# **HD**

Problem: Bigdata

Solution: Hadoop

Storage and processing

Storage in – distributed manner

Data life cycle management :

Aim : : Analytics – for business intelligence –

Distributed data 🡪 Capture 🡪Storage across distributed file system (multiple system) 🡪 Processing

Logic or program is required to procees

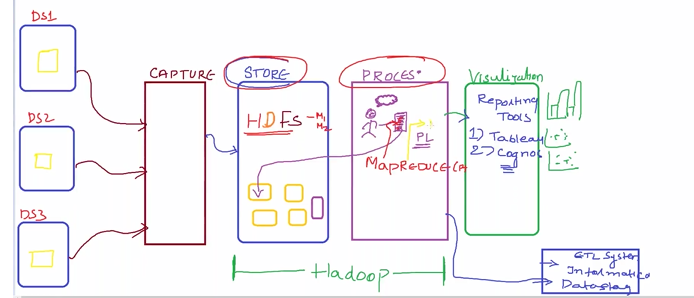
Processed o/p 🡪Vizualization 🡪Reproting tools (tablaue, Bo, cognos,Mstrategy)

Processed data can be send to ETL SYSTEM – RDBMS or any other system

ETL SYSTEMS EX – Informatica, Data Stage,

Storage and processing is impacted stuff 🡪 Hadoop is solution with extended storage and processing

Processing technology 🡪 stoarage (HDFS) + Processing (Mapreduce) is the framework



Mapreduce 🡪

Hive 🡪 by facebook 🡪use sql kind language and in Background 🡪 converts to mapR job is created and run on HDFS

Pig by yahoo 🡪 plain 🡪 scripts genrates Mapr Job 🡪 instead of programing in java

Data sources 🡪 RDBMS 🡪 INJESTION 🡪 capture data to and fro -HDFS – SQL TO HADOOP (SQOOP)

Data import and export to hdfs 🡪 rdbms

Streaming applications

-live data 🡪 coming continuesly 🡪 flume is the technolgy used

Flume ,sqoop, HDFS API

One particular – record access from HDFS – HDFS access pattern will be in sequential data access

Read from the beg till end to extract specific record

Top to bottom read 🡪 sequential data access –

In Rdbms – data can be accessed randomly by querying

Random access capability over HDFS 🡪 achieved with the help of **Nosql databases**

MongoDB

Hbase

Io based processing over top of Hadoop – individual machines –

Faster processing – in memory processing

To manage all the jobs – workflow scheduler is available ie oozie – Hadoop jobs scheduler

Amazon – takes big data 15 billon click streams of customers – for recommendation

Customer scenarios – Big data

Hdfs

Scaling up --.super computer – adding hard ware

Scaling out distributed – hdfs

Dfs – is a s/w – enable as a single system --

1. (quard core , 4g r, 1tb hd) \*100

Types of dfs