Data Analytics vs. Data Analysis

The terms Data Analysis and Data Analytics are often used interchangeably, including in this course.

However it is important to note that there is a subtle difference between the terms and meaning of the words *Analysis* and *Analytics*. In fact some people go far as saying that these terms mean different things and should not be used interchangeably. Yes, there is a technical difference...

The dictionary meanings are:

Analysis - *detailed examination of the elements or structure of something*

Analytics - *the systematic computational analysis of data or statistics*

*Analysis* can be done without numbers or data, such as business analysis psycho analysis, etc. Whereas *Analytics*, even when used without the prefix "Data", almost invariably implies use of data for perfoming numerical manipulation and inference.

Some experts even say that *Data Analysis* is based on inferences based on historical data whereas *Data Analytics* is for predicting future performance. The design team of this course does not subscribe to this view, and you will see why later in the course as you become familiar with the terms like *predictive analytics, prescriptive analytics, etc*.

So in this course we take a more liberal view, and use the terms Data Analysis and Data Analytics to mean the same thing. For example, an earlier video is titled *Defining Data Analysis*, whereas the preceeding video with the viewpoints of several data professionals is titled *What is Data Analytics*. The difference in these titles is not intentional.

Summary and Highlights

In this lesson, you have learned the following information:

A modern data ecosystem includes a network of interconnected and continually evolving entities that include:

Data that is available in a host of different formats, structure, and sources.

Enterprise Data Environment in which raw data is staged so it can be organized, cleaned, and optimized for use by end-users.

End-users such as business stakeholders, analysts, and programmers who consume data for various purposes.

Emerging technologies such as Cloud Computing, Machine Learning, and Big Data, are continually reshaping the data ecosystem and the possibilities it offers. Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts, all play a vital role in the ecosystem for deriving insights and business results from data.

Based on the goals and outcomes that need to be achieved, there are four primary types of Data Analysis:

Descriptive Analytics, that helps decode “What happened.”

Diagnostic Analytics, that helps us understand “Why it happened.”

Predictive Analytics, that analyzes historical data and trends to suggest “What will happen next.”

Prescriptive Analytics, that prescribes “What should be done next.”

The Data Analysis process involves:

Developing an understanding of the problem and the desired outcome.

Setting a clear metric for evaluating outcomes.

Gathering, cleaning, analyzing, and mining data to interpret results.

Communicating the findings in ways that impact decision-making.

In this lesson, you have learned the following information:

The role of a Data Analyst spans across:

Acquiring data that best serves the use case.

Preparing and analyzing data to understand what it represents.

Interpreting and effectively communicating the message to stakeholders who need to act on the findings.

Ensuring that the process is documented for future reference and repeatability.

In order to play this role successfully, Data Analysts need a mix of technical, functional, and soft skills.

Technical Skills include varying levels of proficiency in using spreadsheets, statistical tools, visualization tools, programming and querying languages, and the ability to work with different types of data repositories and big data platforms.

An understanding of Statistics, Analytical techniques, problem-solving, the ability to probe a situation from multiple perspectives, data visualization, and project management skills – all of which come under Functional Skills a Data Analyst needs in order to play an effective role.

Soft Skills include the ability to work collaboratively, communicate effectively, tell a compelling story with data, and garner support and buy-in from stakeholders. Curiosity to explore different pathways and intuition that helps to give a sense of the future based on past experiences are also essential skills for being a good Data Analyst.

In this lesson, you have learned the following information:

A data analyst ecosystem includes the infrastructure, software, tools, frameworks, and processes used to gather, clean, analyze, mine, and visualize data.

Based on how well-defined the structure of the data is, data can be categorized as:

Structured Data, that is data which is well organized in formats that can be stored in databases.

Semi-Structured Data, that is data which is partially organized and partially free form.

Unstructured Data, that is data which can not be organized conventionally into rows and columns.

Data comes in a wide-ranging variety of file formats, such as delimited text files, spreadsheets, XML, PDF, and JSON, each with its own list of benefits and limitations of use.

Data is extracted from multiple data sources, ranging from relational and non-relational databases to APIs, web services, data streams, social platforms, and sensor devices.

Once the data is identified and gathered from different sources, it needs to be staged in a data repository so that it can be prepared for analysis. The type, format, and sources of data influence the type of data repository that can be used.

Data professionals need a host of languages that can help them extract, prepare, and analyze data. These can be classified as:

Querying languages, such as SQL, used for accessing and manipulating data from databases.

Programming languages such as Python, R, and Java, for developing applications and controlling application behavior.

Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, for automating repetitive operational tasks.

Earlier in the course, we examined databases, data warehouses, and big data stores.

Now we’ll go a little deeper in our exploration of data warehouses, data marts, and data lakes;

and also learn about the ETL process and data pipelines.

A data warehouse works like a multi-purpose storage for different use cases.

By the time the data comes into the warehouse, it has already been modeled and structured

for a specific purpose, meaning it is analysis ready.

As an organization, you would opt for a data warehouse when you have massive amounts of

data from your operational systems that needs to be readily available for reporting and analysis.

Play video starting at ::54 and follow transcript0:54

Data warehouses serve as the single source of truth—storing current and historical

data that has been cleansed, conformed, and categorized.

A data warehouse is a multi-purpose enabler of operational and performance analytics.

A data mart is a sub-section of the data warehouse, built specifically for a particular business

function, purpose, or community of users.

The idea is to provide stakeholders data that is most relevant to them, when they need it.

For example, the sales or finance teams accessing data for their quarterly reporting and projections.

Since a data mart offers analytical capabilities for a restricted area of the data warehouse,

it offers isolated security and isolated performance.

The most important role of a data mart is business-specific reporting and analytics.

A Data Lake is a storage repository that can store large amounts of structured, semi-structured,

and unstructured data in their native format, classified and tagged with metadata.

So, while a data warehouse stores data processed for a specific need, a data lake is a pool

of raw data where each data element is given a unique identifier and is tagged with metatags

for further use.

You would opt for a data lake if you generate, or have access to, large volumes of data on

an ongoing basis, but don’t want to be restricted to specific or pre-defined use cases.

Unlike data warehouses, a data lake would retain all source data, without any exclusions.

And the data could include all types of data sources and types.

Data lakes are sometimes also used as a staging area of a data warehouse.

The most important role of a data lake is in predictive and advanced analytics.

Now we come to the process that is at the heart of gaining value from data—the Extract,

Transform, and Load process, or ETL.

ETL is how raw data is converted into analysis-ready data.

It is an automated process in which you gather raw data from identified sources,

extract the information that aligns with your reporting and analysis needs,

clean, standardize, and transform that data into a format that is usable in the context

of your organization; and

load it into a data repository.

While ETL is a generic process, the actual job can be very different in usage, utility,

and complexity.

Extract is the step where data from source locations is collected for transformation.

Data extraction could be through:

Batch processing, meaning source data, is moved in large chunks from the source to the

target system at scheduled intervals.

Tools for batch processing include Stitch and Blendo.

Stream processing, which means source data is pulled in real-time from the source and

transformed while it is in transit and before it is loaded into the data repository.

Tools for stream processing include Apache Samza, Apache Storm, and Apache Kafka.

Transform involves the execution of rules and functions that converts raw data into

data that can be used for analysis.

For example,

making date formats and units of measurement consistent across all sourced data,

removing duplicate data,

filtering out data that you do not need,

enriching data, for example, splitting full name to first, middle, and last names,

establishing key relationships across tables,

applying business rules and data validations.

Load is the step where processed data is transported to a destination system or data repository.

It could be: Initial loading, that is, populating all the data in the repository,

Incremental loading, that is, applying ongoing updates and modifications as needed periodically; or Full refresh, that is, erasing contents of one or more tables and reloading with fresh data.

Load verification, which includes data checks for missing or null values, server performance, and monitoring load failures, are important parts of this process step.

It is vital to keep an eye on load failures and ensure the right recovery mechanisms are in place.

ETL has historically been used for batch workloads on a large scale.

However, with the emergence of streaming ETL tools, they are increasingly being used for real-time streaming event data as well.

It’s common to see the terms ETL and data pipelines used interchangeably.

And although both move data from source to destination, data pipeline is a broader term that encompasses the entire journey of moving data from one system to another, of which

ETL is a subset.

Data pipelines can be architected for batch processing, for streaming data, and a combination of batch and streaming data.

In the case of streaming data, data processing or transformation, happens in a continuous flow.

This is particularly useful for data that needs constant updating, such as data from

a sensor monitoring traffic.

A data pipeline is a high performing system that supports both long-running batch queries and smaller interactive queries.

The destination for a data pipeline is typically a data lake, although the data may also be loaded to different target destinations, such as another application or a visualization tool.

There are a number of data pipeline solutions available, most popular among them being Apache Beam and DataFlow.

In this digital world, everyone leaves a trace.

From our travel habits to our workouts and entertainment,

the increasing number of internets connected devices that we interact with

on a daily basis record vast amounts of data about us

there's even a name for it Big Data. Ernst and

Young offers the following definition: big data refers to the dynamic,

large, and disparate volumes of data being created by people,

tools, and machines. It requires new, innovative and scalable technology to

collect, host, and analytically process the vast

amount of data gathered in order to drive

real-time business insights that relate to consumers,

risk, profit, performance, productivity management,

and enhanced shareholder value. There is no one definition of big data

but there are certain elements that are common

across the different definitions, such as velocity,

volume, variety, veracity, and value. These are the V's of big data

Velocity is the speed at which data accumulates. Data is being generated

extremely fast in a process that never stops.

Near or real-time streaming, local, and cloud-based

technologies can process information very quickly.

Volume is the scale of the data or the increase in the amount of data stored.

Drivers of volume are the increase in data sources,

higher resolution sensors, and scalable infrastructure.

Variety is the diversity of the data. Structured data fits neatly into rows

and columns in relational databases, while

unstructured data is not organized in a predefined way

like tweets, blog posts, pictures, numbers,

and video. Variety also reflects that data comes from different

sources; machines, people, and processes, both internal and external to

organizations. Drivers are mobile technologies social

media, wearable technologies, geo technologies

video, and many, many more. Veracity

is the quality and origin of data and its conformity to facts and accuracy.

Attributes include consistency, completeness, integrity, and ambiguity.

Drivers include cost and the need for traceability.

With the large amount of data available, the debate rages on about the accuracy

of data in the digital age. Is the information real or is it false?

Value is our ability and need to turn data into value.

Value isn't just profit. It may have medical or social benefits,

as well as customer, employee or personal satisfaction.

The main reason that people invest time to understand big data

is to derive value from it. Let's look at some examples of the V's in action.

Velocity. Every 60 seconds, hours of footage are uploaded to YouTube, which is

generating data. Think about how quickly data accumulates

over hours, days, and years. Volume.

The world population is approximately 7 billion people and the vast majority are

now using digital devices. Mobile phones, desktop and laptop

computers, wearable devices, and so on. These devices

all generate, capture, and store data approximately 2.5 quintillion bytes

every day. That's the equivalent of 10 million blu-ray DVDs.

Variety. Let's think about the different types of data.

Text, pictures, film, sound, health data from wearable devices, and many

different types of data from devices connected to the internet of things.

Veracity. Eighty percent of data is considered to be unstructured

and we must devise ways to produce reliable and accurate insights.

The data must be categorized, analyzed, and visualized.

Data scientists, today, derive insights from big data and cope with the

challenges that these massive data sets present.

The scale of the data being collected means that it's not feasible to use

conventional data analysis tools, however, alternative tools that leverage

distributed computing power can overcome this problem.

Tools such as Apache Spark, Hadoop, and its ecosystem

provides ways to extract, load, analyze, and process the data

across distributed compute resources, providing new insights

and knowledge. This gives organizations more ways to connect with their

customers and enrich the services they offer.

So next time you strap on your smartwatch, unlock your smartphone,

or track your workout, remember your data is starting a journey that might take it

all the way around the world, through big data analysis and back to you.

The Big Data processing technologies provide ways to work with large sets of structured,

semi-structured, and unstructured data so that value can be derived from big data.

In some of the other videos, we discussed Big Data technologies such as NoSQL databases

and Data Lakes.

In this video, we are going to talk about three open source technologies and the role

they play in big data analytics—Apache Hadoop, Apache Hive, and Apache Spark.

Hadoop is a collection of tools that provides distributed storage and processing of big

data.

Hive is a data warehouse for data query and analysis built on top of Hadoop.

Spark is a distributed data analytics framework designed to perform complex data analytics

in real-time.

Hadoop, a java-based open-source framework, allows distributed storage and processing

of large datasets across clusters of computers.

In Hadoop distributed system, a node is a single computer, and a collection of nodes

forms a cluster.

Hadoop can scale up from a single node to any number of nodes, each offering local storage

and computation.

Hadoop provides a reliable, scalable, and cost-effective solution for storing data with

no format requirements.

Using Hadoop, you can: Incorporate emerging data formats, such as

streaming audio, video, social media sentiment, and clickstream data, along with structured,

semi-structured, and unstructured data not traditionally used in a data warehouse.

Provide real-time, self-service access for all stakeholders.

Optimize and streamline costs in your enterprise data warehouse by consolidating data across

the organization and moving “cold” data, that is, data that is not in frequent use,

to a Hadoop-based system.

One of the four main components of Hadoop is Hadoop Distributed File System, or HDFS,

which is a storage system for big data that runs on multiple commodity hardware connected

through a network.

HDFS provides scalable and reliable big data storage by partitioning files over multiple

nodes.

It splits large files across multiple computers, allowing parallel access to them.

Computations can, therefore, run in parallel on each node where data is stored.

It also replicates file blocks on different nodes to prevent data loss, making it fault-tolerant.

Let’s understand this through an example.

Consider a file that includes phone numbers for everyone in the United States; the numbers

for people with last name starting with A might be stored on server 1, B on server 2,

and so on.

With Hadoop, pieces of this phonebook would be stored across the cluster.

To reconstruct the entire phonebook, your program would need the blocks from every server

in the cluster.

HDFS also replicates these smaller pieces onto two additional servers by default, ensuring

availability when a server fails, In addition to higher availability, this offers multiple

benefits.

It allows the Hadoop cluster to break up work into smaller chunks and run those jobs on

all servers in the cluster for better scalability.

Finally, you gain the benefit of data locality, which is the process of moving the computation

closer to the node on which the data resides.

This is critical when working with large data sets because it minimizes network congestion

and increases throughput.

Some of the other benefits that come from using HDFS include:

Fast recovery from hardware failures, because HDFS is built to detect faults and automatically

recover.

Access to streaming data, because HDFS supports high data throughput rates.

Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers,

in a single cluster.

Portability, because HDFS is portable across multiple hardware platforms and compatible

with a variety of underlying operating systems.

Hive is an open-source data warehouse software for reading, writing, and managing large data

set files that are stored directly in either HDFS or other data storage systems such as

Apache HBase.

Hadoop is intended for long sequential scans and, because Hive is based on Hadoop, queries.

have very high latency—which means Hive is less appropriate for applications that

need very fast response times.

Hive is not suitable for transaction processing that typically

involves a high percentage of write operations.

Hive is better suited for data warehousing tasks such as ETL, reporting, and data analysis

and includes tools that enable easy access to data via SQL.

This brings us to Spark, a general-purpose data processing engine designed to extract

and process large volumes of data for a wide range of applications, including Interactive

Analytics, Streams Processing, Machine Learning, Data Integration, and ETL.

It takes advantage of in-memory processing to significantly increase the speed of computations

and spilling to disk only when memory is constrained.

Spark has interfaces for major programming languages, including Java, Scala, Python,

R, and SQL.

It can run using its standalone clustering technology as well as on top of other infrastructures

such as Hadoop.

And it can access data in a large variety of data sources, including HDFS and Hive,

making it highly versatile.

The ability to process streaming data fast and perform complex analytics in real-time

is the key use case for Apache Spark.

n this lesson, you have learned the following information:

A Data Repository is a general term that refers to data that has been collected, organized, and isolated so that it can be used for reporting, analytics, and also for archival purposes.

The different types of Data Repositories include:

* Databases, which can be relational or non-relational, each following a set of organizational principles, the types of data they can store, and the tools that can be used to query, organize, and retrieve data.
* Data Warehouses, that consolidate incoming data into one comprehensive storehouse.
* Data Marts, that are essentially sub-sections of a data warehouse, built to isolate data for a particular business function or use case.
* Data Lakes, that serve as storage repositories for large amounts of structured, semi-structured, and unstructured data in their native format.
* Big Data Stores, that provide distributed computational and storage infrastructure to store, scale, and process very large data sets.

ETL, or Extract Transform and Load, Process is an automated process that converts raw data into analysis-ready data by:

* Extracting data from source locations.
* Transforming raw data by cleaning, enriching, standardizing, and validating it.
* Loading the processed data into a destination system or data repository.

Data Pipeline, sometimes used interchangeably with ETL, encompasses the entire journey of moving data from the source to a destination data lake or application, using the ETL process.

Big Data refers to the vast amounts of data that is being produced each moment of every day, by people, tools, and machines. The sheer velocity, volume, and variety of data challenge the tools and systems used for conventional data. These challenges led to the emergence of processing tools and platforms designed specifically for Big Data, such as Apache Hadoop, Apache Hive, and Apache Spark.

# Summary and Highlights

In this lesson, you have learned:

* The process of identifying data begins by determining the information that needs to be collected, which in turn is determined by the goal you seek to achieve.
* Having identified the data, your next step is to identify the sources from which you will extract the required data and define a plan for data collection. Decisions regarding the timeframe over which you need your data set, and how much data would suffice for arriving at a credible analysis also weigh in at this stage.
* Data Sources can be internal or external to the organization, and they can be primary, secondary, or third-party, depending on whether you are obtaining the data directly from the original source, retrieving it from externally available data sources, or purchasing it from data aggregators.
* Some of the data sources from which you could be gathering data include databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys and observation studies.
* Data that has been identified and gathered from the various data sources is combined using a variety of tools and methods to provide a single interface using which data can be queried and manipulated.
* The data you identify, the source of that data, and the practices you employ for gathering the data have implications for quality, security, and privacy, which need to be considered at this stage

# Summary and Highlights

In this lesson, you have learned the following information:

Once the data you identified is gathered and imported, your next step is to make it analysis-ready. This is where the process of Data Wrangling, or Data Munging, comes in. Data Wrangling is an iterative process that involves data exploration, transformation, and validation.

Transformation of raw data includes the tasks you undertake to:

* Structurally manipulate and combine the data using Joins and Unions.
* Normalize data, that is, clean the database of unused and redundant data.
* Denormalize data, that is, combine data from multiple tables into a single table so that it can be queried faster.
* Clean data, which involves profiling data to uncover quality issues, visualizing data to spot outliers, and fixing issues such as missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors, and outliers.
* Enrich data, which involves considering additional data points that could add value to the existing data set and lead to a more meaningful analysis.

A variety of software and tools are available for the Data Wrangling process. Some of the popularly used ones include Excel Power Query, Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python, and R, each with their own set of characteristics, strengths, limitations, and applications.

Summary and Highlights

In this lesson, you have learned the following information:

Data has value through the stories that it tells. In order to communicate your findings impactfully, you need to:

Ensure that your audience is able to trust you, understand you, and relate to your findings and insights.

Establish the credibility of your findings.

Present the data within a structured narrative.

Support your communication with strong visualizations so that the message is clear and concise, and drives your audience to take action.

Data visualization is the discipline of communicating information through the use of visual elements such as graphs, charts, and maps. The goal of visualizing data is to make information easy to comprehend, interpret, and retain.

For data visualization to be of value, you need to:

Think about the key takeaway for your audience.

Anticipate their information needs and questions, and then plan the visualization that delivers your message clearly and impactfully.

There are several types of graphs and charts available for you to be able to plot any kind of data, such as bar charts, column charts, pie charts, and line charts.

You can also use data visualization to build dashboards. Dashboards organize and display reports and visualizations coming from multiple data sources into a single graphical interface. They are easy to comprehend and allow you to generate reports on the go.

When deciding which tools to use for data visualization, you need to consider the ease-of-use and purpose of the visualization. Some of the popularly used tools include Spreadsheets, Jupyter Notebook, Python libraries, R-Studio and R-Shiny, IBM Cognos Analytics, Tableau, and Power BI.

# Summary and Highlights

In this lesson, you have learned the following information:

Data Analyst roles are sought after in every industry, be it Banking and Finance, Insurance, Healthcare, Retail, or Information Technology.

Currently, the demand for skilled data analysts far outweighs the supply, which means companies are willing to pay a premium to hire skilled data analysts.

Data Analyst job roles can be broadly classified as follows:

* Data Analyst Specialist roles - On this path, you start as a Junior Data Analyst and move up to the level of a Principal Analyst by continually advancing your technical, statistical, and analytical skills from a foundational level to an expert level.
* Domain Specialist roles - These roles are for you if you have acquired specialization in a specific domain and want to work your way up to be seen as an authority in your domain.
* Analytics-enabled job roles - These roles include jobs where having analytic skills can up-level your performance and differentiate you from your peers.
* Other Data Professions - There are several other roles in a modern data ecosystem, such as Data Engineer, Big Data Engineer, Data Scientist, Business Analyst, or Business Intelligence Analyst. If you upskill yourself based on the required skills, you can transition into these roles.

There are several paths you can consider in order to gain entry into the Data Analyst field. These include:

* An academic degree in Data Analytics or disciplines such as Statistics and Computer Science.
* Online multi-course specializations offered by learning platforms such as Coursera, edX, and Udacity.
* Mid-career transition into Data Analysis by upskilling yourself. If you have a technical background, for example, you can focus on developing the technical skills specific to Data Analysis. If you do not have a technical background, you can plan to skill your self in some basic technologies and then work your way up from an entry-level position.