.

Data processing:

The collection process provides data that potentially has useful information. You can analyze the extracted information for intelligence that will help you grow your business. This intelligence might, for example, tell you about your user behavior and the relative popularity of your products. The best practice to gather this intelligence is to load your raw data into a data warehouse to perform further analysis.

There are two types of processing workflows to accomplish this: batch processing and real-time processing. The most common forms of processing, online analytic processing (OLAP) and OLTP, each use one of these types. OLAP processing is generally batch-based. OLTP systems are oriented toward real-time processing, and are generally not well suited for batch-based processing. If you decouple data processing from your OLTP system, you keep the data processing from affecting your OLTP workload.

First, let's look at what is involved in batch processing.

Batch Processing:

Extract Transform Load (ETL) — ETL is the process of pulling data from multiple sources to load into data warehousing systems. ETL is normally a continuous, ongoing process with a well-defined workflow. During this process, data is initially extracted from one or more sources. The extracted data is then cleansed, enriched, transformed, and loaded into a data warehouse. For batch ETL, use AWS Glue or Amazon EMR. AWS Glue is a fully managed ETL service. You can create and run an ETL job with a few clicks in the AWS Management Console. Amazon EMR is for big data processing and analysis. EMR offers an expandable, low-configuration service as an easier alternative to running in-house cluster computing.

Extract Load Transform (ELT) — ELT is a variant of ETL, where the extracted data is loaded into the target system first. Transformations are performed after the data is loaded into the data warehouse. ELT typically works well when your target system is powerful enough to handle transformations. Amazon Redshift is often used in ELT pipelines, because it is highly efficient in performing transformations.

Online Analytical Processing (OLAP) — OLAP systems store aggregated historical data in multidimensional schemas. Used widely for query, reporting, and analytics, OLAP systems enable you to extract data and spot trends on multiple dimensions. Because it is optimized for fast joins, Amazon Redshift is often used to build OLAP systems.

Real-time processing

We talked about streaming data earlier, and mentioned Amazon Kinesis Services and Amazon MSK as solutions to capture and store streaming data. You can process this data sequentially and incrementally

.

on a record-by-record basis, or over sliding time windows. Use the processed data for a wide variety of analytics, including correlations, aggregations, filtering, and sampling. This type of processing is called real-time processing.

Information derived from real-time processing gives companies visibility into many aspects of their business and customer activity, such as service usage (for metering or billing), server activity, website clicks, and geolocation of devices, people, and physical goods. This enables them to respond promptly to emerging situations. Real-time processing requires a highly concurrent and scalable processing layer.

To process streaming data in real-time, use AWS Lambda. Lambda can process the data directly from AWS IoT or Amazon Kinesis Data Streams. Lambda enables you to run code without provisioning or managing servers.

Amazon Kinesis Client Library (KCL) is another way to process data from Amazon Kinesis Streams. KCL gives you more flexibility than Lambda to batch your incoming data for further processing. You can also use KCL to apply extensive transformations and customizations in your processing logic.

Amazon Kinesis Data Firehose is the easiest way to load streaming data into AWS. It can capture streaming data and automatically load it into Amazon Redshift, enabling near-real-time analytics with existing BI tools, and dashboards you're already using today. Define batching rules with Kinesis Data Firehose, and it takes care of reliably batching the data and delivering it to Amazon Redshift.

Amazon MSK is an easy way to build and run applications that use Apache Kafka to process streaming data. Apache Kafka is an open-source platform for building real-time streaming data pipelines and applications. With Amazon MSK, you can use native Apache Kafka APIs to populate data lakes, stream changes to and from databases, and power machine learning and analytics applications.

AWS Glue streaming jobs enable you to perform complex ETL on streaming data. Streaming ETL jobs in AWS Glue can consume data from streaming sources like Amazon Kinesis Data Streams and Amazon MSK, clean and transform those data streams in-flight, and continuously load the results into S3 data lakes, data warehouses, or other data stores. As you process streaming data in an AWS Glue job, you have access to the full capabilities of Spark Structured Streaming to implement data transformations, such as aggregating, partitioning, and formatting, as well as joining with other data sets to enrich or cleanse the data for easier analysis.

Six stages of data processing:

1. Data collection

.

Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

2. Data preparation

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as “pre-processing” is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. 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 clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. 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, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

5. Data output/interpretation

The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

6. Data storage

The final stage of data processing is storage. After all of 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. Plus, properly stored data is a necessity for compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

The future of data processing:

.

The future of data processing lies in the cloud. Cloud technology builds on the convenience of current electronic data processing methods and accelerates its speed and effectiveness. Faster, higher-quality data means more data for each organization to utilize and more valuable insights to extract.

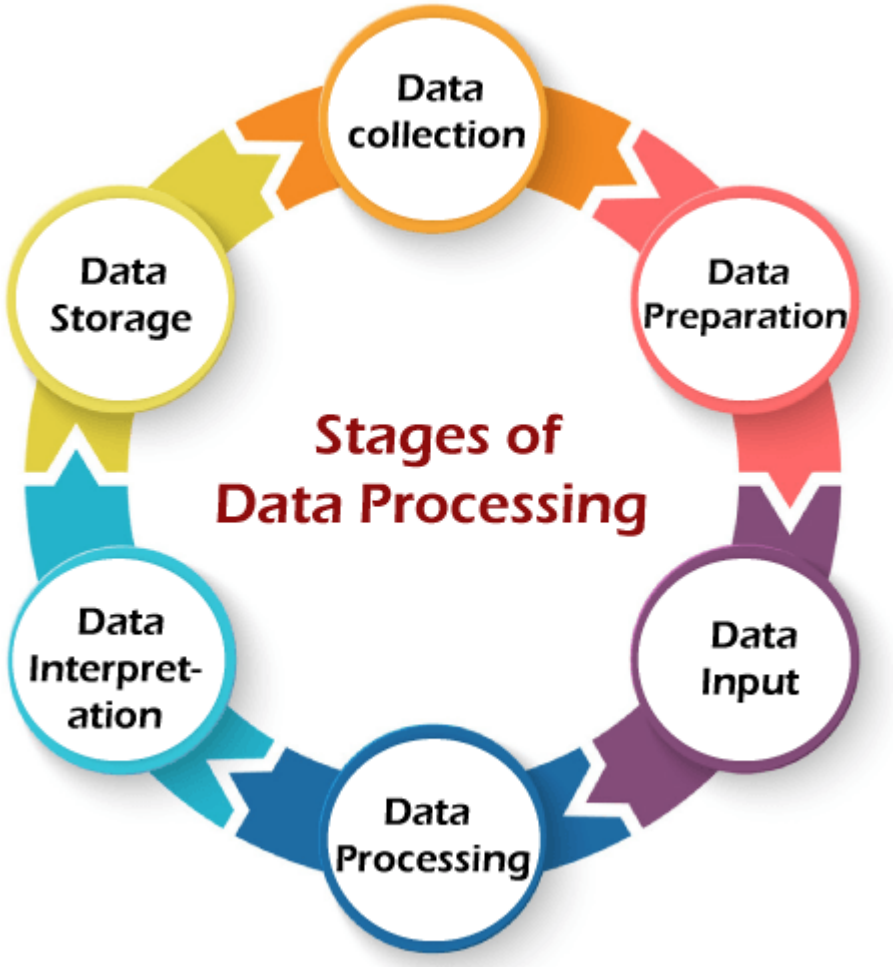
As big data migrates to the cloud, companies are realizing huge benefits. Big data cloud technologies allow for companies to combine all of their platforms into one easily-adaptable system. As software changes and updates (as it does often in the world of big data), cloud technology seamlessly integrates the new with the old.

the benefits of cloud data processing are in no way limited to large corporations. In fact, small companies can reap major benefits of their own. Cloud platforms can be inexpensive and offer the flexibility to grow and expand capabilities as the company grows. It gives companies the ability to scale without a hefty price tag.

From data processing to analytics Big data is changing how all of us do business.

Today, remaining agile and competitive depends on having a clear, effective data processing strategy. While the six steps of data processing won't change, the cloud has driven huge advances in technology that deliver the most advanced, cost-effective, and fastest data processing methods to date.

The data processing consists of the following six stages.



.

The collection of raw data is the first step of the data processing cycle. The raw data collected has a huge impact on the output produced. Hence, raw data should 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 behavior, etc.

2**. Data Preparation**

Data preparation or data cleaning is the process of sorting and filtering the raw data to remove unnecessary and inaccurate data. Raw data is checked for errors, duplication, miscalculations, or missing data and transformed into a suitable form for further analysis and processing. This ensures that only the highest quality data is fed into the processing unit.

3. Data Input

In this step, the raw data is converted into machine-readable form and fed into the processing unit. This can be in the form of data entry through a keyboard, scanner, or any other input source.

4. Data Processing

In this step, the raw data is subjected to various data processing methods using machine learning and artificial intelligence algorithms to generate the desired output. This step may vary slightly from process to process depending on the source of data being processed (data lakes, online databases, connected devices, etc.) and the intended use of the output.

5. Data Interpretation or Output

The data is finally transmitted and displayed to the user in a readable form like graphs, tables, vector files, audio, video, documents, etc. This output can be stored and further processed in the next data processing cycle.

7:**Data Storage**

The last step of the data processing cycle is storage, where data and metadata are stored for further use. This allows quick access and retrieval of information whenever needed. Effective proper data storage is necessary for compliance with GDPR (data protection legislation).

Data Processing in Data Mining

1. Manual Data Processing

Data is processed manually in this data processing method. 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 is a low-cost methodology and does not need very many tools. However, it produces high errors and requires high labor costs and lots of time.

.

2. **Mechanical Data Processing**
   Data is processed mechanically through the use of devices and machines. These can include simple devices such as calculators, typewriters, printing press, etc. Simple data processing operations can be achieved with this method. It has much fewer errors than manual data processing, but the increase in data has made this method more complex and difficult.

3.Electronic Data Processing

Data processing is processed with modern technologies using data processing software and programs. The software gives a set of instructions to process the data and yield output. This method is the most expensive but provides the fastest processing speeds with the highest reliability and accuracy of output.

Types of Data Processing

There are different types of data processing based on the source of data and the steps taken by the processing unit to generate an output. There is no one size fits all method that can be used for processing raw data.

Data Processing in Data Mining

Batch Processing: In this type of data processing, data is collected and processed in batches. It is used for large amounts of data. For example, the payroll system.

Single User Programming Processing: It is usually done by a single person for his personal use. This technique is suitable even for small offices.

Multiple Programming Processing: This technique allows simultaneously storing and executing more than one program in the Central Processing Unit (CPU). Data is broken down into frames and processed using two or more CPUs within a single computer system. It is also known as parallel processing. Further, the multiple programming techniques increase the respective computer's overall working efficiency. A good example of multiple programming processing is weather forecasting.

Real-time Processing: This technique facilitates the user to have direct contact with the computer system. This technique eases data processing. This technique is also known as the direct mode or the interactive mode technique and is developed exclusively to perform one task. It is a sort of online processing, which always remains under execution. For example, withdrawing money from ATM.

Online Processing: This technique facilitates the entry and execution of data directly; so, it does not store or accumulate first and then process. The technique is developed to reduce the data entry errors, as it validates data at various points and ensures that only corrected data is entered. This technique is widely used for online applications. For example, barcode scanning.

Time-sharing Processing: This is another form of online data processing that facilitates several users to share the resources of an online computer system. This technique is adopted when results are needed swiftly. Moreover, as the name suggests, this system is time-based. Following are some of the major advantages of time-sharing processing, such as:

Several users can be served simultaneously.

.

Coding

```
List<T> data = CreateData(…)
//where T is some known datatype,
//     CreateData is some function which returns a collection of instances of T

ProcessList(data)
//where ProcessList performs the required processing on the generated data
//NOTE: Some explanation is provided as comments

Public delegate Task ProduceAsync<TP>
(IProducerBuffer<TP> buffer, CancellationToken token);

//accepts buffer and cancellation token as inputs and returns a Task
//   where TP is the datatype of item produced by producer
//   and IProducerBuffer is an interface to our Buffer implementation
//we add CancellationToken as an input parameter in order to support
//      interruptible pipeline feature
//In this way, by simply supplying CancellationToken.None to the pipeline
//      we can create uninterruptible pipeline.
//IDisposable to avail Dispose method to perform resource clean-up
Public interface IProducer<TP> : IDisposable
{
   //to perform some pre-processing initialization
   Task InitAsync();

   //actual data generating method
   Task ProduceAsync(IProducerBuffer<TP> buffer, CancellationToken token);
}
//Implementation
```

.

//==================================

//=========Dispose USAGE=============

//==================================

//during App Shutdown Or after network close

//saved_instace.Dispose();

Ls -l | grep key | less (3 operations with 2 pipes)

*_____*_____*_____*