Assignment 2: Map-Reduce

1. Lab Assignments: Due Dates & Grading

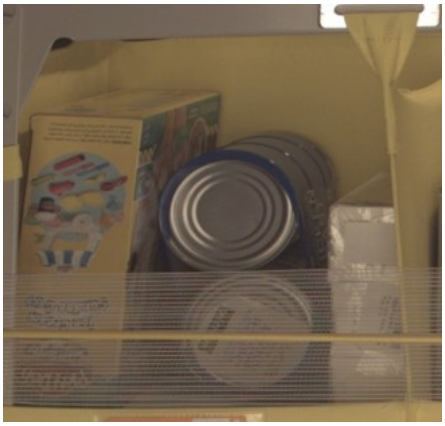
The grading for lab assignments is up to 4 points of the total grading of the course: Lab assignments (4 points) and Exam (6 points). The partial grading and due date of Lab Assignments are:

LAB GRADING (4 points)		
Assignment 1 (2 points)	**AWS Architecture for Tagged Files**	
	Delivery	20th October 20:00 (*faitic*)
	Assessment	Monday, 23th October. 11:00 (*T110*)
Assignment 2 (2 points)	**Map Reduce on Distributed Data**	
	Delivery	24th November 20:00 (*faitic*)
	Assessment	Monday, 27th November. 11:00 (*T110*)
EXAM (6 points)		

2. The Data

You will be working with **Amazon Bin Image Dataset,** which contains images and metadata from bins of a pod in an operating Amazon Fulfilment Centre. The bin images in this dataset are captured as robot units carry pods as part of normal Amazon Fulfilment Centre operations. Amazon uses a random storage scheme where items are placed into accessible bins with available space, so the contents of each bin are random, rather than organized by specific product types. Thus, each image may show only one type of product or a diverse range of products. Occasionally, items are misplaced while being handled, so the contents of some bin images may not match the recorded inventory of that bin.

Over 500,000 JPEG images and corresponding JSON metadata files describing items are available in the "aft-vbi-pds" S3 bucket in the US East Region. Images are located in the *bin-images* directory, and metadata for each image is located in the *metadata* directory. The bin image contains multiple object categories and various number of instances. The corresponding metadata exist for each bin image and it includes the object category identification (*Amazon Standard Identification Number*, ASIN), quantity, size of objects, weights, and names. Images and their associated metadata share simple numerical unique identifiers. See an example of image and metadata below.



```
{"BIN_FCSKU_DATA": {

 "B00CFQWRPS": {
    "asin": "B00CFQWRPS",
    "height": {
    "unit": "IN",
    "value": 2.399999997552
     },
    "length": {
       "unit": "IN",
       "value": 8.199999991636
        },
    "name": "Fleet Saline Enema, 7.8
Ounce (Pack of 3)",

    "normalizedName": "(Pack of 3) Fleet
Saline Enema, 7.8 Ounce",

     "quantity": 1,
     "weight": {
        "unit": "pounds",
        "value": 1.8999999999999997
        },
      "width": {
         "unit": "IN",
         "value": 7.199999992656
         }
   },
```

```
"ZZXI0WUSIB": {
    "asin": "B00T0BUKW8",
     "height": {
        "unit": "IN",
         "value": 3.99999999592
         },
      "length": {
          "unit": "IN",
          "value": 7.899999991942001
          },
       "name": "Kirkland Signature
Premium Chunk Chicken Breast Packed in
Water, 12.5 Ounce, 6 Count",

        "normalizedName": "Kirkland
Signature Premium Chunk Chicken Breast
Packed in Water, 12.5 Ounce, 6 Count",

          "quantity": 1,
          "weight": {
             "unit": "pounds",
              "value": 5.7
               },
          "width": {
             "unit": "IN",
             "value": 6.49999999337
              }
  },
```

```
"ZZXVVS669V": {
   "asin": "B00C3WXJHY",
   "height": {
      "unit": "IN",
      "value": 4.330708657
      },
    "length": {
       "unit": "IN",
       "value": 11.1417322721
       },
    "name": "Play-Doh Sweet Shoppe
Ice Cream Sundae Cart Playset",

     "normalizedName": "Play-Doh
Sweet Shoppe Ice Cream Sundae Cart
Playset",

      "quantity": 1,
         "weight": {
         "unit": "pounds",
         "value": 1.4109440759087915
            },
       "width": {
          "unit": "IN",
           "value": 9.448818888
           }
        }
    },


"EXPECTED_QUANTITY": 3
}
```


3. Part I: Apache Hadoop locally in pseudo-distributed mode (1 point)

Using the metadata of Amazon Bin Image dataset, we expect you to create the corresponding mappers and reducers to accomplish the following tasks:

- ✓ Global statistics:
  - Expected number of items.
  - Average *height, length, width and weight.*
  - Maximum and minimum length of name.
  - Maximum and minimum length of normalized name.
- ✓ For each ASIN in the dataset:
  - Expected number of items.
  - TRUE if all items of the ASIN have the same metadata (height, length, width, weight, name and normalized name).

The student should use the CLI interface or any other interface to S3 in order **to obtain a local copy** of the metadata information in Amazon **Bin Image Dataset.** The **result should be a *json* file** with this information entitled "*StatisticsBinImage.json*" (see an example format in Section 6). The student can **freely decide about the programming language**. To deploy the processing infrastructure, the student should set up a **single-node Hadoop in their computer and run the job with pseudo-distributed Operation** where each Hadoop daemon runs in a separate process (emulating separate nodes in a cluster).

4. Part II: Apache Hadoop in Amazon EC2 pseudo-distributed mode (0'5 points)

In Part II, we expect you to use the same **data** to run a Hadoop job in an Amazon EC2 instance to **repeat the Part I** but using the data **available at Amazon S3** as a bucket and running the job in Amazon EC2.

Running Hadoop in EC2 instances is not different from running Hadoop in your local computer, but you should be aware of the remote configuration of EC2 instances. **We recommend this unofficial tutorial:**

[https://letsdobigdata.wordpress.com/2014/01/13/setting-up-hadoop-multi-node-cluster-on-amazon-ec2-part-1/].

5. Part III: Comparison (0'5 points)

The student should compare the performance of Part I and Part II. The student is free to set up an alternative running environment in Amazon EC2 which improves performance. For that comparison, the metrics from the job history (either stored on the JobTracker or on HDFS) can be used; also third-party packages.

6. Submission

Please, note that you have to deliver your results using the AWS S3 service.

- Part I. Upload the code and a result file with the results to *faitic*.
- Part II: AMI & Results should be in a bucket named "*AmazonBinImageStatistics*"
- Part III: Upload Report on performance comparison to *faitic*.

7. Example result file


```
{
  "StatisticsBinImage":{
    "General":{
      "Expecteditems":1,
      "Averageheight":{
        "unit":"IN",
        "value":2.399999997552
      },
      "Averagelength":{
        "unit":"IN",
        "value":8.199999991636
      },
      "Averageweight":{
        "unit":"IN",
        "value":8.199999991636
      },
      "Averagewidth":{
        "unit":"IN",
        "value":8.199999991636
      },
      "MinimumName":100,
      "MaximunName":10000,
      "MinimumNormalizedName":200,
      "MaximumNormalizedName":2000
    },
    "B00CFQWRPS":{
      "items":150,
      "identical":"TRUE"
    },
    "ZZXI0WUSIB":{
      "items":2,
      "identical":"TRUE"
    },
    "ZZXVVS669V":{
      "items":2,
      "identical":"FALSE"
    }
  }
}
```