Explain about data architecture design.

**Data architecture Design:**

Data is one of several architecture domains that form the pillars of an enterprise architecture or solution architecture.

**Data architecture** is the models, policies, rules, and standards that govern which data is collected and how it is stored, arranged, integrated, and put to use in data systems and in organizations.

A data architecture aims to set data standards for all its data systems as a vision or a model of the eventual interactions between those data systems.

 For example, Data integration should be dependent upon data architecture standards since data integration requires data interactions between two or more data systems.

 A data architecture describes the data structures used by a business and its computer applications software.

Data architectures address data in storage, data in use, and data in motion; descriptions of data stores, data groups, and data items; and mappings of those data artifacts to data qualities, applications, locations, etc.

.

Essential to realizing the target state, data architecture describes how data is processed, stored, and used in an information system.

 Data Architecture provides criteria for data processing operations to make it possible to design data flows and also control the flow of data in the system.

The data architect is typically responsible for defining the target state, aligning during development and then following up to ensure enhancements are done in the spirit of the original blueprint.

During the definition of the target state, the data architecture breaks a subject down to the atomic level and then builds it back up to the desired form. The data architect breaks the subject down by going through three traditional architectural stages:

The data architecture is formed by dividing into three essential models and then are combined:



* **Conceptual model**
It is a business model which uses Entity Relationship(ER) model for relation between entities and their attributes.
* **Logical model**
It is a model where problems are represented in the form of logic such as rows and columns of data, classes, xml tags and other DBMS techniques.
* **Physical model**
Physical models hold the database design like which type of database technology will be suitable for architecture.

A data architect is responsible for all the design, creation, manage, deployment of data architecture and defines how data is to be stored and retrieved, other decisions are made by internal bodies.

Explain about data architecture .

Data architecture is a discipline that documents an organization's data assets, maps how data flows through its systems and provides a blueprint for managing data. The goal is to ensure that data is managed properly and meets business needs for information.

While data architecture can support operational applications, it most prominently defines the underlying data environment for business intelligence (BI) and advanced analytics initiatives. Its output includes a multilayer framework for data platforms and [data management](https://www.techtarget.com/searchdatamanagement/definition/data-management) tools, as well as specifications and standards for collecting, integrating, transforming and storing data.

Ideally, data architecture design is the first step in the data management process. But that usually isn't the case, which creates inconsistent environments that need to be harmonized as part of a data architecture. Also, despite their foundational nature, data architectures aren't set in stone and must be updated as data and business needs change. That makes them an ongoing concern for data management teams.

Data architecture goes hand in hand with [data modeling](https://www.techtarget.com/searchdatamanagement/definition/data-modeling), which creates diagrams of data structures, business rules and relationships between data elements. They're separate data management disciplines, though. In an article on [how data modeling and data architecture differ](https://www.techtarget.com/searchdatamanagement/tip/Data-modeling-vs-data-architecture-Whats-the-difference), David Loshin, president of consultancy Knowledge Integrity Inc., distinguished between modeling's micro focus on data assets and data architecture's broader macro perspective.

This guide to data architecture further explains what it is, why it's important and the business benefits it provides. You'll also find information on data architecture frameworks, best practices and more. Throughout the guide, there are hyperlinks to related articles that cover the topics in more depth.

**How have data architectures evolved?**

In the past, most data architectures were less complicated than they are now. They mostly involved structured data from transaction processing systems that was stored in relational databases. Analytics environments consisted of a [data warehouse](https://www.techtarget.com/searchdatamanagement/definition/data-warehouse), sometimes with smaller data marts built for individual business units and an operational data store as a staging area. The transaction data was processed for analysis in batch jobs, using traditional extract, transform and load (ETL) processes for data integration.

Starting in the mid-2000s, the adoption of [big data technologies in businesses](https://www.techtarget.com/searchdatamanagement/The-ultimate-guide-to-big-data-for-businesses) added unstructured and semistructured forms of data to many architectures. That led to the deployment of [data lakes](https://www.techtarget.com/searchaws/definition/data-lake), which often store raw data in its native format instead of filtering and transforming it for analysis upfront -- a big change from the data warehousing process. The new approach is driving wider use of ELT data integration, an alternative to ETL that inverts the load and transform steps.

The increased use of [stream processing](https://www.techtarget.com/searchdatamanagement/definition/stream-processing) systems has also brought real-time data into more data architectures. Many architectures now support artificial intelligence and machine learning applications, too, in addition to the basic BI and reporting driven by data warehouses. The shift to cloud-based systems further adds to the complexity of data architectures.

Another emerging architecture concept is the [data fabric](https://www.techtarget.com/searchdatamanagement/definition/data-fabric), which aims to streamline data integration and management processes. It has a variety of potential [use cases in data environments](https://www.techtarget.com/searchdatamanagement/feature/The-top-6-use-cases-for-a-data-fabric-architecture)



Data ingestion is the process of transporting data from one or more sources to a target site for further processing and analysis. This data can originate from a range of sources, including data lakes, IoT devices, on-premises databases, and SaaS apps, and end up in different target environments, such as cloud data warehouses or data marts.

Data ingestion is a critical technology that helps organizations make sense of an ever-increasing volume and complexity of data. To help businesses get more value out of data ingestion, we’ll dive deeper into this technology. We’ll cover types of data ingestion, how data ingestion is done, the difference between data ingestion and ETL, data ingestion tools, and more

A **staging area**, or **landing zone**, is an intermediate storage area used for data processing during the [extract, transform and load (ETL)](https://en.wikipedia.org/wiki/Extract%2C_transform%2C_load) process. The data staging area sits between the data source(s) and the data target(s), which are often [data warehouses](https://en.wikipedia.org/wiki/Data_warehouse), [data marts](https://en.wikipedia.org/wiki/Data_mart), or other data repositories.[[1]](https://en.wikipedia.org/wiki/Staging_%28data%29#cite_note-Oracle_Ref-1)

Data staging areas are often transient in nature, with their contents being erased prior to running an ETL process or immediately following successful completion of an ETL process. There are staging area architectures, however, which are designed to hold data for extended periods of time for archival or troubleshooting purpose

A Data Hub is a data exchange with frictionless data flow at its core. It can be described as a solution consisting of different technologies: Data Warehouse, Engineering, and Data Science. It’s rather a technology, but an approach to more effectively determine where, when, and for whom data needs to be mediated, shared, and then linked and/or persisted. Endpoints, which can be applications, processes, people, or algorithms, interact with the hub, potentially in real time, to provide data to or receive data from the hub.

Data profiling is the process of examining, analyzing, and creating useful summaries of data. The process yields a high-level overview which aids in the discovery of [data quality](https://www.talend.com/resources/what-is-data-quality/) issues, risks, and overall trends. Data profiling produces critical insights into data that companies can then leverage to their advantage.

More specifically, data profiling sifts through data to determine its legitimacy and quality. Analytical algorithms detect dataset characteristics such as mean, minimum, maximum, percentile, and frequency to examine data in minute detail. It then performs analyses to uncover metadata, including frequency distributions, key relationships, foreign key candidates, and functional dependencies. Finally, it uses all of this information to expose how those factors align with your business’s standards and goals.

Data profiling can eliminate costly errors that are common in customer databases. These errors include null values (unknown or missing values), values that shouldn’t be included, values with unusually high or low frequency, values that don’t follow expected patterns, and values outside the normal range.

**Data cleansing** or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate [records](https://en.wikipedia.org/wiki/Storage_record) from a record set, [table](https://en.wikipedia.org/wiki/Table_%28database%29), or [database](https://en.wikipedia.org/wiki/Database) and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the [dirty](https://en.wikipedia.org/wiki/Dirty_data) or coarse data. Data cleansing may be performed [interactively](https://en.wikipedia.org/wiki/Interactively) with [data wrangling](https://en.wikipedia.org/wiki/Data_wrangling) tools, or as [batch processing](https://en.wikipedia.org/wiki/Batch_processing) through [scripting](https://en.wikipedia.org/wiki/Script_%28computing%29).[[2]](https://en.wikipedia.org/wiki/Data_cleansing#cite_note-2)

After cleansing, a [data set](https://en.wikipedia.org/wiki/Data_set) should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different [data dictionary](https://en.wikipedia.org/wiki/Data_dictionary) definitions of similar entities in different stores. Data cleaning differs from [data validation](https://en.wikipedia.org/wiki/Data_validation) in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

### Why are data architectures important?

A well-designed data architecture is a crucial part of the data management process. It supports data integration and data quality improvement efforts, as well as data engineering and data preparation. It also enables effective [data governance](https://www.techtarget.com/searchdatamanagement/definition/data-governance) and the development of internal data standards. Those two things, in turn, help organizations ensure that their data is accurate and consistent.

A data architecture is also the foundation of a data strategy that supports business goals and priorities. In an article on [key data strategy components](https://www.techtarget.com/searchdatamanagement/tip/6-key-components-of-a-successful-data-strategy), Donald Farmer, principal of consultancy TreeHive Strategy, wrote that "a modern business strategy depends on data." That makes data management and analytics too important to leave to individuals, Farmer said. To manage and use data well, an organization needs to [create a comprehensive data strategy](https://www.techtarget.com/searchdatamanagement/tip/Developing-an-enterprise-data-strategy-10-steps-to-take), underpinned by a strong data architecture. 

These are the four main phases of developing a data strategy, according to Donna Burbank of consulting firm Global Data Strategy.

### What are the characteristics and components of a data architecture?

In an article on the [principles of modern data architectures](https://www.techtarget.com/searchdatamanagement/tip/5-principles-of-a-well-designed-data-architecture), Farmer stressed the importance of including both data governance and regulatory compliance processes and the growing need to support multi-cloud environments. He concluded by noting that data's potential business value will be wasted if a data architecture doesn't make it available for analytics uses.

"It's a cliché of modern data management that data is a business asset," Farmer wrote. "But data that just sits there is only a cost center, requiring maintenance without providing any business benefits."

Other common characteristics of well-designed data architectures include the following:

* a business-driven focus that's aligned with organizational strategies and data requirements;
* flexibility and scalability to enable various applications and meet new business needs for data; and
* strong security protections to prevent unauthorized data access and improper use of data.

From a purist's point of view, data architecture components don't include platforms, tools and other technologies. Instead, a data architecture is a conceptual infrastructure that's described by a set of diagrams and documents. Data management teams then use them to guide technology deployments and how data is managed.

Some examples of those components, or artifacts, are as follows:

* data models, data definitions and common vocabularies for data elements;
* [data flow diagrams](https://www.techtarget.com/searchdatamanagement/definition/data-flow-diagram-DFD) that illustrate how data flows through systems and applications;
* documents that map data usage to business processes, such as a CRUD matrix -- short for create, read, update and delete;
* other documents that describe business goals, concepts and functions to help align data management initiatives with them;
* policies and standards that govern how data is collected, integrated, transformed and stored; and
* a high-level architectural blueprint, with different layers for processes like data ingestion, data integration and data storage.



Follow these principles to help put your data architecture on the right track.

### What are the benefits of a data architecture?

Ideally, a well-designed data architecture helps an organization [develop effective data analytics platforms](https://www.techtarget.com/searchdatamanagement/feature/Building-a-strong-data-analytics-platform-architecture) that deliver useful information and insights. In companies, those insights improve strategic planning and operational decision-making, potentially leading to better business performance and competitive advantages. They also aid in various other applications, such as the diagnosis of medical conditions and scientific research.

Data architecture also helps improve data quality, streamline data integration and reduce data storage costs, among other benefits. It does so by taking an enterprise view compared to domain-specific data modeling or focusing on architecture at the database level, according to Peter Aiken, a data management consultant and associate professor of information systems at Virginia Commonwealth University.

"When we look at it from a data architecture perspective, we have the greater value

potential, and that's because we're looking at broad use [of data] across all of the databases," Aiken said during a Dataversity webinar in May 2021.

### What are the risks of bad data architecture design?

One data architecture pitfall is too much complexity. The dreaded "spaghetti architecture" [is evidence of that](https://data-sleek.com/what-is-spaghetti-architecture-and-how-to-avoid-it/), with a tangle of lines representing different data flows and point-to-point connections. The result is a ramshackle data environment with incompatible [data silos](https://www.techtarget.com/searchdatamanagement/definition/data-silo) that are hard to integrate for analytics uses. Ironically, data architecture projects often aim to bring order to existing messy environments that developed organically. But if not managed carefully, they can create similar problems.

Another challenge is getting universal agreement on standardized data definitions, formats and requirements. Without that, it's hard to create an effective data architecture. The same goes for putting data in a business context. Done well, data architecture "captures the business meaning of the data required to run the organization," Aiken said in the Data versity webinar. But failing to do so may create a disconnect between the architecture and the strategic data requirements it's supposed to meet.

### Data architecture vs. data modeling

Data modeling focuses on the details of specific data assets. It creates a visual representation of data entities, their attributes and how different entities relate to each other. That helps in scoping the data requirements for applications and systems and then designing database structures for the data, a process that's done through a progression of conceptual, logical and physical data models.

Data architecture takes a more global view of an organization's data to create a framework for data management and usage. But, as consultant Loshin wrote in his article comparing the two, data modeling and data architecture complement each other. Data models are a crucial element in data architectures, and an established data architecture simplifies data modeling, said Loshin, who also is director of the Master of Information Management program at the University of Maryland's College of Information Studies.

Rick Sherman, managing partner at consulting firm Athena IT Solutions, separately explained [seven techniques for modeling data](https://www.techtarget.com/searchdatamanagement/tip/7-data-modeling-techniques-and-concepts-for-business), including the entity-relationship, dimensional and graph modeling approaches that are most used now. He also outlined a set of data modeling best practices, including these recommendations:

* Gather both business and data requirements upfront, before building models.
* Develop data models iteratively and incrementally to make the process manageable.
* Use data models as a tool for communicating with business users about their needs.
* Manage data models just like any other type of application code.



Data management teams typically build these three types of data models in a phased process.

### Data architecture vs. information architecture and enterprise architecture

In a second article, Sherman described the [difference between data architecture and information architecture](https://www.techtarget.com/searchdatamanagement/tip/Data-architecture-vs-information-architecture-How-they-differ) in enterprise applications. "Information is data in context," he wrote. "An information architecture defines the context that an enterprise uses for its business operations and management." A data architecture that delivers high-quality, reliable data is the foundation for the information architecture, he added.

Meanwhile, data architecture is commonly viewed as a subset of [enterprise architecture](https://www.techtarget.com/searchcio/definition/enterprise-architecture) (EA), which aims to create an organizational blueprint for an organization in four domains. EA also encompasses the following:

* business architecture, which involves business strategy and key business processes;
* application architecture, which focuses on individual applications and their relationships to business processes; and
* technology architecture, which includes IT systems, networks and other technologies that support the other three domains.

### What data architecture frameworks are available?

Organizations can use standardized frameworks to design and implement data architectures instead of starting completely from scratch. These are three well-known framework options:

**DAMA-DMBOK2.** The DAMA Guide to the Data Management Body of Knowledge is a data management framework and reference guide created by DAMA International, a professional association for data managers. Now [in its second edition](https://datacrossroads.nl/2018/06/25/dama-dmbok-in-a-nutshell/) and commonly known as DAMA-DMBOK2, the framework addresses data architecture along with other data management disciplines. The first edition was published in 2009, and the second one became available in 2017.

**TOGAF.** Created in 1995 and updated several times since then, TOGAF is an enterprise architecture framework and methodology that includes a section on data architecture design and roadmap development. It was developed by The Open Group, and TOGAF initially stood for The Open Group Architecture Framework. But it's now referred to simply as the TOGAF standard.

**The Zachman Framework.** This is an ontology framework that uses a 6-x-6 matrix of rows and columns to describe an enterprise architecture, including data elements. It doesn't include an implementation methodology; instead, it's meant to serve as the basis for an architecture. The framework was originally developed in 1987 by John Zachman, an IBM executive who retired from the company in 1990 and founded a consulting firm called Zachman International.

### Key steps for creating a data architecture

Data management teams must work closely with business executives and other end users to develop a data architecture. If they don't, it may not be in tune with business strategies and data requirements. Engaging with senior execs to get their support and meeting with users to understand their data needs are two of the [nine data architecture planning steps](https://www.techtarget.com/searchdatamanagement/tip/9-steps-to-a-dynamic-data-architecture-plan) that consultant Loshin listed in an article.

Among other steps, he also recommended that organizations do the following:

* evaluate data risks based on data governance directives;
* track data flows, as well as data lifecycle and data lineage info;
* document and appraise the existing data management technology infrastructure; and
* scope out a roadmap for the data architecture deployment projects.

Another article, by technology writer George Lawton, provides tips on [building a cloud-based architecture](https://www.techtarget.com/searchdatamanagement/tip/How-to-build-a-successful-cloud-data-architecture) for data management and analytics. It also outlines potential challenges that data management teams can face in the cloud, including data security requirements, regulatory compliance mandates and data gravity issues that can complicate migrations of data sets.

Explain data processing and its steps.

## Data processing:

Data processing is a process of converting raw facts or data into a meaningful information

 

**IMPORTANCE OF DATA PROCESSING**

* Whether you use the internet to learn about a certain topic, complete financial transactions online, order food, online shopping etc., data is being generated every single second.
* The use of social media, online shopping and video streaming services have all added to the increase in the amount of data.
* A study by Domo estimates that 1.7MB data is created every second for every human being on the planet in 2020.
* Data in its raw form is not useful to any organization.
* In order to utilize and get insights from such a huge amount of data - data processing comes into play.
* Data processing is crucial for organizations to create better business strategies and increase their competitive edge.
* By converting the data into a readable format like graphs, charts, and documents, employees throughout the organization can understand and use the data.

## Stages of Data Processing

Data processing consists of following 6 stages:

 

1. **Data Collection:** The raw data is collected from various sources like excel file, database, text file, and unorganised data such as audio clips, images, GPRS and video clips.

The type of raw data collected has a huge impact on the output produced. Hence, it can be gathered from defined and accurate sources so that the subsequent findings are valid and usable.

Raw data can include monetary figures, website cookies, profit/loss statements of a company, user behaviour, etc.

**2. Data Preparation:** Once the data is collected, it then enters the stage, Data preparation, often referred to as “pre-processing”.

 In this stage the raw data is cleaned up by checking for errors, duplication, miscalculations or missing data.

Then transformed into a suitable form for the following stage of data processing.

The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

**3. Data Input:** The prepared data is translated into a machine readable form by using a CRM (Customer Relationship Management) like Salesforce or a data warehouse like Redshift.

Data input is the first stage in which raw data begins to take the form of usable information.

**4. Processing:** During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms.

The process may vary slightly depending on

* The source of data being processed (like data lakes, social networks, connected devices etc.)
* Its intended use (like examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

**5. Data Interpretation**: The data is converted into videos, graphs, images and plain text. The non-data scientists find this data very helpful. Members of a company can start analysing this data and applying it into their projects.

**6. Data Storage:**  After the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

**DEFINITION:**

Data processing is the method of collecting raw data and translating it into usable information in a step-by-step process by a team of data scientists and data engineers in an organization and stored for future purpose.

In simple words the raw data is collected, filtered, sorted, processed, analysed, stored, and then presented in a readable format**.**

**TOOLS:**

The most commonly used tools for data processing are Storm, Hadoop, HPCC, Statwing, Qubole and CouchDB. The output is worthwhile information various file formats like a chart, audio, table, graph, image, vector file depending on software or application necessary

**DIFFERENT TYPES OF OUTPUT**

The different types of output files in data processing are –

* **Plain Text File** – The text file is the simplest format of a data file. It will be exported as Notepad or WordPad files.
* **Table/Spreadsheet** – the data is represented in columns and rows that helps in quick analysis and understanding of data. Tables/ Spreadsheet allows numerous operations like sorting & filtering in descending/ascending order and statistical operations.
* **Charts and graphs** – The most common features in almost all software is the graphs and charts format. This format enables easy analysis of data by just a glance.
* **Maps/Vector or Image File** – The requirement to store and analyse spatial data and export data can be fulfilled by this image and map formats.
* Specialised software can process software specific file formats.

 **DIFFERENT DATA PROCESSING METHODS**

The three prominent data processing methods are as follows:

* **Manual Data Processing:** Data is processed manually. The entire procedure of data collecting, filtering, sorting, calculation and alternative logical operations is all carried out with human intervention without using any electronic device or automation software. It’s a low-priced methodology and needs very little to no tools; however, it produces high errors and requires high labour prices and much of your time.
* **Mechanical Data Processing:** Data is processed using machines and simple devices such as typewriters, calculators, printing press, etc. Simple data processing operations can be accomplished by this method. There are fewer errors compared to manual data processing, but the only drawback is that this method cannot be utilized with the increase of data.
* **Electronic Data Processing:**  Data processing softwares and programs are used to process data. A series of instructions are given to the software to process the data and produce the desired output. It is more expensive but provides faster processing with the highest reliability and accuracy.

**TYPES OF DATA PROCESSING**

The types of data processing are as below:

* **Batch Processing:** The collection and processing of data is done in batches where there is a huge quantity of data.

 E.g., the payroll system.

**Real-time processing:** For a small quantity of data, real-time processing is done where data can be processed within seconds of data input.

E.g., withdrawing money from ATM

* **Online Processing:** As and when data is available, it is automatically entered in the CPU. This is useful for processing of data continuously.

E.g., barcode scanning

* **Multiprocessing:** This also goes by the name parallel processing, where data is fragmented into small frames and processed in two CPUs within a single computer system.

E.g., weather forecasting

* **Timeshared Processing:** Allocates computer resources and data in time slots to several users simultaneously.

**WHY WE SHOULD USE DATA PROCESSING**

In the modern era, most of the work relies on data, therefore collection of large amounts of data for different purposes like academic, scientific research, institutional use, personal and private use, for commercial purposes and lots more. The processing of this data collected is essential so that the data goes through all the above-stated steps and gets sorted, stored, filtered, presented in the required format and analyzed.

The amount of time consumed and the intricacy of processing will depend on the required results. In situations where large amounts of data are acquired, the necessity of processing to obtain authentic results with the help of data processing in data mining and data processing in data research gets inevitable.

Examples of Data Processing

Data processing occurs in our daily lives whether we may be aware of it or not. Here are some real-life examples of data processing:

* A stock trading software that converts millions of stock data into a simple graph
* An e-commerce company uses the search history of customers to recommend similar products
* A digital marketing company uses demographic data of people to strategize location-specific campaigns
* A self-driving car uses real-time data from sensors to detect if there are pedestrians and other cars on the road

**Explain about data Quality and Preprocessing**

 **DATA PREPROCESSING**

**Data Quality**

Data has quality if it satisfies the requirements of its intended use.

Two different users may have very different assessments of the quality of a given database.

There are many factors comprising data quality.

These include:

 Accuracy,

 Completeness,

 Consistency,

 Timeliness,

 Believability,

 Interpretability.

**Poor Data :**

**Incomplete -** lacking attribute values or certain attributes of interest, or containing only aggregate data,

For example, want to include whether each item purchased was advertised as on sale for analysis, but the information has not been recorded.

**Inaccurate or noisy -** containing errors, or values that deviate from the expected.

**Inconsistent –** containing discrepancies in the department codes used to categorize items.

**REASONS FOR INACCURATE OR INCORRECT DATA**

* The data collection instruments used may be faulty.
* There may have been human or computer errors occurring at data entry.
* Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal

Information, e.g., by choosing the default value ‘January 1’ displayed for birthday (This is known as disguised missing data.)

* Errors in data transmission can also occur.
* Technology limitations, such as limited buffer size for coordinating synchronized data transfer and consumption.
* Inconsistencies in naming conventions or data codes used
* Inconsistent formats for input fields, such as date.

**REASONS FOR INCOMPLETE DATA**

* Attributes of interest may not always be available.
* Other data may not be included simply because they were not considered important at the time of entry.
* Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions.
* Data that were inconsistent with other recorded data may have been deleted.
* The recording of the history or modifications to the data may have been overlooked.

**Factors that affect Data Quality:**

**Missing data**, particularly for tuples with missing values for some attributes, may need to be inferred.

**Timeliness** also affects data quality.

**Believability** reflects how much the data are trusted by users.

**Interpretability** reflects how easy the data are understood.

**Major Tasks in Data Pre-processing**

 In this section, we look at the major steps involved in data pre-processing, namely,

**I .Data cleaning**

**II. Data integration**

**III. Data reduction**

**IV. Data transformation**

 Data Cleaning

 Data Integration

Data Transformation

 22, 32,100, 59, 48-🡪20.02,0.32,1.00, 0.59, 0.48

Data Reduction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|   |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**I. Data Cleaning**

Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning or data cleansing routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

**I.1 Missing Values:**

 Missing values can be filled in using the following ways.

**1. Ignore the tuple:** This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably. By ignoring the tuple, we do not make use of the remaining attributes values in the tuple that could have been useful to the task at hand.

 **2. Fill in the missing value manually:** In general, this approach is time consuming and may not be feasible given a

large data set with many missing values.

 **3. Use a global constant to fill in the missing value:** Replace all missing attribute values by the same constant, such as a label like “Unknown” or −∞. If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of “Unknown.” Hence, although this method is simple, it is not foolproof.

 **4. Use a measure of central tendency for the attribute (such as the mean or median) to fill in the missing value:** Measures of central tendency indicate the “middle” value of a data distribution. For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median.

 For example, suppose that the data distribution regarding the income of customers is symmetric and the average income is $56,000,use this value to replace the missing value for income.

 **5. Use the attribute mean or median for all samples belonging to the same class as the given tuple:** For example, if classifying customers according to credit risk, we may replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.

**6. Use the most probable value to fill in the missing value:** This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

**I.2 Noisy Data**

Noise is a random error or variance in a measured variable. It can be smoothed by using different techniques.

Given a numeric attribute such as, say, price, how can we “smooth” out the data to remove the noise?

 Let’s look at the following data smoothing techniques:

 **1. Binning:**

Binning methods smooth a sorted data value by consulting its “neighborhood,” that is, the values around it. The sorted values are

 Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

 Partition into (equal-frequency) bins:

 Bin 1: 4, 8, 15

 Bin 2: 21, 21, 24

 Bin 3: 25, 28, 34

 **Smoothing by bin means:**

Bin 1: 9, 9, 9

 Bin 2: 22, 22, 22

 Bin 3: 29, 29, 29

**Smoothing by bin boundaries:**

 Bin 1: 4, 4, 15

 Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

Binning methods for data smoothing.

Data values are sorted and distributed into a number of “buckets,” or bins.

**Note:** Because binning methods consult the neighbourhood of values, they perform local smoothing.

Above figure illustrates some binning techniques.

In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values).

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

For example, the mean of the values 4, 8, and 15 in Bin 1 is 9.

Therefore, each original value in this bin is replaced by the value 9.

Similarly, **smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median.

In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing.

**2. Regression:**

 Data smoothing can also be done by conforming data values to a function, a technique known as regression.

 **Linear regression** involves finding the “best” line to fit two attributes (or variables), so that one attribute can be used to predict the other.

**Multiple linear regression** is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

 **3. Outlier analysis:**

Outliers may be detected by clustering, for example

,

Figure : A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster centroid is marked with a “+”, representing the average point in space for that cluster. Outliers may be detected as values that fall outside of the sets of clusters.

where similar values are organized into groups, or “clusters.” Intuitively, values that fall outside of the set of clusters may be considered outliers.

Many methods for data smoothing are also methods for data discretization (a form of data transformation) and data reduction.

 **II. Data Integration**

The merging of data from multiple data stores. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set which leads to improve the accuracy and speed of the subsequent mining process.

The semantic heterogeneity and structure of data pose great challenges in data integration. Following are the ways to handle this problem:

**II.1** **The Entity Identification Problem:**

 How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identification problem.

For example, how can the data analyst or the computer be sure that *customer id* in one database and *cust number* in another refer to the same attribute? Metadata for each attribute include the name, meaning, data type, and range of values permitted for the attribute, and null rules for handling blank, zero, or null values . Such metadata can be used to help avoid errors in schema integration.

**II.2 Redundancy and Correlation Analysis:**

An attribute such as *annual revenue*, may be redundant if it can be “derived” from another attribute or set of attributes. Some redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other. For nominal data, we use the *χ*2 (*chi- square*) test. For numeric attributes, we can use the *correlation coefficient* and *covariance*, both of which access how one attribute’s values vary with those of another.

**III. Data Reduction**

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

Data reduction strategies include *dimensionality reduction*, *numerosity reduc- tion*, and *data compression*.

**III.1Dimensionality reduction**

 The process of reducing the number of random variables or attributes under consideration. Dimensionality reduction methods include *wavelet transforms* and *principal components analysis* which transform or project the original data onto a smaller space.

 *Attribute subset selection* is another method in which irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.

**III.2 Numerosity reduction**

 This method replaces the original data volume by alternative, smaller forms of data representation. These techniques may be parametric or nonparametric.

 For *parametric methods*, only the data parameters need to be stored, instead of the actual data. Regression and log-linear models are examples. *Nonparametric methods* include *histograms,* *clustering*, *sampling*, *data cube aggregation, and compression are used*.

## **IV. Data Transformation**

The data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

### IV.1 Overview of Data Transformation Strategies

In *data transformation*, the data are transformed or consolidated into forms appropriate for mining. Strategies for data transformation include the following:

**IV.1 Smoothing:**

Works to remove noise from the data. Such techniques include binning, regression, and clustering.

**IV.2 Attribute construction (or *feature construction*):**

 New attributes are constructed and added from the given set of attributes to help the mining process.

**IV.3 Aggregation:**

Summary or aggregation operations are applied to the data.

 For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple levels of abstraction.

**IV.4 Normalization**:

The attribute data are scaled so as to fall within a smaller range, such as −1*.*0 to 1*.*0, or 0*.*0 to 1*.*0.

**IV.5 Discretization:** The raw values of a numeric attribute (such as *age*) are replaced by interval labels e.g., 0-10, 11-20, and so on or conceptual labels e.g., *youth, adult*, and *senior* .

# Understanding various Sources of Data

Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.

In the process of big data analysis, “Data collection” is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as “knowledge”. Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc. The main goal of data collection is to collect information-rich data.

Data collection starts with asking some questions such as what type of data is to be collected and what the source of collection is. Most of the data collected are of two types known as “qualitative data“ which is a group of non-numerical data such as words, sentences mostly focus on behaviour and actions of the group.Another one is “quantitative data” which is in numerical forms and can be calculated using different scientific tools and sampling data.

# The actual data is then further divided mainly into two types known as:

* + **Primary data**

# Secondary data



1. **Sources of Primary data:**

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys.

The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

# Interview method:

The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

# Survey method:

`

The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analysing data. Examples are online surveys or surveys through social media polls.

# Observation method:

The observation method is a method of data collection in which the researcher keenly observes the behaviour and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behaviour towards the products. The data obtained will be sent for processing.

# 4 .Experimental method:

The experimental method is the process of collecting data through performing **experiments,**

research, and investigation.

The most frequently used experiment methods are CRD, RBD, LSD, and FD.

**CRD- Completely Randomized design** is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.



# CRD: Experiment treated as a single unit

**RBD- Randomized Block Design** is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.

# RBD :Experiment divided into small units called BLOCKS.

**LSD – Latin Square Design** is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is a balanced arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Interchange of any rows or columns

will not cause any disturbance in arrangement .Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.

# LATIN SQUARE DESIGN: Experiment divided into NxN squares

* + **FD- Factorial design** is an experimental design that allows the experimenter to test two or more independent variables simultaneously. It also measures interaction effects of the variables and analyses the impacts of each of the variables.

# Factorial design can be depicted with a numbering in terms of levels of each of the

**factors.**

# Source of Secondary data:

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data sources and it has two types of sources named internal sources and external sources.

# 1 .Internal sources:

These types of data can easily be found within the organization

# Accounting Resources:

This gives so much information which can be used by market researcher

* **Internal Experts:** These are the people who are heading various departments. They can give the idea of how a particular thing is working.
* **Miscellaneous Reports:** Information we can get from operational reports. The cost and time consumption is less in obtaining internal sources.

If the data available within organization are unsuitable and inadequate, the marketer should extend the research to the external resources.

# 2.External sources:

The data which can’t be found at internal organizations and can be gained through external third party resources is external source data.

External data can be divided into following classes:

# a.Government publications:

Government resources provide a rich pool ofdata at free of cost on internet websites.

* + **Registrar general of India**-> generates demographic data like gender, age, occupation, etc.
	+ **Central Statistical Organization**-> publishes National Accounts Statistics such as estimates of national income for several years, growth rate, etc.
	+ **Ministry of Commerce and Industries**->provides information on wholesale price index related to food, power, fuel, food grains etc.
	+ **Planning Commission**-> provides statistics of Indian Economy.
	+ **RBI**-> Provides information on banking savings and investments, currency and finance report.
	+ **Labour Beuro**-> provides information on skilled,unskilled ,white collared jobs, etc.
	+ **National Sample Survey**->This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.
	+ **Department of Economic Affairs**-> It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.
	+ **State Statistical Abstract**-> This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

The cost and time consumption is more because this contains a huge amount of data.

**b.Non-Government Publications**- These includes publications of various industrial and trade associations, such as

* Indian Cotton Mill Association
* Various chambers of commerce
* Bombay Stock Exchange
* Various Associations of Press Media.
* Export Promotion Council.
* Confederation of Indian Industries (CII)
* Small Industries Development Board of India(SIDBI)
* Different Mills like - Woollen mills, Textile mills, etc.

# c. Syndicate Services-

These services are provided by certain organizations which collect and tabulate the marketing information on a regular basis for a number of clients who are the subscribers to these services.

So the services are designed in such a way that the information suits the subscriber. These syndicate services provide data from both household as well as institutions. **d.International Organizations-**

These includes

* The International Labour Organization (ILO)- It publishes data on the total and active population, employment, unemployment, wages and consumer prices
* The Organization for Economic Co-operation and Development (OECD) - It publishes data on foreign trade, industry, food, transport, and science and technology.
* The International Monetary Fund (IMF) - It publishes reports on national and international foreign exchange regulations.

# Other sources:

* **SENSORS:** Sensor data is the output of a device that detects and responds to some type of input from the physical environment. The output may be used to provide information or input to another system or to guide a process. Sensors can be used to detect just about any physical element.
	+ **Photosensor-** detects the presence of visible light, infrared transmission (IR) and/or ultraviolet (UV) energy. Ex; card readers, remote controls.
	+ **Lidar-**Light Detection and Ranging is a laser-based method of detection, range-finding and mapping devices, typically uses a low-power, eye-safe pulsing laser working in conjunction with a camera.

A Lidar instrument consists of a laser, a scanner, a specialized GPS receiver.

Airplanes and helicopters are the most commonly used platforms for acquiring Lidar data over broad areas.

* + **Charge-coupled device (CCD)–** is a light-sensitive device that stores and displays the data for an image in such a way that each pixel is converted into an electrical charge, the intensity of which is related to a color in the color spectrum.
	+ **Smart grid sensors**- can provide real-time data about grid conditions, detecting outages, faults, load and triggering alarms.
	+ [Wireless Sensor Networks-](https://searchdatacenter.techtarget.com/definition/sensor-network) combine specialized transducers with a communications infrastructure for monitoring and recording conditions at diverse locations. Commonly monitored parameters include temperature, humidity, pressure, wind direction and speed, illumination intensity, vibration intensity, sound intensity

# 2.SIGNALS:

Signal is **the real pattern, the repeatable process that we hope to capture and describe**. It is the information that we care about. The signal is what lets the model generalize to new situations. The noise is everything else that gets in the way of that.

# Separating signal from noise

When we are building a model, we are making the assumption that our data has two parts, signal and noise. Signal is the real pattern, the repeatable process that we hope to capture and describe. It is the information that we care about. The signal is what lets the model generalize to new situations.

It’s easy to picture the difference between signal and noise if you imagine listening to your favorite playlist in the middle of winter while there is a heater running nearby. The music is the signal. That’s the thing that you want to track and absorb. The heater fan is noise. It is additional variation piled on top of the signal. And if it gets too loud, it becomes impossible to follow the flow of the signal.

# The challenge:

It is the goal of models to describe the signal, despite the noise.A perfect model describes the signal exactly, and ignores all of the noise. If a model fails to capture all of the signal, that type of error is called **bias**.

If a model captures of some of the noise, that type of error is called **variance**.

Too much bias in our model means that it will perform poorly in all situations because it hasn’t captured the signal well. You may also hear this called **underfitting**.

This was the case when we fit a straight line to our temperature data. It didn’t capture the underlying pattern well, and because of that, had a much higher error then the rest of our candidates.