307401
Big Data and Data Warehouses

Introduction to Big Data

What is Big Data?

- Big Data refers to datasets that are so large, diverse, and fast-changing that they exceed the capabilities of traditional data processing technologies and methods.
- These datasets require advanced tools and techniques to capture, store, manage, and analyze effectively.

Motivations for Big Data

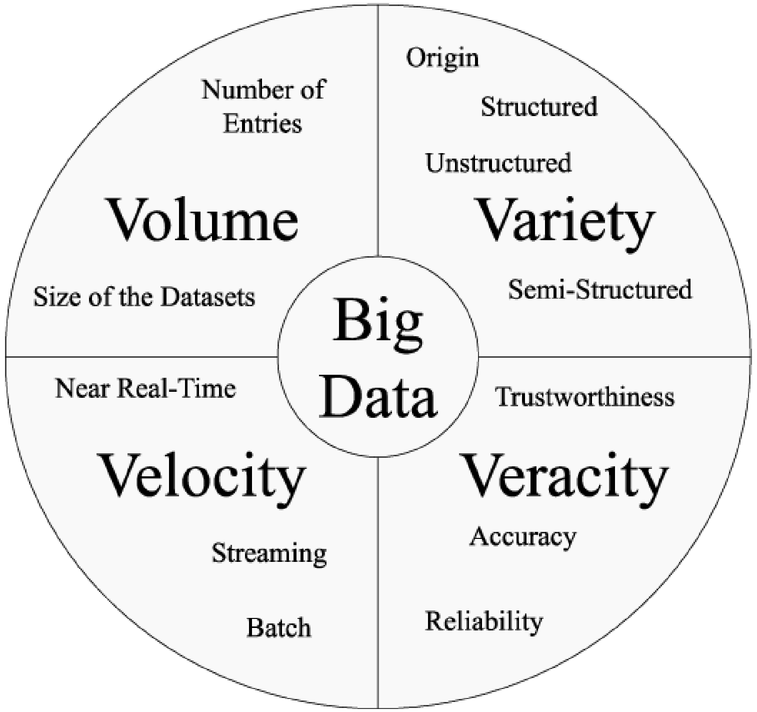
- **Global Connectivity:**
  - Internet penetration, Mobile device usage
- **Technological Advancements:**
  - Increased computing power, Improved storage solutions
- **Data Generation Sources:**
  - Proliferation of IoT devices, Social media explosion
- **Economic and Business Needs:**
  - Enhanced decision-making, Personalized marketing
- **Scientific and Research Advancements:**
  - Genome sequencing, Astronomical data
- **Regulatory and Compliance Requirements:**
  - Data retention policies, Transparency and accountability
- **User Expectations:**
  - On-demand services, Enhanced user experience

Historical Evolution

- **Pre-1970s**: Mainframe computers, batch processing, expensive storage, and limited analysis.
- **1970s**: Introduction of RDBMS and SQL, revolutionizing structured data management.
- **1980s**: Birth of data warehousing, concept of ETL, integration of data from multiple sources.
- **1990s**: Growth of OLAP, data warehouse appliances, enhanced data analysis capabilities.
- **2000s**: Emergence of Big Data, limitations of traditional systems, introduction of Hadoop and NoSQL databas
- **2010s**: Development of Spark, real-time data processing technologies, cloud computing.
- **2020s**: Integration of AI and ML, rise of data lakes, edge computing, advanced analytics.

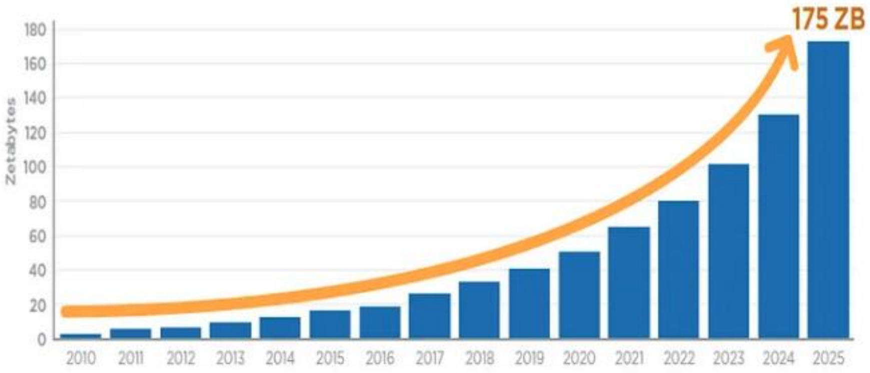
Challenges in Processing Big Data - The 5 Vs of Big Data

1. Volume: The sheer size of data generated every second.
2. Velocity: The speed at which new data is generated and processed.
3. Variety: The different types of data.
4. Veracity: The accuracy and trustworthiness of the data.
5. Value: The potential insights and benefits that can be derived from data.



Volume

- Volume refers to the amount of data generated and stored. This data is often measured in petabytes or exabytes.
- Large data volumes requires scalable storage solutions and distributed processing power to manage and analyze.
- Every day, 2.5 quintillion bytes of data are created (IBM).



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

- **Examples**:
  - Social Media: Facebook generates 4 petabytes of data per day from user interactions.
  - E-commerce: Amazon handles millions of back-end operations daily, along with queries from more than half a billion customer accounts.
  - IoT Devices: Connected devices are expected to generate 79.4 zettabytes of data by 2025.

Velocity

- Velocity refers to the speed at which data is generated, collected, and processed.
- Velocity requires real-time data processing and analytics capabilities.
- "90% of the world’s data has been generated in the last two years" (SINTEF).
- **Examples**:
  - Financial Markets: Stock exchanges generate millions of transactions per second that need to be processed in real-time for trading decisions.
  - Online Advertising: Real-time bidding platforms process terabytes of data within milliseconds to serve targeted ads.
  - Sensors and IoT: Autonomous vehicles generate and process data in real-time to make driving decisions.

Variety

- Variety refers to the different types of data: structured, semi-structured, and unstructured.
- Variety requires advanced data integration techniques and diverse analytical tools.
- **Examples**:
    - Structured Data: Traditional databases with tables and columns, such as customer information in a CRM system.
    - Semi-Structured Data: JSON, XML files, log files from servers.
    - Unstructured Data: Text from social media posts, images, videos, emails, and audio files. For example, 95 million photos and videos are shared on Instagram daily.

Veracity

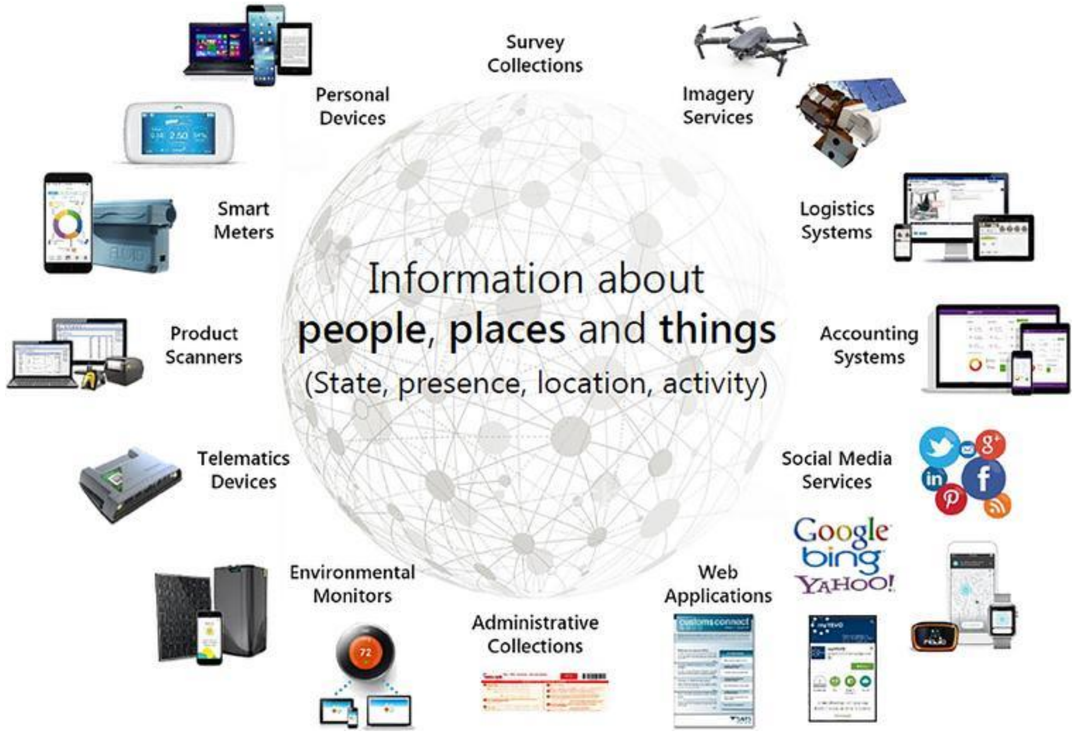
- Veracity refers to the accuracy, reliability, and trustworthiness of data.
- Veracity requires robust data cleaning, validation, and governance to ensure data accuracy and reliability.
- Challenges: Inconsistencies, biases, noise.
- **Examples**:
  - Social Media Data: User-generated content can contain noise, biases, and inaccuracies that need to be filtered.
  - Product ratings on an online marketplace might be skewed by fake reviews posted by automated bots or competitors.
  - A temperature sensor in a smart home system might report incorrect values due to dust accumulation or hardware faults.
  - Web server logs might contain duplicated or corrupted entries due to server crashes or improper shutdowns, leading to inaccurate traffic analysis.

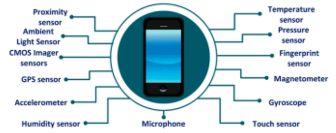
Value

- Value refers to the insights and benefits derived from analyzing Big Data.
- The ultimate goal of Big Data is to extract meaningful insights that drive better decision-making and innovation.
- **Examples**:
  - Retail: Walmart uses Big Data analytics to optimize its supply chain, reducing costs and improving efficiency.
  - Healthcare: Predictive analytics helps in early diagnosis of diseases, personalized treatment plans, and improved patient outcomes.
  - Marketing: Companies like Netflix use data analytics to recommend personalized content, enhancing customer experience and increasing retention.

Sources of Big Data

- **Social Media**:
  - Platforms like Facebook, Twitter, Instagram.
  - Examples: Posts, tweets, photos, videos, comments.
- **Sensors and IoT Devices**:
  - Smart homes, wearables, industrial sensors.
  - Examples: Temperature readings, fitness data, machinery monitoring.
- **Business Applications**:
  - Financial, Retail, Health, Government systems and application.
  - Examples: Banks Transactions, Purchase history, account activity.
- **Data Logs and Machine Data**:
  - Generated by servers, applications, and network devices.
  - Examples: Web server logs, application logs, network traffic logs.
- **Public Data**:
  - Government reports, weather data, research publications.
  - Examples: Census data, meteorological data, scientific datasets.





Importance and Applications of Big Data

Real-Time Stock Market Analysis

- **Goldman Sachs** uses big data to perform real-time stock market analysis, enabling high-frequency trading and informed investment decisions.
- **Data Volume:** Goldman Sachs processes terabytes of data daily, including market transactions, historical stock prices, and economic indicators.
- **Data Velocity:** The stock market generates data at an extremely high velocity, with thousands of trades executed every second that need to be analyzed instantly to capitalize on market opportunities.

- **Conventional Limitation:** Traditional financial analysis systems cannot keep up with the high frequency and vast amount of data, leading to delays in trading decisions and missed opportunities.
- **Big Data Parallelism:** Big data platforms utilize distributed computing and parallel processing to analyze real-time data streams from the stock market, enabling Goldman Sachs to execute trades within milliseconds and make timely investment decisions based on comprehensive and up-to-date information.

Customer Behavior Analysis in Retail

- **Walmart** uses big data to analyze customer behavior, optimizing inventory management and personalized marketing.
- **Data Volume:** Walmart processes over 2.5 petabytes of data every hour from transactions, customer interactions, and inventory records.
- **Data Velocity:** With millions of transactions occurring daily across thousands of stores worldwide, Walmart generates data in real-time that needs to be processed instantly to make timely decisions.

- **Conventional Limitation:** Traditional retail analytics systems cannot handle the massive volume and real-time processing requirements, leading to delayed insights and suboptimal inventory management.
- **Big Data Parallelism:** Big data systems allow Walmart to process vast amounts of data in parallel across distributed computing resources, enabling real-time analytics to forecast demand, optimize stock levels, and personalize marketing efforts effectively.

Personalized Medicine and Genomic Analysis:

- **The Mayo Clinic** uses big data for personalized medicine by analyzing genomic data to tailor treatments to individual patients.
- **Data Volume:** The Mayo Clinic processes petabytes of genomic data, which includes sequencing information from millions of patients' DNA.
- **Data Variety:** Genomic data is highly diverse, including information on gene sequences, protein interactions, and patient medical histories.

- **Conventional Limitation:** Traditional computing methods struggle to handle the vast, complex, and varied datasets required for genomic analysis, often resulting in prolonged processing times and less precise treatment recommendations.
- **Big Data Parallelism:** Big data platforms use parallel processing and advanced algorithms to analyze vast genomic datasets quickly and accurately, enabling the Mayo Clinic to develop personalized treatment plans based on an individual's genetic makeup, thereby improving treatment efficacy and patient outcomes.

Predictive Maintenance in Manufacturing

- **General Electric (GE)** uses big data for predictive maintenance to improve equipment reliability and reduce downtime in its manufacturing plants.
- **Data Volume:** GE processes terabytes of data daily from sensors embedded in machinery, capturing information such as temperature, vibration, and pressure.
- **Data Velocity:** The sensors generate data in real-time, with thousands of data points per second that need to be analyzed continuously to detect potential equipment failures.

- **Conventional Limitation:** Traditional maintenance systems rely on scheduled inspections and reactive repairs, which can miss early signs of equipment degradation and lead to unplanned downtime.
- **Big Data Parallelism:** Big data platforms enable GE to analyze real-time sensor data using machine learning algorithms and distributed computing. This allows for the early detection of anomalies and prediction of equipment failures before they occur, facilitating proactive maintenance and minimizing downtime.

Fleet Management in Transportation

- **UPS** uses big data for fleet management to optimize delivery routes, reduce fuel consumption, and improve delivery times. It can use data to predict demand, weather, traffic and improve efficiency and customer satisfaction.
- **Data Volume:** UPS processes over 16 petabytes of data annually from GPS devices, vehicle sensors, customer orders, and traffic information.
- **Data Velocity:** The fleet generates data in real-time, with continuous updates on vehicle locations, speed, engine performance, and traffic conditions that need to be analyzed instantly.

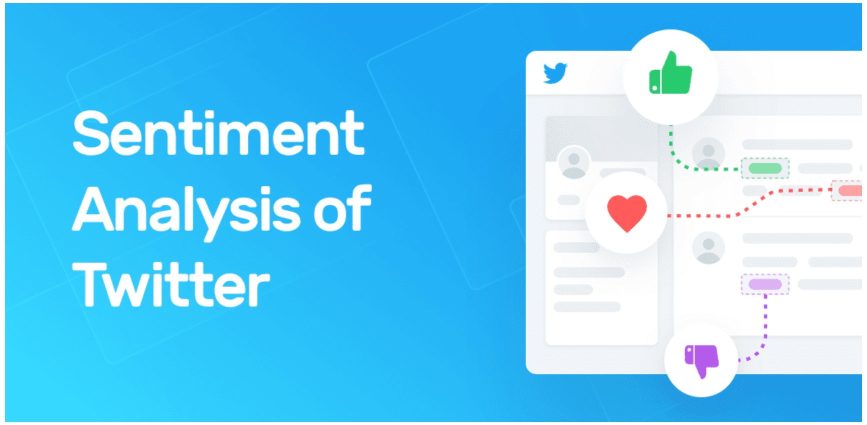
- **Conventional Limitation:** Traditional fleet management systems cannot handle the massive volume of data and lack the real-time processing capabilities needed to optimize routes dynamically and respond to changing conditions on the fly.
- **Big Data Parallelism:** Big data platforms allow UPS to analyze real-time data from thousands of vehicles simultaneously, using advanced algorithms to optimize delivery routes, predict maintenance needs, and adjust for traffic conditions in real-time. This leads to reduced fuel consumption, lower operational costs, and improved delivery times.

Smart Grid Management in the Energy Sector

- **Duke Energy** uses big data for smart grid management to enhance energy distribution efficiency and reliability.
- **Data Volume:** Duke Energy processes terabytes of data daily from smart meters, grid sensors, weather stations, and consumer usage patterns.
- **Data Velocity:** Smart meters and grid sensors generate data in real-time, with updates every few seconds, providing continuous streams of information on energy consumption, grid status, and environmental conditions.

- **Conventional Limitation:** Traditional energy management systems struggle to process and analyze the vast amounts of real-time data needed to optimize grid performance and respond to issues quickly, leading to inefficiencies and slower response times to outages or demand spikes.
- **Big Data Parallelism:** Big data platforms enable Duke Energy to analyze real-time data from millions of smart meters and sensors simultaneously, using parallel processing and advanced analytics to optimize energy distribution, predict and prevent outages, and balance supply with demand dynamically. This results in a more efficient, reliable, and responsive energy grid.

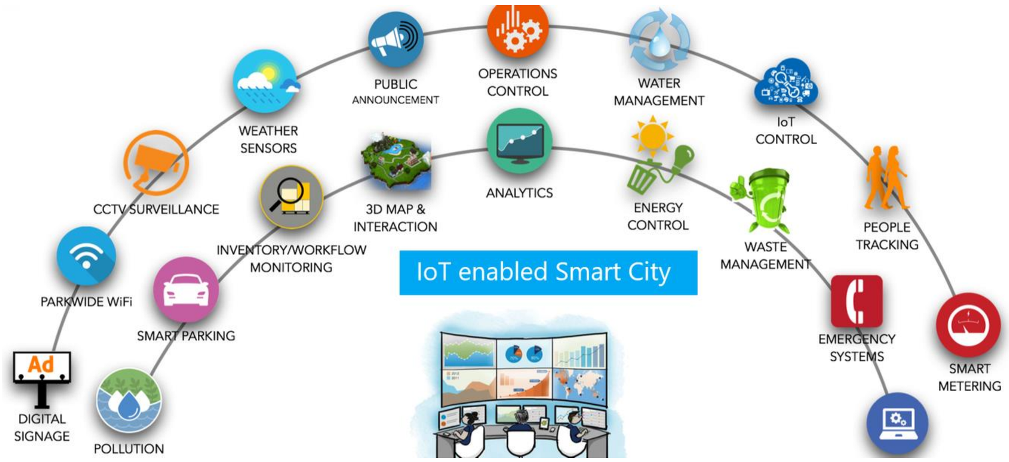
Sentiment Analysis and Trend Prediction in Social Media



- **Twitter** uses big data to perform sentiment analysis and predict trends, helping businesses and advertisers understand public opinion and market trends.
- **Data Volume:** Twitter processes over 500 million tweets daily, amounting to terabytes of data that include text, images, videos, and user interactions.
- **Data Velocity:** The platform generates data at an extremely high velocity, with thousands of tweets being posted every second, requiring real-time analysis to capture trending topics and shifts in public sentiment.
- **Conventional Limitation:** Traditional text analysis systems cannot keep up with the high speed and sheer volume of social media data, leading to delayed insights and missed opportunities to capitalize on emerging trends.
- **Big Data Parallelism:** Big data platforms allow Twitter to analyze vast amounts of real-time data using distributed computing and natural language processing (NLP) algorithms. This enables the platform to provide instant insights into public sentiment, identify trending topics as they emerge, and deliver targeted content to users.

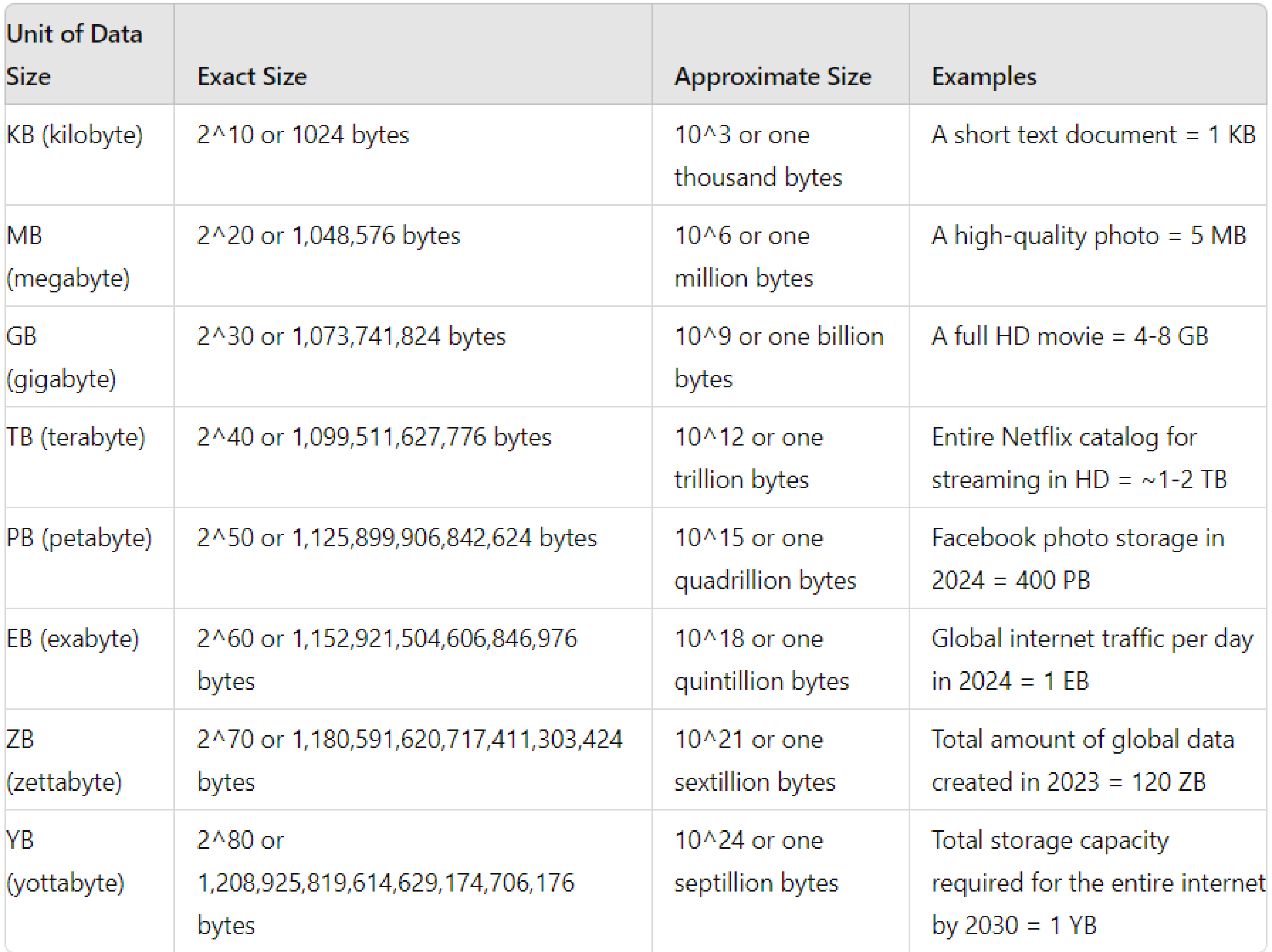
Smart City Infrastructure Management using IoT

- **Barcelona** uses big data and IoT for smart city infrastructure management to improve urban living and resource efficiency.
- **Data Volume:** Barcelona collects petabytes of data annually from thousands of IoT sensors deployed across the city, monitoring everything from traffic flow and air quality to energy usage and waste management.
- **Data Velocity:** IoT sensors generate data continuously in real-time, providing constant updates on various parameters such as traffic congestion, pollution levels, and energy consumption.



- **Conventional Limitation:** Traditional urban management systems cannot handle the vast amounts of high-velocity data from numerous sensors, leading to delayed responses and inefficient resource utilization.
- **Big Data Parallelism:** Big data platforms enable Barcelona to process and analyze real-time data streams from IoT sensors using distributed computing and advanced analytics. This allows for dynamic traffic management, real-time monitoring of environmental conditions, efficient energy distribution, and timely waste collection, significantly enhancing the city's operational efficiency and quality of life for its residents.

Data Sizes and Data Units



Unit of Data Size	Exact Size	Approximate Size	Examples
KB (kilobyte)	2^10 or 1024 bytes	10^3 or one thousand bytes	A short text document = 1 KB
MB (megabyte)	2^20 or 1,048,576 bytes	10^6 or one million bytes	A high-quality photo = 5 MB
GB (gigabyte)	2^30 or 1,073,741,824 bytes	10^9 or one billion bytes	A full HD movie = 4-8 GB
TB (terabyte)	2^40 or 1,099,511,627,776 bytes	10^12 or one trillion bytes	Entire Netflix catalog for streaming in HD = ~1-2 TB
PB (petabyte)	2^50 or 1,125,899,906,842,624 bytes	10^15 or one quadrillion bytes	Facebook photo storage in 2024 = 400 PB
EB (exabyte)	2^60 or 1,152,921,504,606,846,976 bytes	10^18 or one quintillion bytes	Global internet traffic per day in 2024 = 1 EB
ZB (zettabyte)	2^70 or 1,180,591,620,717,411,303,424 bytes	10^21 or one sextillion bytes	Total amount of global data created in 2023 = 120 ZB
YB (yottabyte)	2^80 or 1,208,925,819,614,629,174,706,176 bytes	10^24 or one septillion bytes	Total storage capacity required for the entire internet by 2030 = 1 YB

This table that maps **hardware specifications (RAM & CPU)** to **data size limits** and the **best-suited technologies** for processing

Hardware Specs	Data Size Range	Recommended Technology	Notes
Low-end Laptop (4GB RAM, Dual-Core CPU)	**< 100MB**	**Excel, Pandas (small data)**	Suitable for small datasets like CSVs, spreadsheets.
Mid-range Laptop/Desktop (8GB RAM, Quad-Core CPU)	**100MB – 1GB**	**Pandas, SQLite, Polars**	Pandas works but may slow down, Polars is more efficient.
High-end Desktop (16GB RAM, 6-8 Core CPU)	**1GB – 10GB**	**Pandas (optimized), Dask, PostgreSQL**	Dask allows parallel processing; SQL databases handle structured data well.
Workstation (32GB RAM, 8-12 Core CPU)	**10GB – 100GB**	**Dask, Polars, PySpark (local mode), PostgreSQL**	Dask and Polars can handle larger-than-memory datasets efficiently.
High-performance Workstation (64GB+ RAM, 12+ Core CPU)	**100GB – 500GB**	**PySpark (local cluster), DuckDB, BigQuery (cloud SQL)**	Need distributed computing for efficiency.
Small Hadoop Cluster (3+ Nodes, 128GB+ RAM)	**500GB – 1TB**	**Apache Spark, Hadoop (MapReduce, Hive), Databricks**	For large-scale batch processing and analytics.
Enterprise-Scale Cluster (10+ Nodes, 1TB+ RAM)	**1TB+**	**Apache Spark, Hadoop, Google BigQuery, Azure Databricks**	Cloud-based or distributed computing required for scalability.