Introduction to

TEKsystems
Own change

Content

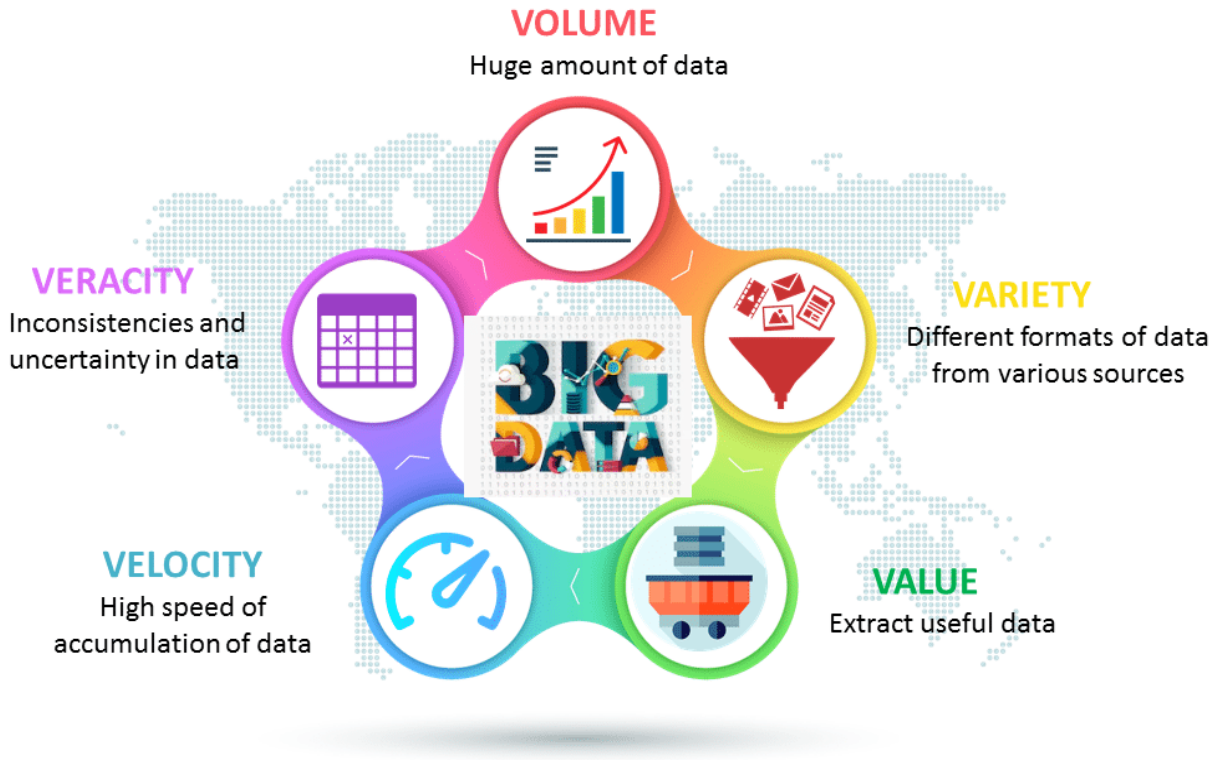
- Introduction
- What is Big Data & Hadoop
- Characteristic of Big Data
- Why Hadoop
- History
- Big Data sources
- Components of Big Data
- Benefits of Big Data
- Risks of Big Data
- Application of Big Data
- Use Cases
- How Big Data Impact IT
- Future of Big Data

What is Big Data & Hadoop

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The challenges include capture, curation, storage, search, sharing, analysis and visualization.

Hadoop is a framework written in Java that :

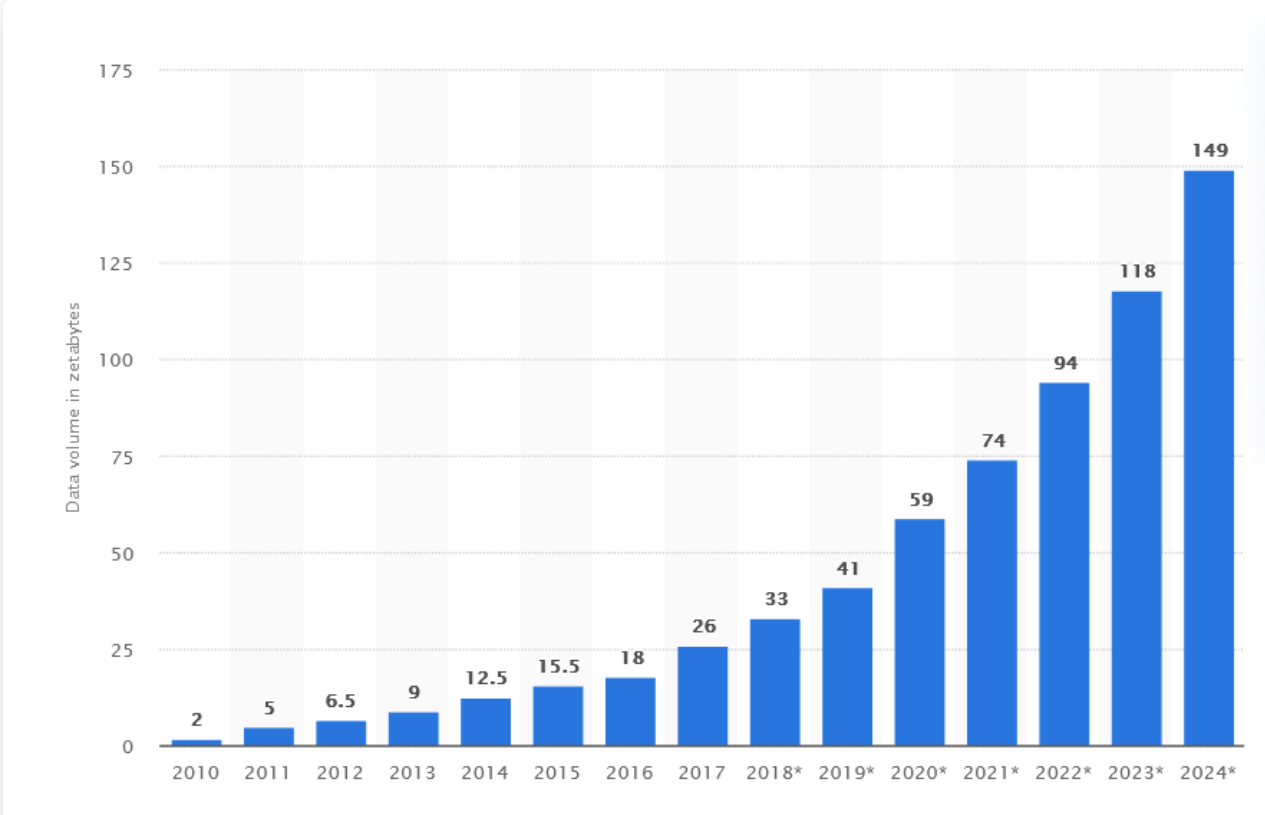
- Enables distributed storage and processing of large volumes of data
- Processes the data using the MapReduce programming model.
- Provides high availability by handling any failure efficiently.



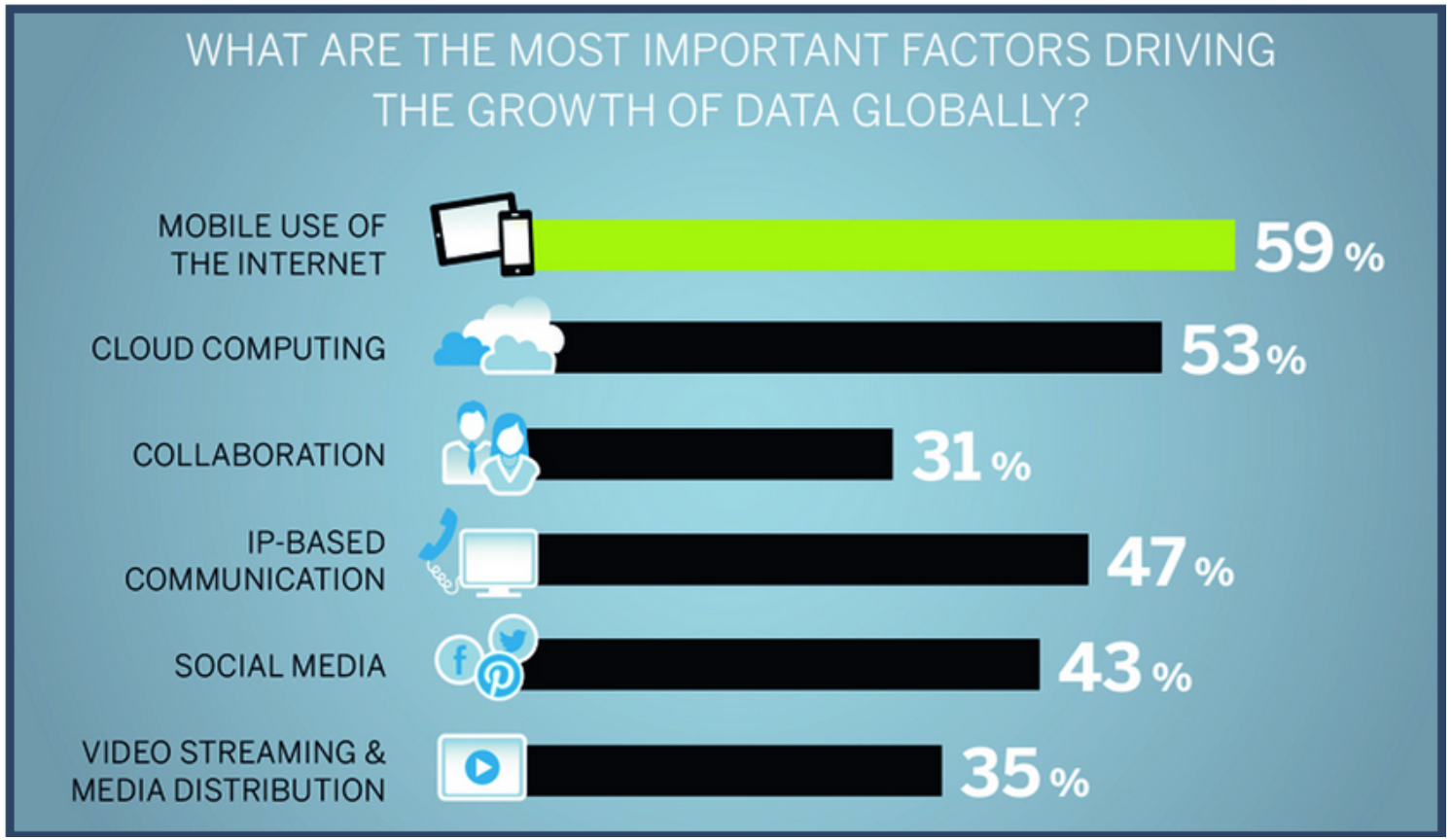
Characteristic of Big Data

Volume

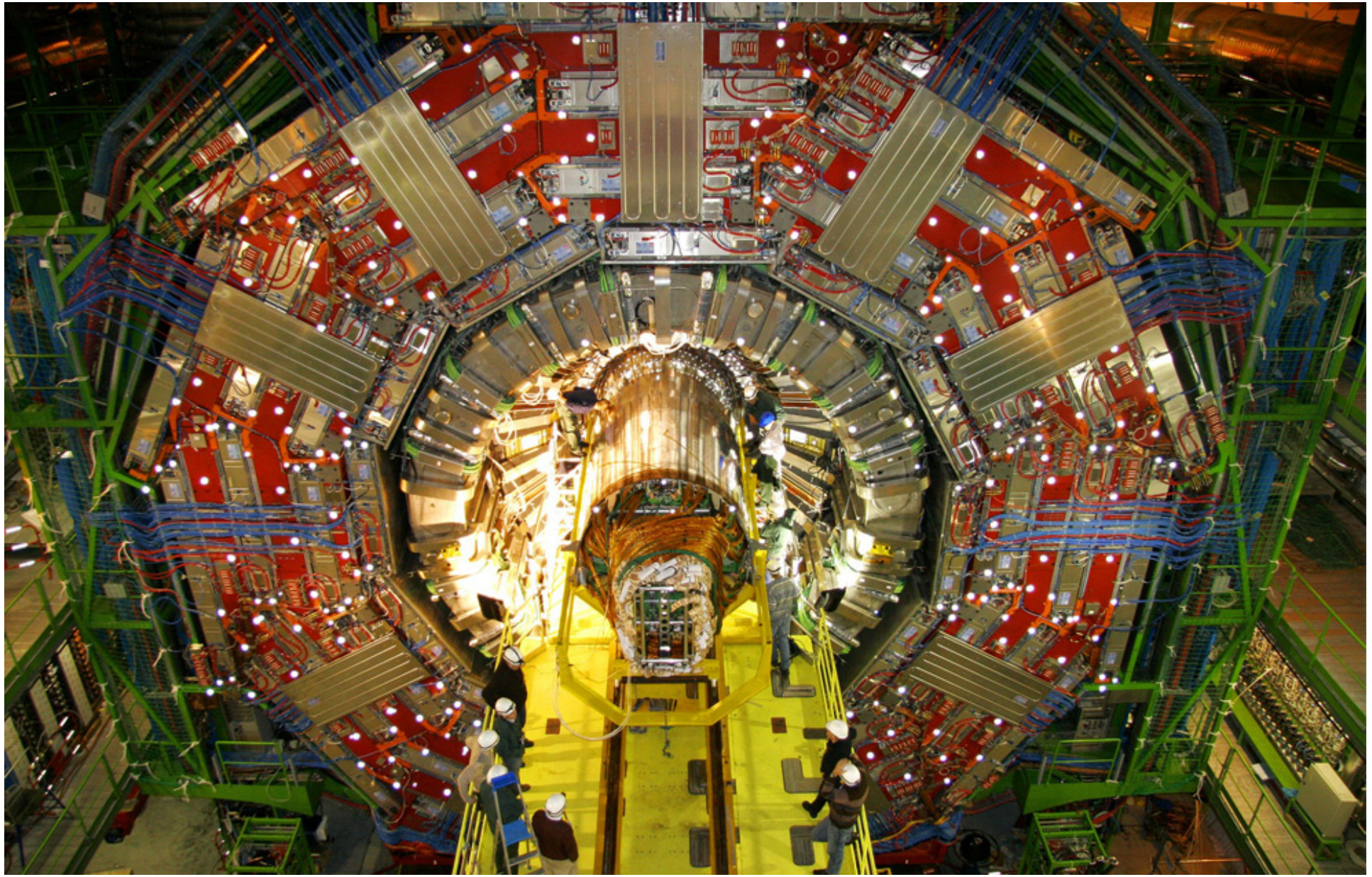
- Volume of data will grow to 44 ZB by the end of 2020.
- Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data
- 247 billion email messages sent each day



Velocity



- 90% of all data has been created in the last two years
- Today, Facebook ingests 500 terabytes of new data every day
- Twitter users send over half a million tweets every minute.
- YouTube users upload 48 hours of new video every minute of the day.
- WhatsApp users exchange up to 65 billion messages daily
- 571 new websites are created every minute of the day.
- Internet users generate about 2.5 quintillion bytes of data each day.

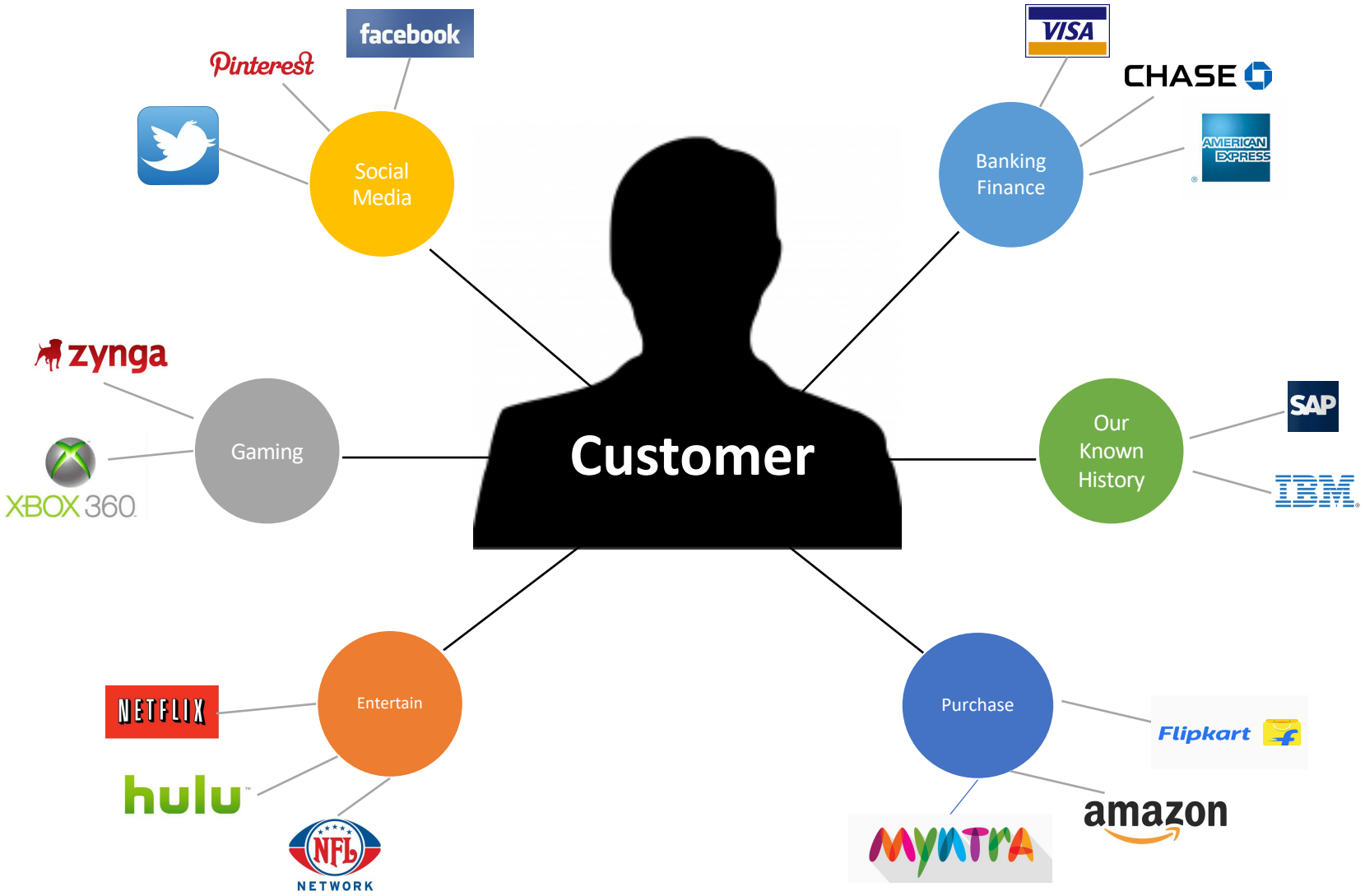


CERN’s Large Hadron Collider (LHC) generates 15 PB a year

Variety

- Big Data isn't just numbers, dates and strings. Big Data is also geospatial data, 3D data, audio and video, and unstructured text, including log files and social media.
- 80-90% of the data we generate today is unstructured
- Various machine output data medical records, game console records
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.

A Single View to the Customer



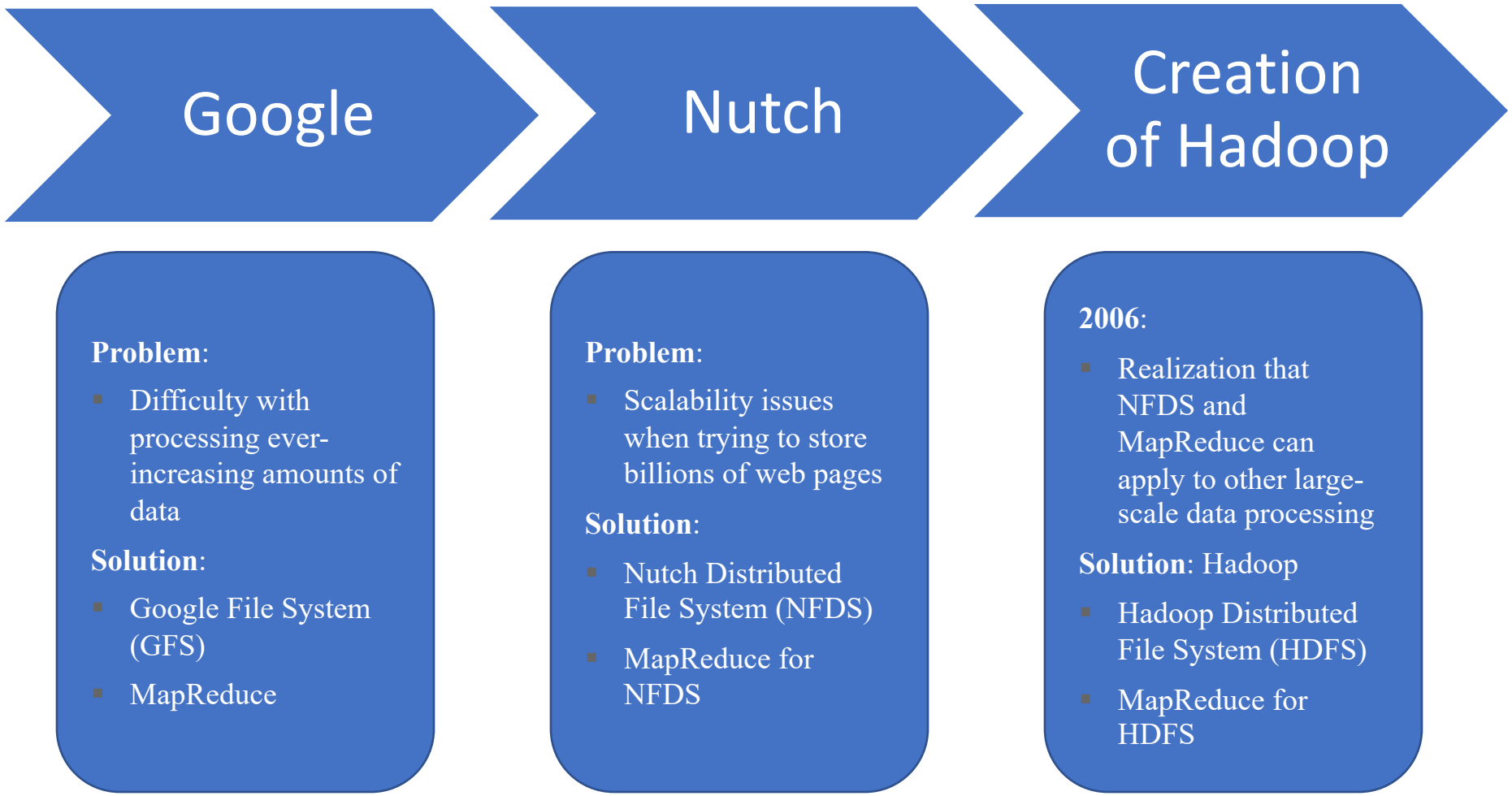
Veracity

- Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Veracity in data analysis is the biggest challenge when compares to things like volume and velocity.
- Data in Doubt: Uncertainty due to data inconsistency & incompleteness, uncleansed, untrusted, ambiguities, deception.

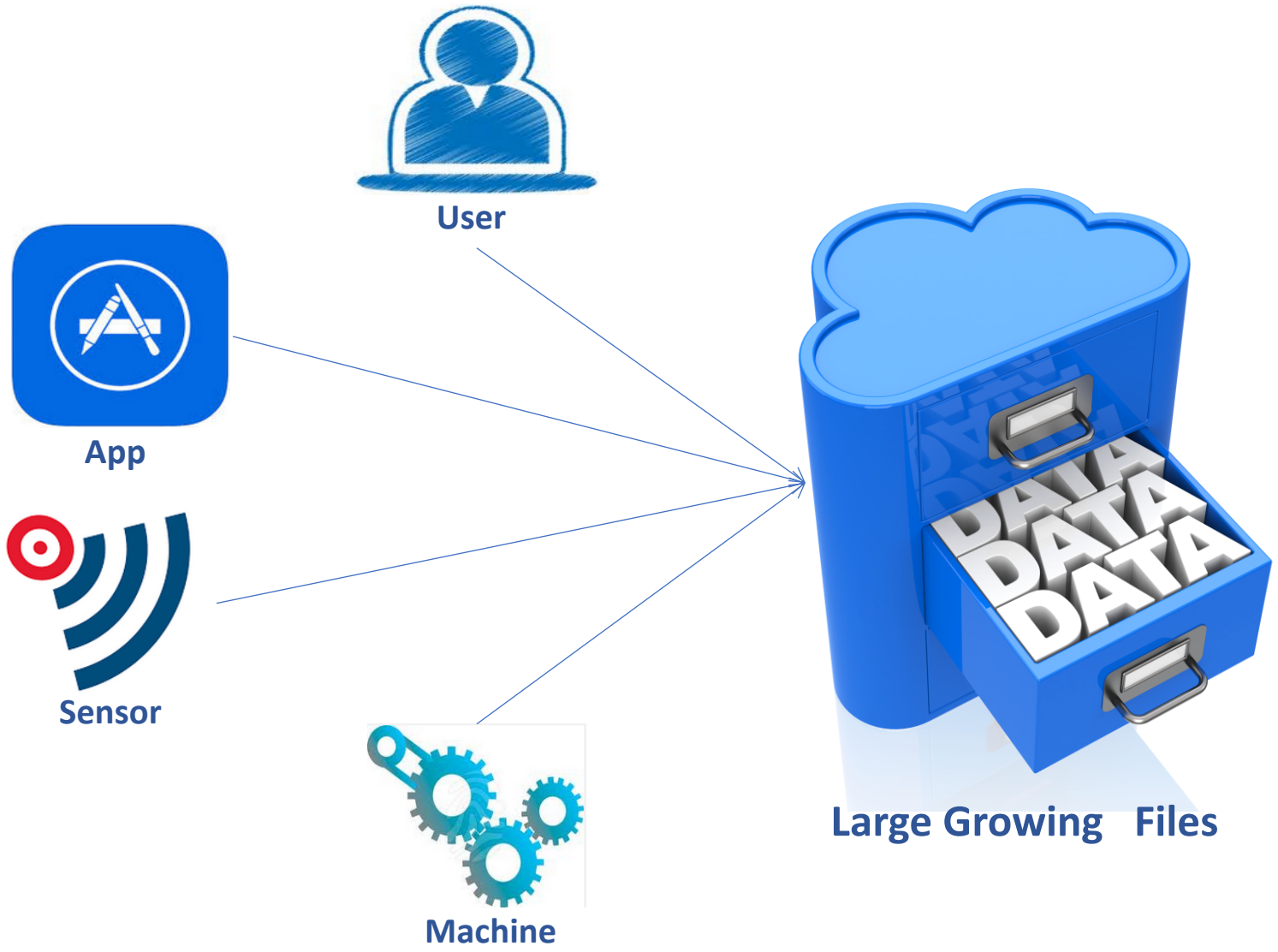
Why Hadoop

- **Open-source software:** Broad and open network of developers create and manage the programs
- **Distributed:** Data is divided and stored across multiple computers, and computations can be run in parallel across multiple connected machines
- **Massive storage:** Hadoop framework can store huge amounts of data by breaking the data into blocks and storing it on clusters of lower-cost commodity hardware
- **Faster processing:** Hadoop processes large amounts of data in parallel across clusters
- **Low cost:** The open-source framework is free and uses commodity hardware to store large quantities of data.
- **Scalability:** You can easily grow your system simply by adding more nodes
- **Storage flexibility:** Unlike traditional relational databases, you don’t have to preprocess data before storing it. And that includes unstructured data like text, images and videos
- **Fault Tolerance:** Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail.

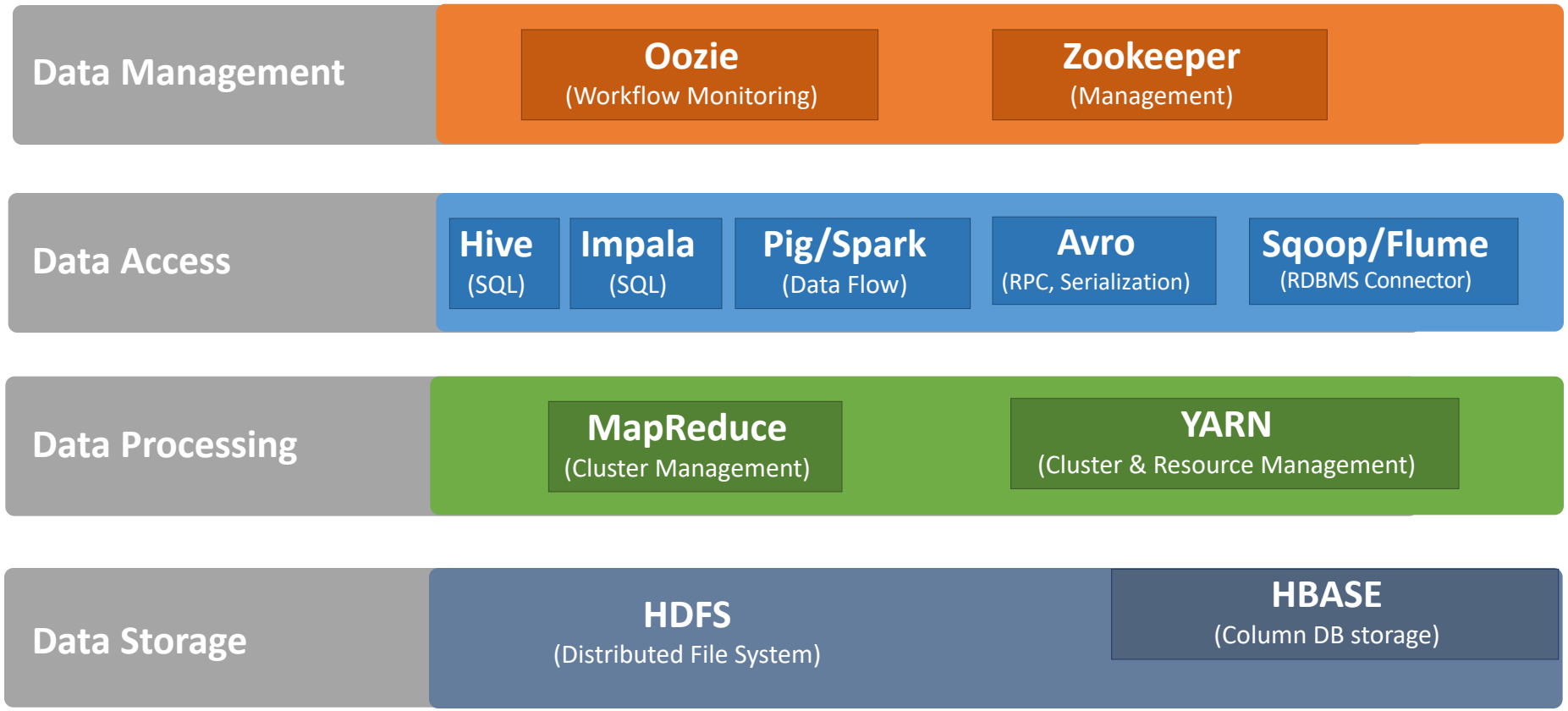
History


Google
Nutch
Creation of Hadoop
Problem:
Difficulty with processing ever-increasing amounts of data
Solution:
Google File System (GFS)
MapReduce
Problem:
Scalability issues when trying to store billions of web pages
Solution:
Nutch Distributed File System (NFDS)
MapReduce for NFDS
2006:
Realization that NFDS and MapReduce can apply to other large-scale data processing
Solution: Hadoop
Hadoop Distributed File System (HDFS)
MapReduce for HDFS

Big Data sources



Components of Big Data



Benefits of Big Data

- **Cost reduction:** Open-source framework & commodity hardware to store
- **Faster, better decision making:** Because of its architecture
- **New products and services:** Open source developers are creating components everyday to fulfill the needs
- **Reliable:** The data is distributed among the hard drive with several copy of it
- **Better Analysis:** Helps you to use more granule level data, sentimental analysis & predictive analysis are cherry on the top
- **Scalability:** You can easily grow your system simply by adding more nodes
- **Customize your website in real time:** personalize the content or look and

  feel of the website in real time to suit each consumer depending on their

  interest, age, sex, location

Risks of Big Data

- **Discrimination:** The use of predictive analytics can now be used by the government and companies to make determinations about our ability to fly, to obtain a job, a clearance, or a credit card. Which will have a negative impact on individuals directly inhibits freedom of association
- **Catastrophic data breaches:** There are numerous reports of Big Data analytics being used to expose personal details. (Target, Home Depot, eBay, Sony)
- **Government exemptions:** Personally Identifiable Information (PII) including name, any aliases, race, sex, date and place of birth, Social Security number, passport and driver's license numbers, address, telephone numbers, photographs, fingerprints, financial information like bank accounts, employment and business information and more are easily available

Application of Big Data

Smarter Healthcare

Homeland Security

Traffic Control

Manufacturing

Multi-channel sales

Telecom

Trading Analytics

Search Quality

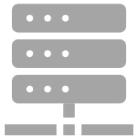
Use Cases

Recommendation Engines

- If you've been on YouTube, Netflix, Spotify, or other online media services, you may have noticed those "recommended for you" videos, movies, or music
- Big data, with its scalability and power to process massive amounts of both structured (eg. video titles users search for, music genre they prefer) and unstructured data (eg. user viewing/listening patterns)
- This can enable companies to analyze billions of clicks and viewing data from you and other users like you for the best recommendations
- Over time, through machine learning and predictive analytics, the recommendations become better tailored to the user's taste

Log Analytics

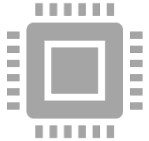
Log data is a fundamental foundation of many business big data applications.



With the exponential growth of business activities and transactions, log data can be a huge.

Open source log analytics tools can provide you the ability collect, process, and analyze massive log data

Log search capabilities and big data analytics has enabled organizations to discover insights for more agile operations.



Widely used for various business goals, from IT system security and network performance, to market trends and e-commerce personalization

E-Commerce Personalization

A powerful search and big data analytics platform allows e-commerce companies to:

- Clean and enrich product data for a better search experience on both desktops and mobile devices
- Use predictive analytics and machine learning to predict user preferences through log data, then personalize products in a most-likely-to-buy order that maximizes conversion

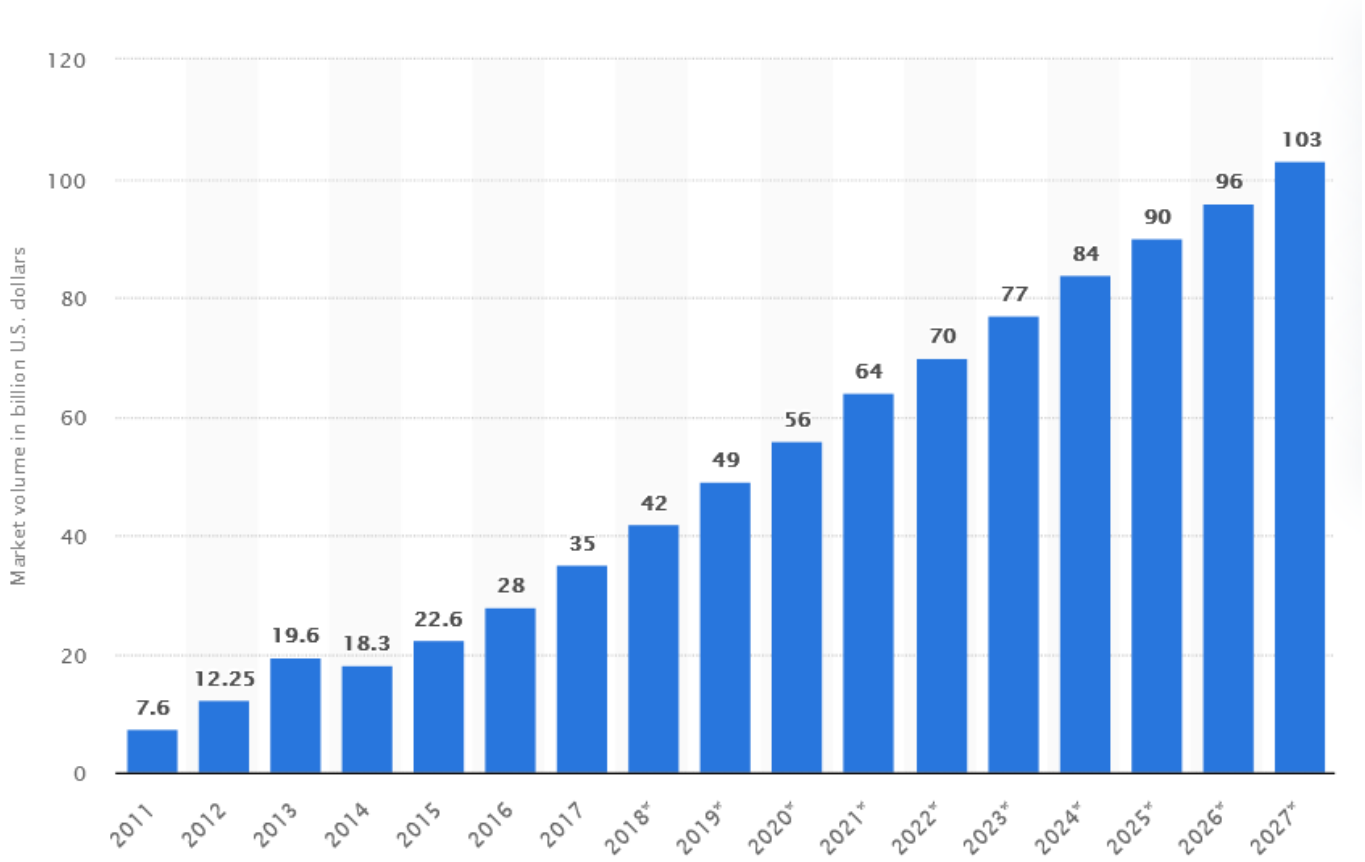
How Big Data Impact IT

- Big data is a troublesome force presenting opportunities with challenges to IT organizations.
- 95% of businesses cite the need to manage unstructured data as a problem for their business
- Since 2012, big data has created **8 million jobs** in the US alone and 6 million more worldwide.
- India will require a minimum of 1 lakh data scientists in the next couple of years in addition to data analysts and data managers to support the Big Data space.
- Leading technology vendors

Yahoo! Google Amazon Apache Facebook SAP

EMC Cloudera Hortonworks Oracle IBM Teradata

Future of Big Data



- **The big data analytics market is set to reach $77 billion by 2023.**
- **95% of businesses cite the need to manage unstructured data as a problem for their business.**
- **97.2% of organizations are investing in big data and AI.**
- **Using big data, Netflix saves $1 billion per year on customer retention**
- **In 2020, the big data market is expected to grow by 14%**

Any Questions?

TEKsystems
Own change

THANK YOU