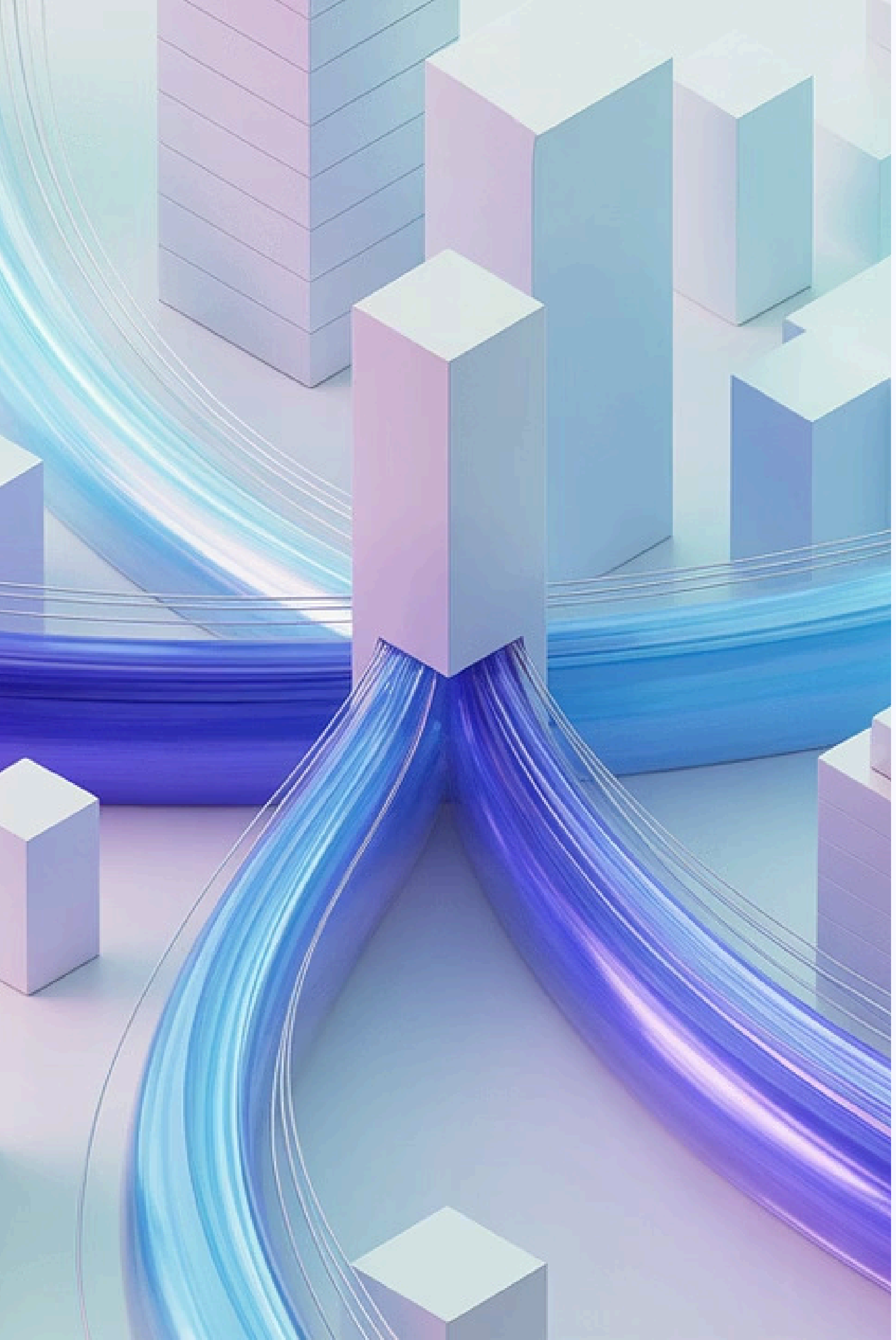


Big Data Concepts & Terminology for Enterprise-Scale Applications

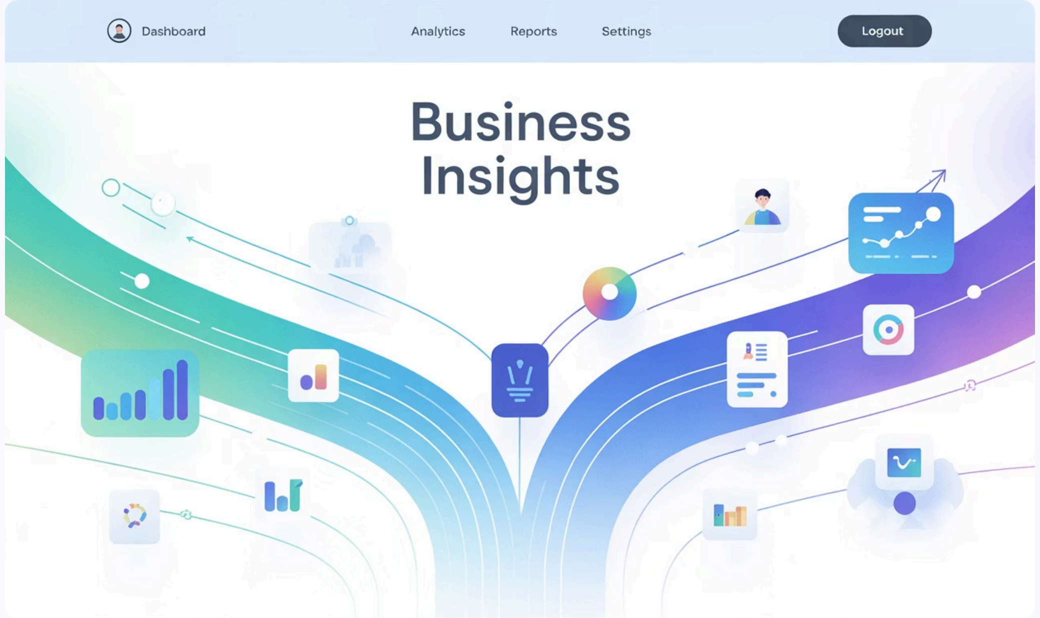
Why It Matters, Who Uses It, How It Works

J **by Jemael Nzihou**

Why Do We Care About Big Data Today?

- Explosion of digital data: IoT, social, transactions
- Businesses need insights faster than ever
- Data-driven decisions = competitive advantage

Real-world impact – e.g., personalized shopping, fraud detection.



What is Big Data?

- Definition: Data too big/complex for traditional systems
- The *6 Vs*: Volume, Velocity, Variety, Veracity, Value, Virality
- Structured vs Unstructured Data

Volucity
Volume
Veracity
6 VS
of Big Data
Veracity
Value
Virality

Who Uses Big Data?

Retail

Amazon, Walmart

Finance

Banks, fintech

Healthcare

Patient data, genomics

Manufacturing

Predictive maintenance

Beneficiaries:

Organizations, governments, consumers

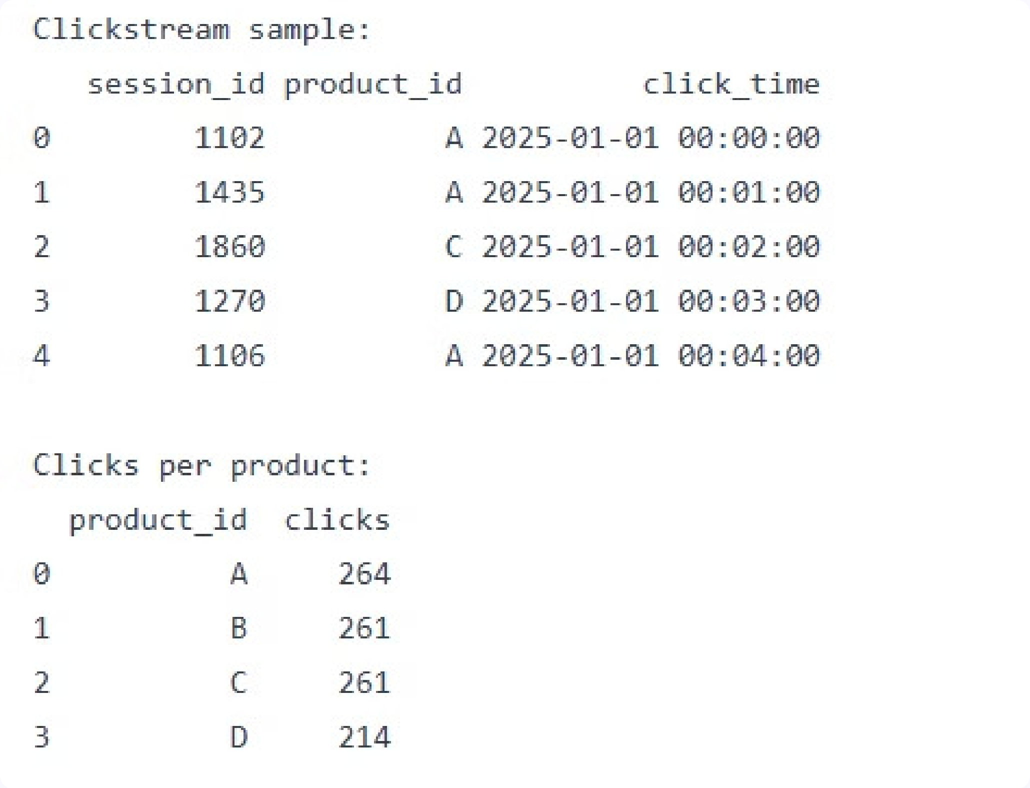
When & Why Is Big Data Needed?

When:

- Huge scale: petabytes, billions of records
- Real-time streaming (web clicks, sensors)
- Complex data: text, video, logs

Why:

- Better decisions
- Faster reactions
- Innovation



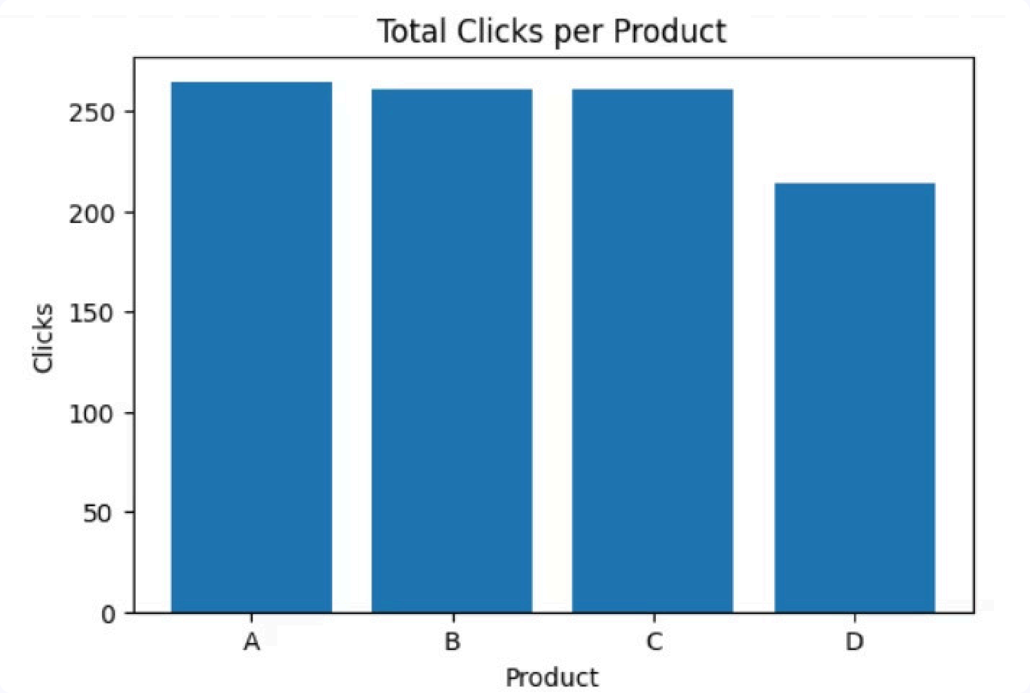
```
Clickstream sample:
   session_id product_id          click_time
0        1102          A 2025-01-01 00:00:00
1        1435          A 2025-01-01 00:01:00
2        1860          C 2025-01-01 00:02:00
3        1270          D 2025-01-01 00:03:00
4        1106          A 2025-01-01 00:04:00

Clicks per product:
  product_id  clicks
0          A     264
1          B     261
2          C     261
3          D     214
```

E-Commerce Data in Real-Time

- This structured data demonstrates how e-commerce sites record user interactions as they happen
- Each log entry captures a specific action from users
- The first five clicks come from different session IDs, showing multiple users browsing various products (A, C, D) at slightly different timestamps
- This simple table illustrates two key big data concepts: **volume** and **velocity**
- Now imagine this table expanding to millions of rows per hour in a major platform like Amazon
- That's the scale of big data in action!

Customer interest level



For Product D showing lower engagement:

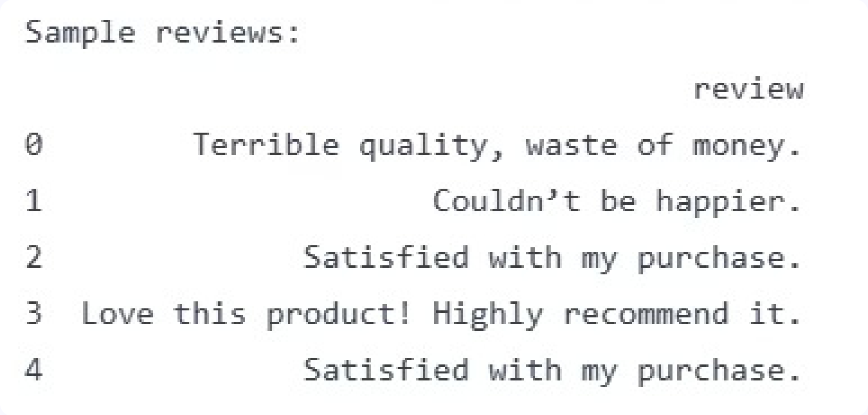
- Evaluate placement optimization on the website
- Develop targeted marketing campaigns or special promotions
- Investigate potential barriers (inventory issues, pricing concerns, or presentation problems)

For high-performing Products A - C:

- Prioritize as featured listings in prominent positions
- Develop strategic cross-selling opportunities with complementary items
- Create value-adding bundle offers to increase average order value

These data-driven insights enable immediate tactical adjustments to maximize revenue potential across your product portfolio.

A mix of positive and negative sentiment:



```
Sample reviews:
                                    review
0         Terrible quality, waste of money.
1                      Couldn’t be happier.
2               Satisfied with my purchase.
3   Love this product! Highly recommend it.
4               Satisfied with my purchase.
```

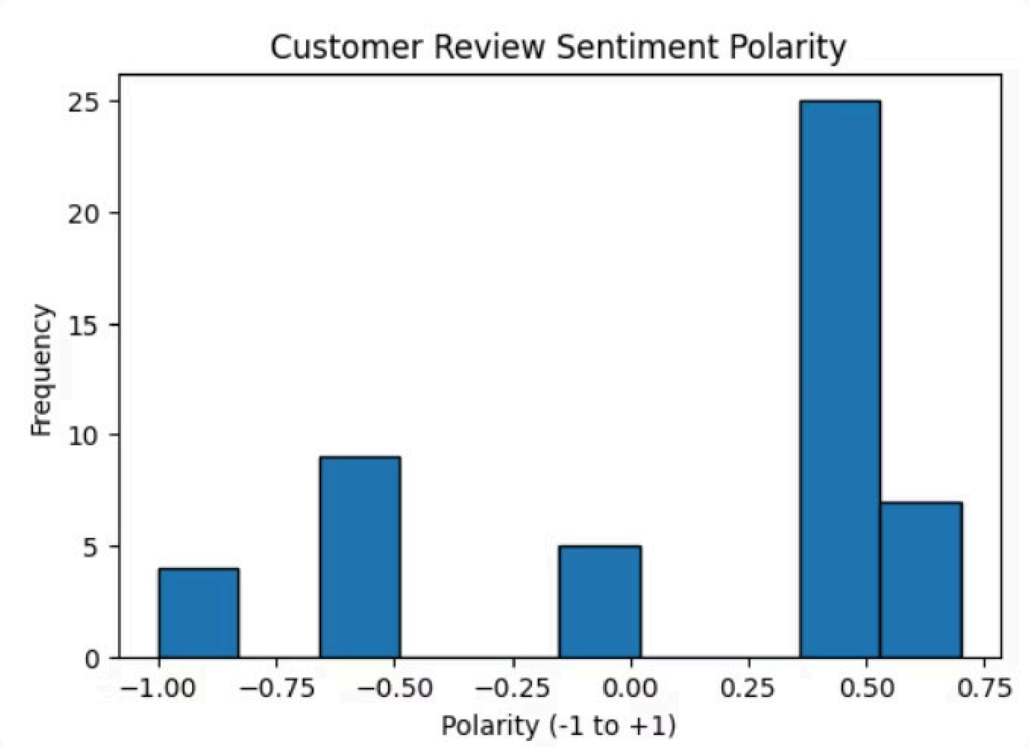
Positive-->

- "Couldn't be happier."
- "Satisfied with my purchase."
- "Love this product! Highly recommend it."

Negative-->

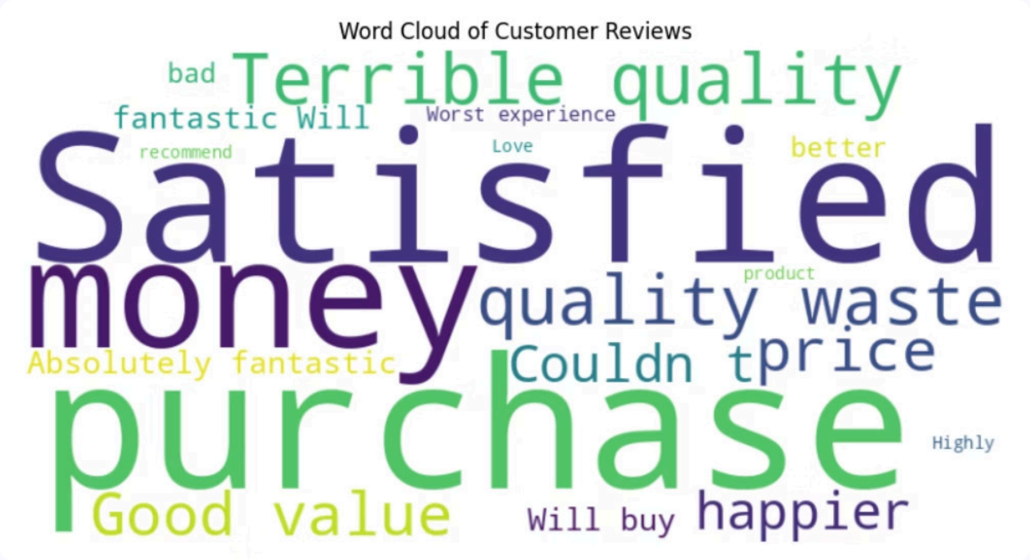
- "Terrible quality, waste of money."

Such reviews directly show veracity (truthfulness) challenges: some feedback is glowing, some is critical – both must be analyzed.



sentiment polarity

- Most reviews have positive polarity around +0.5 – indicating that many customers left satisfied or happy feedback
- A smaller number of reviews fall into strongly negative polarity ranges (-1.0 to -0.5)
- There's a small peak near zero, representing neutral or mixed reviews



Word Cloud Analysis

Most prominent words:

- "Satisfied", "purchase", "money" – suggest that many customers mention satisfaction and the value they get for their money.

Positive signals:

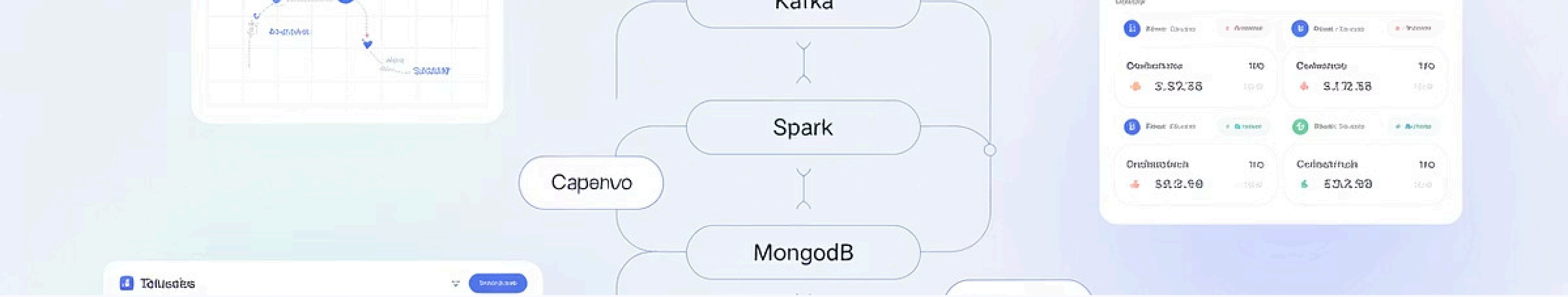
- Words like "Satisfied", "Good value", "Absolutely fantastic", "recommend", "Love", "happier" stand out.

Negative signals:

- Words like "Terrible quality", "waste", "Worst experience", "bad" also appear, highlighting mixed feedback.

Common Big Data Tools

Tool	Purpose	Example
Hadoop	Batch processing & storage	Facebook
Spark	In-memory, fast analytics	Netflix
Kafka	Real-time streams	LinkedIn
NoSQL	Flexible storage	Instagram
Tableau/Power BI	Visualization	Walmart dashboards



Hands-On Example

E-Commerce Customer Behavior

Kafka streams click data

Spark processes in real-time

MongoDB stores sessions/reviews

Tableau dashboard: top products, sentiment, drop-offs

How Big Data Solves Decision-Making

1. Data Ingestion

2. Storage

3. Processing (batch/real-time)

4. Analytics (descriptive, predictive)

5. Visualization (dashboards)

6. Actionable Decision

Core Math & Analytics

Statistics:

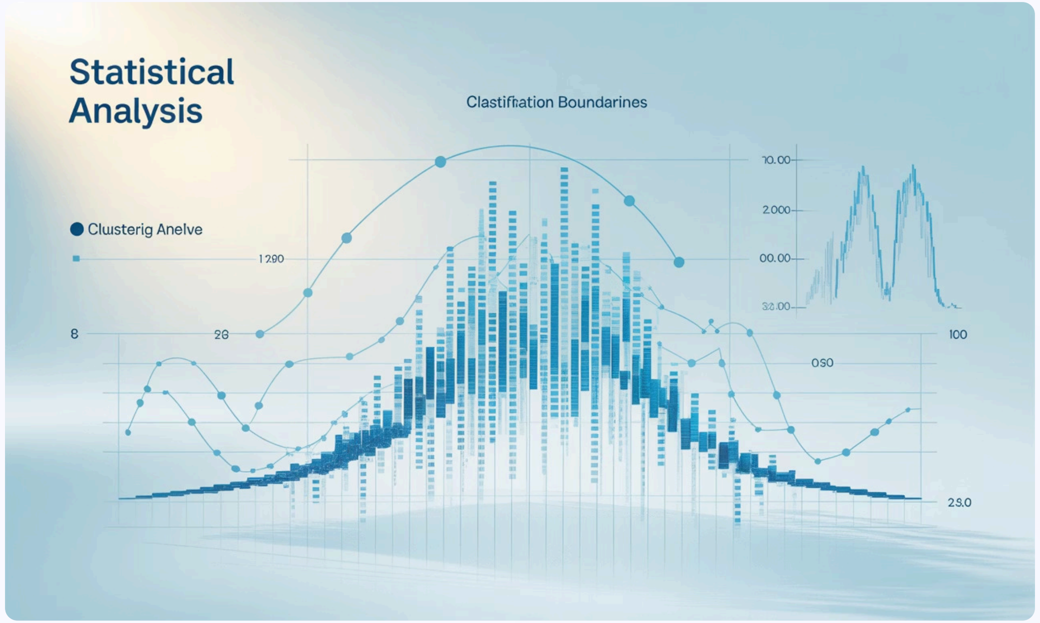
Mean, variance, correlation

Predictive models:

- Regression: $y = \beta_0 + \beta_1 x$
- Classification: logistic regression
- Clustering: K-means

Optimization:

Gradient descent



Alternatives to Big Data Tools

- Data warehouse: Snowflake, Redshift
- Edge computing for IoT
- Relational DB + ETL for smaller orgs
- Analytics-as-a-Service (Google Analytics)

Example: Small retailer uses Google Analytics + Looker Studio instead of Hadoop.

Summary

- Big data turns massive data into value.
- Tools & math make it scalable & actionable.
- Impacts daily life, industries, and jobs.
- Key for digital transformation & innovation.

References

- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing https://doi.org/10.1016/j.is.2014.07.006
- Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM, 57*(7), 86-94. https://doi.org/10.1145/2611567

References (Continued)

- Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*(2), 171-209. **https://doi.org/10.1007/s11036-013-0489-0**
- Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and good practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)* (pp. 404-409). IEEE. https://doi.org/10.1109/IC3.2013.6612229