 列出了所综述的 NoSQL 数据源。

Table 1. Access control techniques on NoSQL data sources
表 1. NoSQL 数据源上的访问控制技术

数据源	数据模型	访问控制技术	最细粒度(数据)	是否支持上下文	主要用途
Cassandra	Bigtable [5]模型	K-VAC [14]	单元	是	提供各种数据访问控制粒度
		Shalabi 等[15]	单元	否	加密执行访问控制策略
HBase		K-VAC [14]	单元	是	提供各种数据访问控制粒度
		Huang L 等[16]	列	是	提供非常灵活的授权
MongoDB	文档模型[13]	Mem [17]	文档	否	实施隐私策略，细粒度访问控制
		ConfinedMem [18]	域	是	将细粒度访问控制升级
Redis	键－值[5]	基于键的访问控制[9]	键	否	限制用户对键的访问，提高安全性

NoSQL 数据库中的数据模型多种多样，访问控制功能也多种多样，这为数据库的安全管理带来挑战。因此，近年来产生了为多种 NoSQL 数据库统一提供访问控制功能的技术。由于基于属性的访问控制模型的灵活性，Colombo 等[19]为支持 SQL++的 NoSQL 数据库提出了统一的基于属性的访问控制技术。Gupta 等[8]为所有 NoSQL 数据库提出了统一的基于属性的访问控制方案。

在大数据存储中，文件系统需要支持分布式的大规模的数据存储。HDFS 是工业界大数据处理中常用的文件系统，它与 Linux 文件系统类似，采用纯文本数据模型。纯文本数据模型核心组件包括文件、记录，因此对应的数据保护粒度可以包括文件级、记录级。HDFS 对文件及文件夹的授权模型与 POSIX 模型类似，HDFS 本身不支持细粒度的访问控制[20]。

为了实现大数据分布式场景下进程与进程之间的数据传递，开发人员专门研究用于分布式场景下的消息中间件。消息中间件中的消息数据模型核心组件是主题，因此对应的数据保护粒度为主题级。Kafka

是工业界大数据处理中常用的消息中间件，它采用访问控制列表在不同的主题之间对不同的用户实施隔离[21]。

3.2. 访问控制技术——数据处理

根据文献[4] [22] [23]，由于大数据计算平台经常处理敏感数据，并且用户使用数据的上下文不同(比如，由于分布式环境，使用数据的地点不同)，大数据计算平台需要细粒度及上下文感知的访问控制。由于 HDFS 文件系统只提供粗粒度的访问控制功能，基于传统的基于角色的访问控制模型，科研人员已经为基于 MapReduce 的计算框架(比如 Hadoop)提出许多细粒度的访问控制技术，比如 GuardMR [24]支持细粒度的非结构化数据访问控制。流数据计算在实际应用中越来越受到重视，其中会涉及到大量实时的敏感数据(比如手机用户的位置信息)，SparkXS [25]为 Spark 中的非结构化流数据处理设计了一种基于属性的访问控制技术，可以灵活地基于数据的各种属性进行访问控制。由于 Spark SQL (可以执行 Hive 查询)引擎缺乏访问控制功能，不能满足 Spark SQL 用户间在列级/行级数据上的隔离，HDP [26]为 Spark SQL 引入了数据保护粒度为列级/行级的访问控制，并且可以实施 Hive 中的访问控制策略。Databricks [27]进一步为 Spark SQL 引入数据保护粒度为单元级的访问控制。以上访问控制技术均为静态的访问控制，即，它们通过直接实施访问控制策略阻止敏感数据进入数据处理逻辑。

在实际应用中，数据处理中存在着各种动态属性，比如用户的数据处理逻辑会随着应用场景的改变而改变，数据的敏感度随着时间的推移可能会改变，处理数据的用户可信度可能会改变。因此，除了上面介绍的静态访问控制技术，大数据计算平台也需要动态的访问控制技术。为了统一地支持 Spark 上的结构化数据分析引擎，GuardSpark++ [3]为 Spark 引入基于数据处理目的的访问控制机制；GuardSpark++基于启发式规则动态分析用户处理数据的目的，根据用户处理数据的目的可以保证用户不能直接看到受保护的敏感数据，比如：如果用户使用敏感数据的目的是统计分析则允许使用敏感数据，如果是直接输出敏感数据，访问控制机制阻止用户看到敏感数据。GuardSpark++中最细的数据保护粒度为单元级，并且可以根据用户使用数据的地点等上下文信息授权数据处理目的。Kumar 等[28]为 Hadoop 提出了一种基于数据敏感度的访问控制方案，但是该方案需要数据拥有者介入。Idar 等[29]，Belhadaoui 等[30]为 Hadoop 提出了基于动态数据敏感度的访问控制方案(D2SAC)，他们的方案不需要数据拥有者介入。TDACS [31]将基于属性的访问控制与区块链结合实现动态地估计数据用户的可信度，并根据动态的用户可信度实施控制。

3.3. 访问控制技术——统一

为了统一地为多种数据源提供访问控制服务，Apache 社区产生了 Apache Sentry 以及 Apache Ranger [20]两大中间件，它们可以统一地为 Hadoop 生态中的 HDFS、Hive 等数据源提供访问控制服务。但是不同的数据源需要不同的插件，这影响了这两种中间件在实际中的应用，现阶段它们支持的数据源个数都很有限。

为了统一地管理数据源以及数据处理方面的访问控制，近些年 Hadoop 生态圈中产生了新型的访问控制技术。HeAC [32]将 Apache Sentry 与 Apache Ranger 中的访问控制模型，以及 Hadoop 中原生访问控制特性形式化(基于客体属性(tags)为各种数据源系统中的各种数据客体授权访问)。OT-RBAC [33]是一种定制的基于角色的访问控制模型，与 HeAC 中的客体 tags 整合。在 HeAC 与 OT-RBAC 的基础之上，HeABAC [34]依赖于基于属性的访问控制支持多租户下的数据访问隔离。HBD-Authority [35]通过基于属性的访问控制中的 PIP、PDP、PEP 等组件进一步定制访问控制模型以加强 Hadoop 生态圈中访问控制技术的鲁棒性。Anisetti 等[36]通过基于属性的访问控制为数据的整个生命周期实施访问控制，可以统一地实现数据

源以及数据处理方面的访问控制。

3.4. 访问控制技术——数据处理测量

站在数据处理逻辑的角度，以上介绍的访问控制机制具有以下共同点：它们都通过干预数据处理逻辑实施访问控制。但是存在两种干预机制：① 在数据处理逻辑从数据源读取数据之后直接过滤敏感数据，② 在数据处理逻辑中插入过滤器过滤受保护数据。

这两种干预方式各有优缺点。在“允许用户使用敏感数据，但是敏感数据不能直接泄露给用户”的场景下，机制①不能适用，但是这种机制可以有效阻止数据处理测量工具从数据处理逻辑执行过程中抓取敏感数据。机制②可以控制数据处理逻辑的输出(比如，如果数据处理逻辑可能直接输出敏感数据，则插入敏感数据过滤器阻止直接输出敏感数据)。在保证敏感数据不直接泄露给用户的前提下，机制②允许用户尽可能地使用数据(包括敏感数据)，但是它不能有效阻止数据处理测量工具抓取敏感数据。学术界还没有访问控制技术用于确保用户进行安全的数据处理测量(即，基于访问控制技术阻止用户抓取到敏感数据)。下面，本文给出一种解决方案。

数据处理测量工具所做的数据操作从本质上讲包括：1) 将数据处理逻辑中的目标数据抓取到测量工具中，2) 处理抓取到的数据，3) 信息输出，完成测量。为了防止敏感数据通过数据处理测量工具泄露出去，我们不应该直接通过干预数据处理逻辑的方式实施访问控制，因为如果我们直接干预数据处理逻辑来控制数据测量，那么数据处理逻辑的数据使用/输出结果会因此而受到影响。因此，我们的设计方案如图2所示。BanHunt方案将数据分析任务分为两个模块：1) 数据处理模块(DPM)，2) 数据测量模块(DMM)。BanHunt 中有一个核心的访问监控器(用于实施访问控制模型)，该访问监控器根据 DPM 中的数据处理逻辑以及 DMM 中的数据操作决定是否过滤抓取到的敏感数据。该访问监控器不会影响 DPM 中的数据处理逻辑。

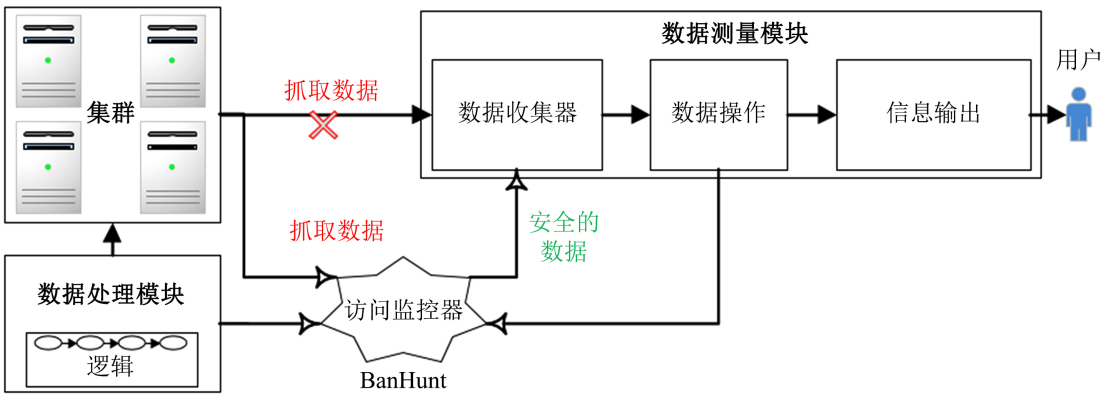

Figure 2. Secure data processing measurement based on access control
图 2. 基于访问控制实现安全数据处理测量

4. 结束语

本文从数据源、数据处理、数据处理测量的角度综述并讨论了现有的大数据访问控制技术。现阶段学术界缺乏对安全数据处理测量的研究，本文给出了基于访问控制技术实现安全数据处理测量的解决方案。通过梳理现有访问控制技术，我们可以发现不同的应用场景会需要不同的数据模型，也就从根本上影响到访问控制保护数据的能力，不同的访问控制需求也会产生不同的访问控制机制；在系统应用中，细粒度访问控制以及基于属性的访问控制越来越受重视。

最后，我们从以下几方面对大数据访问控制领域进行展望。

统一的访问控制：1) 为了促进组合不同的数据处理方式，大数据计算平台向流批一体方向发展，因此需要有对应的访问控制机制统一地为流计算、批处理实施访问控制，此外也需要专用的访问控制机制(比如专用的面向流数据处理的访问控制机制)。2) 在现实大数据处理中，不同的大数据计算平台可能会被组合使用，这种场景需要统一地为各种大数据计算平台实施访问控制，比如怎样将 Apache Spark，Apache Hadoop，Apache Flink 平台上的数据处理逻辑统一地交给访问控制机制施加控制。

动态的访问控制：在大数据时代，数据应用场景、数据用户属性、数据拥有者的属性、数据本身的属性等都很容易变化，因此访问控制机制也要有足够的能力去适应这些变化。动态的访问控制技术可以用于应对这些变化，但是现阶段用于大数据的动态访问控制技术研究还处于起步状态，需要学术界进一步研究。

区块链与访问控制：在安全领域，区块链可以以不同的方式使用，从根本上加强大数据的安全性和私密性[37] [38]，比如，Khalil 等[39]提出一种利用区块链技术的安全数据存储框架：通过管理大数据的元数据和策略，避免外部人员维护数据安全和隐私，提供更安全的大数据存储。访问控制领域也可以与区块链结合。在数据共享场景下，为增强数据的安全性，Ding 等[40]提出了一种基于区块链的多维用户授权和基于角色的访问控制机制。BEAAS [41]基于以太坊区块链能够验证访问控制数据以及云服务做出的访问决策是否在授权的状态下完成。通过访问控制技术与区块链技术的组合加强大数据安全的研究还处于起步状态，需要学术界进一步研究。

参考文献

[1] 维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 大数据时代: 生活, 工作与思维的大变革[M]. 杭州: 浙江人民出版社, 2013.

[2] Nazari, E., Shahriari, M.H. and Tabesh, H. (2019) BigData Analysis in Healthcare: Apache Hadoop, Apache Spark and Apache Flink. *Frontiers in Health Informatics*, **8**, e14. https://doi.org/10.30699/fhi.v8i1.180

[3] Xue, T., Wen, Y., Luo, B., Zhang, B.Y., Zheng, Y., Hu, Y.F., Li, Y.J. and Li, G. (2020) GuardSpark++: Fine-Grained Purpose-Aware Access Control for Secure Data Sharing and Analysis in Spark. *ACSAC*'20: *Annual Computer Security Applications Conference*, Austin, 7-11 December 2020, 582-596. https://doi.org/10.1145/3427228.3427640

[4] Colombo, P. and Ferrari, E. (2019) Access Control Technologies for Big Data Management Systems: Literature Review and Future Trends. *Cybersecurity*, **2**, Article No. 3. https://doi.org/10.1186/s42400-018-0020-9

[5] Kleppmann, M. (2017) Designing Data-Intensive Applications: The Big Ideas behind Reliable, Scalable, and Maintainable Systems. O'Reilly Media, Inc., Sebastopol.

[6] Chambers, B. and Zaharia, M. (2017) Spark: The Definitive Guide: Big Data Processing Made Simple. O'Reilly Media, Inc., Sebastopol.

[7] Goel, K. and Hofstede, A.H.M.T. (2021) Privacy-Breaching Patterns in NoSQL Databases. *IEEE Access*, **9**, 35229-35239. https://doi.org/10.1109/ACCESS.2021.3062034

[8] Gupta, E., *et al*. (2021) Attribute-Based Access Control for NoSQL Databases. *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, Virtual Event, 26-28 April 2021, 317-319. https://doi.org/10.1145/3422337.3450323

[9] Alotaibi, A.A., Alotaibi, R.M. and Hamza, N. (2019) Access Control Models in NoSQL Databases: An Overview. *Journal of King Abdulaziz University*: *Computing and Information Technology Sciences*, **8**, 1-9. https://doi.org/10.4197/Comp.8-1.1

[10] Sun, Z.H. and Huo, Y.X. (2021) The Spectrum of Big Data Analytics. *Journal of Computer Information Systems*, **61**, 154-162. https://doi.org/10.1080/08874417.2019.1571456

[11] 王珊, 萨师煊. 数据库系统概论(第 4 版) [M]. 北京: 高等教育出版社, 2006.

[12] Elmasri, R. and Navathe, S.B. (2021) Fundamentals of Database System. Global Edition, Pearson, London.

[13] Balusamy, B., Abirami, R.N., Kadry, S. and Gandomi, A.H. (2021) NoSQL Database. In: *Big Data*: *Concepts*, *Technology*, *and Architecture*, John Wiley & Sons, Hoboken, 53-81. https://doi.org/10.1002/9781119701859.ch3

[14] Kulkarni, D. (2013) A Fine-Grained Access Control Model for Key-Value Systems. *Proceedings of the Third ACM Conference on Data and Application Security and Privacy*, San Antonio, 18-20 February 2013, 161-164. https://doi.org/10.1145/2435349.2435370

[15] Shalabi, Y. and Gudes, E. (2017) Cryptographically Enforced Role-Based Access Control for NoSQL Distributed Databases. In: Livraga, G. and Zhu, S., Eds., *Data and Applications Security and Privacy XXXI*, Springer, Cham, 3-19. https://doi.org/10.1007/978-3-319-61176-1_1

[16] Huang, L., Zhu, Y., Wang, X., *et al*. (2019) An Attribute-Based Fine-Grained Access Control Mechanism for HBase. In: Hartmann, S., *et al*., Eds., *Database and Expert Systems Applications*, Springer, Cham, 44-59. https://doi.org/10.1007/978-3-030-27615-7_4

[17] Colombo, P. and Ferrari, E. (2015) Enhancing MongoDB with Purpose-Based Access Control. *IEEE Transactions on Dependable & Secure Computing*, **14**, 591-604.

[18] Colombo, P. and Ferrari, E. (2016) Towards Virtual Private NoSQL Datastores. 2016 *IEEE* 32*nd International Conference on Data Engineering* (*ICDE*), Helsinki, 16-20 May 2016, 193-204. https://doi.org/10.1109/ICDE.2016.7498240

[19] Colombo, P. and Ferrari, E. (2017) Towards a Unifying Attribute Based Access Control Approach for NoSQL Datastores. *IEEE International Conference on Data Engineering*, San Diego, 19-22 April 2017, 709-720. https://doi.org/10.1109/ICDE.2017.123

[20] Begum, G., Huq, S.Z.U. and Siva Kumar, A.P. (2021) Security Features in Hadoop—A Survey. In: Saini, H.S., Sayal, R., Govardhan, A. and Buyya, R., Eds., *Innovations in Computer Science and Engineering*, Springer, Singapore, 269-276. https://doi.org/10.1007/978-981-33-4543-0_29

[21] Gopalakrishnan, A.A., *et al*. (2021) HACS: Access Control for Streaming Data across Heterogeneous Communication Models. 2021 *IEEE World AI IoT Congress* (*AIIoT*), Seattle, 10-13 May 2021, 109-114. https://doi.org/10.1109/AIIoT52608.2021.9454185

[22] Awaysheh, F.M., *et al*. (2020) Next-Generation Big Data Federation Access Control: A Reference Model. *Future Generation Computer Systems*, **108**, 726-741. https://doi.org/10.1016/j.future.2020.02.052

[23] Odugu, N.K. (2021) A Fine-Grained Access Control Survey for the Secure Big Data Access. *Turkish Journal of Computer and Mathematics Education* (*TURCOMAT*), **12**, 4180-4186.

[24] Ulusoy, H., Colombo, P., Ferrari, E., Kantarcioglu, M. and Pattuk, E. (2015) GuardMR: Fine-Grained Security Policy Enforcement for MapReduce Systems. *Proceedings of the* 10*th ACM Symposium on Information, Computer and Communications Security*, Singapore, 14 April-17 March 2015, 285-296. https://doi.org/10.1145/2714576.2714624

[25] Preuveneers, D. and Joosen, W. (2015) SparkXS: Efficient Access Control for Intelligent and Large-Scale Streaming Data Applications. 2015 *International Conference on Intelligent Environments*, Prague, 15-17 July 2015, 96-103. https://doi.org/10.1109/IE.2015.21

[26] Introducing Row/Column Level Access Control for Apache Spark. https://blog.cloudera.com/row-column-level-control-apache-spark/

[27] Databricks (2021) Data Governance Model. https://docs.databricks.com/security/access-control/table-acls/object-privileges.html

[28] Ashwin Kumar, T.K., *et al*. (2017) Content Sensitivity Based Access Control Framework for Hadoop. *Digital Communications and Networks*, **3**, 213-225. https://doi.org/10.1016/j.dcan.2017.07.007

[29] Idar, H.A., *et al*. (2018) Dynamic Data Sensitivity Access Control in Hadoop Platform. 2018 *IEEE* 5*th International Congress on Information Science and Technology*, Marrakech, 21-27 October 2018, 105-109. https://doi.org/10.1109/CIST.2018.8596381

[30] Ait Idar, H., Belhadaoui, H. and Filali, R. (2021) A Conceptual Model for Dynamic Access Control in Hadoop Ecosystem. In: Saeed F., *et al*., Eds., *Advances on Smart and Soft Computing*, Springer, Singapore, 421-430. https://doi.org/10.1007/978-981-15-6048-4_37

[31] Yang, M. (2020) TDACS: An ABAC and Trust-Based Dynamic Access Control Scheme in Hadoop. https://arxiv.org/abs/2011.07895

[32] Gupta, M., Patwa, F. and Sandhu, R. (2017) POSTER: Access Control Model for the Hadoop Ecosystem. *Proceedings of the* 22*nd ACM on Symposium on Access Control Models and Technologies*, Indianapolis, 21-23 June 2017, 125-127. https://doi.org/10.1145/3078861.3084164

[33] Gupta, M., Patwa, F. and Sandhu, R. (2017) Object-Tagged RBAC Model for the Hadoop Ecosystem. In: Livraga, G. and Zhu, S., Eds., *Data and Applications Security and Privacy XXXI*, Springer, Cham, 63-81. https://doi.org/10.1007/978-3-319-61176-1_4

[34] Gupta, M., Patwa, F. and Sandhu, R. (2018) An Attribute-Based Access Control Model for Secure Big Data Processing

in Hadoop Ecosystem. *Proceedings of the Third ACM Workshop on Attribute-Based Access Control*, Tempe, 21 March 2018, 13-24. https://doi.org/10.1145/3180457.3180463

[35] Chen, C.W., Elsayed, M.A. and Zulkernine, M. (2020) HBD-Authority: Streaming Access Control Model for Hadoop. 2020 *IEEE 6th International Conference on Dependability in Sensor*, *Cloud and Big Data Systems and Application*, Nadi, 14-16 December 2020, 16-25. https://doi.org/10.1109/DependSys51298.2020.00012

[36] Anisetti, M., *et al*. (2021) Dynamic and Scalable Enforcement of Access Control Policies for Big Data. *Proceedings of the 13th International Conference on Management of Digital Ecosystems*, Virtual Event, 1-3 November 2021, 71-78. https://doi.org/10.1145/3444757.3485107

[37] Alsulbi, K., Khemakhem, M., Basuhail, A., *et al*. (2021) Big Data Security and Privacy: A Taxonomy with Some HPC and Blockchain Perspectives. *International Journal of Computer Science and Network Security*, **21**, 43-55.

[38] Deepa, N., *et al*. (2020) A Survey on Blockchain for Big Data: Approaches, Opportunities, and Future Directions. https://arxiv.org/abs/2009.00858

[39] Alsulbi, K.A., *et al*. (2021) A Proposed Framework for Secure Data Storage in a Big Data Environment Based on Blockchain and Mobile Agent. *Symmetry*, **13**, Article No. 1990. https://doi.org/10.3390/sym13111990

[40] Ding, Y., *et al*. (2020) Blockchain-Based Access Control Mechanism of Federated Data Sharing System. *IEEE International Conference on Parallel & Distributed Processing with Applications*, *Big Data & Cloud Computing*, *Sustainable Computing & Communications*, *Social Computing & Networking*, Exeter, 17-19 December 2020, 277-284. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00060

[41] Kumar, R., Palanisamy, B. and Sural, S. (2021) BEAAS: Blockchain Enabled Attribute-Based Access Control as a Service. 2021 *IEEE International Conference on Blockchain and Cryptocurrency*, Sydney, 3-6 May 2021, 1-3. https://doi.org/10.1109/ICBC51069.2021.9461151