

1 / 10

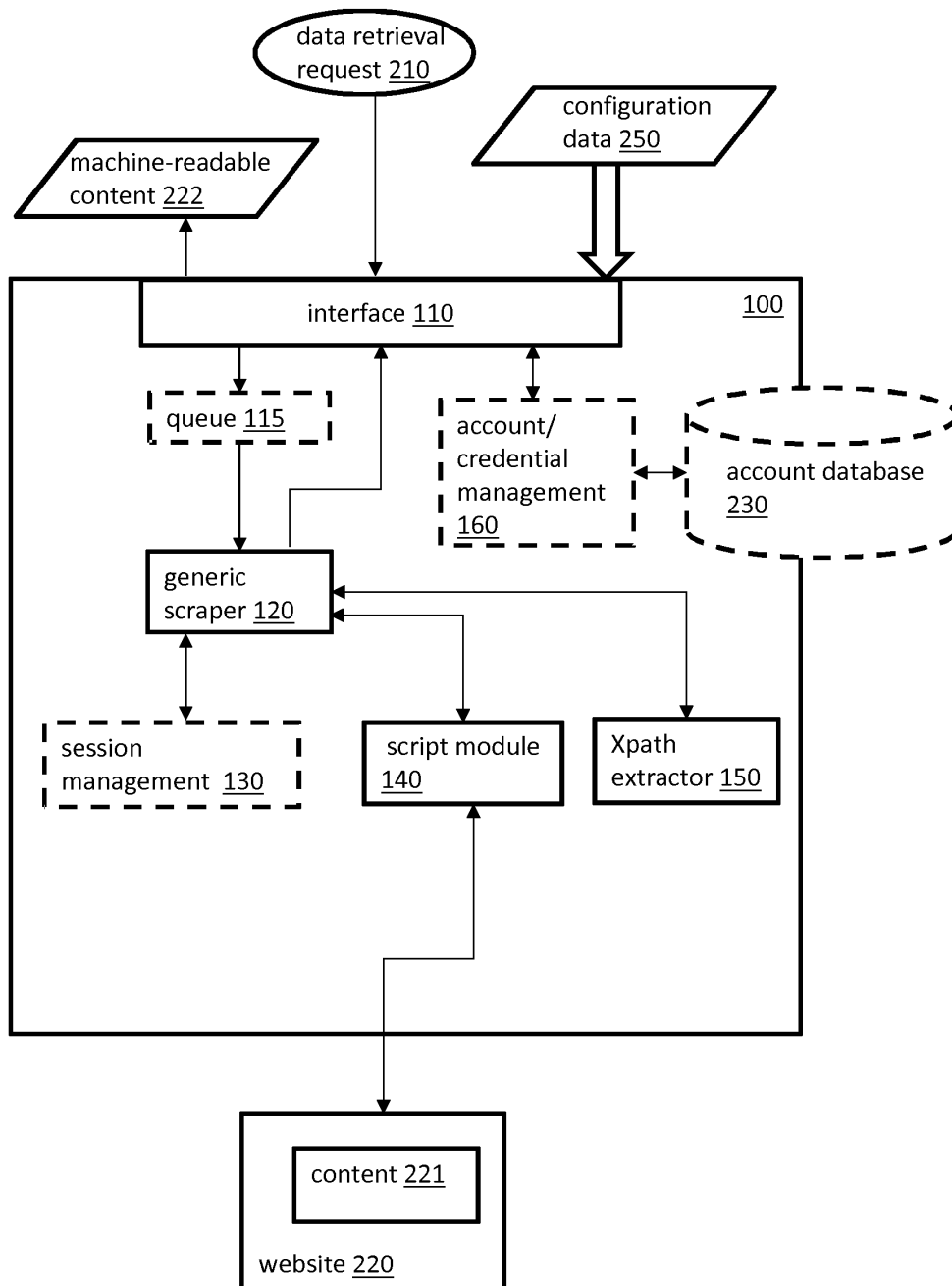


FIG. 1

2 /10

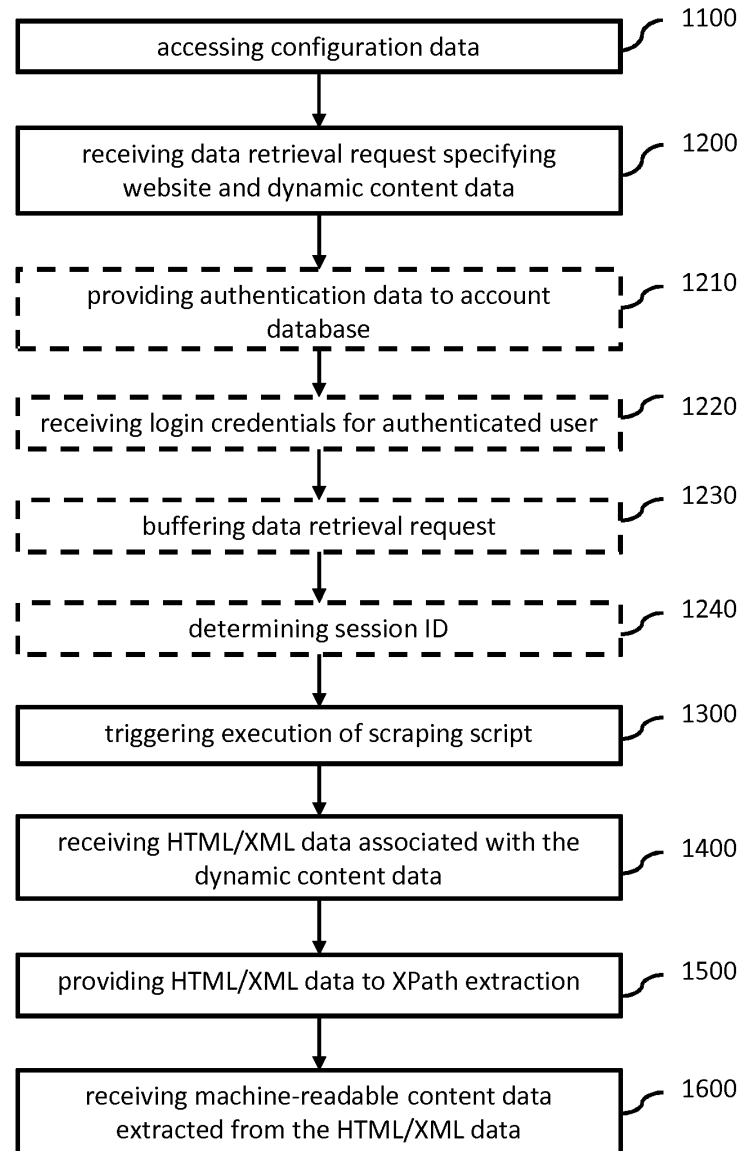
1000

FIG. 2A

3 /10

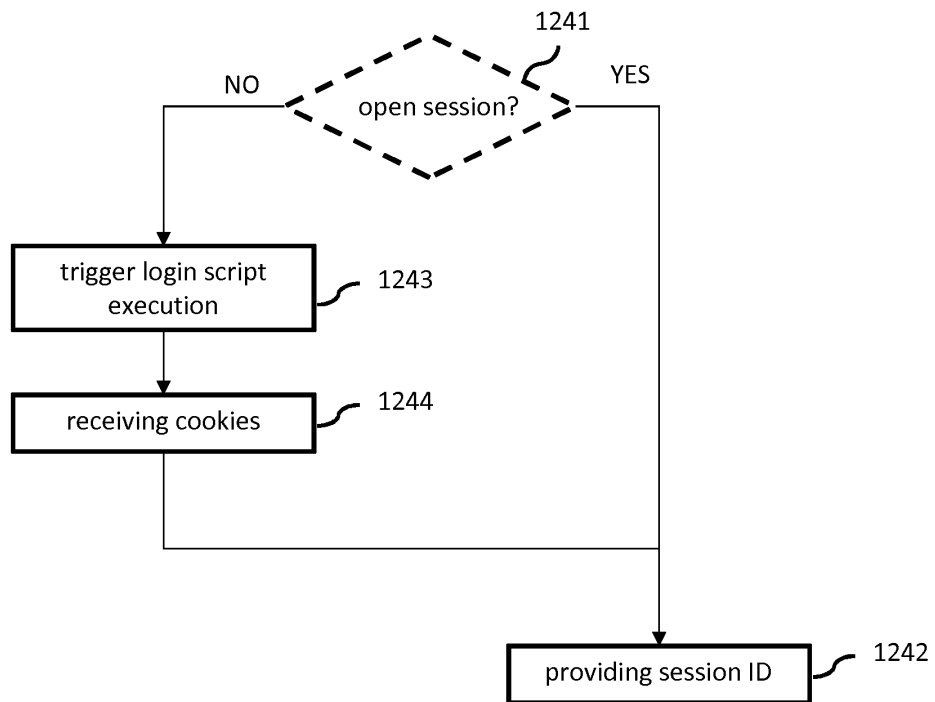
1240

FIG. 2B

4 /10

2000

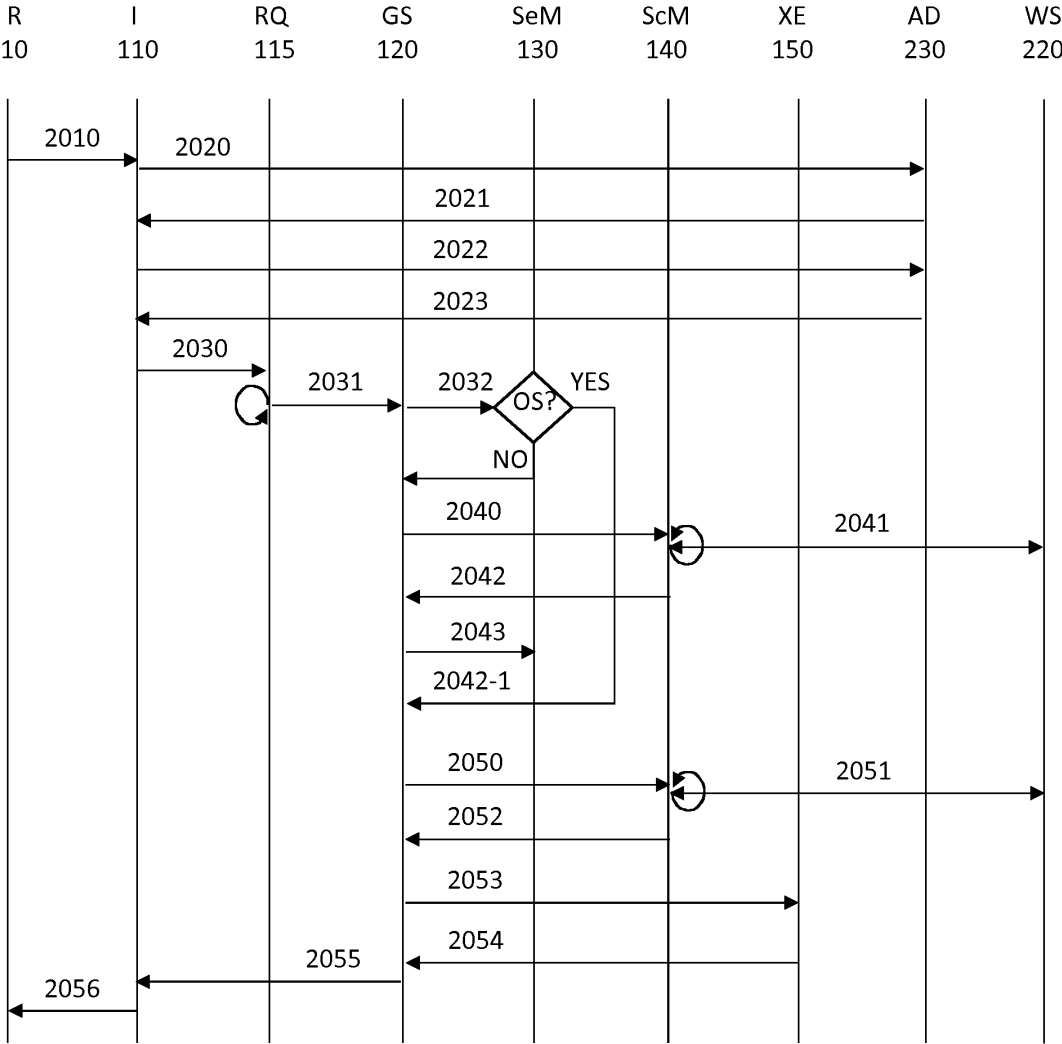


FIG. 3

5 /10

```

{
  "de.usi.screenscrapers.HttpScriptVerticle": {
    "instances": 5,
    "siteTimeOut": 10000
  },
  "de.usi.screenscrapers.RestInterfaceVerticle": {
    "instances": 1,
    "timeout": 50000
  },
  {
    "de.usi.screenscrapers.SessionManagementVerticle": {
      "instances": 1
    }
  },
  "de.usi.screenscrapers.CredentialManagementVerticle": {
    "instances": 1
  },
  "credentialManagement": {
    "dbUser": "<username>",
    "dbPassword": "<password>",
    "connectionUrl": "<server address>",
    "driverClass": "com.microsoft.sqlserver.jdbc.SQLServerDriver",
    "encryptionConfig": {
      "keyStoreFilename": "keystore.jks",
      "keyStorePassword": "<password>",
      "keyAlias": "<alias>",
      "keyPassword": "<password>"
    }
  },
  "cacheValidityMins": 15
},
"accountWebsite": {
  "port": 9999
},
},
....
}

```

400

401

402

403

404

405

FIG. 4A

6 /10

```

{
  ...
  "de.usi.screenscrapers.DomExtractorVerticle": {
    "instances": 5,
    "xpathConfigs": {
      "wikipedia": "wikipedia-robust.json",
    }
  },
  "server": {
    "port": 9090,
    ...
    "screenScrapperUrl": "/ScreenScrapper",
    ...
  },
  "scrapers": [
    {
      "class": "de.usi.screenscrapers.GenericScraperVerticle",
      "pathAndId": "wikipedia",
      "maxParallelRequests": 5,
      "instances": 1,
      "worker": false,
      "requiredParams": ["searchstring"],
      "scripts": {
        "scrape": [
          {
            "url": "https://de.wikipedia.org/",
            "method": "get",
            "checkSuccessXPath": "//span[text()='Willkommen bei Wikipedia']"
          },
          {
            "urlFromXPath": "//form[@id='searchform']/@action",
            "method": "post",
            "formData": {
              "search": "%searchstring%"
            },
            "checkSuccessXPath": "//title/contains(text(), '%searchstring%')"
          }
        ]
      }
    }
  ]
}

```

410

411

412

413

FIG. 4B

7 /10

450

```

"scrapers": [
  {
    "class": "de.usi.screenscrapers.GenericScraperVerticle",
    "pathAndId": "wikipedia",
    "credentialAlias": "wikipedia",
    ...
    "scripts": {
      "login": [
        {
          "url": "https://de.wikipedia.org/w/index.php?title=Spezial:Anmelden",
          "checkSuccessXPath": "///body//h1[@id='firstHeading'][text()='Anmelden']"
        },
        {
          "method": "post",
          "urlFromXPath": "///form[@name='userlogin']/@action",
          "formDataFromXPath": "///form[@name='userlogin']",
          "formData": {
            "wpName": "%username%",
            "wpPassword": "%password%"
          },
          "checkSuccessXPath": "///title[.='Object moved']"
        },
        {
          "urlFromHeader": "Location",
          "checkSuccessXPath": "///span[text()='Willkommen bei Wikipedia']"
        }
      ],
    },
  },
]

```

451

FIG. 4C

8 /10

420

```

{
  "foundStatus": {
    "statement": "if ((//h1[id='firstHeading'] != '') then 'page found' else 'not found'",
    "type": "string"
  },
  "isFound": {
    "statement": "if ((//h1[id='firstHeading'] != '') then 'true' else 'false'",
    "type": "boolean"
  },
  "pageTitle": {
    "statement": "(//title/text())[1]",
    "type": "string"
  },
  "pageNote": {
    "statement": "(//div[@class='mw-parser-output']/child::table/*[1]//tr/td[not(descendant::img)]/text())[1]",
    "type": "string"
  },
  "pageText": {
    "statement": "string-join(//div[@class='mw-parser-output']/*[not(self::table)]/text())",
    "type": "string"
  },
  "pageLinks": {
    "statement": "//h2[child::span[@id='Weblinks']]/following-sibling::div[1]/descendant::a/@href",
    "type": "string"
  }
}

```

FIG. 4D

9 /10

430

```
{
  "foundStatus": {
    "statement": "if (((//body//div/h1 != '')) then 'page found' else 'not found'",
    "type": "string"
  },
  "isFound": {
    "statement": "if (((//body//div/h1 != '')) then 'true' else 'false'",
    "type": "boolean"
  },
  "pageTitle": {
    "statement": "//title/text()",
    "type": "string"
  },
  "pageNote": {
    "statement": "//body/div/div/div/div/table/tbody/tr/td[2]/text()[1]",
    "type": "string"
  },
  "pageText": {
    "statement": "string-join(//body/div/div/div/div/*/text())",
    "type": "string"
  },
  "pageLinks": {
    "statement": "//*[@id='mw-content-text']/div/div[3]/b/a/@href",
    "type": "string"
  }
}
```

FIG. 4E

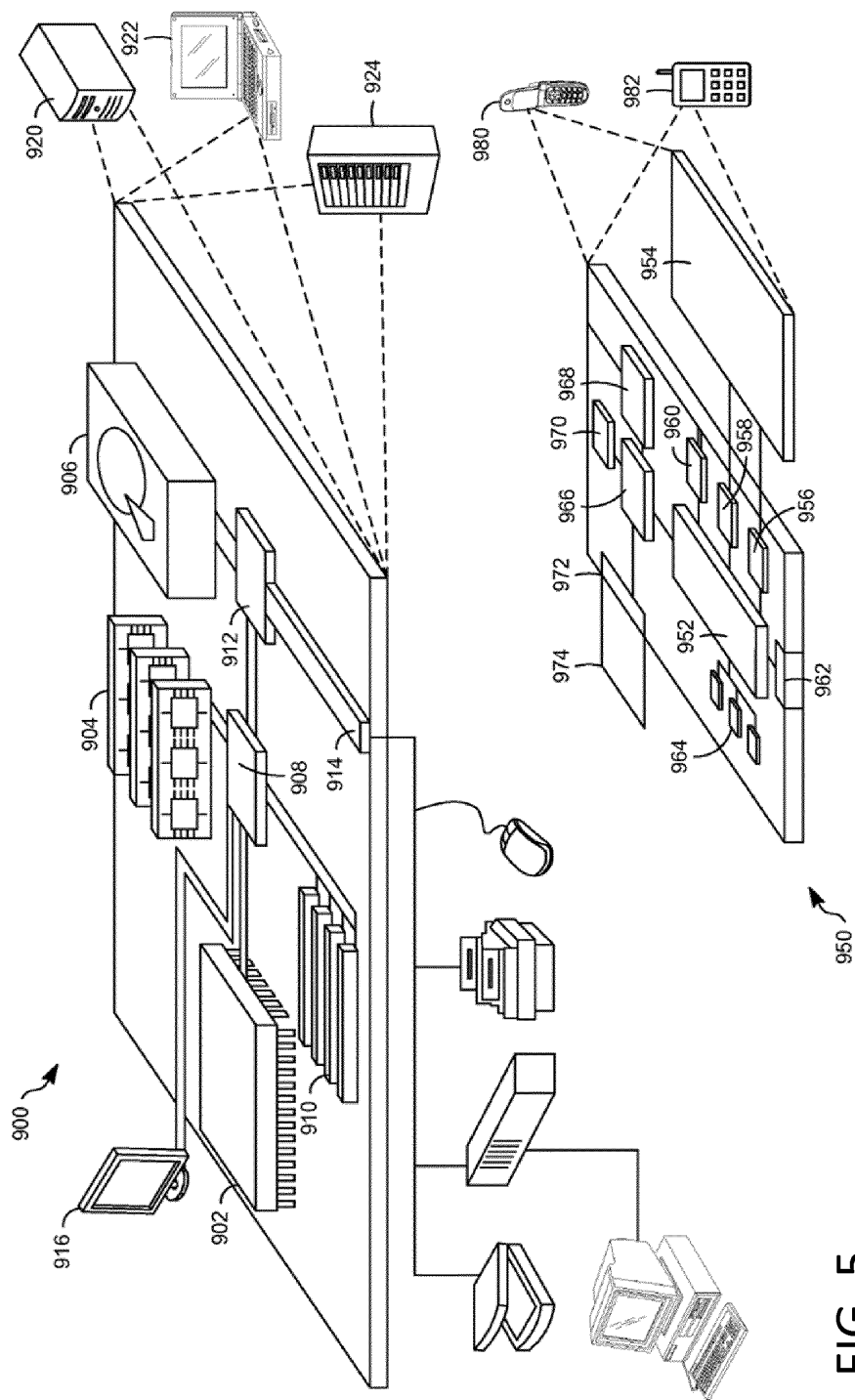


FIG. 5