RÉALISATION D'UN ÉMULATEUR DU SERVICE TEXTRACT DE AWS

SOUTENANCE DE STAGE 2A

JALAL IZEKKI

Bordeaux INP
ENSEIRB
MATMECA

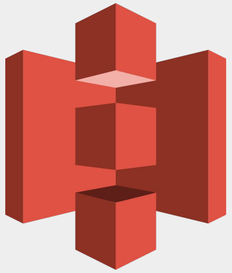
PLAN

1. Introduction & contexte
2. Problématique
3. Solution proposée
4. Conclusion

INTRODUCTION ET CONTEXTE

AWS Services

EC2
virtual
machines



S3
object
storage

Lambda
serverless
computing

CloudWatch
monitoring &
logging

Cognito
user sign-up
& sign-in

aws

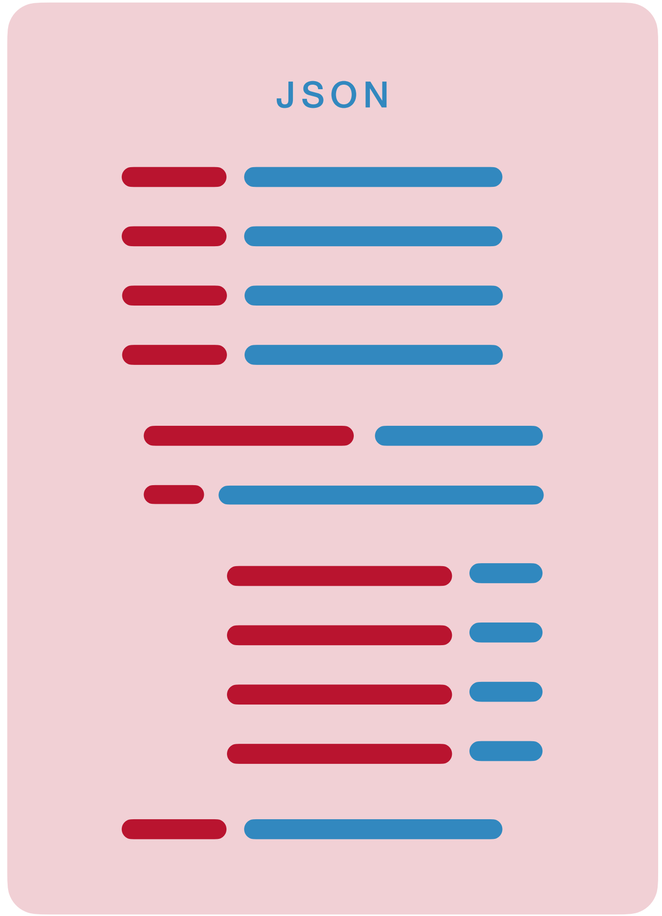
IAM
identity &
access

SNS
message
delivery

SQS
message
queuing

Textract
OCR

IMAGE JPEG / PNG OU PDF



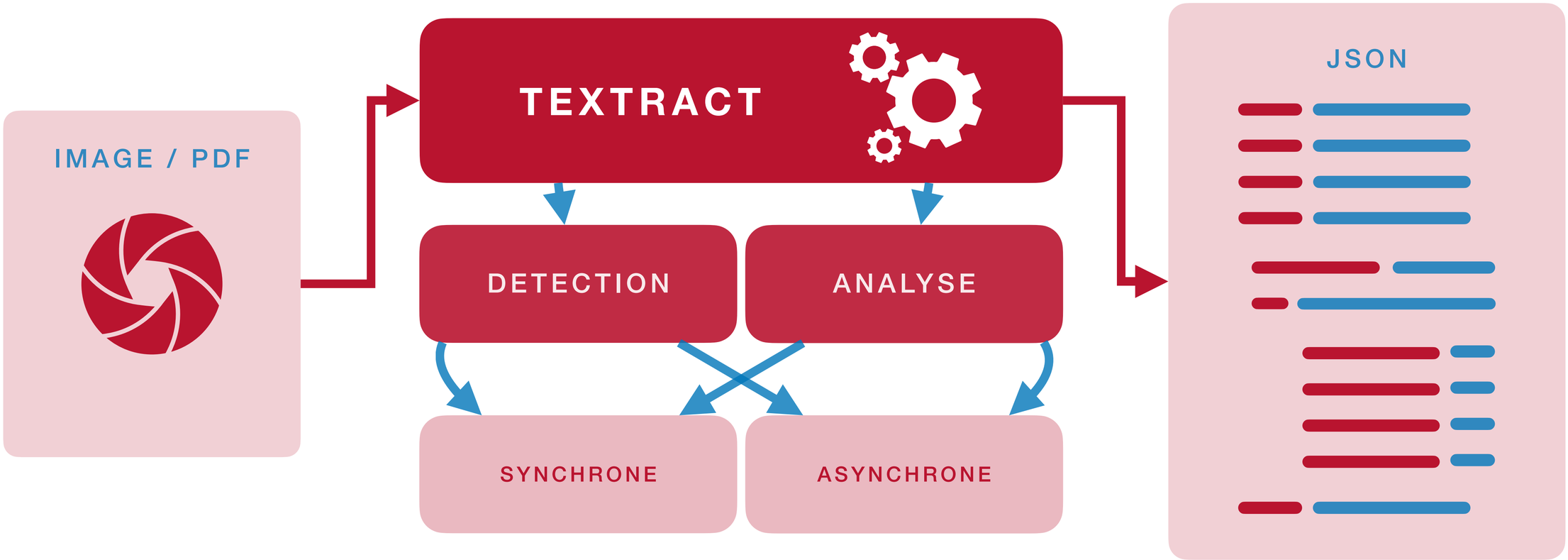
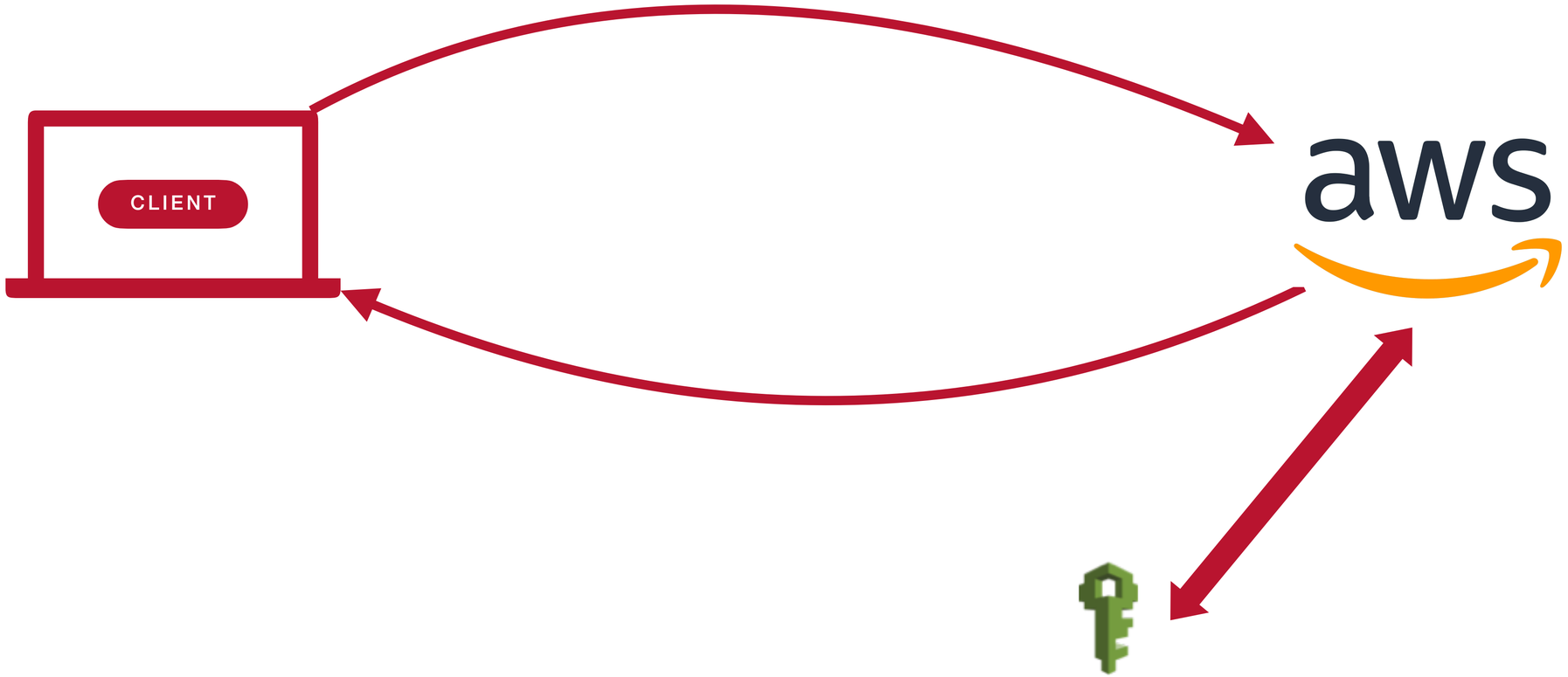


IMAGE / PDF
TEXTRACT
DETECTION
ANALYSE
SYNCHRONE
ASYNCHRONE
JSON

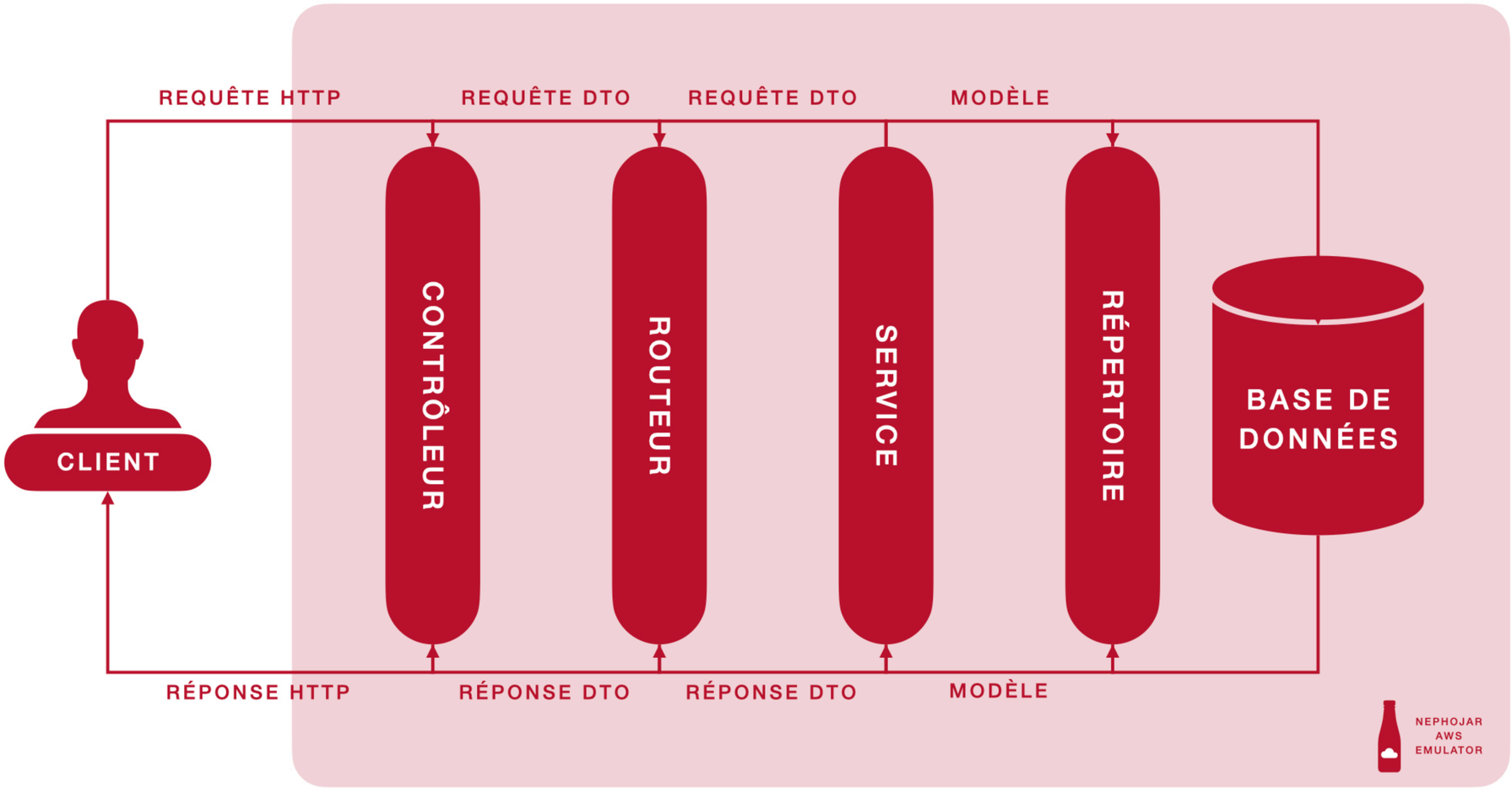
PROBLÉMATIQUE



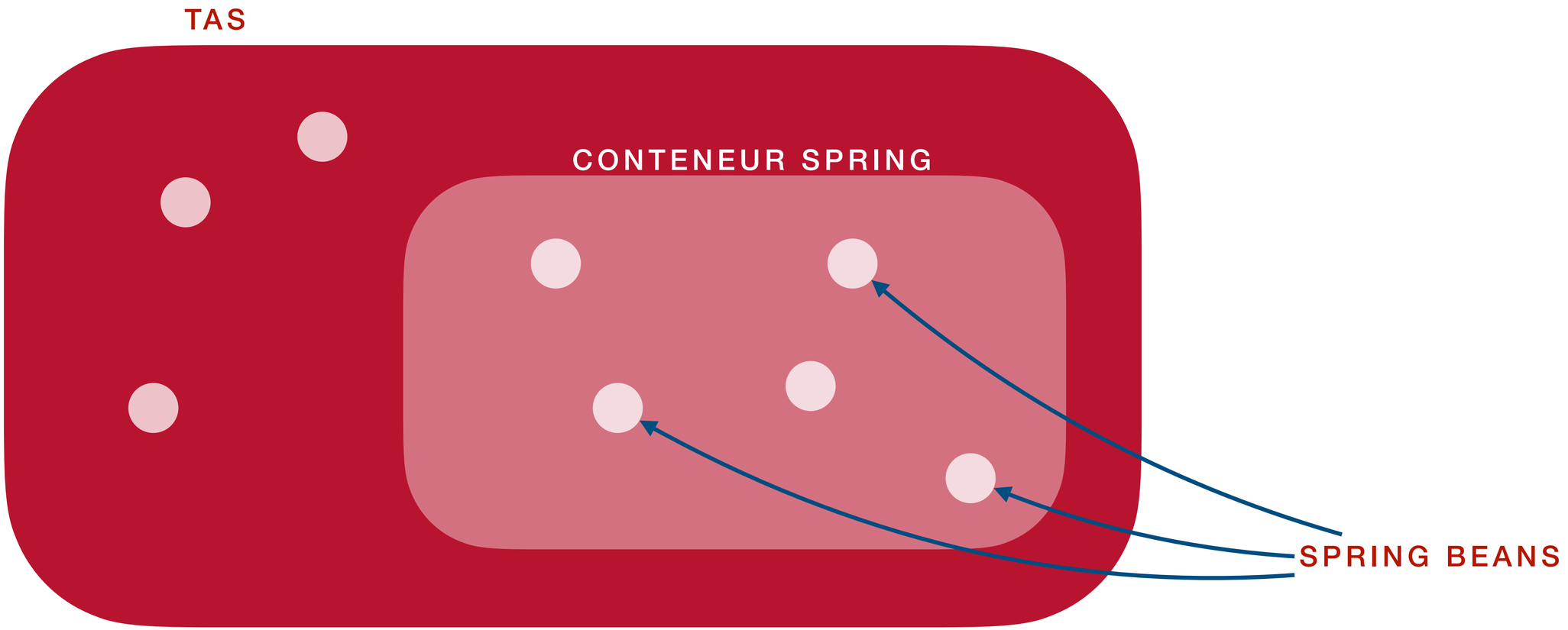
AWS IAM
VÉRIFICATION DES PERMISSIONS

SOLUTION PROPOSÉE

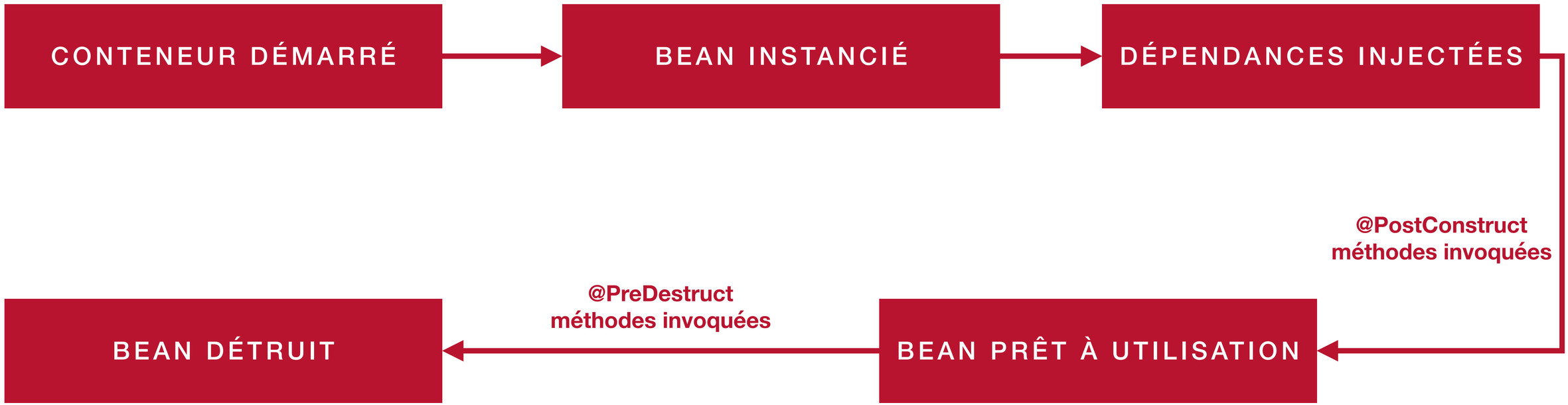
Structure d’un projet Spring



Fonctionnement de Spring



Cycle de vie d’un Spring Bean



Syntaxe d’une requête DetectDocumentText

```
{
   "Document": {
      "Bytes": blob,
      "S3Object": {
         "Bucket": "string",
         "Name": "string",
         "Version": "string"
      }
   }
}
```

Syntaxe d’une requête AnalyzeDocument

```
{
   "Document": {
      "Bytes": blob,
      "S3Object": {
         "Bucket": "string",
         "Name": "string",
         "Version": "string"
      }
   },
   "FeatureTypes": [ "string" ],
   "HumanLoopConfig": {
      "DataAttributes": {
         "ContentClassifiers": [ "string" ]
      },
      "FlowDefinitionArn": "string",
      "HumanLoopName": "string"
   }
}
```

Syntaxe d’une requête StartDocumentTextDetection

```
{
   "ClientRequestToken": "string",
   "DocumentLocation": {
      "S3Object": {
         "Bucket": "string",
         "Name": "string",
         "Version": "string"
      }
   },
   "JobTag": "string",
   "KMSKeyId": "string",
   "NotificationChannel": {
      "RoleArn": "string",
      "SNSTopicArn": "string"
   },
   "OutputConfig": {
      "S3Bucket": "string",
      "S3Prefix": "string"
   }
}
```

Syntaxe d’une requête GetDocumentTextDetection

```
{
   "JobId": "string",
   "MaxResults": number,
   "NextToken": "string"
}
```

Syntaxe d’une requête StartDocumentAnalysis

```
{
   "ClientRequestToken": "string",
   "DocumentLocation": {
      "S3Object": {
         "Bucket": "string",
         "Name": "string",
         "Version": "string"
      }
   },
   "FeatureTypes": [ "string" ],
   "JobTag": "string",
   "KMSKeyId": "string",
   "NotificationChannel": {
      "RoleArn": "string",
      "SNSTopicArn": "string"
   },
   "OutputConfig": {
      "S3Bucket": "string",
      "S3Prefix": "string"
   }
}
```

Syntaxe d’une requête GetDocumentAnalysis

```
{
   "JobId": "string",
   "MaxResults": number,
   "NextToken": "string"
}
```

Outil de reconnaissance de caractères

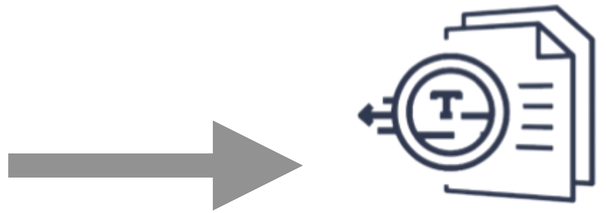
Outil de traitement d’images

Amazon.com, Inc. is located in Seattle, WA

It was founded July 5th, 1994 by Jeff Bezos

Amazon.com allows customers to buy everything from books to blenders

Seattle is north of Portland and south of Vancouver, BC.



Textract
OCR

```
{
   "BlockType": "PAGE",
   "Confidence": null,
   "Text": null,
   "TextType": null,
   "RowIndex": null,
   "ColumnIndex": null,
   "RowSpan": null,
   "ColumnSpan": null,
   "Geometry": {
      "BoundingBox": {
         "Width": 1.0,
         "Height": 1.0,
         "Left": 0.0,
         "Top": 0.0
      },
      "Polygon": [
         {
            "X": 0.0,
            "Y": 0.0
         },
         {
            "X": 1.0,
            "Y": 0.0
         },
         {
            "X": 1.0,
            "Y": 1.0
         },
         {
            "X": 0.0,
            "Y": 1.0
         }
      ]
   },
   "Id": "u2yu00m7-883z-h2ou-izna-m883zh2ouizm",
   "Relationships": [
      {
         "Type": "CHILD",
         "Ids": [
            "u2yu00m6-883z-h2ou-izfa-m883zh2ouize",
            "u2yu00m6-883z-h2ou-izha-m883zh2ouizg",
            "u2yu00m6-883z-h2ou-izja-m883zh2ouizi",
            "u2yu00m6-883z-h2ou-izla-m883zh2ouizk"
         ]
      }
   ],
   "EntityTypes": null,
   "SelectionStatus": null,
   "Page": null
},
```

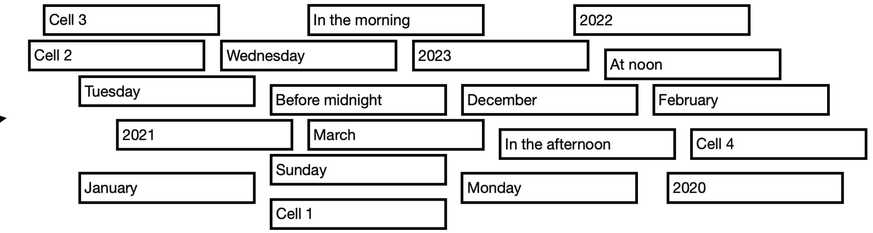
```
{
   "BlockType": "LINE",
   "Confidence": 90.2963,
   "Text": "Amazon.com, Inc. is located in Seattle, WA",
   "TextType": null,
   "RowIndex": null,
   "ColumnIndex": null,
   "RowSpan": null,
   "ColumnSpan": null,
   "Geometry": {
      "BoundingBox": {
         "Width": 0.51079136,
         "Height": 0.060215052,
         "Left": 0.0647482,
         "Top": 0.20430107
      },
      "Polygon": [
         {
            "X": 0.0647482,
            "Y": 0.20430107
         },
         {
            "X": 0.5755396,
            "Y": 0.20430107
         },
         {
            "X": 0.5755396,
            "Y": 0.26451612
         },
         {
            "X": 0.0647482,
            "Y": 0.26451612
         }
      ]
   },
   "Id": "u2yu00m6-883z-h2ou-izfa-m883zh2ouize",
   "Relationships": [
      {
         "Type": "CHILD",
         "Ids": [
            "u2yu00ak-883z-h2ou-ixfa-m883zh2ouixe",
            "u2yu00ak-883z-h2ou-ixha-m883zh2ouixg",
            "u2yu00ak-883z-h2ou-ixja-m883zh2ouixi",
            "u2yu00ak-883z-h2ou-ixla-m883zh2ouixk",
            "u2yu00ak-883z-h2ou-ixna-m883zh2ouixm",
            "u2yu00ak-883z-h2ou-ixpa-m883zh2ouixo"
         ]
      }
   ],
   "EntityTypes": null,
   "SelectionStatus": null,
   "Page": null
},
```

```
{
   "BlockType": "WORD",
   "Confidence": 84.5368,
   "Text": "Amazon.com,",
   "TextType": "PRINTED",
   "RowIndex": null,
   "ColumnIndex": null,
   "RowSpan": null,
   "ColumnSpan": null,
   "Geometry": {
      "BoundingBox": {
         "Width": 0.15930113,
         "Height": 0.047311828,
         "Left": 0.0647482,
         "Top": 0.2172043
      },
      "Polygon": [
         {
            "X": 0.0647482,
            "Y": 0.2172043
         },
         {
            "X": 0.22404933,
            "Y": 0.2172043
         },
         {
            "X": 0.22404933,
            "Y": 0.26451612
         },
         {
            "X": 0.0647482,
            "Y": 0.26451612
         }
      ]
   },
   "Id": "u2yu00ak-883z-h2ou-ixfa-m883zh2ouixe",
   "Relationships": null,
   "EntityTypes": null,
   "SelectionStatus": null,
   "Page": null
},
```

Year	Month	Day	Time
2020	January	Monday	In the morning
2021	February	Tuesday	At noon
2022	March	Wednesday	In the afternoon
2023	December	Sunday	Before midnight

Year	Month	Day	Time
2020	January	Monday	In the morning
2021	February	Tuesday	At noon
2022	March	Wednesday	In the afternoon
2023	December	Sunday	Before midnight

Year	Month	Day	Time
2020	January	Monday	In the morning
2021	February	Tuesday	At noon
2022	March	Wednesday	In the afternoon
2023	December	Sunday	Before midnight



Year	Month	Day	Time
2020	January	Monday	In the morning
2021	February	Tuesday	At noon
2022	March	Wednesday	In the afternoon
2023	December	Sunday	Before midnight

COCLUSION

Améliorations possibles du projet

- Implémentation de la détection des formulaires
  - Soit par la même procédure suivie pour la détection des tableaux
  - Soit par l'entraînement d’un modèle avec des méthodes de l'apprentissage profond
- Amélioration de la détection des tableaux
  - Pour mieux analyser les tableaux ayant une structure plus complexe
- Amélioration des tests des méthodes asynchrones
  - En ajoutant plus de tests pour couvrir tous les cas d’utilisation

MERCI POUR VOTRE ATTENTION

Bordeaux INP
ENSEIRB
MATMECA