

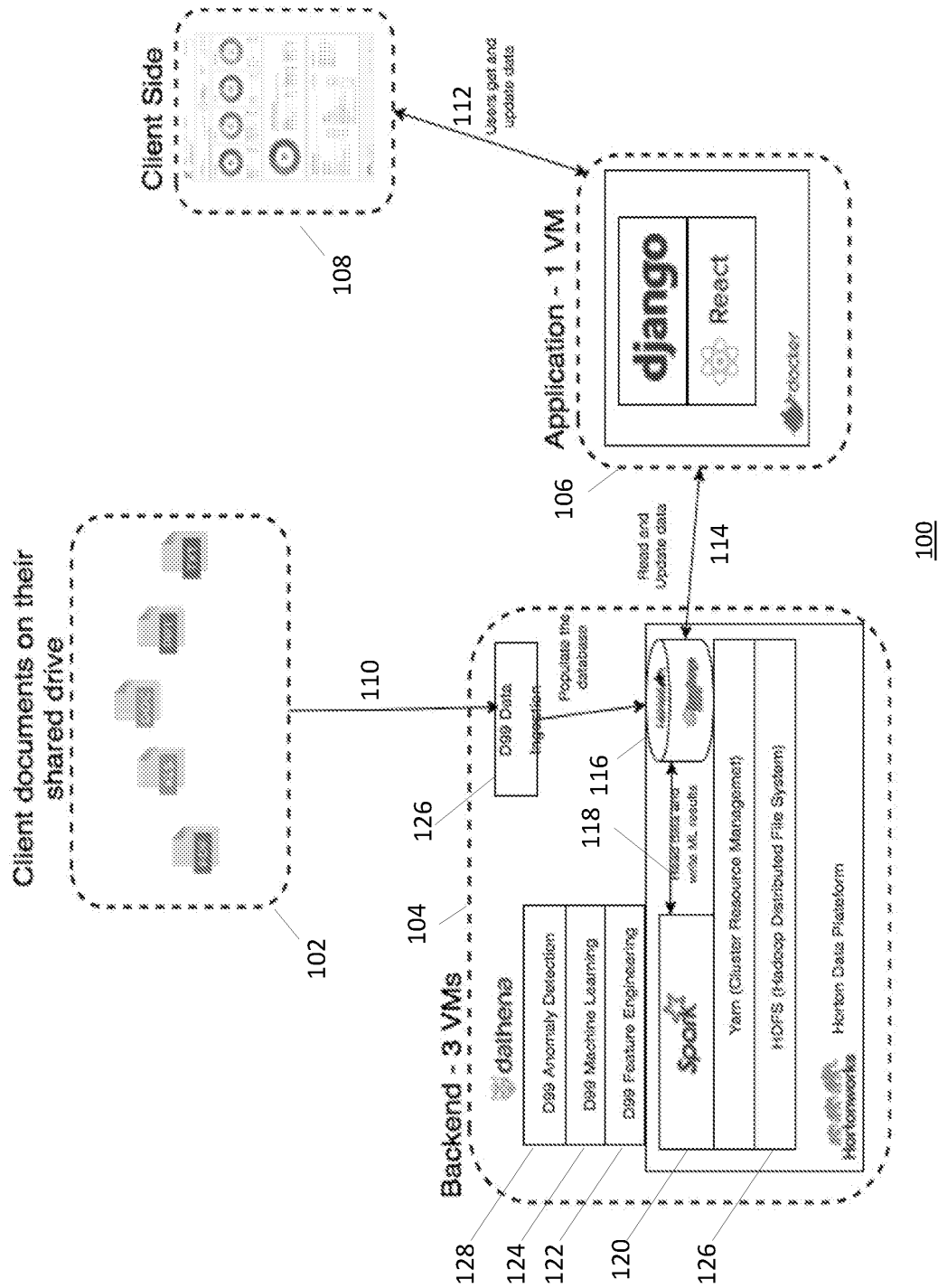


Fig. 1A

DOCMETA	
+ key: Document hash	
Information	
+ remotePath: Path in the drive	
+ size: Document size	
+ has_exception: "true" or "false"	
+ docType: Document type (Excel, Word, ...)	
+ encrypted: "true" "false"	
+ filename: last part of the remotePath	
+ has_datatypes: list of datatypes with ,	
+ hash: file's content sha256 hash	
+ is_train: 0 or 1	
+ language: short language name	
+ lastModified: date of last modification	
+ lastModifier: username of the last modifier	
+ owner: username of the owner	
+ owner_nifs: username of the NIFS owner -	
+ section_index: section index (DEPRECATED)	
Classification	
+ proposedCategory: Category from clustering	
+ reviewedCategory: Category from user review	
+ predictedCategory: Category from classification	
+ validatedCategory: Category from user validation	
+ proposedConfidentiality: Confidentiality from clustering	
+ reviewedConfidentiality: Confidentiality from user review	
+ predictedConfidentiality: Confidentiality from classification	
+ validatedConfidentiality: Confidentiality from user validation	
+ cluster_id	
+ distance: Distance to centroid (cluster)	
+ accuracy: Accuracy of the proposition	

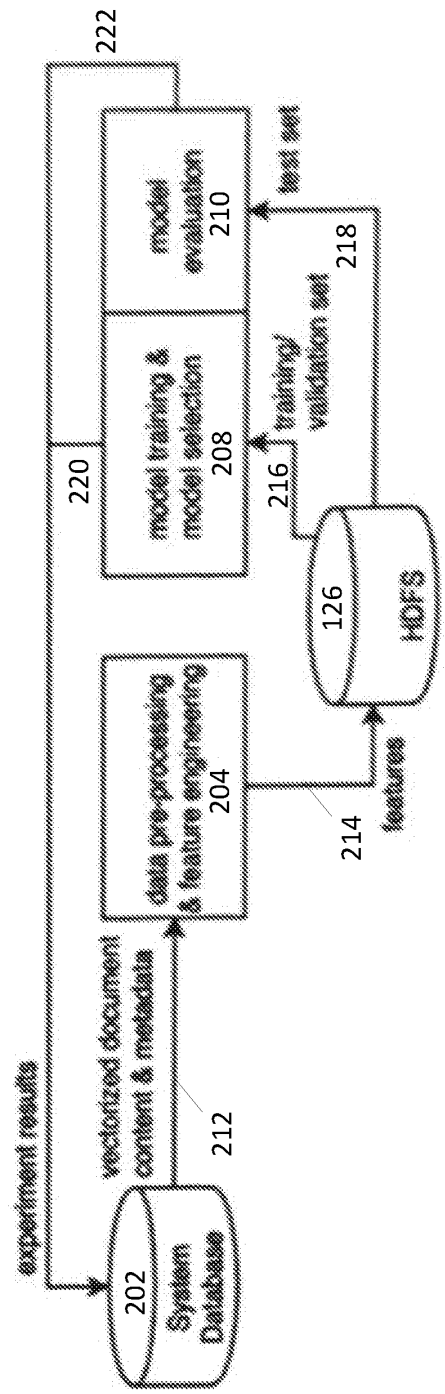
Fig. 1B

USER-DOC
+ key: hash + file: document key + user: user CN + access: group giving access

Fig. 1C

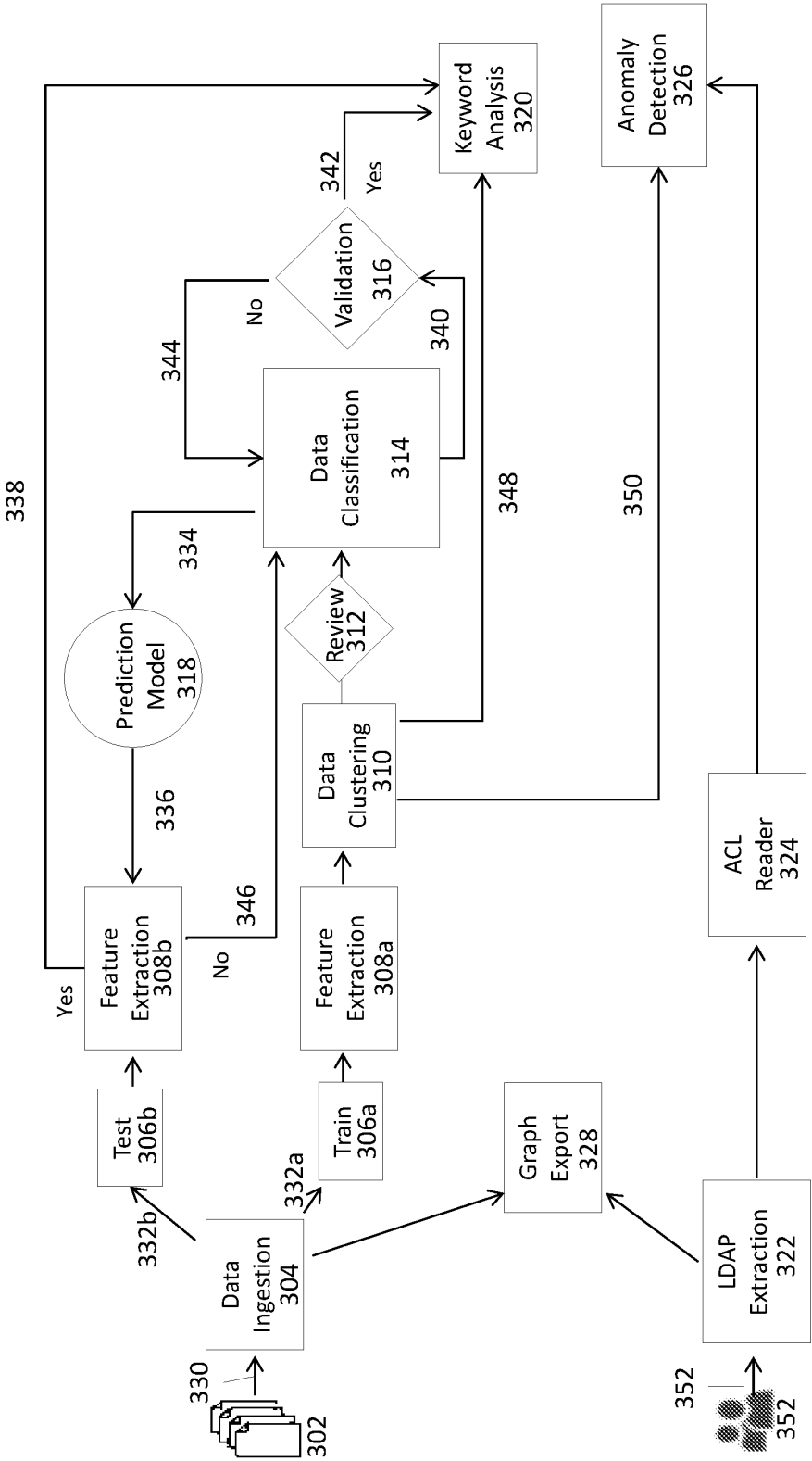
Table BOW
Row key: hashed file content (binary)
Column family 0
size (int)
indices (int array)
values (double array)
label (int)

Fig. 1D



200

Fig. 2



300
Fig. 3

