BDA
Chp-1

* Data Mining :- It's about extract the Data and discover some patterns & Model.

There are different Types of Models:-
1. Statistical Modeling
2. Machine learning
3. computational Approaches to Modeling
4. summarization
5. Feature extraction.

1] Statistical Modeling :-
Statistical Modeling that is an Underlying distribution from which the visible Data is drawn.

2] Machine learning :-
Machine learning practitioners use the data as a training set to train an algorithm of one of many types used by machine learning, such as BayesNet Support vector machines.

3] Computational Approaches to Modeling :-
To solve complex problem we use computational Model. There are many different Approach to modeling Data

4] Summarization:-

One of the most intresting forms of summarization is the pageRank Idea, which made Google succesfuly and which we will know. In the form of web mining the entire complex structure of the web is summarized by a single number for each page.

5] Feature extraction:-

The typical feature based model looks for the most extreme examples of phenomenon and represent the data

Some of the important kinds of feature extraction from large-scale data that is:

1. frequent Items
2. similar items.

* Statistical limits on Data Mining :-

① Total Information Awarness :-
- It was conjectured that the information needed to predict and foil the attack was available in data, but that there was then no way to examine the data and detect suspicious events. This response was a program called TIA.

② Bonferroni's Principle :-
- A theorem of statistics known as the Bonferroni correction gives a statistically sound way to avoid most of these bogus.

- Bonferroni's principle help us avoid treating random occurances as if they were real.

example : 1. There are one billion people who might be evil-doers.

② Everyone goes a hotel on day in 100.

③ we shall examine hotel records for 1000 days.

* Hash-Functions:-

- A hash Function h takes a hash-key value as an argument & produces a bucket number as a results.
- The bucket number is an integer.
- Hash key can be of any type.
- If hash key drawn Randomly then h will send equal numbers of hash keys each of the Buckets
- Uniform Distribution.

$h(x) = x$

- Σ of value & divide by Numbers of Bucket.

* Indexes:-

- It is an a data structure that makes it efficient objects given the value of one or more element of objects.
- It store the records. It allow to find record quickly through Index number.

Secondary Storage :-

— It is Important, when dealing with large scale Data, that we have a Good understanding of the difference in time taken to perform computation.

— Disks are Organized into blocks.

Thursday

January

JAN
M T W T F S S M T W T F
1 2 3 4 5 6 7 8 9 10 11 12
15 16 17 18 19 20 21 22 23 24 25 26
29 30 31

Chp-2

mmds.org

1 to 32 slide (Revise of BDA)

* Big Data Workflows :-

Understanding workflows & effect of big data is following :-

1. Identify the big data sources you need to use.
2. Map the big data types to your workflow data types.
3. Ensure that you have the processing speed & storage access to support your workflow.
4. Select the data store.

Do not plan for fresh ventures before finishing what is at hand.

* SPARK :- It is a multi-language for executing data engineering, data science and machine learning on single node machine or clusters.

- It provides high level APIs in Java, Scala, Python & R and an optimized engine that supports general execution.

* RDD :- The main abstraction Spark provides is a resilient distributed dataset (RDD). which is collection of element partitioned across the nodes of the cluster that can be operated on in parallel.

- RDD automatically recover from node failures.

Their are 2 ways to create RDD :-

① parallelizing an existing collection in your driver program.

② referencing a dataset in an external storage system.

* RDD Operation in pyspark :-

The RDD supports two types of operations :-

(1) Transformations :-

- Transformations are the process which are used to creat a new RDD.
- Few of transformations are given below :-

1. map
2. flatmap
3. filter
4. distinct
5. reduceBykey
6. mapPartitions
7. sortBy.

14 Sunday

(2) Actions :-

- Actions are the processes which are applied on an RDD.

-

Monday 13

Pair Actions and Transformations:

1. collect
2. collectAsMap
3. reduce
4. countByKey
5. countByValue
6. take
7. first

Various Operations in RDD:-

1. count()
2. reduce(f)
3. collect(f)
4. map
5. filter(f)
6. foreach(f)
7. join
8. CC

* Create RDD Using SparkContext.parallelize()

Code:-

```
#Create RDD from parallelize
data = [1,2,3,4,5,6,7,8,9,10]
rdd = Spark.SparkContex.Parallelize(data)
```

* Create RDD using SparkContext.textFile()

```
#Create RDD from external Data
rdd2 = spark.SparkContext.textFile("/Path/textfile.txt")
```