Manoj Kumar

Introduction to Big Data: Understanding the Basics

Big Data

Big Data is a revolution in this technical age. According to reports, 90% of the current data was generated in the past two years.

Big Data is defined as vast amounts of data measured in terabytes or petabytes. It helps companies to understand their products and services better and generates valuable insights.

What is Big Data:

Big Data refers to large and often complex datasets that are too big for traditional relational databases to handle. These datasets need special tools and methods to perform operations. Big Data is made up of structured, semi-structured, and unstructured data, including audio, video, photos, websites, and more.

That how Huge Big Data is!

Manoj Kumar

Types of Big Data

Structured Data: This is highly organized data that follows a specific format for storage and processing. This type of data is easy to retrieve and includes information like mobile numbers, employee details, and salaries. Data stored in relational database management is an example of structured data.

Unstructured Data: Unstructured data is highly unorganized and does not follow a specific format. This type of data cannot be stored in relational databases and cannot be analyzed until it is transformed into a structured format. 80% of data in an organization is unstructured, and it is available in multiple formats such as images, audio, video, social media posts, and more.

Semi-structured Data: Semi-structured data is a combination of structured and unstructured data that doesn't have a specific format but has identifying characteristics. For example, images may contain metadata or tags, but the information within has no structure. XML or JSON files are examples of semi-structured data.

5 V's in Big Data

Manoj Kumar

Volume: The massive amount of data that is growing at an exponential rate. The data can range from terabytes to petabytes and beyond. It needs to be stored in distributed systems like Hadoop and MongoDB, as traditional databases cannot handle such large amounts of data.

Velocity: The speed at which data is generated, stored, and analyzed. The velocity of data generation is particularly high in real-time applications such as social media, which produces audio, videos, posts, etc. at a rapid pace.

Variety: The different types of data that are used daily. The data can be structured, unstructured, or semi-structured, collected from diverse sources. The need for analyzing and processing technologies that can handle different data formats (text, audio, videos, etc.) has become increasingly important.

Veracity: The quality or trustworthiness of available data. Veracity deals with the accuracy and certainty of the data analyzed. The reliability of the data is essential, especially in the case of data generated in real-time, such as on social media platforms, where the data can be subject to errors, typos, and other inaccuracies.

Value: The need to convert raw data into something valuable, in order to extract useful information. The data is considered valuable if it results in a meaningful return on investment.

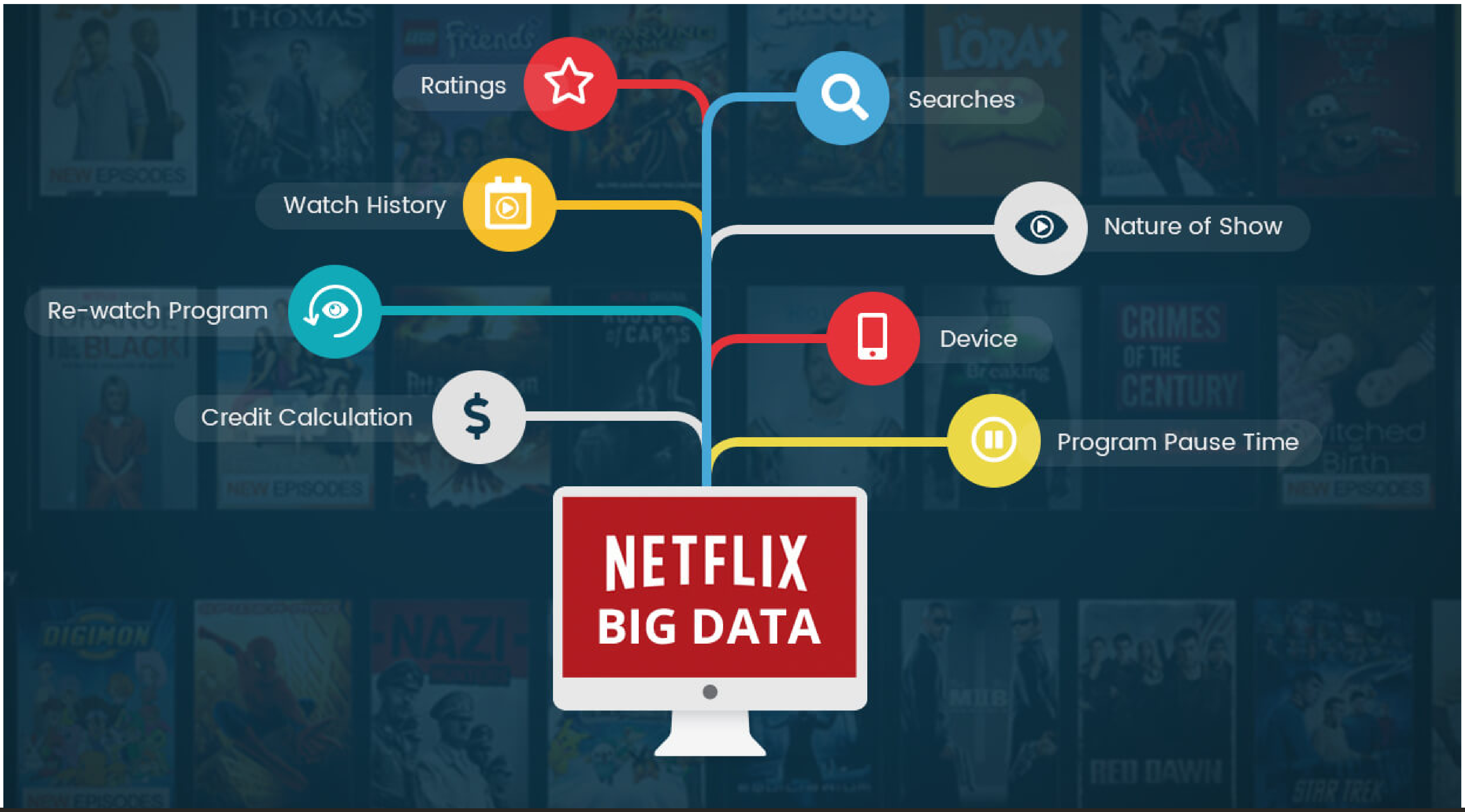
Use Case : Netflix

Manoj Kumar

Netflix analyzes our data about what we're watching or searching, extracting the data points from that, like what titles customers watch, what genre they like, how often playback stopped, ratings are given, etc., and feeding that to its recommendations system.

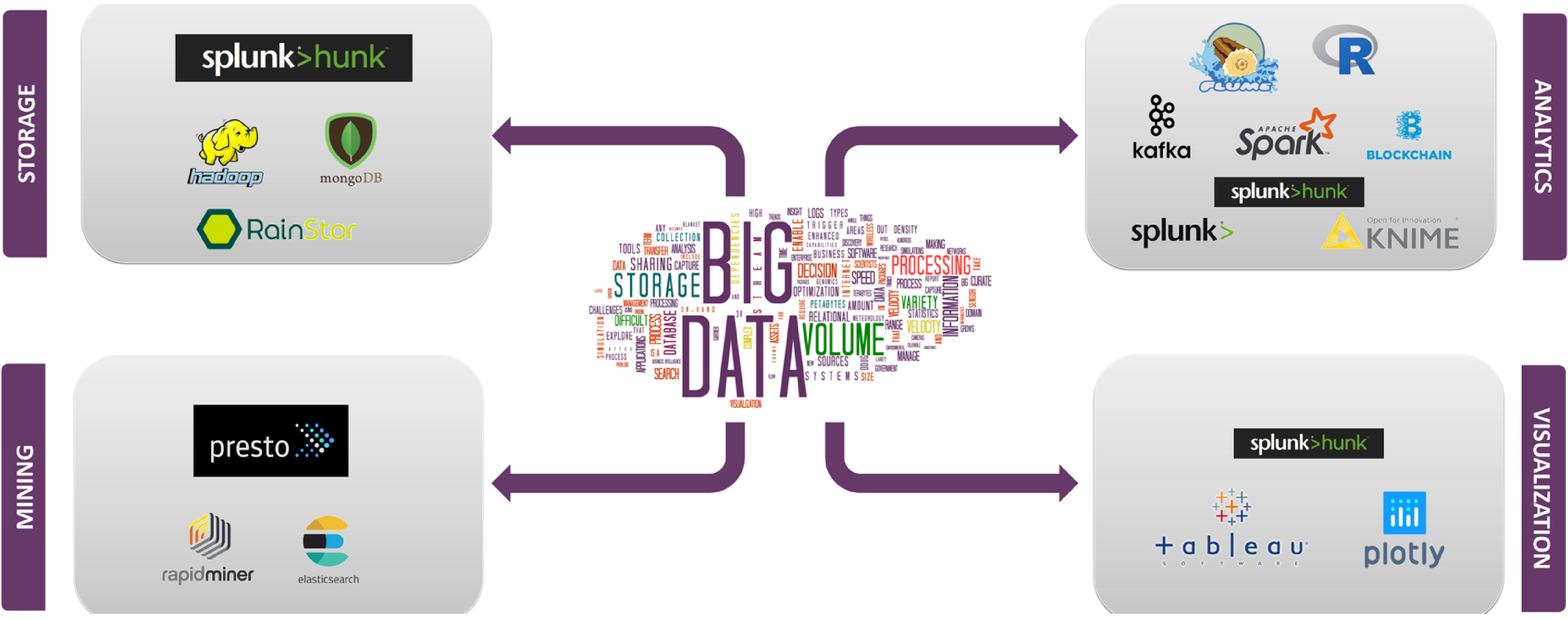
This will make decisions smooth and firm in terms of knowing the customers' needs rather than assuming them (what most companies do).

The major data structures used in this process include Hadoop, Hive, Pig, and other traditional business intelligence.



Manoj Kumar

Top Tools in Big Data



STORAGE

ANALYTICS

MINING

VISUALIZATION

Manoj Kumar

Looking for Real-World Experience in Data Analytics/BI?

IM me on LinkedIn to know more
Or
click here to book a call

https://www.linkedin.com/in/mk-analytics