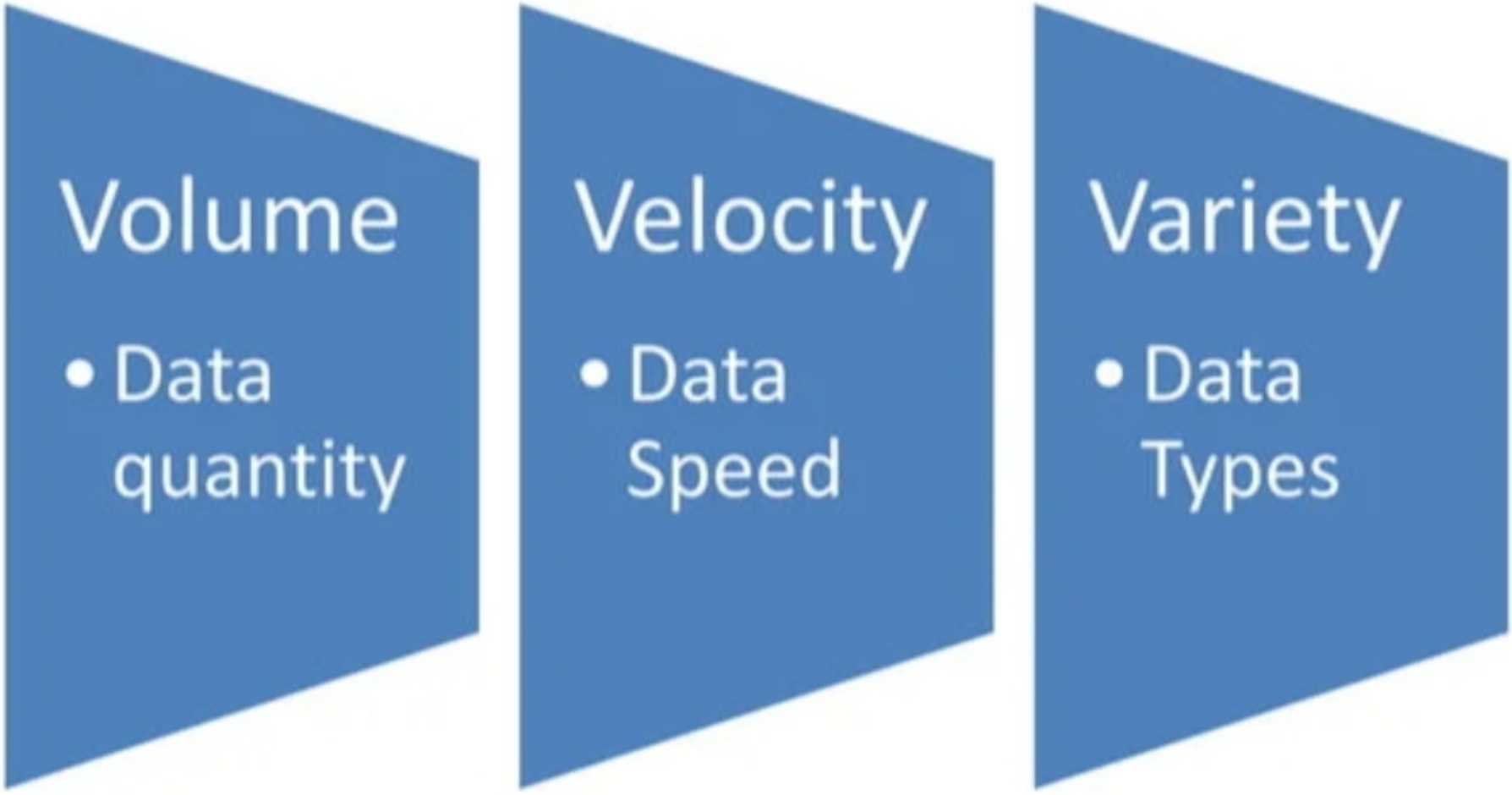
What is BIG DATA?

- 'Big Data' is similar to 'small data', but bigger in size
- but having data bigger it requires different approaches:
  - Techniques, tools and architecture
- an aim to solve new problems or old problems in a better way
- Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

Three Characteristics of Big Data V3s



1st Character of Big Data
Volume

- A typical PC might have had 10 gigabytes of storage in 2000.

- Today, Facebook ingests 500 terabytes of new data every day.

- Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.

- The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video.

2nd Character of Big Data

Velocity

- Clickstreams and ad impressions capture user behavior at millions of events per second
- high-frequency stock trading algorithms reflect market changes within microseconds
- machine to machine processes exchange data between billions of devices
- infrastructure and sensors generate massive log data in real-time
- on-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

3rd Character of Big Data

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.

- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.

- Big Data analysis includes different types of data

Storing Big Data

❖ **Analyzing your data characteristics**

- Selecting data sources for analysis
- Eliminating redundant data
- Establishing the role of NoSQL

❖ **Overview of Big Data stores**

- Data models: key value, graph, document, column-family
- Hadoop Distributed File System
- HBase
- Hive

Processing Big Data

❖ **Integrating disparate data stores**

- Mapping data to the programming framework
- Connecting and extracting data from storage
- Transforming data for processing
- Subdividing data in preparation for Hadoop MapReduce

❖ **Employing Hadoop MapReduce**

- Creating the components of Hadoop MapReduce jobs
- Distributing data processing across server farms
- Executing Hadoop MapReduce jobs
- Monitoring the progress of job flows

The Structure of Big Data

- Structured
  - Most traditional data sources
- Semi-structured
  - Many sources of big data
- Unstructured
  - Video data, audio data

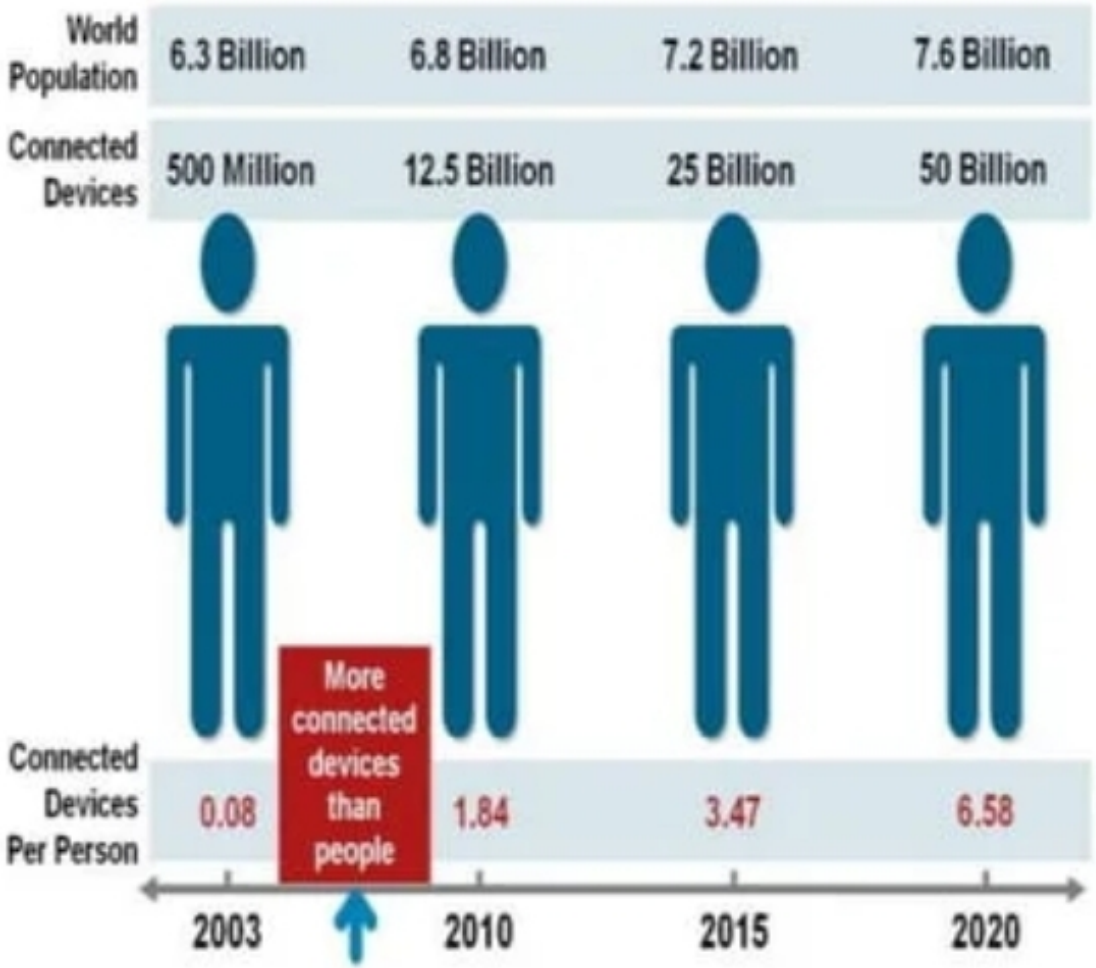
Why Big Data

- Growth of Big Data is needed
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data(different data types)
  - Every day we create 2.5 quintillion bytes of data; 90% of the data in the world today has been created in the last two years alone

Why Big Data

- FB generates 10TB daily
- Twitter generates 7TB of data Daily
- IBM claims 90% of today's stored data was generated in just the last two years.

Figure 1. The Internet of Things Was "Born" Between 2008 and 2009



Source: Cisco IBSG, April 2011

How Is Big Data Different?

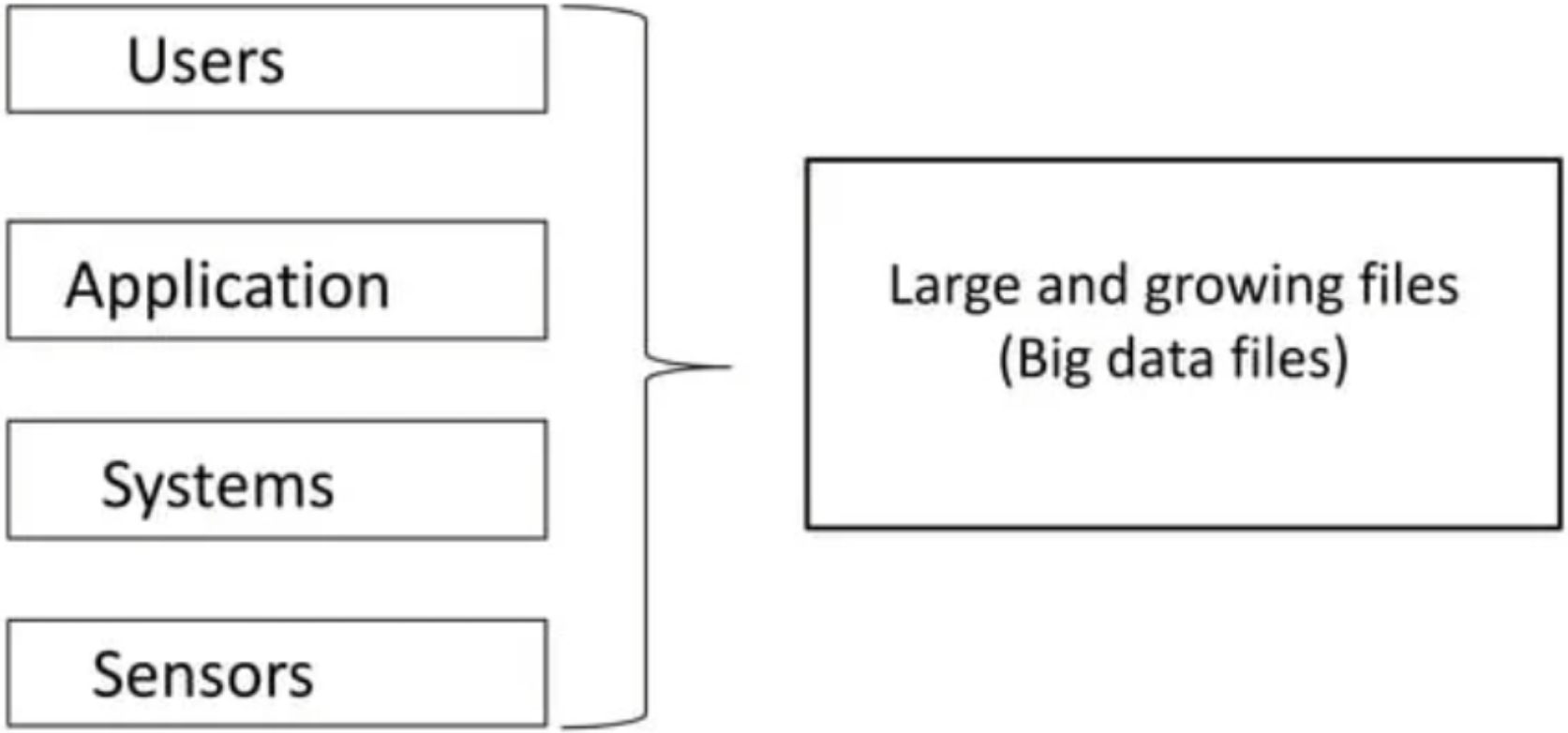
1) Automatically generated by a machine
   (e.g. Sensor embedded in an engine)

2) Typically an entirely new source of data
   (e.g. Use of the internet)

3) Not designed to be friendly
   (e.g. Text streams)

4) May not have much values
   - Need to focus on the important part

Big Data sources



Data generation points Examples

- Mobile Devices
- Microphones
- Readers/Scanners
- Science facilities
- Programs/ Software
- Social Media
- Cameras

Big Data Analytics

- Examining large amount of data
- Appropriate information
- Identification of hidden patterns, unknown correlations
- Competitive advantage
- Better business decisions: strategic and operational
- Effective marketing, customer satisfaction, increased revenue

Application Of Big Data analytics

Smarter Healthcare

Homeland Security

Traffic Control

Manufacturing

Multi-channel sales

Telecom

Trading Analytics

Search Quality

How Big data impacts on IT

- Big data is a troublesome force presenting opportunities with challenges to IT organizations.

- By 2015 4.4 million IT jobs in Big Data ; 1.9 million is in US itself
- 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.

Benefits of Big Data

- Real-time big data isn't just a process for storing petabytes or exabytes of data in a data warehouse, It's about the ability to make better decisions and take meaningful actions at the right time.

- Fast forward to the present and technologies like Hadoop give you the scale and flexibility to store data before you know how you are going to process it.

- Technologies such as MapReduce,Hive and Impala enable you to run queries without changing the data structures underneath.

Benefits of Big Data

- Our newest research finds that organizations are using big data to target customer-centric outcomes, tap into internal data and build a better information ecosystem.
- Big Data is already an important part of the $64 billion database and data analytics market
- It offers commercial opportunities of a comparable scale to enterprise software in the late 1980s
- And the Internet boom of the 1990s, and the social media explosion of today.

Future of Big Data

- $15 billion on software firms only specializing in data management and analytics.
- This industry on its own is worth more than $100 billion and growing at almost 10% a year which is roughly twice as fast as the software business as a whole.
- In February 2012, the open source analyst firm Wikibon released the first market forecast for Big Data , listing $5.1B revenue in 2012 with growth to $53.4B in 2017
- The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020.