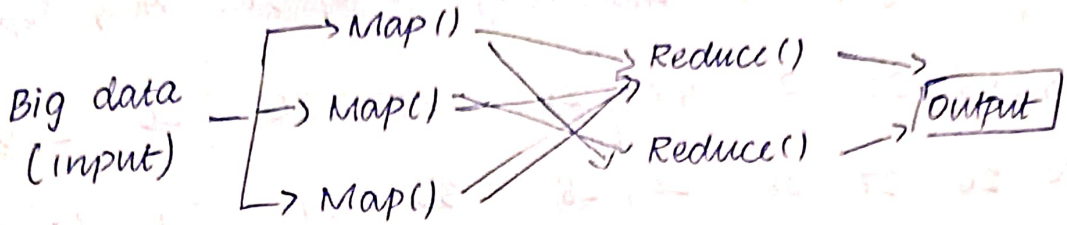
Map reduce in hadoop:



(Hello,1), (I,1), (Am,1), (your,1), (AI,1), (Assistant,1)
(How,1), (can,1), (I,1), (help,1), (you,1)
(How,1), (can,1), (I,1), (assist,1), (you,1)
(Are,1), (you,1), (am,1), (engineer,1)
(Are,1), (you,1), (looking,1), (coding,1)
(Are,1), (you,1), (looking,1), (for,1), (interview,1), (questions,1)
(what,1), (are,1), (you,1), doing,1), (these,1), (days,1)
(what,1), (are,1), (your,1), (Strength,1)

hello - [1]
I -> [1,1,1]
am -> [1]
your -> [1,1]
AI -> [1]
Assistant -> [1]

How -> [1,1]
can -> [1,1]
help -> [1]
you -> [1,1,1,1,1,1]
assist -> [1]
are -> [1,1,1,1,1]
an -> [1]

engineer -> [1]
looking -> [2]
coding -> [1]
for -> [1]
interview -> [1]
question -> [1]
what -> [1,1]
doing -> [1]
these -> [1]
days -> [1]
Strength -> [1]

hello -> 1, I : 3, am : 1, your : 2, AI : 1, Assistant : 1, How : 2,
can -> 2, help : 1, you : 6, assist : 1, are : 5, an 1,
engineer -> 1, looking : 2, coding 1, for : 1, interview : 1,
question : 1, what : 2, doing : 1, these : 1, day : 1,
Strength : 1

Types of digital data:

i) Unstructured
ii) Semi-structured
iii) Structured

Characteristics of structured data:

* Data conforms to a data model and has easily identifiable structure.
* Data is stored in the form of rows and columns
* Easy to query data.

Sources:

* SQL database
* Spreadsheet such as excel

Advantage:

* Ensuring security to data is easy.
* Easily scalable in case there is an increment of data.

Unstructured data:

It refers to information that doesn't follow a predefined data model

eg: text, images, audio, video, mixed format.

Characteristics:

* No fixed rows / columns
* 80-90% of enterprise data
* Require advance technique to analyze
* Stored as files and objects

How to proceed:

* Natural language processing
* Machine learning / Deep learning
* Metadata extraction
* Vector embeddings

Semi-Structured

It doesn't follow a rigid table schema, still contains tags, keys or markers

* It usually text based with a flexible structure

Example:

JSON
XML
YAML
HTML
MongoDB, firebase

Characteristics:

* No fixed schema, but self-describing
* Uses keys, tags or attributes
* Schema can evolve over time
* Easier to parse than unstructured data

Advantage:

* Flexibility
* scalability
* Interoperability

challenges:

* Data consistency
* Query complexity
* validation

General process:

I) Ingest the data
II) Parse the structure
III) Extract relevant fields
IV) Clean & normalize
V) Store / analyze

Big data:

It is a high-volume, velocity and a variety of information assets that demand cost-effective, innovative forms of information processing for enhanced insights and decision making.

history:

Early data management (60s-80s)

* data stored on mainframe and magnetic tapes
* Use of hierarchical and network databases
* Data volume were small and highly structured
* transactions.

II) Relational database (80-90s)

Into Relational Database Management System

* SQL become standard
* Data stored in rows and columns with fixed schemas

limitation:

* could not scale well for massive or unstructured data

3) Data Explosion (90-2000s)

* Rapid growth of the internet, email and e-commerce
* Emergence of unstructured and semi-structured
* Traditional systems struggled with scale
* Term Big Data started gaining attention

4) Big data Technologies:

Google:

* Google file system (GFS)
* Map reduce

Apache Hadoop:

* HDFS for Storage
* Mapreduce for processing
* Enabled distributed storage and processing

5) NoSQL & Real time processing (2010-2015)

→ Rise of NoSQL databases (MongoDB, cassandra)
→ Need to handle
* High velocity data
* Flexible schemas

→ Apache Spark

Cloud & Big Data Ecosystem. (2015-Present)

* cloud platform (AWS, Azure, GCP) offer
  -> Scalable storage
  -> Managed big data services

Integration with:

* AI & ML
* Streaming platforms
* Data lakes

CISCO

6 V

i) Volume:

Amount of data generated and stored

* Ranges from terabytes to petabytes and beyond
* comes from social media, sensors, transaction, video

eg: facebook stores petabytes of user data daily

ii) Velocity:

Speed at which data is generated, processed and analysed.

eg: stock market trades

iii) variety:

different types and formats of data

* Structured (tables)
* Semi-structured (JSON, XML)
* Unstructured (text, image, videos)

iv) Veracity

Quality, accuracy, and trustworthiness of data

5) Value.

Usefulness of data in decision making

6) Variability:

Inconsistency and changing meaning of data

Uses of big data

i) customer experience analysis

ii) Predictive failure analysis

Benefit of big data it

i) New platform

ii) Faster processing

iii) Cost reduction.

business:

i) fraudd detection

ii) customer relationship management

iii) Behavioral Analysis

iv) Financial risk analysis

Enterprise:

i) Enterprise search

ii) revenue assurance

iii) Equipment monitoring.

iv) Data reduction

v) pricing optimization

vi) Legal discovery

challenges:

I) sharing and accessing data

-> inaccessibility of datasets from external sources
-> substantial challenges.
-> multiple difficulties.
-> timely and upto date.
->

II) privacy security:

-> Sensitive, conceptual, technical as well as legal significena

III) Analytical

Lifecycle phases of Big data Analysis

Stage 1: Business case evaluation
Stage 2: Identification of data
Stage 3: Data filtering
Stage 4: Data extraction
Stage 5: Data aggregation
Stage 6: Data analysis
Stage 7: Visualization of data
Stage 8: Final analysis result

Data analytics:

* Hadoop.
* MongoDB
* Talend
* Cassandra
* Spark
* STORM
* Kafka

BDA Industry Application:

* ecommerce
* Marketing
* Education
* Healthcare
* Media and entertainment
* Banking
* Telecommunication
* Government

History of Hadoop:

Doug Cutting and Michael Cafarella in 2005

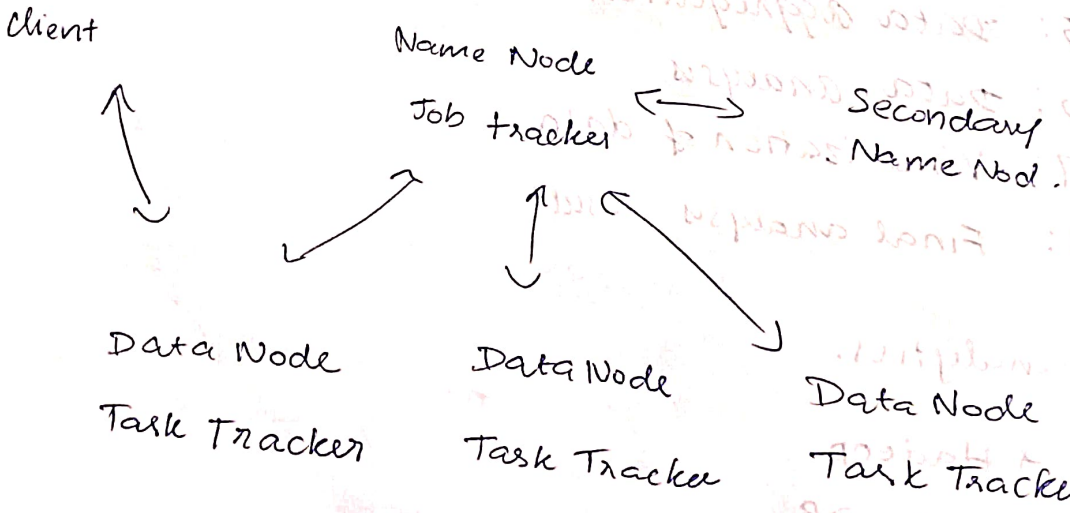
* Hadoop distributed File System.
* Data from Map reduce

Components:

* hdfs
* mapreduce
* Yarn

Hadoop version

1. Hdfs, mapreduce.



Hadoop version 2:

YARN is added in this version with hdfs and mapreduce.

v-3:

multiple name nodes:

erasure coding, use of GPU hardware and docker

* Economically feasible.
* Easy to use
* open source
* Fault tolerance
* scalability
* Distributed Proce
* Locality of Data.

Apache Hadoop:

* collection of open-source.
* massive amount of data

Components:

Hadoop common
Mapreduce
YARN
HDFS
Ozone

Analyzing Data with Unix tools:

I/O DATA / PROGRAM 1 / PROGRAM 2 / PROGRAM 3

Map reduce: — Mapping
\ reducing

i) Split
ii) Suffling
iii) sort
iv) reduce

Rack awareness