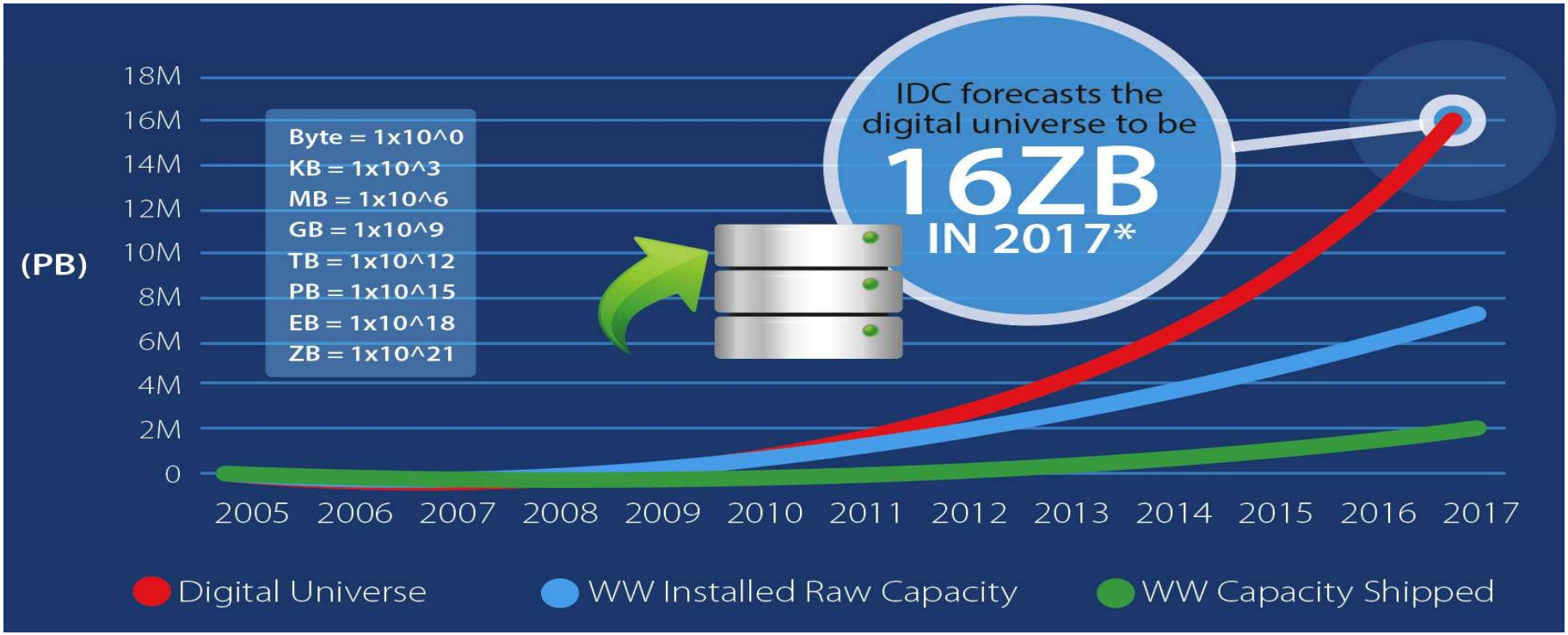
Introduction to Big Data

Tassadaq Hussain
www.tassadaq.pakistansupercomputing.com

Information Future Trend

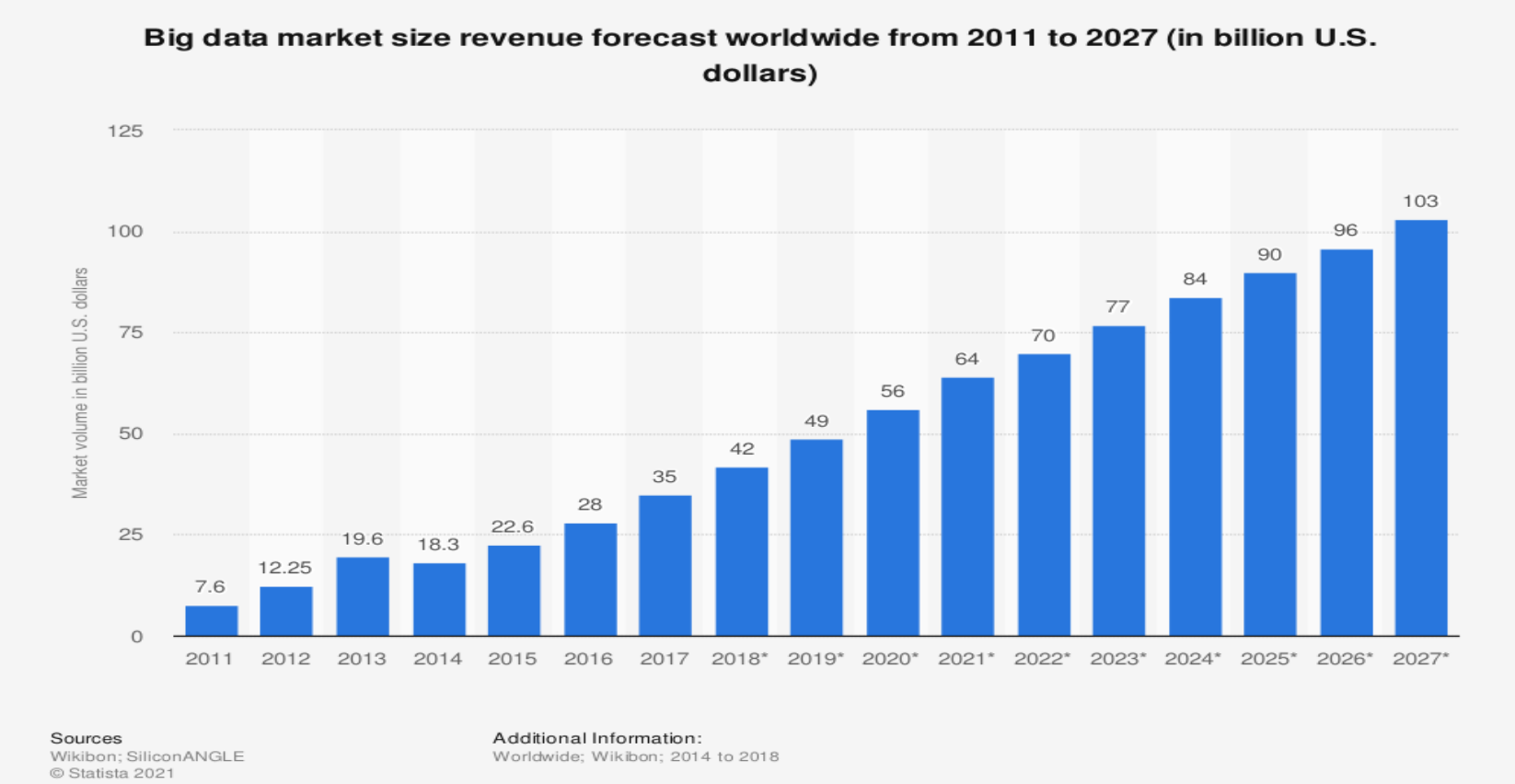
- Information Age
- Information doubling after every 18 months



Technology Research Gartner Inc. states that information data volume doubles after every 18 months.

Why Big Data

- Enormous generation of data
- New strategies to deal with the data
- Data management.



Types of Big Data

- STRUCTURED DATA
  - It particularly suited to further analysis because they are less complex with defined length, semantics, and format.
- UNSTRUCTURED DATA
  - lack a predefined data format and do not fit well into the traditional relational database systems
- SEMI-STRUCTURED DATA
  - combination of both structured and unstructured data. They still have the data organized in chunks, with similar chunks grouped together. However, the description of the chunks in the same group may not necessarily be the same.

Five V’s of Big Data

- VOLUME
- VELOCITY
  - Represents the generation and processing of in-flight transitory data within the elapsed time limit.
- VARIETY
  - Reveals heterogeneity of the data with respect to its type (structured, semi-structured, and unstructured), representation, and semantic interpretation.
- VERACITY
  - Relates to the uncertainty of data within a data set. As more data are collected, there is a considerable increase in the probability that the data are potentially inaccurate or of poor quality.
- VALUE
  - importance to Big Data analytics, because data will lose their meaning without contributing significant value

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

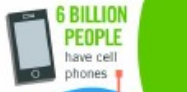
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Data Representation

- Databases
- Datasets
- Data types
- Data Structure

Some Applications

- Disease Patterns
- Shopping Patterns
- Sensor and Intelligent devices Data analytics
- Social Network associations and suggestions
- Predictive analytics
- Crime investigation

Steps

- Identify a problems (rice classification, human diagnosis etc.)
  - Present application in simple words
- Collect information
  - Signals, Tabular, Images or Videos of different classes of data
  - Clean and label them into folder as per number of classifications

Data Set information

- Features
  - Color
  - Behavior
  - Pattern
  - Shape
  - Correlation
  - etc