رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

IBM
What is Data Science?

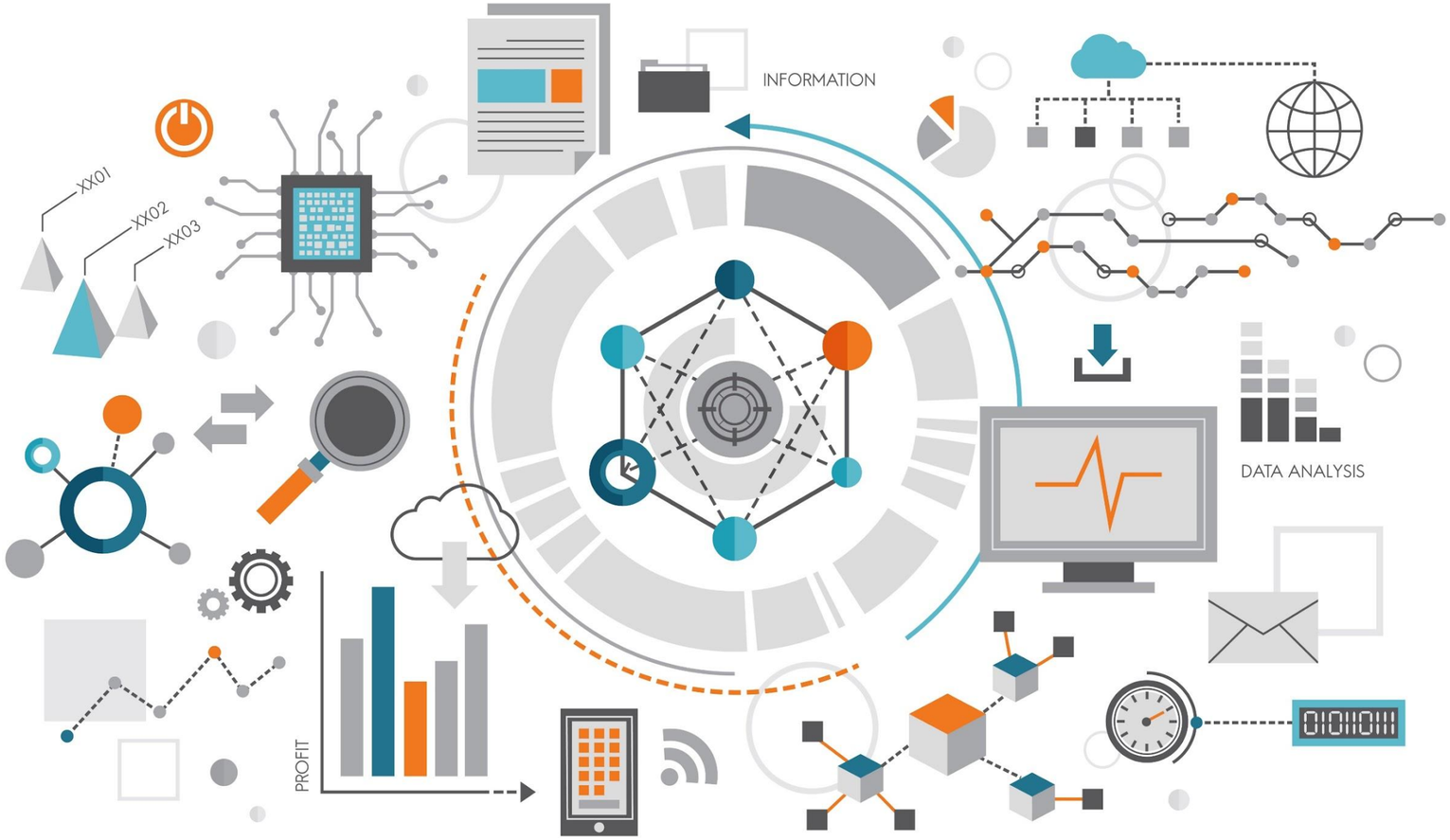
9/3/2024

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Module II: Data science Topics

How Bia Data is Drivina Diaital Transformation?




رواد مصر الرقمية

Understanding Digital Transformation

- **Digital Transformation:** Overhauling business operations to leverage new technologies effectively.
- **Integration of Digital Technology:** Fundamental changes in operations and value delivery across all areas.
- **Driven by Data Science and Big Data:** Utilizing vast data resources for competitive advantage and innovation.
- **Industry Examples:** Netflix, Houston Rockets, and Lufthansa embracing digital transformation for success.
- **Core Organizational Change:** Digital transformation affects businesses fundamentally and culturally.
- **Example Case:** Houston Rockets' use of Big Data to revolutionize basketball strategy.

Key Aspects of Digital Transformation

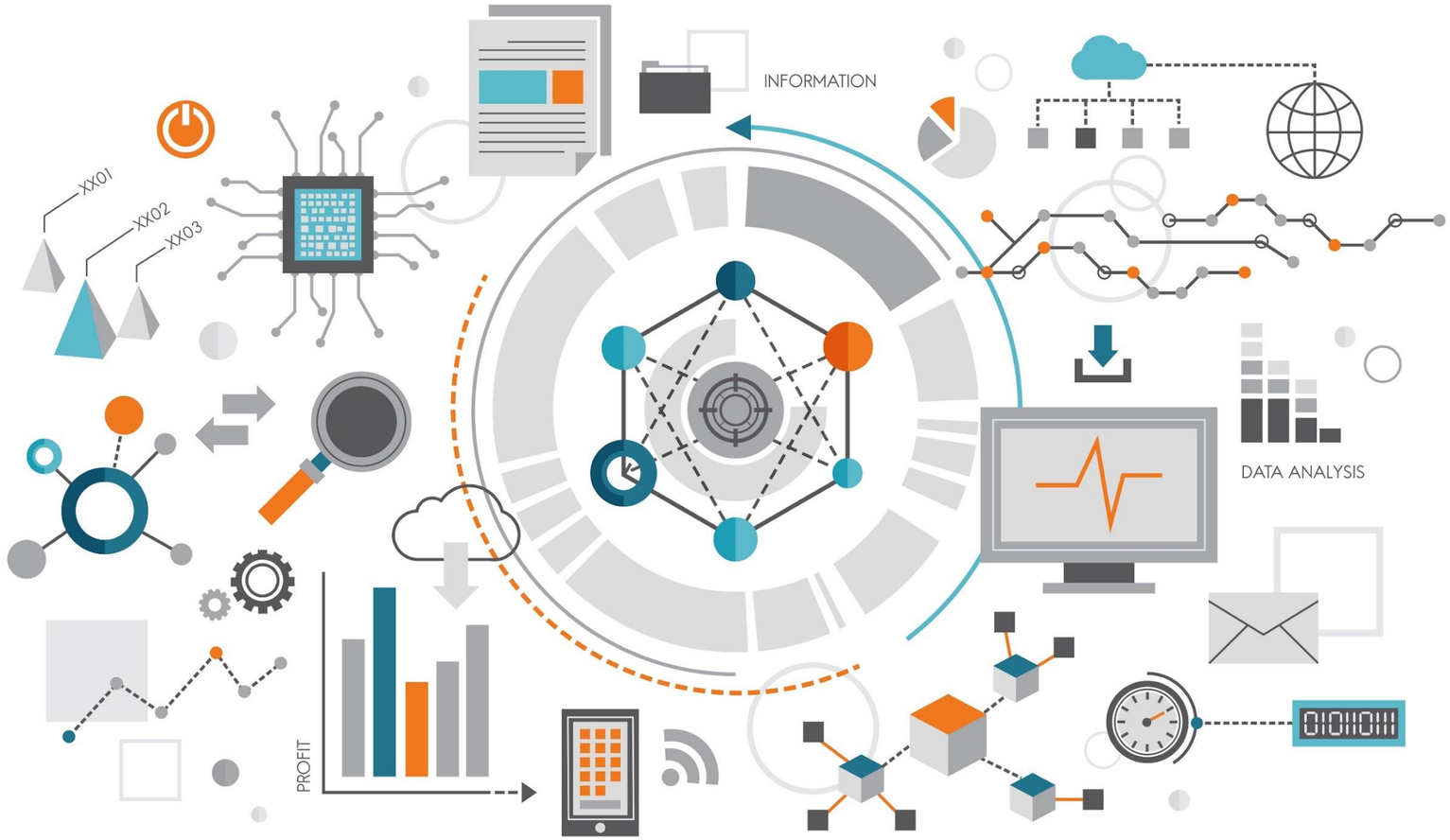
- **Process Improvement:** In-depth analysis leads to enhancements in operations and workflows.
- **Organizational Culture:** Requires fundamental changes in approaches to data, employees, and customers.
- **Leadership Involvement:** Decision-makers at top levels crucial for successful implementation.
- **Executive Support:** CEO, CIO, and emerging role of Chief Data Officer pivotal in guiding transformation.
- **Whole Organization Approach:** Success relies on support from all levels and departments.
- **New Mindset:** Navigating challenges of digital transformation necessitates adopting a forward-thinking perspective.

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Module II: Data science Topics

Introduction to Cloud



Understanding Cloud Computing

- **Cloud Computing:** Delivery of on-demand computing resources over the Internet.
- **Examples:** Online web apps, secure business applications, cloud-based storage platforms.
- **User Benefits:** Cost-effectiveness, access to latest application versions, collaborative work.
- **Essential Characteristics:** On-demand self-service, broad network access, resource pooling.
- **Characteristics Continued:** Rapid elasticity, measured service for transparent payment based on usage.
- **Transformation Impact:** Cloud computing changes how organizations consume compute services.

Cloud Deployment and Service Models

- **Deployment Models:** Public, private, and hybrid clouds based on infrastructure ownership.
  - **Public Cloud:** Leverages services over the open internet, shared by multiple companies.
  - **Private Cloud:** Infrastructure provisioned exclusively for a single organization.
  - **Hybrid Cloud:** Seamless integration of public and private clouds.
- **Service Models:** Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS).
  - **IaaS (Infrastructure as a Service (IaaS):** Access to physical computing resources without managing them.
  - **PaaS (Platform as a Service):** Access to **hardware and software tools** for application development and deployment.
  - **SaaS (Software as a Service):** Centralized hosting and licensing of software on a subscription basis.

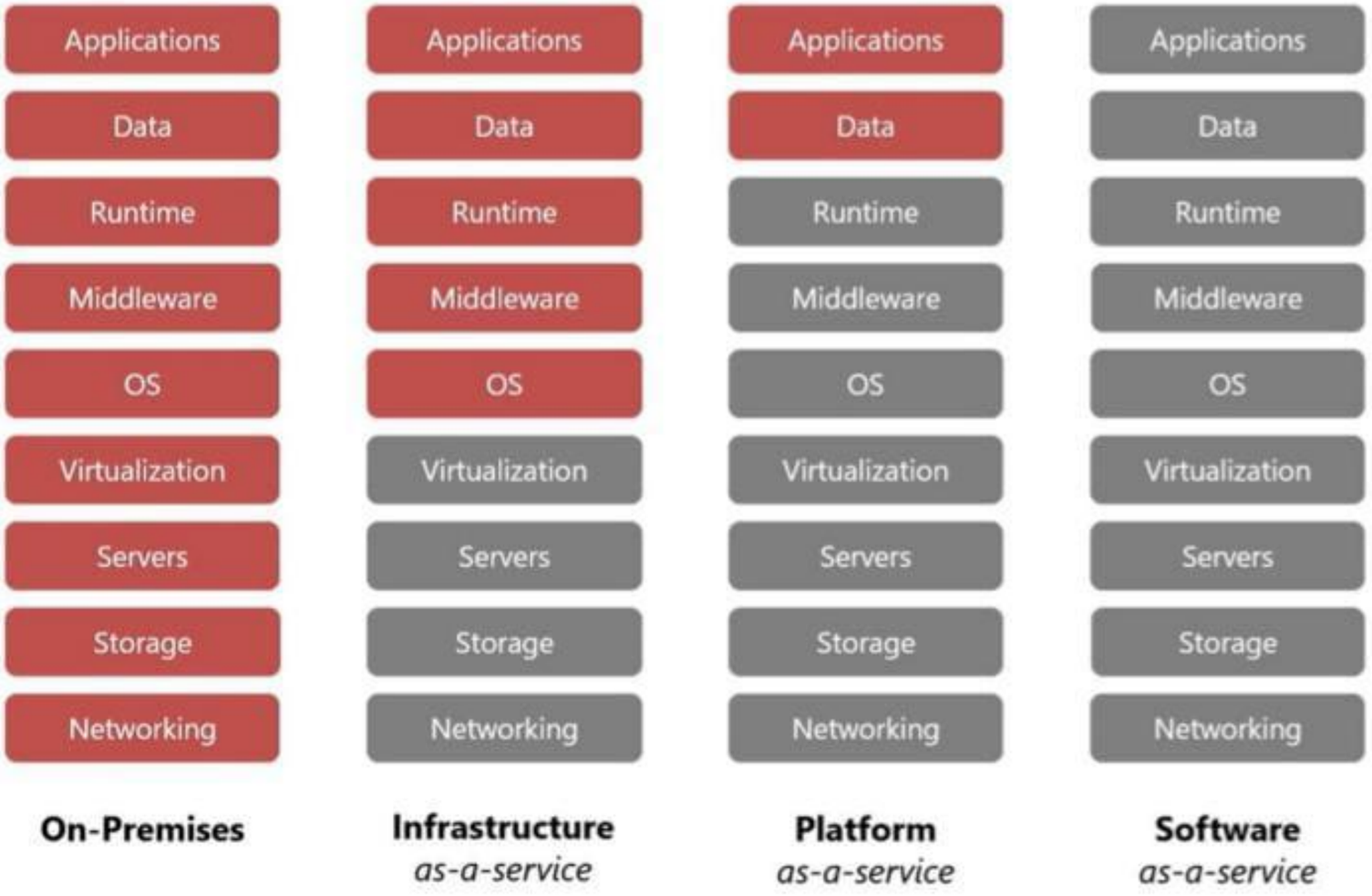
Benefits of Cloud Computing for Data Scientists

- **Centralized Storage**:
  - Eliminate physical storage limits by securely storing vast amounts of data in the Cloud.
- **Access to High-Performance Machines**:
  - Deploy complex data analytics on advanced computing machines without needing expensive hardware.
- **Scalable Algorithm Deployment**:
  - Run advanced algorithms at scale using cloud-based platforms designed for data-intensive tasks.
- **Collaborative Work**:
  - Enable global teams to work on the same data simultaneously, enhancing productivity and collaboration.
- **Instant Access to Technologies**:
  - Quickly leverage open-source technologies like Apache Spark, TensorFlow, and more, without setup delays.
- **Up-to-Date Tools and Libraries**:
  - Benefit from the latest tools, libraries, and frameworks automatically updated by cloud providers, reducing maintenance overhead.

Cloud Accessibility and Collaboration

- **Anytime, Anywhere Access:** Cloud technologies accessible from various devices globally.
- **Enhanced Collaboration:** Simultaneous data access enables easier collaboration among teams.
- **Pre-Built Environments:** Big tech companies offer Cloud platforms like IBM Cloud, AWS, Google Cloud.
- **IBM Skills Network Labs:** Access to tools like Jupyter Notebooks and Spark clusters for data science projects.
- **Productivity Enhancement:** Cloud dramatically enhances productivity for data scientists with practice.
- **Global Availability:** Cloud services available across different time zones, fostering collaboration and innovation.

Delivery Models



Types of Services

- Compute
- Networking
- Storage
- Databases
- Web
- IoT / Event
- Big Data / Analytics
- Identity
- AI
- Monitoring
- DevOps

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

Module II: Data science Topics

Introduction to Big Data Clusters

What is Hadoop?

Introduction to Big Data Clusters

- Traditional data processing involved bringing data to the computer and running programs on it.
- Big data clusters, pioneered by Larry Page and Sergey Brin, distribute and replicate data across thousands of computers for parallel processing.
- Map and reduce processes enable handling large datasets and scaling linearly with server additions.
- Hadoop, a popular big data architecture, emerged as Yahoo adopted Google's approach in 2008.

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

Module II: Data science Topics

Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark

Introduction to Big Data Processing Technologies

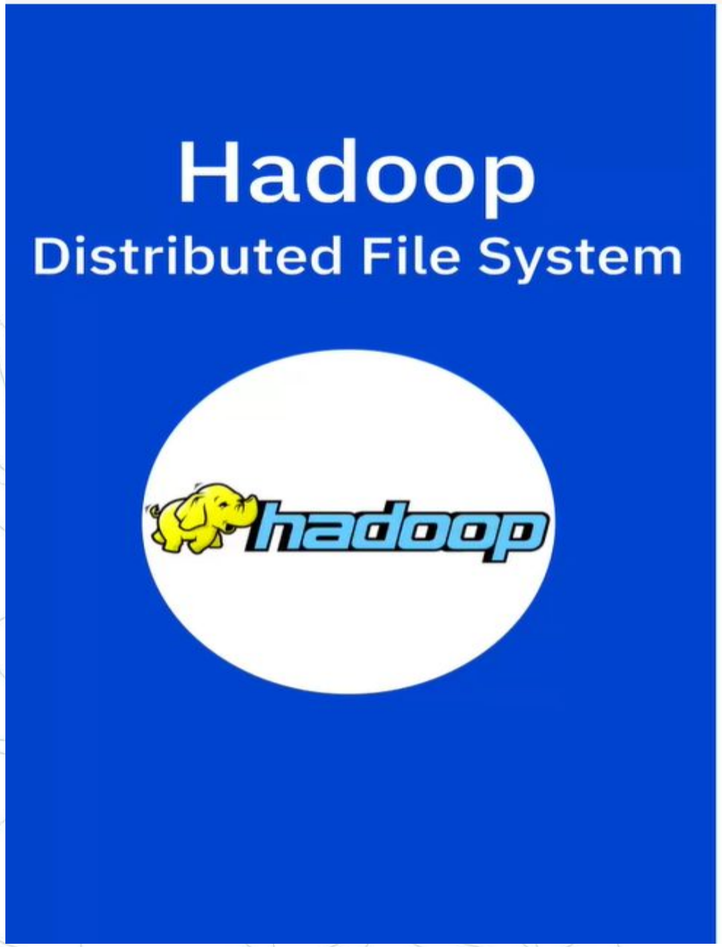
- Big Data processing enables handling large sets of structured, semi-structured, and unstructured data.

- **Overview of Apache Hadoop, Apache Hive, and Apache Spark in Big Data Analytics:**
  - **Hadoop**:
    - Offers distributed storage and processing capabilities for large datasets.
  - **Hive**:
    - Serves as a data warehouse on top of Hadoop for querying and analyzing data.
  - **Spark**:
    - A distributed analytics framework designed for complex real-time data analytics.

These technologies provide scalable, reliable, and cost-effective solutions for big data storage and processing.

Understanding Hadoop and HDFS

- Hadoop facilitates distributed storage and processing across clusters of computers.
- HDFS partitions files over multiple nodes, allowing parallel access and computations.
- Replication of file blocks ensures fault tolerance and availability in case of node failures.
- Data locality minimizes network congestion and increases throughput by moving computations closer to data nodes.
- Benefits of using HDFS include fast recovery, access to streaming data, scalability, and portability.
- Hadoop's ability to consolidate and optimize data storage across the organization enhances enterprise data warehouse efficiency.

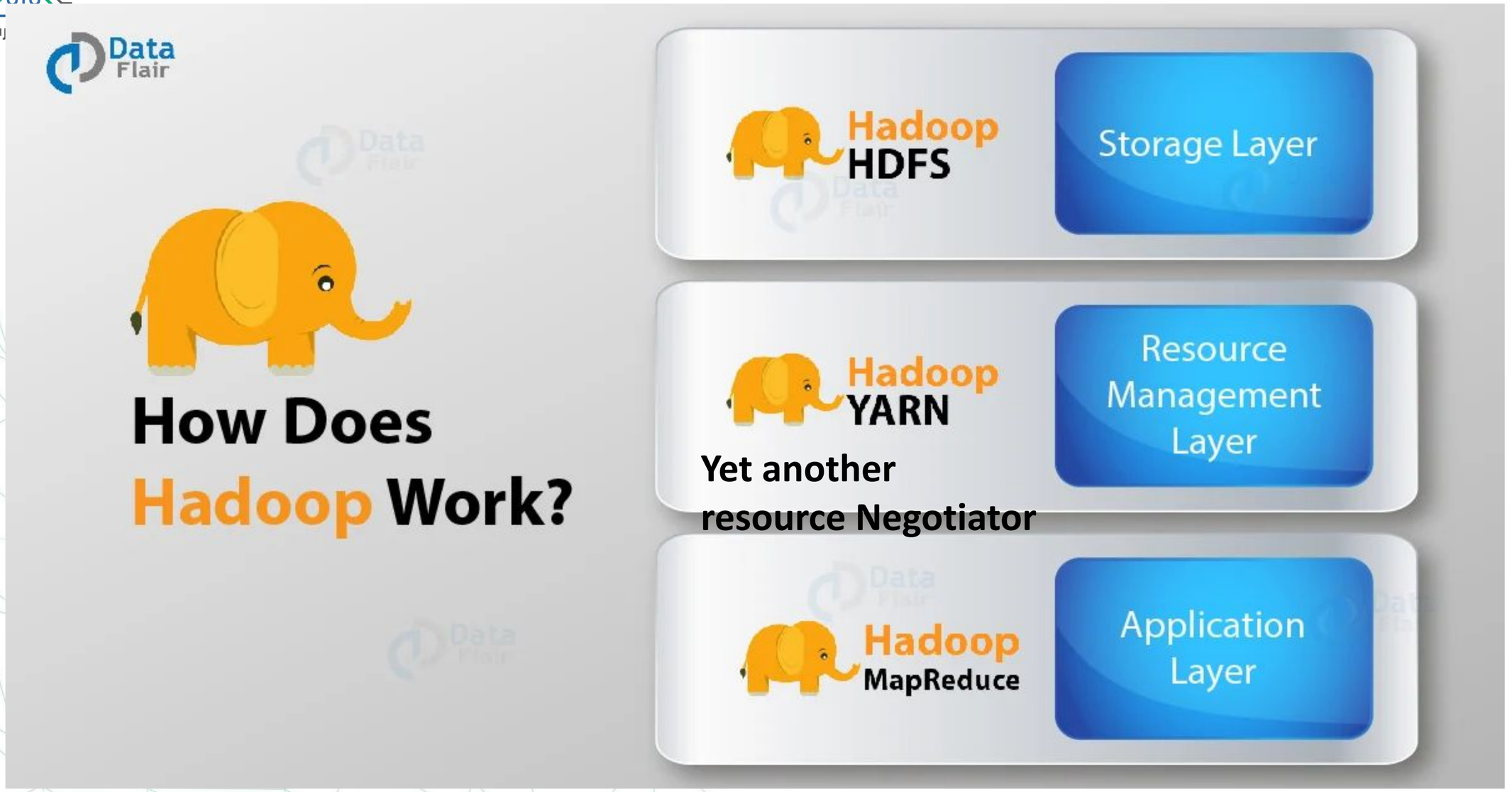


Hadoop Distributed File System, or HDFS, is a storage system for big data that runs on multiple commodity hardware connected through a network.

Provides scalable and reliable big data storage by partitioning files over multiple nodes

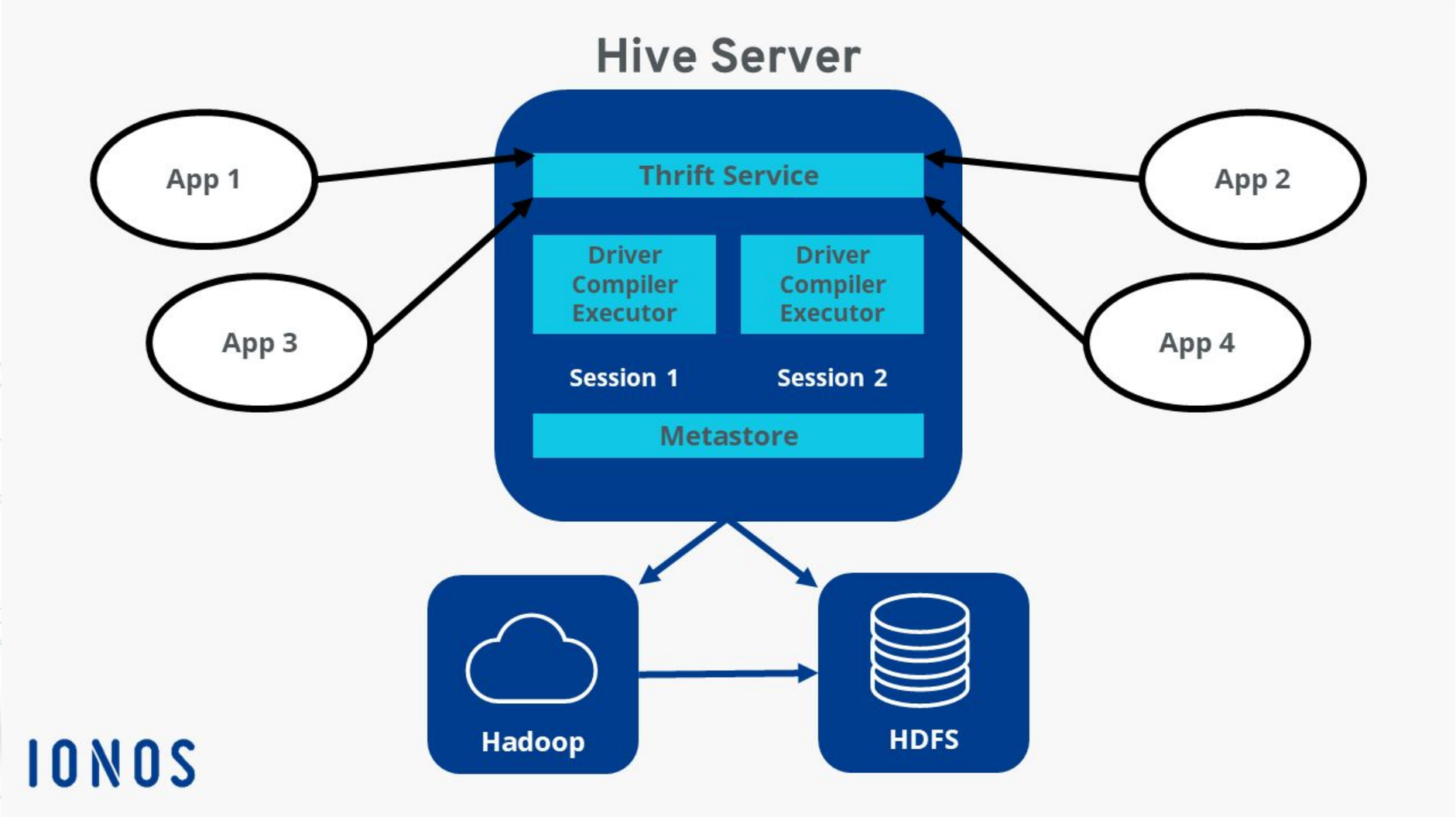
Splits large files across multiple computers, allowing parallel access to them

Replicates file blocks on different nodes to prevent data loss


Data Flair
How Does Hadoop Work?
Hadoop HDFS
Storage Layer
Hadoop YARN
Resource Management Layer
Yet another resource Negotiator
Hadoop MapReduce
Application Layer


HIVE
Apache Hive
a data warehouse for data query and analysis


APACHE
Spark
Apache Spark
a distributed analytics framework for complex,
real-time data analytics


Hive Server
App 1
App 2
App 3
App 4
Thrift Service
Driver
Compiler
Executor
Driver
Compiler
Executor
Session 1
Session 2
Metastore
Hadoop
HDFS
IONOS

Integration of Hadoop, Hive, and Spark

1. Hadoop is the core framework for big data storage and processing:

- It consists of main components:
    - **HDFS (Hadoop Distributed File System):** A distributed file system that stores large volumes of data across multiple nodes.
    - **YARN (Yet Another Resource Negotiator):** A resource manager that schedules and manages tasks across the cluster.
    - **MapReduce** processing model:
        i. **Map Phase:** Breaks input data into chunks and transforms them into key-value pairs in parallel.
        ii. **Shuffle & Sort Phase:** Groups all key-value pairs by keys and sorts them.
        iii. **Reduce Phase:** Aggregates values for each key to produce the final result.

        MapReduce jobs are executed over data stored in HDFS, often as the default execution engine.

2. Hive is a SQL-based query engine built on top of Hadoop:

- Allows users to write queries in **HiveQL** (similar to SQL).
- Translates those queries into **MapReduce** or **Spark** jobs under the hood.
- Reads from and writes to **HDFS** seamlessly.

3. Spark is a high-performance distributed processing engine (Real-Time Streams):

- Can run on top of **YARN** using Hadoop's cluster resources.
- Efficiently reads from and writes to **HDFS**.
- Supports **Hive queries** through **Spark SQL**, enabling faster in-memory processing.
- Outperforms MapReduce by processing data directly in memory, reducing I/O overhead.

Practical Workflow:

- Raw data is stored in **HDFS**.
- Analysts write queries using **Hive**.
- The queries are executed using either **MapReduce** or **Spark**, depending on configuration.
- **YARN** handles resource allocation and task scheduling across the nodes.

Introduction to Apache HBase

- **What is HBase?**
  - A distributed, scalable NoSQL database that runs on top of HDFS.
  - Designed to handle large amounts of sparse, unstructured data in real-time.
- **Key Features**:
  - **Column-Oriented Storage**: Ideal for querying specific columns efficiently.
  - **NoSQL**: Flexible data storage without relying on predefined schemas.
  - **Real-Time Data Access**: Supports fast read/write operations on large datasets.

HBase in the Hadoop Ecosystem

- **Scalability**:
  - Horizontally scales across thousands of servers, managing petabytes of data.
- **Integration with Hadoop**:
  - Uses HDFS for fault-tolerant storage.
  - Supports MapReduce for processing large-scale data.
- **Use Cases**:
  - Web analytics, log data analysis, and applications needing fast access to big data.
- **Why HBase?**
  - Real-time access and processing, complementing Hadoop's batch processing capabilities.

Introduction to Data Mining

- **What is Data Mining?**
  - The process of discovering patterns, trends, and useful insights from large datasets.
  - Involves techniques from statistics, machine learning, and database systems.
- **Purpose of Data Mining**:
  - To transform raw data into actionable information.
  - Helps in decision-making, predictive analytics, and identifying hidden patterns.

Key Techniques in Data Mining

Classification:

- Assigning data to predefined categories based on its attributes.
- Commonly used in spam detection and fraud detection.

Clustering:

- Grouping similar data points together based on shared characteristics.
- Used for market segmentation and customer profiling.

Association Rule Learning:

- Identifying relationships between variables in large datasets.
- Commonly used for market basket analysis (e.g., "customers who buy X also buy Y").

Applications of Data Mining

Business Intelligence:
- Helps companies understand customer behavior, market trends, and product performance.

Healthcare:
- Used for diagnosing diseases, predicting patient outcomes, and optimizing treatment plans.

E-commerce and Retail:
- Analyzes customer purchase patterns to optimize sales strategies and personalized recommendations.

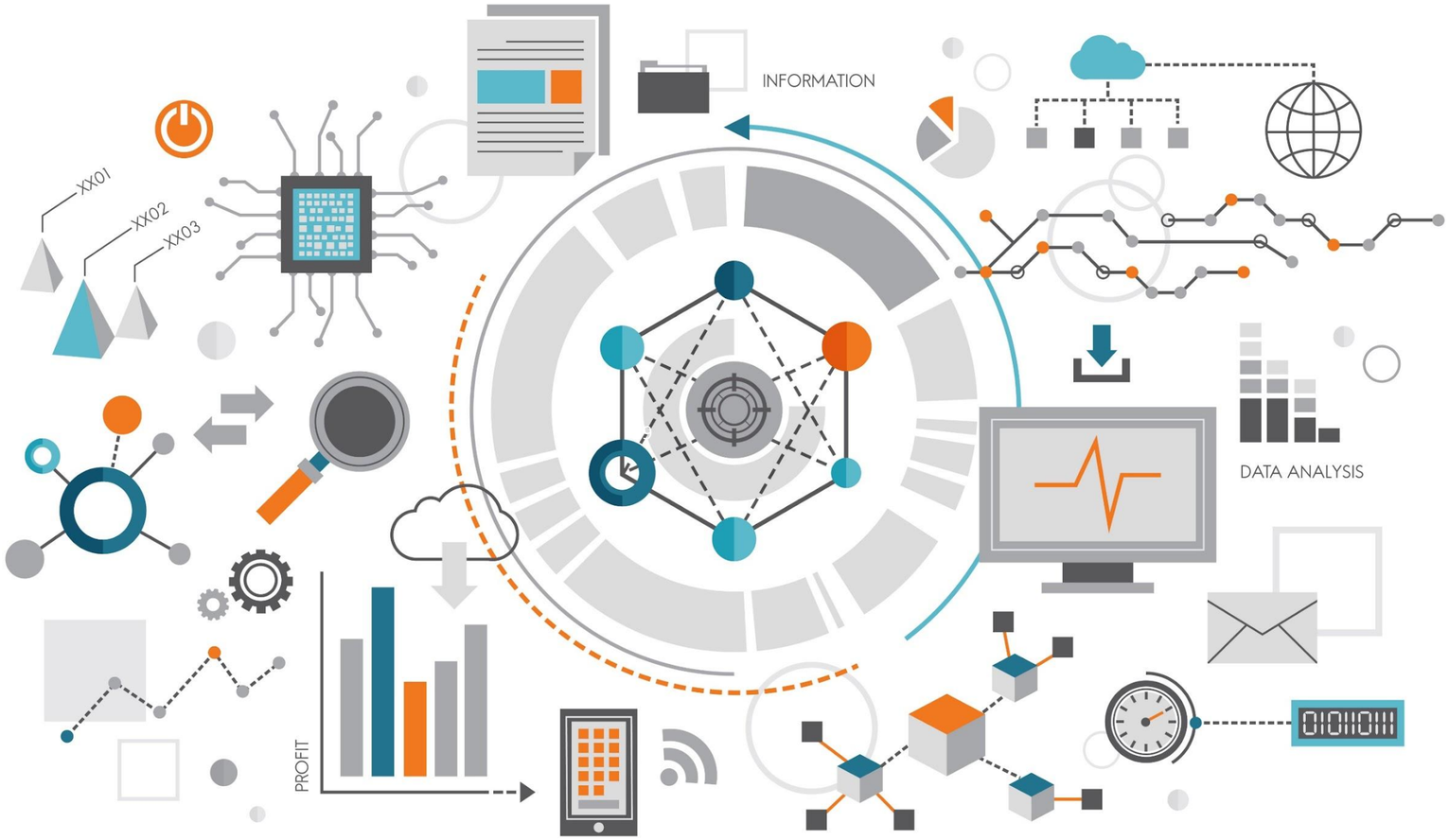
Finance:
- Detects fraudulent transactions, assesses credit risk, and manages investment portfolios.

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Module II: Data science Topics

Lesson Summary: Big Data and Data Mining



Fundamentals of Big Data and Cloud Computing

- Big data impacts various societal aspects, including business operations and sports.
- Understanding key attributes and challenges associated with big data is crucial.
- Big data drives digital transformation by necessitating fundamental changes in business approaches.
- The five characteristics of big data include value, volume, velocity, variety, and veracity.
- Cloud computing enables access to on-demand computational resources via the internet.
- Cloud computing features on-demand access, network accessibility, resource pooling, elasticity, and measured service.

Leveraging Cloud Technologies for Big Data Processing

- Cloud computing addresses scalability, collaboration, accessibility, and software maintenance challenges.
- Instant access to technologies and updated versions without installation is a benefit of cloud computing.
- Popular open-source tools for big data processing include Apache Hadoop, Hive, and Spark.
- Hadoop provides distributed storage and processing across computer clusters.
- Hive serves as a data warehouse for large datasets stored in Hadoop File System (HDFS) or Apache HBase.
- Spark is a versatile data processing engine suitable for various applications, leveraging cloud advantages for big data mining.

Data mining process

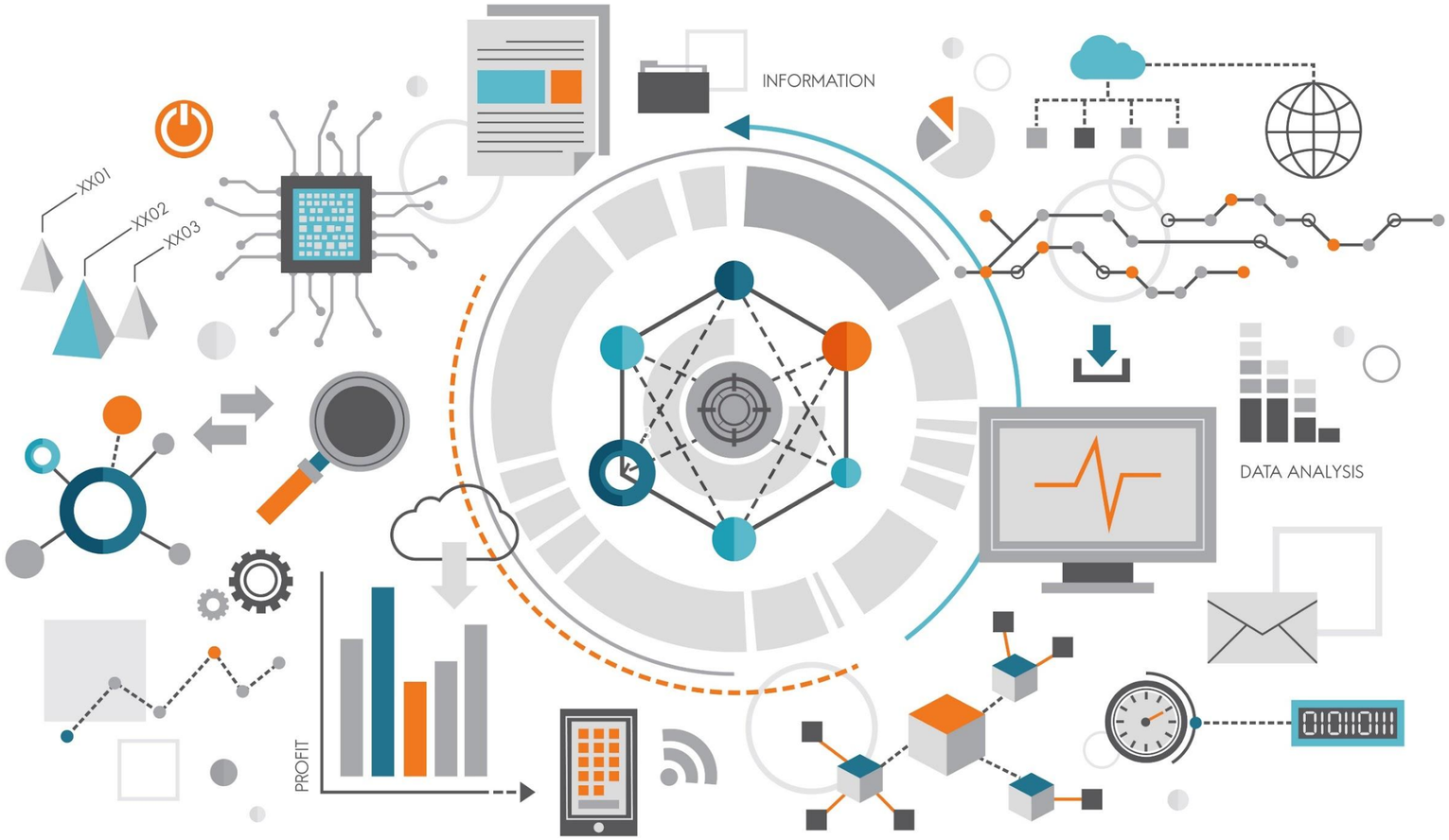
1. Goal set	→	Identify key questions
2. Select data	→	Identify data sources
3. Preprocess	→	Clean the data
4. Transform	→	Determine storage needs
5. Data mine	→	Determine methods and analyze
6. Evaluate	→	Assess outcomes, share results

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Module II: Data science Topics

Generative AI and Data Science

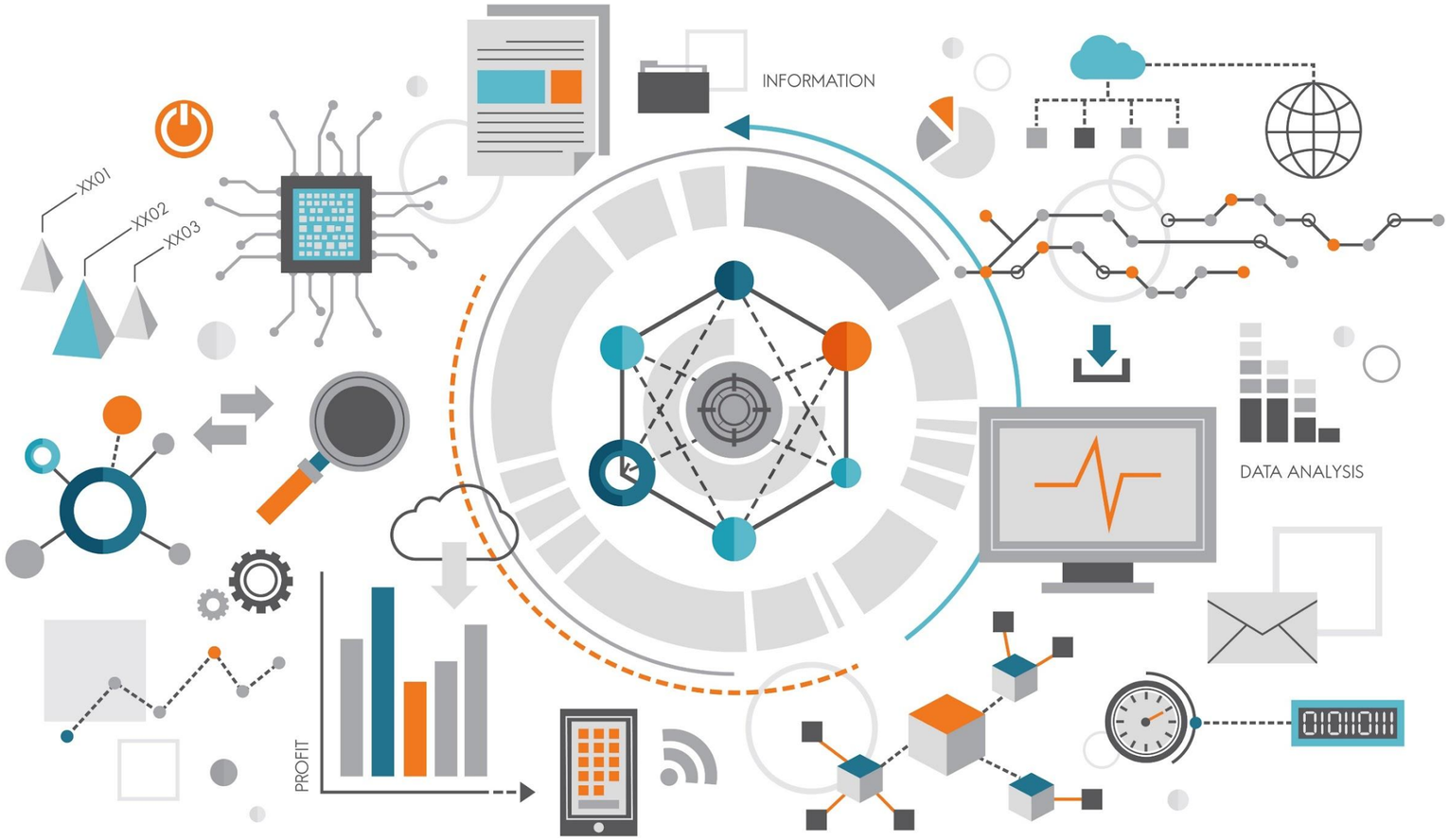


Understanding Generative AI

- Generative AI creates new data rather than analyzing existing datasets.
- Models like GANs and VAEs are foundational to generative AI.
- Generative AI mimics human creations in images, music, language, and more.
- Applications span diverse industries, from content creation to healthcare and gaming.
- Examples include GPT-3 for text generation and medical image synthesis.
- Generative AI aids in fashion design, game development, and creating artworks.

Module II: Data science Topics

Applications of NLP in Data Science



1. Chatbots

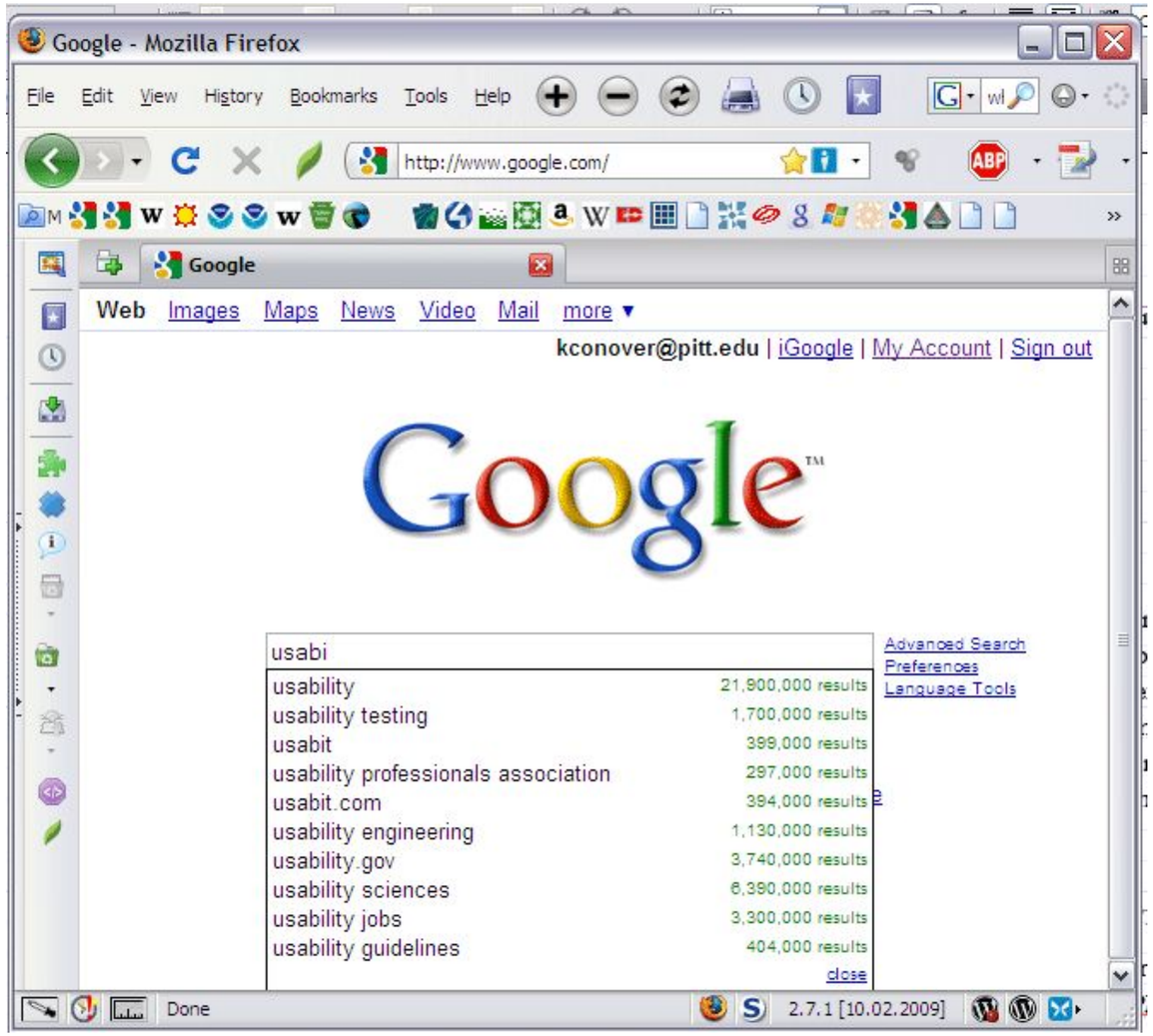
Chatbots are a form of artificial intelligence that are programmed to interact with humans in such a way that they sound like humans themselves.

2. Autocomplete in Search Engines

Have you noticed that search engines tend to guess what you are typing and automatically complete your sentences?

For example, On typing “game” in Google, you may get further suggestions for “game of thrones”, “game of life” or if you are interested in math then “game theory”.

All these suggestions are provided using autocomplete that uses Natural Language Processing to guess what you want to ask.

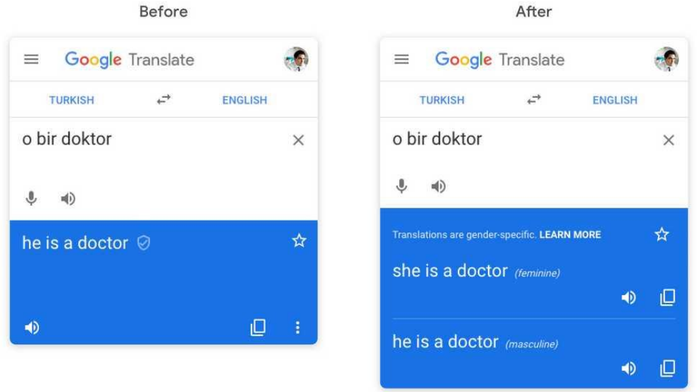


3. Voice Assistants

These days voice assistants are all the rage! Whether its Siri, Alexa, or Google Assistant,

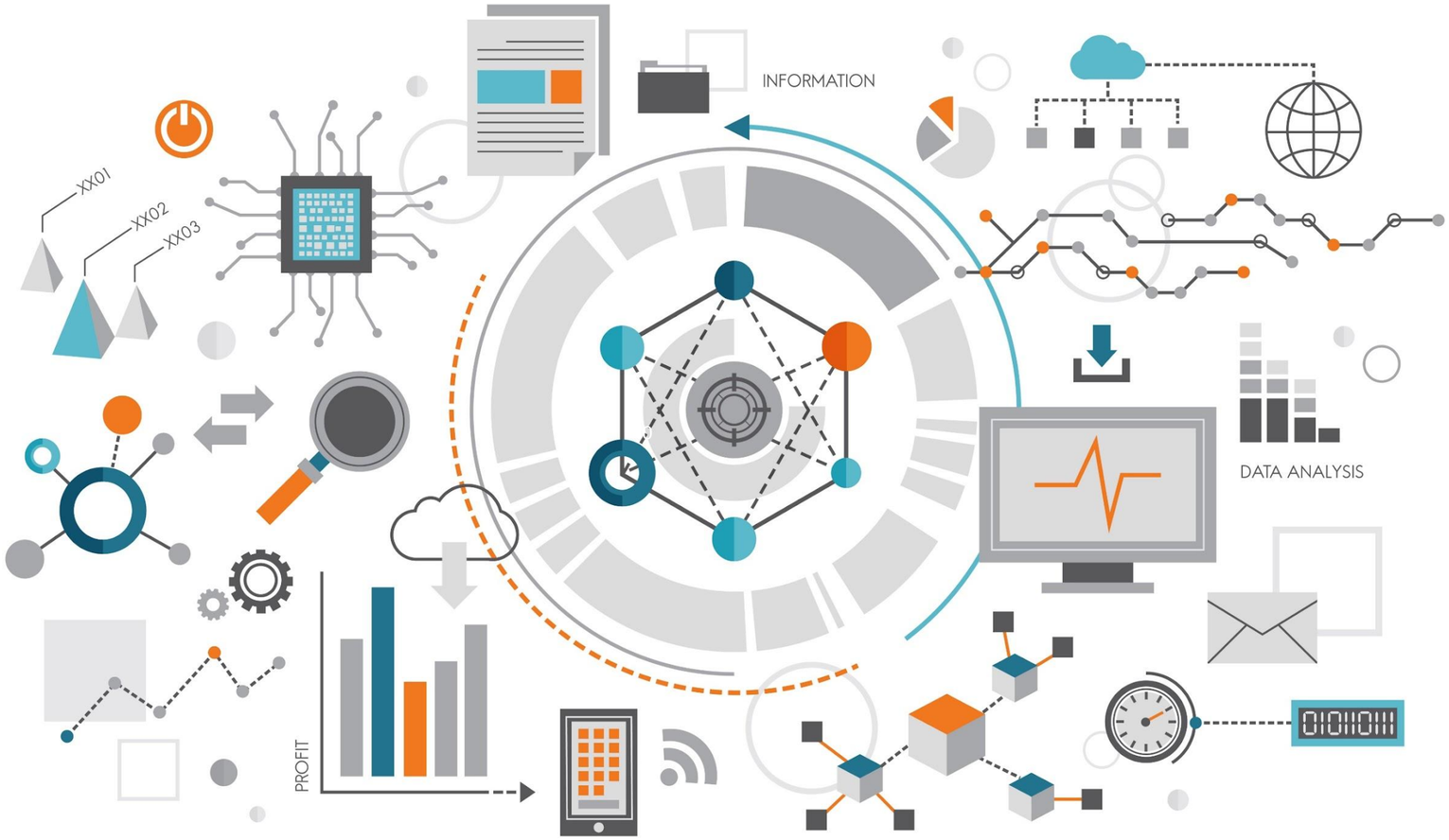
almost everyone uses one of these to make calls, place reminders, schedule meetings, set alarms, surf the internet, etc.

4. Language Translator
5. Sentiment Analysis
6. Grammar Checkers
7. Email Classification and Filtering



Module II: Data science Topics

Applications of Machine Learning



Applications of Machine Learning

- Recommender systems are significant applications of machine learning.
- Predictive analytics utilizes techniques like decision trees and Bayesian analysis.
- Understanding precision versus recall and overfitting is crucial in applying machine learning.
- Machine learning finds applications in various sectors, including fintech.
- Recommendations in fintech mirror those in platforms like Netflix or Facebook.
- Fraud detection, particularly in retail banking, is a critical area for machine learning.

Machine Learning in Fintech

- Machine learning models analyze previous transactions to identify fraudulent activities.
- Real-time decision-making in fraud detection is essential for timely intervention.
- Machine learning enhances risk management and security measures in fintech.

رواد مصر الرقمية

Enhanced Fraud Detection

Generative AI models can simulate various fraudulent scenarios to improve detection algorithms, making fraud prevention systems more robust and responsive.

10/1/2024

رواد مصر الرقمية

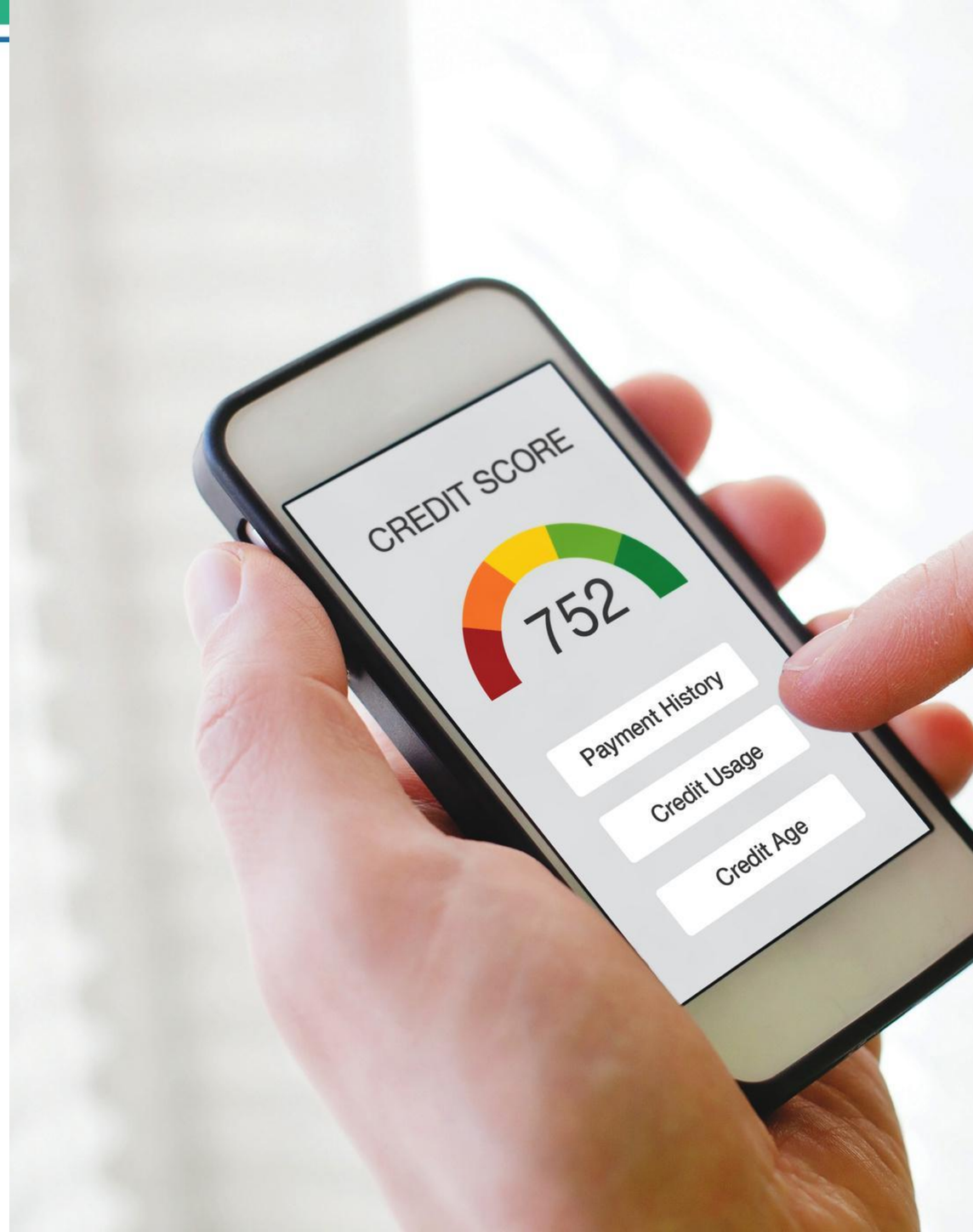
Risk Assessment And Credit Scoring:

Generative AI is reshaping risk assessment and credit scoring in the banking sector.

By creating detailed simulations of financial scenarios, generative AI tools provide deeper insights into credit risks.

This helps financial institutions improve the accuracy of their credit-scoring models, leading to smarter lending decisions.

10/1/2024



رواد مصر الرقمية

Document Processing Automation:

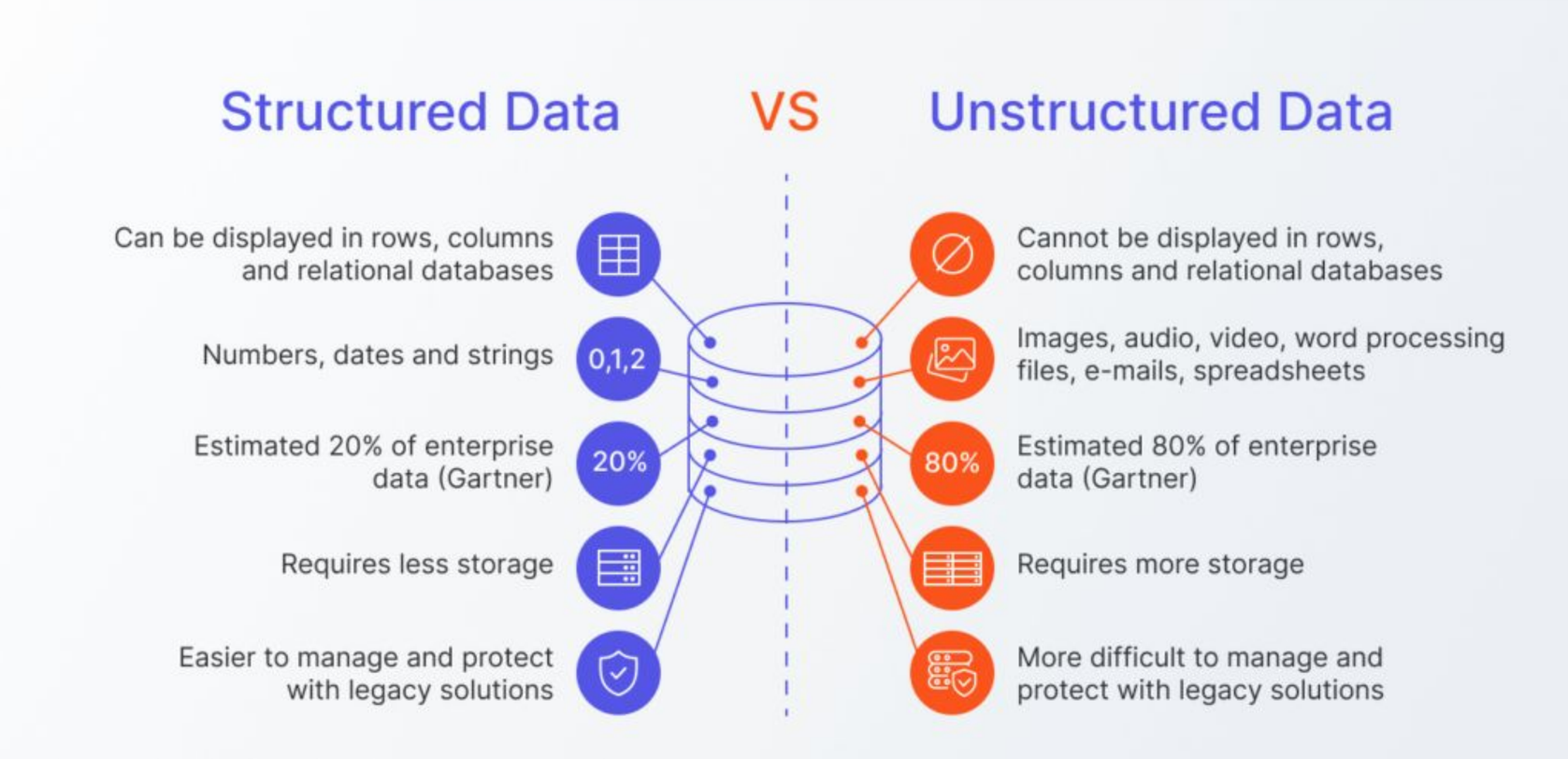
Generative AI excels in automating the generation and processing of complex banking documents, reducing errors, and increasing efficiency.

10/1/2024

Module III: Data Literacy for Data Science

Understanding Data

Understanding Data Structures

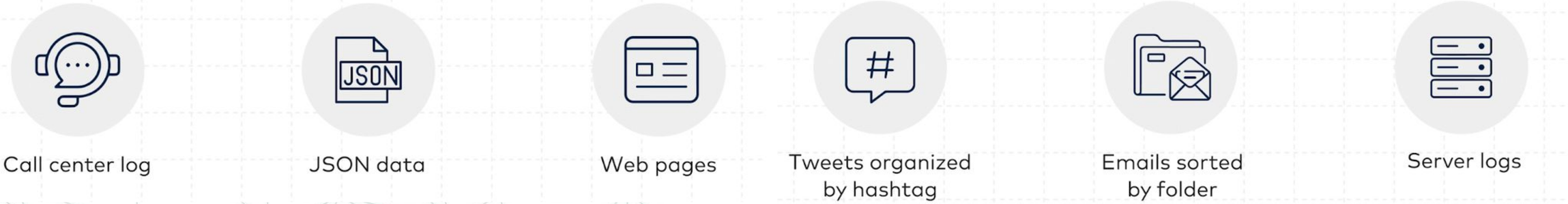


Understanding Data Structures

- **Structured Data:**
  - Well-defined structure or data model.
  - Stored in databases with rows and columns.
  - Examples: SQL databases, spreadsheets, online forms.

Understanding Data Structures

- **Semi-structured Data:**
  - Has some organizational properties but no fixed schema.
  - Uses tags and metadata for grouping and hierarchy.
  - Examples: XML, JSON, emails, TCP/IP packets, Zipped files.



Understanding Data Structures

- **Unstructured Data:**
  - Lacks identifiable structure, not stored in rows/columns.
  - Includes web pages, social media feeds, multimedia files, and documents.
  - Stored in files for a manual analysis or NoSQL databases for analysis tools.

- **Databases**
  a. Relational Databases
  b. NoSQL database (JSON or XML) (DynamoDB)

****

- Data Management (Architectures)
  - Data Warehouses >> Clean data (ETL) not Raw data
  - 2) Data Marts subset of DWH
  - Data Lakes >> Raw data (no ETL)
- ETL (Extract, Transform, Load)

Module III: Data Literacy for Data Science

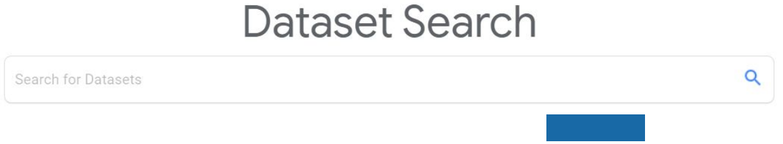
Data Sources

Ways to acquire datasets for machine learning

- Public Datasets Repositories
- APIs and Web Scraping
- Data Providers and Marketplaces
- Government and Research Institutions
- Data Collaboration
- Data Generation

Public Datasets Repositories

- UCI Machine Learning Repository:
  - Trusted source of datasets for research and experimentation.
  - Here is the link: UCI Machine Learning Repository
- Kaggle Datasets:
  - Kaggle's datasets are available on their website.
  - Here is the link: Kaggle Datasets
- GitHub:
  - GitHub hosts numerous datasets.
  - Visit GitHub and use relevant keywords to search for datasets in repositories.
- Data.gov:
  - Data.gov is the U.S. government's open data portal,
  - Here is the link: Data.gov
- Google Dataset Search:
  - Google Dataset Search is a specialized search engine for datasets.
  - Here is the link: Google Dataset Search



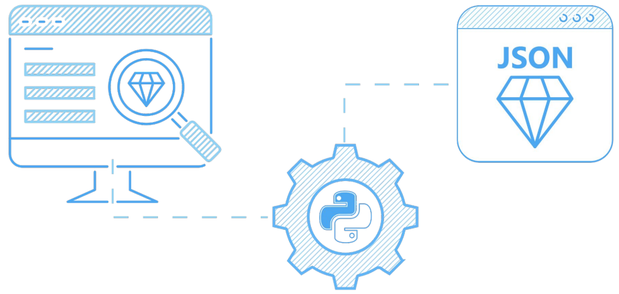
APIs and Web Scraping

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

- Accessing Data Through APIs
  - APIs allow programmable access to data from various online sources.
  - Examples of APIs include social media APIs (e.g., Twitter API), weather APIs (e.g., OpenWeatherMap API), and financial data APIs (e.g., Alpha Vantage API).
  - APIs provide structured data in real-time and are suitable for up-to-date information retrieval.

- Web Scraping Techniques
  - Web scraping involves extracting data from websites by parsing HTML or other structured formats.
  - It's useful for collecting data from websites that do not offer APIs or for custom data extraction needs.
  - Python libraries like BeautifulSoup and Scrapy are popular for web scraping.



Data Providers and Marketplaces

- AWS Data Exchange
  - Amazon Web Services (AWS) Data Exchange is a platform where you can find and subscribe to datasets.
  - It offers a diverse selection of datasets, including financial, healthcare, and geospatial data.

- Quandl
  - Quandl is a popular data provider known for its financial and economic datasets.

- DataMarket
  - DataMarket is a data marketplace that provides access to a variety of datasets.
  - It offers datasets related to demographics, economics, and social sciences.

Government and Research Institutions

- Government Agencies
  - Many government agencies provide access to datasets related to public policy, demographics, economics, and more.
  - Examples include the U.S. Census Bureau, Bureau of Labor Statistics, and Centers for Disease Control and Prevention (CDC).
- Research Institutions
  - Research institutions conduct studies and experiments that generate valuable datasets.
- Universities
  - Universities worldwide contribute to data science and research by publishing datasets covered a wide range of domains, including environmental studies, astronomy, and biology.

Data Sources Overview

- Relational Databases:
  - Organize data in structured tables (SQL Server, Oracle, MySQL).
  - Used for internal applications and business activities.
- Flatfiles and XML Datasets:
  - Flat files store data in plain text format (CSV, spreadsheet files).
  - XML files use tags to mark up data with hierarchical structures.
- APIs and Web Services:
  - Provide data access for multiple users or applications.
  - Examples: Twitter, Facebook, Stock Market APIs, Data Lookup APIs.

Advanced Data Sources

- Web Scraping:
  - Extracts data from unstructured web sources.
  - Used for product details, sales leads, forum posts, and more.
  - Popular tools: BeautifulSoup, Scrapy, and Selenium.

- Data Streams and Feeds:
  - Aggregates constant streams of data from various sources.
  - Applications in real time flights, surveillance, social media for sentiment analysis.
  - Popular technologies: Kafka, Spark, and Apache Storm.

- RSS (Really Simple Syndication) Feeds:
  - Captures updated data from online forums and news sites.
  - Streamed to user devices using feed readers for real-time updates.

Module III: Data Literacy for Data Science

Viewpoints: Working with Varied Data Sources and Types

Challenges in Working with Data Sources

- Diverse data formats require adapting data handling methods.
- SQL is crucial for data movement, structuring, and security.
- Migrating data between relational databases faces vendor changes and versioning challenges.
- Flexibility is key when working with various data sources.
- Evaluating multiple solutions is necessary for consistent and performant data movement.

Relational Databases and Alternatives

- Relational databases struggle with unstructured data like logs, XML, and JSON.
- Heavy write-intensive applications such as IoT pose challenges for relational databases.
- Alternatives like Google BigTable, Cassandra, and HBase gain popularity for specific data handling needs.
- Data engineers deal with standard formats (CSV, JSON, XML) and proprietary formats.
- Data integration spans relational databases, NoSQL databases, and Big Data repositories.

Handling Complex Data Formats

- Log data's unstructured nature demands custom parsing tools.
- XML data's resource intensity challenges efficient data handling.
- JSON's popularity stems from its simplicity and usage in RESTful APIs.
- Apache Avro gains traction for its efficient data storage capabilities.
- Import/export differences between Db2 and SQL Server
- present integration challenges.

Module III: Data Literacy for Data Science

Data Collection and Organization

Understanding Data Repositories

- A data repository encompasses organized data used for business operations and analysis.
- It includes small to large database infrastructures with one or more databases.
- Types of repositories include databases, data warehouses, and big data stores.
- Databases are designed for input, storage, retrieval, and modification of data.
  - Relational databases (RDBMS) organize data into tables with SQL for querying.
  - Non-relational databases (NoSQL) offer flexibility, speed, and scalability for big data.

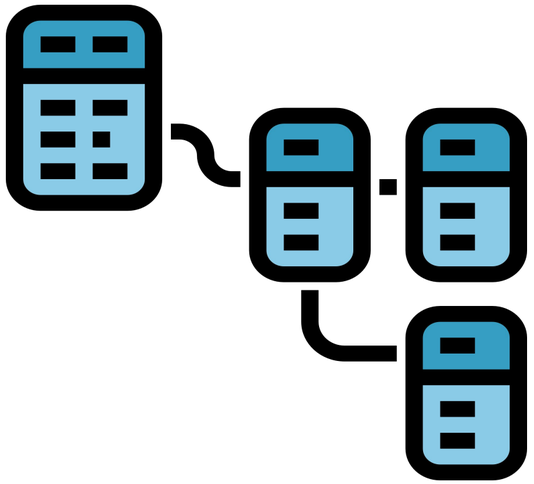
Advanced Data Repository Concepts

- Data warehouses consolidate data from various sources for analytics and BI.
- The ETL process (Extract, Transform, Load) cleans and integrates data into warehouses.
- Data Marts and Data Lakes are subsets of warehouses for specific purposes.
- Both relational and non-relational repositories are used in data warehousing.
- Big Data Stores handle distributed storage and processing of large datasets.
- Repositories enhance data isolation and reporting efficiency and serve as archives.

Module III: Data Literacy for Data Science

Relational Database Management System

Introduction to Relational Databases



- **What is a Relational Database?**
  Organized collection of data in tables.
  Tables are linked based on common data.
  Each table has rows (records) and columns (attributes).

- **Key Concepts**
  Table Example: Customer table with Company ID, Name, Address, Phone.
  Linking Tables: Relating customer and transaction tables via Customer ID.

- **Advantages of Relational Databases**
  Data Organization: Structured storage and retrieval of large volumes.
  Data Integrity: Minimized redundancy, consistent data types.
  Querying Power: Uses SQL for efficient data processing and retrieval.

Applications and Limitations

- **Use Cases of Relational Databases**
  - **OLTP:** Online Transaction Processing for fast, frequent data transactions.
  - **Data Warehouses:** Analyzing historical data for business intelligence.
  - **IoT Solutions:** Lightweight database for collecting and processing IoT data.

- **Limitations of Relational Databases**
  - **Data Type Limitations:** Not suitable for semi-structured or unstructured data.
  - **Schema Requirements:** Need identical schemas for data migration.
  - **Field Length Limitations:** Data fields have length restrictions.

- **Conclusion**
  - Despite limitations, relational databases remain essential for structured data management and common business applications.

Module III: Data Literacy for Data Science

NoSQL

Introduction to NoSQL Databases

- **What is NoSQL?**
  - Non-relational database design for flexible data storage.
  - Flexible schemas for scalability, performance, and ease of use.

- **Key Concepts**
  - Flexible Schemas: Not limited by fixed row/column structures.
  - Data Models: Four common types - Key-value store, Document-based, Column-based, and Graph-based.

- **Types of NoSQL Databases**
  - Key-value Store: Ideal for user session data and real-time recommendations.
  - Document-based: Flexible indexing for eCommerce and CRM platforms.
  - Column-based: Efficient for time-series and IoT data storage.
  - Graph-based: Visualizing and analyzing interconnected data.

Advantages and Differences

- **Advantages of NoSQL**
  - Scalability: Distributed systems for large data volumes.
  - Cost-Effective: Scale-out architecture with low-cost hardware.
  - Agility: Simplified design for better control and scalability.

- **Key Differences from Relational Databases**
  - Schema Flexibility: NoSQL allows schema-agnostic data storage.
  - Cost Considerations: Lower maintenance costs compared to RDBMS.
  - ACID Compliance: Relational databases offer transaction reliability.

- **Conclusion**
  - NoSQL databases offer scalability, cost-effectiveness, and flexibility, making them valuable for modern applications despite differences from traditional RDBMS.

Module III: Data Literacy for Data Science

Data Marts, Data Lakes, ETL, and Data Pipelines

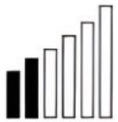
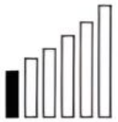
ETL = Extract > Transform > Load

Understanding Data Warehouses, Data Marts, and Data Lakes

- **Data Warehouse Overview**
  - Multi-purpose storage for analysis-ready data.
  - Single source of truth for historical and current data.

- **Data Mart Definition**
  - Sub-section of a data warehouse for specific business functions.
  - Provides isolated security and performance for targeted analytics.

- **Data Lake Concept**
  - Storage for structured, semi-structured, and unstructured data.
  - Retains raw data without predefined use cases.

	Most Important Use Group & Use-Cases	Time-to-Market Questions & Solutions	Cost Implementation & Ownership	Users (# & Types)	Data Growth Volume & Variety
Data Lake	Predictive & Advanced Analytics	 Weeks - Months	$$$$$		
Data Warehouse	Multi-Purpose Enabler of Operational & Performance Analytics	 Hours - Days	$$$$$		
Data Mart	Line of Business Specific Reporting & Analytics	 Minutes - Hours	$$$$$		

Exploring ETL Process and Data Pipelines

- **ETL Process Explanation**
  - Extract: Collecting raw data from various sources.
  - Transform: Cleaning, standardizing, and converting data for analysis.
  - Load: Transporting processed data to a data repository.

- **Types of ETL**
  - Batch Processing: Scheduled transfers in large chunks.
  - Stream Processing: Real-time data processing before loading.

- **Data Pipelines Overview**
  - Broader term including ETL for data movement.
  - Supports batch and streaming data processing for various applications.

Module III: Data Literacy for Data Science

Viewpoints: Considerations for Choice of Data Repository

Factors in Choosing a Data Repository

Factor	Why It's Important	Impact on Repository Choice
Data Type	Determines how the data is stored and queried	- Structured → Relational DB / Data Warehouse - Semi-structured → NoSQL - Unstructured → Data Lake
Data Volume	Affects performance and scalability needs	- Small/Medium → RDBMS / NoSQL - Huge volumes → Data Lake / Hadoop
Access Speed	Real-time vs batch processing requirements	- Real-time → HBase / NoSQL - Batch → Data Warehouse
Usage Purpose	Operational vs Analytical use	- Operational systems → RDBMS - BI & reporting → Data Warehouse - AI/ML → Data Lake
Tool Integration	Depends on compatibility with BI, ML, or data pipelines	- BI tools → Warehouse - ML tools → Lake - High flexibility → NoSQL
Security & Governance	Controls access, encryption, auditing	- High compliance → RDBMS / DW - Requires add-ons → Data Lake / NoSQL
Cost & Infrastructure	Budget and technical environment	- Cloud-based lakes are cheaper for large data - DW needs more investment but gives high performance

Module III: Data Literacy for Data Science

Data Integration Platforms

Data Integration Overview

- **Definition:** Gartner defines data integration as the practice, techniques, and tools for ingesting, transforming, and provisioning data across various types.
- **Usage Scenarios:** Includes data consistency, master data management, data sharing, migration, and consolidation.
- **Analytics and Data Science:** Involves accessing, transforming, merging, ensuring data quality, governance, and delivering integrated data for analytics.
- **Example:** Extracting customer data from sales, marketing, and finance systems for unified analytics.

Data Integration Capabilities

- **Modern Solutions:** Offer extensive connectors, open-source architecture, batch, and continuous processing, integration with Big Data sources, and additional functionalities.
- **Portability:** Supports cloud models, including single cloud, multi-cloud, or hybrid environments.
- **Market Overview:** Various platforms and tools available, including IBM's offerings, Talend, SAP, Oracle, Denodo, and others.
- **Evolution:** Data integration evolves with technology advancements and increasing data complexity in business decision-making.

Questions & Answers

Thank you!