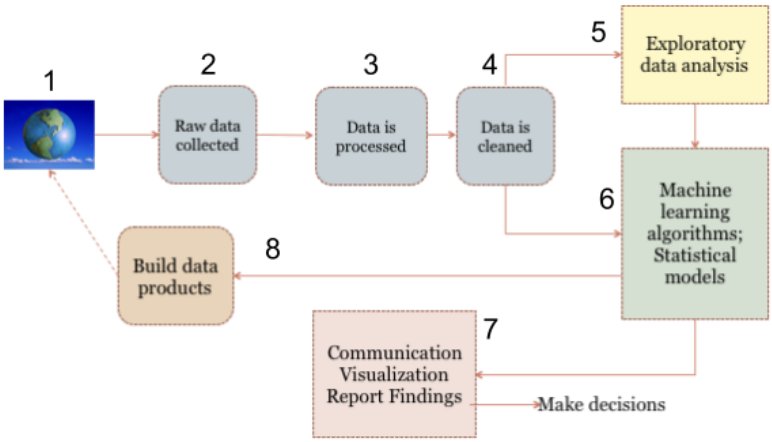
Question 1 - Ethics in Data Science [20 Points]



The diagram to the right shows the data science pipeline we've referred to over the course of the semester. Each stage of the pipeline has been labeled with a number, which you can use in the following questions related to ethics in data science.

a) Name and define (**in at most one sentence**) 4 of the 6 types of colloquial biases discussed in class. [12 points]

1. **Historical Bias: Bias in the data that occurs due to pre-existing biases present in society.**
2. **Representation Bias: Biases which occur when a particular group is under-represented in the dataset.**
3. **Measurement Bias: Bias that occurs when there is a mismatch between the variable that you want to measure, vs the variable you are actually using.**
4. **Aggregation Bias: Biases that occur when you attempt to use one model across multiple groups and that model does not work well for some of the groups present.**
5. **Evaluation Bias: Biases that occur when evaluating the model often due to a poor choice of test dataset, or an evaluation metric that does not treat the groups present in the model fairly.**
6. **Deployment Bias: Biases that occur when our deployment environment does not match the environment in which the model was built/trained/evaluated.**

b) For each of the 4 types of bias mentioned above, name one of the stages in the data science pipeline where this type of bias may show up. [4 points]

1. **May show up in step 1 & 2**
2. **May show up in step 2, 3, or 4**
3. **May show up in 2 or 6 (either the wrong data was collected, or the wrong features were selected)**
4. **May show up in step 6**
5. **May show up in step 6**
6. **May show up in step 7 or 8**

c) Choose ONE of the 4 types of bias mentioned above and briefly describe a solution to correct for that type of bias, and which stage(s) of the pipeline this fix would involve. [4 points]

1. **Cannot "fix", must instead be aware of it throughout the process. One fix would be to include human involvement when products are deployed in steps 7 or 8.**
2. **Fix is to collect more data on the underrepresented groups (stage 2), and/or to choose different sampling methods (stage 3, 4).**
3. **Fix is to either collect different data (stage 2) if possible that more closely matches what you are attempting to measure, or choose features that more closely match when building your models (stage 6)**
4. **Using EDA (stage 5), you can determining which groups are represented in the data, and then during stage 6 more carefully choose and apply models for each group.**
5. **During stage 6, when evaluating the models, ensure that test data is independent of training data, and make sure to use an evaluation metric that is appropriate, or try out multiple evaluation metrics.**
6. **Try to ensure that your model is tested/developed (stage 6) in a way that more closely matches your planned deployment environment.**

Question 2 - Distributed Systems [30 Points]

a. List 2 distinct benefits that Spark has over MapReduce [6 points]

In memory caching provides better performance on iterative applications
Richer set of operations improves productivity
Can be more easily used in all parts of the DIC pipeline

b. Briefly explain how HIVE improves programmer productivity when compared to MapReduce [2 points]

When you have tabular data, HIVE provides a higher level abstraction for querying that data that does not require the programmer to write raw map reduce. The richer operations make it easier to write more powerful/complex queries in a much simpler/more succinct way.

c. In **one sentence** explain the primary way fault-tolerance is achieved in:

i. HDFS [4 points]

Data is split into blocks and each block is REPLICATED on multiple nodes

ii. Spark [4 points]

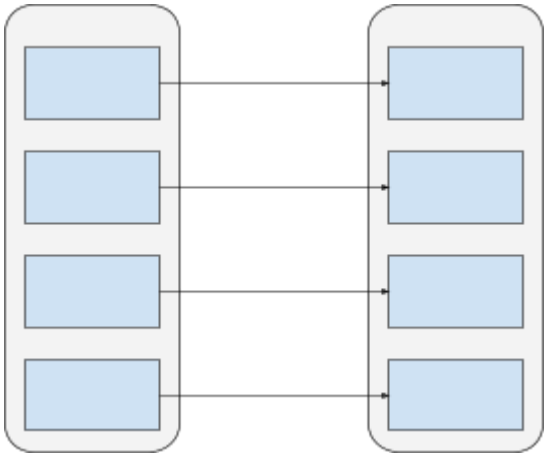
The DAG of computation is stored so that if data is lost it can easily be RECOMPUTED

d. Explain the difference between a transformation and an action in Spark [4 points]

Transformations are evaluated lazily – they do not trigger any computation; they just build the DAG.
Actions are what actually triggers computation.

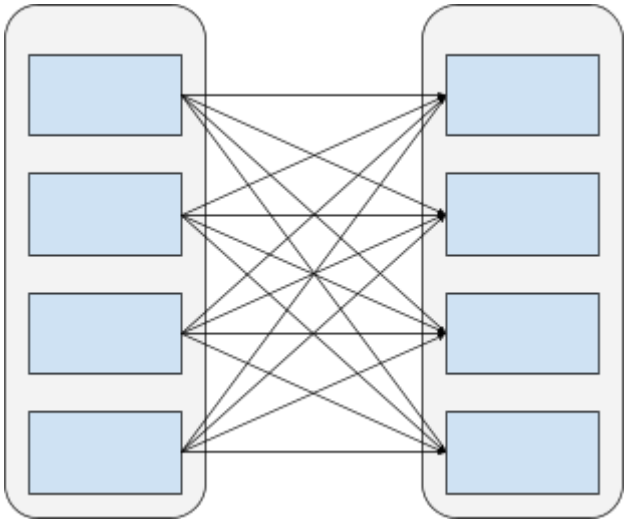
e. Name one transformation that results in a ***narrow*** dependency and draw a DAG showing how this transformation acts on the RDDs and their partitions [5 points]

map



f. Name one transformation that results in a ***wide*** dependency and draw a DAG showing how this transformation acts on the RDDs and their partitions [5 points]

reduceByKey

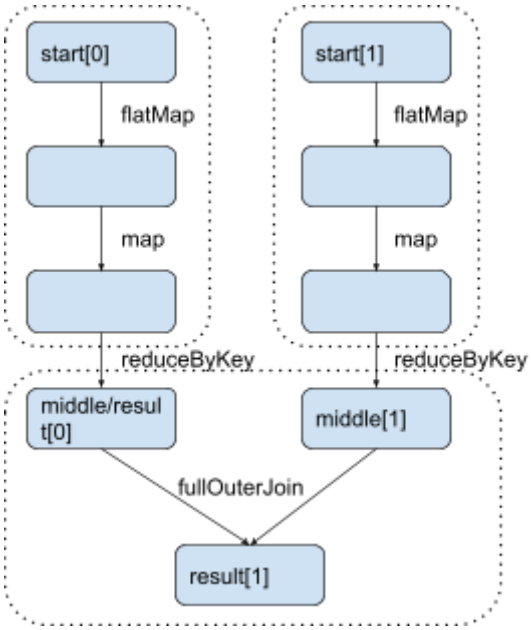


Question 3 - Spark [40 Points]

For Question 3, refer to the following Spark code. Assume that the number of files in the **`files`** list is 2, and that **`fullOuterJoin`** is a transformation resulting in a narrow dependency.

```
1  result = None
2  for file in files:
3      start = sc.textFile(file)
4      middle = start.flatMap(lambda x: x.split(' ')) \
5                    .map(lambda x: (x, 1)) \
6                    .reduceByKey(lambda a,b: a + b)
7
8      if result == None:
9          result = middle
10     else:
11         result = result.fullOuterJoin(middle)
12 result.collect()
```

a. Draw the lineage graph for the RDD **`result`** right before the call to **`result.collect()`**. Include nodes for all intermediate RDDs, even if they are unnamed. [10 points]



b. How many stages will your DAG be broken up into for execution? [10 points]
Draw the stage boundaries in your diagram.
3 stage (see above diagram for boundaries)

c. Identify in the above code (**excluding line 11**) one instance of the following:

i. A transformation that results in a wide dependency [4 points]
Line 6 redueByKey

ii. A transformation that results in a narrow dependency [4 points]
Line 5 map (or line 4 flatMap)

iii. An action [4 points]
Line 12 collect

d. How many "jobs" will the above code run? [4 points]
1 job

e. What algorithm is the above code an implementation of? [4 points]
WordCount

Question 4 - Classifiers [15 Points]

a) Name the three classification algorithms we described in lecture. For each algorithm state whether it is a structural classifier, or statistical classifier. [6 points]

kNN structural
Naive Bayes statistical
Log Regression statistical

b) In the second half of the class we used two motivating examples when discussing classifiers: spam email classification, and ad click probability. State what was used as the features for each example, and give one reason why k-NN may not be effective for these two use cases. [4 points]

Spam - Words were the features
Ad clicks - URLs visited were the features

k-NN would not be effective because it does not work well for large number of features

c) Naive bayes is an exceedingly simple algorithm. Name one reason why it can still be effective for something like spam classification. [2 points]

Simple algorithms can still be very effective when given a large amount of data.

d) How can we adapt binary classifiers to multi-class problems? [3 points]

Run the binary classifier on each class, then pick the class that has the highest probability.

Question 6 - Misc [10 Points]

a) What are the four Vs of data intensive computing? [2 points]
Volume Veracity Variety Velocity

b) Name two sources of uncertainty in data. [2 points]
Randomness due to the collection process
Randomness in the process generating the data

c) Briefly explain why fault tolerance in distributed systems is important. [2 points]
When we run on larger and larger systems, the probability that a fault will occur approaches 100%

d) Name one concrete example of a data intensive application that you interact with on a frequent basis. [2 points]
Google search, almost any social media application, netlfix, amazon (or other recommendation systems), etc.

e) Name two solutions you could employ if the dataset you are working on becomes too big to store/process in a timely manner. [2 points]
Sample the data, or use distributed computing (ie Hadoop/MR/Spark/HIVE/etc)

Scrap Paper

Scrap Paper

Scrap Paper