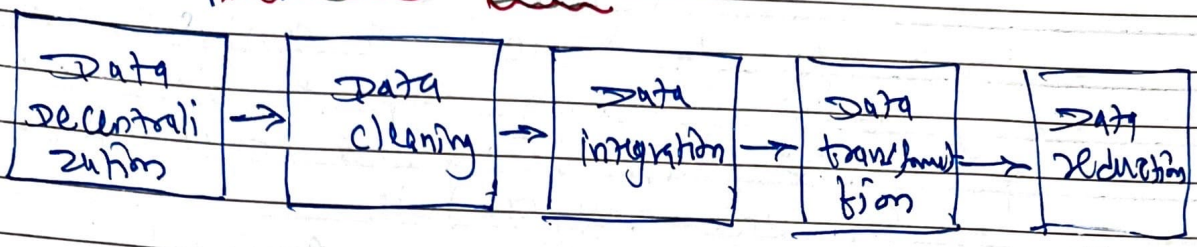
Date :
Page : 1

UNIT I

Data preparation :-

⇒ Data preparation is the process of cleaning and transforming raw data prior to processing and Analysis.

Data preparation steps.



How to approach new data science problem?
or

Data preparation process

→ Following tasks.

1. **Data collection :-**
   gathered data from operational systems, data warehouses, and various data sources. Data scientist, BI team, data professionals collect data.

2. **Data discovery and profiling :-**
   To explore collected data to better understand what it contains and what need to be done.
   Data profiling identifies patterns relationships and other attributes in data. as well as inconsistencies anomalies, missing values and other issues can be addressed.

3 Data cleansing
identifies data errors and issues are corrected to create complete and accurate dataset. eg. removed or fixed, missing values.

4. Data structuring :—
The data needs to be modified and organized to met the analytics requirements. eg. csv (comma separated file) .xlsx, .avro
This data in various format must accessible to BI and analytics tools.

5. Data transformation and enrichment :—
Data must be transformed into a unified and usable format. eg. creating new fields or column, aggregate values.

6. Data validation and publishing :—
Data to validate its consistency, completness and accuracy.
The prepared data is then stored in a data warehouse or data repository.

What are the challenges of data preparation?

→ The 80/20 rule is often applied to analytics applications with about 80% of the work said to be devoted to collecting and preparing data and only 20% to analyzing it.

Data preparation challenges:-

1. Inadequate or nonexistent data profiling:—
   If data isn't properly profiled, errors, anomalies and other problems might not be identified, which can result in flawed analytics.
2. missing or incomplete data:—
   Data sets often have missing values and other forms of incomplete data, such issues need to be assessed as possible errors and addressed if so.
3. Invalid data values:—
   misspellings, other typos and wrong numbers are examples of invalid entries that frequently occur in data and must be fixed to ensure analytics accuracy.
4. Name and address standardization:—
   Names and addresses may be inconsistent in data from different systems, with variations that can affect views of customers and other entities.
5. Inconsistent data across enterprise systems:—
   Other inconsistencies in data sets drawn from multiple source systems, such as different terminology and unique identifiers, are also a pervading issue in data preparation efforts.

6. Data enrichment:-
Deciding how to enrich a dataset eg what to add to it, is a complex task that requires a strong understanding of business needs and analytics goals.

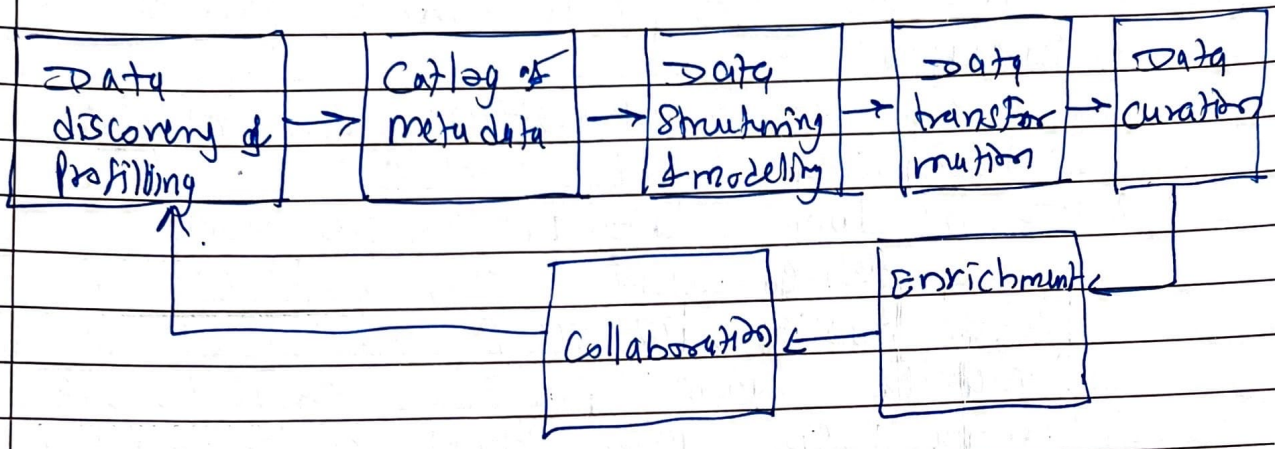
7. maintaining and expanding data prep. process:-
Data preparation work often becomes a recurring process that need to be sustained and enhanced on an ongoing basis.

Data preparation tools used in company

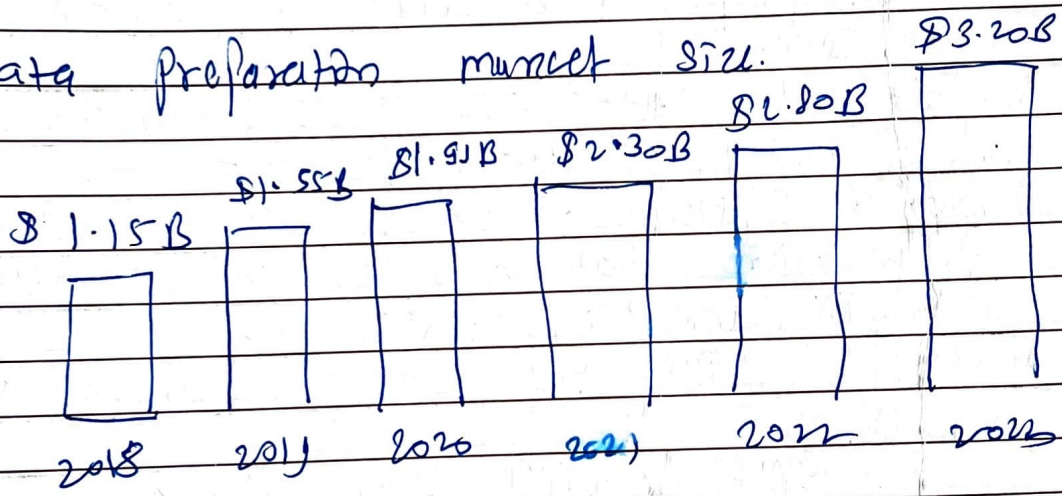
1. Altair
2. Boomi
3. Datameer
4. DataRobot
5. IBM
6. informatica
7. microsoft
8. precisely
9. SAP
10. SAS
11. Tableau
12. Talend
13. Tamr
14. Tibco Software.

Date :
Page : 1

Core Features of self service data preparation tools



Data preparation market size.



08/09/2022
..day.

How to get started on data preparation.

1 Think of data preparation as part of data analysis: —

Data preparation and analysis are "two sides of the same coin".

2. Define what data preparation success means:

Desired data accuracy levels and other data quality metrics should be set as goals balanced against projected costs to create a data prep. plan that's appropriate to each use case.

3 Prioritize data sources based on the application:

Resolving differences in data from multiple source systems is an important element of data preparation that also should be based on the planned analytics use case.

4 Use the right tools for the job and your skill level: —

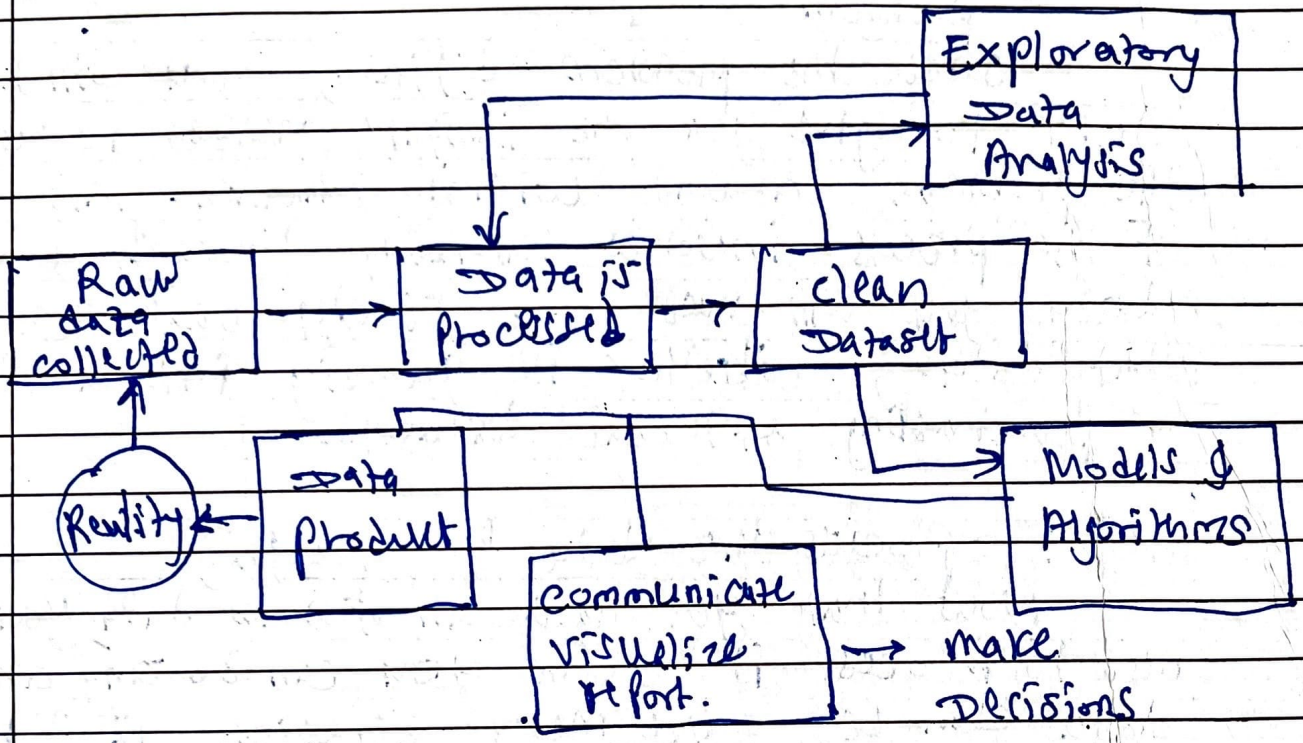
Self service data preparation tools aren't the only option available — other tools and technologies can also be used, depending on your skills and data needs.

5. Be prepared for failures when preparing data: —

Error handling capabilities need to be built into the data preparation process to prevent it from going awry or getting bogged down when problems occurs.

6. Keep an eye on data preparation costs: — The cost of software licenses, preprocessing and storage resources, and the people involved in preparing data should be watched closely to ensure that they don't get out of hand.

Explain ⇒ata science process.



Step I: Frame the problem.-
To define exactly what is problem.
eg Sales person of VP.
you should ask questions like the following.
1. Who are the customers.
2. why are they buying our product.
3. How do we predict if a customer is going to buy our product.
4. What is different from segments who are performing well and those that are performing below expectations.
5. How much money will we lost if we don't actively sell the product to these people

Step2 Collect the raw data needed for your problem.

once the problem defined, you will need data to give you the insights needed to turn the problem around with a solution. This part of the process involves thinking through what data you will need and finding ways to get that data., whether its querying internal databases or purchasing external datasets.

Step3:- Process the data for analysis

Now that you have all of raw data, you will need to process it before you can do any analysis data can be quite messy; especially if it hasn't been well maintained. you will see errors that will corrupt your analysis, values set to null through they really are zero, duplicate values, & missing value.

Step4:- Explore data.

Step5:- perform in depth analysis

Step6:- Communicate results of the analysis

Difference between Data science and business intelligence

Data science	Business intelligence
1. It is a field that uses mathematics, statistics & various other tools to discover the hidden pattern in the data.	1. It is basically a set of technologies applications & processes that are used by the enterprises to business data analysis.
2. It focuses on the future.	2. It focuses on the past and present.
3. It deals with both structured as well as unstructured data.	3. It mainly deals only with structured data.
4. Data science is much more flexiable as data sources can be added as per requirement.	4. It is not flexible as in case of business intelligence data sources need to be pre planned.
5. It makes use of the scientific method.	5. It makes use of the analytic method
6. It has a higher complexity in comparision to business intelligence.	6. It is much simpler when compared to data science.
7. It is expertise is data scientst.	7. It is expertise is the business user.
8. It deals with the questions of what will happen and what if	7. It deals with the question of what happened.
9. The data to be used to disseminated in real time clusters.	9. Data warehouse is utilized to hold data.

10 The ETL [Extract-Load-Transform] process is generally used for the integration of data for data science applications.

11 IT IS tools are SAS, Bigml, MATLAB, Excel.

12 reduce risk and increase income.

13 greater business value achieved.

14. Hadoop IT utils for data storage.

The ETL process is generally used for the integration of data for BI application.

11. Insight squared, Sales Analytics, Klipfolio, Knowhtspot, Cyfe, TIBCO, Spotfire

12. Helps in performing root cause analysis on a failure or to understand the current state.

13 Offer business value

14 No tools.

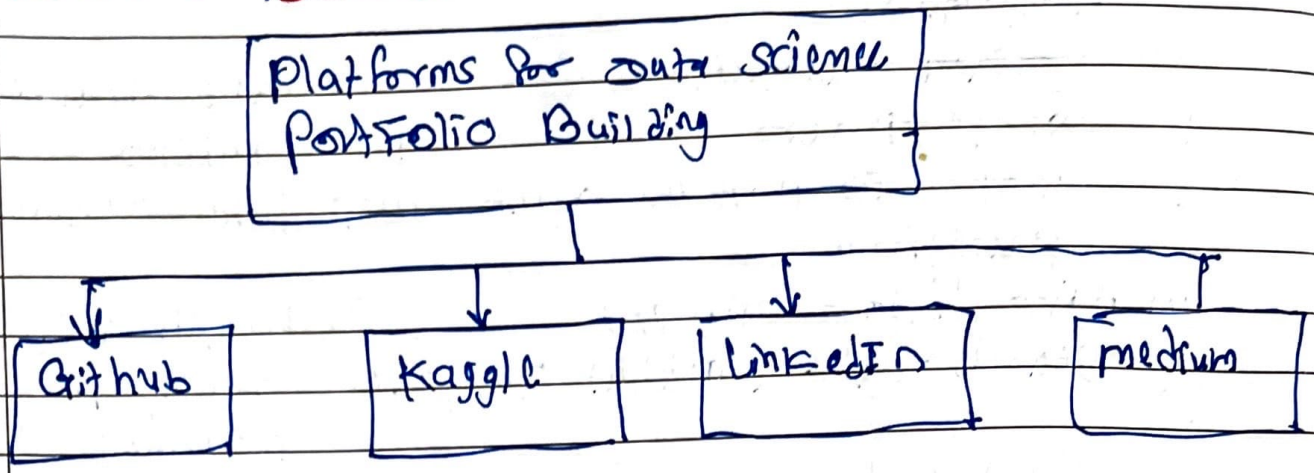
Difference between data science and data Analytics.

	Data science	Data Analytics
1	scope - Macro	① micro
2.	goal - To ask the right question.	② To find actionable data.
③	Major field - ML, AI search engine corporate analytics	③ Health care, gaming, travel, industries with immediate data needs
④	Visiy Big data - yes	yes.

Difference between data analyst and data scientist.

	Data analyst	Data Scientist.
①	Functional maths, statistics	① Advanced statistical, predictive analysitics.
②	basic fluency in R, Python, SQL	② Advanced oop.
③	SAS, Excel, BI: SW tools.	③ Hadoop, mysql, Tensorflow, Spark.
④	Analytical thinking, data visualisation.	④ ML, data modility

List out platforms for data science portfolio building



1 Github :-

Github is a very useful platform for displaying your data science projects. As a data science, Github should serve as the 1st platform that you use as a repository of completed project throughout your data science journey. These projects could include projects from weekly assignments. This platform enables you to share your code with other data scientists or data science aspirants.

2 Kaggle :-

Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals. Kaggle allows users to find and publish data sets explore and build models in web based data science environment. work with other data scientists and ML engineers. It is important to build a nice with other data science aspirants who can serve as team members for kaggle challenge competitions.

As you participate in kaggle competitions you can showcase your completed projects, including your datasets, Jupyter notebook and project reports on your public profile.

3. LinkedIn :-

LinkedIn is a very powerful for showcasing your skills and for networking with other data science professionals and organizations. LinkedIn is now one of the most famous platforms for posting data science jobs and for recruiting data scientists.

4. Medium :-

Medium is now considered one of the fastest growing platforms for portfolio building and for writing. If you are interested in using this platform for portfolio building, the first step would be to create a medium account. You can create a free account or a number account.

Why data science portfolio is important.

1. A portfolio helps you showcase your data science skills.
2. A portfolio enables you to mix with other data science professionals and leaders in the field.
3. A portfolio is good for book keeping. You can use it to keep a record of your completed projects, including datasets, codes, and sample of files.
4. By building a portfolio and networking with other data science is a field that is ever changing due to advances in technology.
5. A portfolio increases your chances of getting a job. I have had numerous opportunities from LinkedIn. For instance recruiters reaching out to me for job opportunities in data science

Data formats in data science using Python.

1. comma - seperated values (CSV)
2. XLSX
3. zip
4. plan text (txt)
5. JSON
6. XML
7. HTML
8. images
9. Hierarchical Data format
10. PDF
11. DOCX
12. MP3
13. MP4.

File format is a standard way in which information is encoded for storage in a file. The file format specifies whether the file is a binary or ASCII file. It shows how the information is organized.

Why should a data scientist understand different file formats?

→ The files you will come across will depend on the application you are building. For eg. image processing system, you need image file as input. and .gif. So you will mostly see files in .jpeg, .gif or .png format.

As a data scientist, you need to understand the underlying structure of various file formats. Unless you understand the underlying structure of the data.

you will not be able to explore it. also, at times you need to make decisions about how to store data.

Choosing the optimal file format for storing data can improve the performance of your models in data processing.

Data Parsing

Data parsing is the process of taking data in one format and transforming it to another format. They are commonly used in compilers when we need to parse computer code and generate machine code.

This happens all the time when developers write code that gets run on HTML. Parsers are also present in SQL engines. SQL engines parse a SQL query, execute it, and return the results.

In the case of web scraping, this usually happens after data has been extracted from a web page via web scraping. Once you have scraped data from the web, the next step is making it more readable and better for analysis so that your team can use the results effectively.

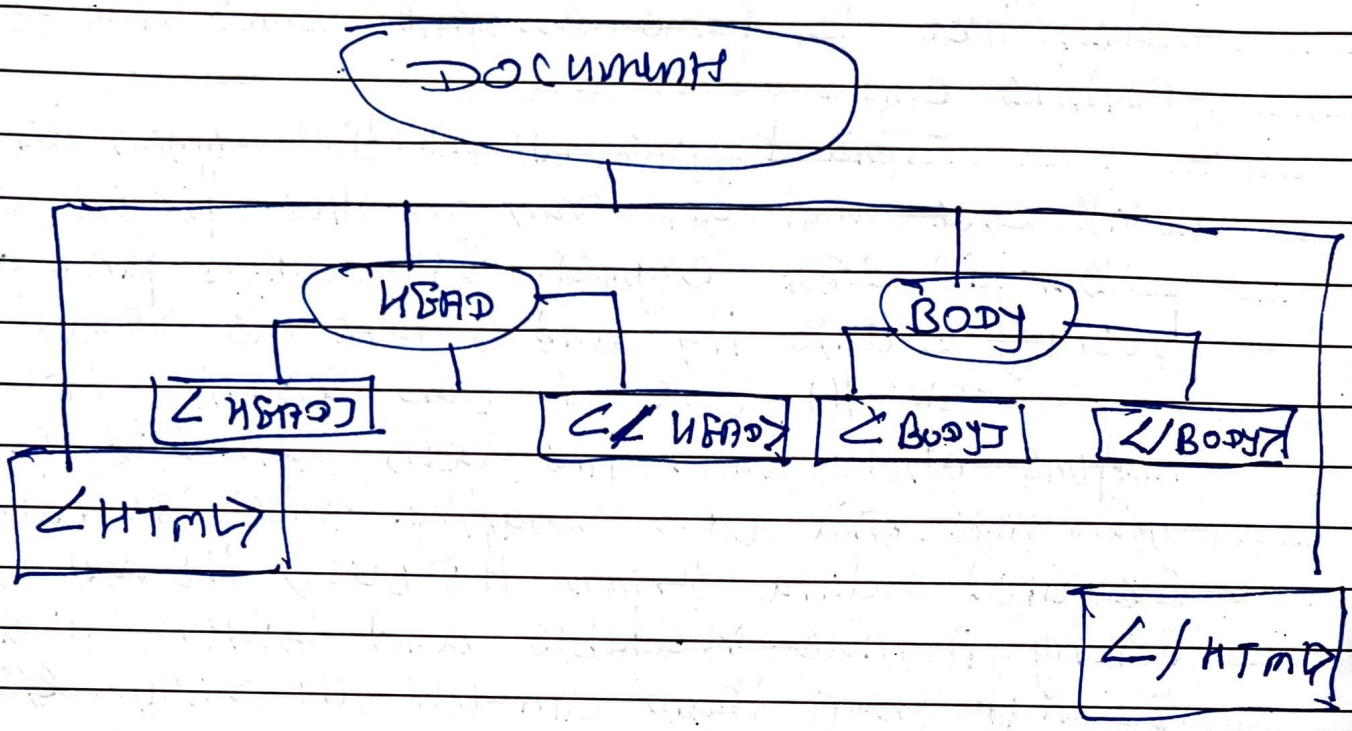
A good data parser isn't constrained to particular formats. Your should be able to input any data type and output a different datatype.

How to build a data parser

What type of data parser you choose, a good parser will figure out what information from an HTML string is useful and based on pre-defined rules. There are usually two steps to the parsing process, lexical analysis and syntactic analysis.

Lexical analysis is the 1st step in data parsing. It basically creates tokens from a sequence of characters that come into the parser as a string of unstructured data. like HTML

The parser makes the tokens by using lexical units like keywords and delimiters. It also ignores irrelevant information like whitespaces and comments.



After the parser has separated the data between lexical units and the irrelevant information it discards all of the irrelevant information and parses the relevant information to the next step.

The next part of the data parsing process is syntactic analysis. This is where parse tree building happens. The parser takes the relevant tokens from the lexical analysis and arranges them into a tree. Any further irrelevant tokens like semicolons and curly braces, are added to the nesting structure of the tree.

Once the parse tree is finished, they you are left with relevant information in a

) Handling missing data:

structured format that can be saved in any file type. There are several different ways to build parser from creating one programmatically to using existing tools. It depends on your business needs, how much time you have, what your budget is, and a few other factors.

HTML parsing libraries / Tools

HTML parsing libraries are great for adding automation to your scraping flow. You can connect many of these libraries to your web scraper via API calls and parse data as you receive it.

1. Scrapy or BeautifulSoup

These are libraries written in Python. BeautifulSoup is a python library for pulling data out of HTML and XML files. Scrapy is a data parser that can also be used for web scraping. When it comes to web scraping with Python, there are a lot of options available and it depends on how hands on you want to be.

2. Cheerio:-

If you are used to working with JavaScript. Cheerio is a good option. It parses markup and provides an API for manipulating the resulting data structure. You could also use Puppeteer. This can be used to generate screenshots and PDFs of specific pages that can be saved and further parsed with other tools. There are many other JavaScript based web scrapers & web parsers.

3 Jsoup :-

For those that works primarily with Java, there are options for you as well. Jsoup is one option. It allows you to work with real world HTML through its API for fetching URL's and extracting and manipulating data. It acts as both a web scraper and a web parser. It can be challenging to find other Java options that are open source, but it's definitely worth a look.

4. Nokogiri :-

There is an option for Ruby as well. Take a look at Nokogiri. It allows you to work with HTML and HTML with Ruby. It has an API similar to the other packages in other languages that lets you query the data you have retrieved from web scraping. It adds an extra layer of security because it treats all documents as untrusted by default. Date parsing in Ruby can be tricky as it can be harder to find gems you can work with.

Data Transformation

Raw data is difficult to track or understand. That is why it needs to be pre-processed before retrieving any information from it. Data transformation is a technique used to convert the raw data into a suitable format that efficiently eases data and retrieves strategic information. Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form.

Data transformation is an essential data preprocessing technique that must be performed on the data before data mining to provide patterns that are easier to understand.

Data transformation changes the format, structure or values of the data and converts them into clean, usable data. Data may be transformed at two stages of the data pipeline for data analytics projects.

Data integration, migration, data warehousing data wrangling may all involve data transformation. Data transformation increases the efficiency of business and analytic processes. and it enables businesses to make better data driven decisions. during the data transformation process, an analyst will determine the structure of the data. This could mean that data transformation may be:

1 Constructive:-

The data transformation process adds, copies or replicates data.

2. Destructive :- The system deletes fields or records
3. Aesthetic :- The transformation standardizes the data to meet requirements or parameters.
4. Structural :- The database is reorganized by renaming, moving or combining columns.

Data transformation Techniques.

There are several data transformation techniques that can help structure and clean up the data before analysis or storage in a data warehouse.

1. Data Smoothing
2. Attribute construction
3. Data generation
4. Data Aggregation
5. Data Discretization
6. Data Normalization.

I Data Smoothing :-

Data Smoothing is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

The concepts behind the data Smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of

data which can often be difficult to digest. for finding pattern that they wouldn't see otherwise. We have been how the noise is removed from the data using the techniques such as binning, regression, clustering.

1 Binning :—

This method splits the sorted data into the no. of bins and smoothens the data values in each bin considering the neighborhood values around it.

2 Regression :—

This method identifies the relation among two dependent attributes so that if we have one attribute, it can be used to predict the other attribute.

3. Clustering :—

This method groups similar data values and form a cluster. The values that lie outside a cluster are known as outliers.

II Attribute construction

In the attribute construction method, the new attributes consult the existing attributes to construct a new data set that eases data mining. new attributes are created and applied to assist the mining process from the given attributes. This simplifies the original data and makes the mining more efficient.

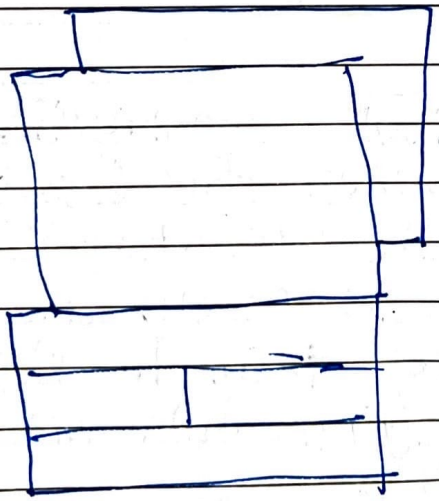
Eg. data represent in form of graph.

III Data Aggregation:—

Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity & quality of the data used.

gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results. The collection of data is useful for everything from decisions concerning financing or business strategy of the product - pricing, operations & marketing strategies.

For ex., we have data set of sales reports of an enterprise that has quarterly sales of each year. We can aggregate the data to get the enterprise's annual sales report.



IV Data Normalization:-

Normalizing the data refers to scaling the data values to a much smaller range such as [-1, 1] or [0.0, 1.0]. There are different methods to Normalize the data.

Consider that we have a numeric attribute A and we have n number of observed values for attribute A that are $v_1, v_2, \ldots v_n$.

1. min-max normalization:-

This method implements a linear transformation on the original data. Let us consider that we have min_A and max_A as the minimum and maximum value observed for attribute A and v_i is the value for attribute A that has to be normalized.

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new_{max_A} - new_{min_A}) + new\_min_A$$

min = 8,200 max = 89,800
income (0.0, 1.0) mul value = 87360

$$= \frac{73600 - 1200}{98000 - 1200}(1.0 - 0.0) + 0.0 = 0.716$$

2 Z-Score normalization

This method normalizes the value for attribute A using the mean and standard deviation.

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

Here $\bar{A}$ and σ_A are mean & Standard deviation.

mean = $ 54000
Standard = $ 16000
Normalize value = $ 73,000

$$\frac{73600 - 54000}{16000} = 1.225$$

3. decimal Scaling :-
This method normalizes the value of attribute A by moving the decimal point in the value. This movement of a decimal point depends on the maximum absolute value of A. -

$$v_i' = \frac{v_i}{10^j}$$

j = Smallest integer
max = (|v_i'|) < 1

attribute range = -986 to 917
max range A = 986
divide attribute A by 1000 i.e. j = 3
So
-986 Normalized to -0.986
917 -"- to 0.917

V. Data Discretization :-

This is a process of converting continous data into a set of data intervals. Continous attribute values are substituted by small interval labels. This makes the data easier to study and analyze. If a data mining task handles a continous attribute, then its discrete value can be replaced by constant quality attributes. This improves the efficiency of the task.

This method is also called a data reduction Mechanism as it transforms a large dataset into a set of categorical data. Discretization also uses decision tree based algorithms to produce short, compact, and accurate results when using discrete values.

Data discretization can be classified into two types: Supervised discretization, where the class information is used. and Unsupervised discretization, which is based on which direction the process proceeds. i.e. top down splitting strategy.

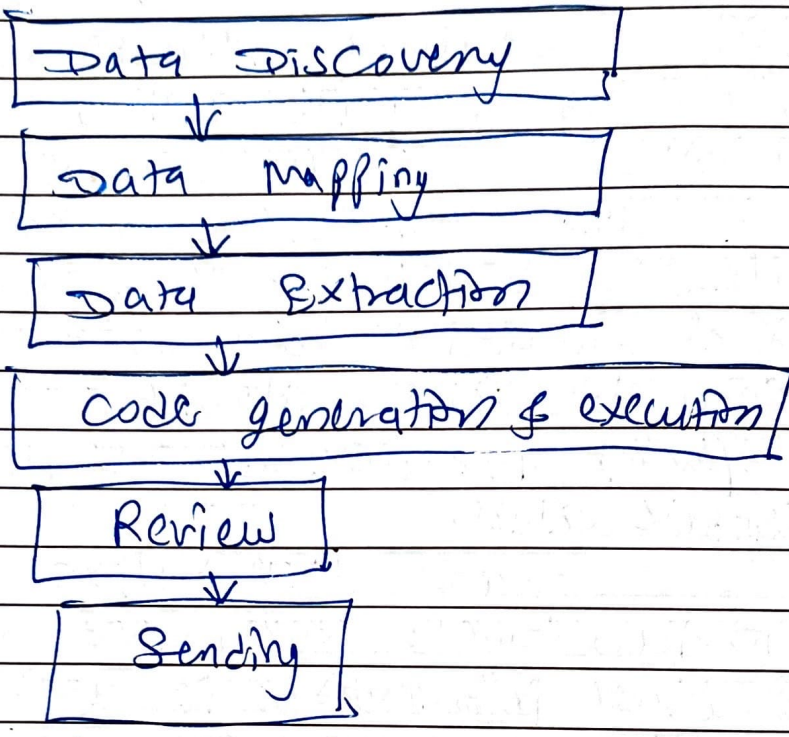
VI. Data generalization :-

It converts low level data attributes to high level data attributes using concept hierarchy. This conversion from a lower level to a higher conceptual level is useful to get a clear picture of the data. Data generalization can be divided into 2 approaches.

1. Data cube process (OLAP) approach.
2. Attribute - oriented induction (AOI) approach.

Data Transformation Process

The entire process for transforming data is known as ETL (Extract, Load, Transform). Through the ETL process, analysts can convert data to its desired format.



Data transformation process

I **Data Discovery:** —

During the 1st stage, analyst work to understand and identify data in its source format. To do this, they will use data profiling tools. This step helps analysts decide what they need to do to get data into its desired format.

II **Data mapping:** —

During this phase, analysts perform data mapping to determine how individual fields are modified, mapped, filtered, joined, and aggregated. Data mapping is essential to many

data processes, and one misstep can lead to incorrect analysis and ripple through your entire organization.

III Data Extraction: —
During this phase, analysts extract the data from its original source. These may include structured sources such as databases or streaming sources such as customer log files from web applications.

IV Code generation and execution: —
Once the data has been extracted analysts need to create a code to complete the transformation often analysts generate code with the help of data transformation platforms or tools.

V Review: — After transforming the data analysis need to check it to ensure everything has been formatted correctly.

VI Sending: — The final step involves sending the data to its target destination. The target might be a data warehouse or a database that handles both structured and unstructured data.

Advantages:
1 Better organization — humans & computers to use.
2. improved data quality — risk and cost of bad data
3 perform Faster queries
4 better data management
5 more use out of data

⇒ Difference between

	⇒ Data cleaning	⇒ Data transformation
1	process of detecting and removing corrupted or inaccurate records from a record set, table or database.	process of converting data from one format or structure into another format or structure
2	Helps to clean the dataset and improve the data consistency.	Helps to make data processing easier.

What are scalability types in a real time system?

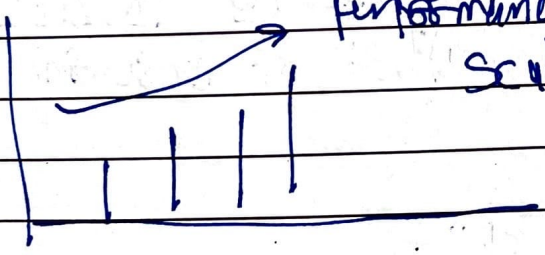
An application could be small or huge, depending on the project requirements. Multiple forms of scalability must be considered.

1. Scalability of performance
2. Scalability of maintenance
3. Scalability of cost.

1. Scaability of performance:-

The most frequent scalability consideration is performance. We want our system to became faster ~~than~~ and more responsive as additional users data, and features are added. if an application has 1000 client connections, it will have effectively scaled performance when its response times.

An application with 1000 client connections that has effectively scaled performance will the same response time, or close to the same response time, as it does with 50,000 client connections.



many performance variables will have an infact on our real time application. As with typical online apps, the data store will surely be the source of performance issues as the program expands. real time applications are affected by

performance considerations, although non real time apps may not be affected.

For eg. In a real time application, you will need to communicate information about a high no. of real time connections between your application's server.

II Scalability of maintainance:-

The developer of an application should be concerned with maintenance scalability. When we add new features, fix bugs or maintain the uptime of program over time. We call it maintainence.

Developer must spend more time adding features or identifying existing problems in an application. maintainance is a difficult issue to optimize since it is easy to be ignorant of issues that will arise in the future. We may use a new approach or technology that we believe will make future adjustments easier, but in reality, the reverse will occur.

Utilizing programming best practices and clearly defined boundaries in an application is a tried and true method of ensuring future maintainance.

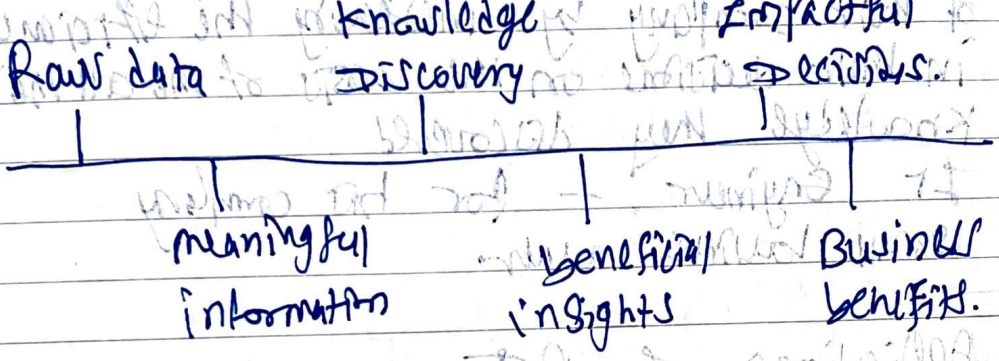
III Scalability of cost:-

It's easy to overlook the cost of something because, as developers, we are frequently isolated from the financial cost of our apps. eg. Server resources can be saved or spent.

Date :
Page : 1

Business Intelligence (BI)

BI uses a set of processes, technologies, and tools to transform raw data into meaningful information and then transform information to provide knowledge. Then afterwards some beneficial insights can be extracted manually and by some slow then the decision makers can make an impactful decision of the basis of insights.



Features Business intelligence

1. Fact based decision making
2. 360 degrees perspective on your business
3. Virtual team members on the same page.
4. measurement for creating KPI. (Key performance indicator on the basis of historic data.
5. identify the benchmark and then set the benchmarks for different processes.
6. BI systems can use to identify the market trends and also to spot Business problems that need to be identified and solved
7. BI is used for data visualization.
8. BI is used for increasing quality of decision making.
9.

Types of users of Business Intelligence

I] Analyst (⇒ data Analyst or business Analyst)
They are the statistician of the company they used BI on the basis of historical data priority stored in the system.

II] Head or manager of the company.
Head of the company used business intelligence used to increase the profitability of their company by increasing the efficiency in their decisions on the basis of the all the knowledge they discovered

III] IT Engineer — for his company

IV] small business owner

Applications of BI

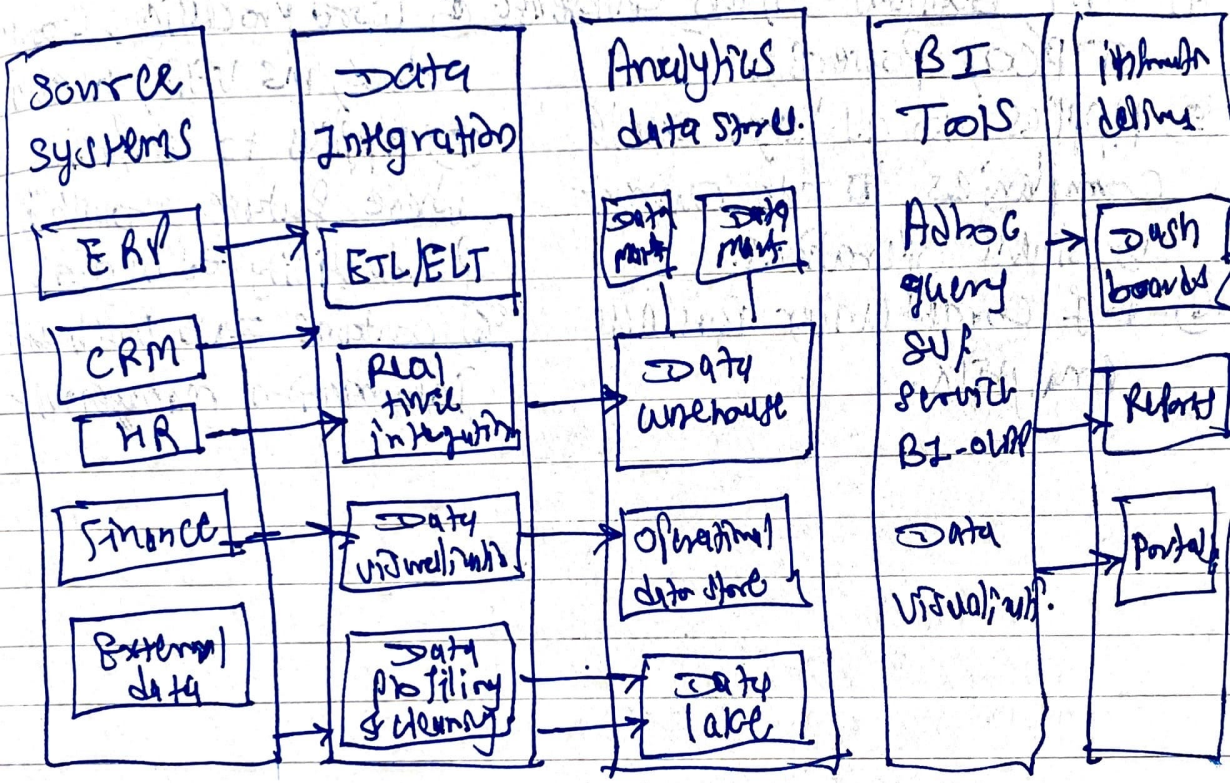
1. In decision making of the company by decision makers of the org.
2. In data mining while extracting knowledge
3. In operational Analystics and operational management
4. In predictive Analystics
5. In prescriptive Analystics
6. making structured data from the unstructured data.
7. In decision support system.
8. In Execution information system.

Date :
Page :

Advantages of BI

1. Identify ways to increase profit
2. understand the market trends
3. optimize business operations
4. Analyze customer behaviour
5. compare performance with competitors.

Business Intelligence architecture.



Business Intelligence

① Analyses past and present to drive current business needs
② To run current business operations.
③ BI for current business operations.
④ SAP Business objects, qliksense, TIBCO, PowerBI
⑤ Applies to all large scale companies to run current business operations.
⑥ BI comes under Business Analytics.

Business Analytics.

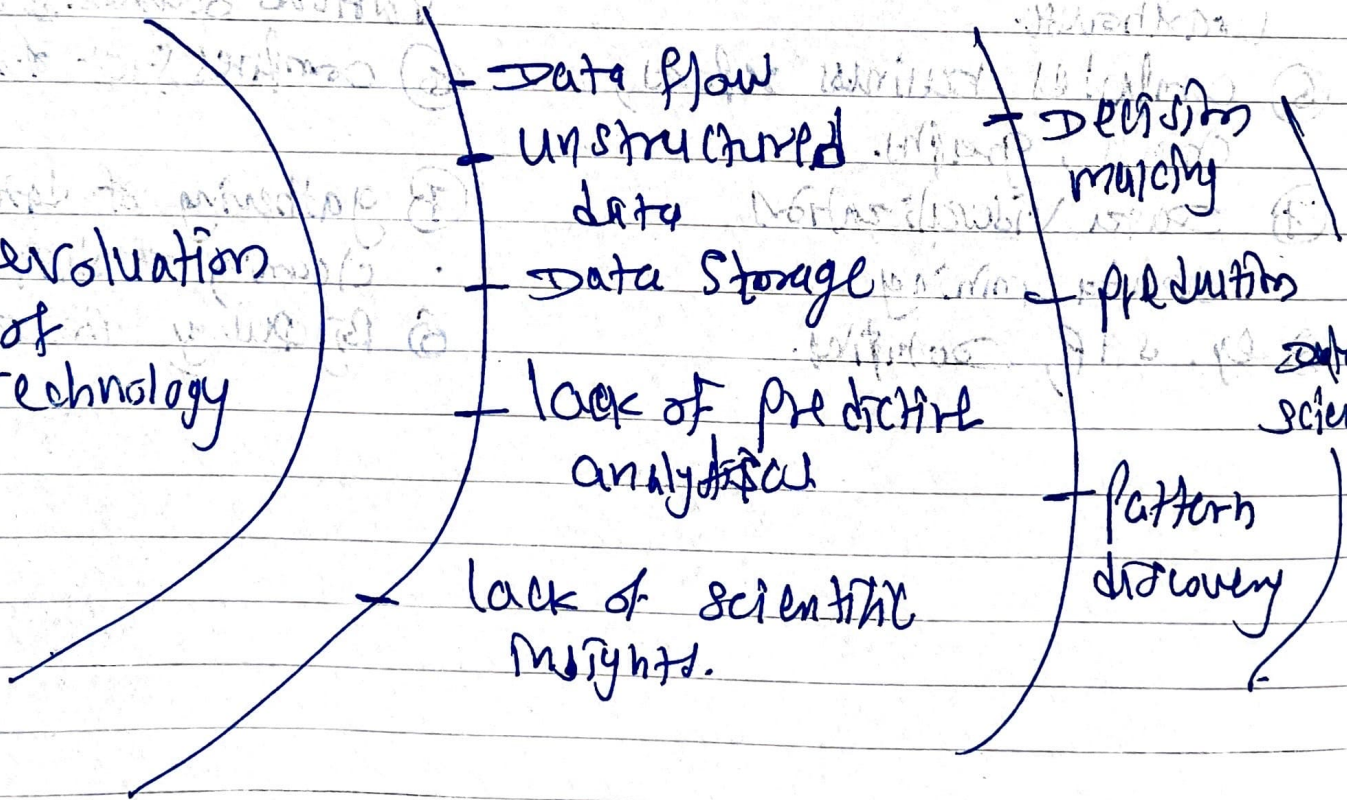
① Analyses past data to drive current business
② To change business operations and improve productivity
③ BA for future business operations
④ word processing, Google docs, MS visio, MS office
⑤ Applies to companies where future growth & productivity is its goal.
⑥ contains into warehouse information management

Business Intelligence	Data warehouse.
① It is a set of tools and methods to analyze data and discover, extract and formulate actionable information that would be ~~for~~ useful for business decisions.	① It is a system for storage of data from various sources in an orderly manner as to facilitate business minded heads and wills.
② It is a decision support system.	② data storage system.
③ Server at the frontend	③ Server at the back end.
④ The aim of BI is to enable users to make informed data driven decisions.	④ to provide users of BI structured
⑤ collect data from data warehouse.	⑤ Collect data from various sources. e.g. org.
⑥ comprises business reports, charts, graphs.	⑥ comprises fact & tables
⑦ Data visualization Data mining	⑦ gathering of data · cleaning of data
⑧ Eg. SAP, Tableau.	⑧ BigQuery, Amazon

Difference between Business Intelligence and Data science.

Business Intelligence	Data science
① looking backward	① looking forward
② Structured data, mostly SQL, data warehouse	② Structured and unstructured data. SQL, NoSQL
③ approach to statistics & visualization	③ Statistics, ML, & graph.
④ Emphasis on past & present	④ Analysis & neuro linguistic prog.
⑤ Tools Pentaho, Microsoft BI, QlikView	⑤ R, Tensorflow

Need for Data science



Types of data science job

1. Data scientist
2. Data Analyst
3. machine learning expert
4. Data Engineer
5. Data Architect
6. Data Administrator
7. Business Analyst
8. Business Intelligence manager

Data science life cycle