**Evolution of Big Data**

◦The concept of Big Data has evolved over decades, driven by changes in data generation, storage, and processing technologies.

🔹 a) Pre-Digital Era (Before 1980s)

◦Data was generated manually: paperwork, logs, registers.

◦Analysis was limited, slow, and based on small samples.

◦Storage was expensive and capacity was minimal.

🔹 b) Digital Revolution (1980s–2000s)

◦ Introduction of computers led to structured data storage in relational databases (RDBMS).

◦Use of spreadsheets, ERP, CRM systems began.

◦Data was mostly numeric, structured, and stored in tables.

◦However, limitations in handling volume and variety emerged

🔹 c) Internet and Web 2.0 Era (2000s onwards)

◦Explosion of unstructured data: emails, blogs, videos, social media.

◦Massive data generation from IoT devices, smartphones, GPS, and sensors.

◦Traditional systems failed to process such large and diverse data efficiently.

🔹 d) Big Data Era (2010 onwards)

◦Emergence of Big Data technologies like Hadoop, Spark, NoSQL databases.

◦Focus shifted to: ◦Distributed storage (e.g., Hadoop HDFS)

◦Parallel processing (MapReduce, Spark)

◦Real-time analytics (Kafka, Flink)

◦Enabled analysis of massive, varied, and fast-moving datasets.

Era Focus Data Type Technology

Pre-1980s Manual records Paper-based Human effort

1980s–2000s Structured DBMS RDBMS SQL, OLTP

2000–2010 Web 2.0 Unstructured XML, JSON

2010 onwards Big Data S.D. , U.D. Hadoop, Spark,NoSQL

**Big Data Analytics**

**Overview of Business Intelligence:**  Business Intelligence is the talk of a new changing and growing world that can be defined as a set of concepts and methodologies to improve decision-making in business through the use of facts and fact-based systems.

**Goal of Business Intelligence** is to improve decision-making in business ideas and analysis. Business Intelligence is not just a concept; it’s a group of concepts and methodologies. Business Intelligence uses analytics and gut feelings for making decisions.

**Definition:** Business intelligence refers to a collection of mathematical models and analysis methods that utilize data to produce valuable information and insight for making imp decisions.

**Main Components of Business Intelligence System:**

**1. Data Source:** The first step is gathering and consolidating data from an array of primary and secondary sources. These sources vary in origin and format, consisting mainly of operational system data but also potentially containing unstructured documents like emails and data from external providers.

**2. Data Mart / Data Warehouse:** Through the utilization of extraction and transformation tools, also known as extract, transform, load (ETL), data is acquired from various sources and saved in databases designed specifically for business intelligence analysis. These databases, commonly known as data warehouses and data marts, serve as a centralized location for the gathered data.

**3.Data Exploration:** The third level of the pyramid offers essential resources for conducting a passive analysis in business intelligence. These resources include query and reporting systems, along with statistical methods. These techniques are referred to as passive because decision makers must first develop ideas or establish criteria for data extraction before utilizing analysis tools to uncover answers and confirm their initial theories. For example, a sales manager might observe a decrease in revenues in a particular geographic region for a specific demographic of customers. In response, she could utilize extraction and visualization tools to confirm her hypothesis and then use statistical testing to validate her findings based on the data.

**4.Data Mining**: The fourth level, known as active business intelligence methodologies, focuses on extracting valuable information and knowledge from data.delveir into various techniques such as mathematical models, pattern recognition, machine learning, and data mining.

**5.Optimization:** As you ascend the pyramid, you'll encounter optimization models that empower you to choose the most optimal course of action among various alternatives, which can often be quite extensive or even endless. These models have also been effectively incorporated in marketing and logistics.

**6.Decisions:** At last, the pinnacle of the pyramid reflects the ultimate decision made and put into action, serving as the logical end to the decision-making process. Despite the availability and effective utilization of business intelligence methodologies, the decision still lies in the hands of the decision makers, who can incorporate informal and unstructured information to fine-tune and revise the suggestions and outcomes generated by mathematical models.

**Process Used in Business Intelligence**

BI uses a set of processes, technologies, and tools to transform raw data into meaningful info and then transform information to provide knowledge. Then afterwards some beneficial insights can be extracted manually and by some software then the decision-makers can make an impact decision on the basis of insights. BI provides accurate information in the right and ethical format to the decision-makers of the organization.

Some Imp features of BI are: ◦ Fact-based decision making. ◦ 360 degrees perspective on your business. ◦ Measurement for creating KPI (Key Performance Indicators) on the basis of historic data fed into the system. ◦ Identify the benchmark and then set the benchmarks for different processes. ◦ Identify market trends and also to spot business problems that need to be identified and solved.

**Types of Users of Business Intelligence**

1. Analyst (Data Analyst or Business Analyst): They are the statistician of the company, they used BI on the basis of historical data priorly stored in the system.

2. Head or Manager of the Company: Head of the company uses Business Intelligence used to increase the profitability of their company by increasing the efficiency in their decisions on the basis of all the knowledge they discovered.

3. Small Business Owners: Can be used by a small businessman because it is quite affordable too.

4. Government Officials: In the decision-making of the government.

**Applications of Business Intelligence**

◦ In Decision Making of the company . ◦ In Data Mining while extracting knowledge.

◦ In Operational Analytics and operational management. ◦ In Predictive Analytics.

◦ In Prescriptive Analytics.. ◦ In Executive Information System (EIS)

**Big Data Analytics**

It refers to the process of examining large and complex data sets to uncover hidden patterns, correlations, market trends, customer preferences, and other useful business information.

It involves advanced analytic techniques (like statistical analysis, machine learning, data mining) applied to Big Data technologies (like Hadoop, Spark, NoSQL databases) to gain insights that were previously impossible using traditional data processing methods.

An example of big data analytics can be found in the healthcare industry, where millions of patient records, medical claims, clinical results, care management records and other data must be collected, aggregated, processed and analyzed. Traditional data analysis methods can't support this level of complexity at scale, leading to the need for big data analytics systems.

Big data analytics is also commonly used for accounting, decision-making, predictive analytics and many other purposes. The data found in big data analytics varies greatly in type, quality and accessibility, presenting significant challenges but also offering tremendous benefits.

**Characteristics of Big Data Analytics**

1. Volume

◦ Refers to the massive amount of data generated from various sources like sensors, social media, transactional systems, IoT devices, etc.

◦ Data size ranges from terabytes (TB) to petabytes (PB) and exabytes (EB).

◦ Traditional databases cannot handle this volume efficiently.

1. Velocity

◦ Describes the speed at which data is generated, collected, and processed.

◦ Data streams in at unprecedented rates, requiring real-time or near real-time processing.

◦ Examples: Social media feeds, stock market transactions, live sensor data. .

1. Variety

◦ Represents the different types and formats of data:

◦ Structured Data: Relational databases.

◦ S.S.D.: XML, JSON. ◦ Unstructured Data: Text, images, videos, social media posts.

◦ Managing and analyzing this diverse data is a key challenge.

1. Veracity

◦ Refers to the uncertainty, trustworthiness, and quality of data.

◦ Data may be incomplete, inconsistent, or noisy, making it challenging to ensure accuracy.

◦ Data cleansing and validation are crucial to ensure reliable analytics.

5. Value

● The ultimate goal of Big Data Analytics is to extract meaningful value from data.

● Insights derived must benefit the organization in terms of decision-making, efficiency, and competitiveness.

● Value creation involves turning raw data into actionable insights

**Classification of Analytics**

1. **Basic Analytics**

Basic analytics refers to the initial and most fundamental use of data for generating reports and gaining simple insights from historical data. It often involves manually prepared spreadsheets, simple summaries, and basic charts — primarily focused on describing past events. It is the starting point for many organizations on their data journey, particularly in teams or departments that are not heavily invested in advanced analytics tools or data science.

🔹 Purpose: The goal of basic analytics is to: ● Understand past performance ● Track operational metrics ● Support basic decision-making ● Share performance summaries with management or stakeholders

It answers questions like: ● How many sales did we make last month? ● What were our website visits last week? ● How many customers do we have today compared to last year? This type of analysis helps monitor business health at a high level but does not provide deep insights or predictions

🔹 Characteristics:

1. Manual or Semi-Automated Reports:

○ Reports are often generated using Excel or basic tools.

○ Data might be manually pulled from different systems and compiled.

○ There’s little to no automation — updates are often done weekly or monthly.

2. Limited Scope (Departmental):

○ Often focused on individual teams or departments (e.g., marketing, sales, HR).

○ Not integrated across the whole organization.

○ Limited to key metrics relevant to the specific function.

3. Spreadsheet-Based or Basic Tools:

○ Most common tool is Microsoft Excel.

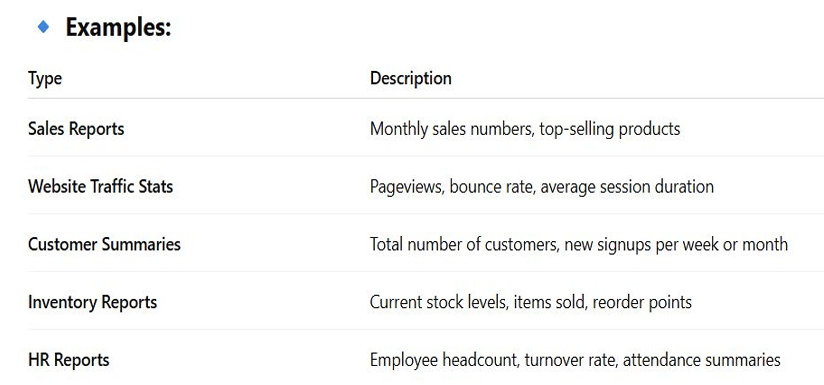
○ May also include Google Sheets, or basic database queries via SQL.

○ Visualization may include simple bar charts, line graphs, or tables.

4. Descriptive Insights Only (Not Predictive):

○ Basic analytics tells you what happened — not why it happened or what might happen next.

○ No predictive modeling, forecasting, or AI is involved



**2. Operationalized Analytics**

🔹 Definition: Analytics that are embedded into business processes and workflows to improve day-to-day operations.

🔹 Purpose: To support real-time or near-real-time decision-making within regular operations

🔹 Examples: ● Inventory dashboards updated hourly ● Automated alerts for process failures ● Real-time customer service metrics

Key Features

🔹 Automated Dashboards and Alerts: Visual reports and notifications that update in real time without manual effort, helping teams respond quickly to changes.

🔹 Integrated into Daily Tools: Analytics is built directly into systems like ERP or CRM, so users access insights within the tools they already use.

🔹 Used Regularly by Teams: Managers and staff rely on these insights daily to monitor operations and make decisions.

🔹 May Include Diagnostic Insights: Helps identify the root cause of issues, not just show what’s happening.

**3. Advanced Analytics**

🔹Definition: Advanced analytics involves the use of sophisticated techniques like machine learning, predictive modeling, and AI to analyze data, forecast future outcomes, and simulate different scenarios.

🔹 Purpose: To provide deep, forward-looking insights that support strategic planning and decision-making — going beyond what happened, to what will or should happen.

🔹 Examples: ● Churn Prediction Models: Forecast which customers are likely to leave

● Fraud Detection Systems: Identify suspicious transactions using machine learning

● AI Recommendation Engines: Suggest products based on user behavior

🔑 Key Characteristics

1. Involves data scientists using tools like Python, R, or AI platforms

○ Requires technical expertise to build models using programming languages and machine learning tools.

2. Works with both structured and unstructured data (e.g., text, images)

○ Analyzes traditional data (tables) and complex data (social media, images, etc.).

3. Predicts future trends (e.g., customer churn, sales forecasts)

○ Uses historical data to forecast outcomes, identify risks, or spot opportunities.

4. Enables prescriptive analytics, suggesting best actions to take

○ Goes a step further by recommending actions based on predictions, often using optimization or simulation.

**4. Monetized Analytics**

🔹 Definition: Monetized analytics refers to using data and insights as a source of revenue, not just for internal decision-making.

🔹 Purpose: The goal is to turn data into a marketable asset—either by selling insights, offering data-driven services, or creating new revenue-generating products.

📌 Examples: ● A telecom company selling anonymized customer behavior data to marketers.

● A retail platform offering premium analytics dashboards to partner brands.

● A SaaS company charging clients for AI-driven insights or reports.

**Characteristics of Monetized Analytics**

**1. Data products or services offered externally**

● Organizations don’t just use analytics internally; they package and sell data-driven products or services to customers, partners, or even other businesses.

● These can be in various forms such as:

○ Data sets (e.g., aggregated market data, anonymized customer insights)

○ Analytics platforms or dashboards that provide insights for a fee

○ API access to real-time analytics or predictions

● This turns data into a direct source of revenue, like a traditional product.

Example: A financial firm selling real-time stock market analytics to investors.

**2. Analytics-driven innovation (new services, pricing models)**

● Monetized analytics encourages to innovate based on insights derived from data.

● This innovation can lead to:

○ New services tailored to customer needs identified through data patterns

○ Dynamic pricing models that adjust prices based on demand or customer behavior

○ Personalized offers or product recommendations that increase sales

● The insights from analytics fuel business model innovation beyond traditional offerings.

Example: A ride-sharing app uses predictive analytics to implement surge pricing during peak hours, maximizing revenue.

**3. Competitive differentiation using data**

● Companies leverage monetized analytics as a strategic asset to stand out from competitors.

● By effectively using data to:

○ Deliver unique insights or services unavailable elsewhere

○ Enhance customer experience through data-driven personalization

○ Optimize operations and reduce costs based on analytic insights

● Data becomes a source of competitive advantage that is hard for others to replicate. Example: An e-commerce company that uses advanced customer analytics to recommend products better than competitors, increasing customer loyalty and sales.

**Importance of big data analytics**

1. Reactive - Business Intelligence

● Definition: Reacts to past events by analyzing historical data after the fact.

● Focus: Understanding what happened and why it happened based on existing data.

● Use: Producing reports, dashboards, and alerts once an event or issue occurs.

● Example: Generating monthly sales reports to analyze why sales dropped last month.

2. Reactive - Big Data Analytics

● Definition: Uses big data tools and technologies to analyze large volumes of data after events have occurred.

● Focus: Dealing with large, complex datasets retrospectively to explain outcomes or trends.

● Use: Post-event analysis using big data technologies like Hadoop or Spark.

● Example: After a marketing campaign, analyzing social media data and sales to evaluate effectiveness.

3. Proactive Analytics

● Definition: Uses data and analytics to anticipate potential future events and take action before problems or opportunities occur.

● Focus: Predicting outcomes and enabling preventive or opportunity-driven decision-making.

● Use: Forecasting, early warning systems, and scenario planning.

● Example: Predicting customer churn and engaging those customers with retention offers before they leave.

4. Proactive Big Data Analytics

● Definition: Applies big data analytics in a forward-looking way, often in real-time or near real-time, to predict and influence future events.

● Focus: Combining real-time big data streams with predictive and prescriptive analytics to guide timely actions.

● Use: Real-time fraud detection, dynamic pricing, or supply chain optimization before disruptions occur.

● Example: Monitoring IoT sensor data to predict equipment failures and schedule maintenance proactively.

**Basic terminologies in big data environment**

**In-memory analytics:**

◦ In-memory analytics is a data processing technique where all data is stored and analyzed directly in a computer’s main memory (RAM) rather than being read from traditional disk-based storage.

◦ This allows for much faster data access and analysis because RAM is significantly faster than hard drives or even SSDs.

◦ By keeping data in memory, systems can perform complex calculations, queries, and visualizations in real time or near real time. This is especially useful for businesses that need to analyze large volumes of data quickly, such as in financial services, retail, or telecommunications.

◦ Tools like SAP HANA, Power BI, Qlik, and Tableau often use in-memory analytics to allow users to explore data interactively and make rapid, data-driven decisions. Overall, in-memory analytics improves performance, speeds up decision-making, and enables organizations to gain faster insights from their data.

**In-Database Processing**

In-Database Processing is a technique where data analysis tasks—like data transformations, statistical calculations, and even machine learning—are performed directly within the database system itself, rather than moving the data out to a separate application or server for processing. Traditionally, data is extracted from a database and sent to another system (like a Python script or BI tool) for processing. This causes:

● Slow performance due to data movement.

● Higher network and storage usage.

● Security risks from copying sensitive data.

With in-database processing, the computation happens where the data lives—inside the database engine—so it's much faster, more secure, and scalable.

Key Benefits

● Faster performance: No time wasted transferring data.

● Lower latency: Real-time insights are more achievable.

● Scalability: Leverages the database’s optimized infrastructure.

● Security: Keeps sensitive data within the database boundaries.

● Efficiency: Ideal for large datasets and frequent queries.

**Symmetric Multiprocessing (SMP)**

Symmetric Multiprocessing (SMP) is a computer architecture where two or more identical processors are connected to a shared main memory and operate under a single operating system. In this setup, each processor has equal access to memory, I/O devices, and the OS, allowing them to perform independent or parallel tasks efficiently. This architecture is commonly found in modern servers, desktops, and high-performance computing systems.

**🖥 SMP Architecture Overview**

● Multiple CPUs: Usually on the same motherboard.

● Shared Memory: All processors access the same RAM.

● Single OS Instance: Manages and schedules tasks across all CPUs.

● Shared I/O Bus: All processors share I/O resources (e.g., disk, network).

**Benefits of SMP**

● Increased performance for multi-threaded or multi-process applications.

● Better resource utilization as the load is spread across CPUs.

● Scalability (to a point)—adding more processors can improve throughput.

● Simplicity in programming compared to distributed systems.

**Massively Parallel Processing (MPP)**

MPP is a high-performance computing architecture designed to handle large-scale data workloads by using many independent processors or nodes that work in parallel. Each node in an MPP system operates with its own CPU, memory, and storage—this is known as a shared-nothing architecture. MPP is ideal for big data analytics, data warehousing, and any workload requiring the processing of huge volumes of data quickly and efficiently.

**⚙ How MPP Works (Simplified Steps)**

1. Query submitted: A user or application sends a query.

2. Coordinator node: Breaks the query into smaller subtasks.

3. Worker nodes: Each processes its part of the data in parallel.

4. Results merged: Outputs from all nodes are combined and returned.

✅ Advantages

● Extreme performance for large datasets ● Efficient use of hardware through parallelism

● Elastic scaling as data volume grows.