(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 February 2013 (07.02.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/019869 A3

(51) International Patent Classification:
H04L 29/06 (2006.01)

(21) International Application Number:
PCT/US20 12/049 174

(22) International Filing Date:
1 August 2012 (01 .08.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/5 13,747 1 August 201 1 (01.08.201 1) US

(71) Applicant *(for all designated States except US):* **ACTIFIO, INC.** [US/US]; 225 Wyman St., Waltham, MA 0245 1 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only):* **ABERCROMBIE, Philip J.** [US/US]; 79 Winn St., Belmont, MA 02478 (US). **MUTALIK, Madhav** [US/US]; 8 Graystone Way, Southborough, MA 01772 (US). **PROVENZANO, Christopher, A.** [US/US]; 95 Liberty Ave., Somerville, MA 02144 (US).

(74) Agents: **SAJI, Michael, Y.** et al; Wilmer Cutler Pickering Hale and Dorr LLP, 60 State Street, Boston, MA 02109 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available):* AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, 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, KR, 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) Designated States *(unless otherwise indicated, for every kind of regional protection available):* ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (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).

Published:

— *with international search report (Art. 21(3))*

(88) Date of publication of the international search report:
13 June 2013

(54) Title: DATA FINGERPRINGTING FOR COPY ACCURACY ASSURANCE

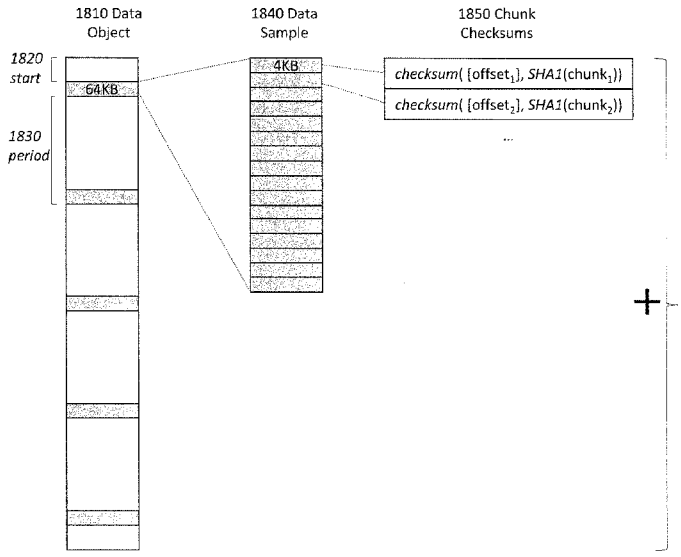

FIG. 18

(57) Abstract: Systems and methods are disclosed for efficiently and quickly creating a data fingerprint to identify or characterize contents of a data object by using a selection function to select a plurality of non-contiguous regions from the data object, the selected regions each having a small number of bytes relative to the number of bytes in the data object and being distributed throughout the data object so that the selected regions comprise a sparse subset of the data of the data object yet provide a significant probability of including bytes that change if the data object were modified; and performing a hash operation on the data contained in the plurality of regions to produce a fingerprint based on the sparse subset of the data object. The data fingerprint thereby efficiently provides an indication of the contents of the data object, so that comparing data fingerprints for two data objects can determine if the data objects are different if the corresponding fingerprints are different.

WO 2013/019869 A3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US12/49174

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - H04L 29/06 (2013.01)
USPC - 713/167
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8): G06F 12/14, 21/00, 17/15; H04L 29/06 (2013.01)
USPC: 713/167, 190; 708/422; 707/999.101

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
MicroPatent (US-G, US-A, EP-A, EP-B, WO, JP-bib, DE-CB, DE-A, DE-T, DE-U, GB-A, FR-A); DialogPRO; IEEE/IEEEXplore; Google/Google Scholar; IP.com; Search Terms: data, object, fingerprint, hash, one way function, non-contiguous, discontinuous, detach, disconnect, disjoin, nonsuccessive, sampling, binary, segment, subsegment, portion, subset, shingle, byte, chunk

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2007/0130188 A1 (MOON, H., et al.) June 7, 2007, abstract, figures 1A, 1B, 3A, paragraphs [0014], [0019]-[0021], [0048], [0055], [0058], [0059], [0063], [0066], [0074]	1-6
A	US 2011/0040819 A1 (LI, K., et al.) February 17, 2011, abstract, figures 2, 4, 6, paragraphs [0023], [0024], [0027], [0031], [0035], [0036]	1-6

☐ Further documents are listed in the continuation of Box C.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
31 January 2013 (31.01.2013)	08 FEB 2013

Name and mailing address of the ISA/US	Authorized officer:
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US12/49174

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. Ill Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Group I: Claims 1-6; Group II: Claims 7-12; Group III: Claims 13-19; Group IV: Claims 20-27

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fee must be paid.
Group I: Claims 1-6 are directed toward a method of efficiently and quickly creating a data fingerprint to identify or characterize contents of a data object, the method comprising: using a selection function to select a plurality of non-contiguous regions from the data object, the selected regions each having a small number of bytes relative to the number of bytes in the data object and being distributed throughout the data object so that the selected regions comprise a sparse subset of the data of the data object yet provide a significant probability of including bytes that change if the data object were modified; and performing a hash operation on the data contained in the plurality of regions to produce a fingerprint based on the sparse subset of the data object, the data fingerprint thereby efficiently providing an indication of the contents of the data object, so that comparing data fingerprints for two data objects can determine if the data objects are different if the corresponding fingerprints are different.
-"-Continued Within the Extra Page- ***-

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Claims 1-6

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

-Continued from Box No. III: Observations where unity of invention is lacking-

Group II: Claims 7-12 are directed toward a method for checking the data integrity of a data object copied between storage pools in a storage system by comparing data fingerprints of data objects, the method comprising: scheduling a series of successive copy operations over time for copying a data object from a source data store to a target data store; generating a partial fingerprint of the data object at the source data store using a data fingerprinting operation that creates a fingerprint from a subset of data of the data object, the subset being selected by a dynamic function that selects different subsets in response to a selection parameter that changes with each successive copy operation; sending the partial fingerprint of the data object to the target data store; sending any new data contents for the data object to the target data store; and creating a partial fingerprint of the data object at the target data store and comparing it to the partial fingerprint sent to the target data store to determine if they differ and thereby indicate that the data object at the target data store differs from the corresponding data object at the source data store, thereby allowing incremental verification that the copy of the data object at the target data store is the same as at the source data store.

Group III: Claims 13-19 are directed toward a system for copying a data object to a target storage pool using a hybrid of storage pools, in which at least one of the storage pools of the hybrid is particularly efficient at identifying data that should be used for copying the data object to the target storage pool, and at least one of the storage pools of the hybrid is particularly efficient at retrieving the data that should be sent to the target storage pool, the system comprising: a performance storage pool for storing data and having relatively high performance for retrieving stored data; a deduplicating storage pool for storing deduplicated data, said deduplicating storage pool further storing metadata about data objects in the system and which has relatively high performance for identifying and specifying differences in a data object over time; a controller for, in response to a command to copy a data object to the target storage pool, causing the deduplicating storage pool to identify and specify differences between a first version of a data object at a first instant in time and a second version of said data object at a second instant in time, for causing the specification of differences in the data object to be provided to the target storage pool, for causing the performance storage pool to retrieve any data specified in the differences specification that is not already stored at the target storage pool, and for causing the data retrieved by the performance storage pool to be provided to the target storage pool, thereby synchronizing the target storage pool to have the second version of the data object.

Group IV: Claims 20-27 are directed toward an asynchronous data replication system for providing a remote copy of data, in which remote replication is provided with reduced bandwidth requirements by copying deduplicated differences in business data from a local storage site to a remote, backup storage site, the system comprising: a local performance storage pool for storing data; a local deduplicating storage pool for storing deduplicated data, said local deduplicating storage pool further storing metadata about data objects in the system and which has metadata analysis logic for identifying and specifying differences in a data object over time; a remote performance storage pool for storing a copy of said data, available for immediate use as a backup copy of said data to provide business continuity to said data; a remote deduplicating storage pool for storing deduplicated data, said remote deduplicating storage pool being in communication with said local deduplicating storage pool and with said remote performance storage pool; and a controller for, in response to a remote replication command, causing the local deduplicating storage pool to identify and specify differences between a first version of a data object at a first instant in time and a second version of said data object at a second instant in time, for causing the specification of differences in the data object to be provided to the remote deduplicating storage pool, for causing the local performance storage pool to retrieve any data specified in the differences specification that is not already stored at the remote deduplicating storage pool, for causing the data retrieved by the local performance storage pool to be provided to the remote deduplicating storage pool, and for causing the remote deduplicating storage pool to provide a synchronized copy of said data to said remote performance storage pool; thereby synchronizing the remote performance storage pool to have the second version of the data object.

The common technical feature shared by Groups I, II, III and IV is a method of efficiently and quickly creating a data fingerprint to identify or characterize contents of a data object. However, this common feature is previously disclosed by US 2011/0022840 A1 (Stefan). Stefan discloses a method of efficiently and quickly creating a data fingerprint to identify or characterize contents of a data object (a method for detecting data modification including hash and fingerprint indicators, abstract).

Since the common technical feature is previously disclosed by the Stefan reference, this common feature is not special and so Groups I, II, III and IV lack unity.