



BSM 562

BIG DATA

Lecture 1

Kevser Ovaz Akpınar, PhD

Agenda

- Grading scale
- Materials
- Course description
- Intro



Grading Scale

- Midterm 20%
 - Assignments 20 %
 - Quiz 10%
 - Final 50%
-
- Midterm and final will be comprehensive and cover the material from lectures and assignments.




Materials

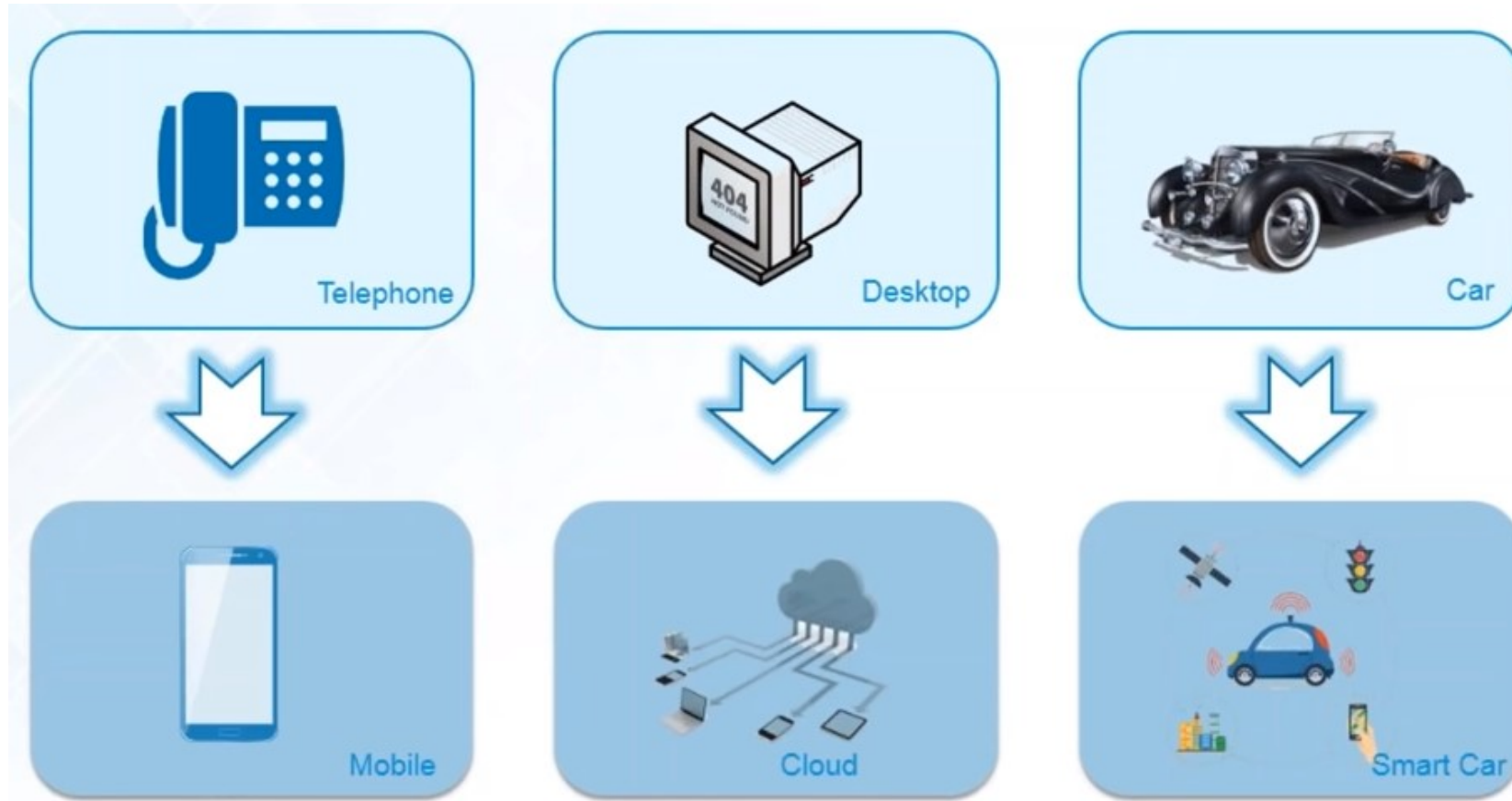
- Stanford University, Computer Science Course
 - CS246: Mining Massive Datasets
 - CS345A: Data MiningAnand Rajaraman and Jeffrey David Ullman
Available online at: <http://infolab.stanford.edu/~ullman/mmds.html>
- Brigham Young University, Big Data Science & CAPSTONE Course
Available online at: <http://bigdata.cs.byu.edu/>

Book

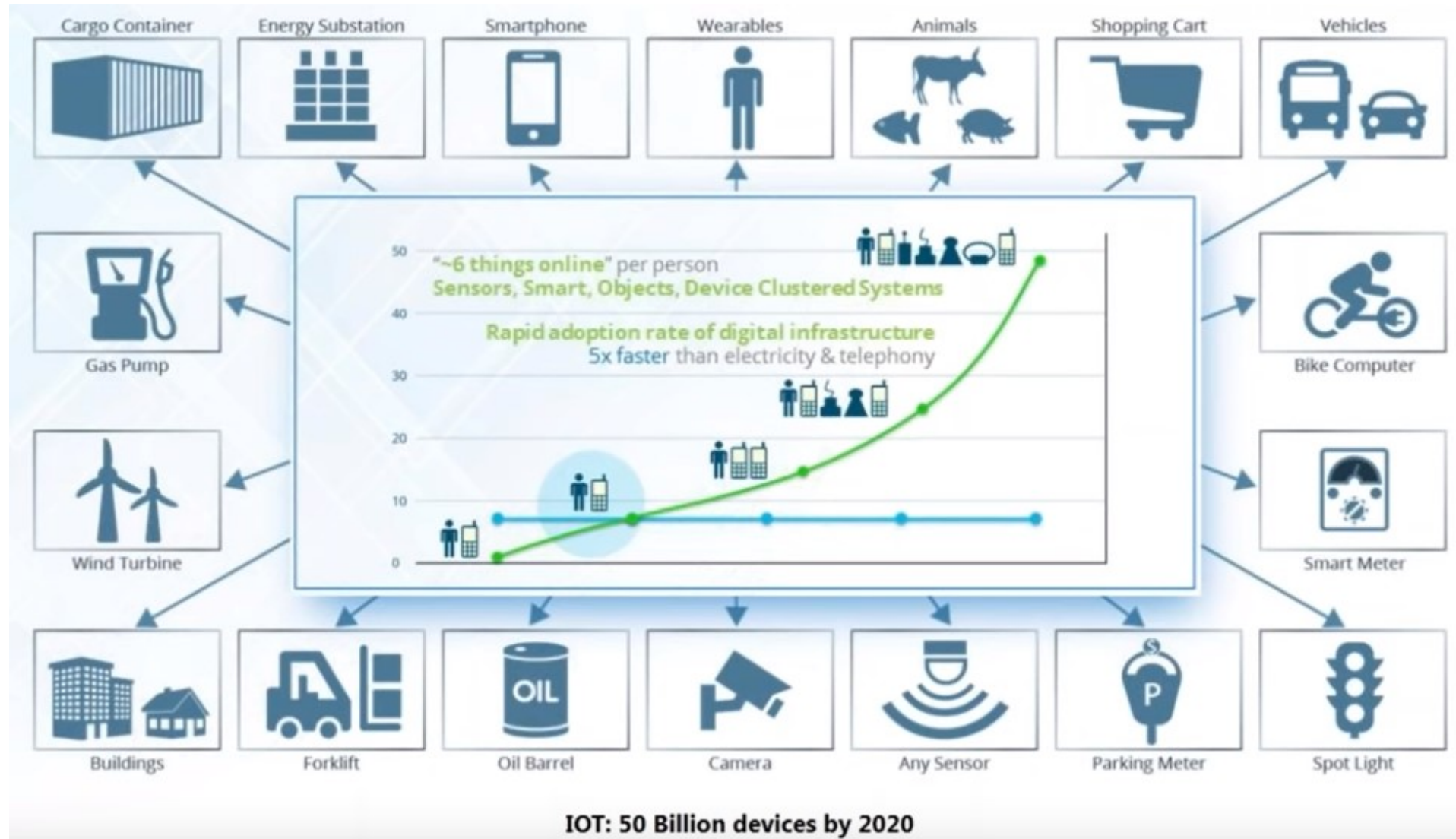
- 1- Hands-On Big Data Modeling, James Lee, Tao Wei, Suresh Kumar Mukhiya
- 2- An Introduction to Data Science (2013) 3rd edition , Jeffrey Stanton
- 3- Big Data Fundamentals, Thomas Erl, Wajid Khattak, Paul Buhler
- 4- R Programming for Data Science(2016), Roger D. Peng.
- 5- Big Data: Principles And Best Practices of Scalable Real-Time Data Systems, Nathan Marz with James Warren
- 6- Mining of Massive Datasets 2nd edition

Week	Description
1	Introduction to Big Data 
2	Python(Syntax, Dictionaries, Pandas, BeutifulSoup)
3	Text Mining with R
4	Big Data Adoption and Planning Considerations
5	Storing & Analyzing Big Data
6	Storage Technology & Analysis Techniques
7	Midterm 
8	Hadoop Implementations on Cloudera
9	Pig & Hive
10	MongoDb with Java in Big data Enrironments
11	Kafka & Zookeeper
12	Spark
13	Spark
14	Projects Usually with 3 options 

Evaluation of Technologies



IoT



WHAT

IS

BIG

DATA

?

3 Vs of Big Data

- **Volume:** Variety of sources, which may include business transactions, social media, and even information from sensors of organizations. Storing these data had been a problem in the past. But new technologies (such as Hadoop) have made this an easy task.
- **Velocity:** Velocity typically indicates the rapid speed at which data is transferred or received. This can be understood just by visualizing the amount of data—in terms of likes, comments, video uploads, tags etc.—that is handled by the social networking sites like Facebook in just one hour.
- **Variety:** Data can be seen in any type of formats. It can be in structured format, like the numeric data in traditional databases, or in unstructured format, such as, text, email, video, audio, or data from some financial transactions.



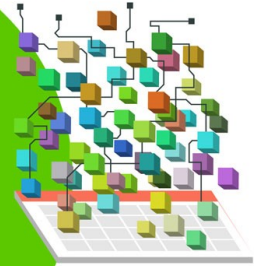
40 ZETTABYTES

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



Volume SCALE OF DATA

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



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

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



As of 2011, the global size of
data in healthcare was
estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**
are shared on Facebook
every month



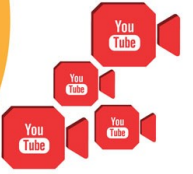
Variety DIFFERENT FORMS OF DATA

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



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



The New York Stock Exchange
captures

**1 TB OF TRADE
INFORMATION**
during each trading session



By 2016, it is projected
there will be

**18.9 BILLION
NETWORK
CONNECTIONS**

— almost 2.5 connections
per person on earth

Velocity ANALYSIS OF STREAMING DATA



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



**1 IN 3 BUSINESS
LEADERS**

don't trust the information
they use to make decisions



Poor data quality costs the US
economy around

\$3.1 TRILLION A YEAR

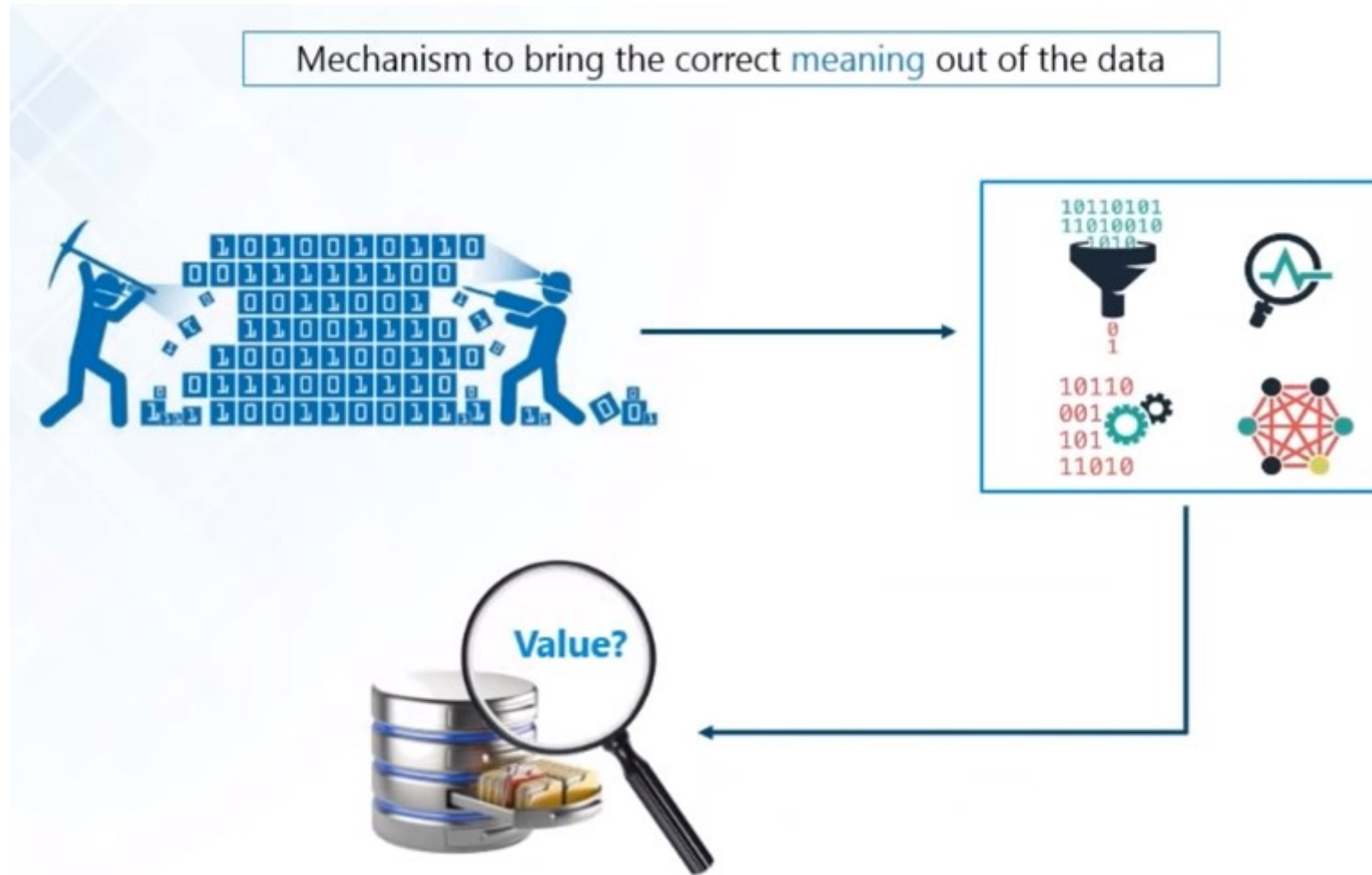


**27% OF
RESPONDENTS**

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

Veracity UNCERTAINTY OF DATA

5th V -> VALUE



6th V -> Visualization

- **Visualization:** Intelligibility of the data



Small Glims About Big Data

- Youtube users upload > 48 hours of video every minute
- Twitter 12 terabytes of Tweets every day
- Yahoo 60 PB, eBay: 40 PB of data
- Facebook: 1.39 billion monthly active users, 1.13 trillion Likes (Aug-2015)



Making Use of Big Data

- 1) Reduction in cost
- 2) Reduced production time
- 3) Development of new products
- 4) Smart decision-making

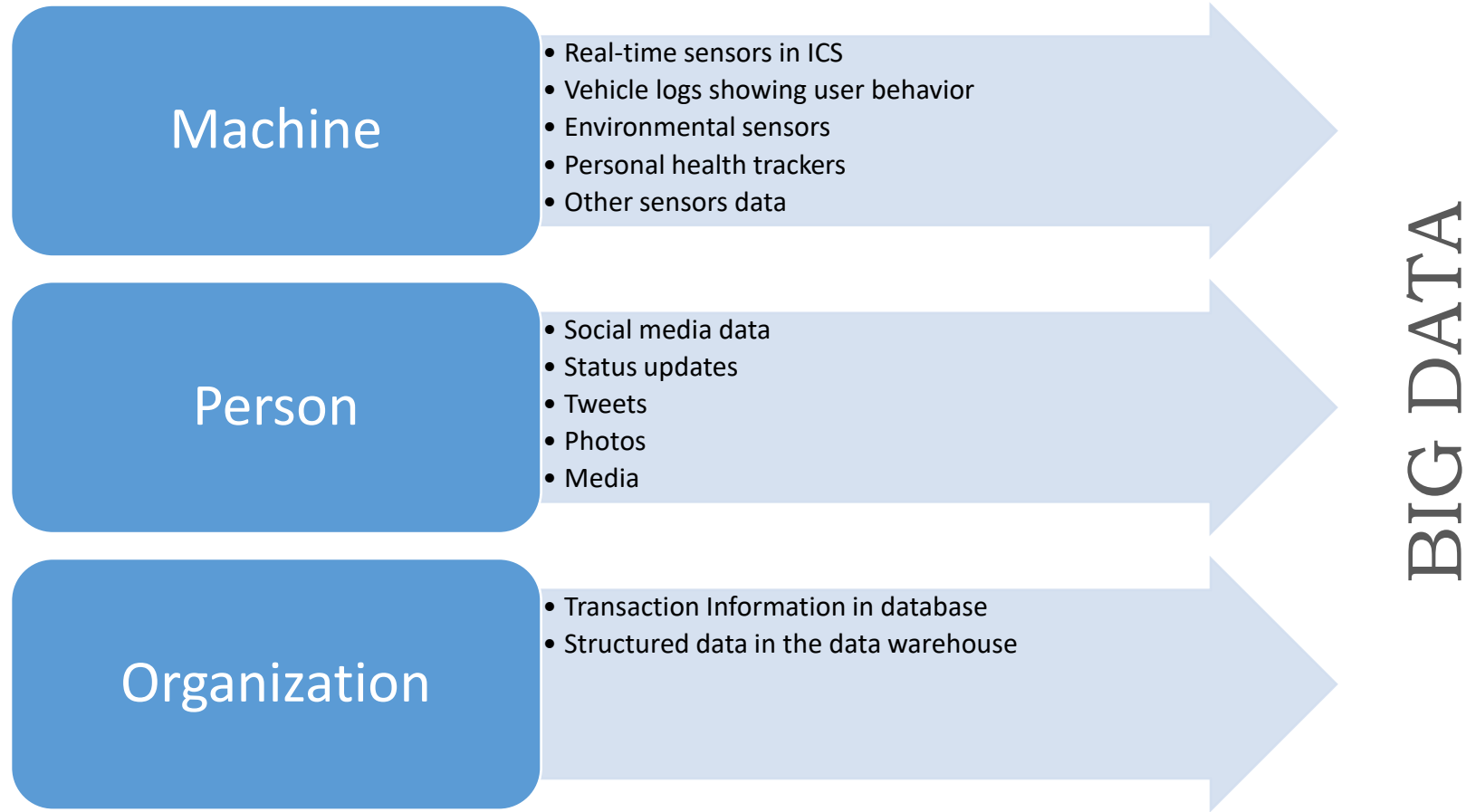


Where Does Big Data Come From

- Computer Generated
 - Application server logs (web sites, games, internet)
 - Sensor data (weather, atmospheric science, astronomy, smart grids)
 - Images/videos (traffic, security cameras, military surveillance)
- Human Generated
 - Blogs/reviews/emails/pictures/scientific research/medical records
 - Social graphs: facebook, contacts, twitter



Big Data Sources



Big Data Analytics

- Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses collecting, cleansing, organizing, storing, analyzing and governing data.



The symbol used to represent data analytics

Big Data Analytics Lifecycle

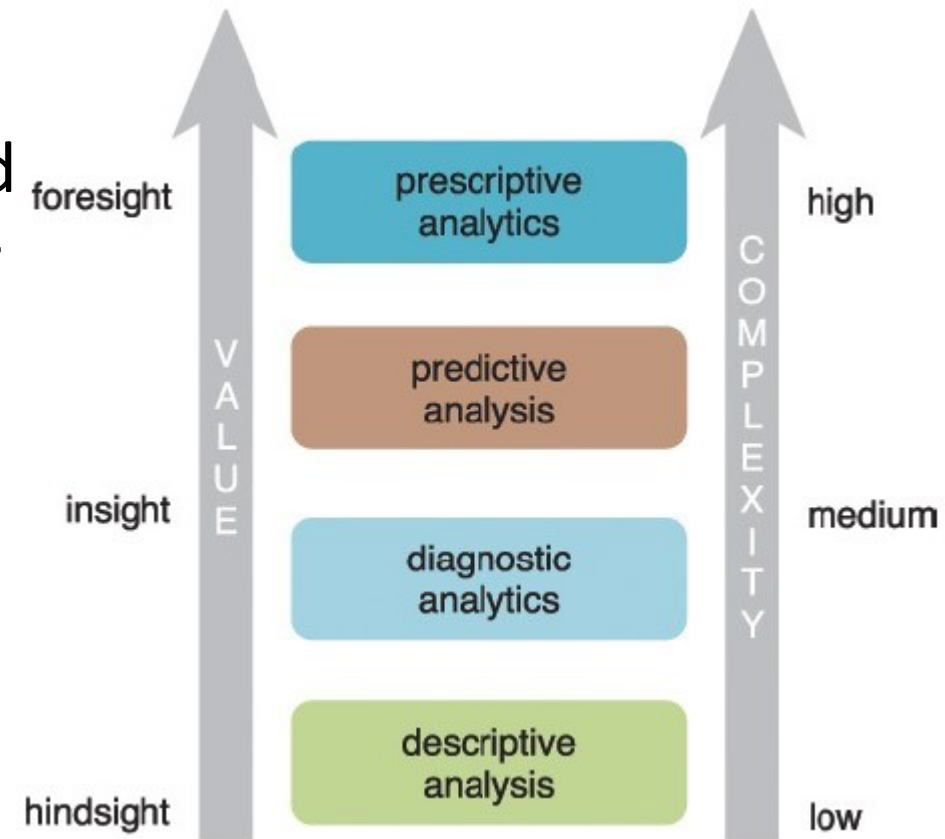
- Identifying
- Procuring
- Preparing
- Analyzing



Big Data Analytics Categories

Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone. There are four general categories of analytics that are distinguished by the results they produce:

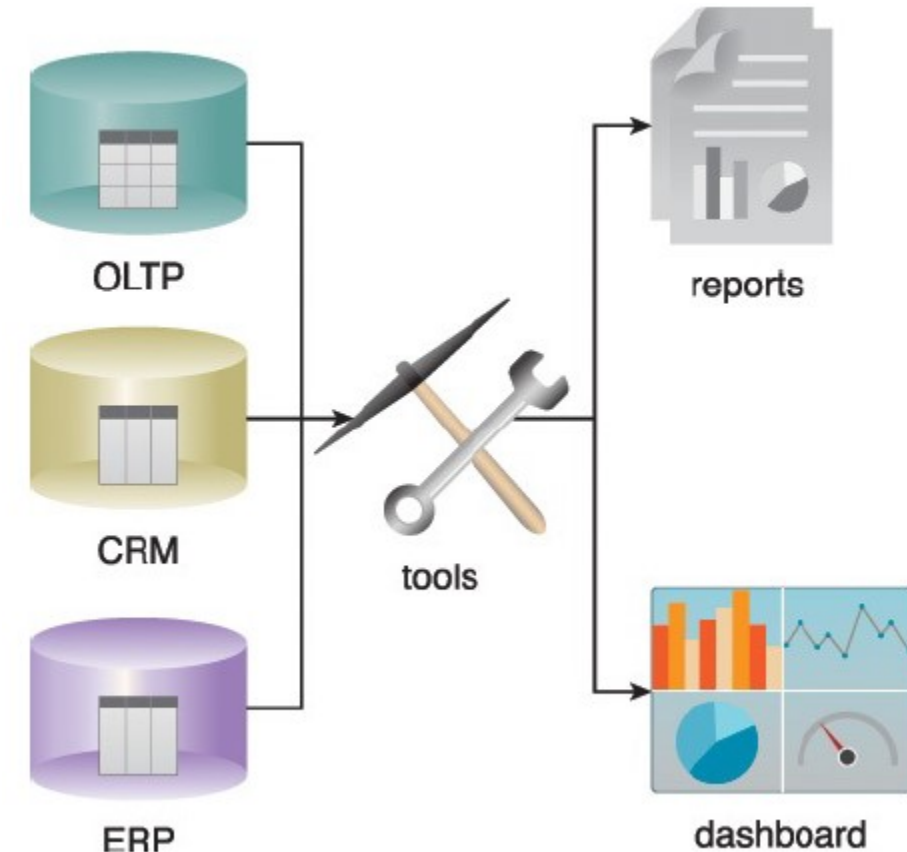
- descriptive analytics
- diagnostic analytics
- predictive analytics
- prescriptive analytics



Big Data Analytics Categories

1- Descriptive Analytics: Answer questions about events that have already occurred.

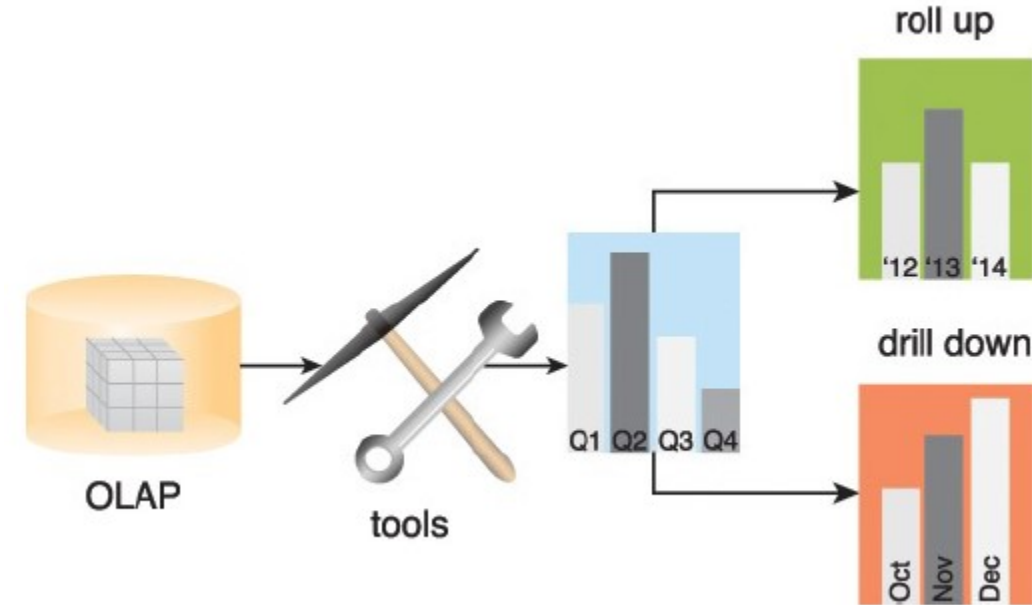
- What was the sales volume over the past 12 months?
- What is the number of support calls received as categorized by severity and geographic location?
- What is the monthly commission earned by each sales agent?



Big Data Analytics Categories

2- Diagnostic Analytics: Determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.

- Why were Q2 sales less than Q1 sales?
- Why have there been more support calls originating from the Eastern region than from the Western region?
- Why was there an increase in patient re-admission rates over the past three months?

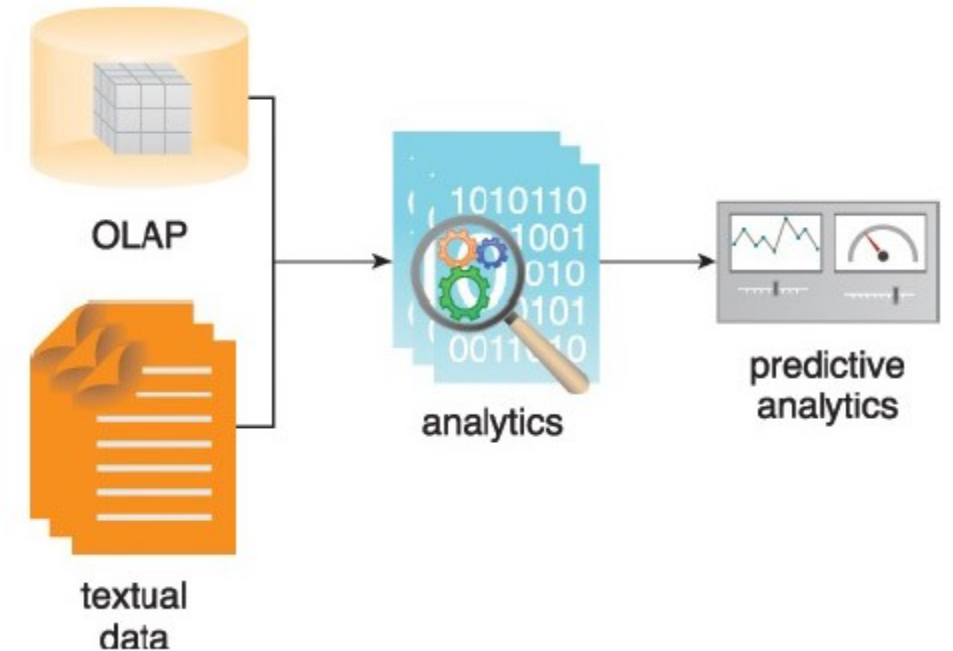


Diagnostic analytics can result in data that is suitable for performing drilldown and roll-up analysis

Big Data Analytics Categories

3- Predictive Analytics: Determine the outcome of an event that might occur in the future.

- What are the chances that a customer will default on a loan if they have missed a monthly payment?
- What will be the patient survival rate if Drug B is administered instead of Drug A?
- If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?

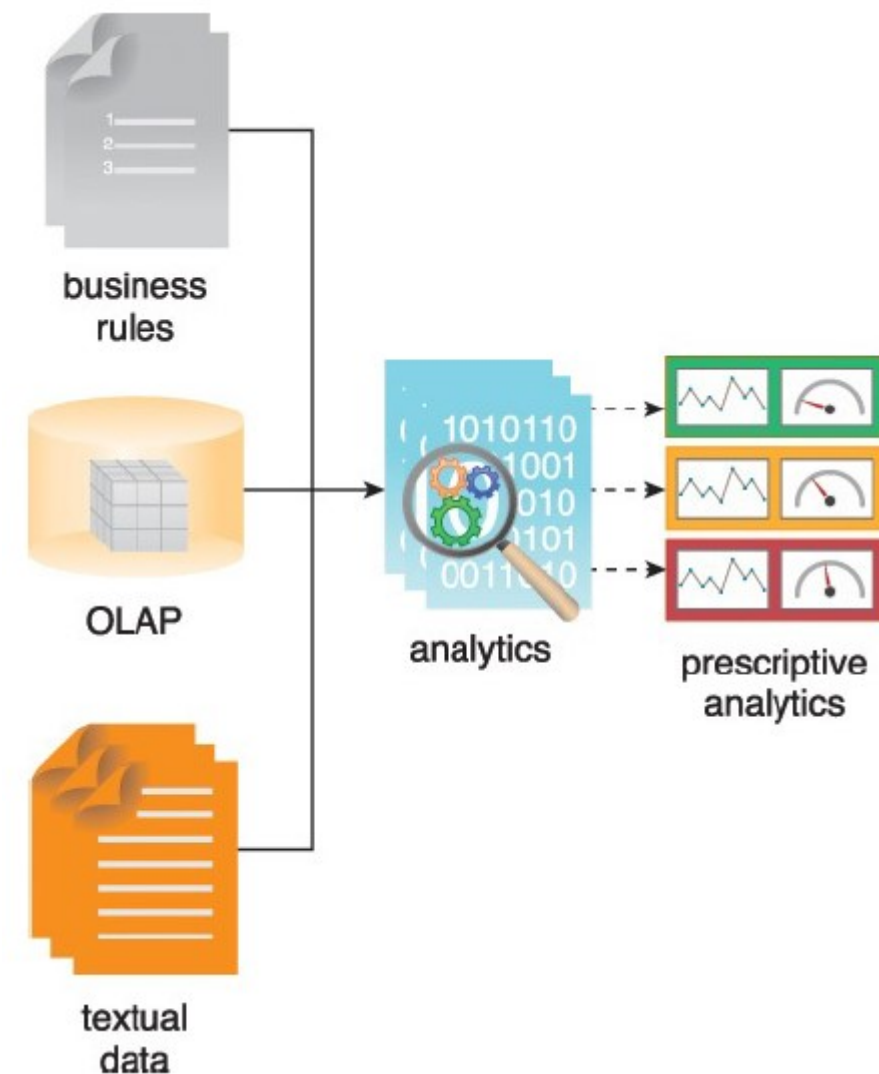


Big Data Analytics Categories

4- Prescriptive Analytics: It is build upon the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow, but why.

- Among three drugs, which one provides the best results?
- When is the best time to trade a particular stock?

Prescriptive analytics provide more value than any other type of analytics and correspondingly require the most advanced skillset, as well as specialized software and tools.



Big Data Analytics Categories



Descriptive

What happened in my business?

Comprehensive, accurate and effective visualization



Diagnostic

Why it has happened in my business?

Ability to drill down to the root cause



Predictive

What will happen in future based on past trends?

Historical patterns being used to predict specific outcomes using algorithms



Prescriptive

What should be done ?

Applying advanced analytical algorithms to make specific recommendations and strategies.

Big Data Applications

- Education
- Healthcare
- Government
- Entertainment and Media
- Weather
- Transportation
- Banking



Big Data in Education

•Customized and Dynamic Learning Programs

Customized programs and schemes to benefit individual students can be created using the data collected on the bases of each student's learning history. This improves the overall student results.

•Reframing Course Material

Reframing the course material according to the data that is collected on the basis of what a student learns and to what extent by real-time monitoring of the components of a course is beneficial for the students.

•Grading

New advancements in grading systems have been introduced as a result of a proper analysis of student data.

•Career Prediction

Appropriate analysis and study of every student's records will help understand each student's progress, strengths, weaknesses, interests, and more. It would also help in determining which career would be the most suitable for the student in future



[IntelliPaat]

Big Data in Healthcare

- No unnecessary diagnosis -> treatment cost reduce
- Epidemic prediction and help to decide preventative measures
- Detection of diseases in early stage
- Medical results of past medicines -> evidence-base more accurate prescription



[IntelliPaat]

Big Data in Government

Welfare Schemes

- In making faster and informed decisions regarding various political programs
- To identify areas that needs attention
- To stay up to date in the field of agriculture by keeping track of all existing land and livestock
- To overcome national challenges such as unemployment, terrorism, energy resources exploration..

Cyber Security

- Deceit recognition
- To catch tax evaders



[IntelliPaat]

Big Data in Entertainment and Media

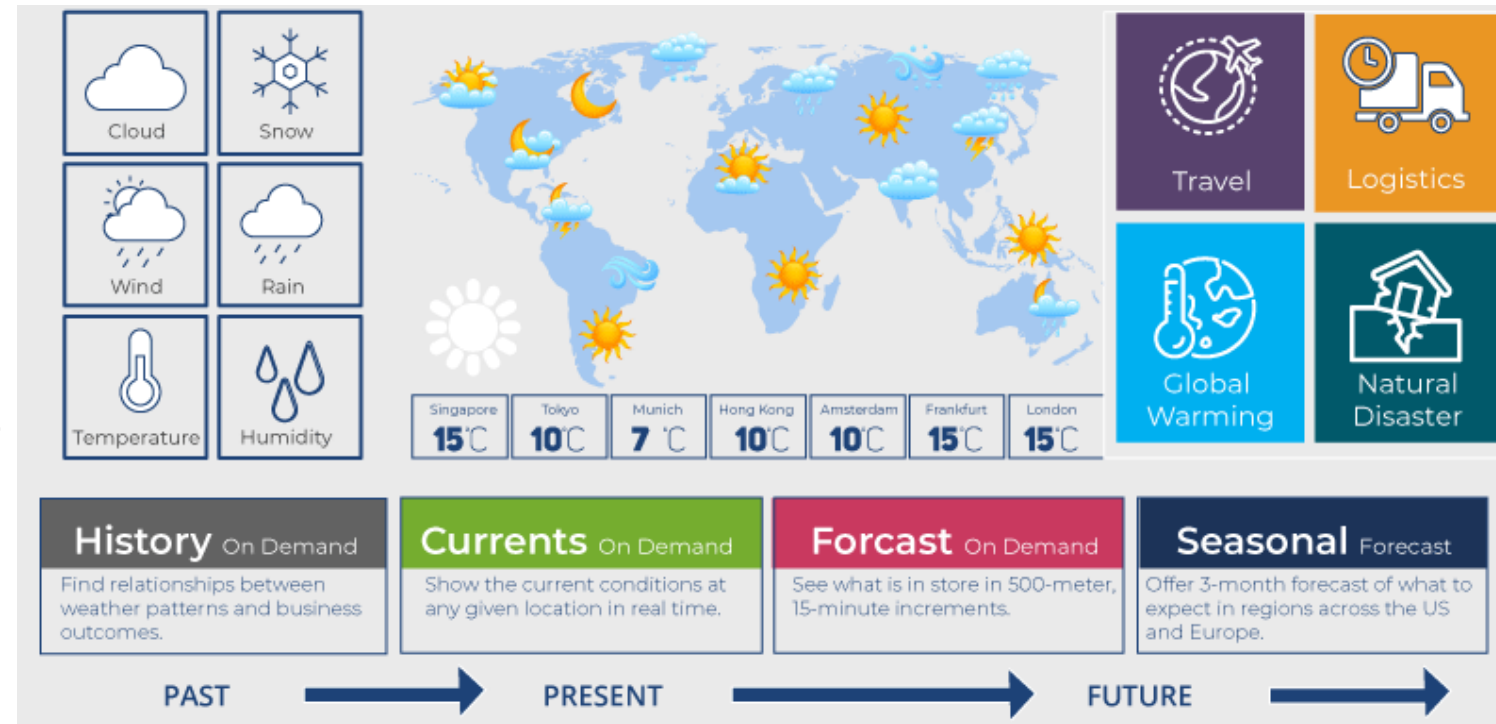
- Predicting the interests of audiences
- Effective targeting of the advertisements
- Optimized or on-demand scheduling of media streams in digital media distribution platforms
- Getting insights from customer reviews



[IntelliPaat]

Big Data in Weather

- Forecasting
- To study global warming
- In understanding the patterns of natural disasters
- To make necessary preparations in the case of crises
- To predict the availability of usable water



[IntelliPaat]

Big Data in Transportation

- Route planning:** To estimate users' needs on different routes and on multiple modes of transportation and then utilize route planning to reduce their wait time

- Congestion management and traffic control:** Real-time estimation of congestion and traffic patterns. E.g. using Google Maps to locate the least traffic-prone routes

- Safety level of traffic:** To use the real-time processing of big data and predictive analysis to identify accident-prone areas and help reduce accidents



[IntelliPaat]

Big Data in Banking

- Misuse of credit/debit cards
- Venture credit hazard treatment
- Business clarity
- Customer statistics alteration
- Money laundering
- Risk mitigation



[IntelliPaat]

Big Data Challenges

- Capturing data
- Storing data
- Searching data
- Sharing data
- Transferring data
- Analysis of the previously stored data
- Presentation



Big Data Challenges

Storing exponentially growing huge datasets

- Data generated in past 2 years is more than the previous history in total
- By 2020, total digital data will grow to 44 zettabytes approximately
- By 2020, about 1.7MB of new info will be created every second for every person



Big Data Challenges

Processing data having complex structure

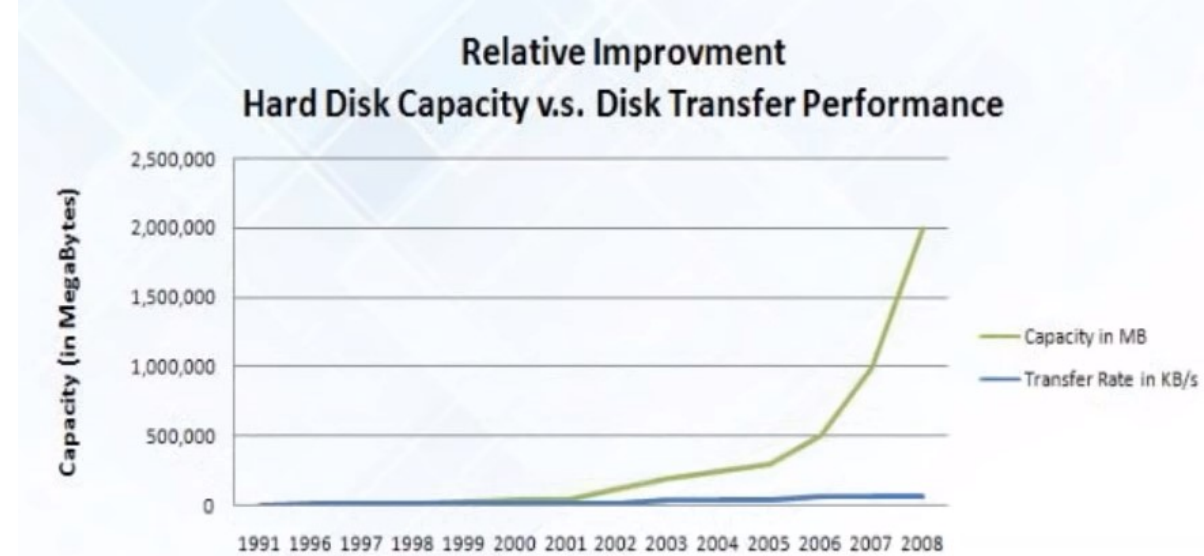
- Structured, semi structured, unstructured



Big Data Challenges

Processing data faster

- The data is growing at much faster rate than that of disk read/write speed
- Bringing huge amount of data to computation unit becomes a bottleneck



Types of Big Data

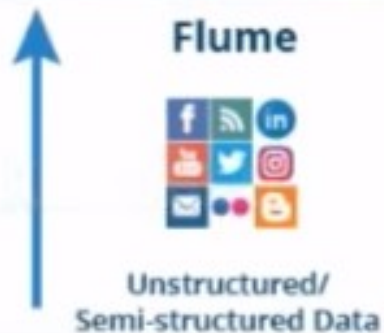
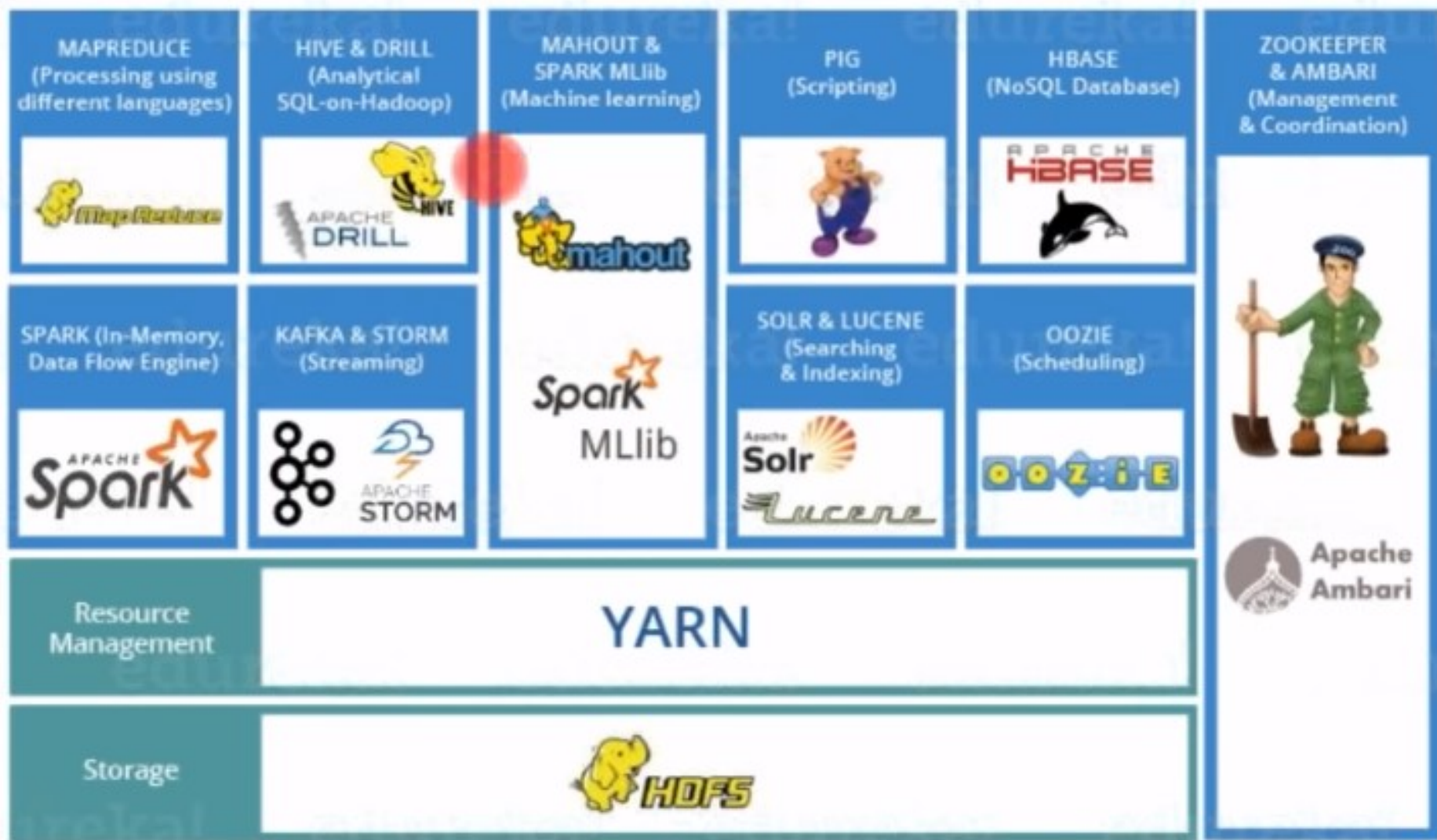
- Structured Data
- Unstructured Data
- Semi-structured Data



How is Big Data Processed?

- Parallel processing
- Interconnected systems





THANK YOU

QUESTIONS

ADDITIONAL REFERENCES

<https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/>

<https://www.edureka.co/hadoop>

<https://www.greycampus.com/opencampus/big-data-developer>

<https://www.datasciencecentral.com/profiles/blogs/what-is-big-data>

What Does Big Data Analytics Mean Today?, Tableau Conference, 2018

Big Data Tutorial For Beginners, Edureka