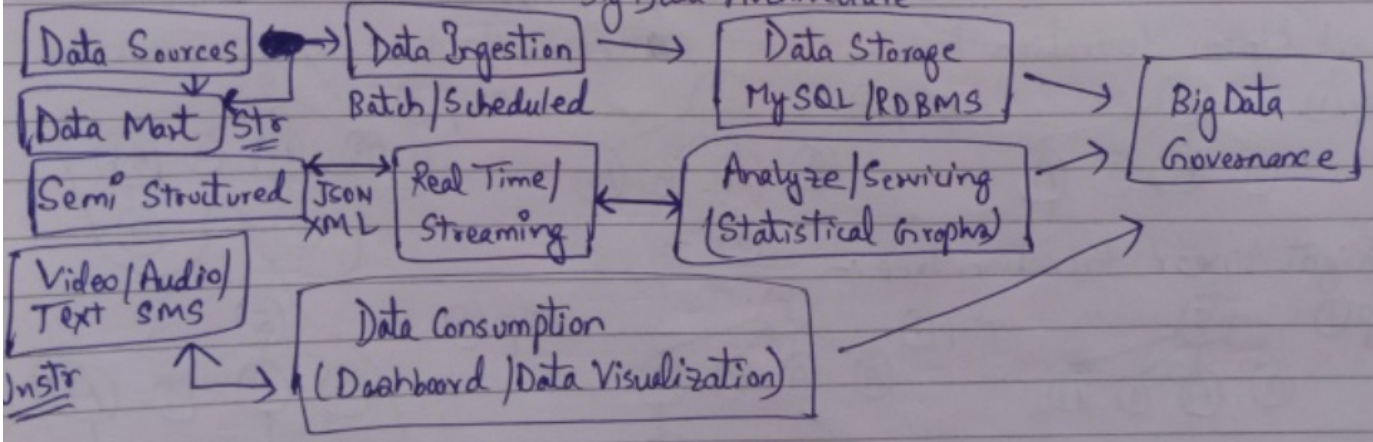
Left linear

BD

Characteristics of Big Data [5 V's] :-

1. ~~Velocity~~ Veracity :- Data is uncertain. Entries incomplete. NULL value possible.
3. Variety :- Data stored from different sources, eg: video, picture, voice.
2. Volume :- Large no. of files.
4. Velocity :- Previous years, exponential increase in collection of data.
5. Value :- Useful data to make some analysis, valuable data.

Big Data Architecture



BD

Components of Big Data :- / Flow of Data :-

1. Capturing the data (Raw collection of facts or figure)
2. Storing the data (Structured, Semi-structured or unstructured)
3. Data Ingestion / Processing / Analysis (Outliers, error data checking)
4. Data Visualization

Fields of Big Data :-

1. Predictive Analysis
2. Machine Learning / Artificial Intelligence {Pattern Recognition}
3. Natural Language Procesing [NLP].
4. Computer Vision (AI)

Technologies related to Big Data :-

1. Apache HADOOP :- It is opensource framework. used in distributed computing of large data set.
2. Apache SPARK :- used along with HADOOP, open source procesing engine.
3. Apache FLINK :- Open source stream procesing framework, user friendly API, livestreaming of real time data.
4. PRESTO :- Open source SEQUEL engine, supports interactive analysis of huge data set. Different systems are there for different analysis.
5. DRUID :- Open source Analytical data storage designed for queries on event based data. eg: log files.

(Stored in accumulator) (Accumulator - A)

BD

(*) Features of Big Data :-

(1.) Data Preparation :- used before iterating the data of the model, capturing the data construction of model used at

(2.) Data Exploration :- visualize the insights like pictorial/graphical representation of data.

(3.) Scalability :- uses less network gear, use less energy (memory + processing should be fast) performance not degraded.

(4.) Supports various types of Analytics :- Numerical values,

IDS: Intrusion Detection System } use of certain firewalls. for security purpose.

statistical values by pictorial representation. (pie chart, graphs, tabular charts) Categorical analysis.

(5.) Version Control :- particular version gets enhanced.

(6.) Data Management :- Capture, Storing in cost effective / space / Time.

(7.) Data Integration :- Collection of data and storing in database (integrating)

(8.) Data Governance :- Data should be accurate, available, reliable i.e. data should be easily governable.

(9.) Security :-

(10.) Visualization :- Data driven environment.

(*) Application of Big Data :-

→ Monitoring then targetting :-

→ Recommendation of Application :-

→ Healthcare (IOT) → Smart Traffic System. (Ola, Uber, Google Map) →

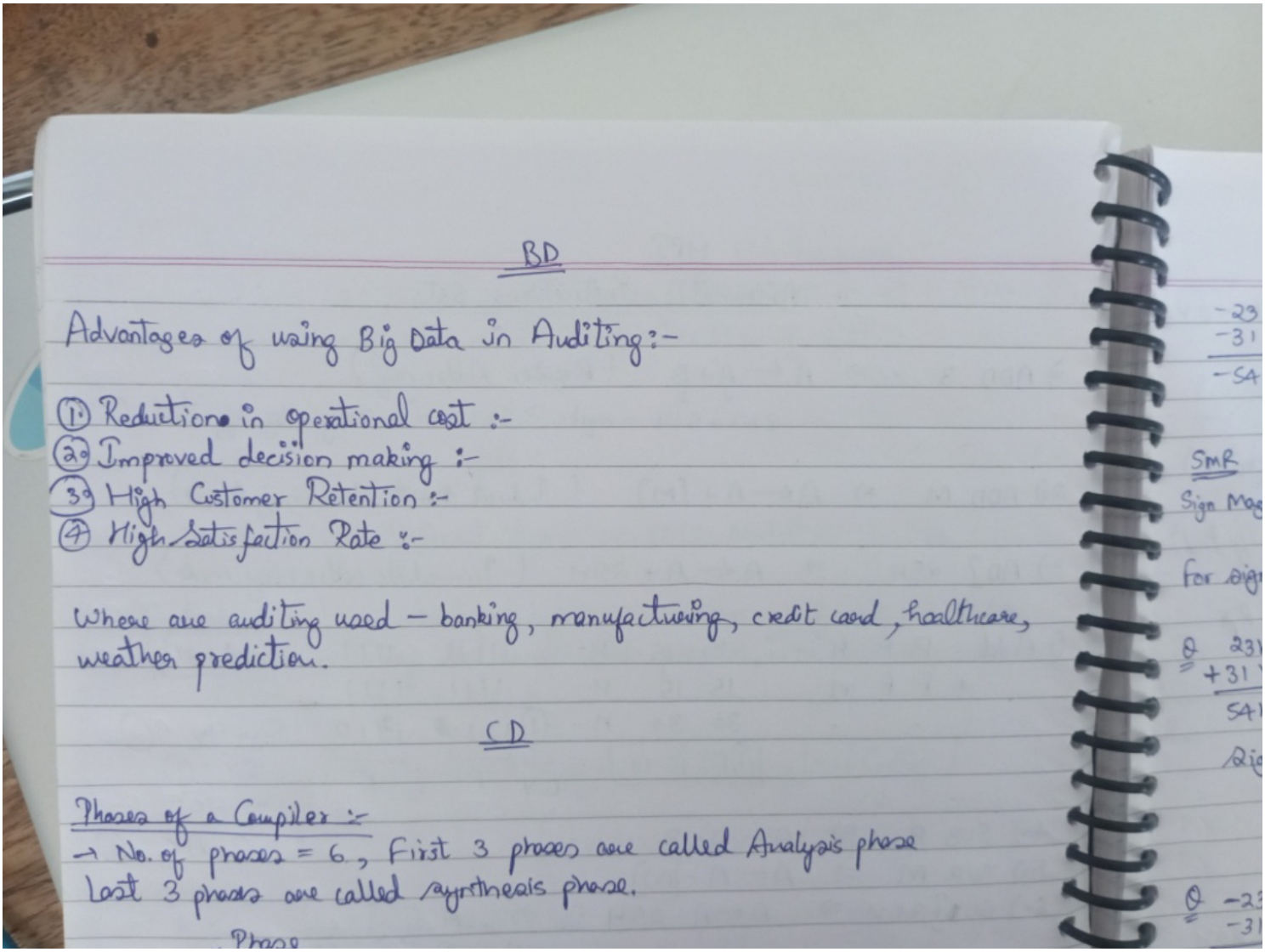
Pg

WAP to enter a number and check whether it is a perfect number.

```
print ("Enter a number")
num = int (input ())
for i in range (1, num) :
    if (num % i == 0)
        sum += i
if (sum == num)
    print ("Yes")
else
    Print ("No")
```

Q

```
i = 11
while (i <= 5) :
    print (i)
    i = i + 1
else :
    print ("Hello")
```



BD

Advantages of using Big Data in Auditing :-

1. Reductions in operational cost :-
2. Improved decision making :-
3. High Customer Retention :-
4. High Satisfaction Rate :-

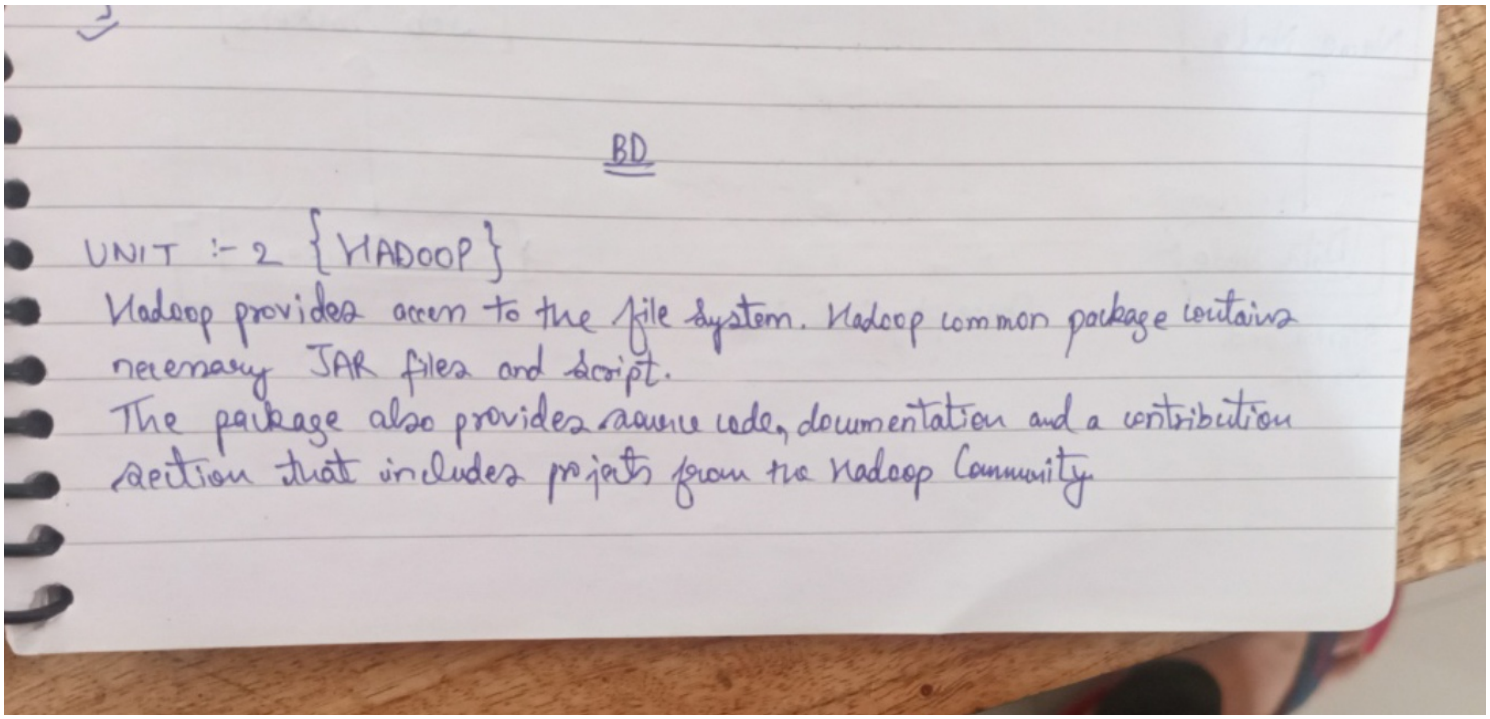
Where are auditing used – banking, manufacturing, credit card, healthcare, weather prediction.

CD

Phases of a Compiler :-

→ No. of phases = 6, First 3 phases are called Analysis phase
Last 3 phases are called synthesis phase.

Phase



BD

UNIT :- 2 {HADOOP}

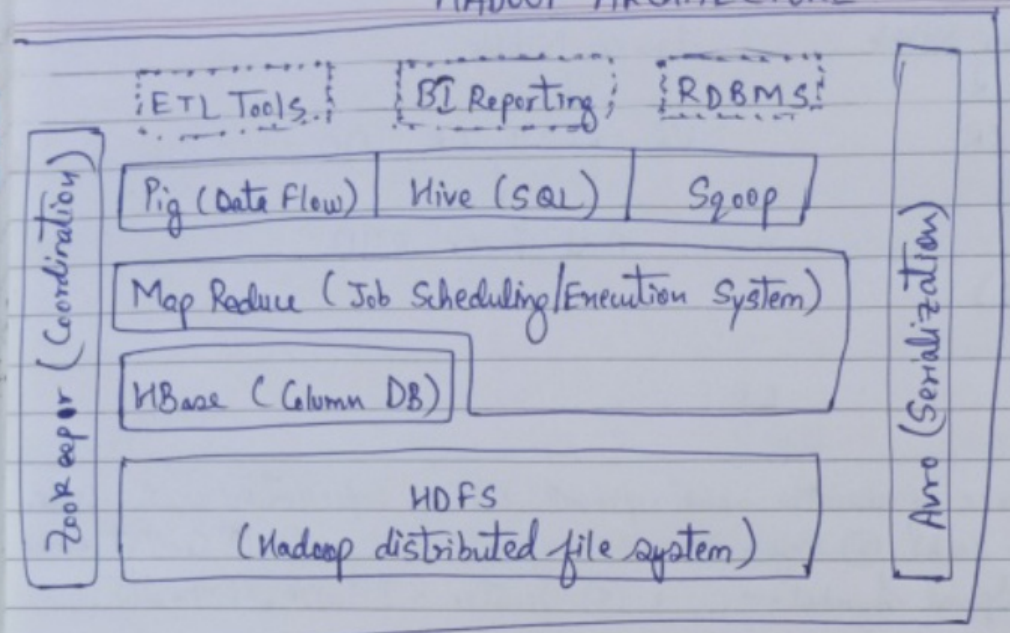
Hadoop provides access to the file system. Hadoop common package contains necessary JAR files and script.

The package also provides source code, documentation and a contribution section that includes projects from the Hadoop Community.

Extract transport Load (ETL) Business Intelligence (BI)

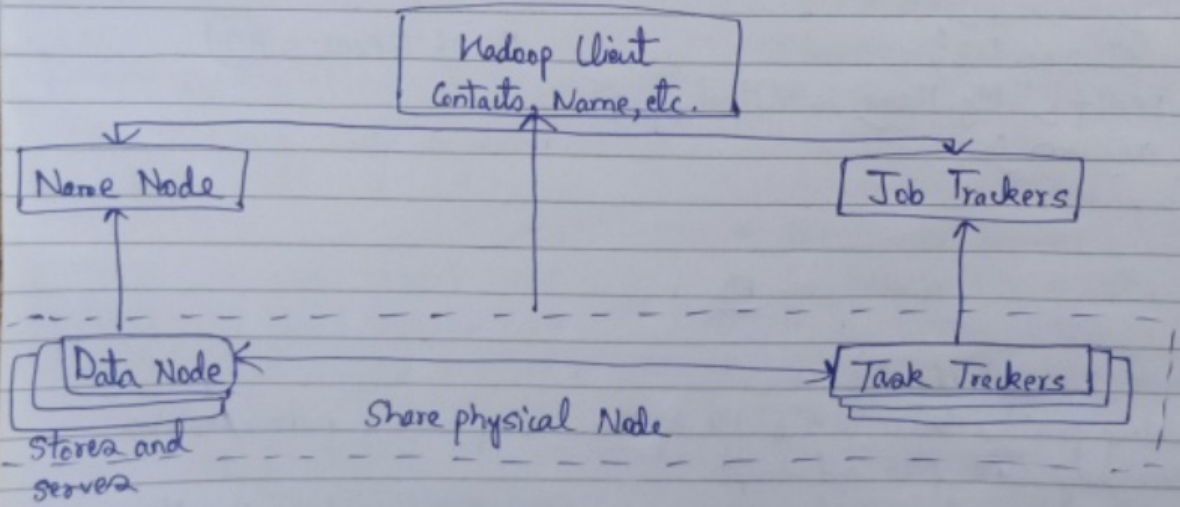
Command Line Interface = Sqoop.

HADOOP ARCHITECTURE

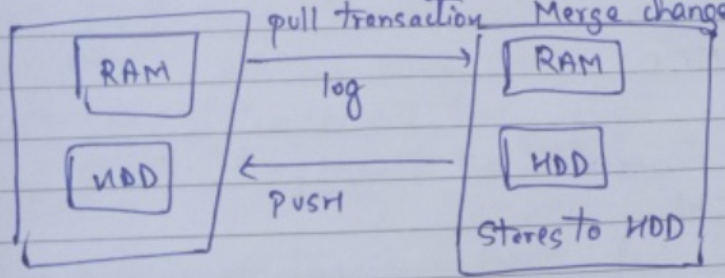


Q What is HDFS? → Distributed file system → Traditional heirarchical file organization. → Single namespace for entire cluster. → Write once - read many access model → Aware of network topology

High Level Architecture



HDFS Name Node Architecture :-



Primary name Node

Secondary Name Node.

MPI

Logical Instructions

1) ANA R — $A \leftarrow A \wedge R$ (Register) — AND operation

2) ANA M — $A \leftarrow A \wedge [M]$ (Indirect)

AND Accumulator

$A \leftarrow A \wedge 25H$ (Immediate)