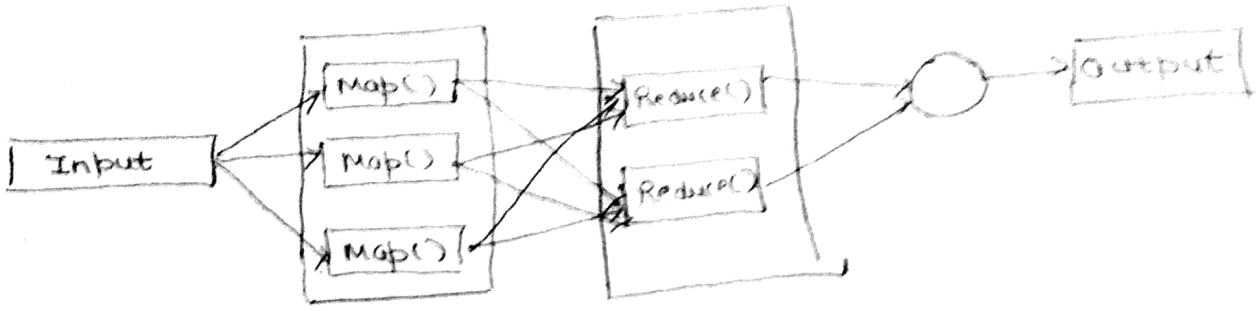
UNIT-3

Map Reduce

It is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment.



1. Two phases – Map & Reduce. Reducer phase takes place after mapper phase
2. Map Job – Block of data is read and processed to produce key-value pairs
3. O/P of Mapper (key-value pairs) is I/P to Reducer

By Reducer (shuffle stage) – receives key-value pair from multiple map jobs; aggregates those key-value pairs into smaller set of tuples (key-value pairs) which is final O/P.

• Advantages of Map-Reduce

1. Parallel Processing – We are dividing the job among multiple nodes and each node works simultaneously.
   - Data is processed by multiple machines instead of a single one and hence time taken to process data gets reduced.
2. Data Locality – Instead of moving data to processing unit, we are moving processing unit to data (moving huge data to PU is costly & time taking; Node can get overburdened)
   - Data is distributed among multiple nodes & all the nodes work with their part of data in parallel.
   - No chance of a node getting overburdened

Scanned by CamScanner

• Terminology Related with Map-Reduce

1. **PayLoad** – Applications implementing the Map-Reduce functions
2. **Mapper** – maps the I/P key-value pairs to a set of intermediate key-value pair
3. **NameNode** – Node that manages the HDFS
4. **DataNode** – Data is presented in advance before any processing takes place
5. **MasterNode** – Node where JobTracker Runs and accept job request from clients
6. **SlaveNode** – Node where Map-Reduce program runs
7. **JobTracker** – Schedules jobs and tracks the assign job to Task Tracker
8. **Task Tracker** – ~~sched~~ Tracks the task & reports status to Job Tracker
9. **Job** – A program is an execution of a mapper & reducer across a dataset
10. **Task** – An execution of a Mapper / Reducer on slice of data
11. **Task Attempt** – A particular instance of an attempt to execute a task on Slave Node.

Scanned by CamScanner

HADOOP
- Open source platform
- provides analytical technologies + computational boost to work with large volume of data
- Apache open source mechanism written in JAVA.
- allows distributed processing of large datasets
- client - server model.
  - master → responsible for data-distribution among clients
  - Slaves - carry out computational task.

HAPOOP common - Java libraries & utilities required by other components necessary files to start HADOOP.

HDFS - cluster of storage solutions

Hadoop Yarn - Platform for job scheduling & resource management.

Hadoop Map-Reduce - perform mathematical computations.

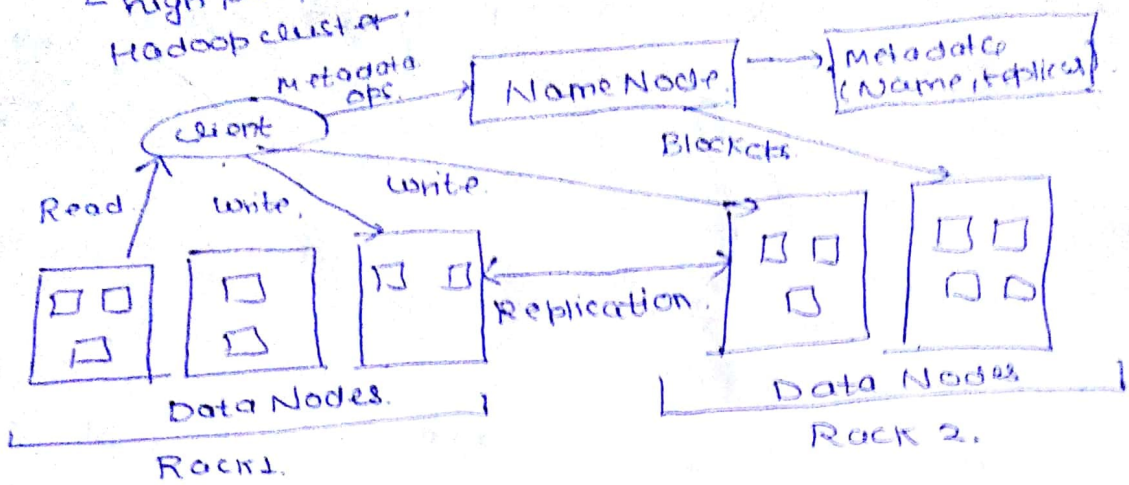
Hadoop Map-Reduce
- software platform - easily writing of applications which process big amounts of data
- limits amount of communication b/w processes as each individual record is processed by a task in isolati

Single Master Job Tracker
- resource management
- scheduling job component-task on slaves.

Slave Tracker
- executes task as directed by master
- provides task status info to master

Scanned by CamScanner

- **Hadoop Distributed File System (HDFS)**
  - primary data storage system.
  - employs a NameNode and DataNode architecture
  - high performance access to data across highly scalable Hadoop cluster.



It follows master-slave architecture

NameNode – commodity H/W that contains GNU/Linux OS & NN software
- system having NN acts as master server,

→ Manage file system namespace.
→ Regulate client's access to files.
→ Execute file system operations – renaming, closing & opening files/dir

DataNode – commodity H/W having GNU/Linux O.S & DN software.
- For every node in cluster, there will be data node

→ Perform read-write operations on file systems.
→ Operations → block creation, deletion, replication according to instructions in namenode.

Scanned by CamScanner

Goals of HDFS

Fault Detection & Recovery

HDFS have large no. of commodity H/w, failure is frequent
↳ should have mechanism for quick & automatic fault detection & recovery.

2. **Huge Datasets** – HDFS should have 100s of nodes per cluster to manage large datasets.

3

3. **Highly Fault-Tolerant** – File system replicates each piece of data multiple time & distributes copies to each node. As a result, data on nodes that crash can be found elsewhere within a cluster.

Scanned by CamScanner

UNIT-4

Analytic Process – technique for analyzing & organizing complex decisions, based on mathematics & psychology

Business Understanding – Problems to be solved / Decisions to be made.
Once the objectives are identified & determined
↳ identify the goals.

Data collection – Data is collected from different sources.
(Structured, Semi-structured / Unstructured).

Data Preparation – Remove unnessary / unwanted data.

Data Modelling – A model is created to analyze relationship b/w different objects of data.

Data Evaluation – Results are obtained from different test-cases.
Data is evaluated & rechecked.

Deployement – constantly checked for errors & data is maintain

- **Characteristics of Big Data**

1. Programmatic – Analysts use code to handle raw data.
to manipulate & explore it
2. Data Driven → BD uses high volume of data to drive analysis
critical characteristic for big data analytics.
3. Iterative → iterations on models provides more computational powers.
4. Use Many Attributes → used for describing relationship b/w entities.

Scanned by CamScanner

- **Categories of Big Data Analytics.**

1. **Prescriptive** – reveals what actions should be taken. (most valuable)
   - results in rules / recommendation for next steps.

2. **Predictive** – analysis of scenarios what might happen – predictive forecast.

3. **Diagonstic**
   - looks at past performance & determine what happened & why.
   - Result – an analytic dashboard.

4. **Descriptive** – what is happening now based on incoming data.
   - use real-time dashboard or email reports.

- **Types of Big Data Analytics**

1. **Basic Analysis**
   - Explore data whose value is unknown (data may be useful or not)
   - Used when lot of disparate data needs to be analyzed.
   - Visualization of simple statistics.

   Only used for small amount of data.

Slicing & Dicing – breaking down your data into smaller sets – easier to explore.

Basic Monitoring – monitor large volume of data in real time.

Anomaly Identification – identify anomalies – an event where observations differ from what u expected
- gives a clue – that something is going wrong

---

2. **Advanced Analytics** – complex analysis of structured / unstructured da
   - sophisticated statistical models – ML, Neural Network, Text Analysis are used

Map Reduce

Predictive modeling – statistical / data mining soln. containing algos.
↓
to determine future outcome.

Text Analytics
- Analyze unstructured text
- extract relevant info
- transform to structured info that can be leveraged in many ways.
- Used computational linguistics / statistics etc.

Scanned by CamScanner

3. **Operational Analytics** - Type of Buisness analytics.
   - Focuses on improving existing operations.
   - have different software & packages
     - which offers various models for showing what happens within buisness.

---

4. **Monetizing Analytics** - can be used to optimize your buisness for better decision making & thus increasing revenues.
   - Able to assemble unique data set that is valuable to company.

   Eg Credit card providers use their assembled data to provide value-added services.

Text Analysis - Analyzing unstructured data, extracting relevant info and transforming it into structured info that can be leverged in various way.

Deriving info from text sources. unstructure & semi-structured. ↓ web logs, blogs, email, social media.

- Statiscal Models + Linguistic Theories (NLP + ML. Stats).
  are used to capture patterns in human language such that machines can understand text & perform analytic task.

Entity extraction (named entities) - identifies an item / person / geographical location / contact info / date / time / currencies.
E.g X is identified as a person in a text to be analyzed

Facts extraction (Relationship)
- Relationship b/w two entities.
- statement about something that exist / happened

E.g X is the CEO of company Y.

Events - Usually contain a time dimension - causes facts to change
E.g- Status of a sales process.

Concepts - identify an event / process / trend / behaviour.
- set of words / phrases → indicating a particular idea / topic to which a customer

Scanned by CamScanner

Sentiments: identify viewpoints / emotions in the underlying text.
- Techniques used – ML / NLP

Social Media Analytics

Gathering data across the internet – from various blogs / social medias, news articles / online text forums etc.

This huge stream of data – is analyzed (often using text analytics)
- to make better business decisions.
- mine customer sentiments to support marketing & customer service activities.

- Used by health agencies – to identify public health treats.
- Used by govt – to look terrorist activities.

Level 1 : Internal / within Enterprise – does not understand usefulness of social media.
Level 2 : Monitoring – minimal engagement – no intention to address any issue being discussed.
Level 3 : Broadcast – interaction with 1/2 – social media – broadcasts marketing message.
Level 4 : Viral – limited engagement
Level 5 : Campaign – engages with community – run contests.
Level 6 : Collabrative – begins listening / analyzing / interacting / provides feedback
Level 7 : People Powered – Top level engagement
- monitor conversation around their profile.
- use SM – to actively locate & address problems

Scanned by CamScanner

Text Analytics Tools for Big Data

Attensity
- orginal text analytics company - began developing & selling products - 1990s
- World's largest NLP development group
- Provides several engines for text analytics
  - Auto-classification
  - Entity-extraction
  - Exhaustive Extraction
    - Attensity's flagship technology that automatically extracts facts from parsed text & organize this info

→ Company is focused on social and multichannel analytics and routing it to buisness users for engagement.

→ Developed a grid computing system - provides high performance capabilities for processing massive amount of real-time data

→ Uses Hadood framework (MapReduce, HDFS, HBase) → to store.

Clarabridge
- pure-play text analytics vendor. (Deals with unstructured data)
- Goal - help companies to take better buisness decisions based on customers experiences & issues.
  - Helping everyone in organisation to take actions & collaborate in real time.

Includes
- real time sentiment determination.
- classification of customer feedback data / text
- staging verbatim for future processing in clarabridge

Features - Single-click root cause analysis.
- identify what is causing a change in volume of text feeds, sentiment etc.

Scanned by CamScanner

Open Text — Candian-based company.
- best known for — its leadership in enterprise information management (EIM)

Visican — managing, securing, extracting value from unstructure data of enterprises.

Semantic Middleware →
- Enables real-time analytics with high accuracy on large datasets across languages, formats and industry domains.
- Semantics can be exposed at different levels and work with different technologies to address buisness issues.

Streaming Data

Streaming data is useful when analytics need to be done in real time - when data is in motion
- analytic computing platform focused on speed. these application requires
- continous stream of unstructured data

→ Data is continuosly analyzed and transformed in memory before it is stored on disk.

→ Processing streams of data — by processing 'time windows' of the data in a memory across several clusters of servers

Streaming data can be used in following scenarios.
- collecting info about movement around a secure site.
- To be able to react to an event — that needs immediate response. (patient's medical condition
- Real time ~~collect~~ calculation of cost that depend on various variables - usage & available resour

Scanned by CamScanner

Complex Event Processing (CEP)

→ Streams & CEP both are used to manage data in motion but the uses of both are quite different.

→ CEP is a technique for - tracking, analyzing & processing data as event happens.
- Info is then processed & communicated based on business rules & processes.

Idea — To be able to establish a correlation b/w stream of information & match the resulting pattern with defined behaviors.

→ It is advanced approach based on simple event processing that collects & combines data from different relevant resources to discover events & patterns. that can result in actio

E.g. Credit card companies uses CEP for better fraud managem
- when a pattern of fraud emerges - company shuts down credit
- whether - reporting applications / sales management applicatio

- Difference b/w CEP and streams.

Streaming Data	CEP.
Applied to analyze vast amount of data in real time.	Focused on - solving specific. Use case based on events & action
Streaming Data applications - manage a lot of data & process it at a high rate of speed.	Does not manage as much data. Runs on less complex hardware.
Managed in highly distributed clustered environment.	

Scanned by CamScanner

Operationalizing Big Data

Making Big Data a part of your business process

- Planning an integration strategy.
  - Data (tradional / big data) – needs to be integrated as a seamless part:

Patient eg.

1. Need to treat patient
2. Diagonis & Test
3. Results of test Past history
4. Start treatment monitor patient

Integrating Big Data

- Understand the problem
- Identify the process involved.
- Identify the info required to solve problem
- Gather data, process it & analyze the results

Applying Big Data Within Your Organization

(Not complete).

1. Figuring the Economics of Big Data

To look at various methods for putting big data to work for your org cost may vary due to size of org, purchasing power, vendor relations

B.D economics should be analyzed in following areas.

- Identification of data type & sources.
- Technology changes / New technologies for bB
- New talent acquistion / upgrades to existing.

2. Identification of data type & sources.

Scanned by CamScanner

Security & Governance For Big Data Environments

→ If collecting data from unstructured data sources, social media sites.
  └ make sure that viruses / bogus links are not buried in content

Data Protection options

1. Data anonymization
   - remove all data that can be uniquely tied to an individual (person's name / SNO / Credit card No.)
   - can protect personal identification

2. Tokenization
   - protects sensitive data by replacing with random tokens / alias values that mean nothing to someone who gets unauthorized access to this data.
   - can protect credit card info, passwords, personal info

3. Cloud Database Control
   - access controls are built into the db.
   - protects whole db so that each piece of data doesn't need to be encrypted.

Data Governance

① Visibility
- You may not have control how your data is used and controlled.
- No control over your visibility into your resources that are running outside your control.

② Unvetted Employees
- company should go through extensive background check of all of its employees
- trust on cloud-provider.

Scanned by CamScanner

(3) Auditing your big data process

- claim that you are meeting the rules necessary to support the operations of business.
- Need to show logs / evidence - that data you are using is secure & clean.
- Need to explain the sources of data.

(4) Putting the Right organizational structure in Place.

(5) Preparing for management of risk.

(6) Setting the right governance & quality policies.

(7) Developing a well Governed & secure Big Data Environment.

Scanned by CamScanner