Analysis and Estimation of Endangered Wildlife Species for Protection by using Big Data Techniques

DIVYANSH ARYA

VIVEK SINGHAL

ASHWANI SINGHAL

Abstract:

Big Data Analytics is a process to uncover some hidden patterns and information by applying big data techniques, by using this technique we can analyze all the data and can get the significant value from it. We know that in current scenario data can be in any form like in the form of some written text, in the from of audio files or in the form of images and it could be of any form so by applying modern big data techniques like Hadoop, MapReduce and No SQL database we can store and process these data very efficiently. Nowadays big data is very important, it is very helpful for the business organizations because by using this technique they can understand the customer needs better, they can understand the new market strategies, by using this technique they can take help from social media to understand the customer behaviour better. Now coming to the benefit of big data in wildlife conservation and protection, we all know that wildlife plays an important role in balancing the environment and provides stability to different natural processes of nature and wildlife conservation is the practice of protecting wild plant and animal species and their habitats., so by keeping all these things in mind, by using Hadoop, machine learning and No SQL databases we will predict which endangered species will be able to recover in how much time period with the help of data of each species. With this project environmentalists will be able to know and plan the steps to take for the coming years for conservation and will also be able to get the information whether the species is sustainable in an environment or not.

1. Introduction

Basically there are 5 foundational V’s of big data which help us to understand it better -volume: the amount of data is stored, velocity: the speed of data and processing, variety: the different types of data like structured, unstructured and semi structured, veracity: This context is equivalent to quality. We have all the data, but are we missing something? is this data “clean” and accurate? Do they really have something to offer? and the last is value: This refers to the ability to transform a tsunami of data into business. As we know the time is changing rapidly and so as environment. Environment is badly effected due to global warming. Water bodies have become polluted. Pollution gives birth to acid rain and we know water is very essential for everyone including humans and animals. Therefore, in this so-called modern environment we cannot rule out the dangerous state that many species have gone into. Many species around the globe are endangered and some very close to extinction. Government is also taking so many initiatives to improve this critical emergency. It is making use of data, but the actions and decisions can be improved if we know the time it would take to save each species provided that necessary steps are taken. This will set an approximate deadline that the conservation programs will ensure to achieve to save a species. Natural life and forests depend much on wildlife and its proper balance. Some decades this balance was pretty much maintained but as the world grows and development activities expand, animals are losing their habitat and lacking food. Causing them to starve to death or get weak enough to become someone's prey. So, we try and use the existing data about species and predict what length of time will be required to save a species from extinction. Now since number of species is large then the dataset will also be very large and with traditional or old techniques it is not possible to analyse the data accurately and that will take time also so, here big data comes into role, as we know big data is a technique to process and find hidden patterns from large data set. So, with the help of this modern technique we can analyze and process large datasets very efficiently.

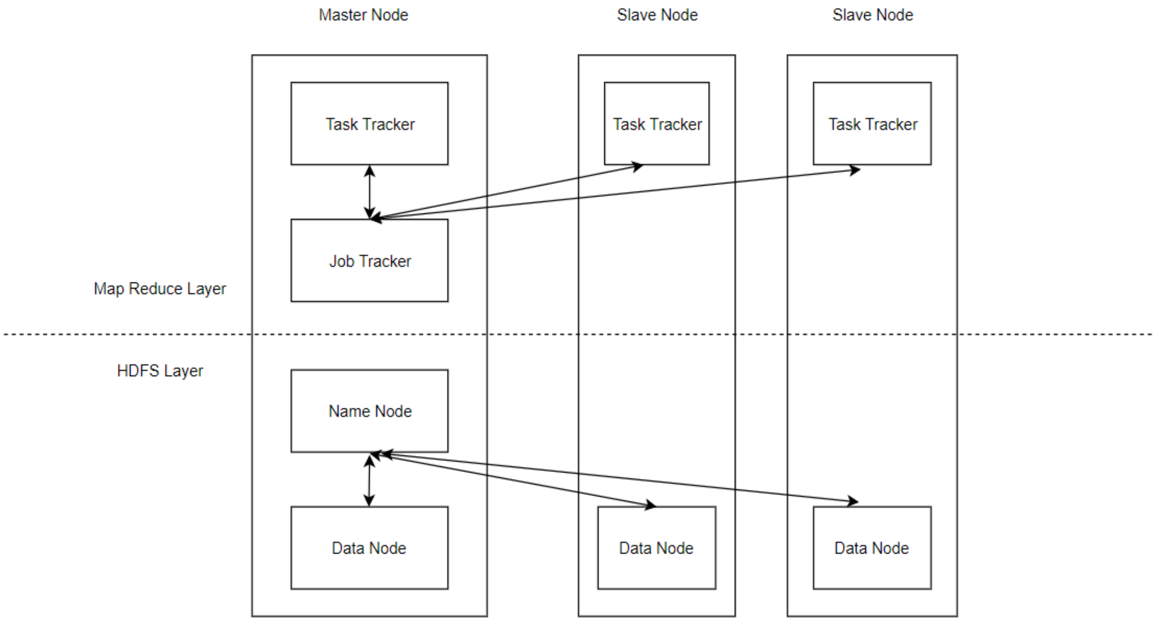
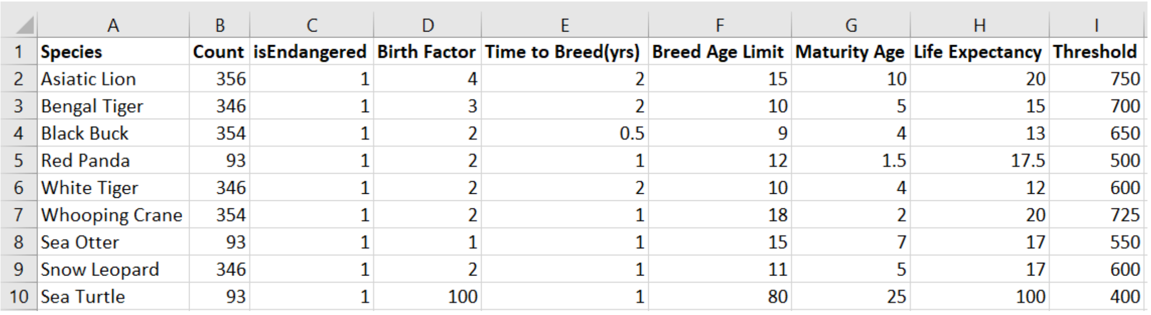


Figure:1.1 Hadoop Architecture

2. Literature Review

In modern world amount of users are very much and there are huge amount of data exist so, normal techniques are not useful to process huge amount of data so here bigdata comes into picture, bigdata is more real in comparison to another techniques for processing huge data. Massively data processing, scale out architectures are unit compatible for big data applications. There are lot of datasets available on the internet and all the information that provides the required values to our project's specifications. There are different datasets available for each type of species , based on geographical location, endangerment status, Class of animals, size and type of data also. Data provided can be private, public, research or surveyed. For technical and scientific feasibility, we created our own sample dataset to depict the working of our model and display how it uses the data provided to predict the conservation time of the species. The wildlife is changing at a critical rate and needs quick actions to be taken for its preservation and hence calls for urgent interventions at every level. We have collected the dataset from the various website. Our dataset consists of the data of species name and the attributes we require in order to predict the outcome. Further each species has its own dataset which contains more information about it like number of males, females, female age array, species total count.

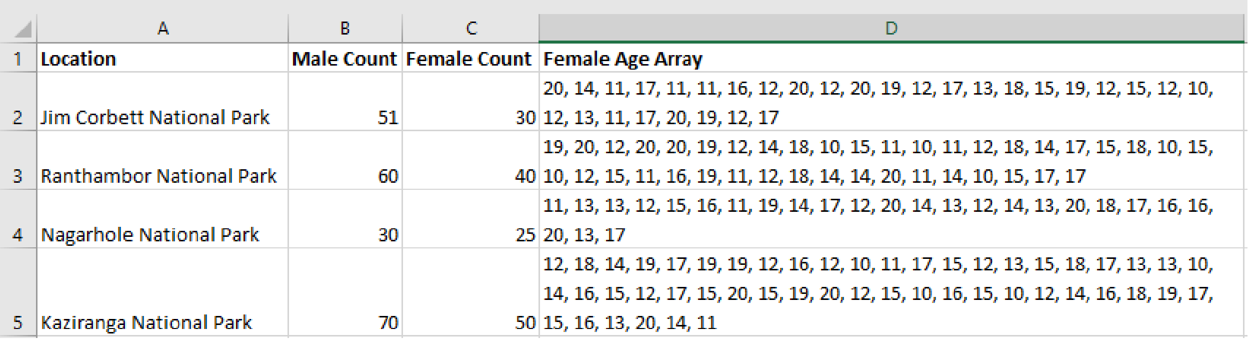


	A	B	C	D	E	F	G	H	I
1	Species	Count	isEndangered	Birth Factor	Time to Breed(yrs)	Breed Age Limit	Maturity Age	Life Expectancy	Threshold
2	Asiatic Lion	356	1	4	2	15	10	20	750
3	Bengal Tiger	346	1	3	2	10	5	15	700
4	Black Buck	354	1	2	0.5	9	4	13	650
5	Red Panda	93	1	2	1	12	1.5	17.5	500
6	White Tiger	346	1	2	2	10	4	12	600
7	Whooping Crane	354	1	2	1	18	2	20	725
8	Sea Otter	93	1	1	1	15	7	17	550
9	Snow Leopard	346	1	2	1	11	5	17	600
10	Sea Turtle	93	1	100	1	80	25	100	400

Figure: 2.1 Sample Species Dataset

One of the most important aspect of analyzing a data requires cleaning of dataset. We took help from many datasets and removed the columns which created redundancy and only included those which were required. We used "pandas" which is a famous library of python that takes data (like a CSV or TSV file, or a SQL database) and creates a python object with rows and columns called data frame that looks very similar to table in a statistical software. It is mainly used for data manipulation and data analysis.

We used group by function of pandas to group species and calculate mean of their selected attributes. Further we added a new column named as “Threshold” which shows the target count of species we wish to achieve.



	A	B	C	D
1	**Location**	**Male Count**	**Female Count**	**Female Age Array**
2	Jim Corbett National Park	51	30	20, 14, 11, 17, 11, 11, 16, 12, 20, 12, 20, 19, 12, 17, 13, 18, 15, 19, 12, 15, 12, 10, 12, 13, 11, 17, 20, 19, 12, 17
3	Ranthambor National Park	60	40	19, 20, 12, 20, 20, 19, 12, 14, 18, 10, 15, 11, 10, 11, 12, 18, 14, 17, 15, 18, 10, 15, 10, 12, 15, 11, 16, 19, 11, 12, 18, 14, 14, 20, 11, 14, 10, 15, 17, 17
4	Nagarhole National Park	30	25	11, 13, 13, 12, 15, 16, 11, 19, 14, 17, 12, 20, 14, 13, 12, 14, 13, 20, 18, 17, 16, 16, 20, 13, 17
5	Kaziranga National Park	70	50	12, 18, 14, 19, 17, 19, 19, 12, 16, 12, 10, 11, 17, 15, 12, 13, 15, 18, 17, 13, 13, 10, 14, 16, 15, 12, 17, 15, 20, 15, 19, 20, 12, 15, 10, 16, 15, 10, 12, 14, 16, 18, 19, 17, 15, 16, 13, 20, 14, 11

Figure 2.2 Asiatic Lion Dataset Sample

Data visualization is very important to represent the features of data in graphical form to understand complicated relationship in data. Standardizing the data is essential need before visualization. We used MinMaxScaler of sci-kit learn library to standardize the dataset. We plotted bar graphs using matplotlib library.

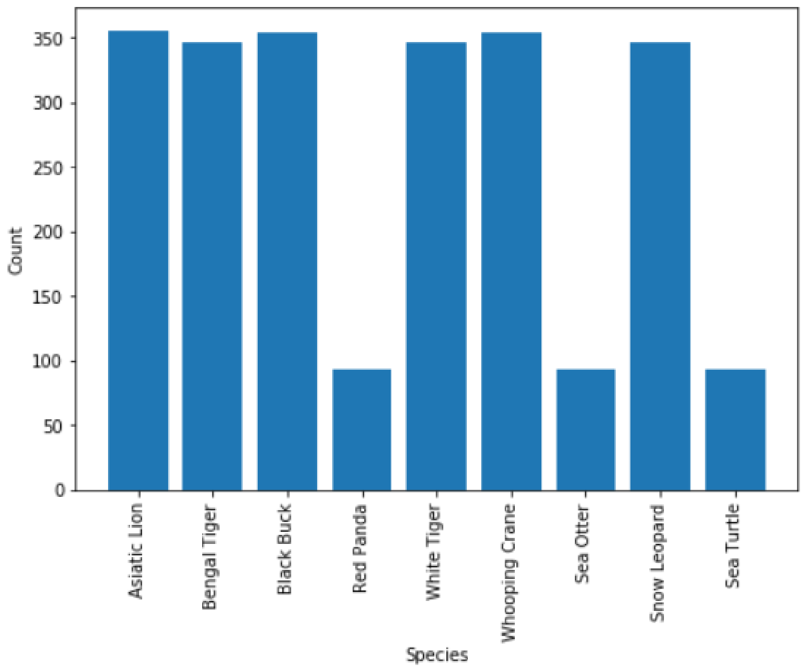


Figure 2.3 Species Name v/s Count

The trend of social media is increasing rapidly. Almost all the peoples are active on social media platforms like Facebook, Instagram, twitter etc. On these platforms we post text messages, images, audio files, video files. More than millions of people use these platforms it means more than billon of messages and to store these messages there is a database which is present. These messages can be stored in RDBMS. In this project first we imported XLSX file into python jupyter notebook and cleaned it. Then saved the cleaned data as our new dataset.

3. Future Prospective

The time is changing, life has come to smart phones and tablets from desktops and laptops and everyone own smart phones. As of now this application will be available for wildlife workers as a windows application but the android application of this project will also be made by which some problems like portability will get reduced and will be easy to operate. Android application should be based on location based service, by this service workers will do less efforts, they won't need to type or select anything, they will just open up the application and on the basis of their location and choices data will be fetched from database and results will get display.

In future, on this project MLlib will be applied which is a apache spark's scalable machine learning library by using this library many algorithms like regression, clustering, collaborative and classification will get implemented on this application by which it will give better results and will predict more results.

References:

[1] http://www.animalinfo.org/species

[2] A web platform and that index research Datasets and Conservation Projects on Wildlife: https://www.systemanaturae.org/datasets/