BIG DATA 101- IBM

DESCRIPTION

Module 1 - What is Big Data?

- Characteristics of Big Data
- What are the V's of Big Data?
- The Impact of Big Data

Module 2 - Big Data - Beyond the Hype

- Big Data Examples
- Sources of Big Data
- Big Data Adoption

Module 3 - The Big Data and Data Science

- The Big Data Platform
- Big Data and Data Science
- Skills for Data Scientists
- The Data Science Process

Module 4 - Big Data Use Cases

- Big Data Exploration
- The Enhanced 360 View of a Customer
- Security and Intelligence
- Operations Analysis

Module 5 - Processing Big Data

- Ecosystems of Big Data
- The Hadoop Framework

1. 1 CHARACTERISTICS OF BIG DATA

What is Big Data?

"The basic idea behind the phrase **'Big Data'** is that everything we do is increasingly leaving a digital trace (or data), which we can use and analyze to become smarter The driving forces in this brave new world are access to ever-increasing volumes of data and our ever-increasing technological capability to mine that data for commercial insights."

Bernard Marr

"**Big Data** refers to the dynamic, large and disparate volumes of data being created by people, tools and machines; it requires new, innovative and scalable technology to collect, host and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management and enhanced shareholder value."

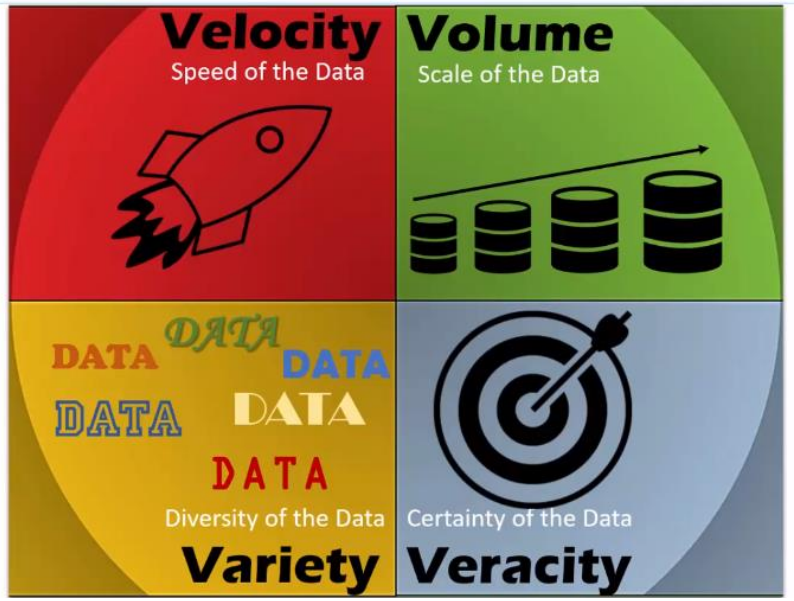
EY

"**Big Data** is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

Gartner

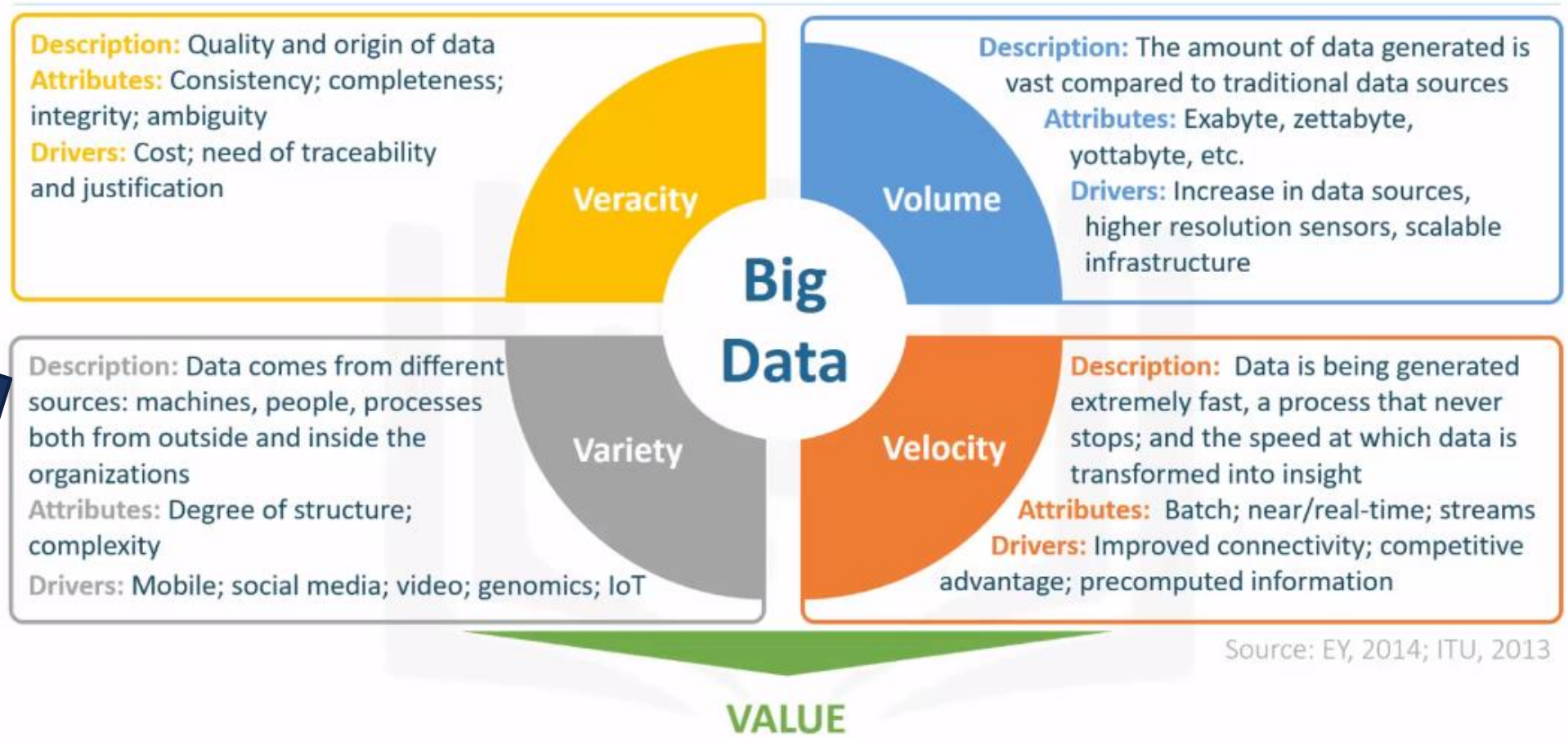
"**Big Data** is a collection of data from traditional and digital sources inside and outside your company that represents a source for ongoing discovery and analysis."

Lisa Arthur



1.2 WHAT ARE THE V'S OF BIG DATA?

80% of data is considered to be unstructured and we must devise ways to produce reliable and accurate insights.
The data must be categorized, analyzed and visualized.



Source: EY, 2014; ITU, 2013

I.3 THE IMPACT OF BIG DATA

Recommendation Engines

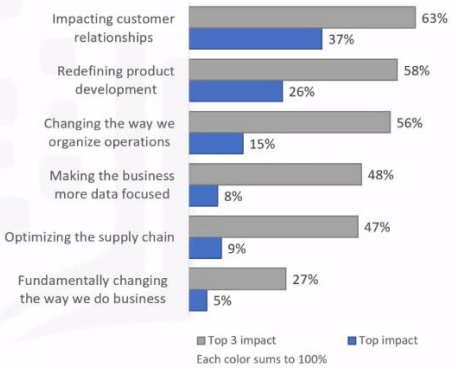
amazon.com

Amazon's recommendations are based on what the user has bought in the past, the items he/she has in the virtual shopping cart, items he/she has rated and liked, and what other customers have viewed and purchased.

NETFLIX Spotify

Big Data will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus

Big Data will fundamentally change the way businesses compete and operate



Online streaming – Netflix

Data:
>> Time of day when movies/shows are watched
>> Record of when users pause, rewind and fast forward
>> User ratings
>> Searches
>> Type of device

HOUSE of CARDS

E-commerce in China

Facts:
>> E-commerce accounts for more than 13% of China's total retail sales of consumer goods
>> About 90% of Internet users and 70-80% of consumers as a whole are shopping online in top tier cities
>> Chinese consumers often visiting 4 to 5 sites before reaching a purchase decision

淘宝网 Taobao.com

The Internet of Things (IoT)

Virtual personal assistants

Google Now

Siri knows what users mean when they ask her questions, she knows where they are, what time they are talking about and can use this information to look for restaurants of a particular type of food and check whether there reservations are available.

Google Now makes recommendations before users ask for them, especially when it is linked up to the user's calendar and location sensing is enabled on the user's phone. Google Now knows where the user is and where he/she needs to be, it can tell users about things like traffic or the weather before you even ask for it.

2. BIG DATA - BEYOND THE HYPE

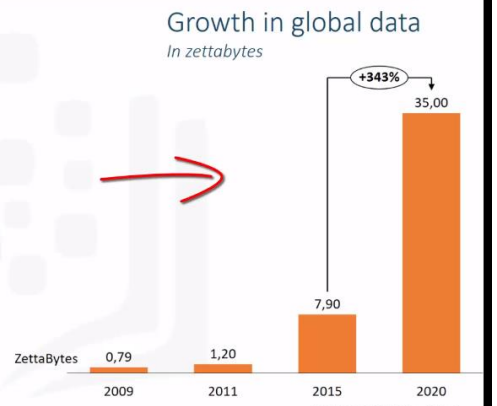
2

Why is everyone talking about Big Data?

Facts:

- More data has been created in the past two years than in the entire previous history of the human kind.
- By 2020, about 1.7 megabytes of new information will be created every second for every human being in the world.
- By 2020, the data we create and copy will reach around 35 zettabytes, up from only 7.9 zettabytes today.

Sources: McKinsey report – Digital era, data growth, estimates and, Gartner, IDC estimates



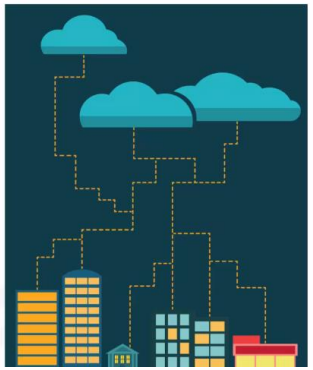
35 zettabytes by 2020!

- 10% by machines.
- Emerging markets to produce most of the world's data.

Advances in cloud computing have contributed to the increasing potential of Big Data

According to McKinsey (2013), the emergence of cloud computing has highly contributed to the launch of the Big Data era:

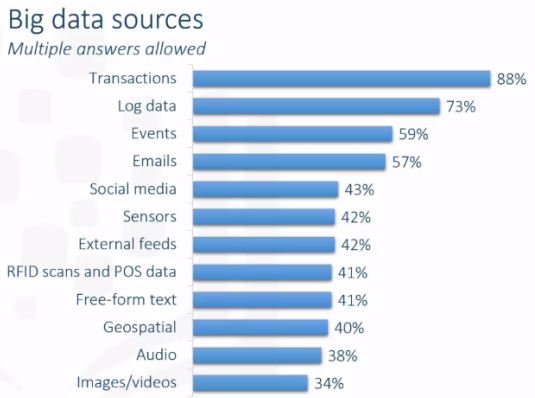
- Cloud computing allows users to access highly scalable computing and storage resources through the Internet.
- By using cloud computing, companies can use server capacity as needed and expand it rapidly to the large scale required to process big data sets and run complicated mathematical models.
- Cloud computing lowers the price to analyze big data as the resources are shared across many users, who pay only for the capacity they actually utilize.

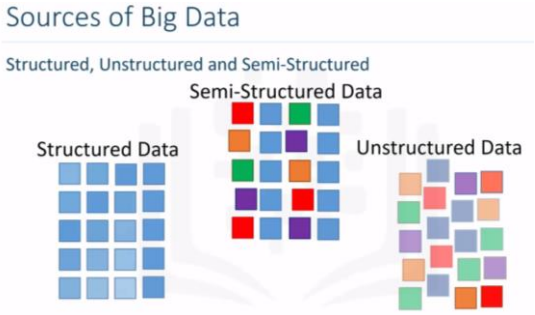


Where is all the data coming from?

There are three major sources of Big Data:

- People-generated data
- Machine-generated data
- Business-generated data





3. THE BIG DATA AND DATA SCIENCE

3

Key Aspects of a Big Data Platform

Integration

Analysis

Visualization

Workload Optimization

Security

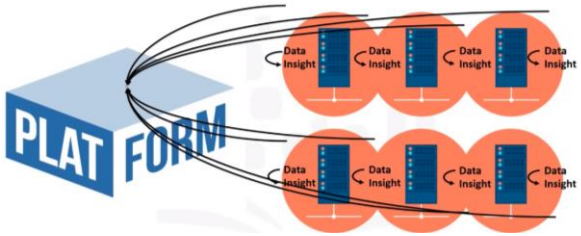
Governance

Big Data Skills

Learn about the different applications involved in a Big Data workflows, storage, analytics, and more.

Governance for Big Data

Integration



Big Data Skills – Demand

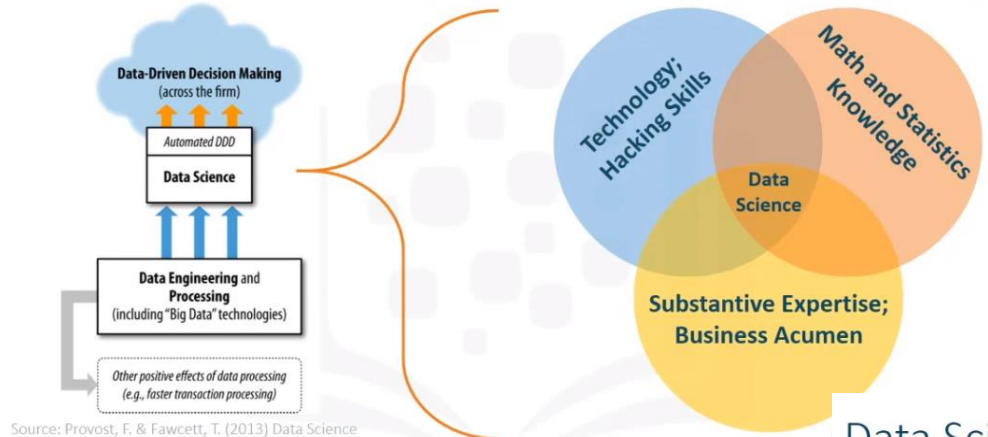
Big Data skills are in high demand.

Security and Governance

Visualization

3.1 THE DATA SCIENCE PROCESS

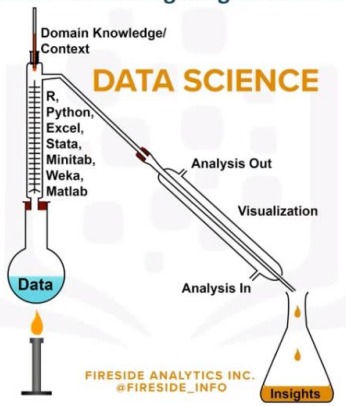
How does Big Data relate to Data Science?



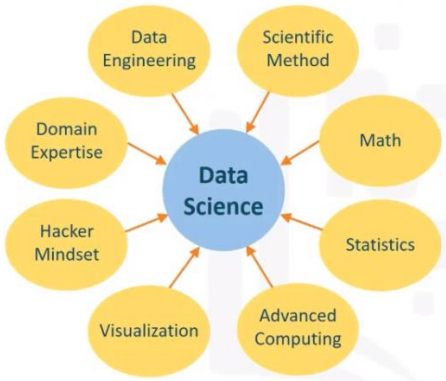
Source: Provost, F. & Fawcett, T. (2013) Data Science for Business. Sebastopol, CA: O'Reilly Media

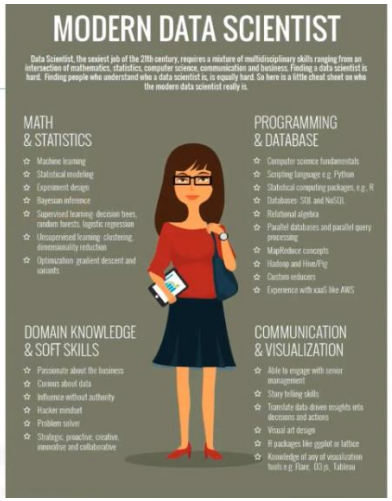
How can we make sense of Big Data?

Data science is the process of distilling insights from data to inform decisions.



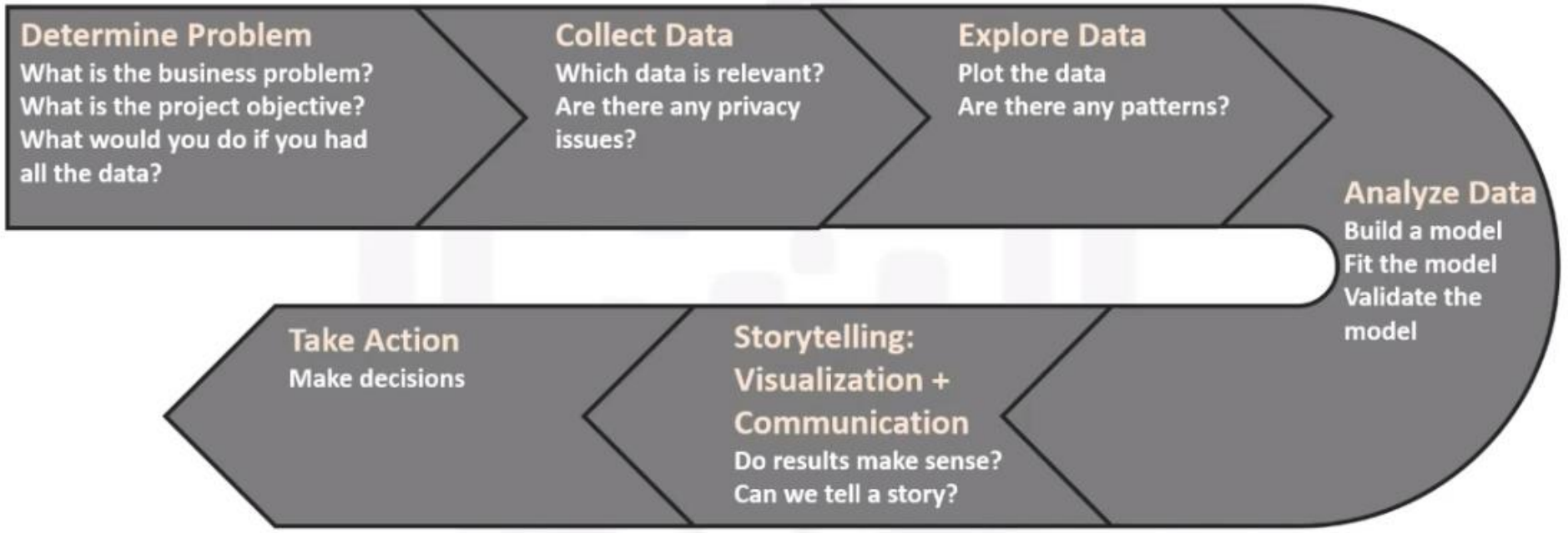
Data Scientist Skills



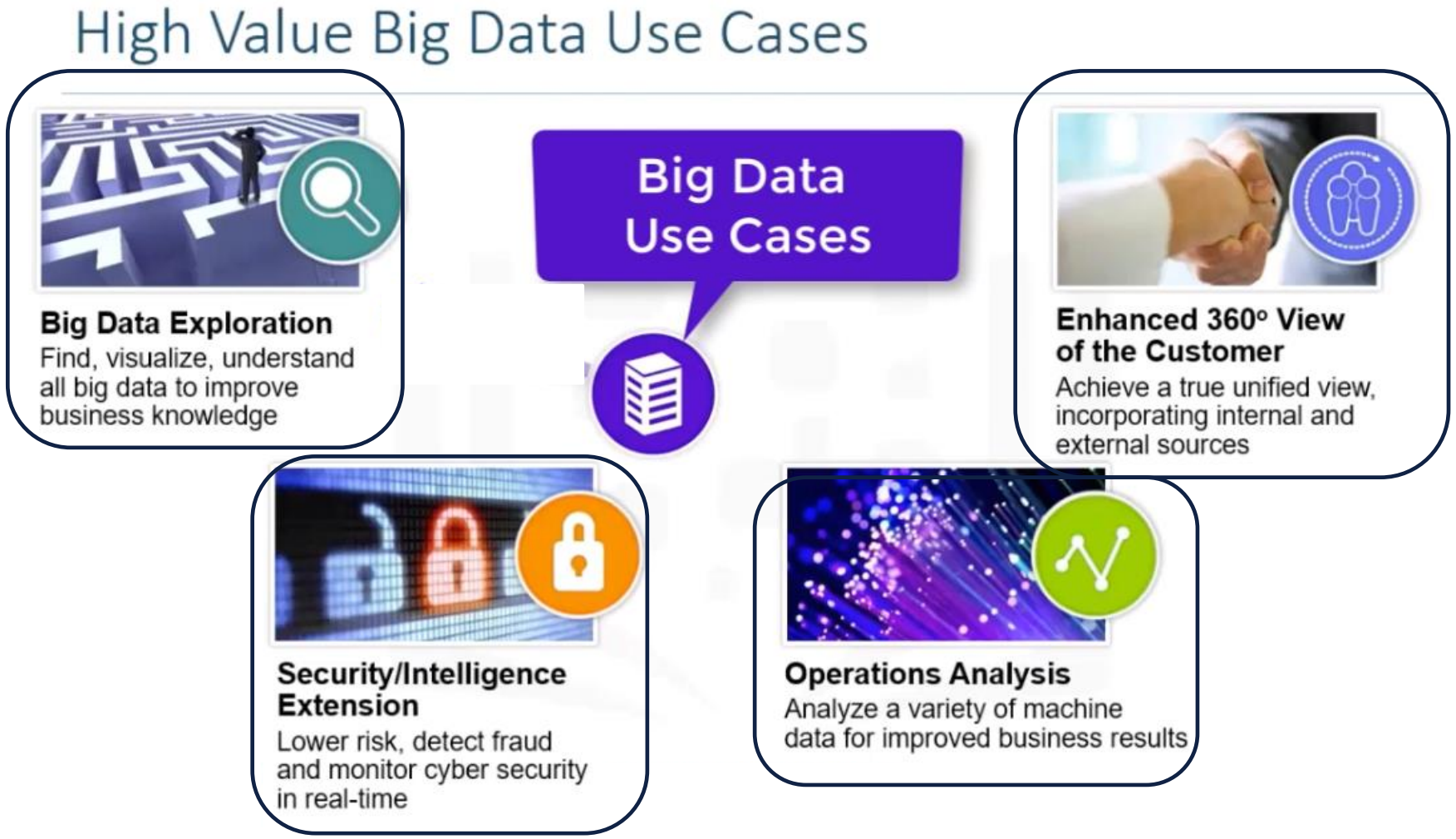


How can we make sense of Big Data?

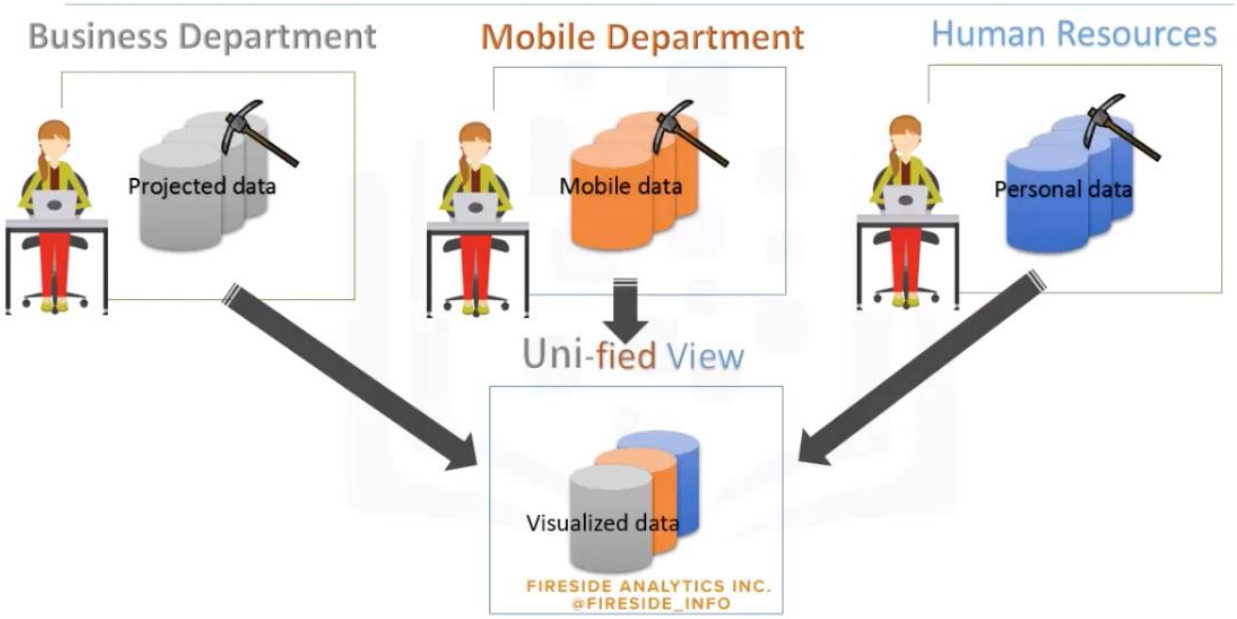
The data science process:



4. BIG DATA USE CASES



4.1 BIG DATA EXPLORATION



Example

Reducing Traffic Congestion

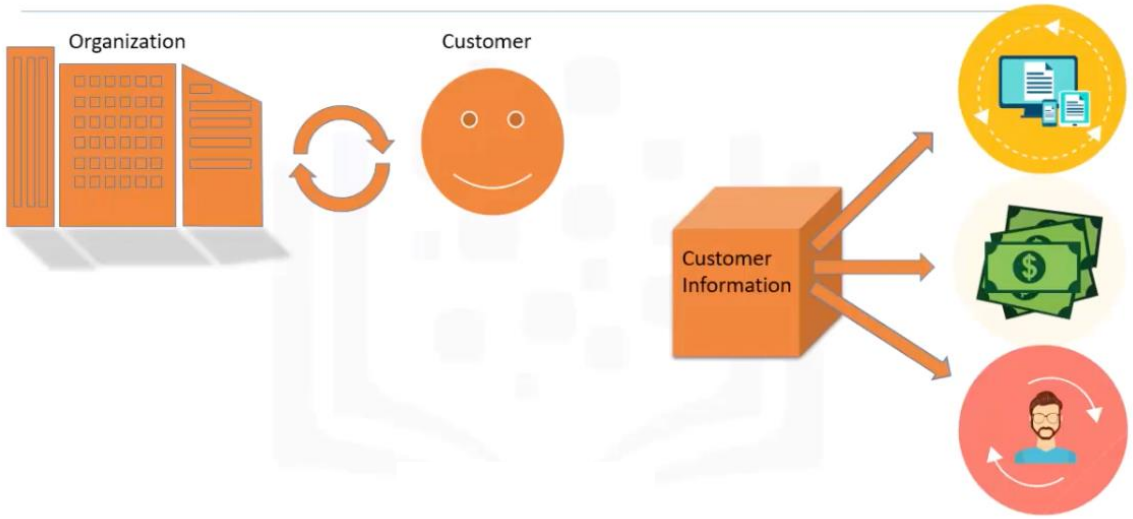
- Deployed real-time Smarter Traffic system to predict and improve traffic flow
- Analyzed streaming real-time data gathered from cameras at entry/exit to city, GPS data from taxis and trucks, and weather information

Results

- Enables ability to analyze and predict traffic faster and more accurately than ever before
- Provides new insight into mechanisms that affect a complex traffic system
- Smarter, more efficient, and more environmentally friendly traffic

4.2 THE ENHANCED 360 VIEW OF A CUSTOMER

Enhanced 360° View of the Customer



Example

Retail: Improving Customer Relations

- Prepare for increased customer rush times
- Stock up customer preferences
- Plan for customer spending habits (sales, coupons, product exposure)

Results

- Increased revenue
- Increased efficiency
- Build customer loyalty

4.3 SECURITY AND INTELLIGENCE

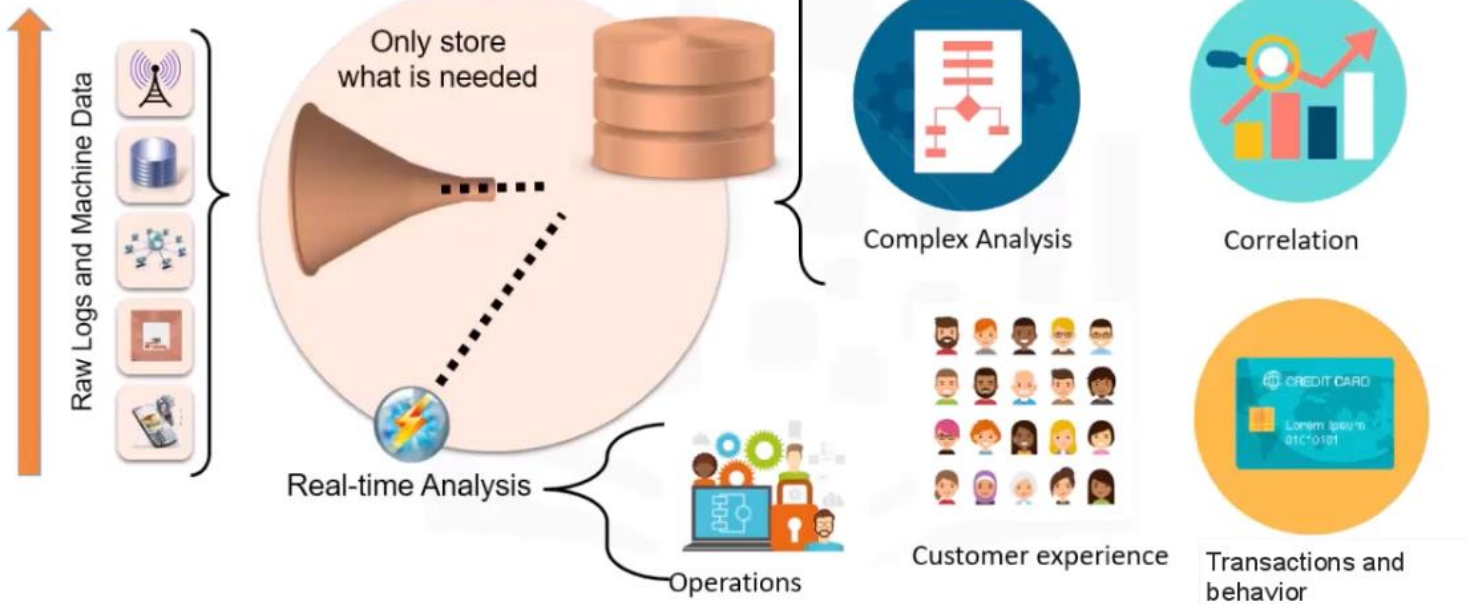
Improving Security

Analyze under-leveraged data:

- Emails
- Social media
- Search results
- Prepare for threats

The growing number of high-tech crimes,
cyber-based terrorism, espionage, computer intrusions,
and major cyber fraud cases,
poses a real threat to every individual and organization.
To meet these security challenges,
businesses are using big data technologies
to change and enhance their cyber security
and intelligence activities, how?
By processing and analyzing new data types,
such as social media, emails,
and analyzing hours and hours of video footage.
Analyzing data in motion, and at rest,
can help find new associations,
or uncover patterns and facts to significantly improve
intelligence, security, and law enforcement.

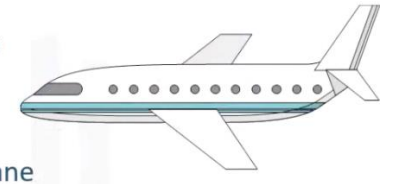
4.4 OPERATIONS ANALYSIS



Examples

Aviation: Improving Security

- Collect and analyze massive amounts of data from a plane's turbine, sensors, and GPS.
- Visualize data with complex big data analysis



Results

- Real-time Visibility into Operations of the plane
- Improved customer experience
- Increased efficiency of plane through optimization of fuel usage
- Real-time analysis for irregularities that may occur

Walmart

Objective

"We want to know what every product in the world is. We want to know who every person in the world is. And we want to have the ability to connect them together in a transaction."
– **Neil Ashe, CEO of Global E-commerce at Walmart**

Data

Walmart collects 2.5 petabytes of information from 1 million customers every hour.

Big Data Applications

Mining sales data is helping Walmart find patterns that can be used to provide personalized product recommendations and promotional offers.

Walmart is working to better understand the context of what its customers are saying and doing by **connecting in-store and online customer behavior** using data from sources such as clickable actions on its website, contact information from its e-club, and point-of-sale transactions. This provides Walmart the context it needs to send tailored offers and messages at scale.

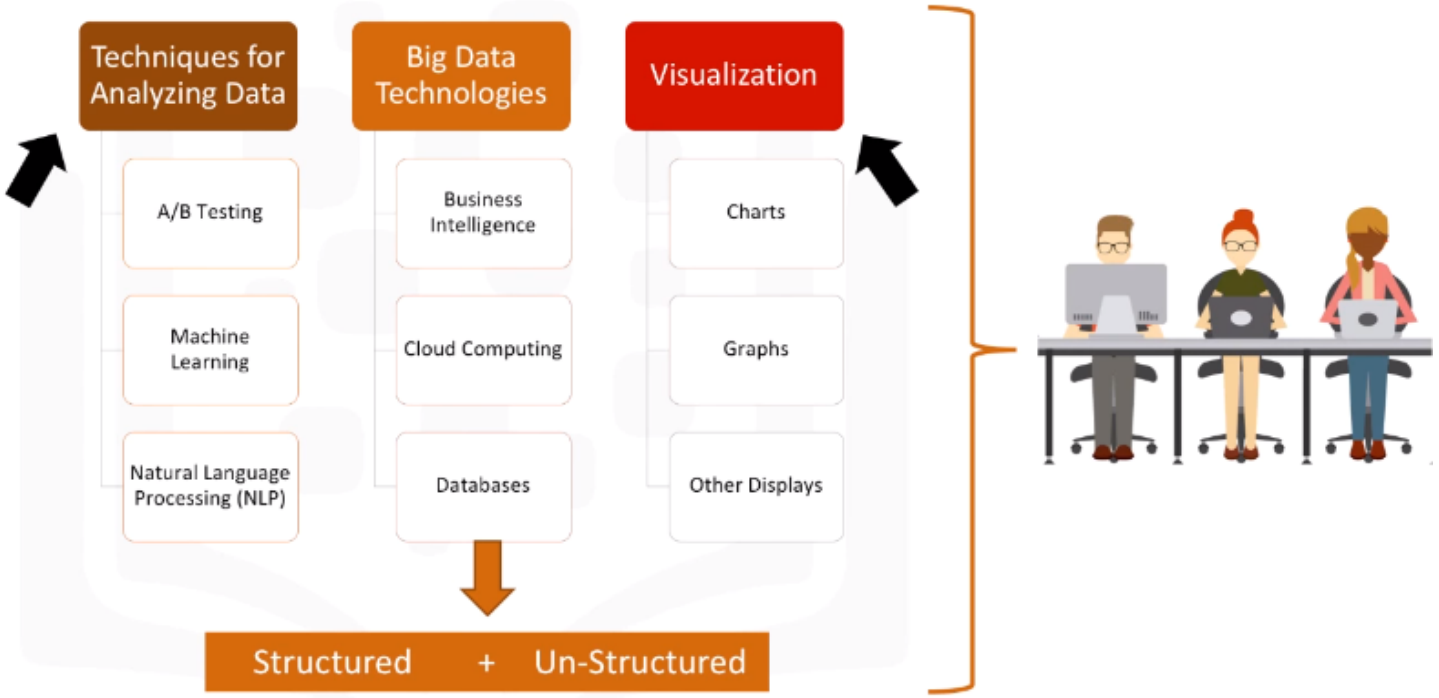
Walmart's system for **personalized email marketing** is more sophisticated than that of most of its retail competitors, Amazon being and exception.

5. PROCESSING BIG DATA

5

Components and Ecosystems of Big Data

Source: "Big data: The next frontier for innovation, competition, and productivity" McKinsey Global Institute, June, 2011



5.1 BIG DATA TECHNOLOGIES

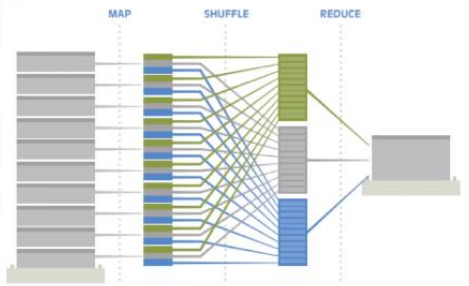
There are a number of vendors in this space

cloudera
Data Scientist Workbench
EMC²
Microsoft
amazon web services™
BIG DATA UNIVERSITY
Hortonworks
coursera
MAPR
IBM
DATASTAX

What is the Hadoop framework?

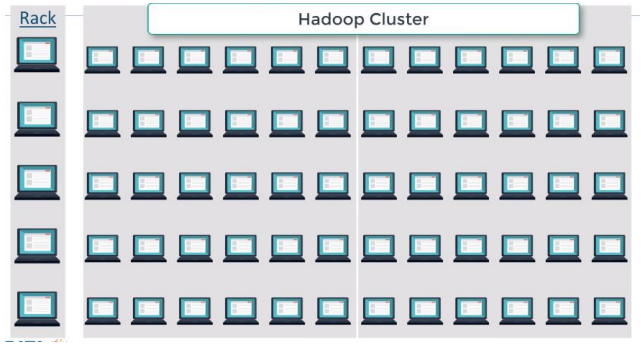
Hadoop is an open-source software framework used to store and process huge amounts of data. It is implemented in several distinct, specialized modules:

1. Storage, principally employing the Hadoop File System (HDFS)
2. Resource management and scheduling for computational tasks
3. Distributed processing programming model based on MapReduce
4. Common utilities and software libraries necessary for the entire Hadoop platform



Source: The Executive's Guide To Big Data & Apache Hadoop

5.2 THE HADOOP FRAMEWORK



Source: IBM Watson Analytics - www.ibm.com/analytics/us/en/technology/hadoop

- Apache™ Hadoop® is a highly scalable storage platform designed to process very large data sets across hundreds to thousands of computing nodes that operate in parallel
- Hadoop provides a cost effective storage solution for large data volumes with no format requirements
- MapReduce, the programming paradigm that allows for this massive scalability, is the heart of Hadoop

The term MapReduce refers to two separate and distinct tasks that Hadoop programs perform:

- *Map & Reduce*

Hadoop Strategy

Source: IBM Watson Analytics - www.ibm.com/analytics/us/en/technology/hadoop

1. Choosing recommended distributions
2. Maturing the environment with modernized hybrid architectures
3. Adopting a data lake strategy based on Hadoop technology

"Data lakes are a method of storing data that keep vast amounts of raw data in their native format and more horizontally to support the analysis of originally disparate sources of data."

Data Warehouses

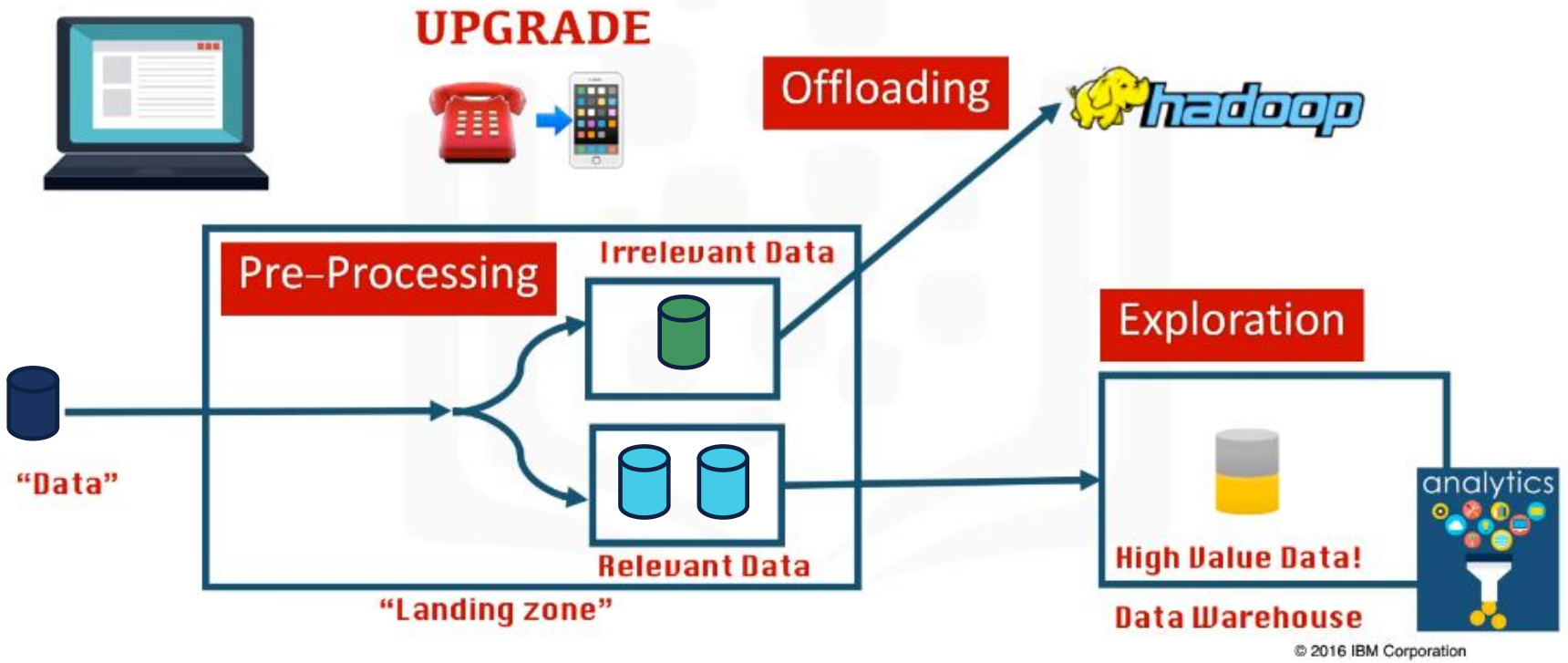
Source: IBM Watson Analytics - www.ibm.com/analytics/us/en/technology/hadoop

"Big Data is best thought of as a platform"

Data Warehouses

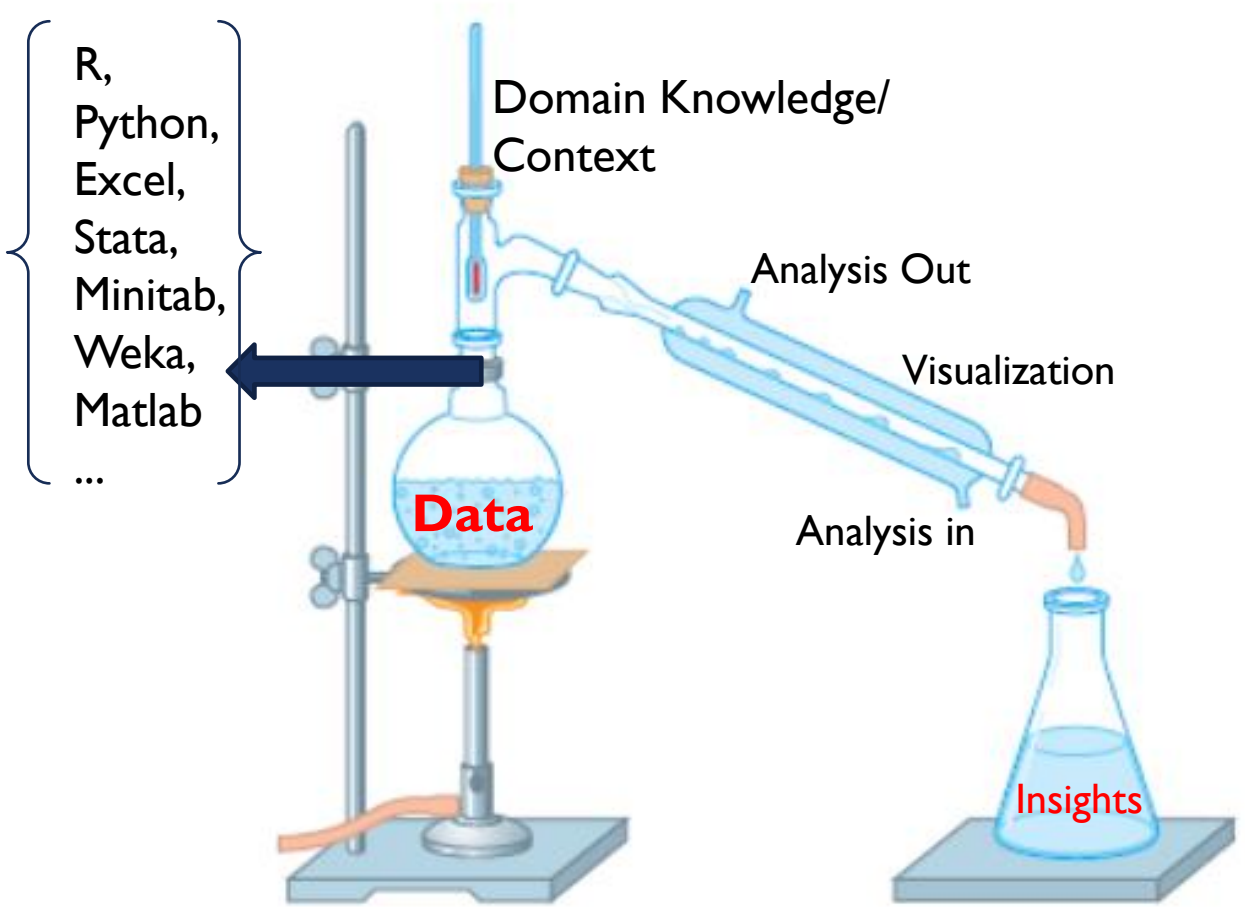
- Deliver deep insight with advanced in-database analytics & operational analytics
- Provide online analytic processing (or OLAP)

Data Warehouse Modernization



DATA SCIENCE

Is the process of distilling insights from data to inform decisions.



THANKS