

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2012년 6월 28일 (28.06.2012)



(10) 국제공개번호
WO 2012/086920 A2

- (51) 국제특허분류:
G06F 15/16 (2006.01) H04L 9/32 (2006.01)
- (21) 국제출원번호: PCT/KR2011/008224
- (22) 국제출원일: 2011년 10월 31일 (31.10.2011)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보:
10-2010-0134842 2010년 12월 24일 (24.12.2010) KR
- (71) 출원인 (US 을(를) 제외한 모든 지정국에 대하여): (주) 케이티 (KT CORPORATION) [KR/KR]; 경기도 성남시 분당구 정자동 206, 463-711 Gyeonggi-do (KR).
- (72) 발명자: 겸
- (75) 발명자/출원인 (US 에 한하여): 김미점 (KIM, Mi-Jeom) [KR/KR]; 경기도 안양시 동안구 관양동 1608-2, 현대 아이스페이스 606 호, 431-060 Gyeonggi-do (KR). 김효민 (KIM, Hyo-Min) [KR/KR]; 서울시 서초구 우면동 40-1 103 호, 137-900 Seoul (KR). 이어형 (LEE, Eo-Hyung) [KR/KR]; 경기도 용인시 수지구 죽전 1 동 현대홈타운 3 차 1 단지아파트 723 동 102 호, 448-160 Gyeonggi-do (KR). 황진경 (HWANG, Jin-Kyung) [KR/KR]; 경기도 과천시 중앙동 삼성테미안 APT 1108 동 705 호, 427-010 Gyeonggi-do (KR).

(74) 대리인: 리앤목특허법인 (Y.P.LEE, MOCK & PARTNERS); 서울시 서초구 서초동 1575-1, 137-875 Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

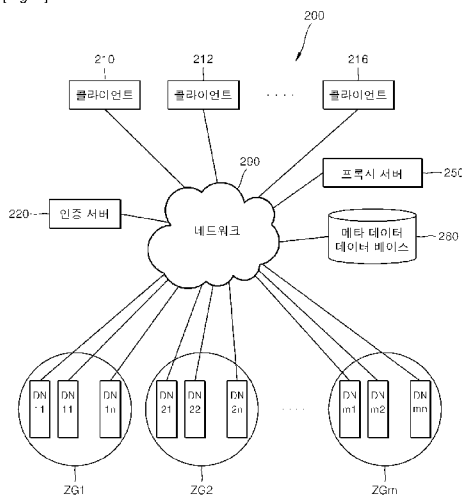
(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[다음 쪽 계속]

(54) Title: DISTRIBUTED STORAGE SYSTEM HAVING CONTENT-BASED OVERLAP PREVENTION FUNCTION, METHOD FOR STORING OBJECT THEREOF, AND STORAGE MEDIUM READABLE BY COMPUTER

(54) 발명의 명칭: 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템 및 그 오브젝트 저장 방법 및 컴퓨터에 의하여 독출가능한 저장 매체

[Fig. 2]



210, 212, 216 ... Client
220 ... Authentication server
250 ... Proxy server
280 ... Metadata database
290 ... Network

(57) Abstract: Disclosed are a distributed storage system having a content-based overlap prevention function and a method for storing an object. The distributed storage system comprises: an authentication server for authenticating a client; a plurality of data nodes, each of the data nodes storing at least one object; a metadata database for storing metadata, which includes the unique information of the object and the unique information of the data node that stores the object; and a proxy server for replying to a request for storing the object by the authenticated client to store a target object, and for referencing the metadata to provide to the client the list of the unique information of the target data node to which the target object is to be stored. The proxy server decides a content-specific index, which is decided by the content of the target object, and judges whether the target object overlaps with pre-stored objects by using the content-specific index that is decided. According to the present invention, an overlap prevention action can be performed efficiently.

(57) 요약서: 중복 방지 기능을 가지는 분산 저장 시스템 및 오브젝트 저장 방법이 개시된다. 분산 저장 시스템은 클라이언트를 인증하기 위한 인증 서버, 각각 적어도 하나의 오브젝트들을 저장하는 복수 개의 데이터 노드들, 오브젝트의 고유 정보 및 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 저장하는 메타데이터 데이터베이스,

[다음 쪽 계속]

**공개:**

- 국제조사보고서 없이 공개하며 보고서 접수 후 이를 별도 공개함 (규칙 48.2(g))

및 대상 오브젝트를 저장하고자 하는 인증된 클라이언트의 오브젝트 저장 요청에 응답하여, 메타데이터를 참조하여 대상 오브젝트가 저장될 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공하는 프락시 서버를 포함한다. 프락시 서버는 클라이언트로부터 저장 요청이 수신되면, 대상 오브젝트의 내용에 의하여 결정되는 내용 특이적 인덱스를 결정하고, 결정된 내용 특이적 인덱스를 이용하여 대상 오브젝트가 기저장된 오브젝트들과 중복되는지 여부를 판단한다. 본 발명에 의하면 중복 방지 동작이 효율적으로 수행될 수 있다.

명세서

발명의 명칭: 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템 및 그 오브젝트 저장 방법 및 컴퓨터에 의하여 독출가능한 저장 매체

기술분야

- [1] 본 발명은 오브젝트 스토리지 시스템에서 오브젝트의 중복 제거 또는 중복 방지(deduplication) 동작을 효율적으로 하기 위한 내용 기반(content_based) 오브젝트 저장 기술에 관한 것으로서, 특히, 중복 방지된 오브젝트들을 지역별로 그룹핑된 데이터 노드들과 클라이언트의 위치 정보에 기반하여 선택된 대상 데이터 노드에 저장함으로써, 오브젝트를 신뢰성있게 저장할 수 있는 분산 저장 시스템에 관한 것이다.

배경기술

- [2] 클라우드 컴퓨팅은 인터넷 망을 통해 여러 가지 IT(Information Technology) 자원들을 분배하여 서비스하는 개념으로 가장 일반적인 서비스 분류로는 첫째, 하드웨어 인프라를 서비스로 제공하는 IaaS(Infrastructure As A Service), 둘째, 응용 개발 및 실행 플랫폼을 서비스로 제공하는 PaaS(Platform As A Service), 그리고 마지막으로 어플리케이션을 서비스로 제공하는 SaaS(Software As A Service)로 나눈다.
- [3] IaaS에는 많은 서비스 카테고리가 있는데, 대표적으로 연산 자원을 가상 머신 형태로 제공하는 연산 서비스와 저장 서비스가 있다. 이러한 분산 저장 시스템은 클라우드 스토리지 서비스를 제공하는데, 이를 이용하면 저사양의 하드웨어를 활용하여 공용의 저장소 풀(storage pool)을 만들어 탄력적이고 유연한 사용량을 적시에 만족하는 특징을 가진다. 이를 위해 널리 이용되는 단순하고 강력한 오브젝트 기반 스토리지 기법에서는 물리적인 저장 공간 관리 기능을 저장 장치 자체에서 직접 수행하게 한다. 그러므로, 저장 장치의 성능이 향상되고 손쉽게 저장 장치의 용량을 확장할 수 있다. 또한, 플랫폼과 독립적으로 데이터를 안전하게 공유할 수 있는 특징을 가진다.
- [4] 도 1은 종래 기술에 의한 분산 저장 시스템을 개념적으로 나타내는 도면이다.
- [5] 도 1에 도시되는 오브젝트 스토리지 시스템은 클라이언트의 인증을 처리하는 인증 서버, 클라이언트의 요구사항을 처리하는 프락시 서버(proxy server)(또는 마스터 서버), 오브젝트들의 물리적인 위치를 포함하는 메타데이터를 저장하는 메타데이터 데이터베이스, 실제 오브젝트 저장 및 관리를 담당하는 데이터 노드, 그리고 데이터의 복제 등을 관리하는 복제 서버(replicator server) 등을 포함한다. 클라이언트는 초기에 인증 서버를 통해 인증을 받으며, 인증이 완료된 후에는 프락시 서버에게 원하는 오브젝트를 관리하는 데이터 노드의 정보를 요청한다. 클라이언트의 요청에 대해 프락시 서버는 메타데이터를 참조하여 해당하는

데이터 노드에게 원하는 동작 요청을 전달하고, 데이터 노드는 동작을 수행한 결과를 프락시 서버를 통해 클라이언트에게 전달한다. 또는, 데이터 노드는 프락시 서버를 거치지 않고 직접 클라이언트에게 응답을 제공할 수도 있다. 이 경우 지연이나 데이터 트래픽이 감소하는 효과를 기대할 수 있으나 모든 데이터 노드가 클라이언트 인터페이스를 가져야 함으로 데이터 노드의 복잡도가 증가할 수 있다.

- [6] 오브젝트 저장소는 데이터의 안전성과 높은 가용성을 위해 데이터를 복제(replication)하며 이런 복제본을 레플리카(replica)라고 한다. 널리 이용되는 분산 저장 시스템은 일반적으로 2개 내지 3개의 복제본을 가지지만, 오브젝트의 중요도에 따라 더 많은 복제본을 가지기도 한다. 오브젝트의 복제본은 서로 동기를 유지해야 하며 이는 보통 별도의 복제 서버(replicator server)에서 처리한다.
- [7] 데이터 복제와 상반되는 개념으로 같은 내용의 여러 개의 오브젝트를 중복하여 저장하는 요청이 있을 때 하나의 오브젝트만 저장하는 것이 중복 제거 또는 중복 방지(deduplication) 기술이다. 예를 들어, 최신의 인기 영화 파일은 많은 사람들이 오브젝트 스토리지에 저장하고자 할 수 있다. 이런 경우 하나의 오브젝트만 유지하고(물론 복제본은 존재) 그 뒤에 같은 내용의 오브젝트를 업로드하는 요구가 발생할 때(다른 클라이언트에 의해서라도) 오브젝트에 대한 위치 정보를 저장한 메타데이터만 별도로 유지하고 같은 내용의 오브젝트 자체는 다시 저장하지 않음으로 해서 경제성을 향상시킨다.
- [8] 그런데, 종래 기술에 의한 중복 방지 기법은 오브젝트의 논리적인 이름에 기반하여 모든 데이터 노드에 대하여 동일한 논리적인 이름이 존재하는지를 확인한다. 그러므로, 종래 기술에 의한 물리적 위치 매핑 방법에 따르면 중복 방지를 위해 기존의 모든 오브젝트를 다 검사해야 하기 때문에, 너무 많은 부하를 요구한다.
- [9] 그러므로, 중복 방지 기법을 효율적으로 지원하기 위한 오브젝트의 효율적 분산 저장 방법이 절실히 요구된다. 또한, 이러한 중복 방지 기법 기법을 구현하는데 필요한 메타데이터의 구조를 제공하는 것이 절실히 요구된다.

발명의 상세한 설명

기술적 과제

- [10] 본 발명의 목적은, 클라우드 스토리지 서비스를 위한 오브젝트 스토리지 시스템에서 중복 제거를 위한 내용 기반의 오브젝트 저장 기법을 제공하는 것이다.
- [11] 또한, 본 발명의 목적은 오브젝트 중복 방지 동작을 효율적으로 수행할 수 있는 메타데이터의 구조를 제공하는 것이다.

과제 해결 수단

- [12] 상기와 같은 목적들을 달성하기 위한 본 발명의 일면은, 복수 개의

클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템(distribution storage system)에 관한 것이다. 본 발명의 일면에 의한 분산 저장 시스템은 클라이언트를 인증하기 위한 인증 서버, 각각 적어도 하나의 오브젝트들을 저장하는 복수 개의 데이터 노드들, 오브젝트의 고유 정보 및 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 저장하는 메타데이터 데이터베이스, 및 대상 오브젝트(target object)를 저장하고자 하는 인증된 클라이언트의 오브젝트 저장 요청에 응답하여, 메타데이터를 참조하여 대상 오브젝트가 저장될 대상 데이터 노드(target data node)의 고유 정보의 목록을 클라이언트에게 제공하는 프락시 서버를 포함하며, 프락시 서버는 클라이언트로부터 저장 요청이 수신되면, 대상 오브젝트의 내용에 의하여 결정되는 내용 특이적 인덱스(content-specific index)를 결정하고, 결정된 내용 특이적 인덱스를 이용하여 대상 오브젝트가 기저장된 오브젝트들과 중복되는지 여부를 판단하며, 중복되지 않은 대상 오브젝트에 대해서만 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공하도록 구성되고, 클라이언트는 제공된 대상 데이터 노드의 고유 정보의 목록을 이용하여 대상 오브젝트를 저장하도록 구성된다. 특히, 프락시 서버는 대상 오브젝트의 소정 부분에 소정 해쉬 함수를 적용한 결과를 내용 특이적 인덱스로서 결정하도록 구성된다. 더 나아가, 프락시 서버는 대상 오브젝트의 최초 소정 길이를 입력으로써 이용하는 MD5, SHA1, SHA256, SHA384, RMD128, RMD160, RMD256, RMD320, HAS160, 및 TIGER 해쉬 함수 중 어느 하나를 이용하여 내용 특이적 인덱스를 결정하도록 구성된다. 특히, 본 발명에 따르면, 메타데이터는 사용자 ID, 디렉토리 ID, 오브젝트 ID, 및 내용 특이적 인덱스 중 적어도 하나를 포함하는 오브젝트 테이블 및 내용 특이적 인덱스 및 오브젝트의 복제본이 저장된 데이터 노드의 ID를 포함하는 복제본 위치 테이블을 포함한다. 더 나아가, 데이터 노드들은 지역(zone)별로 그룹핑되며, 프락시 서버는 동일한 오브젝트는 동일한 지역 그룹(zone group)에 속한 데이터 노드들 중 오직 하나에만 저장되도록 대상 데이터 노드의 고유 정보의 목록을 결정하도록 구성된다. 특히, 본 발명에 의한 분산 저장 시스템은 데이터 노드들 및 클라이언트의 위치 관계에 기반하여 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하고, 선택된 지역 그룹 및 클라이언트 간의 거리에 기반하여 지역 그룹별 우선 순위를 결정하는 위치 인식 서버(location-aware server)를 더 포함하고, 프락시 서버는 위치 인식 서버가 선택한 지역 그룹 당 하나의 대상 데이터 노드를 결정하고, 결정된 대상 데이터 노드들의 목록을 이용하여 메타데이터 데이터베이스를 갱신하며, 대상 데이터 노드들의 목록 및 지역 그룹별 우선 순위를 클라이언트에게 전송하도록 구성되고, 클라이언트는 가장 높은 지역 그룹별 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터 노드에 대상 오브젝트를 저장함으로써, 우선 순위에 따라 순차적으로 더 낮은 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터

노드들에 대상 오브젝트가 복제되는 복제 동작이 수행되도록 야기하도록 더욱 구성된다. 뿐만 아니라, 프락시 서버는 동일한 지역 그룹에 포함되는 데이터 노드들의 가용 저장 용량 및 오브젝트 저장 내역을 고려하여 동일한 지역 그룹에 포함되는 데이터 노드들에게 우선 순위를 부여하고, 가장 높은 우선 순위를 가지는 데이터 노드를 대상 데이터 노드로서 결정하도록 더욱 구성된다. 본 발명에 의하면, 오브젝트의 고유 정보는 오브젝트의 ID, 크기, 데이터 타입, 및 작성자 중 적어도 하나를 포함하고, 데이터 노드의 고유 정보는 데이터 노드의 ID, IP(Internet Protocol) 주소, 및 물리적 위치 중 적어도 하나를 포함한다. 특히, 메타데이터는 데이터 노드들의 사용량, 각 지역 그룹에 속한 데이터 노드들의 목록, 대상 오브젝트에 대한 지역 그룹별 우선 순위, 및 동일한 지역 그룹에 속한 데이터 노드들 간의 우선 순위 중 적어도 하나를 더 포함한다.

- [13] 상기와 같은 목적들을 달성하기 위한 본 발명의 다른 면은 복수 개의 클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템에 오브젝트를 분산 저장하는 방법에 관한 것이다. 분산 저장 방법은, 클라이언트를 인증하는 단계, 프락시 서버가 대상 오브젝트를 저장하고자 하는 인증된 클라이언트의 오브젝트 저장 요청을 수신하는 단계, 프락시 서버가 대상 오브젝트의 내용에 의하여 결정되는 내용 특이적 인덱스를 결정하는 내용 특이적 인덱스 결정 단계, 결정된 내용 특이적 인덱스를 이용하여 프락시 서버가 대상 오브젝트가 기저장된 오브젝트들과 중복되는지 여부를 판단하는 단계, 및 프락시 서버가, 중복되지 않은 대상 오브젝트에 대해서만 오브젝트의 고유 정보 및 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 참조하여 대상 오브젝트가 저장될 대상 데이터 노드(target data node)를 결정하는 대상 데이터 노드 결정 단계, 프락시 서버가, 결정된 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공하는 단계, 및 클라이언트가 대상 오브젝트를 목록에 포함된 대상 데이터 노드에 저장하는 단계를 포함한다. 더 나아가, 내용 특이적 인덱스 결정 단계는 프락시 서버가 대상 오브젝트의 소정 부분에 소정 해쉬 함수를 적용한 결과를 내용 특이적 인덱스로서 결정하는 단계를 포함한다. 특히, 내용 특이적 인덱스 결정 단계는 프락시 서버가 대상 오브젝트의 최초 소정 길이를 MD5, SHA1, SHA256, SHA384, RMD128, RMD160, RMD256, RMD320, HAS160, 및 TIGER 해쉬 함수 중 어느 하나에 적용하여 내용 특이적 인덱스를 결정하는 단계를 포함한다. 또한, 메타데이터는 사용자 ID, 디렉토리 ID, 오브젝트 ID, 및 내용 특이적 인덱스 중 적어도 하나를 포함하는 오브젝트 테이블 및 내용 특이적 인덱스 및 오브젝트의 복제본이 저장된 데이터 노드의 ID를 포함하는 복제본 위치 테이블을 포함한다. 더 나아가, 대상 데이터 노드 결정 단계는 프락시 서버가, 동일한 오브젝트가 동일한 지역 그룹(zone group)에 속한 데이터 노드들 중 오직 하나에만 저장되도록 대상 데이터 노드의 고유 정보의 목록을 결정하는 단계를 포함한다. 바람직하게는, 대상 데이터 노드 결정 단계는 위치 인식

서버가 데이터 노드들 및 클라이언트의 위치 관계에 기반하여 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하고, 선택된 지역 그룹 및 클라이언트 간의 거리에 기반하여 지역 그룹별 우선 순위를 결정하는 단계, 및 프락시 서버가 위치 인식 서버가 선택한 지역 그룹 당 하나의 대상 데이터 노드를 결정하는 단계를 포함한다. 또는, 대상 데이터 노드 결정 단계는 프락시 서버가, 동일한 지역 그룹에 포함되는 데이터 노드들의 가용 저장 용량 및 오브젝트 저장 내역을 고려하여 동일한 지역 그룹에 포함되는 데이터 노드들에게 우선 순위를 부여하는 단계, 및 프락시 서버가, 가장 높은 우선 순위를 가지는 데이터 노드를 대상 데이터 노드로서 결정하는 단계를 포함한다.

발명의 효과

- [14] 본 발명에 의하여, 본 발명은 클라우드 스토리지 서비스에서 요구하는 복제와 중복 방지 기능을 동시에 효율적으로 지원할 수 있다.
- [15] 또한, 본 발명에 의하면 중복 방지 동작을 수행할 때 오브젝트 내용의 일부를 입력으로 한 해수 함수의 결과값이 동일한 오브젝트에 대해서만 중복 체크를 함으로 인해 시간과 오버헤드를 현격하게 줄일 수 있다.
- [16] 더 나아가, 본 발명에 의하면 데이터 노드들을 지역별로 그룹핑하고, 복제본들이 상이한 지역에 분산되어 저장되도록 하기 때문에, 하나의 지역에 네트워크 문제가 발생했을 때에도 다른 지역에 저장된 복제본을 독출할 수 있어서 더욱 신뢰성 있는 서비스가 가능하게 된다.

도면의 간단한 설명

- [17] 도 1은 종래 기술에 의한 분산 저장 시스템을 개념적으로 나타내는 도면이다.
- [18] 도 2는 본 발명의 일면에 의한 중복 방지 기능을 가지는 분산 저장 시스템의 일 실시예를 개념적으로 나타내는 도면이다.
- [19] 도 3은 본 발명의 다른 면에 의한 중복 방지 기능을 가지는 분산 저장 시스템의 오브젝트 저장 방법을 개념적으로 나타내는 흐름도이다.
- [20] 도 4는 본 발명에 적용될 수 있는 해시 함수의 특징을 설명하기 위한 표이다.
- [21] 도 5a 및 도 5b는 본 발명에 의한 분산 저장 시스템에서 이용하는 메타데이터에 포함되는 테이블들을 예시하는 도면이다.
- [22] 도 6은 본 발명의 일면에 의한 중복 방지 기능을 가지는 분산 저장 시스템의 다른 실시예를 개념적으로 나타내는 도면이다.

발명의 실시를 위한 최선의 형태

- [23] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.
- [24] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게

설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

- [25] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.
- [26] 도 2는 본 발명의 일면에 의한 중복 방지 기능을 가지는 분산 저장 시스템의 일 실시예를 개념적으로 나타내는 도면이다.
- [27] 도 2에 도시된 분산 저장 시스템(200)은 네트워크(290)에 연결되는 복수 개의 클라이언트들(210, 212, 216) 및 데이터 노드들(DN11-DN1n, DN21-DN2n, DNm1-DNmn)을 포함한다. 또한, 도 2에 도시된 분산 저장 시스템(200)은 인증 서버(220), 프락시 서버(250) 및 메타데이터 데이터베이스(280)를 더 포함한다.
- [28] 인증 서버(220)는 클라이언트를 인증하고, 데이터 노드들(DN11-DN1n, DN21-DN2n, DNm1-DNmn)은 각각 적어도 하나의 오브젝트들을 저장한다. 또한, 메타데이터 데이터베이스(280)는 오브젝트의 고유 정보 및 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 저장한다.
- [29] 클라이언트들(210, 212, 216) 중 제1 클라이언트(210)가 오브젝트를 데이터 노드들(DN11-DN1n, DN21-DN2n, DNm1-DNmn) 중 하나에 저장하려고 시도하는 경우에 대해서 설명한다. 우선, 인증된 클라이언트(210)는 대상 오브젝트(target object)를 저장하기 위하여 프락시 서버(250)에 클라이언트의 오브젝트 저장 요청을 송신한다. 본 발명에 의한 프락시 서버(250)는 동작 요청이 있을 때 모든 대상 오브젝트를 저장하는 것이 아니라, 대상 오브젝트가 이미 데이터 노드들(DN11-DN1n, DN21-DN2n, DNm1-DNmn) 중 하나에 기저장되었는지를 판단한다. 이러한 중복 방지 동작을 수행하기 위하여, 프락시 서버(250)는 우선 대상 오브젝트의 내용에 의하여 결정되는 내용 특이적 인덱스(content-specific index)를 결정하고, 결정된 내용 특이적 인덱스를 이용하여 대상 오브젝트가 기저장된 오브젝트들과 중복되는지 여부를 판단한다. 만일, 대상 오브젝트가 이미 데이터 노드 중 하나에 기저장되어 있다면, 프락시 서버(250)는 동작 요청을 무시한다. 따라서, 동일한 오브젝트가 불필요하게 많은 데이터 노드에 저장되어 시스템 자원을 낭비하는 것을 방지할 수 있다. 만일, 대상 오브젝트가 기저장된 오브젝트들과 상이하다면, 프락시 서버(250)는 이와 같은 중복되지 않은 대상 오브젝트에 대해서만 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공한다. 그러면, 클라이언트(210)는 제공된 대상 데이터 노드의 고유 정보의 목록을 참조하여 대상 데이터 노드를 식별한 후, 대상 데이터 노드의 IP 주소 등을 이용하여 대상 오브젝트를 해당 대상 데이터 노드에 저장한다.
- [30] 특히, 프락시 서버는 대상 오브젝트의 소정 부분(예를 들어 대상 오브젝트의

최초 65 메가바이트)에 해쉬 함수를 적용하고, 그 결과를 해당 대상 오브젝트에 대한 내용 특이적 인덱스로서 결정한다. 본 명세서에서 내용 특이적 인덱스는 중복되는 대상 오브젝트들을 쉽게 찾아내기 위하여 이용되는 모든 정보일 수 있다. 프락시 서버(250)가 이용하는 해쉬 함수에 대해서는 도 4를 이용하여 상세히 후술된다. 본 발명에 의한 분산 저장 시스템(200)에 포함되는 프락시 서버(250)는 내용 특이적 인덱스를 이용하여 오브젝트의 동일 여부를 판단하기 때문에, 대상 오브젝트와 동일한 오브젝트이지만 다른 사용자에게 의하여 다른 명칭이 부여된 오브젝트도 대상 오브젝트와 같은 오브젝트라고 쉽게 판단할 수 있다.

[31] 본 명세서에서 '대상 오브젝트(target object)'란 클라이언트가 저장하고자 하는 오브젝트나 데이터 노드로부터 조회하고자 하는 관심 대상인 오브젝트를 의미한다. 또한, '대상 데이터 노드(target data node)'란 여러 개의 데이터 노드 중에서 대상 오브젝트가 저장된 데이터 노드를 의미한다. 그리고, 본 명세서에서 '우선 순위'란 특정 대상 오브젝트를 저장하기에 어떤 지역 그룹 또는 데이터 노드가 다른 지역 그룹 또는 데이터 노드에 비하여 더 적합한지 판단하여 매긴 순위를 나타낸다. 우선 순위에는 특정 지역 그룹이 다른 지역 그룹에 비하여 가지는 우선 순위 및 동일한 지역 그룹 내에 속한 데이터 노드들 간의 우선 순위가 포함될 수 있다. 또한, 우선 순위는 클라이언트가 직접 어느 대상 오브젝트에 관련한 특정 지역 및 데이터 노드에 대한 선호도에 기반하여 매길 수도 있고, 또는 프락시 서버 또는 위치 인식 서버에 의하여 자동으로 결정될 수도 있다. 이에 대해서는 명세서의 해당 부분에서 상세히 후술된다.

[32] 또한, 다시 도 2를 참조하면, 데이터 노드들(DN11-DN1n, DN21-DN2n, DNm1-DNmn)은 제1 내지 제m 지역 그룹(ZG1, ZG2, ZGm) 중 어느 하나에 포함된다는 것을 알 수 있다. 도 2에 도시되는 지역 그룹(ZG1, ZG2, ZG3)들은 효과적인 복제본의 분산 저장을 위하여, 각각 지역적으로 인접한 데이터 노드들을 그룹핑함으로써 정의된다. 또한, 동일한 지역 그룹에 속하는 데이터 노드들은 동일한 오브젝트를 저장하지 않도록 구성된다. 즉, 하나의 오브젝트의 복제본은 다른 지역 그룹에 속하는 데이터 노드들에 분산되어 저장되기 때문에 두 개의 복제본이 어느 하나의 지역 그룹에 속한 두 개의 데이터 노드에 공통적으로 저장되지 않는다. 이를 메타데이터의 관점에서 보면, 오브젝트의 물리적인 위치를 나타내는 메타데이터에서 하나의 오브젝트의 복제본들은 다른 지역 그룹에 속한 데이터 노드들에 매핑된다. 그러므로, 어느 특정한 지역 그룹이 전체 네트워크 상에 문제가 생기는 등의 물리적인 피해를 입은 경우라도 복제본이 다른 지역 그룹에 속한 데이터 노드에 분산되어 저장되어 있으므로 신뢰성을 높일 수 있다.

[33] 본 발명에서, 지역 그룹은 하나의 데이터 센터가 될 수도 있고 좀 더 좁은 지역으로는 하나의 서버 랙이 될 수도 있다. 지역 그룹이 설정되면, 해당 지역 그룹 내에 속한 데이터 노드들을 그 지역 그룹에 속하는 것으로 메타데이터에

등록된다. 그러면, 오브젝트의 복제본들은 다른 지역 그룹에 속한 데이터 노드에 복제된다.

[34] 데이터 노드를 지역 그룹으로 그룹핑함으로써 얻어지는 장점은 다음과 같다.

[35] 1) 본 발명에서, 모든 클라이언트들(210, 212, 216) 및 데이터 노드들(DN11-DN1n, DN21-DN2n, DNm1-DNmn)들은 네트워크(290)를 통하여 서로 통신한다. 즉, 클라이언트 각각 및 데이터 각각 간에는 가상 채널(virtual channel)이 존재한다. 그러나, 이러한 가상 채널은 클라이언트 및 데이터 노드의 모든 쌍에 대해 반드시 동일한 조건을 가지는 것은 아니다. 예를 들어, 가상 채널의 통신 환경은 클라이언트 및 데이터 노드 간의 물리적 거리에 따라 달라질 수 있다. 클라이언트 및 데이터 노드 간의 물리적 거리가 멀수록 오브젝트는 더 많은 릴레이 노드 또는 게이트웨이를 통하여 전달되기 때문에 송수신 시간이 오래 걸린다. 또한, 가상 채널의 통신 환경은 네트워크 트래픽의 양 및 가상 채널을 구성하는 네트워크 자원의 성능에 따라서도 달라질 수 있다. 가상 채널을 통해 전달되는 트래픽의 양이 많을수록 가상 채널 상에서 전송 충돌(transmission collision)이 발생할 가능성이 높아지며, 네트워크 자원의 성능이 높을수록 가상 채널의 송수신 속도는 빨라진다. 그러므로, 본 발명에서는 이와 같은 가상 채널의 통신 환경을 고려하여 클라이언트 및 데이터 노드 간에 가장 최적의 가상 채널을 선택한다. 최적의 가상 채널을 선택하기 위하여 본 발명에 의한 분산 저장 시스템에서는 클라이언트 및 지역 그룹 간의 물리적 거리를 참조할 수 있다. 따라서, 저장한 오브젝트를 포함하는 클라이언트로부터 가장 가까운 거리에 위치한 지역 그룹에 속한 데이터 노드에 오브젝트를 저장함으로써 오브젝트의 업로드 시간을 최소화할 수 있다.

[36] 2) 또한, 본 발명에 의한 분산 저장 시스템은 오브젝트를 복제할 때 동일한 지역 그룹에 속하는 데이터 노드로 복제되지 않도록 한다. 따라서, 저장할 대상 오브젝트는 여러 개의 지역 그룹에 분산 저장된다. 일반적으로, 네트워크 장애가 발생하면 인접 지역의 데이터 노드들의 동작도 불가능한 경우가 많다. 예를 들어, 어느 데이터 센터에 여러 개의 데이터 노드들이 존재하고, 이 데이터 센터가 하나의 지역 그룹으로 설정된다고 가정한다. 이러한 가정은 본 발명을 용이하게 설명하기 위한 것으로 본 발명을 한정하는 것이 아님은 명백하다. 갑작스런 정전 등의 사고로 해당 데이터 센터가 동작 불능의 상태에 빠지는 경우가 발생할 수 있다. 이 경우, 본 발명에 의한 분산 저장 시스템은 대상 오브젝트를 해당 데이터 센터의 데이터 노드들 중 오직 하나의 대상 데이터 노드에만 저장하고, 그 복제본은 다른 지역 그룹에 속하는 대상 데이터 노드에 저장한다. 그러므로, 데이터 센터의 모든 데이터 노드들이 장애를 일으키더라도, 다른 지역 그룹에 속하는 대상 데이터 노드로부터 원하는 대상 오브젝트를 용이하게 조회할 수 있다.

[37] 이상과 같이, 본 발명에 의한 분산 저장 시스템(200)은 오브젝트의 실제 물리적 위치를 매핑하는 메타데이터가 오브젝트의 논리적 이름에 기반한 것이 아니라

오브젝트의 내용에 기반하여 구성된다. 그러므로, 중복 방지 동작을 수행하기 위하여 대상 오브젝트가 이미 저장된 것인지에 대한 판단이 용이하게 이루어진다.

발명의 실시를 위한 형태

- [38] 도 3은 본 발명의 다른 면에 의한 중복 방지 기능을 가지는 분산 저장 시스템의 오브젝트 저장 방법을 개념적으로 나타내는 흐름도이다.
- [39] 우선, 분산 저장 시스템에 포함되는 클라이언트를 인증 서버가 인증한다(S310). 인증이 성공하면, 프락시 서버가 대상 오브젝트(target object)를 저장하고자 하는 인증된 클라이언트의 오브젝트 저장 요청을 수신한다(S320). 만일 오브젝트 저장 요청이 수신되지 않으면 동작 요청이 수신될 때까지 대기한다.
- [40] 프락시 서버가 동작 요청을 수신하면, 프락시 서버는 대상 오브젝트의 내용을 이용하여 내용 특이적 인덱스를 결정한다(S330). 내용 특이적 인덱스가 결정되면, 프락시 서버는 결정된 내용 특이적 인덱스를 이용하여 대상 오브젝트가 기저장된 오브젝트들과 중복되는지 여부를 판단한다(S340).
- [41] 대상 오브젝트가 기저장된 오브젝트와 중복된다고 판단되면, 프락시 서버는 해당 저장 요청을 무시하고, 다음 동작 요청이 있을 때까지 대기한다. 반면에, 중복 여부 판단결과 중복되는 오브젝트가 없는 것으로 확인되면, 프락시 서버는 중복되지 않은 대상 오브젝트가 저장될 대상 데이터 노드(target data node)를 결정한다(S350). 대상 데이터 노드를 결정하기 위하여, 프락시 서버는 데이터 노드들의 로드 밸런싱을 위해 각 데이터 노드의 스토리지 용량을 고려하여 각 데이터 노드의 가중치 값을 사전에 결정할 수 있다. 그러면, 프락시 서버는 가중치 값을 참조하여 가장 가중치가 높은 데이터 노드를 먼저 대상 데이터 노드로서 할당한다. 이를 통하여, 데이터 노드들 간의 로드 밸런싱이 달성된다.
- [42] 대상 데이터 노드가 결정되면, 프락시 서버는 결정된 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공하고(S360), 클라이언트는 대상 오브젝트를 목록에 포함된 대상 데이터 노드에 저장한다(S370).
- [43] 전술한 바와 같이, 본 발명에 의한 프락시 서버는 오브젝트를 업로드할 때 해쉬 결과값이 같은 오브젝트들만을 대상으로 중복 여부를 판단한다. 그러므로, 중복 방지 동작을 효율적으로 수행할 수 있다. 즉, 본 발명에 의하면 오브젝트를 업로드할 때 해쉬 함수를 대상 오브젝트에 적용해 얻어지는 결과값을 보고 결과값이 같은 데이터 노드의 해당 폴더내의 오브젝트들이 대상 오브젝트와 동일한지 여부를 판단하면 족하다. 해쉬 알고리즘 자체의 특성상 내용이 다르면 거의 결과값이 중복되는 경우는 아주 희박하기 때문에 그 만큼 같은 오브젝트를 찾는 일은 효율적으로 수행되게 되고 더불어 중복 방지는 쉬워지게 된다.
- [44] 도 4는 본 발명에 적용될 수 있는 해쉬 함수의 특징을 설명하기 위한 표이다.
- [45] 해쉬 함수(hash function)는 임의의 길이의 입력 메시지를 고정된 길이의 출력값으로 압축시키는 함수이다. 해쉬 함수는 데이터의 무결성 검증 및

메세지의 인증에 사용되며, 일방향성 및 강한 충돌 회피성이라는 두 가지 성질을 만족해야 한다. 해쉬 함수를 사용할 경우, 주어진 조건을 만족하는 임의의 입력 메세지를 찾는 것이 계산적으로 불가능해진다.

- [46] 본 발명에 의한 프락시 서버는 도 4에 도시된 해쉬 알고리즘에 따르는 해쉬 함수를 이용하여 내용 특이적 인덱스를 생성한다. 도 4에는 각 알고리즘의 출력 길이, 블록 크기, 라운드 수 및 엔디언(Endianness)이 나열된다. 엔디언은 컴퓨터의 메모리와 같은 1차원의 공간에 여러 개의 연속된 대상을 배열하는 방법을 뜻한다.
- [47] 도 4에는, MD5, SHA1, SHA256, SHA384, RMD128, RMD160, RMD256, RMD320, HAS160, 및 TIGER 해쉬 함수 등이 소개되지만, 이는 열거적인 의미로 해석되어야 하고, 본 발명을 한정하는 것이 아니다.
- [48] MD5는 널리 사용된 해쉬 알고리즘이지만, 충돌 회피성에서 문제점이 있다는 분석이 있으므로 기존의 응용과의 호환으로만 사용하고 흔히 사용되는 않는다. SHA1은 DSA에서 사용하도록 되어 있으며 많은 인터넷 응용에서 디폴트 해쉬 알고리즘으로 사용된다.
- [49] 또한, SHA256, SHA384, SHA512는 (Advanced Encryption Standard)의 키 길이인 128, 192, 256 비트에 대응하도록 출력 길이를 늘린 해쉬 알고리즘이다. RMD128, RMD160은 RIPE 프로젝트의 RIPEMD나 MD4, MD5를 대신하기 위하여 디자인된 해쉬 알고리즘이다. 128 비트의 출력을 내는 RMD128은 역시 충돌 회피성에서 문제점이 있다. 이에 비하여, RMD160은 효율성은 떨어지지만 안전성을 높인 것으로 많은 인터넷 표준들에서 널리 채택되고 있다. RMD256과 RMD320은 각각 RMD128과 RMD160을 확장한 것이다.
- [50] 그리고, HAS160은 국내 표준 서명 알고리즘 KCDSA를 위하여 개발된 해쉬 함수이다. MD5와 SHA1의 장점을 취하여 디자인 되었다. TIGER는 64 비트 프로세서에 최적화되어서 64 비트 프로세서에서는 매우 빠르다.
- [51] 이와 같이, 본 발명에 의한 프락시 서버는 대상 오브젝트에 다양한 해쉬 함수를 적용하여 얻은 결과를 내용 특이적 인덱스로서 이용한다.
- [52] 도 5a 및 도 5b는 본 발명에 의한 분산 저장 시스템에서 이용하는 메타데이터에 포함되는 테이블들을 예시하는 도면이다.
- [53] 도 5a는 메타데이터에 포함되는 오브젝트 테이블을 예시하며, 도 5b는 복제본 위치 테이블을 예시한다. 오브젝트 테이블은 오브젝트의 사용자 ID(User ID), 디렉토리 ID(Directory ID), 오브젝트 ID(Object ID), 및 내용 특이적 인덱스(index)를 항목으로 포함한다. 그리고, 복제본 위치 테이블은 인덱스별 복제본의 위치를 항목으로 포함한다.
- [54] 프락시 서버는 도 5a와 같은 오브젝트 테이블을 만들어 각 오브젝트의 ID와 오브젝트의 일부 내용으로 해쉬 알고리즘을 적용한 결과값을 인덱스 컬럼에 저장한다. 각 오브젝트는 사용자 ID, 디렉토리 ID 그리고 오브젝트 ID로 구분 가능하다. 해쉬 알고리즘으로 예를 들어 MD5를 쓴다면 MD5는 임의의 길이의

메시지를 입력 받아 128 비트짜리 고정 길이의 출력 값을 생성함으로 인덱스 컬럼을 128 비트로 설정할 수 있다. 입력값으로는 오브젝트 내용에서 처음 64메가 바이트로 할 수 있다. 이것은 본 발명을 용이하게 설명하기 위한 것으로서, 본 발명을 한정하는 것이 아님은 명백하다.

- [55] 도 5b는 복제본의 개수가 3개인 경우를 가정한 경우이고 실제 복제본의 개수에 따라 복제본 위치 테이블의 컬럼 수를 조정하면 된다. 복제본 위치 테이블의 첫째 컬럼에서는 인덱스를 순서대로 저장하고 그 뒤의 컬럼들은 실제 복제본이 위치하는 데이터 노드의 ID 들이다. 예를 들어 도 5a의 오브젝트 테이블에서 mjkim 사용자의 Movies 디렉토리 밑의 Ants 오브젝트는 처음 64메가 바이트로 MD5 해쉬 알고리즘을 적용시 그 결과값이 24356이 나와서 그 값으로 인덱스 값을 가진다. 24356의 인덱스 값을 도 5b의 복제본 위치 테이블에서 찾으면 24, 52, 9의 데이터 노드의 ID와 매치가 된다. 즉, mjkim의 Ants 파일은 데이터 노드 24, 52 그리고 9에 존재하게 된다. 데이터 노드에서는 실제로 오브젝트 데이터를 저장할 때 인덱스 값을 키로 하여 쉽게 오브젝트 탐색이 가능하도록 한다. 예를 들어 인덱스 값 별로 폴더를 만들 수 있다. 같은 인덱스를 가지는 오브젝트들은 같은 데이터 노드의 같은 폴더에 저장되게 되는 것이다. 그러면 중복 방지 동작이 더욱 빠르게 수행될 수 있다.
- [56] 도 6은 본 발명의 일면에 의한 중복 방지 기능을 가지는 분산 저장 시스템의 다른 실시예를 개념적으로 나타내는 도면이다.
- [57] 도 6은 본 발명의 일면에 의한 분산 저장 시스템의 다른 실시예를 개념적으로 나타내는 도면이다.
- [58] 도 6에 도시된 분산 저장 시스템(600)은 네트워크(690)에 연결되는 복수 개의 클라이언트들(610, 612, 616) 및 데이터 노드들(DN11-DN1n, DN21-DN2n, DNm1-DNmn)을 포함한다. 또한, 도 6에 도시된 분산 저장 시스템(600)은 인증 서버(620), 프락시 서버(650), 위치 인식 서버(660), 복제 서버(670) 및 메타데이터 데이터베이스(680)를 더 포함한다.
- [59] 도 6에 도시된 클라이언트들(610, 612, 616), 인증 서버(620) 및 메타데이터 데이터베이스(680)의 구성 및 동작은 도 2에 도시되는 대응 구성 요소들의 그것과 유사하다. 그러므로, 명세서의 간략화를 위하여 반복적인 설명이 생략된다. 예를 들어, 도 6에 포함되는 프락시 서버(650)는 오브젝트 저장 요청이 있을 경우 대상 오브젝트에 해쉬 함수를 적용한 결과를 내용 특이적 인덱스로서 결정하고, 결정된 내용 특이적 인덱스를 이용하여 대상 오브젝트가 기저장된 대상 오브젝트와 동일한지 여부를 판단할 수 있다. 이하, 대상 오브젝트가 기저장된 오브젝트들과는 상이한 것일 경우에 대해서 설명한다.
- [60] 도 6에 도시된 분산 저장 시스템(600)에 포함되는 위치 인식 서버(660)는 지역 그룹 또는 대상 데이터 노드를 자동으로 선택하기 위하여 이용된다. 인증된 클라이언트가 대상 오브젝트를 저장할 대상 데이터 노드에 대해서 프락시 서버(650)에 질의하면, 프락시 서버(650)는 가장 유리한 지역 그룹에 대해서 위치

인식 서버(660)에 문의한다.

- [61] 위치 인식 서버(660)는 여러 가지 방법으로 클라이언트의 위치를 파악할 수 있는데, 일반적으로는 클라이언트의 IP 주소로 클라이언트의 물리적 위치를 파악할 수 있다. 위치 인식 서버(660)는 프락시 서버(650)의 요청에 따라 클라이언트의 기본 복제본 개수만큼의 지역 그룹을 선정한 후 선정된 지역 그룹 리스트를 프락시 서버(650)에게 전송한다. 위치 인식 서버(660)는 물리적으로 프락시 서버(650) 내에 통합되어 구현될 수 있다.
- [62] 위치 인식 서버(660)에 의하여 결정된 지역 그룹 각각에 속하는 대상 데이터 노드를 결정하는 동작은 프락시 서버(650) 또는 위치 인식 서버(660)에 의하여 수행될 수 있다. 위치 인식 서버(660)가 대상 데이터 노드까지 결정하는 경우, 위치 인식 서버(660)는 메타데이터 데이터베이스(680)를 참조하여 선택된 지역 그룹 내에서 대상 오브젝트를 가지고 있는 클라이언트와 가장 근접한 대상 데이터 노드를 선택할 수 있다. 반면에, 프락시 서버(650)가 대상 데이터 노드를 선택한다면, 프락시 서버(650)는 로드 밸런서(load balancer, 655)를 이용하여 각 지역 그룹에 속하는 데이터 노드들의 상태를 점검하고, 이 중에서 최적 조건을 가지는 데이터 노드를 대상 데이터 노드로서 선택할 수 있다. 로드 밸런서(655)는 프락시 서버(650)에 포함되는 것으로 도시되었으나, 이는 본 발명을 한정하는 것이 아님이 이해되어야 한다.
- [63] 또한, 프락시 서버(650)는 각 지역 그룹 내 데이터 노드들의 정보를 메타데이터에 관리하고, 데이터 노드들의 로드 밸런싱을 위해 각 데이터 노드의 스토리지 용량을 고려하여 각 데이터 노드의 가중치 값을 사전에 결정한다. 현재까지 각 데이터 노드의 오브젝트 저장 내역과 데이터 노드의 가중치 값을 고려해 요청 클라이언트의 데이터 노드를 선정함으로써 지역 그룹 내 데이터 노드간 로드 밸런싱을 유지한다.
- [64] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.
- [65] 또한, 본 발명에 따르는 방법은 컴퓨터로 읽을 수 있는 기록 매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함할 수 있다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있으며, 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 분산 컴퓨터 시스템에 의하여 분산 방식으로 실행될 수 있는 컴퓨터가 읽을 수 있는 코드를 저장할 수 있다.
- [66] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

산업상 이용가능성

- [67] 본 발명은 클라우드 스토리지 서비스를 제공할 수 있는 오브젝트 스토리지에서 중복 제거(deduplication)를 효율적으로 지원하기 위한 기술에 적용될 수 있다.

청구범위

[청구항 1]

복수 개의 클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템(distribution storage system)에 있어서,
 상기 클라이언트를 인증하기 위한 인증 서버;
 각각 적어도 하나의 오브젝트들을 저장하는 복수 개의 데이터 노드들;
 상기 오브젝트의 고유 정보 및 상기 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 저장하는 메타데이터 데이터베이스; 및
 대상 오브젝트(target object)를 저장하고자 하는 인증된 클라이언트의 오브젝트 저장 요청에 응답하여, 상기 메타데이터를 참조하여 상기 대상 오브젝트가 저장될 대상 데이터 노드(target data node)의 고유 정보의 목록을 상기 클라이언트에게 제공하는 프락시 서버를 포함하며, 상기 프락시 서버는,
 상기 클라이언트로부터 상기 저장 요청이 수신되면, 상기 대상 오브젝트의 내용에 의하여 결정되는 내용 특이적 인덱스(content-specific index)를 결정하고, 결정된 내용 특이적 인덱스를 이용하여 상기 대상 오브젝트가 기저장된 오브젝트들과 중복되는지 여부를 판단하며, 중복되지 않은 대상 오브젝트에 대해서만 상기 대상 데이터 노드의 고유 정보의 목록을 상기 클라이언트에게 제공하도록 구성되고, 상기 클라이언트는,
 제공된 상기 대상 데이터 노드의 고유 정보의 목록을 이용하여 상기 대상 오브젝트를 저장하도록 구성되는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 2]

제1항에 있어서, 상기 프락시 서버는,
 상기 대상 오브젝트의 소정 부분에 소정 해시 함수를 적용한 결과를 상기 내용 특이적 인덱스로서 결정하도록 구성되는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 3]

제2항에 있어서, 상기 프락시 서버는,
 상기 대상 오브젝트의 최초 소정 길이를 입력으로써 이용하는 MD5, SHA1, SHA256, SHA384, RMD128, RMD160, RMD256, RMD320, HAS160, 및 TIGER 해시 함수 중 어느 하나를 이용하여 상기 내용 특이적 인덱스를 결정하도록 구성되는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 4]

제3항에 있어서, 상기 메타데이터는,

사용자 ID, 디렉토리 ID, 오브젝트 ID, 및 상기 내용 특이적 인덱스 중 적어도 하나를 포함하는 오브젝트 테이블 및
 상기 내용 특이적 인덱스 및 상기 오브젝트의 복제본이 저장된 데이터 노드의 ID를 포함하는 복제본 위치 테이블을 포함하는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 5]

제1항에 있어서,
 상기 데이터 노드들은 지역(zone)별로 그룹화되며,
 상기 프락시 서버는, 동일한 오브젝트는 동일한 지역 그룹(zone group)에 속한 데이터 노드들 중 오직 하나에만 저장되도록 상기 대상 데이터 노드의 고유 정보의 목록을 결정하도록 구성되는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 6]

제5항에 있어서,
 상기 분산 저장 시스템은, 상기 데이터 노드들 및 상기 클라이언트의 위치 관계에 기반하여 상기 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하고, 선택된 지역 그룹 및 상기 클라이언트 간의 거리에 기반하여 상기 지역 그룹별 우선 순위를 결정하는 위치 인식 서버(location-aware server)를 더 포함하고,
 상기 프락시 서버는,
 상기 위치 인식 서버가 선택한 지역 그룹 당 하나의 대상 데이터 노드를 결정하고,
 결정된 대상 데이터 노드들의 목록을 이용하여 상기 메타데이터 데이터베이스를 갱신하며,
 상기 대상 데이터 노드들의 목록 및 상기 지역 그룹별 우선 순위를 상기 클라이언트에게 전송하도록 구성되고,
 상기 클라이언트는, 가장 높은 지역 그룹별 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터 노드에 상기 대상 오브젝트를 저장함으로써, 상기 우선 순위에 따라 순차적으로 더 낮은 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터 노드들에 상기 대상 오브젝트가 복제되는 복제 동작이 수행되도록 야기하도록 더욱 구성되는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 7]

제6항에 있어서, 상기 프락시 서버는,
 동일한 지역 그룹에 포함되는 데이터 노드들의 가용 저장 용량 및 오브젝트 저장 내역을 고려하여 동일한 지역 그룹에 포함되는 데이터 노드들에게 우선 순위를 부여하고,

가장 높은 우선 순위를 가지는 데이터 노드를 상기 대상 데이터 노드로서 결정하도록 더욱 구성되는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 8]

제1항에 있어서,
상기 오브젝트의 고유 정보는 상기 오브젝트의 ID, 크기, 데이터 타입, 및 작성자 중 적어도 하나를 포함하고,
상기 데이터 노드의 고유 정보는 상기 데이터 노드의 ID,
IP(Internet Protocol) 주소, 및 물리적 위치 중 적어도 하나를 포함하는 것을 특징으로 하는 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 9]

제5항에 있어서,
상기 메타데이터는, 상기 데이터 노드들의 사용량, 각 지역 그룹에 속한 데이터 노드들의 목록, 대상 오브젝트에 대한 지역 그룹별 우선 순위, 및 동일한 지역 그룹에 속한 데이터 노드들 간의 우선 순위 중 적어도 하나를 더 포함하는 것을 특징으로 하는 중복 방지 기능을 가지는 분산 저장 시스템.

[청구항 10]

복수 개의 클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템(distribution storage system)에 오브젝트를 분산 저장하는 방법에 있어서,
상기 클라이언트를 인증하는 단계;
프락시 서버가 대상 오브젝트(target object)를 저장하고자 하는 인증된 클라이언트의 오브젝트 저장 요청을 수신하는 단계;
상기 프락시 서버가 상기 대상 오브젝트의 내용에 의하여 결정되는 내용 특이적 인덱스를 결정하는 내용 특이적 인덱스 결정 단계;
결정된 내용 특이적 인덱스를 이용하여 상기 프락시 서버가 상기 대상 오브젝트가 기저장된 오브젝트들과 중복되는지 여부를 판단하는 단계; 및
상기 프락시 서버가, 중복되지 않은 대상 오브젝트에 대해서만 상기 오브젝트의 고유 정보 및 상기 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 참조하여 상기 대상 오브젝트가 저장될 대상 데이터 노드(target data node)를 결정하는 대상 데이터 노드 결정 단계;
상기 프락시 서버가, 결정된 대상 데이터 노드의 고유 정보의 목록을 상기 클라이언트에게 제공하는 단계; 및
상기 클라이언트가 상기 대상 오브젝트를 상기 목록에 포함된 대상 데이터 노드에 저장하는 단계를 포함하는 것을 특징으로

- 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템에서 오브젝트를 저장하기 위한 방법.
- [청구항 11] 제10항에 있어서, 상기 내용 특이적 인덱스 결정 단계는, 상기 프락시 서버가 상기 대상 오브젝트의 소정 부분에 소정 해시 함수를 적용한 결과를 상기 내용 특이적 인덱스로서 결정하는 단계를 포함하는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템에서 오브젝트를 저장하기 위한 방법.
- [청구항 12] 제11항에 있어서, 내용 특이적 인덱스 결정 단계는, 상기 프락시 서버가 상기 대상 오브젝트의 최초 소정 길이를 MD5, SHA1, SHA256, SHA384, RMD128, RMD160, RMD256, RMD320, HAS160, 및 TIGER 해시 함수 중 어느 하나에 적용하여 상기 내용 특이적 인덱스를 결정하는 단계를 포함하는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템에서 오브젝트를 저장하기 위한 방법.
- [청구항 13] 제12항에 있어서, 상기 메타데이터는, 사용자 ID, 디렉토리 ID, 오브젝트 ID, 및 상기 내용 특이적 인덱스 중 적어도 하나를 포함하는 오브젝트 테이블 및 상기 내용 특이적 인덱스 및 상기 오브젝트의 복제본이 저장된 데이터 노드의 ID를 포함하는 복제본 위치 테이블을 포함하는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템에서 오브젝트를 저장하기 위한 방법.
- [청구항 14] 제10항에 있어서, 상기 대상 데이터 노드 결정 단계는, 상기 프락시 서버가, 동일한 오브젝트가 동일한 지역 그룹(zone group)에 속한 데이터 노드들 중 오직 하나에만 저장되도록 상기 대상 데이터 노드의 고유 정보의 목록을 결정하는 단계를 포함하는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템에서 오브젝트를 저장하기 위한 방법.
- [청구항 15] 제14항에 있어서, 상기 대상 데이터 노드 결정 단계는, 위치 인식 서버가 상기 데이터 노드들 및 상기 클라이언트의 위치 관계에 기반하여 상기 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하고, 선택된 지역 그룹 및 상기 클라이언트 간의 거리에 기반하여 상기 지역 그룹별 우선 순위를 결정하는 단계; 및 상기 프락시 서버가 상기 위치 인식 서버가 선택한 지역 그룹 당 하나의 대상 데이터 노드를 결정하는 단계를 포함하는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템에서 오브젝트를 저장하기 위한 방법.
- [청구항 16] 제15항에 있어서, 상기 대상 데이터 노드 결정 단계는,

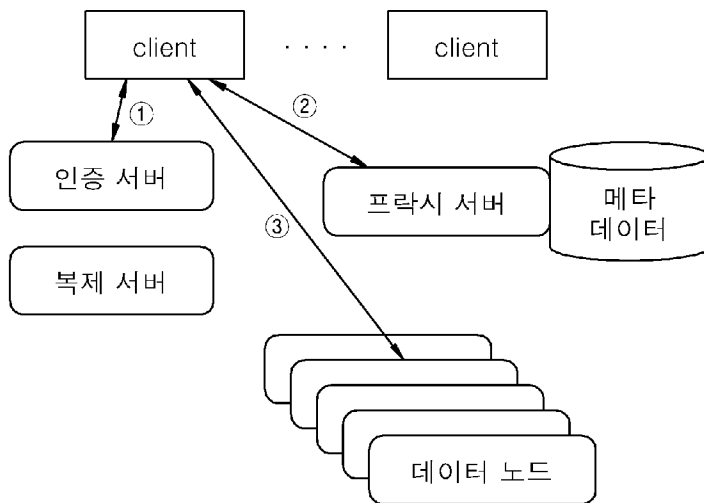
상기 프락시 서버가, 동일한 지역 그룹에 포함되는 데이터 노드들의 가용 저장 용량 및 오브젝트 저장 내역을 고려하여 동일한 지역 그룹에 포함되는 데이터 노드들에게 우선 순위를 부여하는 단계; 및

상기 프락시 서버가, 가장 높은 우선 순위를 가지는 데이터 노드를 상기 대상 데이터 노드로서 결정하는 단계를 포함하는 것을 특징으로 하는 내용 기반 중복 방지 기능을 가지는 분산 저장 시스템에서 오브젝트를 저장하기 위한 방법.

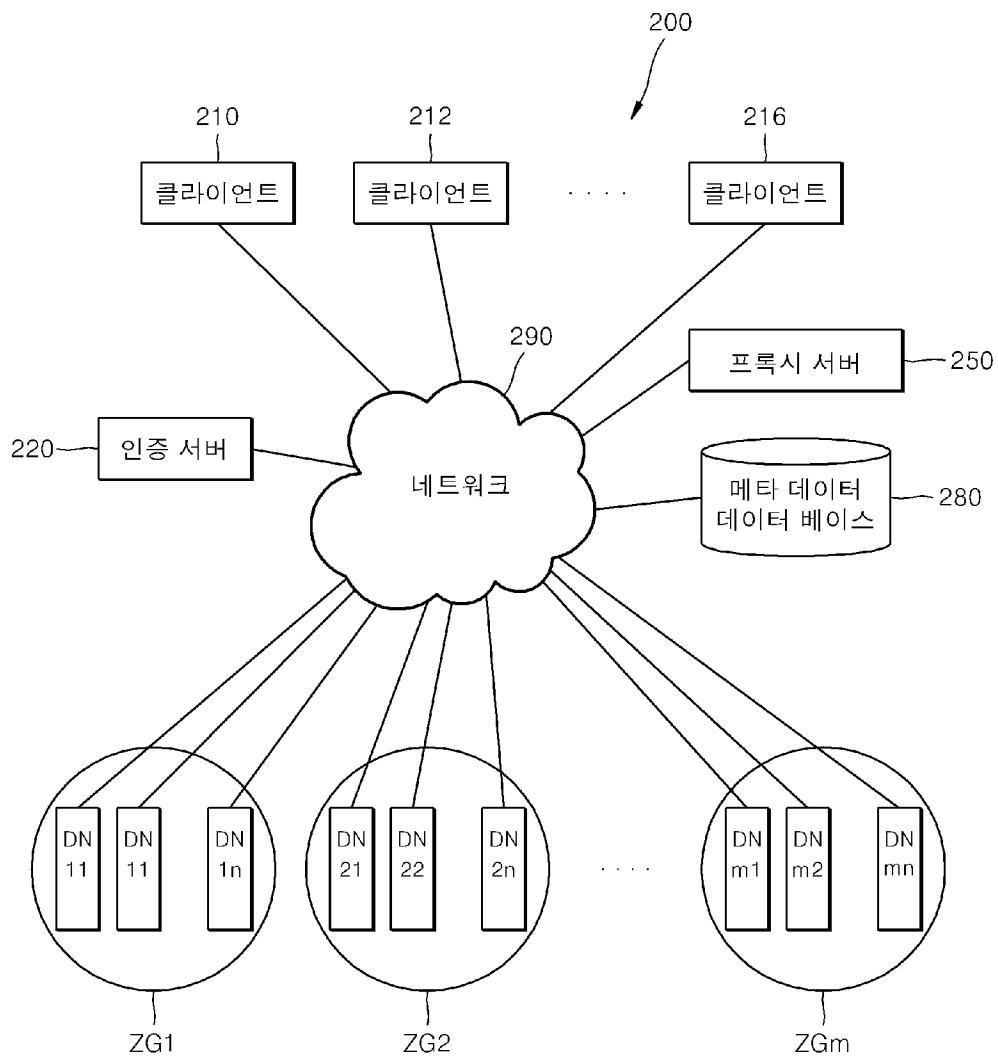
[청구항 17]

제10항 내지 제16항 중 어느 한 항에 따르는 방법을 구현하기 위하여 컴퓨터에 의하여 실행될 수 있는 컴퓨터 프로그램 명령어들을 저장하는 컴퓨터에 의하여 독출가능한 저장 매체.

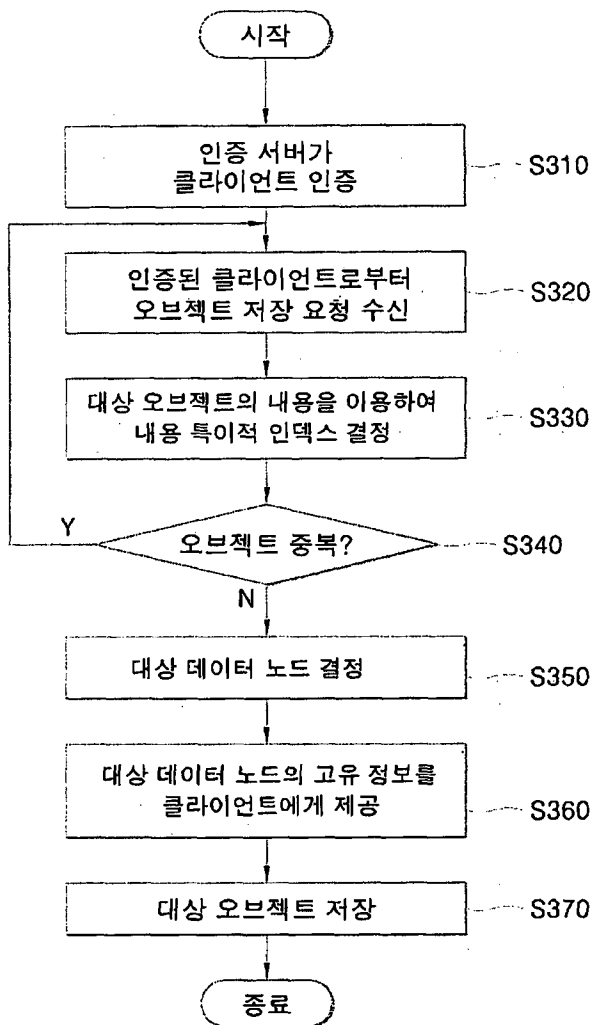
[Fig. 1]



[Fig. 2]



[Fig.3]



[Fig.4]

알고리즘	출력길이	블록의 크기	라운드 수	Endianness
MD5	128	512	64	Little
SHA1	160	512	80	Big
SHA256	256	512	64	Big
SHA384	384	1024	80	Big
SHA512	512	1024	80	Big
RMD128	128	512	128	Little
RMD160	160	512	160	Little
RMD256	256	512	128	Little
RMD320	320	512	160	Little
HAS160	160	512	80	Little
TIGER	192	512	56	Little

[Fig.5a]

User ID	Directory ID	Object ID	index
mjkim	Movies	Ants	24356
cgahn	cars	Lexes	13544

[Fig.5b]

index	Replica 위치 (데이터 노드의 ID)		
1			
⋮			
13544	13	44	2
⋮			
24356	24	52	9
⋮			
2 ^p			

[Fig. 6]

