Apoorva Balasubramanian: 1550424
Rishabh Malhotra: 1526907
Ashish Mahajan: 1526229

PROJECT REPORT

1. OVERVIEW:

Animal shelters in USA have always been working hard to groom and make the animals adoption ready. Yet there are several "*instances*" where despite their valiant efforts the animal is not adopted and has to be transferred or worst put to sleep. Now that the shelters have gone IT and digitalized their data, we have an opportunity to study this data, observe the trends and most importantly can make possible prediction on the outcome for a particular animal coming into the shelter. This will help the shelter identify the weaker ones that have lesser chances of adoption and channelize their efforts on them to improve adoption numbers.

The GOAL of this project is to analyze Animal Shelter Data, observe the existing patterns in the given dataset and predict outcomes for new data by using various machine learning techniques. These insights could help shelters focus their energy on specific animals who need a little extra help finding a new home.

2. ABOUT THE DATASET:

2.1.Animal Shelter Outcomes:

The following dataset is taken from Austin Animal Shelter from October 1st, 2013 to March, 2016 which is used to predict the outcome for each animal.
The data shows that close to 30% of companion animals that come to the adoption shelters end up being unwanted and are euthanized which is the major area of concern.

Following is the snapshot of the dataset:

	A	B	C	D	E	F	G	H	I	J
1	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
2	A671945	Hambone	2/12/2014 18:22	Return_to_owner		Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White
3	A656520	Emily	10/13/2013 12:44	Euthanasia	Suffering	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby
4	A686464	Pearce	1/31/2015 12:28	Adoption	Foster	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White
5	A683430		7/11/2014 19:09	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Cream
6	A667013		11/15/2013 12:52	Transfer	Partner	Dog	Neutered Male	2 years	Lhasa Apso/Miniature Poodle	Tan

2.1.1. SUMMARY:

- *Number of Instances* = 26729
- *Number of Attributes* = 10

2.1.2. ATTRIBUTES:

- AnimalID
- Name
- DateTime
- OutcomeType: represent the status of animals as they leave the Animal Center.

Adoption	Died	Euthanasia	Return to owner	Transfer

- OutcomeSubtype
- AnimalType
- SexuponOutcome
- AgeuponOutcome
- Breed
- Color

2.1.3. PRE-PROCESSING:

We did some pre-processing before classifying with machine learning algorithms:

- DateTime:
  This attribute is spilt into
  - Year
  - Month
  - TimeofDay: Morning, Afternoon, Evening, Night

- SexuponOutcome:
  This attribute is split into:
  - Sex: Male or Female
  - Intact: Neutered, Spayed, Intact
- AgeuponOutcome:
  This attribute is converted into AgeinDays which represents age in terms of days. AgeinDays is further converted to AgeType which classifies an animal as 'baby' or 'adult' depending on age.
- Breed:
  This attribute contains multiple breeds separated by '/' for some instances. The first breed is taken in this case and stored in SimpleBreed.

- Color:
  This attribute contains multiple breeds separated by '/' for some instances. The first color is taken in this case and stored in SimpleColor.
- Attributes that will be considered for training: Year, Month, TimeofDay, OutcomeType, AnimalType, Sex, Intact, AgeType, Breed, Color
- Null Values: Null values in attributes are classified as 'Unknown'

3. MACHINE LEARNING QUESTIONS:

Following are some of the major ML questions that we wanted to answer using this project.

Question 1 →
What could be the potential outcome for a new dog coming to the shelter?

Question 2 →
What could be the potential outcome for a new cat coming to the shelter?

Here are some other interesting questions that could be predicted, but are currently out of the scope of this solution.

Question →
In which season/days are animals more likely to be adopted so that the shelter could make advance plans?

Question →
For a given un-named animal (cat or a dog), does giving the name increase the chances of adoption?

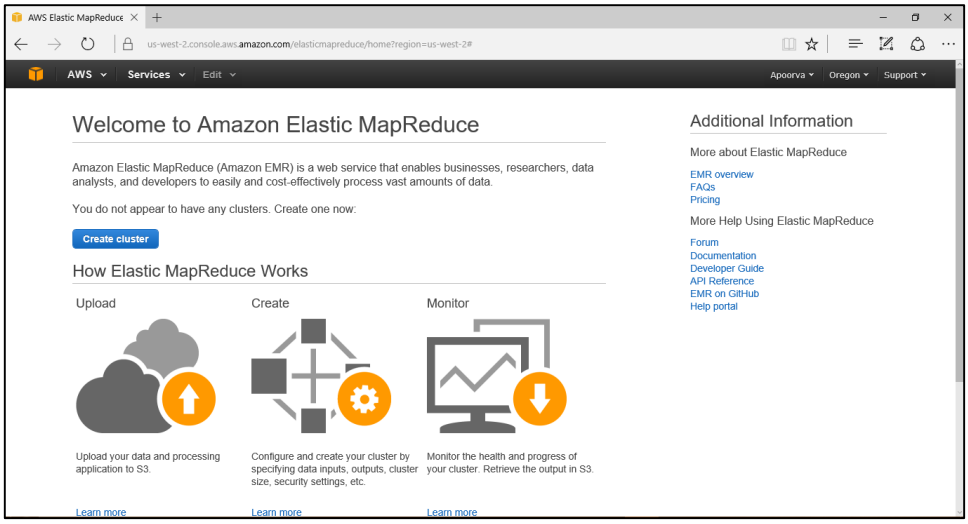
4. Using Hadoop - Map Reduce to ease the analysis

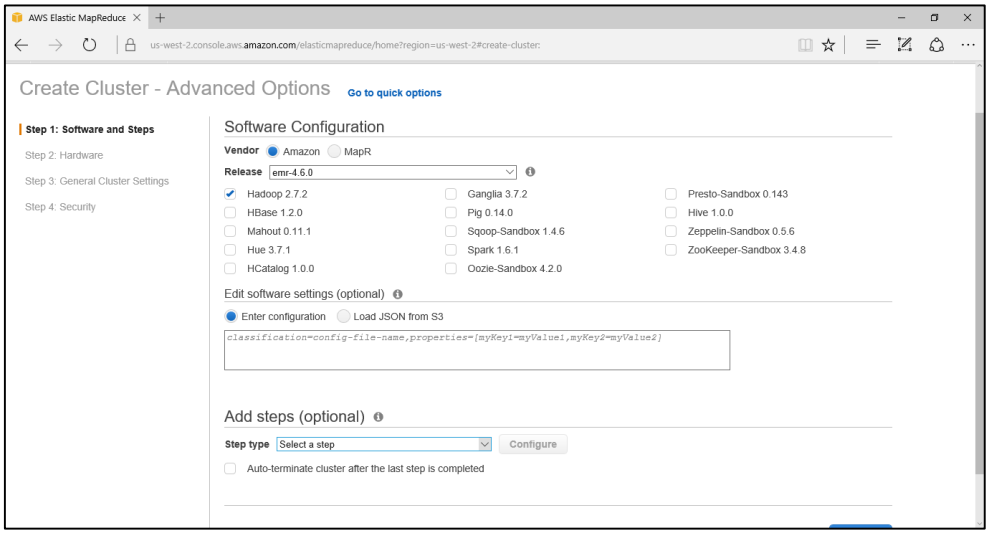
4.1. Hadoop – MapReduce

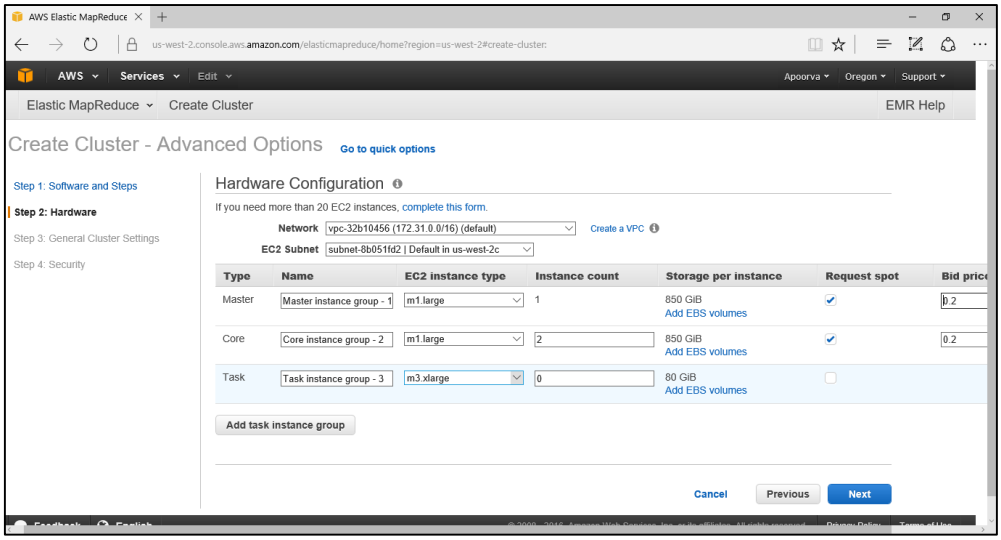
We use Hadoop MapReduce framework for aggregating the large dataset so that we can use the summarized dataset for analysis.

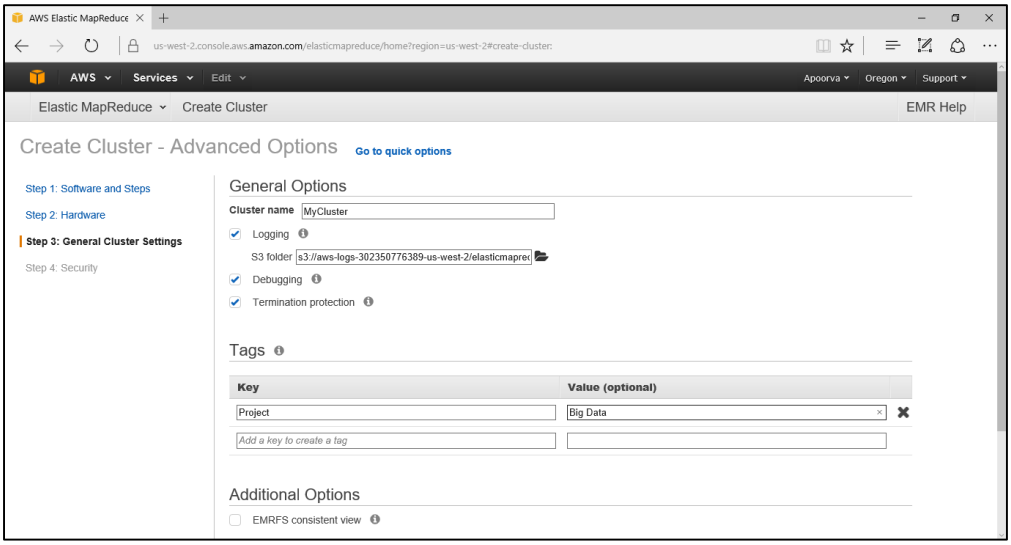
Enabling MapReduce using AWS:

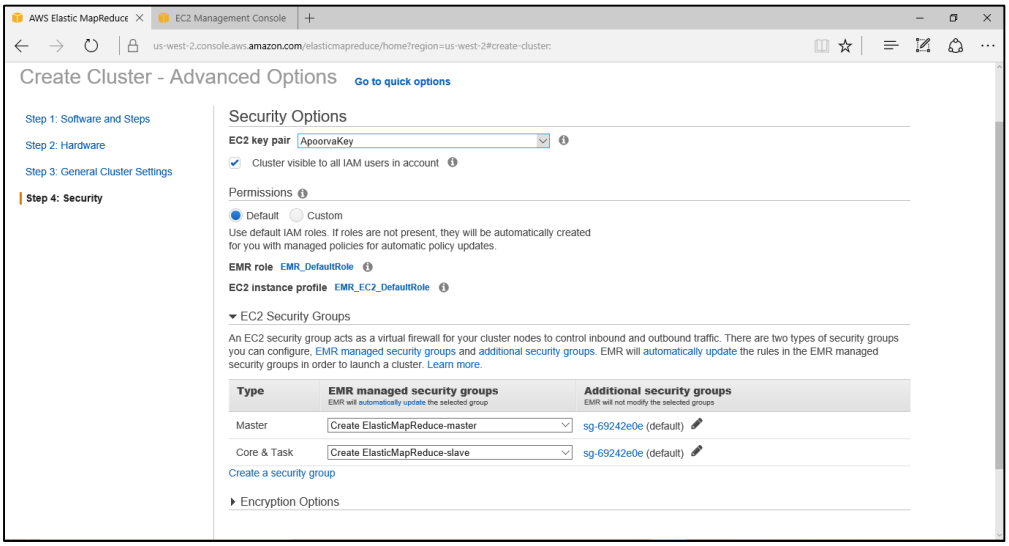
Stepwise process for enabling and running an Amazon Elastic MapReduce job:

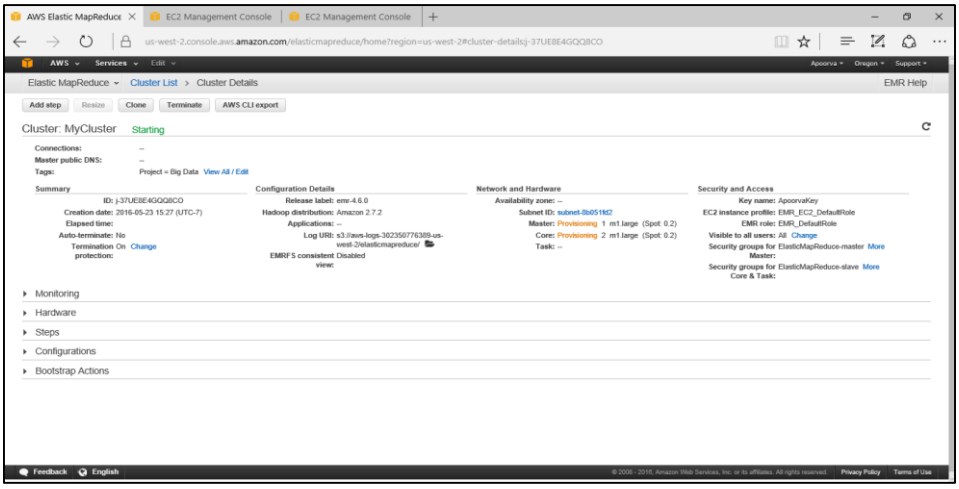

AWS Elastic MapReduce
us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#
AWS
Services
Edit
Apoorva
Oregon
Support
Welcome to Amazon Elastic MapReduce
Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.
You do not appear to have any clusters. Create one now:
Create cluster
How Elastic MapReduce Works
Upload
Upload your data and processing application to S3.
Learn more
Create
Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.
Learn more
Monitor
Monitor the health and progress of your cluster. Retrieve the output in S3.
Learn more
Additional Information
More about Elastic MapReduce
EMR overview
FAQs
Pricing
More Help Using Elastic MapReduce
Forum
Documentation
Developer Guide
API Reference
EMR on GitHub
Help portal

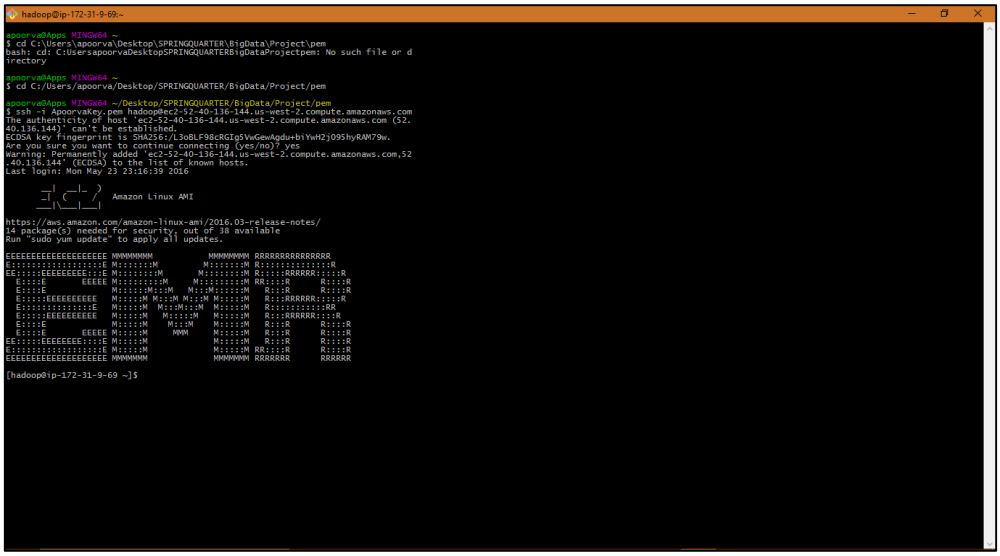

AWS Elastic MapReduce
us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#create-cluster:
Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Software Configuration
Vendor
Amazon
MapR
Release
emr-4.6.0
Hadoop 2.7.2
HBase 1.2.0
Mahout 0.11.1
Hue 3.7.1
HCatalog 1.0.0
Ganglia 3.7.2
Pig 0.14.0
Sqoop-Sandbox 1.4.6
Spark 1.6.1
Oozie-Sandbox 4.2.0
Presto-Sandbox 0.143
Hive 1.0.0
Zeppelin-Sandbox 0.5.6
ZooKeeper-Sandbox 3.4.8
Edit software settings (optional)
Enter configuration
Load JSON from S3
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
Add steps (optional)
Step type
Select a step
Configure
Auto-terminate cluster after the last step is completed

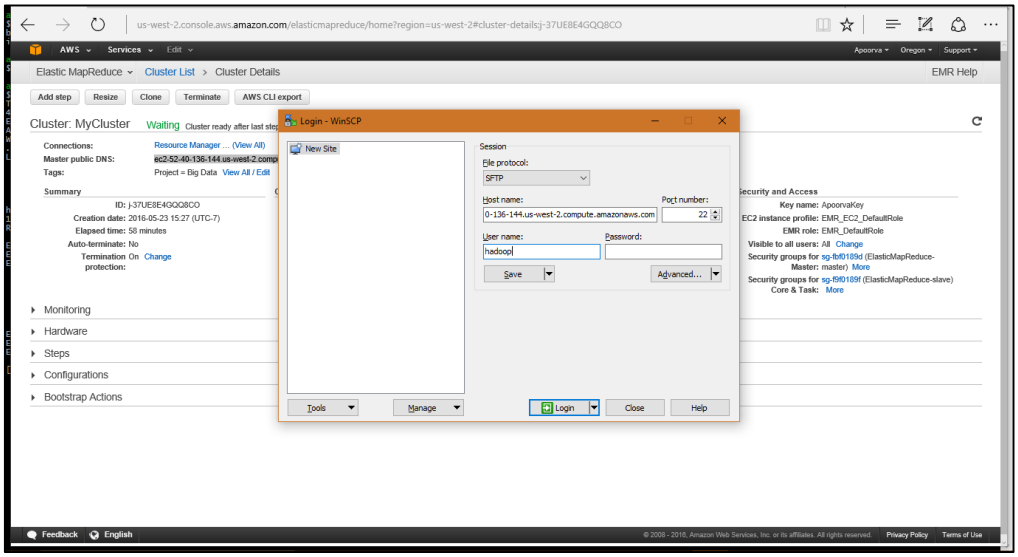

AWS Elastic MapReduce
us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#create-cluster:
AWS
Services
Edit
Apoorva
Oregon
Support
Elastic MapReduce
Create Cluster
EMR Help
Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Hardware Configuration
If you need more than 20 EC2 instances, complete this form.
Network
vpc-32b10456 (172.31.0.0/16) (default)
Create a VPC
EC2 Subnet
subnet-8b051fd2 | Default in us-west-2c
Type
Name
EC2 instance type
Instance count
Storage per instance
Request spot
Bid price
Master
Master instance group - 1
m1.large
1
850 GiB
Add EBS volumes
0.2
Core
Core instance group - 2
m1.large
2
850 GiB
Add EBS volumes
0.2
Task
Task instance group - 3
m3.xlarge
0
80 GiB
Add EBS volumes
Add task instance group
Cancel
Previous
Next

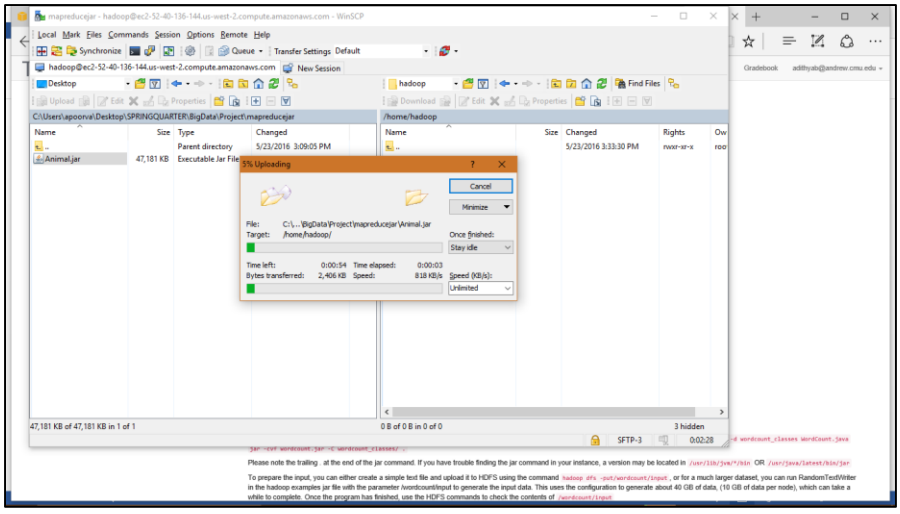

AWS Elastic MapReduce
us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#create-cluster:
AWS
Services
Edit
Apoorva
Oregon
Support
Elastic MapReduce
Create Cluster
EMR Help
Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
General Options
Cluster name
MyCluster
Logging
S3 folder
s3://aws-logs-302350776389-us-west-2/elasticmapred
Debugging
Termination protection
Tags
Key
Value (optional)
Project
Big Data
Add a key to create a tag
Additional Options
EMRFS consistent view

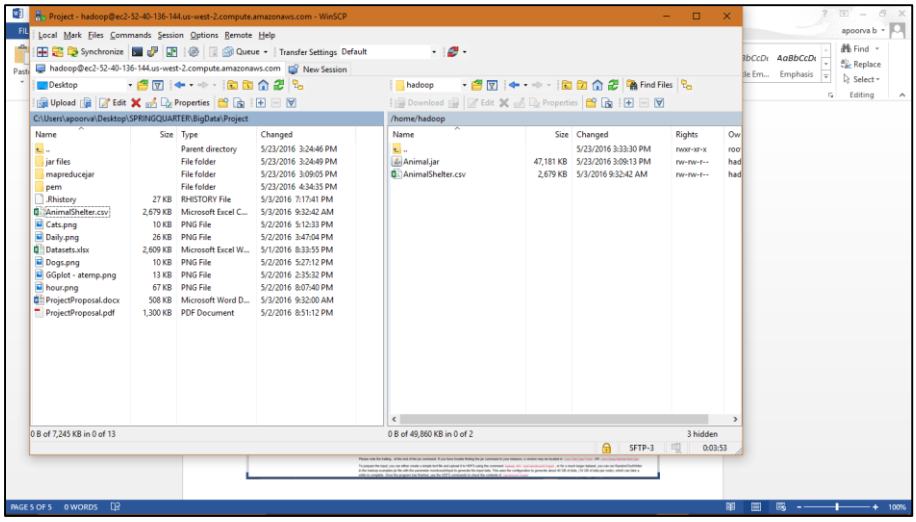

AWS Elastic MapReduce
EC2 Management Console
us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#create-cluster:
Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Security Options
EC2 key pair
ApoorvaKey
Cluster visible to all IAM users in account
Permissions
Default
Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.
EMR role
EMR_DefaultRole
EC2 instance profile
EMR_EC2_DefaultRole
EC2 Security Groups
An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, EMR managed security groups and additional security groups. EMR will automatically update the rules in the EMR managed security groups in order to launch a cluster. Learn more.
Type
EMR managed security groups
EMR will automatically update the selected group
Additional security groups
EMR will not modify the selected groups
Master
Create ElasticMapReduce-master
sg-69242e0e (default)
Core & Task
Create ElasticMapReduce-slave
sg-69242e0e (default)
Create a security group
Encryption Options


AWS Elastic MapReduce
EC2 Management Console
EC2 Management Console
us-west-2.console.aws.amazon.com/elasticmapreduce/home?region=us-west-2#cluster-detailsj-37UE8E4GQQ8CO
Elastic MapReduce
Cluster List
Cluster Details
EMR Help
Add step
Resize
Clone
Terminate
AWS CLI export
Cluster: MyCluster
Starting
Connections:
Master public DNS:
Tags:
Project = Big Data
View All / Edit
Summary
Configuration Details
Network and Hardware
Security and Access
Monitoring
Hardware
Steps
Configurations
Bootstrap Actions
Feedback
English
Privacy Policy
Terms of Use


hadoop@ip-172-31-9-69:~
apoorva@Apps MINGW64 ~
$ cd C:\Users\apoorva\Desktop\SPRINGQUARTER\BigData\Project\pem
bash: cd: C:UsersapoorvaDesktopSPRINGQUARTERBigDataProjectpem: No such file or directory
apoorva@Apps MINGW64 ~
$ cd C:/Users/apoorva/Desktop/SPRINGQUARTER/BigData/Project/pem
apoorva@Apps MINGW64 ~/Desktop/SPRINGQUARTER/BigData/Project/pem
$ ssh -i ApoorvaKey.pem hadoop@ec2-52-40-136-144.us-west-2.compute.amazonaws.com
The authenticity of host 'ec2-52-40-136-144.us-west-2.compute.amazonaws.com (52.40.136.144)' can't be established.
ECDSA key fingerprint is SHA256:/L3oBLF98cRGIg5VwGewAgdu+biYwH2j095hyRAM79w.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-52-40-136-144.us-west-2.compute.amazonaws.com,52.40.136.144' (ECDSA) to the list of known hosts.
Last login: Mon May 23 23:16:39 2016
Amazon Linux AMI
https://aws.amazon.com/amazon-linux-ami/2016.03-release-notes/
14 package(s) needed for security, out of 38 available
Run "sudo yum update" to apply all updates.
[hadoop@ip-172-31-9-69 ~]$


Elastic MapReduce
Cluster List
Cluster Details
EMR Help
Add step
Resize
Clone
Terminate
AWS CLI export
Cluster: MyCluster
Waiting
Login - WinSCP
New Site
Session
File protocol:
SFTP
Host name:
0-136-144.us-west-2.compute.amazonaws.com
Port number:
22
User name:
hadoop
Password:
Save
Advanced...
Tools
Manage
Login
Close
Help
Monitoring
Hardware
Steps
Configurations
Bootstrap Actions


mapreducejar - hadoop@ec2-52-40-136-144.us-west-2.compute.amazonaws.com - WinSCP
Uploading
Cancel
Minimize
Stay idle
Unlimited



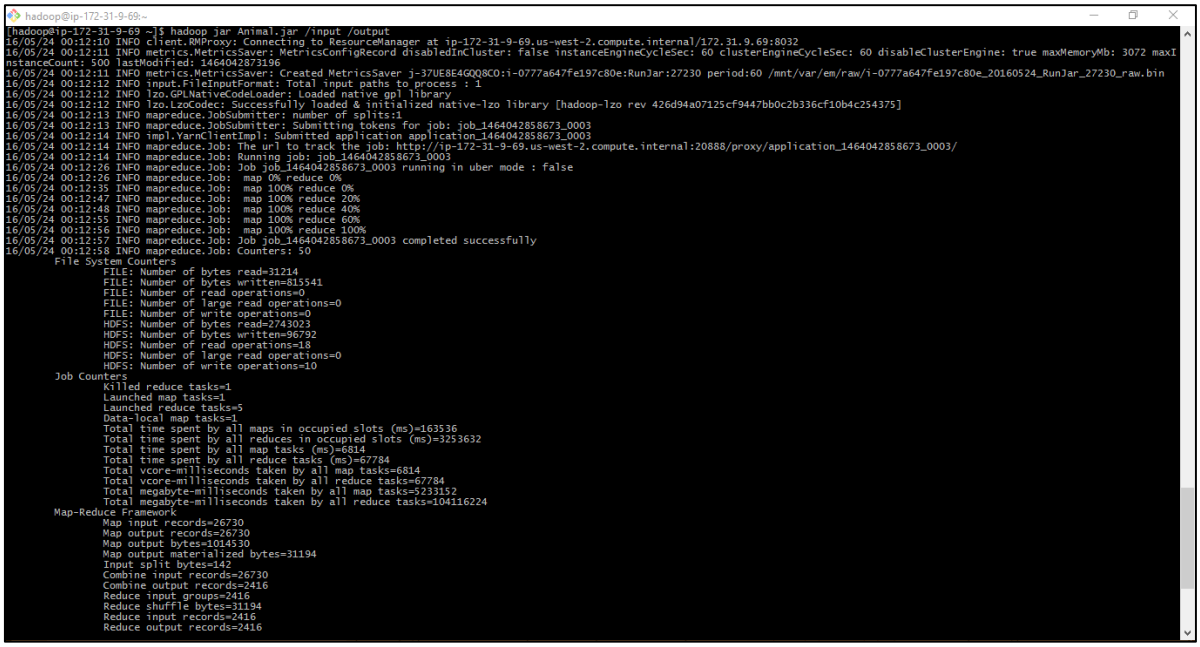
Running the following commands to run our MapReduce program:

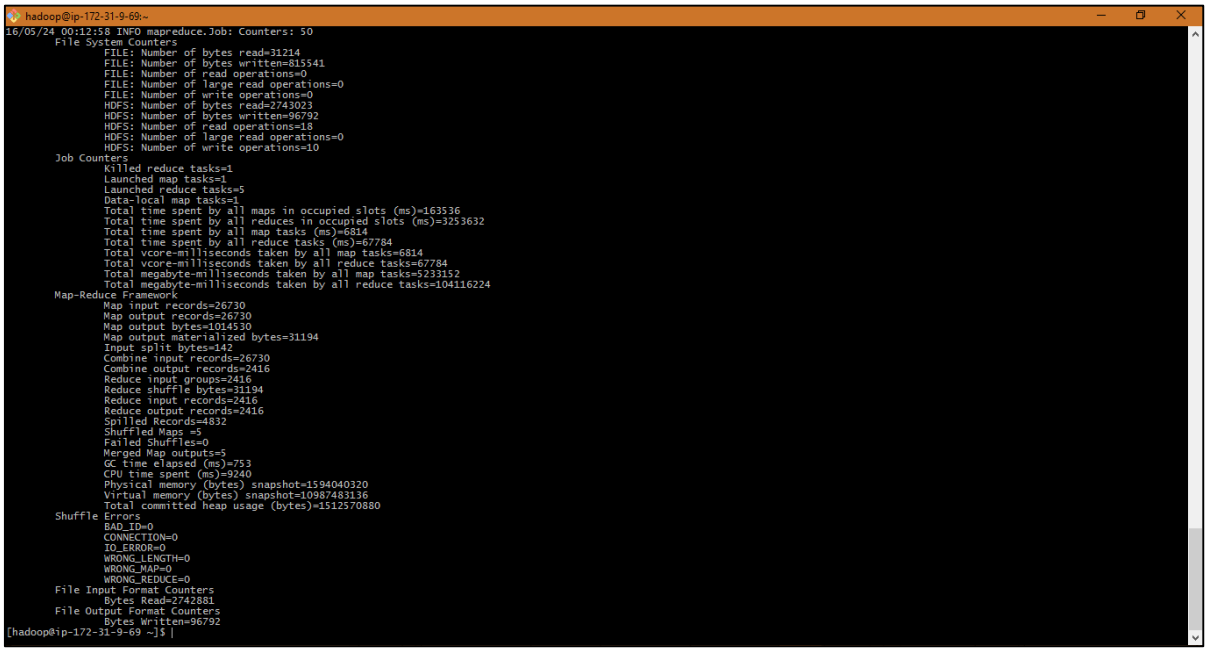
```
[hadoop@ip-172-31-9-69 ~]$ ls -l
total 49864
-rw-rw-r-- 1 hadoop hadoop 48313087 May 23 22:09 Animal.jar
-rw-rw-r-- 1 hadoop hadoop  2742881 May  3 16:32 AnimalShelter.csv
```

Creating an input folder and copying the Animal Shelter dataset:

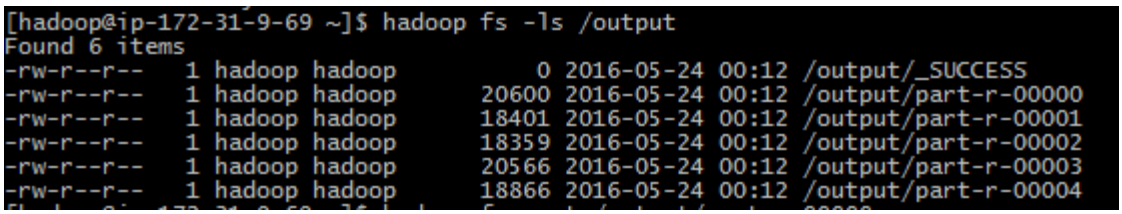
```
[hadoop@ip-172-31-9-69 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hadoop hadoop          0 2016-05-23 22:34 /mnt
drwxrwxrwt   - hdfs   hadoop          0 2016-05-23 22:34 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2016-05-23 22:34 /user
drwxr-xr-x   - hdfs   hadoop          0 2016-05-23 22:34 /var
[hadoop@ip-172-31-9-69 ~]$ hadoop fs -mkdir /input
[hadoop@ip-172-31-9-69 ~]$ hadoop fs -put AnimalShelter.csv /input
[hadoop@ip-172-31-9-69 ~]$ hadoop fs -ls /input
Found 1 items
-rw-r--r--   1 hadoop hadoop    2742881 2016-05-23 23:42 /input/AnimalShelter.csv
```

Running the jar :

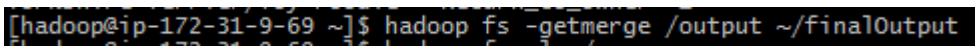




The output:

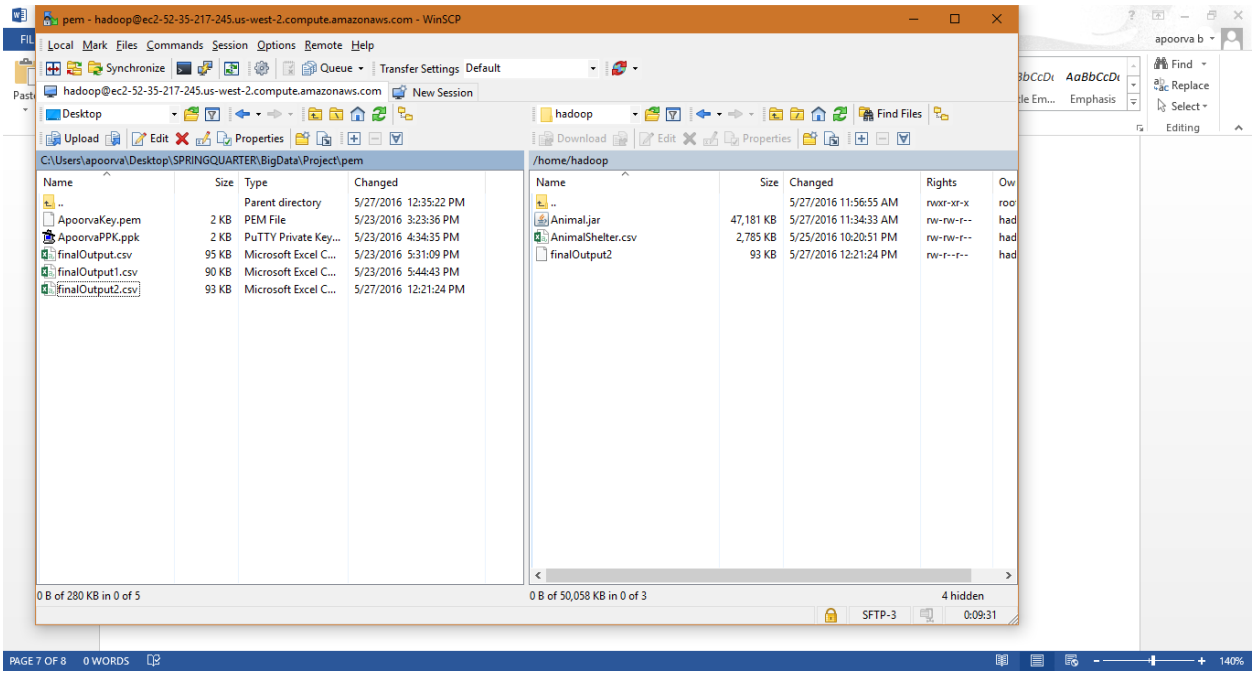


Merging the result:

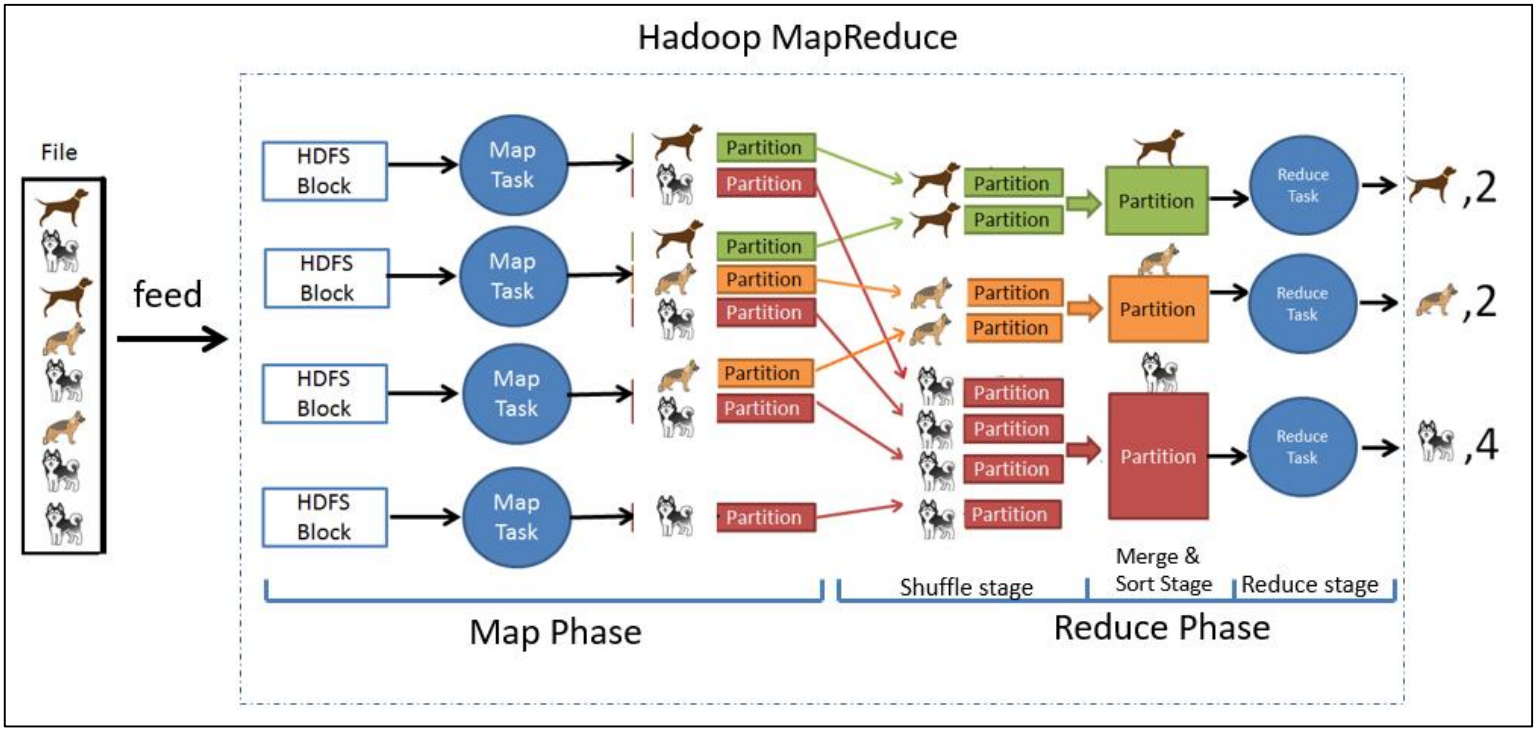


```
[hadoop@ip-172-31-9-69 ~]$ ls -l
total 49964
-rw-rw-r-- 1 hadoop hadoop 48313085 May 23 23:48 Animal.jar
-rw-rw-r-- 1 hadoop hadoop  2742881 May  3 16:32 AnimalShelter.csv
-rw-r--r-- 1 hadoop hadoop    96792 May 24 00:17 finalOutput
```

Transferring to local system:



MapReduce follows the following workflow:-



1. Input Processor: Data from the file is processed into Key/Value pairs which is provided to a Map function.

2. Mapper: The map function outputs a transformed (writable) set of Key/Value pairs, which are subsequently processed by a Reduce function.

```
public static class AnimalMapper extends Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    @Override
    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        String line = value.toString();
        String[] lineArray = line.split(",");
        if (lineArray != null) {
            if (lineArray[8] != null || !lineArray[8].equals("")) {
                    word.set(lineArray[8]+","+lineArray[3]+",");
                    context.write(word, one);
            }
        }
    }
}
```

3. Shuffle and Sort: The next step is the Shuffle and Sort, where the results emitted from each mapper are sorted by key, and partitioned into one of the reducers. This is done when each map task completes to avoid an overload of traffic at the end of the final mapper's operation. This data is written to the local disk and passed into the memory of a waiting reduce task.

4. Reducer: This function is called once for each unique key emitted from the mapper. The Reducer has an iterator for all values for each key. This can be used to aggregate results, and finally returns another output in the desired Output Format.

```
public static class AnimalReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    /**
     * method that iterates all values over a particular key, which is the
     * text assigned in Mapper, and sums up the values
     */

    public void reduce(Text key, Iterable<IntWritable> values, Context context)
            throws IOException, InterruptedException {
        int sum = 0;
        // sum all values by iterating all values over a particular key
        for (IntWritable value : values) {
            sum += value.get();
        }
        context.write(key, new IntWritable(sum));
    }

}
```

In our project, we used to find the outcome of dogs and cats based on Breed type.

	A	B	C
1	Key		Value
2	Abyssinian Mix	Adoption	2
3	Airedale Terrier Mix	Adoption	1
4	Airedale Terrier Mix	Return_to_owner	2
5	Airedale Terrier	Transfer	1
6	Akita	Adoption	2
7	Akita/Labrador Retriever	Transfer	1
8	Alaskan Husky/Australian S	Transfer	1
9	Alaskan Malamute Mix	Return_to_owner	4
10	American Bulldog/Labradc	Transfer	1
11	American Eskimo	Transfer	1
12	American Pit Bull Terrier M	Euthanasia	4
13	American Pit Bull Terrier	Euthanasia	1
14	American Pit Bull Terrier/A	Return_to_owner	1
15	American Pit Bull Terrier/C	Return_to_owner	1
16	American Pit Bull Terrier/F	Adoption	1
17	American Shorthair Mix	Transfer	7

This summarized data can now be easily consumed for further analysis.

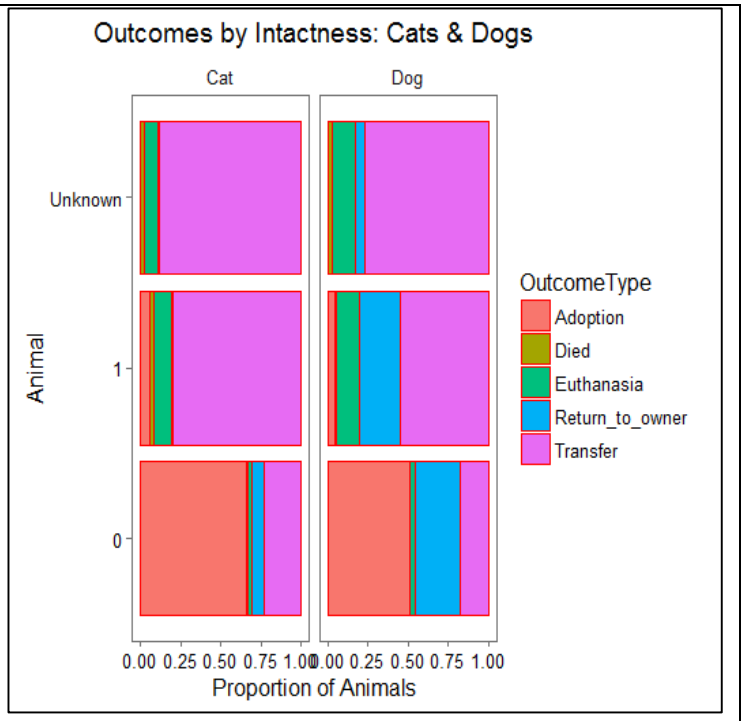
4.2. Resultant Analysis:

Using the resultant aggregated dataset (in a csv format) we can now fit the data into R and perform analysis.

The following points show various analysis made on this dataset:

1. Impact of Intactness on Outcome:

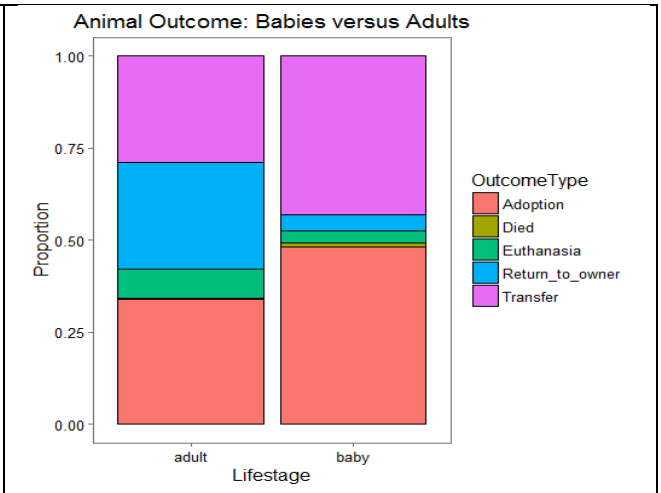
```
# Plot

ggplot(intact, aes(x = Intact, y = num_animals, fill = OutcomeType))
+geom_bar(stat = 'identity', position = 'fill', colour = 'Red')
+facet_wrap(~AnimalType) +
coord_flip() +  labs(y = 'Proportion of Animals', x = 'Animal', title = 'Outcomes by Intactness: Cats & Dogs') +  theme_few()
```



From this graph we can see that adoption rate is more for Spayed or Neutered animals in case of both Cats and Dogs. Intact animals are more likely to be transferred.

2. Baby vs Adult :
   (Age<1 Year is considered as Baby)

```
# Plot

ggplot(train[1:26729, ], aes(x = Lifestage, fill = OutcomeType)) + geom_bar(position = 'fill', colour = 'black') + labs(y = 'Proportion', title = 'Animal Outcome: Babies versus Adults') +theme_few()
```

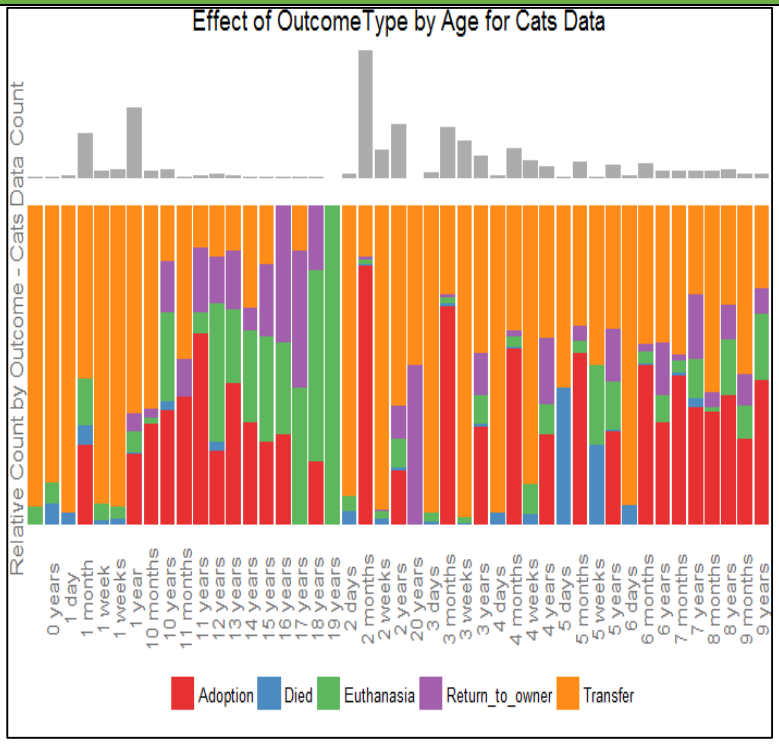


We see that adoptions and transfers are more for baby animals compared to an adult animal. On the other hand, euthanasia is more prominent for adult animal compared to baby animal.

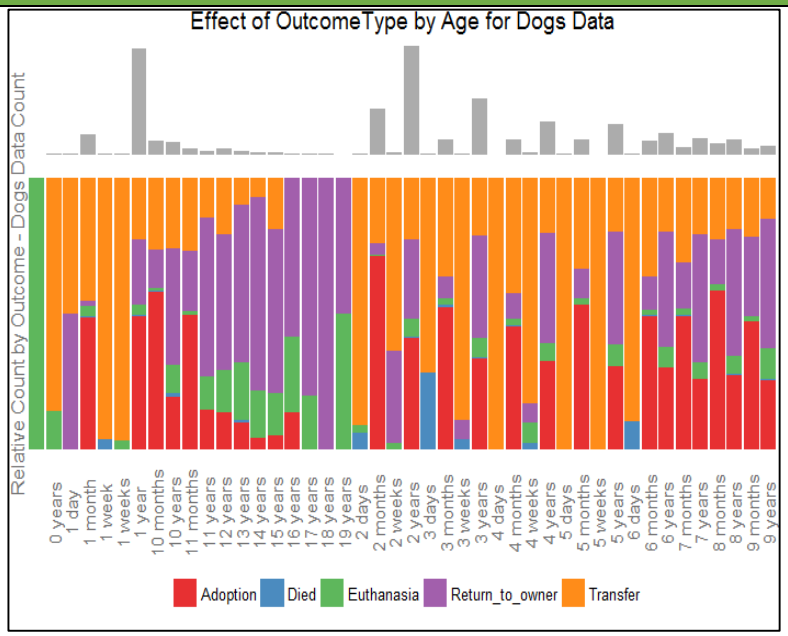
3. Analyzing the impact of age and Animal Type on the Outcome.

CATS

```
#Filtering cat data from the dataset
  > cat_data <- filter(Animal_data, AnimalType == "Cat")

#How does the OutcomeType affect the age upon Outcome
> cat_data %>% count(AgeuponOutcome) %>%
              + ggplot(aes(x = AgeuponOutcome, y = n))
  + geom_bar(stat = "identity", alpha = 0.5)
  + theme_void()
  + ylab("Count")
  +ggtitle("Effect of OutcomeType by Age for Cats Data")
  + theme(axis.title.y = element_text(angle = 90, color = "#73737
  3" )) -> g1
# Relative Count by Outcome – Cats Data
> ggplot(cat_data, aes(x = AgeuponOutcome, fill = OutcomeType)
)
+ geom_bar(stat = "count", position = "fill",  alpha = 0.9)
+ scale_fill_brewer(palette = "Set1")
+ theme_void()
+ ylab("Relative Count by Outcome") +
+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 1
1, color = "#737373"), legend.position = "bottom",axis.title.y = el
ement_text(angle = 90, color = "#737373")) -> g2
# To bring up the above 2 Graphs on the same page
> grid.arrange(g1, g2, ncol = 1, heights = c(1,3))
```



OBSERVATION- Age does not seem to have much impact on adoption rates (unlike for dogs – see below).

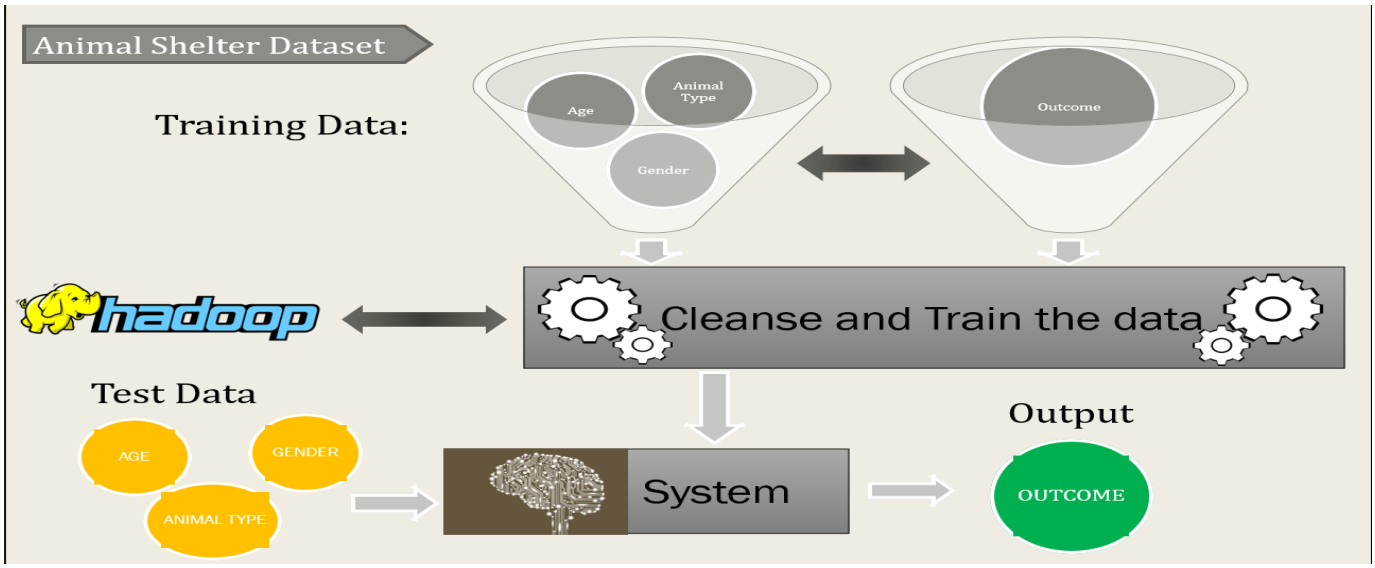
DOGS	
Filtering cat data from the dataset > dog_data <- filter(Animal_data, AnimalType == "Dog") #How does the OutcomeType affect the age upon Outcome > dog_data %>% count(AgeuponOutcome) %>% + ggplot(aes(x = AgeuponOutcome, y = n)) + geom_bar(stat = "identity", alpha = 0.5) + theme_void() + ylab("Count") +ggtitle("Effect of OutcomeType by Age for dogs data") + theme(axis.title.y = element_text(angle = 90, color = "#7373 73")) -> g3 # Relative Count by Outcome – Dogs Data > ggplot(dog_data, aes(x = AgeuponOutcome, fill = OutcomeTy pe)) + geom_bar(stat = "count", position = "fill", alpha = 0.9) + scale_fill_brewer(palette = "Set1") + theme_void() + ylab("Relative Count by Outcome") + + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 11, color = "#737373"), legend.position = "bottom", axis.title.y = element_text(angle = 90, color = "#737373")) -> g4 # To bring up the above 2 Graphs on the same page > grid.arrange(g3, g4, ncol = 1, heights = c(1,3))	

OBSERVATION-

1. The adoption rate is very high for young dogs and puppies. As the age increases the adoption rate falls.
2. Older dogs have higher chances of getting returned to owner.

4.3. SOLUTION ARCHITECTURE & ML TECHNIQUES:

The following diagram shows the solution architecture for predicting Animal Shelter Outcome for the given Dataset:

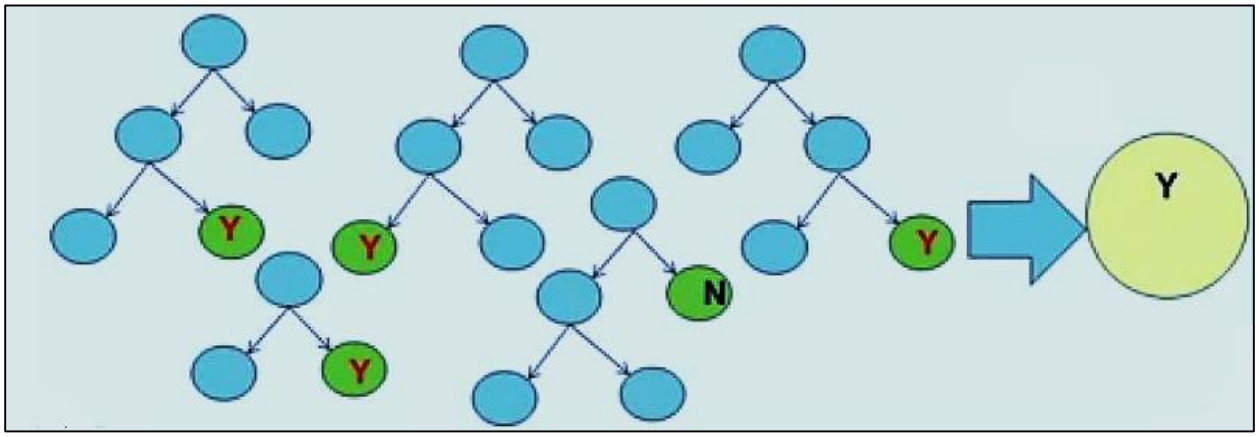


We shall be taking the animal type, age, gender, breed and other suitable attributes as input data which will give us the outcome as output. We will be training this data using *Random Forest algorithm* to predict the outcome. We will use Hadoop for aggregating Raw data for analysis.

5. MACHINE Learning Techniques and Predictions:

- Using Random Forest

A random forest is an ensemble of decision trees which will output a prediction value, in this case "outcome type". Each decision tree is constructed by using a random subset of the training data.



We then consolidate the outcome of each decision tree and choose the most popular outcome

After training the data, we can then pass test data through it, in order to output a prediction.

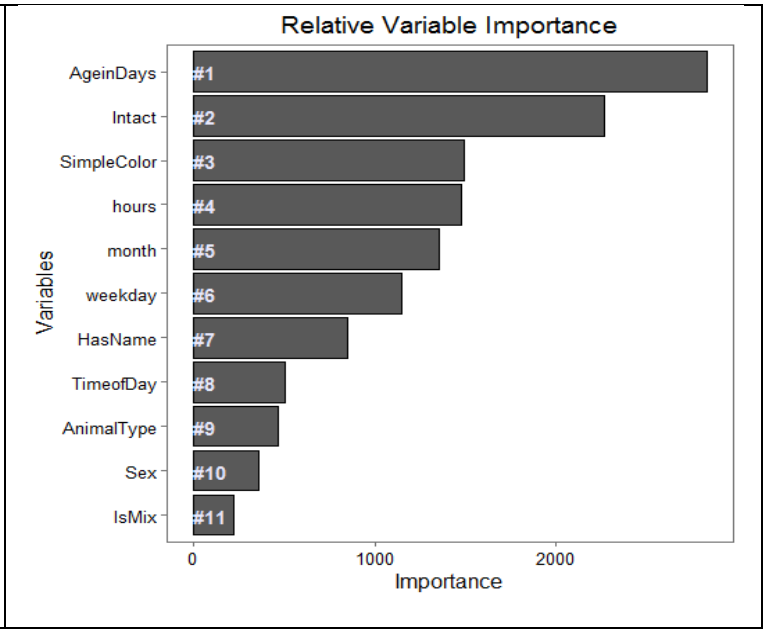
```
# Build the model

rf_mod <- randomForest(OutcomeType ~
AnimalType+AgeinDays+Intact+HasName+hours+weekday+TimeofDay+SimpleColor
+IsMix+Sex+month, data = train, ntree = 600, importance = TRUE)

# Predict using the test set

prediction <- predict(rf_mod, test, type = 'vote')
```

This code outputs:

ID	Adoption	Died	Euthanasi	Return_to	Transfer
A675805	90%	0%	0%	1%	9%
A666170	97%	0%	0%	0%	3%
A669394	57%	0%	1%	18%	24%
A648948	26%	0%	2%	69%	4%
A683385	100%	0%	0%	0%	0%
A719068	2%	0%	1%	2%	95%
A684455	1%	0%	3%	73%	23%

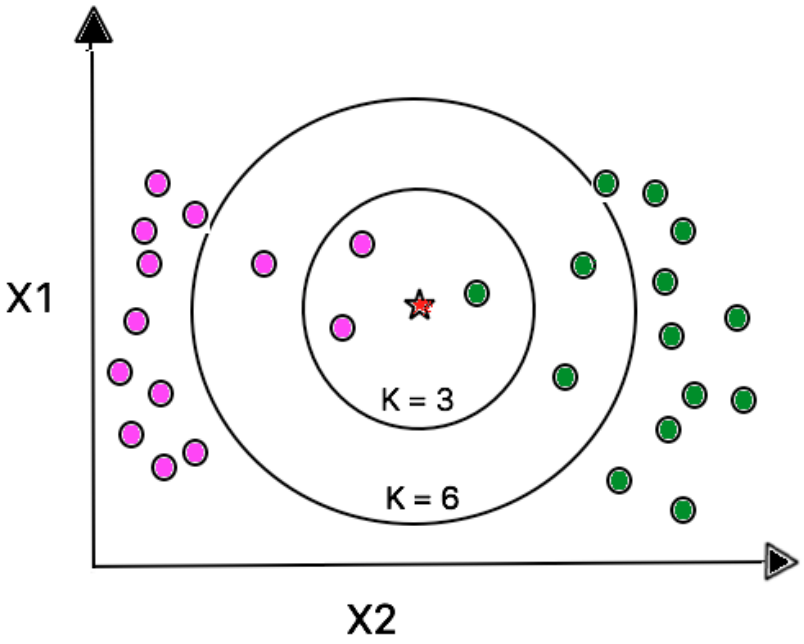
We also calculated the ***relative variable importance*** (which variable has played a major role in predicting the outcome):

```
# Get importance

importance    <- importance(rf_mod)

varImportance <- data.frame(Variables
= row.names(importance), Importance
= round(importance[
,'MeanDecreaseGini'],2))
```



- Using K- Nearest Neighbor

Machine learning algorithms provide methods of classifying objects into one of several groups based on the values of several explanatory variables. Nearest neighbor methods are easily implemented and easy to understand. There is no model associated to them, so errors have to be estimated computationally, but it provides one simple solution to classifying a new object based on known results in a reference set. kNN is a generalization of "if it walks like a chicken, looks like a chicken, and talks like a chicken, it is probably a duck." That is, objects that a close together with respect to the explanatory variables are likely to have the same classification.



In the simplest setting, like the example we will do here, objects can fall into one of two classes, \(A \) or \(B \). We have a set of \(n \) measurements, \(v_1, \dots, v_n \), of any object in question, and assume these are all numeric variables. We use a distance measure, namely Euclidean distance, to manifest when two objects are close with respect to these variables. So, given objects in the domain \(s \), \(r \), which have \(v_i \) measurements \(x_1, \dots, x_n \), \(y_1, \dots, y_n \), respectively, define \[dist(s, r) = \sqrt(\sum_i (x_i-y_i)^2). \] Depending on characteristics of the variables, other distance measures may be more appropriate, but we'll stick with Euclidean distance.

The kNN algorithm begins with a training set of objects for which we know not only the values of the explanatory variables but also the classifications \(A \) and \(B \). To predict the classification of a new object \(q \), the \(k=1 \) version of kNN would proceed by finding the element of the training set with the minimum distance from \(q \), suppose it is \(p_1 \), and predict the classification of \(q \) to be the same as \(p_1 \). If \(p_1 \) is

rather isolated and there are lots of points in the other class almost as close to \(q \) this could be a misclassification. So, we generally pick some larger odd number \(m \) for \(k \); find \(p_1, \dots,p_m \) closest to \(q \) and vote on what the classification of \(q \) should be. A rule of thumb in machine learning is to pick \(k \) near the square root of the size of the training set. In practice this does a good job of telling signal from noise.

This outputs:

ID	Adoption	Died	Euthanasies	Return_to	Transfer
A675805	92%	0%	0%	1%	7%
A666170	98%	0%	1%	0%	1%
A669394	64%	0%	0%	15%	21%
A648948	24%	0%	1%	70%	5%
A683385	8%	0%	0%	5%	87%

6. Analysis:

We have found out that KNN is having good accuracy on the training dataset. We understood that it is because we have all the attributes as categorical and we have one attribute with 382 levels and one with 57 levels after preprocessing. In this scenario, as we are testing on large dataset containing 26729 rows and we found that KNN performs better with our dataset. Random Forest is having comparatively lesser performance because of more levels for the attributes.

7. Conclusion:

Adoption is the most likely outcome for the animals in the given dataset and Euthanasia is the least likely. Hence, we can see that the predictions are analogous to the real world scenario where Euthanasia is uncommon.

8. References:

1. https://cran.r project.org/web/packages/gridExtra/vignettes/arrangeGrob.html
2. https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf
3. http://www.sthda.com/english/wiki/ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page-r-software-and-data-visualization
4. https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
5. https://www.kaggle.com/c/shelter-animal-outcomes/data