MANIPAL INSTITUTE OF TECHNOLOGY
(A constituent Institute of Manipal University)

MANIPAL - 576 104, KARNATAKA, INDIA
MARCH 2017

Rainfall Analysis and Rainstorm Prediction

By

Devansh Ojha
140911314

Devanshi Desai
140911372

Under the Guidance of

Mrs. Anju R
Mrs. Vibha Poora

I. INTRODUCTION

Big data is collection of huge volumes of data that contains both the structured and unstructured data that is difficult to store analyze process, share, visualize and manage with the traditional database and software techniques. Volume of data can be calculated by the amount of transactions. Due to growth increasing need on platforms and in many software industry applications to handle the scalability, accuracy, rate at which enterprises remain to face in a competitive global Market world. Major Big Data challenges are capturing data, storage, transfer, searching, analysis, transfer, presentation. Along with traditional transactional and analytics data stores, we now collect additional data across social media activity, web server log files, financial transactions and sensor data from equipment in the field.

Rainfall data is collected to predict the storm warnings from the hydrological data. This is considered as a research idea as it consumes huge number of records from the distributed system. We aim to manage the data based on spatial temporal characteristics using a Map Reduce Framework. The workload is classified using Naïve Bayes (NB) classifier. The classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Various rainstorm concept prediction is achieved using the big raw rainfall data.

II. PROBLEM DEFINITION

Because of the influence elements of weather is very complicated, Until now the weather forecasting result especially the rain forecasting in a long time is not good enough, it is based on the calculation and prediction, and also the experiences of reporter is playing an important role in it. .Data Ming knowledge are also used in the weather forecasting problem, probabilistic graphical models (Bayesian networks) in Meteorology as a data mining technique. Bayesian networks automatically capture probabilistic information from data using directed acyclic graphs and factorized probability functions.

III. METHODOLOGY

Rainfall data is collected to predict the storm warnings from the hydrological data. This is considered as a research idea as it consumes huge number of records from the distributed system. We aim to manage the data based on spatial temporal characteristics using a Map Reduce Framework. The workload is classified using Naïve Bayes (NB) classifier. The classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Various rainstorm concept prediction is achieved using the big raw rainfall data.

Naive Bayesian is one of the most effective and efficient classification algorithms. Bayesian Theorem is a theorem of probability theory originally stated by the Reverend Thomas Bayesian. The theorem assumes that the probability of a hypothesis is a function of new evidence and previous knowledge. It can be seen as a way of understanding how the probability that a theory is true is affected by a new piece of evidence. It has been used in a wide variety of contexts, ranging from marine biology to the development of "Bayesian" spam blockers for email systems. Naive classifiers have several desirable features: First, they are simple to construct, requiring very little domain background knowledge, as opposed to general Bayesian networks which can require numerous intensive sessions with experts to produce the true dependence structure between features. Second, naive networks have very constrained space and time complexity. Bayesian theorem provides a way to calculate the probability of a hypothesis, here the event Y, given the observed training data, here represented as X (|)() (|) ( ) p X Y pY pY X p X = (1) This simple formula has enormous practical importance in many applications. It is often easier to calculate the probabilities, P(X | Y), P(Y), P(X) when it is the probability P(Y | X) that is required. This theorem is central to Bayesian statistics, which calculates the probability of a new event on the basis of earlier probability estimates derived from empirical data.

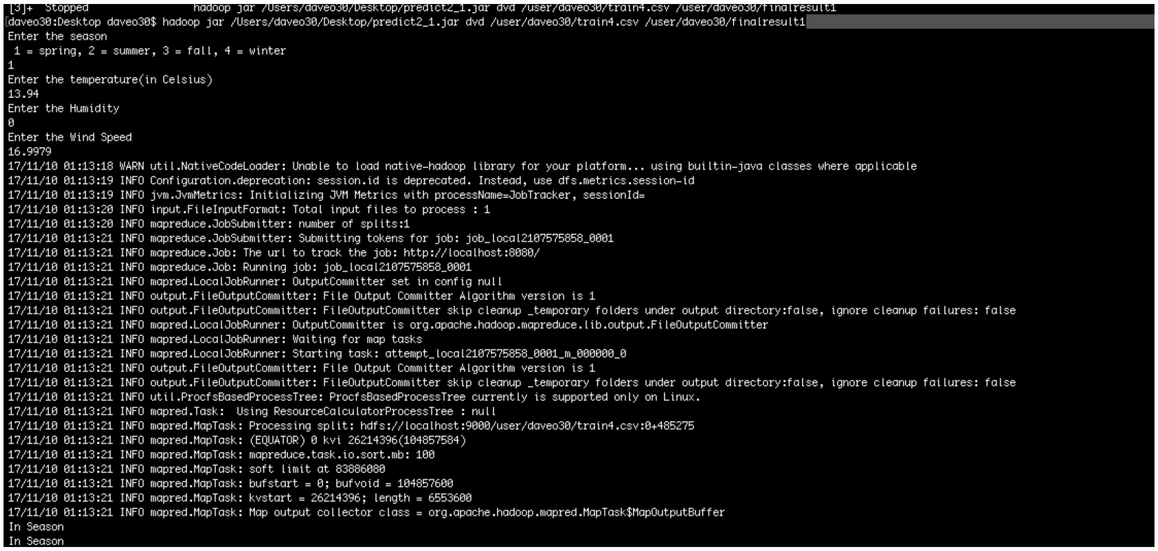
MapReduce is a programming model and an associated implementation for processing and generating rainfall data. Naïve Bayes is used as it is simple to build and fast to make decisions. It efficiently accommodates new data by changing the associated probabilities. The Bayes theorem is central to Bayesian statistics, which calculates the probability of a new event on the basis of earlier probability estimates derived from empirical data.

The computation takes a set of input key/value pairs, and produces a set of output key/value pairs.The MapReduce program consists of two functions, the mapper and reducer. Map process takes caries of portioning the Spatial data of rainfall information.Spatial overlap is handled to reduce function

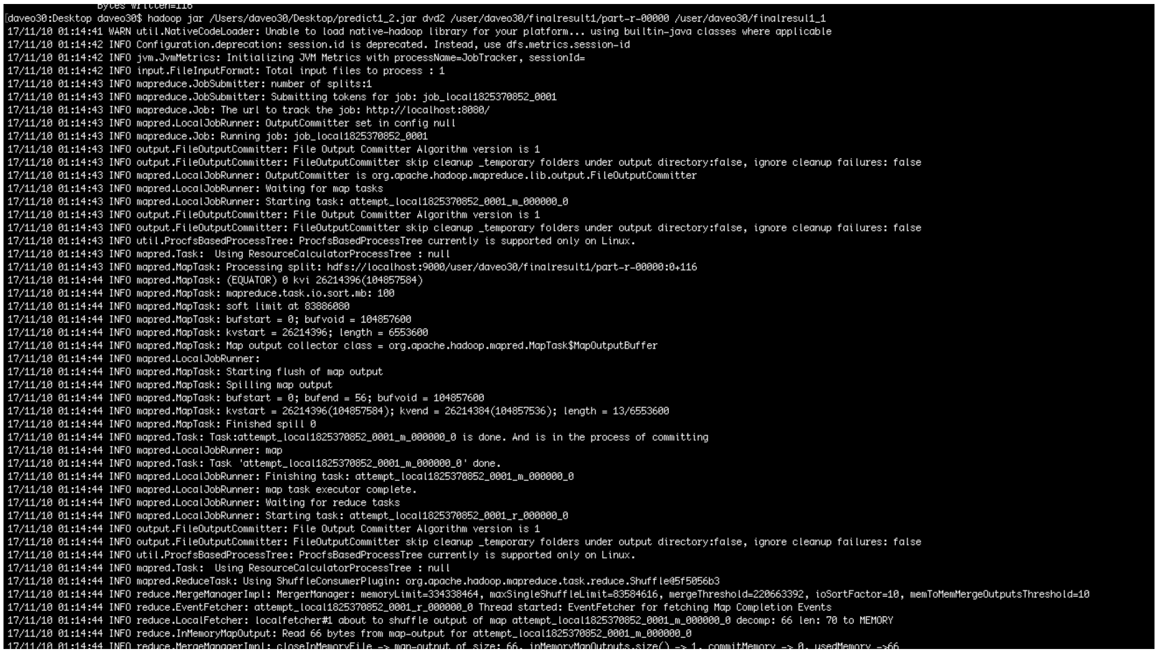
Naive classifiers have several desirable features: First, they are simple to construct, requiring very little domain background knowledge, as opposed to general Bayesian networks which can require numerous intensive sessions with experts to produce the true dependence structure between features. Second, naive networks have very constrained space and time complexity.

The proposed system serves as a tool for predicting rainstorm from a large amount of rainfall data in an efficient manner. The result indicates the proposed system improves the performance in terms of accuracy and efficiency.

IV. RESULTS

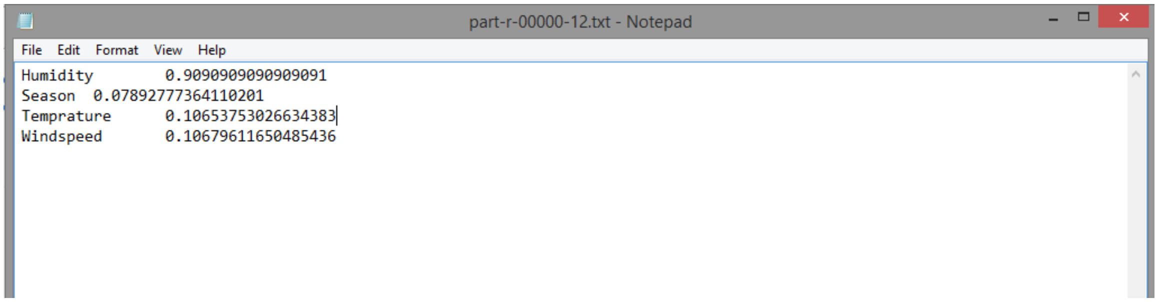


```
[3]+  Stopped                 hadoop jar /Users/daveo30/Desktop/predict2_1.jar dvd /user/daveo30/train4.csv /user/daveo30/finalresult1
daveo30:Desktop daveo30$ hadoop jar /Users/daveo30/Desktop/predict2_1.jar dvd /user/daveo30/train4.csv /user/daveo30/finalresult1
Enter the season
 1 = spring, 2 = summer, 3 = fall, 4 = winter
1
Enter the temperature(in Celsius)
13.94
Enter the Humidity
0
Enter the Wind Speed
16.9979
17/11/10 01:13:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/10 01:13:19 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
17/11/10 01:13:19 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
17/11/10 01:13:20 INFO input.FileInputFormat: Total input files to process : 1
17/11/10 01:13:20 INFO mapreduce.JobSubmitter: number of splits:1
17/11/10 01:13:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2107575858_0001
17/11/10 01:13:21 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
17/11/10 01:13:21 INFO mapreduce.Job: Running job: job_local2107575858_0001
17/11/10 01:13:21 INFO mapred.LocalJobRunner: OutputCommitter set in config null
17/11/10 01:13:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
17/11/10 01:13:21 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
17/11/10 01:13:21 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
17/11/10 01:13:21 INFO mapred.LocalJobRunner: Waiting for map tasks
17/11/10 01:13:21 INFO mapred.LocalJobRunner: Starting task: attempt_local2107575858_0001_m_000000_0
17/11/10 01:13:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
17/11/10 01:13:21 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
17/11/10 01:13:21 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
17/11/10 01:13:21 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
17/11/10 01:13:21 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/daveo30/train4.csv:0+485275
17/11/10 01:13:21 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
17/11/10 01:13:21 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
17/11/10 01:13:21 INFO mapred.MapTask: soft limit at 83886080
17/11/10 01:13:21 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
17/11/10 01:13:21 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
17/11/10 01:13:21 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
In Season
In Season
```



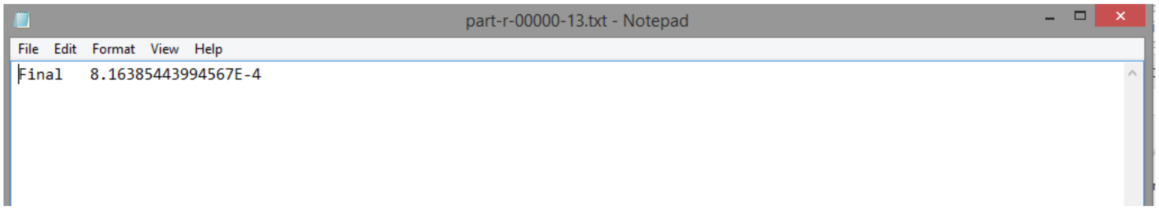
```
                Bytes Written=116
daveo30:Desktop daveo30$ hadoop jar /Users/daveo30/Desktop/predict1_2.jar dvd2 /user/daveo30/finalresult1/part-r-00000 /user/daveo30/finalresul1_1
17/11/10 01:14:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/10 01:14:42 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
17/11/10 01:14:42 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
17/11/10 01:14:42 INFO input.FileInputFormat: Total input files to process : 1
17/11/10 01:14:43 INFO mapreduce.JobSubmitter: number of splits:1
17/11/10 01:14:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1825370852_0001
17/11/10 01:14:43 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
17/11/10 01:14:43 INFO mapred.LocalJobRunner: OutputCommitter set in config null
17/11/10 01:14:43 INFO mapreduce.Job: Running job: job_local1825370852_0001
17/11/10 01:14:43 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
17/11/10 01:14:43 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
17/11/10 01:14:43 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
17/11/10 01:14:43 INFO mapred.LocalJobRunner: Waiting for map tasks
17/11/10 01:14:43 INFO mapred.LocalJobRunner: Starting task: attempt_local1825370852_0001_m_000000_0
17/11/10 01:14:43 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
17/11/10 01:14:43 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
17/11/10 01:14:43 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
17/11/10 01:14:43 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
17/11/10 01:14:43 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/daveo30/finalresult1/part-r-00000:0+116
17/11/10 01:14:44 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
17/11/10 01:14:44 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
17/11/10 01:14:44 INFO mapred.MapTask: soft limit at 83886080
17/11/10 01:14:44 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
17/11/10 01:14:44 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
17/11/10 01:14:44 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
17/11/10 01:14:44 INFO mapred.LocalJobRunner:
17/11/10 01:14:44 INFO mapred.MapTask: Starting flush of map output
17/11/10 01:14:44 INFO mapred.MapTask: Spilling map output
17/11/10 01:14:44 INFO mapred.MapTask: bufstart = 0; bufend = 56; bufvoid = 104857600
17/11/10 01:14:44 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536); length = 13/6553600
17/11/10 01:14:44 INFO mapred.MapTask: Finished spill 0
17/11/10 01:14:44 INFO mapred.Task: Task:attempt_local1825370852_0001_m_000000_0 is done. And is in the process of committing
17/11/10 01:14:44 INFO mapred.LocalJobRunner: map
17/11/10 01:14:44 INFO mapred.Task: Task 'attempt_local1825370852_0001_m_000000_0' done.
17/11/10 01:14:44 INFO mapred.LocalJobRunner: Finishing task: attempt_local1825370852_0001_m_000000_0
17/11/10 01:14:44 INFO mapred.LocalJobRunner: map task executor complete.
17/11/10 01:14:44 INFO mapred.LocalJobRunner: Waiting for reduce tasks
17/11/10 01:14:44 INFO mapred.LocalJobRunner: Starting task: attempt_local1825370852_0001_r_000000_0
17/11/10 01:14:44 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
17/11/10 01:14:44 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
17/11/10 01:14:44 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
17/11/10 01:14:44 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
17/11/10 01:14:44 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@5f5056b3
17/11/10 01:14:44 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
17/11/10 01:14:44 INFO reduce.EventFetcher: attempt_local1825370852_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
17/11/10 01:14:44 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1825370852_0001_m_000000_0 decomp: 66 len: 70 to MEMORY
17/11/10 01:14:44 INFO reduce.InMemoryMapOutput: Read 66 bytes from map-output for attempt_local1825370852_0001_m_000000_0
17/11/10 01:14:44 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 66, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->66
```

Output of the first code,



```
Humidity        0.9090909090909091
Season  0.07892777364110201
Temprature      0.10653753026634383
Windspeed       0.10679611650485436
```

The above output goes as an input into the second program to predict the probability of rainfall to occur.



```
Final   8.16385443994567E-4
```

V. CONCLUSION

Naive classifiers have several desirable features: First, they are simple to construct, requiring very little domain background knowledge, as opposed to general Bayesian networks which can require numerous intensive sessions with experts to produce the true dependence structure between features. Second, naive networks have very constrained space and time complexity.

The proposed system serves as a tool for predicting rainstorm from a large amount of rainfall data in an efficient manner. The result indicates the proposed system improves the performance in terms of accuracy and efficiency.

VI. REFERENCES

- https://link.springer.com/content/pdf/10.1007%2F978-3-642-29387-0_50.pdf
- http://ieeexplore.ieee.org/document/941848/
- https://link.springer.com/chapter/10.1007/978-3-642-29387-0_50
- http://ieeexplore.ieee.org/document/7479954/?section=abstract
- http://www.cs.cmu.edu/~wcohen/10-605/notes/scalable-nb-notes.pdf