1.6 BIGDATA APPLICATIONS

Big data has numerous applications across various industries, including:

- ***Healthcare***: Big data is used in healthcare to improve patient outcomes, reduce costs, and optimize treatment plans. Healthcare providers use big data to analyze patient data, predict disease outbreaks, and improve diagnostic accuracy.

- ***Retail:*** Retailers use big data to analyze customer behavior, optimize pricing and inventory, and personalize the customer experience. Big data helps retailers make data-driven decisions about which products to stock, how to price them, and how to market them.
- ***Finance:*** Big data is used in finance to detect fraud, improve risk management, and predict market trends. Financial institutions use big data to analyze customer behavior, identify potential risks, and optimize investment strategies.
- ***Manufacturing:*** Manufacturers use big data to optimize production processes, improve quality control, and reduce costs. Big data helps manufacturers identify inefficiencies in production processes, track product quality, and optimize supply chain management.
- ***Energy:*** Energy companies use big data to optimize energy production, reduce costs, and improve energy efficiency. Big data helps energy companies predict demand, optimize production processes, and improve energy efficiency.
- ***Transportation*:** Transportation companies use big data to optimize routes, reduce fuel consumption, and improve safety. Big data helps transportation companies track vehicle performance, optimize routing, and improve safety measures.
- ***Marketing:*** Marketers use big data to analyze customer behavior, personalize the customer experience, and optimize marketing campaigns. Big data helps marketers identify target audiences, track customer behavior, and optimize marketing strategies.

BIG DATA TECHNOLOGIES

Big data technologies can be categorized into four main types: data storage, data mining, data analytics, and data visualization [2]. Each of these is associated with certain tools, and you'll want to choose the right tool for your business needs depending on the type of big data technology required.

1. Data storage

Big data technology that deals with data storage has the capability to fetch, store, and manage big data. It is made up of infrastructure that allows users to store the data so that it is convenient to access. Most data storage platforms are compatible with other programs. Two commonly used tools are Apache Hadoop and MongoDB.

- **Apache Hadoop:** Apache is the most widely used big data tool. It is an open-source software platform that stores and processes big data in a distributed computing environment across hardware clusters. This distribution allows for faster data processing. The framework is designed to reduce bugs or faults, be scalable, and process all data formats.
- **MongoDB:** MongoDB is a NoSQL database that can be used to store large volumes of data. Using key-value pairs (a basic unit of data), MongoDB categorizes documents into collections. It is written in C, C++, and JavaScript, and is one of the most popular big data databases because it can manage and store unstructured data with ease.

2. Data mining

Data mining extracts the useful patterns and trends from the raw data. Big data technologies such as Rapidminer and Presto can turn unstructured and structured data into usable information.

- **Rapidminer:** Rapidminer is a data mining tool that can be used to build predictive models. It draws on these two roles as strengths, of processing and preparing data, and building machine and deep learning models. The end-to-end model allows for both functions to drive impact across the organization [3].
- **Presto:** Presto is an open-source query engine that was originally developed by Facebook to run analytic queries against their large datasets. Now, it is available widely. One query on Presto can combine data from multiple sources within an organization and perform analytics on them in a matter of minutes.

3. Data analytics

In big data analytics, technologies are used to clean and transform data into information that can be used to drive business decisions. This next step (after data

mining) is where users perform algorithms, models, and predictive analytics using tools such as Apache Spark and Splunk.

- **Apache Spark:** Spark is a popular big data tool for data analysis because it is fast and efficient at running applications. It is faster than Hadoop because it uses random access memory (RAM) instead of being stored and processed in batches via MapReduce . Spark supports a wide variety of data analytics tasks and queries.
- **Splunk:** Splunk is another popular big data analytics tool for deriving insights from large datasets. It has the ability to generate graphs, charts, reports, and dashboards. Splunk also enables users to incorporate artificial intelligence (AI) into data outcomes.

4. Data visualization

Finally, big data technologies can be used to create stunning visualizations from the data. In data-oriented roles, data visualization is a skill that is beneficial for presenting recommendations to stakeholders for business profitability and operations—to tell an impactful story with a simple graph.

- **Tableau:** Tableau is a very popular tool in data visualization because its drag-and-drop interface makes it easy to create pie charts, bar charts, box plots, Gantt charts, and more. It is a secure platform that allows users to share visualizations and dashboards in real time.
- **Looker:** Looker is a business intelligence (BI) tool used to make sense of big data analytics and then share those insights with other teams. Charts, graphs, and dashboards can be configured with a query, such as monitoring weekly brand engagement through social media analytics.

- This is what makes NoSQL the ideal choice for big data, real-time web apps, customer 360, online shopping, online gaming, Internet of things, social networks, and online advertising applications.

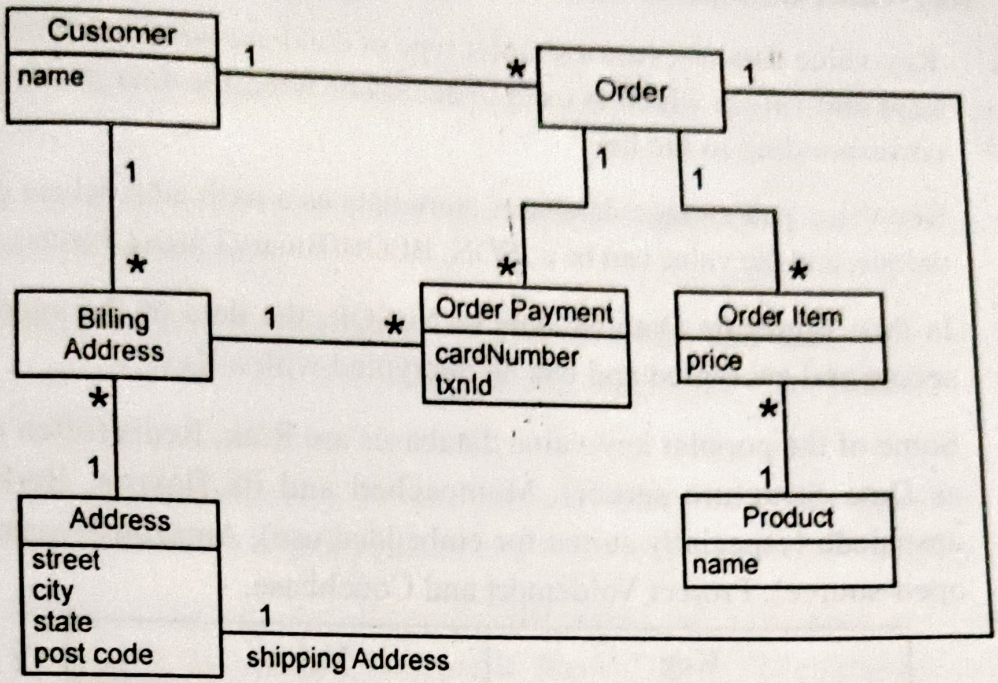
2.2 AGGREGATE DATA MODELS

- Aggregate means a collection of objects that are treated as a unit.
- In NoSQL Databases, an aggregate is a collection of data that interact as a unit. Moreover, these units of data or aggregates of data form the boundaries for the ACID operations.
- Aggregate is a term that comes from **Domain-Driven Design**. In Domain-Driven Design, an aggregate is a collection of related objects that are treated as a unit.
- A data model is the model through which we perceive and manipulate the data.
- The term "data model" often means the model of the specific data in an application.
- Aggregate Data Models in NoSQL make it easier for the Databases to manage data storage over the clusters as the aggregate data or unit can now reside on any of the machines.
- Whenever data is retrieved from the Database all the data comes along with the Aggregate Data Models in NoSQL.
- Aggregate Data Models in NoSQL don't support ACID transactions and sacrifice one of the ACID properties.
- With the help of Aggregate Data Models in NoSQL, OLAP operations can be easily performed on the Database.
- The Aggregate Data Models in NoSQL are majorly classified into 4 Data Models: Key-value pair, Column-oriented, Graph-based and Document-oriented. Every category has its unique attributes and limitations.

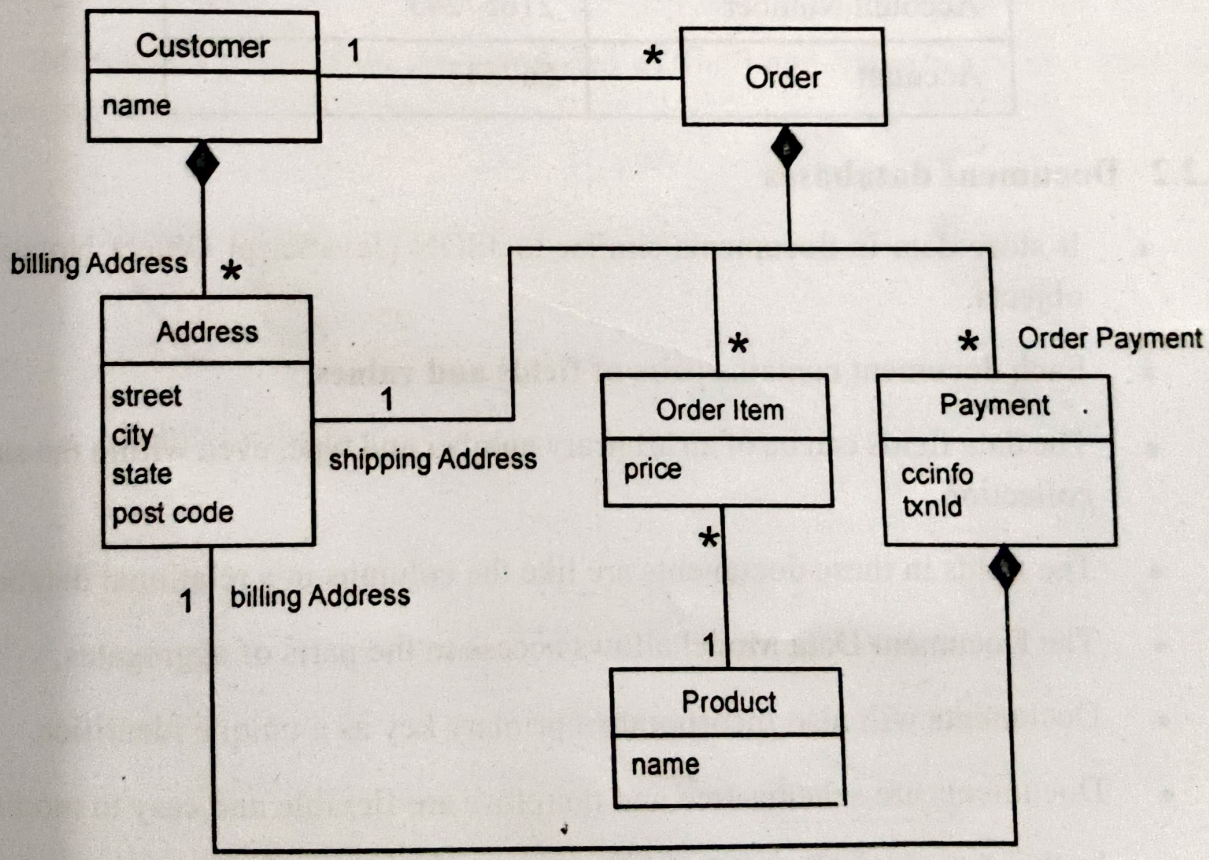
Example of Relations and Aggregates:

Let's assume for building an e-commerce website; we are going to sell items directly to customers over the web, and we will have to store information about users, our product catalog, orders, shipping addresses, billing addresses, and payment data.

Data model for a relational database:



An aggregate data model:



2.2.1 Key-value databases

- Key-value databases are a simpler type of database where each item contains keys and values which is used to access or fetch the data of the aggregates corresponding to the key.
- Key-value pair storage databases store data as a hash table where each key is unique, and the value can be a JSON, BLOB(Binary Large Objects), string, etc.
- In this Aggregate Data Models in NoSQL, the data of the aggregates are secure and encrypted and can be decrypted with a Key.
- Some of the popular key-value databases are Riak, Redis (often referred to as Data Structure server), Memcached and its flavors, Berkeley DB, upscaledb (especially suited for embedded use), Amazon DynamoDB (not open-source), Project Voldemort and Couchbase.

Key	Value
Name	Raman Sharma
Account Number	21657243
Account	567543

2.2.2 Document databases

- It store data in documents similar to JSON (JavaScript Object Notation) objects.
- Each document contains pairs of **fields and values.**
- The data fields can be of an arbitrary number and type, even within the same collection.
- The fields in these documents are like the columns in a relational database.
- The Document Data Model allows access to the parts of aggregates.
- Documents will also incorporate a primary key as a unique identifier.
- Documents are schema-free and therefore are flexible and easy to modify.
- In this Aggregate Data Models in NoSQL, the data can be accessed in an inflexible way.

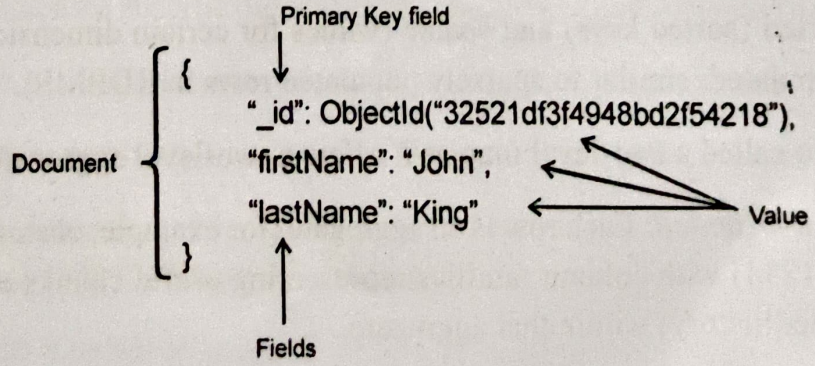
- The Database stores and retrieves documents, which can be XML, JSON, BSON, etc.
- Collections are used to store documents in order to group different kinds of data.
- Documents may have nested documents, various different key-value pairs, or key-array pairs.
- There are some restrictions on data structure and data types of the data aggregates that are to be used in this Aggregate Data Models in NoSQL Database
- Document databases do not support relations. Each document in the document store is independent and there is no relational integrity.
- Some of the popular document databases we have seen are **MongoDB, CouchDB, Terrastore, OrientDB, RavenDB,** and of course the well-known and often reviled Lotus Notes that uses document storage.

Document structure

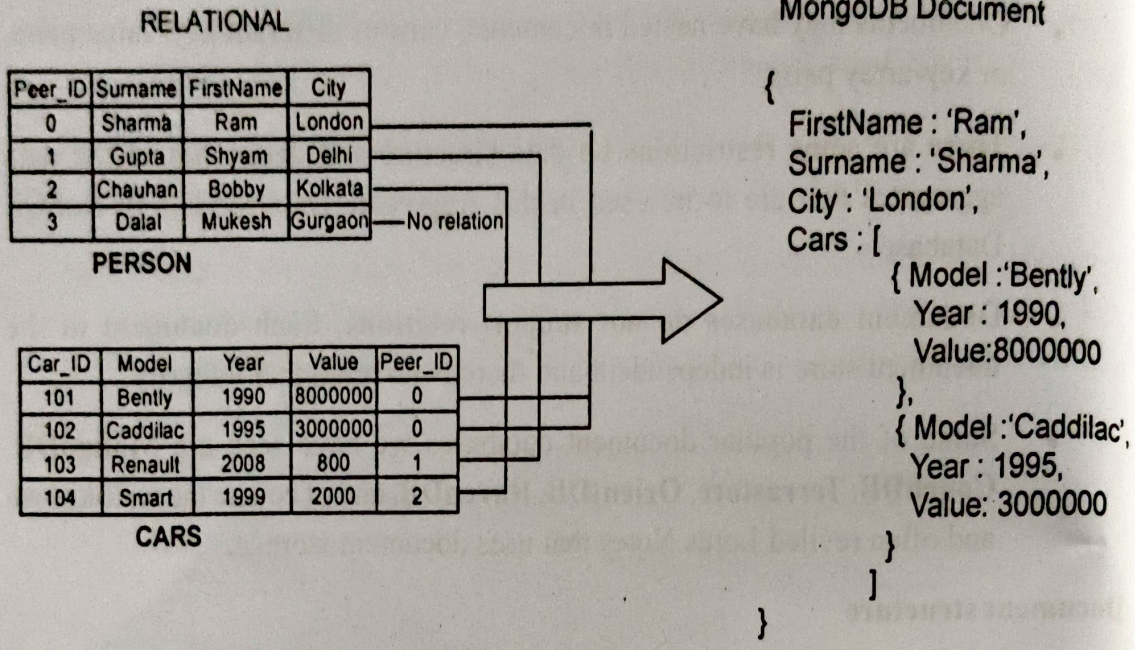
MongoDB documents are composed of field-and-value pairs and have the following structure:

```
{
 field1: value1,
 field2: value2,
 field3: value3,
 ....
 fieldN: valueN
}
```

Example:



In the above example, a document is contained within the curly braces. It contains multiple fields in "field":"value" format. Above, "_id", "firstName", and "lastName" are field names with their respective values after a colon :. Fields are separated by a comma. A single collection can have multiple such documents separated by a comma.



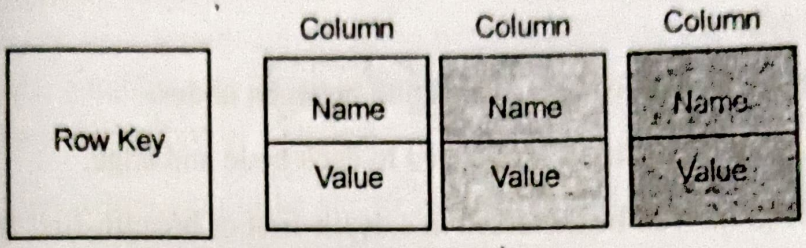
2.2.3 Wide-column stores or Column oriented data model

- It store data in tables, rows, and dynamic columns.
- Column-family databases organize their columns into column families.
- Each column has to be part of a single column family, and the column acts as unit for access, with the assumption that data for a particular column family will be usually accessed together.
- A column family data store is a multi-dimensional key value store (map or associative array) which is persistent (values persist after creation or access), distributed (data is distributed across multiple computing & storage nodes), sorted (sorted keys) and sparse (values for certain dimensions may not be populated, similar to sparsely populated rows in RDBMS).

It is also called a two-level map as it offers a two-level aggregate structure.

- ***Row-oriented:*** Each row is an aggregate (for example, customer with the ID of 1234) with column families representing useful chunks of data (profile, order history) within that aggregate.

- ***Column-oriented:*** Each column family defines a record type (e.g., customer profiles) with rows for each of the records. You then think of a row as the join of records in all column families.



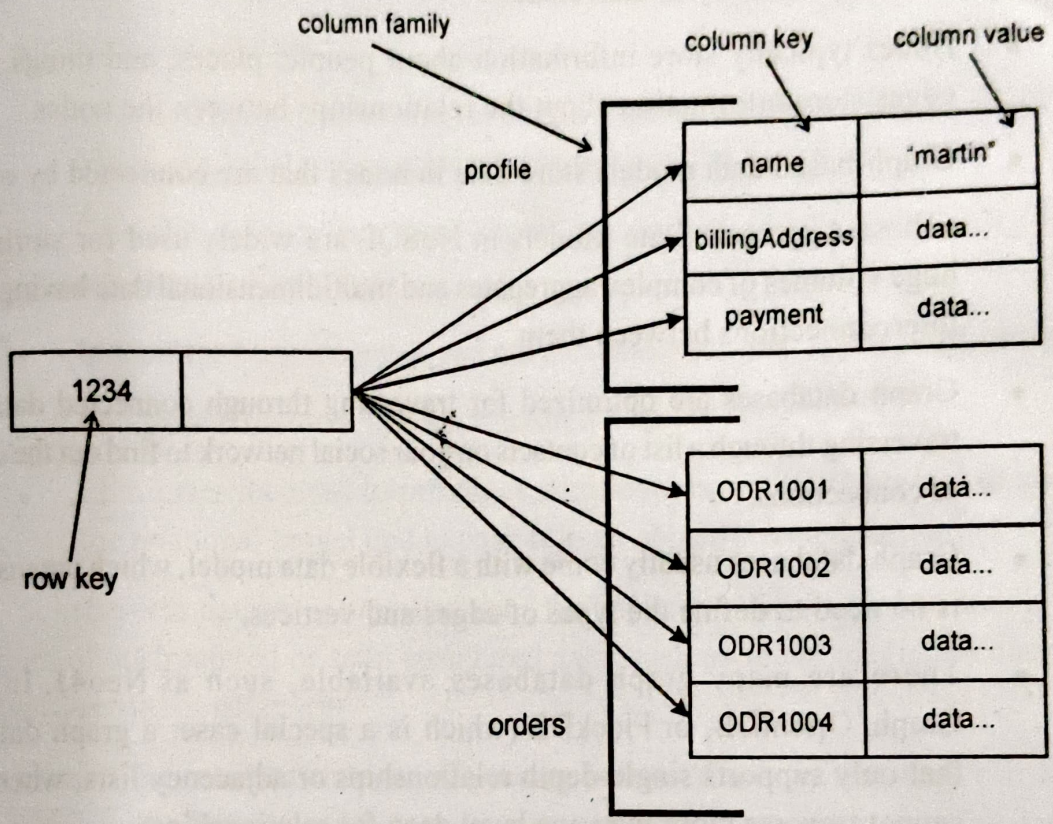


Fig: Representing customer information in a column-family structure

2.2.4 Graph databases

Graph Databases specific purpose is the storage of graph-oriented data structures.

I.e. Graph network database stores data in a graph like structure.

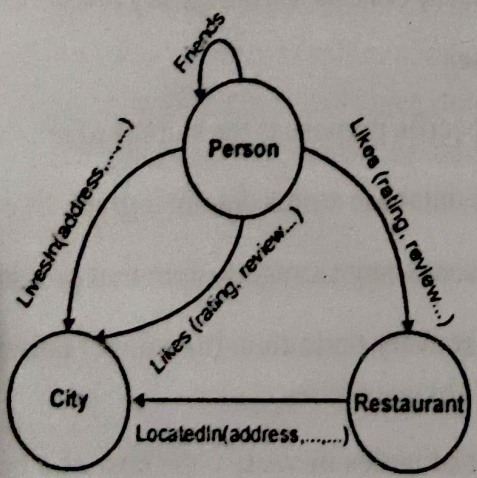
- A graph database is any storage system that provides index-free adjacency.
- This means that every node contains a direct pointer to its adjacent element and no index lookups are necessary.
- As the number of nodes increases, the cost of a hop remains the same.

it store data in nodes and edges.

A graph database maintains both entities and the relationships between them.

- The entity is represented as a node, while the relationships are represented as edges.
- An edge establishes a connection between nodes.
- A unique identifier is assigned to each node and edge.
- The graph can be traversed in a depth-first or breadth-first fashion, starting from any arbitrary or start node.
- Nodes typically store information about people, places, and things, while edges store information about the relationships between the nodes.
- Graph-based data models store data in nodes that are connected by edges.

  •These Aggregate Data Models in NoSQL are widely used for storing the huge volumes of complex aggregates and multidimensional data having many interconnections between them.
- Graph databases are optimized for traversing through connected data, e.g. traversing through a list of contacts on your social network to find out the degree of connections.
- Graph databases usually come with a flexible data model, which means there is no need to define the types of edges and vertices.
- There are many graph databases available, such as Neo4J, Infinite Graph, OrientDB, or FlockDB (which is a special case: a graph database that only supports single-depth relationships or adjacency lists, where you cannot traverse more than one level deep for relationships).



1.12 CROWD SOURCING ANALYTICS

- Crowdsourcing is the collection of information, opinions, or work from a group of people, usually sourced via the Internet.
- Crowdsourcing work allows companies to save time and money while tapping into people with different skills or thoughts from all over the world.
- While crowdsourcing seeks information or work, crowdfunding seeks money to support individuals, charities, or startup companies.
- The advantages of crowdsourcing include cost savings, speed, and the ability to work with people who have skills that an in-house team may not have.
- Crowdsourcing allows companies to farm out work to people anywhere in the country or around the world; as a result, crowdsourcing lets businesses tap into a vast array of skills and expertise without incurring the normal overhead costs of in-house employees.
- Crowdsourcing is becoming a popular method to raise capital for special projects. As an alternative to traditional financing options, crowdsourcing taps into the shared interest of a group, bypassing the conventional gatekeepers and intermediaries required to raise capital.
- Crowdsourcing usually involves taking a large job and breaking it into many smaller jobs that a crowd of people can work on separately.

1.12.1 Crowd Sourcing and Crowd Funding

While crowdsourcing seeks information or workers' labor, crowdfunding instead solicits money or resources to help support individuals, charities, or startups.

People can contribute to crowdfunding requests with no expectation of repayment, or companies can offer shares of the business to contributors.

Advantages

- Crowdsourcing brings together communities around a common project or cause

- Efficient way of solving time-intensive problems
- Deeper engagement by communities, who resonate and build loyalty to the product or solution.

Disadvantages

- Results can be easily skewed based on the crowd being sourced
- Lack of confidentiality or ownership of an idea
- Potential to miss the best ideas, talent, or direction and fall short of the goal or purpose.

1.12.2 Types of Crowdsourcing

Crowdsourcing involves obtaining information or resources from a wide swath of people. In general, we can break this up into four main categories:

Wisdom of the crowd:

- It's a collective opinion of different individuals gathered in a group.
- This type is used for decision-making since it allows one to find the best solution for problems.
- Many brands pay attention to the collective opinion of their customers because they help bring their businesses new ways of thinking, ideas, and strategies.
- As a result, the overall performance of a company improves.

Crowd creation:

- This type involves a company asking its customers to help with new products.
- This way, companies get brand new ideas and thoughts that help a business stand out. For instance, McDonald's is open to new ideas from its consumers.
- The famous fast food company asked customers to create their perfect burgers and submit their ideas to the brand.
- The company released winners' burgers each week, including the creator's short bio.

Crowd voting

- It's a type of crowdsourcing where customers are allowed to choose a winner. They can vote to decide which of the options is the best for them.
- This type can be applied to different situations. Consumers can choose one of the options provided by experts or products created by consumers.
- For instance, if a brand asks its consumers to create a new taste, package, or design of a product, other consumers vote to identify the best one.

Crowdfunding:

- It's when people collect money and ask for investments for charities, projects, and startups without planning to return the money to the owners.
- People do it voluntarily.
- Often, companies gather money to help individuals and families suffering from natural disasters, poverty, social problems, etc.