04 Big Data Intro

Big Data Introduction

1. Definition of Big Data

Big Data refers to collections of datasets that possess one or more of the following characteristics:

- **Huge volume** of data
- **High velocity** of data flow
- **Diverse variety** of data types
- Any combination of the above

These datasets exceed the capacity of traditional storage systems (RDBMS), computing resources, and algorithms to store, process, analyze, and understand data in a cost-effective manner.

Contextual Understanding of "Big"

The concept of "big" in big data is **relative** and depends on the processing capabilities of the system:

Examples:

- A 15GB 4K video file:
  - **Big Data** for a mobile device (limited storage/processing)
  - **Not Big Data** for a high-end laptop
- 100GB 3D graphics rendering:
  - **Big Data** for laptop/desktop machines
  - **Not Big Data** for high-end servers

> **Formula:** Big Data = Huge Volume + High Velocity + Wide Variety

---

2. Why Store and Process Big Data?

Historical Context

- Data storage became viable as **storage devices became cheaper**
- Organizations began collecting and storing more data for future analysis

Business Rationale

- **Historical data** enables learning from past patterns
- **Decision-making** becomes more informed and data-driven
- **Adaptability** to market changes and trends
- **Swift reactions** to future opportunities and challenges

---

3. Applications of Big Data

```
## **WHY SHOULD WE STORE AND PROCESS BIG DATA?**
Ans. We started storing data as storage devices became cheaper.
    - We need historical data to learn, understand, and make decisions for adapting to
changes and reacting swiftly in the future.
- **Applications**
1. _Social Network Analysis (SNA):_ Social network data is rich in content and
relationships that are quite valuable to many third-party business
entities. They use such data for different purposes. For instance,
            - Understanding and targeting customers for marketing.
```

```
            - Detecting online communities, predicting market trends, etc.
2. _E-commerce:_ recommender system (people who like this product may also like another
product in online shopping, friend suggestions
on Facebook), sentiment analysis, marketing, etc.
3. _Banking and Finance:_ stock market analysis, risk/fraud management ,etc.
4. _Transportation:_ logistic optimization, real-time traffic flow optimization, etc.
5. _Healthcare:_ medical record analysis, genome analysis, patient monitoring, etc.
6. _Telecommunication:_ threat detection, violence prediction, etc.
7. Entertainment: animation, 3D video rendering, etc.
8. Forecasting events like disease spread, natural disaster and take proactive measures.
9. Optimizing system (hardware/software) performance.
10. Improving performance in sports.
```

3.1 Social Network Analysis (SNA)

- **Customer targeting** for marketing campaigns
- **Online community detection**
- **Market trend prediction**
- Leveraging rich content and relationship data

3.2 E-commerce

- **Recommender systems** ("Customers who bought this also bought...")
- **Friend suggestions** on social platforms
- **Sentiment analysis** for product reviews
- **Targeted marketing** strategies

3.3 Banking and Finance

- **Stock market analysis** and prediction
- **Risk management** and assessment
- **Fraud detection** and prevention
- **Credit scoring** and loan approvals

3.4 Transportation and Logistics

- **Route optimization** for delivery services
- **Real-time traffic flow** management
- **Supply chain optimization**
- **Fleet management** systems

3.5 Healthcare

- **Medical record analysis** for treatment insights
- **Genome analysis** for personalized medicine
- **Patient monitoring** and predictive healthcare
- **Drug discovery** and development

3.6 Telecommunications

- **Threat detection** and cybersecurity
- **Network optimization**
- **Customer behavior analysis**
- **Service quality monitoring**

3.7 Entertainment and Media

- **3D animation** and rendering

- **Video processing** and streaming optimization
- **Content recommendation** systems
- **Audience analytics**

3.8 Predictive Analytics

- **Disease spread forecasting**
- **Natural disaster prediction**
- **Proactive risk management**
- **System performance optimization**

3.9 Sports Analytics

- **Performance optimization** for athletes
- **Game strategy development**
- **Injury prevention** analysis
- **Fan engagement** insights

---

4. Data Size Reference

Scale Reference: 1 Exabyte can store approximately **11 million 4K resolution movies**

Data Size Units (Ascending Order)

In bytes	Unit	Binary	In bytes	Unit	Binary
1 Bit	0 or 1	-	1024 Kryat byte	1 Amos byte	2^{150} bytes
1 Byte	8 bits	2^{0} bytes	1024 Amos byte	1 Pectrol byte	2^{160} bytes
1024 Bytes	1 Kilo byte	2^{10} bytes	1024 Pectrol byte	1 Bolger byte	2^{170} bytes
1024 Kilo byte	1 Mega byte	2^{20} bytes	1024 Bolger byte	1 Sambo byte	2^{180} bytes
1024 Mega byte	1 Giga byte	2^{30} bytes	1024 Sambo byte	1 Quesa byte	2^{190} bytes
1024 Giga byte	1 Tera byte	2^{40} bytes	1024 Quesa byte	1 Kinsa byte	2^{200} bytes
1024 Tera byte	1 Peta byte	2^{50} bytes	1024 Kinsa byte	1 Ruther byte	2^{210} bytes
1024 Peta byte	1 Exa byte	2^{60} bytes	1024 Ruther byte	1 Dubni byte	2^{220} bytes
1024 Exa byte	1 Zetta byte	2^{70} bytes	1024 Dubni byte	1 Seaborg byte	2^{230} bytes
1024 Zetta byte	1 Yotta byte	2^{80} bytes	1024 Seaborg byte	1 Bohr byte	2^{240} bytes
1024 Yotta byte	1 Bronto byte	2^{90} bytes	1024 Bohr byte	1 Hassiu byte	2^{250} bytes
1024 Bronto byte	1 GeoP byte	2^{100} bytes	1024 Hassiu byte	1 Meitner byte	2^{260} bytes
1024 GeoP byte	1 Sagan byte	2^{110} bytes	1024 Meitner byte	1 Darmstad byte	2^{270} bytes
1024 Sagan byte	1 Pija byte	2^{120} bytes	1024 Darmstad byte	1 Roent byte	2^{280} bytes
1024 Pija byte	1 Alpha byte	2^{130} bytes	1024 Roent byte	1 Coper byte	2^{290} bytes
1024 Alpha byte	1 Kryat byte	2^{140} bytes			

---

5. The Seven V's of Big Data

5.1 Volume

Definition: The amount of data generated and stored

Characteristics:

- More data leads to more accurate decisions
- Processing challenges increase with volume
- I/O bottlenecks become critical limiting factors

Example Challenge: A 1TB dataset on HDD with 32 CPU cores results in most cores remaining idle due to slow HDD I/O rates, actually increasing processing time.

5.2 Velocity

Definition: The speed at which data is generated and processed

Characteristics:

- Streaming data requires **real-time processing**
- Data must be processed before persistent storage
- "The faster, the more revenue" - time-sensitive opportunities

Limitations of Traditional Systems:

- RDBMS requires indexing before data access
- Not suitable for real-time processing requirements

Use Cases:

- Fraud detection in banking
- Threat detection in telecommunications
- Real-time recommender systems
- Live social media analytics

5.3 Variety

Definition: Different types and formats of data

Evolution of Data Types:

- **Traditional:** Documents, logs, transaction files
- **Modern:** Audio, video, images, 3D models, spatial data, temporal data

Data Categories:

Structured Data

- **Format:** Fixed schema, organized in tables
- **Examples:** Banking records, financial transactions
- **Storage:** RDBMS (Relational Database Management Systems)
- **Growth Pattern:** Linear growth

Semi-Structured Data

- **Format:** Partially organized, self-describing
- **Examples:** JSON, XML, YAML, HTML, log files, emails
- **Characteristics:** Has some organizational structure but not rigid schema

Unstructured Data

- **Format:** No predefined structure
- **Examples:** Audio files, video files, text documents, images
- **Growth Pattern:** Exponential growth due to Internet and IoT applications

5.4 Value

Definition: The potential insight and business value extractable from data

Key Principle: "Big data beats better algorithms"

Challenges:

- Extracting **relevant information** from massive datasets
- Information extracted may be proportionally small
- Requires sophisticated **analytics algorithms**
- Questions the **cost-benefit ratio** of data processing

Requirement: Advanced analytics to improve decision-making processes

5.5 Veracity

Definition: The accuracy, authenticity, and trustworthiness of data

Challenges:

- **Public sources** (social networks) may contain inaccurate information
- **User authenticity** is not guaranteed on the Internet
- **Data quality** varies significantly across sources
- **Verification processes** are complex and resource-intensive

Impact: Affects reliability of analysis and decision-making

5.6 Variability

Definition: The dynamic and evolving behavior of data generation sources

Characteristics:

- Data patterns change over time
- Sources may modify their data generation behavior
- Requires **adaptive processing** systems
- **Seasonal variations** in data patterns

5.7 Volatility

Definition: The lifespan and relevance period of data

Key Questions:

- How long is data valid for analysis?
- When does data become irrelevant?
- How long should data be stored?
- What is the **data retention policy**?

Challenge: Determining the point where data loses relevance to current analysis

5.8 Complexity (Bonus V)

Definition: The interconnectedness and relationships between data variables

Characteristics:

- **Unstable number** of variables
- **Complex relationships** between data points
- **Multi-dimensional** data analysis requirements
- **Network effects** and data dependencies

---

6. Practical Example: Insurance Agency Case Study

Scenario

An insurance agency uses big data to decide whether to display insurance advertisements to users booking travel tickets.

Data Sources

- **Social media** activity and profiles
- **Bank transactions** and financial history
- **Web browsing** patterns and behavior
- **Competitor pricing** information

Big Data Characteristics in Action

Volume

- **Historical customer data** accumulated over years
- **Transaction records** from multiple sources
- **Large-scale data storage** requirements

Variety

- **Social media data** (posts, likes, shares, comments)
- **Financial data** (structured transaction records)
- **Behavioral data** (web browsing patterns)
- **Market data** (competitor pricing)

Velocity

- **Real-time click streaming** data
- **Current user activity** monitoring
- **Immediate decision-making** for ad display
- **Live competitor price tracking**

Business Outcome

- **Competitive pricing** strategies
- **Targeted advertising** based on user profiles
- **Real-time personalization** of insurance offers
- **Improved customer acquisition** rates

---

7. Key Takeaways

1. **Big Data is contextual** - what's "big" depends on processing capabilities
2. **Multiple characteristics** - rarely just one "V" but combinations
3. **Business value** - focus on extracting actionable insights
4. **Technology evolution** - traditional systems (RDBMS) have limitations
5. **Real-time processing** - increasingly critical for competitive advantage
6. **Data quality matters** - veracity affects decision reliability
7. **Storage strategies** - volatility determines retention policies
8. **Diverse applications** - spans across all industries and sectors