Harnessing the Power of Big Data Analytics

Dr. Sharon Jones

Let’s collect some data!

PollEv.com/sharonjones821

What is Big Data?

Definition: Big data refers to extremely large data sets that can be analyzed computationally to reveal patterns, trends, and associations.

Key Characteristics:

- **Volume**: Vast amount of data.
- **Velocity**: Speed of data generation and processing.
- **Variety**: Different types of data (structured, unstructured, etc.).
- **Veracity**: Accuracy and trustworthiness of data.

Video Reference

Where Does the Data Come From?

- **Data Generation**: Where does the data come from? (Sensors, devices, social media, etc.)
- **Data Collection**: Methods for capturing and storing large datasets.
- **Data Analysis**: How data is transformed into actionable insights.

Data Generation: ***Where Does the Data Come From?***

Sensors:

- Smart home devices like Nest thermostats collect temperature and energy usage.
- Industrial sensors monitor machine performance in factories to predict maintenance needs.

Devices:

- Smartphones generate location data, app usage statistics, and call logs.
- Wearable fitness trackers like Fitbit record steps, heart rate, and sleep patterns.

Social Media:

- Platforms like Twitter and Instagram produce massive amounts of user-generated content, including posts, likes, and comments.
- Trending hashtags and user sentiment provide insights into public opinion.

E-Commerce Platforms:

- Websites like Amazon record purchase history, clicks, and product reviews.

Data Collection: ***Methods for Capturing and Storing Large Datasets***

APIs (Application Programming Interfaces):

- Example: Twitter API collects tweets based on specific keywords or hashtags.
- Use Case: Brands analyze real-time customer sentiment.

Logs:

- Example: Web server logs track visitor activity on websites.
- Use Case: Companies improve user experience by identifying navigation patterns.

Databases:

- Example: SQL and NoSQL databases like MongoDB and Cassandra store structured and unstructured data.
- Use Case: Retailers manage inventory and sales records efficiently.

Cloud Platforms:

- Example: AWS, Google Cloud, and Azure store massive datasets securely and provide scalable access for analysis.

Data Analysis: ***Transforming Data into Actionable Insights***

Data Cleaning:

- Example: Removing duplicate records or correcting typos in a customer database.
- Outcome: Ensures accuracy in reporting and predictions.

Data Mining:

- Example: Walmart uses data mining to predict demand for products during specific weather events (e.g., pop-tarts before hurricanes).
- Outcome: Optimized supply chain and inventory management.

Visualization:

- Example: Tableau creates a sales heat map showing regional performance.
- Outcome: Identifies high-performing and underperforming regions.

Machine Learning:

- Example: Netflix uses machine learning algorithms to recommend shows and movies based on viewing history.
- Outcome: Increased user engagement and satisfaction.

Big Data Platforms

Apache Spark

Apache Hadoop:

Databricks

Snowflake

Amazon S3

Microsoft Azure

How Does Netflix Use Big Data Platforms

- Netflix stores all of its data on amazon web services S3 storage which enables them to spin multiple Hadoop clusters for different workloads and allow them to access data at the same time.
- The tools like Hive for ad hoc queries and analytics, Pig for ETL and algorithms, Java based mapReduce for complex algorithms are the backbone of Netflix.
- Python is the language for scripting various ETL processes and Pig User Defined Functions.

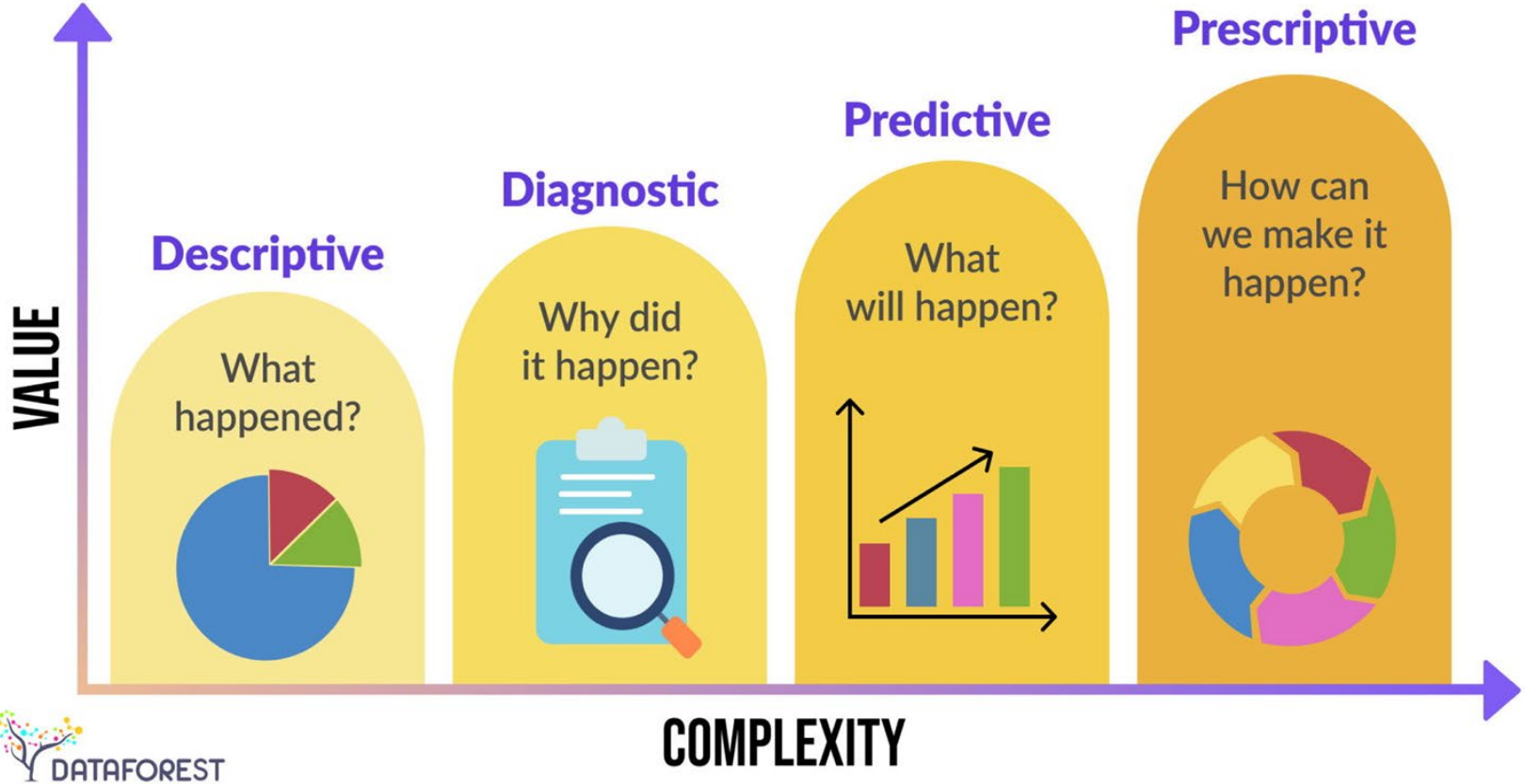
- Netflix uses Amazon's Elastic MapReduce (EMR) for distribution of Hadoop.
    - A Hadoop cluster is a type of computational cluster that uses the Hadoop software framework to process data in a distributed environment.
    - Apache Pig is a high-level tool for performing Extract, Transform, and Load (ETL) operations on large data sets PIG ETL
    - MapReduce is a Java-based, distributed execution framework within the Apache Hadoop Ecosystem.

Case Study

Predictive Analytics & Machine Learning

- **What is Predictive Analytics?**
  - Using historical data to predict future outcomes.
  - Example: Predicting customer behavior.
- **Role of Machine Learning**
  - Automating data analysis processes.
  - Examples of algorithms used: decision trees, neural networks.

TYPES OF DATA ANALYTICS



DATAFOREST


HOW AI
WORKS
MACHINE
LEARNING
C O
D E

Activity

Resources

Data Games

DataBasic.io

YouCubed Data Tells a Story

CodeWizardsHQ