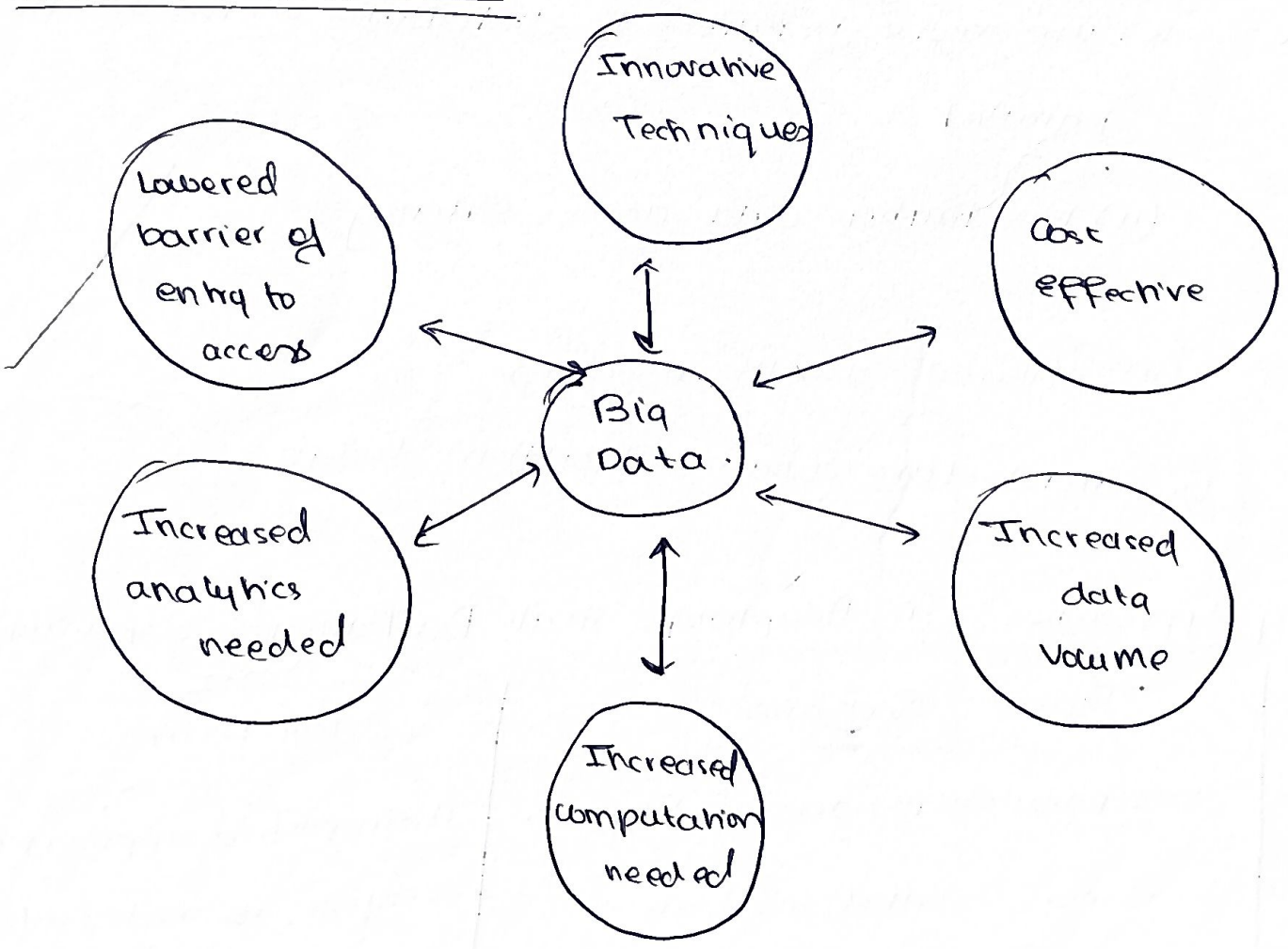
Business Intelligence

Unit - 1

* Definition and Characteristics of Big Data

→ Big data is high volume, high-velocity and high-variety information that demand cost-effective, innovative forms of information processing for enhanced insight & decision-making

* Need for Big Data



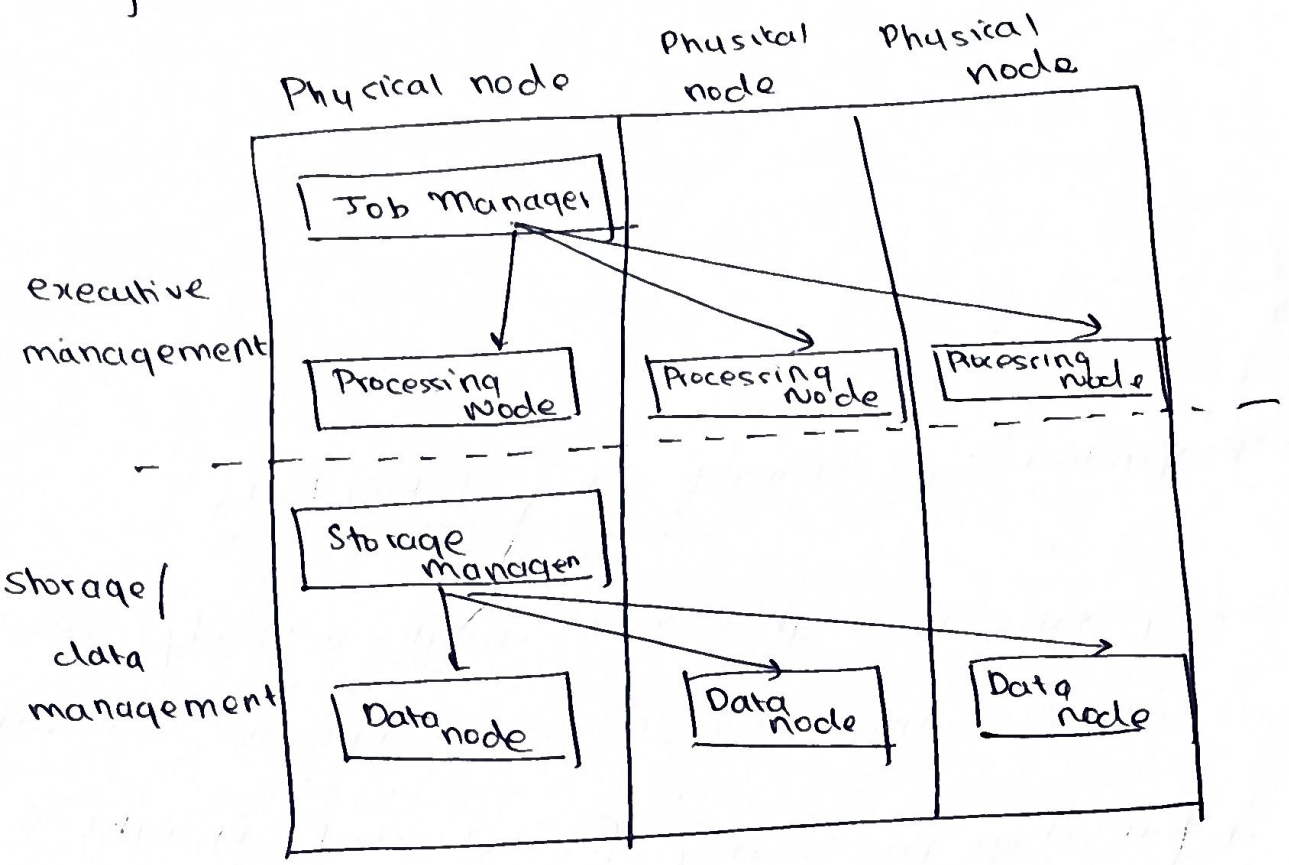
* Key Computing Resources of Big Data

→ Processing capability – CPU, processor or node

→ memory

→ storage

→ node.



* Scalability

(i) Scale Out – (i) use more resources to distribute workload in parallel

(ii) has higher data access latency

(ii) Scale Up – (i) efficient use the resources

(ii) architecture aware algorithm design

* Contrasting Approaches in Adapting High Performance Capabilities

Aspect	Typical Scenario	Big Data
Application Development	uses parallelism developed by developers skilled in high-performance computing, performance optimization & code tuning	A simplified application & execution model on a distributed file system, application programming model and distributed database.

Agent	Typical Scenario	Big Data.
Platform	uses high-cost massively parallel processing computers, utilizing high-bandwidth networks and massive I/O devices	more scalable & elastic virtualized platforms with cloud-based computing services
Data Management	limited to file-based or relational database (RDBMS) using standard row-oriented data layouts	usage of NOSQL can provide in-memory data management for rapid access, columnar layouts to speed query response and graph databases - for social network analytics
Resources	large capital-investment to purchase high-end software	deploy systems like Hadoop on virtualized platforms, for a cloud-based environment. More cost-effective and practical

* Techniques used in BigData

1. Massive parallelism
2. huge data volumes storage
3. data distribution
4. high speed networks
5. high performance computing
6. task and thread management
7. data mining & analytics
8. data retrieval
9. machine learning
10. data visualization

* Why big data now?

→ more data being collected and stored.

→ open source code

→ commodity hardware/cloud

* Apache Hadoop

→ a framework that allows for the distributed processing of large data sets across clusters of computers.

→ designed to scale up from single servers to thousands of machines each offering local computation and storage.

→ Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application

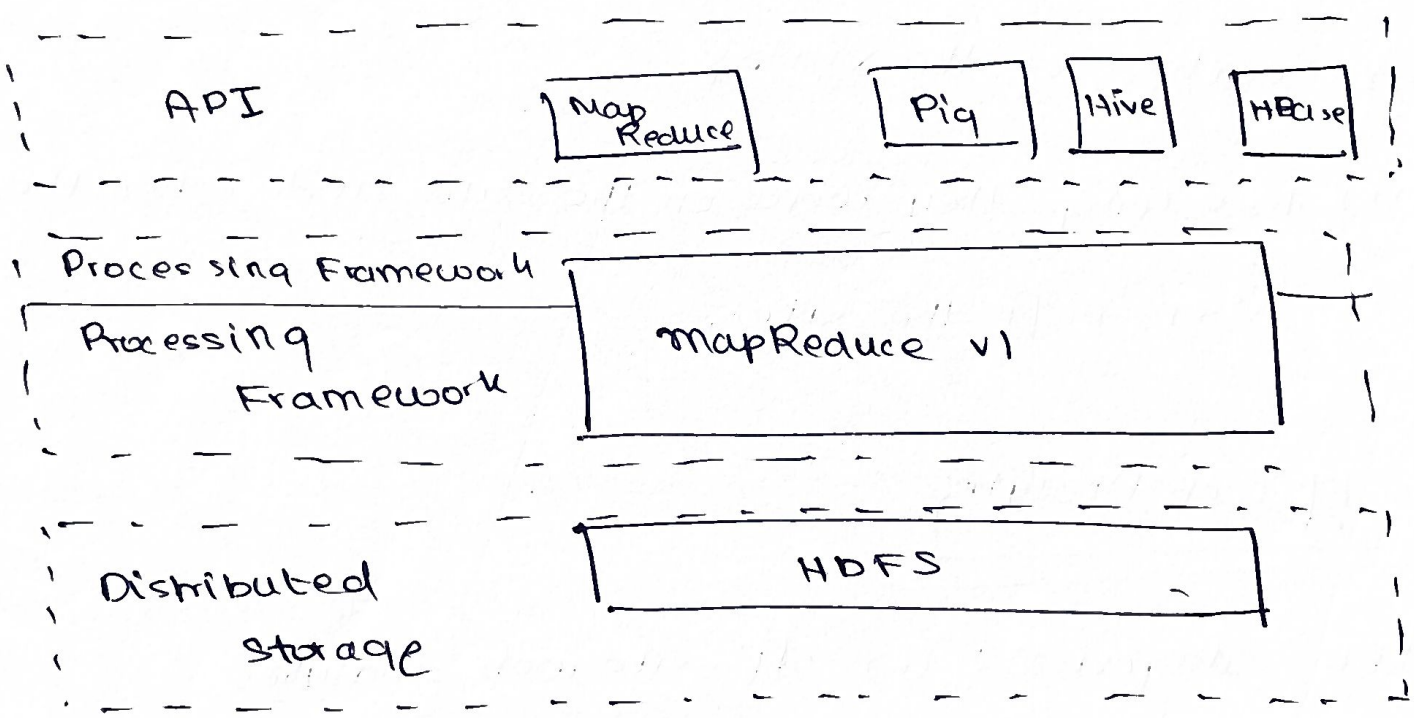
Hadoop has the following modules:

(i) Hadoop Common : utilities that support other Hadoop modules

(ii) Hadoop Distributed File System (HDFS) - a distributed file system that provides high-throughput access to application data

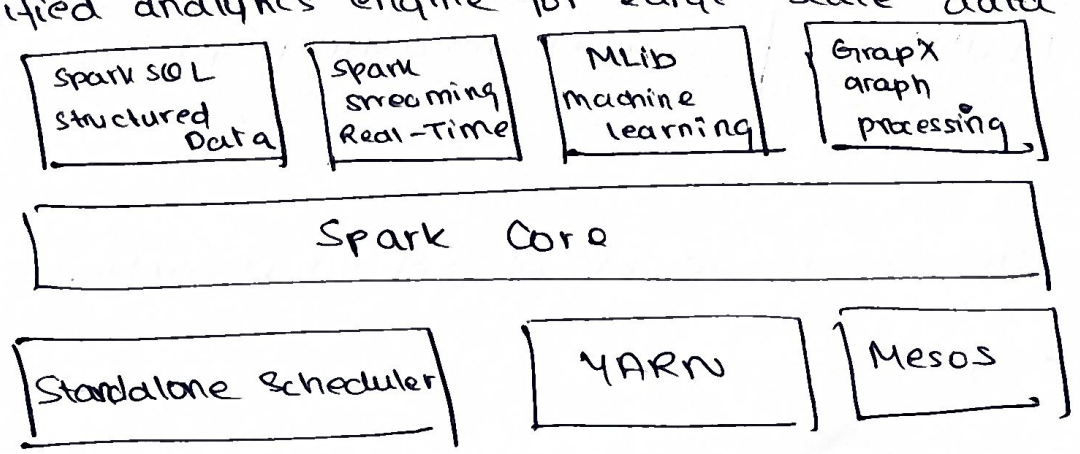
(iii) Hadoop YARN - a framework for job scheduling and cluster resource management.

(iv) Hadoop MapReduce - A YARN-based system for parallel processing of large data sets.



* Apache Spark

→ a unified analytics engine for large-scale data processing



* Web Analytics

→ practice of measuring collecting, analyzing and reporting on Internet / web data.

→ used to monitor whether a site is working properly

→ Web analytics software measure :

(i) web traffic

(ii) how many people visited their site

(iii) how many of those visitors were unique visitors

(iv) how they came to the site

(v) location of the visitors

(vi) how long they were on the site and when the user left the site.

* Categories of Web Analytics

There are two categories : (i) off-site web analytics

(ii) on-site web analytics

A. **Off-site Web Analytics** - web measurement and analysis of website.

→ includes the ~~analysis~~ measurement of potential audience, visibility of the site and comments.

B. **On-site Web Analytics** - measures a visitor's behavior once on the website

→ measures the performance in a commercial context

→ compare it with KPI and stake steps to improve the usage of the website.

* Web Data Collection Methods

A. Log File Analysis - read log files of the web server

B. Page Tagging - uses Javascript, whenever a page is rendered by a web browser or when a mouse click occur, it collect data.

A. Log File Analysis

→ counts the no. of clients requests made to the web server

→ 2 measures : page views and visits

- page views → req. made to server
- visits → sequence of requests from a unique client that expires after a certain amount of time

B. Page Tagging

→ web counters can be used.

→ pass info along, using a small invisible image or Javascript

→ Ajax code would call back to the server & pass info. about the client.

C. Other Analytics

(i) Click Analytics

→ determine the performance of a particular site, w.r.t where the users of the site are clicking

→ A click is logged when it occurs

→ An assumption is that a page view is the result of a click.

(ii) Customer Lifecycle Analytics

→ page views, clicks and other events are tied to an individual visitor instead of storing as separate data points

→ used for website optimization

(iii) Packet Sniffing

→ collects data from the network traffic passing between the web server and the outside world.

* Analytics Terminology

Hit → a request for a file from the web server.

Page view → a request for a file or an event such as a mouse click

event → a discrete action / class of actions

repeat visitor → a visitor that has made at least one previous visit

neco visitor → a visitor that has not made any previous visits

single page visit / singleton → a visit in which only a single page is viewed.

page time viewed → difference between the time of the request to that page and the time of the next recorded request.

* Google Analytics

→ A tracking system - tool suite to monitor and track visitors of the website.

→ 2 areas of Google Analytics : admin area and standard reporting area.

Admin Area - set up goals, add a user or admin, create an advanced segment, manage scheduled reports

Standard Reporting Area - controls / monitors neco visits, page view, page time view

* Google Analytics Metrics

A. **Traffic trend** - analyze traffic using total visit, unique visitors, page views, pages per visit, average visit duration.

B. **Sources** - 3 main traffic sources

(i) referral traffic - from outside web page, linking to your site

(ii) direct traffic - visitor directly typing in your URL

(iii) search traffic - visits coming from search engines like Yahoo| Google.

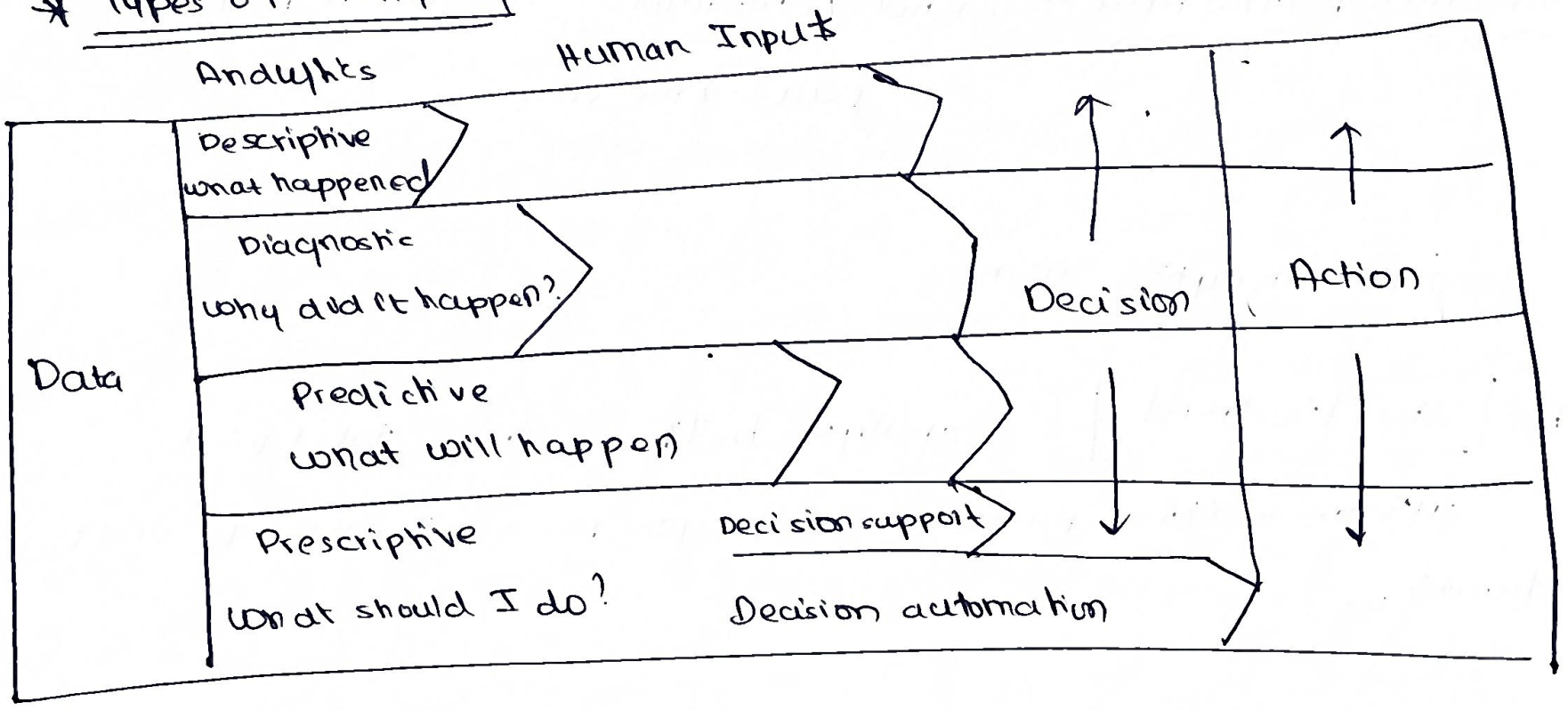
C. **Content** - looking at content based on the top landing pages and top exit pages

Landing page - first page a visitor

Exit page - page where visitors exited on

D. **Conversions** - used to view goals set up and their performance.

* Types of Analytics



A. Descriptive Analysis

→ process of using current and historical data to identify trends and relationships

eg. (i) Traffic and engagement reports
(ii) financial statement analysis
(iii) demand trends
(iv) survey results
(v) progress to goals

B. Diagnostic Analytics

→ process of using data to determine the causes of trends and correlations between variables

→ the next logical step after using descriptive analytics
→ can be done w/ statistical software, hypothesis testing, correlation, regression

eg (i) examining market demand
(ii) examining customer behavior
(iii) identifying technological issues
(iv) improving company culture

C. Predictive Analytics

→ involves the use of data to predict future trends & events.
→ forecast potential scenarios that can help drive strategic decisions

eg (i) Finance - forecasting future cash flow

(ii) determining staffing need. - Hospitality

(iii) Marketing : behavioral targeting

(iv) manufacturing : prevent malfunction

(v) health care : early detection of allergic reactions

D. **Prescriptive Analytics**

→ process of using data to determine an optimal course of action.

eg (i) investment decisions

(ii) product development and improvement

(iii) Read more in sales

A Steps in Predictive Analysis

→ evaluate data on past behaviors & predict the likelihood of future behavior to enable better decisions & outcomes

Steps (i) define project

(ii) data collection

(iii) data analysis

(iv) statistics

(v) modelling

(vi) deployment

(vii) model monitoring

OLAP and OLTP

Online Analytical Processing

→ designed for query and analysis rather than for transaction processing

→ allow for complex calculations, trend analysis and sophisticated data modelling

→ optimized for read-heavy operations - providing fast response times for multidimensional queries.

Operations : aggregation, summarization & reporting

data : large volumes of historical data

users : business analysts, data scientists

eg. analyze sales performance across different regions and time periods.

Online Transaction Processing

→ systems designed to manage transaction-oriented applications.

→ optimized for a large number of short online transactions like insert, update and delete operations.

→ Aim is to ensure data integrity in multi-access environments and provide fast query procesing.

operations : simple queries for insert, update & delete

data: handles real-time data and often has smaller data volumes compared to OLAP

users - front-line employees, clerks and operational staff

eg. an e-commerce application where users place orders & the system needs to handle these transactions efficiently.

* A/B Testing (Split Testing) for Web-based Analytics

→ a method of comparing two versions of a webpage or app against each other to determine which one performs better.

→ used in web analytics to optimize user experience and improve KPI such as conversion rates & user engagement.

Working

① Hypothesis

→ Identify the feature one would like to test / the metric one is focusing on.

② Create Variations

Version A (Control): The original version of the webpage

Version B(Variant): modified version w/ the changes implemented.

③ Random Assignment

→ Randomly assign website users to either the control group or the test group.

④ Data Collection

→ Collect data on how users interact with each version. metrics can be no. of clicks, time spent on page, conversions etc.

⑤ Analysis

→ Analyze which version performs better

→ statistical tests

⑥ Implementation

→ Implement the variant, if it outperforms the control group. Else, iterate on the hypothesis and test new variations.

Benefits of A/B Testing

1. Data-driven decision making
2. Improved user experience
3. Increased conversion rates
4. Reduced risk

V's of Big Data

1. Volume
2. Velocity
3. Variety
4. Veracity - trustworthiness of data
5. ~~Volume~~ Value
   Variability
   Visualization