Lesson 3: Bi Architecture and Data mining concepts

Business intelligence

Def 1

Business intelligence is the process by which enterprises *use strategies* and *technologies* for *analyzing current and historical data,* with the **objective** of *improving strategic decision-making and providing a competitive advantage.*

Def 2

Business intelligence combines *business analytics, data mining, data visualization*, data tools and infrastructure, and best practices to help organizations make more data-driven decisions

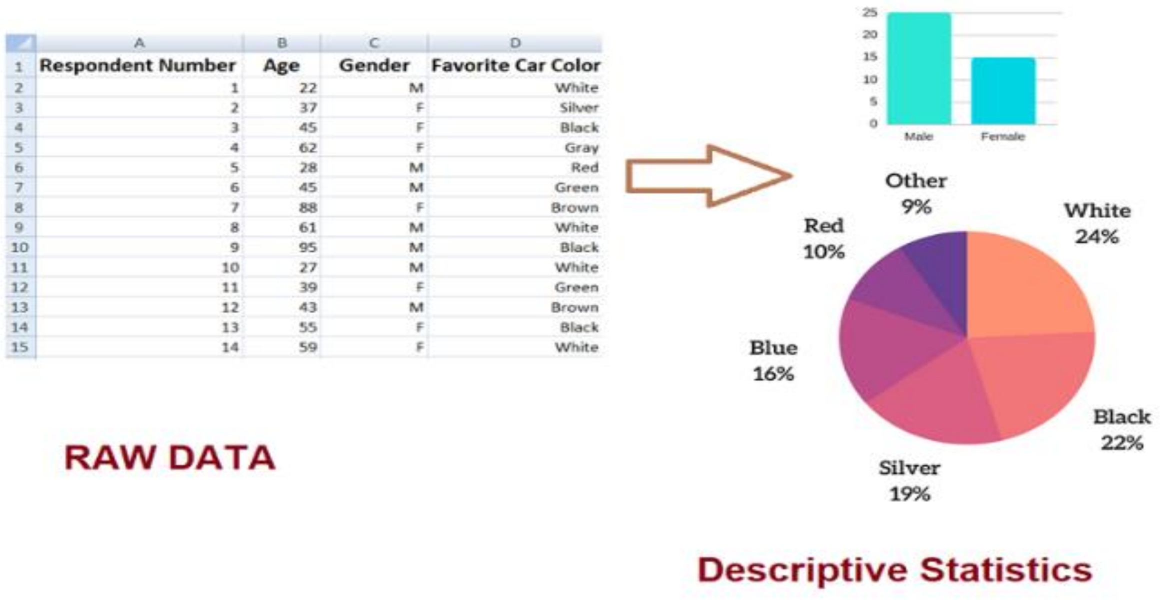
4 types of Business analytics?

1. **Descriptive analytics -** tells *what happened?* in your business in the past week, month or year, presenting it as numbers and visuals in reports and dashboards.
2. **Diagnostic analytics** - is a form of advanced analytics that examines data or content to answer the question, "Why did it happen?" It is characterized by techniques such *as drill-down, data discovery, data mining and correlations.*
3. **Predictive analytics** -determines the potential outcomes of present and past actions and trends. *What will happen?*
4. **Prescriptive analytics** -offers decision support for the best course of action to get desired results. *What should we do?*

Descriptive analytics Example

You've performed a survey to 40 respondents about their favorite car color. And now you have a spreadsheet with the results.

However, this spreadsheet is not very informative and you want to summarize the data with some graphs and charts that can allow you to come up with some simple conclusions (e.g. 24% of people said that white is their favorite color).



	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

Diagnostic Analytics in Healthcare Example

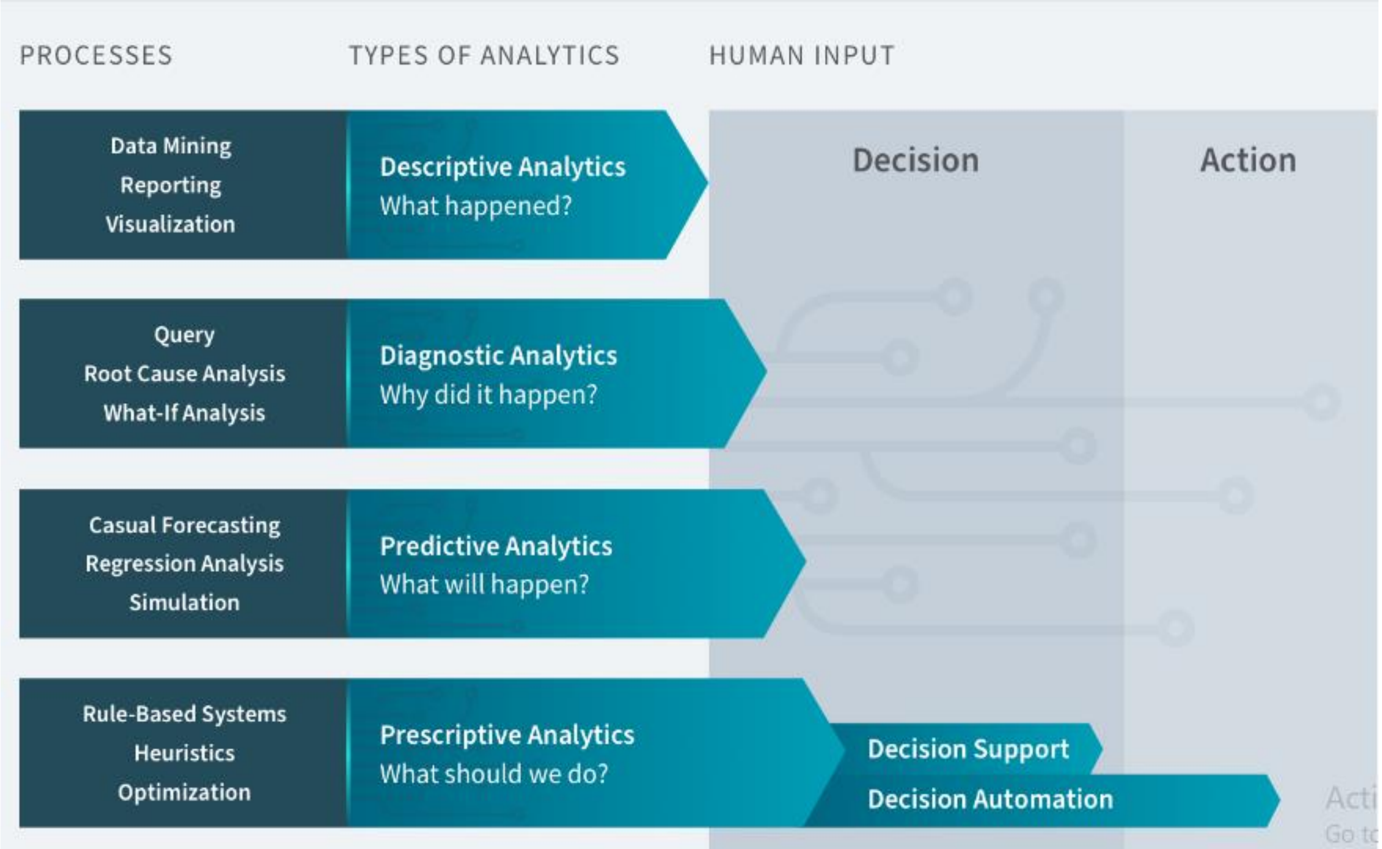
Diagnostic analytics allows us to understand why it happened and answer questions such as: Why did these patients go to the hospital last week?

AI applications of healthcare diagnostics is chatbots.

Eliza chat bots

Predicting Analytics buying behavior Example

One of the biggest uses of predictive analytics is predicting buying behavior in the retail industry. Companies use the tools to learn all about their customers. Companies use advanced analytics to identify buying habits based on previous purchase history.



PROCESSES

TYPES OF ANALYTICS

HUMAN INPUT

Decision

Action

Data Mining
Reporting
Visualization

Descriptive Analytics
What happened?

Query
Root Cause Analysis
What-If Analysis

Diagnostic Analytics
Why did it happen?

Casual Forecasting
Regression Analysis
Simulation

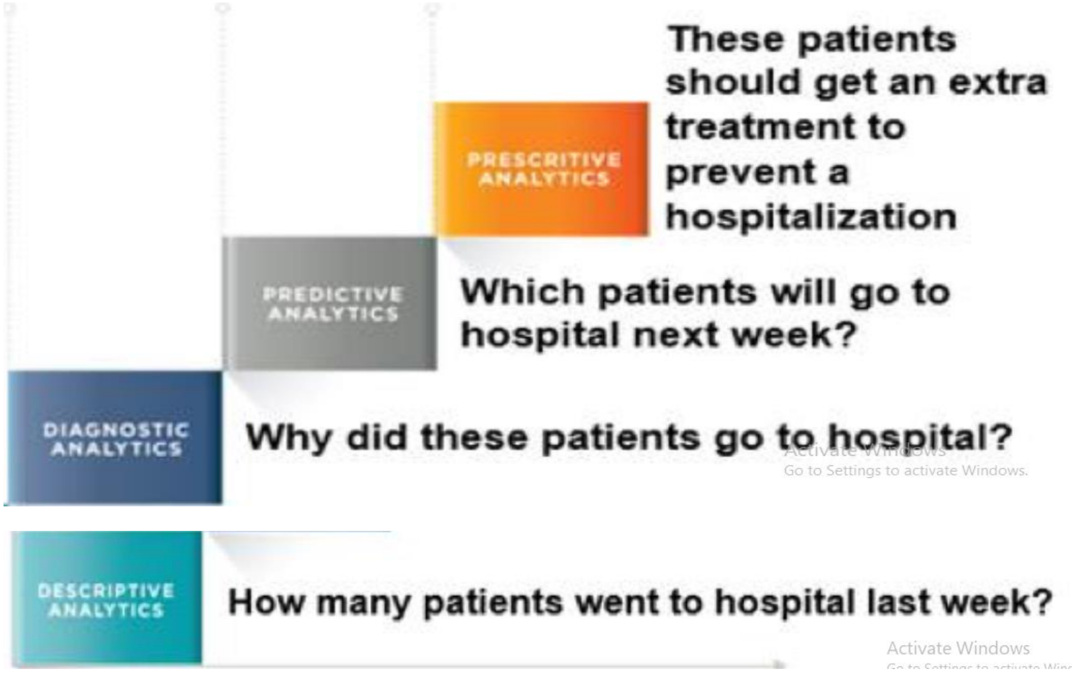
Predictive Analytics
What will happen?

Rule-Based Systems
Heuristics
Optimization

Prescriptive Analytics
What should we do?

Decision Support

Decision Automation



PRESCRIPTIVE ANALYTICS

These patients should get an extra treatment to prevent a hospitalization

PREDICTIVE ANALYTICS

Which patients will go to hospital next week?

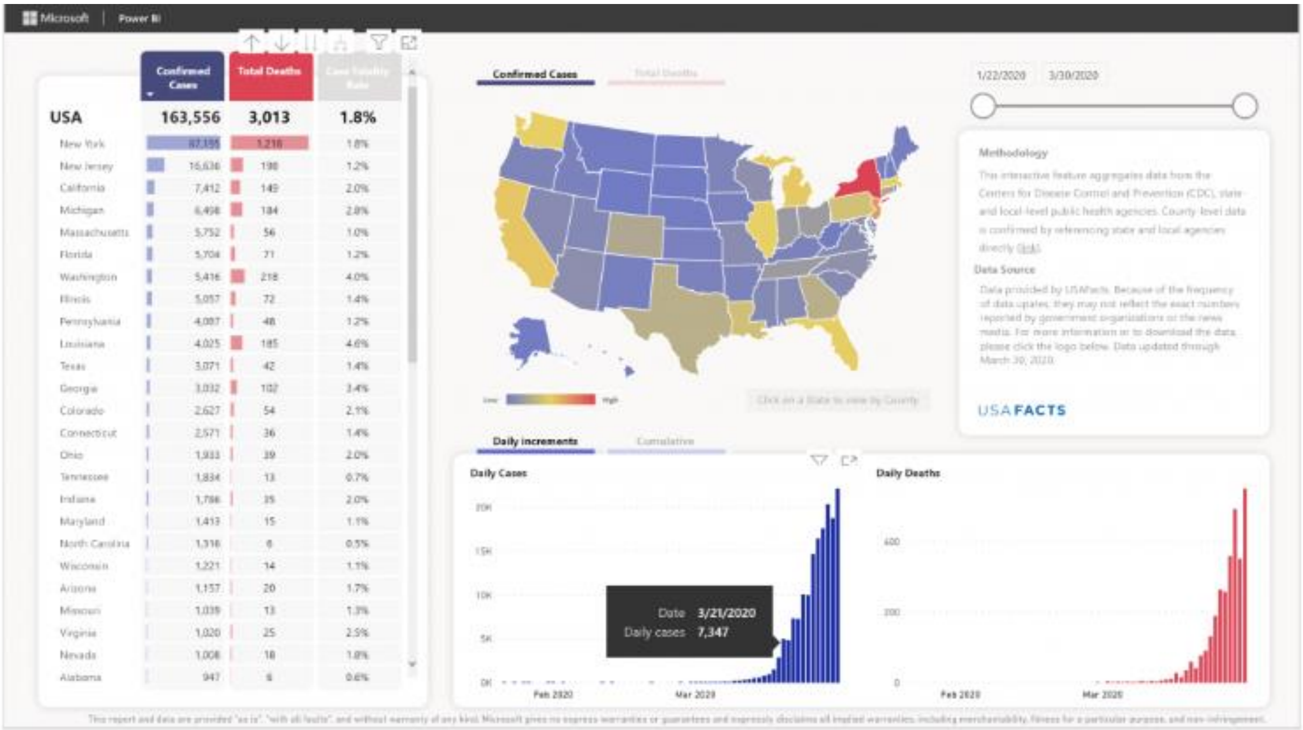
DIAGNOSTIC ANALYTICS

Why did these patients go to hospital?

DESCRIPTIVE ANALYTICS

How many patients went to hospital last week?

Diagrams illustrating data visualization



	Confirmed Cases	Total Deaths	Case Fatality Rate
USA	163,556	3,013	1.8%
New York	67,155	1,218	1.8%
New Jersey	16,636	198	1.2%
California	7,412	149	2.0%
Michigan	6,498	184	2.8%
Massachusetts	5,752	56	1.0%
Florida	5,704	71	1.2%
Washington	5,416	218	4.0%
Illinois	5,057	72	1.4%
Pennsylvania	4,087	48	1.2%
Louisiana	4,025	185	4.6%
Texas	3,071	42	1.4%
Georgia	3,032	102	3.4%
Colorado	2,627	54	2.1%
Connecticut	2,571	36	1.4%
Ohio	1,933	39	2.0%
Tennessee	1,834	13	0.7%
Indiana	1,786	35	2.0%
Maryland	1,413	15	1.1%
North Carolina	1,316	6	0.5%
Wisconsin	1,221	14	1.1%
Arizona	1,157	20	1.7%
Missouri	1,039	13	1.3%
Virginia	1,020	25	2.5%
Nevada	1,008	18	1.8%
Alabama	947	6	0.6%

1/22/2020 3/30/2020

Methodology

This interactive feature aggregates data from the Centers for Disease Control and Prevention (CDC), state and local-level public health agencies. County level data is confirmed by referencing state and local agencies directly (link).

Data Source

Data provided by USAFacts. Because of the frequency of data upates, they may not reflect the exact numbers reported by government organizations or the news media. For more information or to download the data, please click the logo below. Data updated through March 30, 2020.

USAFACTS

Date 3/21/2020
Daily cases 7,347

This report and data are provided "as is", "with all faults", and without warranty of any kind. Microsoft gives no express warranties or guarantees and expressly disclaims all implied warranties, including merchantability, fitness for a particular purpose, and non-infringement.

Fig 1: Power BI releases toolkit for COVID-19 data visualizations

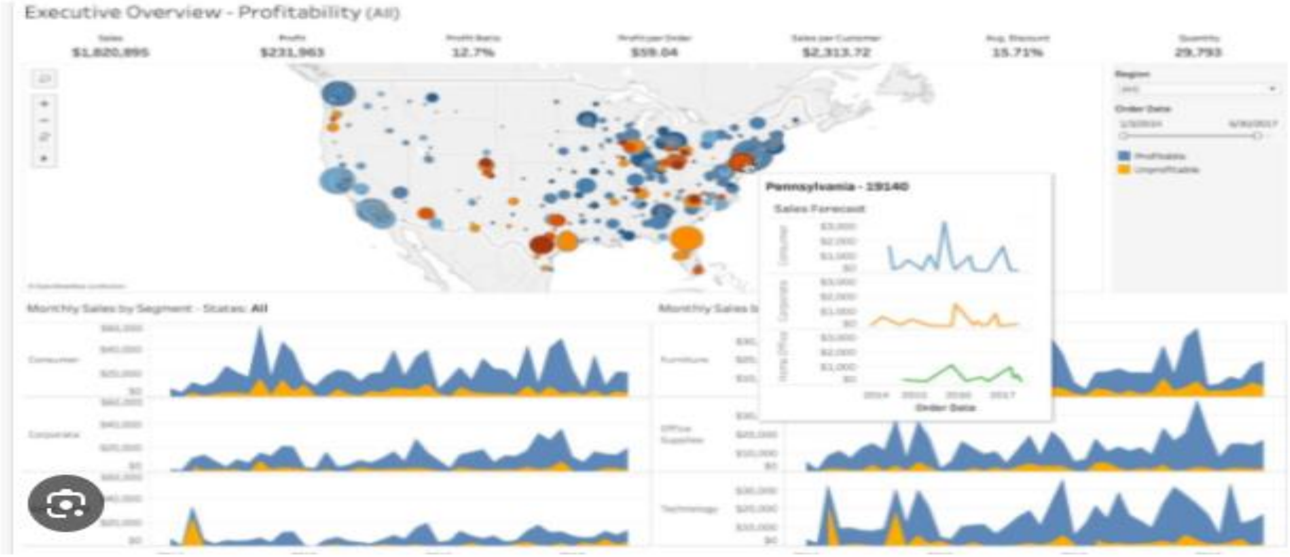
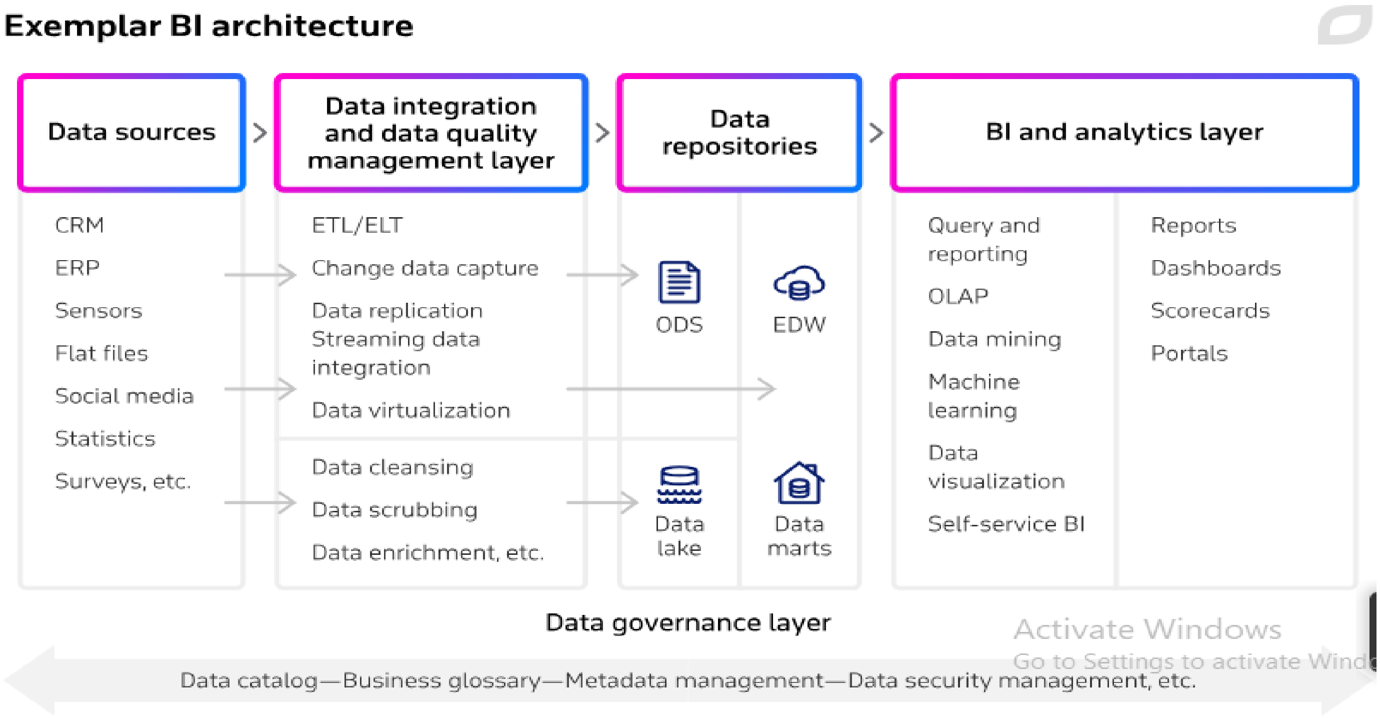


Fig 2: data visualization with Tableau Tool kit

What is BI architecture?

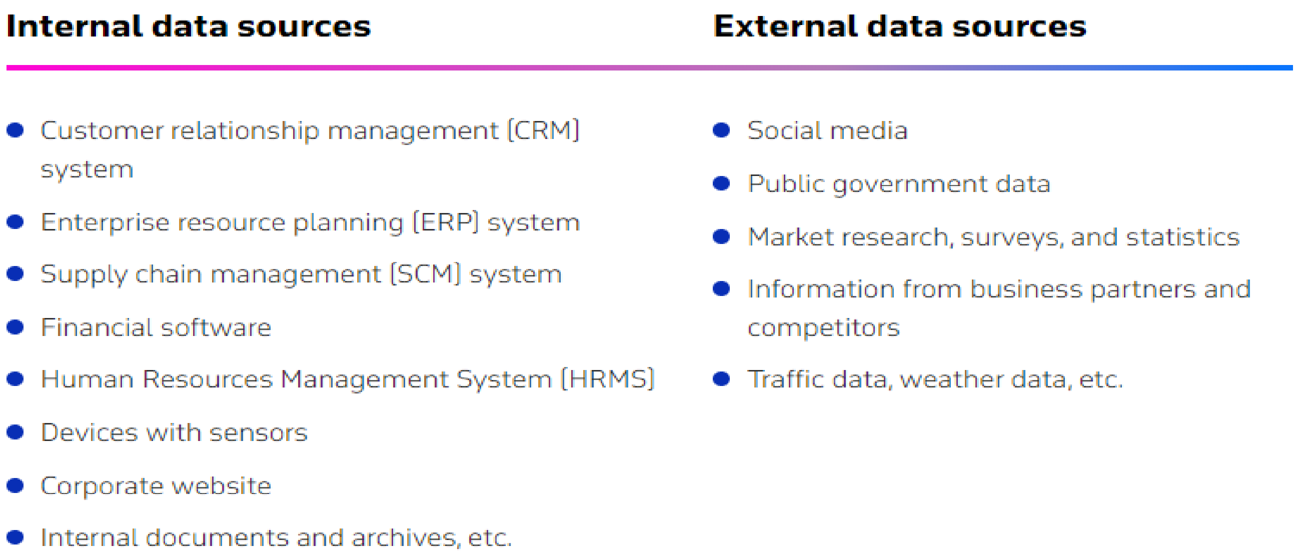
Business intelligence (BI) architecture - is the **infrastructure** a company deploys to support all the stages of the BI process *– from data collection, cleansing, structuring, storage, and analysis to reports and dashboard delivery and insight operationalization*.

5 components to build a business intelligence architecture



1. Data sources

Anything that produces the digital information BI systems consume is considered a data source. Data sources can be internal, with the information captured and maintained within a company, and external, when the information is generated outside the organization.



Internal data sources

- Customer relationship management (CRM) system
- Enterprise resource planning (ERP) system
- Supply chain management (SCM) system
- Financial software
- Human Resources Management System (HRMS)
- Devices with sensors
- Corporate website
- Internal documents and archives, etc.

External data sources

- Social media
- Public government data
- Market research, surveys, and statistics
- Information from business partners and competitors
- Traffic data, weather data, etc.

2. Data integration and data quality management layer

The second step of the BI process is aimed at consolidating datasets from multiple sources for a unified view – that is how the information becomes viable for analytics and operational purposes. There are several data integration methods, the choice of which is dictated by the information type, format, and volumes, as well as the purpose – operational reporting, business analysis, ML use cases, etc.

Extract, Transform, and Load (ETL)

Extract, Transform, and Load (ETL) involves the retrieval of batches of information from the sources of data, conversion into another format/structure, and placement into ultimate storage. While the extract and load parts are rather mechanical, the transformation stage is a complex activity, which involves:

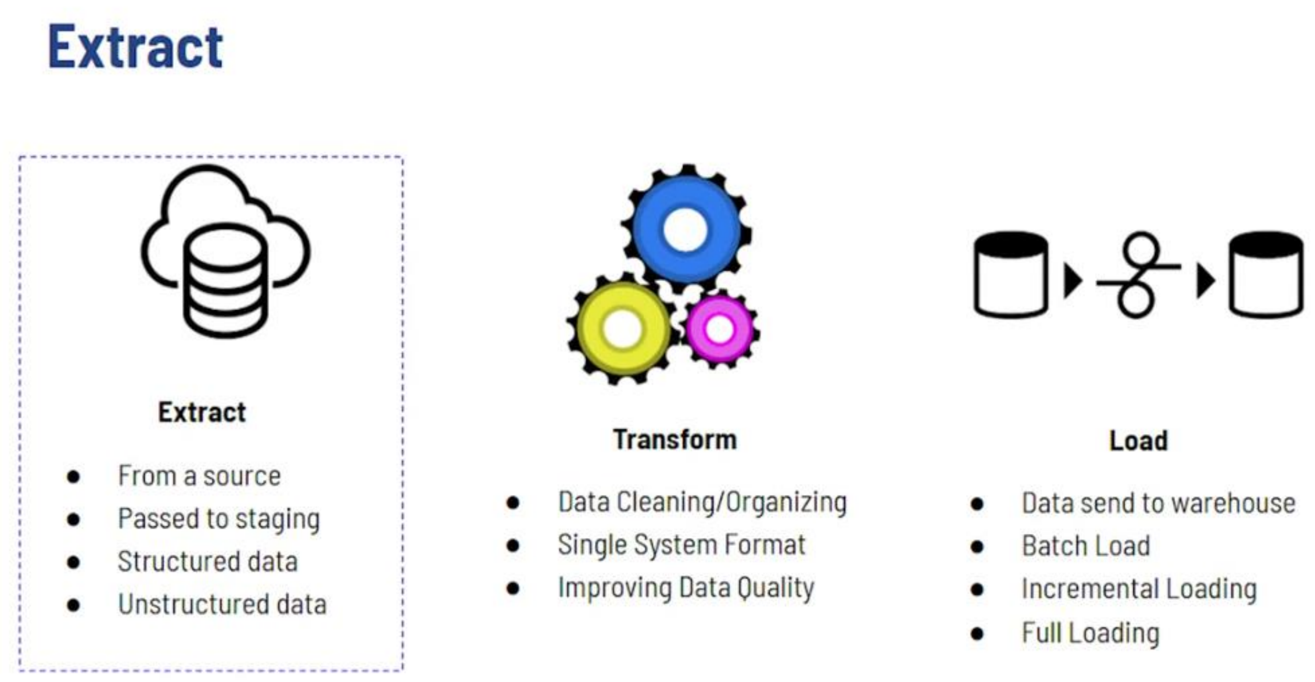
- **Data profiling** – a detailed examination of information, its type, structure, and quality, which would define what kind of transformations are reasonable.
- **Data mapping** – matching the data field of the source to the ultimate one.
- Code creation and actual execution to convert data according to the mapping rules.
- **Data audit** - to ensure the performed transformation is correct and the output data adheres to the set requirement.

What is ETL/ELT and their significance in Business intelligence?

Click the Link and Read the content

https://aws.amazon.com/what-is/etl/#:~:text=Extract%2C%20transform%2C%20and%20load%20(,and%20machine%20learning%20(ML).

What is ETL for Beginners | ETL Non-Technical Explanation



Extract

Extract

- From a source
- Passed to staging
- Structured data
- Unstructured data

Transform

- Data Cleaning/Organizing
- Single System Format
- Improving Data Quality

Load

- Data send to warehouse
- Batch Load
- Incremental Loading
- Full Loading

ETL Tools

APACHE nifi
Informatica
talend

The exact transformations (which are multiple) are defined by business rules. They may be:

Aggregation of several columns into a single one or vice versa, the split of a column into several ones. Encodement of values or translation of the existing ones (‘Male’ to ‘M’, ‘Female’ to ‘F’, or ‘1’ to ‘Female’, ‘2’ to ‘Male’, etc.).

Creation of new calculations (for example, to follow varying accounting rules).

Conversion of low-level data attributes into high level-attributes.

Derivation of new attributes out of the existing ones, etc.

Extract, Load, and Transform

An alternative to the ETL process, the ELT approach implies the transformation happens after data loading. Firstly, this approach saves time and resources, secondly, it better suits the needs of data scientists and data analysts who often want to experiment with raw data and perform a non-trivial transformation. That explains the predominant application of this approach for ML, AI and big data scenarios.

Data replication

The process can take place either in batches or real-time streams and encompasses copying information from the source system to its destination with no or minimal transformations. Data replication is used as an optimal data integration method when a company needs to copy data for backup, disaster recovery, or operational reporting.

Change data capture

This is a real-time data integration method that aims to detect what data changes happened in the source systems and update the destination storage accordingly.

Streaming data integration

This method implies continuous integration of real-time data for operational reporting, real-time analytics, or temporary storage before further processing.

Data virtualization

This data integration method stands apart from the rest as it does not imply the physical movement of information and provides the consolidated data view by creating a unified logical view accessible via a semantic layer.

Data quality management

Data integration and data cleansing are two processes happening in parallel. Data ingested from multiple sources may be inconsistent or data sets may be duplicated. In addition to the problems that occur when data is collected from numerous sources, data may be just of poor quality, with some information missing, irrelevant in terms of time or value, etc. To deal with these issues, the component is structured with the technologies, which automate:

- *Assessment of data quality and identification of data issues (data profiling)*
- *Correction of data errors, data scrubbing (removing duplicate and bad data), data enrichment, etc.*
- *Audit of data quality against the set data quality metrics*
- *Reporting on data quality issues, trends, performed activities, etc.*

3. Data repositories

This component encompasses various repositories that structure and store data for further processing. There are two major data repository groups:

Analytical data stores

Enterprise data warehouse – a unified repository with cleaned, consolidated data. Businesses use different types of databases for this purpose:

relational (stores data in rows)

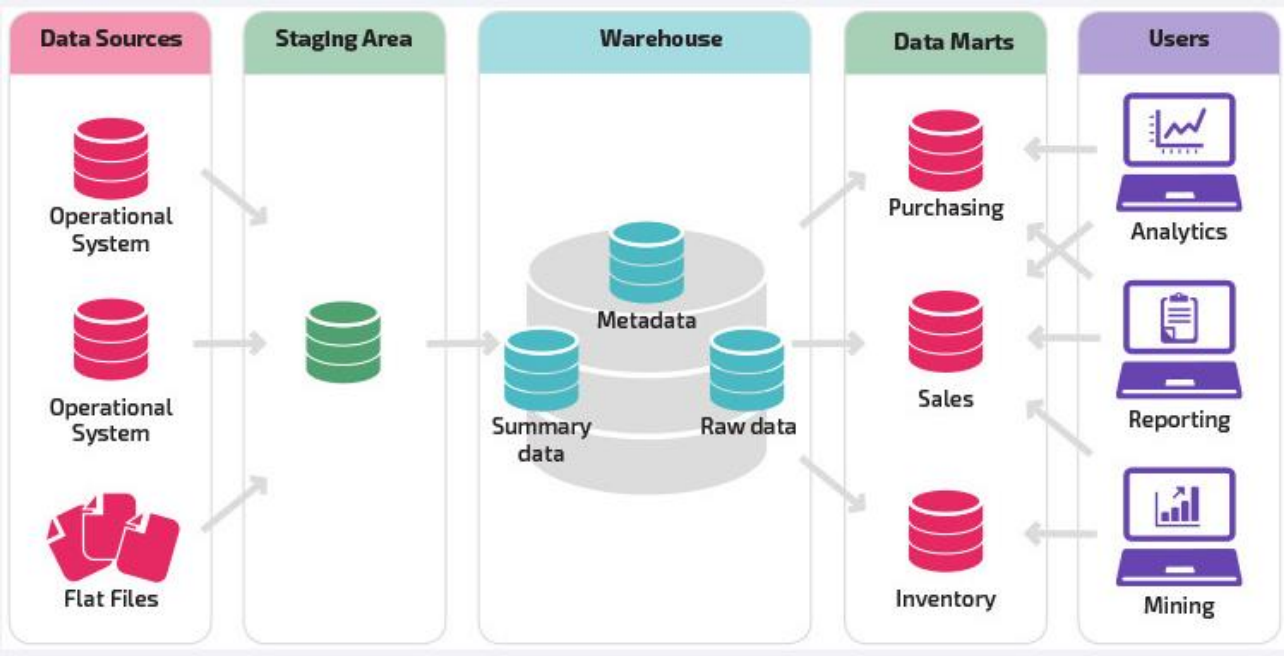
columnar (stores data in columns)

multidimensional (stores data in a data cube format)

Data marts – *a subset of a data warehouse*

- focused on a particular line of business, department, or subject area.

- Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.



Data Mart and Data Warehouse Difference

Data Mart

Focus: A single subject or functional organization area

Data Sources: Relatively few sources linked to one line of business

Size: Less than 100 GB

Normalization: No preference between a normalized and denormalized structure

Decision Types: Tactical decisions pertaining to particular business lines and ways of doing things

Cost: Typically from $10,000 upwards

Setup Time: 3-6 months

Data Held: Typically summarized data

Data Warehouse

Focus: Enterprise-wide repository of disparate data sources

Data Sources: Many external and internal sources from different areas of an organization

Size: 100 GB minimum but often in the range of terabytes for large organizations

Normalization: Modern warehouses are mostly renormalized for quicker data querying and read performance

Decision Types: Strategic decisions that affect the entire enterprise

Cost: Varies but often greater than $100,000; for cloud solutions costs can be dramatically lower as organizations pay per use

Setup Time: At least a year for on-premise warehouses; cloud data warehouses are much quicker to set up

Data Held: Raw data, metadata, and summary data

Data Marts Use Cases

1. Marketing analysis and reporting favor a data mart approach because these activities are typically performed in a specialized business unit, and do not require enterprise-wide data.
2. A financial analyst can use a finance data mart to carry out financial reporting.

Power BI DataMart - What, How and Why?

Video link below how to use BI tool- power Bi

https://www.youtube.com/watch?v=0QD7hXw8ggU

Centralized Data Warehouse Use Cases

1. A company considering an expansion needs to incorporate data from a variety of data sources across the organization to come to an informed decision. This requires a data warehouse that *aggregates data from sales, marketing, store management, customer loyalty, supply chains, etc.*
2. Many factors drive profitability at an insurance company. *An insurance company reporting on its profits needs a centralized data warehouse to combine information from its claims department, sales, customer demographics, investments, and other areas.*

Enterprise Data Warehouses

4. BI and analytics layer

This layer encompasses solutions for accessing and working with data and aimed at data analysts, data scientists, or business users.

This layer naturally reflects the organization's BI maturity and its data analytics objectives: for some companies descriptive and diagnostic analytics capabilities are sufficient enough, others need to run *comprehensive analysis supported with ML and AI via a self-service user interface.*

The portfolio of tools may include:

1. Query and reporting tools to request specific information and create reports with the derived insights. The reports may be delivered to business users via email on a scheduled basis or may be triggered by some events (for example, by a sudden drop in sales). They also may be embedded into applications business users leverage daily for enhanced user experience and quick operationalization.
2. Online Analytical Processing (OLAP) tools *to roll up and roll down, drill down, slice and dice,* etc. business data placed into multidimensional cubes beforehand.
3. Data mining tools to *search for trends, patterns, and hidden correlations* in data.
4. ML and AI tools to create models *that help companies predict future events,* model what-if scenarios, automate analytics-related processes for people without domain background, etc.

5. Data visualization tools to create *dashboards and scorecards* which then can be shared in a secure viewer environment, via a public URL, or through embedding into user applications.

If the solution is equipped with self-service capabilities, business users may not only passively consume the reports and dashboards curated for them by dedicated teams, but also run their analysis, build dashboards and scorecards, edit the existing content, and share their findings with colleagues.

5. Data governance layer

This element is closely intertwined with the other four, as *its major aim is to monitor and govern the end-to-end BI process.* With data governance standards and policies in place, a company controls who accesses the information and how, if the information used for analysis is of proper quality and is safeguarded well, etc. All these policies and standards make up a data management program that can be automated with the data governance tools with capabilities like:

- Data catalogs – capturing data and cataloging it with categories, tags, indexes, etc., which helps both tech and business users know what data is available, where it is maintained, who can access it, what the sensitivity risks are, etc.
- Business glossaries – authoritative sources with the common definitions of business terms for business users from different departments to eliminate any ambiguity.
- Low-code or no-code creation and configuration of data governance workflows and built-in data stewardship functions for data governance teams to manage data-related issues (for example, approve business glossary entries).
- Automated data lineage documentation for data quality management and compliance with data privacy laws.
- Role-based access control for setting user permissions.
- Automated data quality metrics generation, measurement, quality levels monitoring, etc.
- Centralized data policies and standards management (creation, configuration, monitoring of adoption and compliance, etc.), and so on.

Business intelligence teams: core roles and responsibilities

BI program manager

- Defines the scope of the BI program and each BI initiative, as well as timeframes, resources, deliverables, etc.
- Establishes the collaboration of the involved parties (BI teams involved in different BI initiatives)

- Oversees the business intelligence program execution and recommends changes to the existing BI processes based on their analysis, industry trends, new goals, objectives, etc.

BI project manager

- Defines the scope of the BI project, its objectives, stages, deliverables, success metrics, etc.
- Provides project estimates and project scheduling, assesses project risks and suggests solutions
- Oversees project execution, ensuring deadlines and goals are met
- Sets up communication with all involved stakeholders

BI solution architect

- Cooperates with business analysts and business stakeholders to define business requirements, designs appropriate BI solutions, and supervises their development
- Evaluates the existing BI environment, creates new requirements, prioritizes change requests, etc.
- Defines and improves data governance and data security practices

BI developer

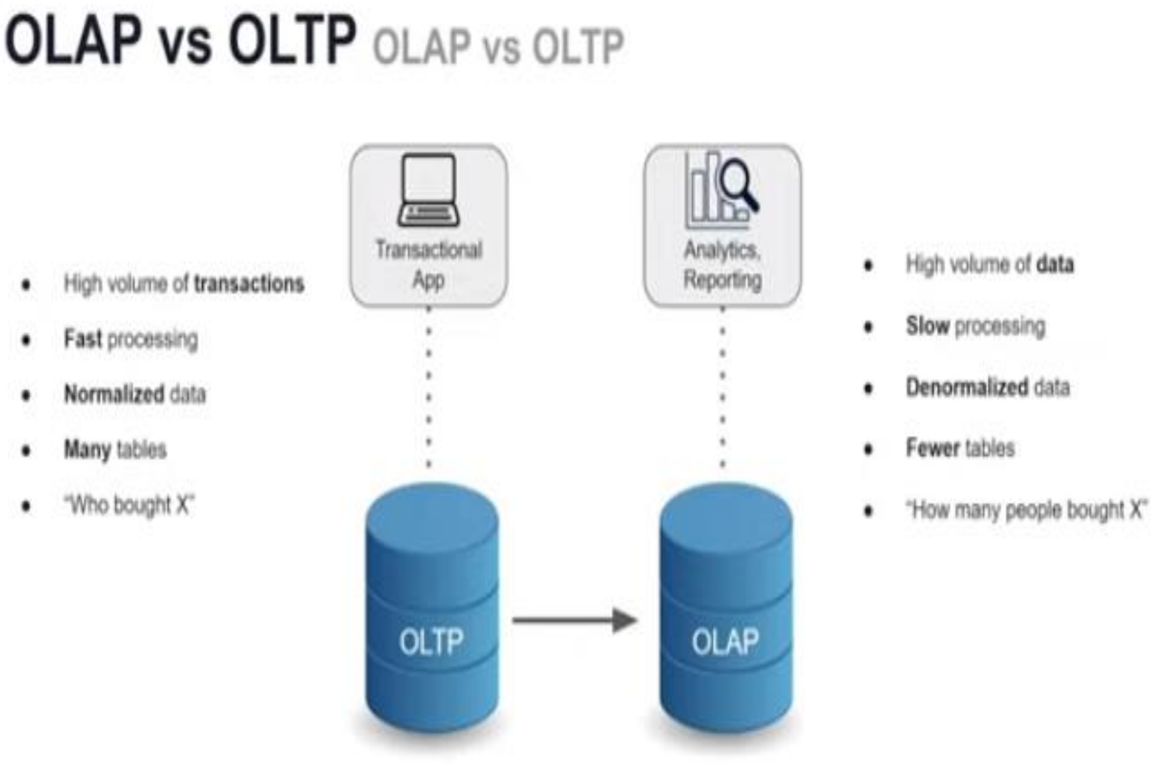
- Develops the BI solution, including data models, ETL/ELT pipelines, reports and dashboards, etc.
- Manages and maintains BI solution components
- Performs troubleshooting, optimizes reports and dashboards, handles data quality issues, etc.

BI analyst

- Reviews and validates business data and develops policies for data management
- Manages master and metadata including creation, update, and deletion
- Gathers business user requirements, supports end-users, and consults leadership

BI systems administrator

- Manages BI systems, monitoring systems performance, availability, backup, updates, etc.
- Installs and configures security settings, user access controls, etc.
- Performs troubleshooting and provides tech support to BI users, etc.



OLAP vs OLTP OLAP vs OLTP

- High volume of **transactions**
- **Fast** processing
- **Normalized** data
- **Many** tables
- "Who bought X"

- High volume of **data**
- **Slow** processing
- **Denormalized** data
- **Fewer** tables
- "How many people bought X"

OLAP VS OLTP a simple explanation in 4 mins

https://www.youtube.com/watch?v=P7hf_emjsRI