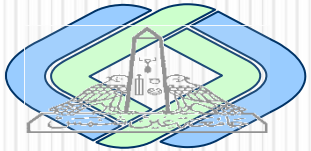
Faculty of Computer
and Information Sciences
*Ain Shams University*



كلية الحاسبات والمعلومات
جامعة عين شمس

# Data Science

**Code:**

## Instructor

**Dr. Abeer M. Mahmoud**
Professor of Computer Science-faculty of Computer and Information Sciences-
Ain Shams University
Abeer.mahmoud@cis.asu.edu.eg

# Data Science and Big Data Analytics v2

# Overall course goal

- The goal of the Data Science and Big Data Analytics course is for you to be able to **immediately participate as a data science team member** on Big Data and other analytics projects.
  - Data scientist p-o-v
  - Open
  - Practical

# Expected background

- Strong mathematical, quantitative capability
- Experience with statistical methods and basic proficiency with a statistical software package, such as R or RStudio, Minitab, Matlab, SAS, or SPSS
- Experience with the conditioning and management of business data including databases
- Basic programming skills, preferably including SQL

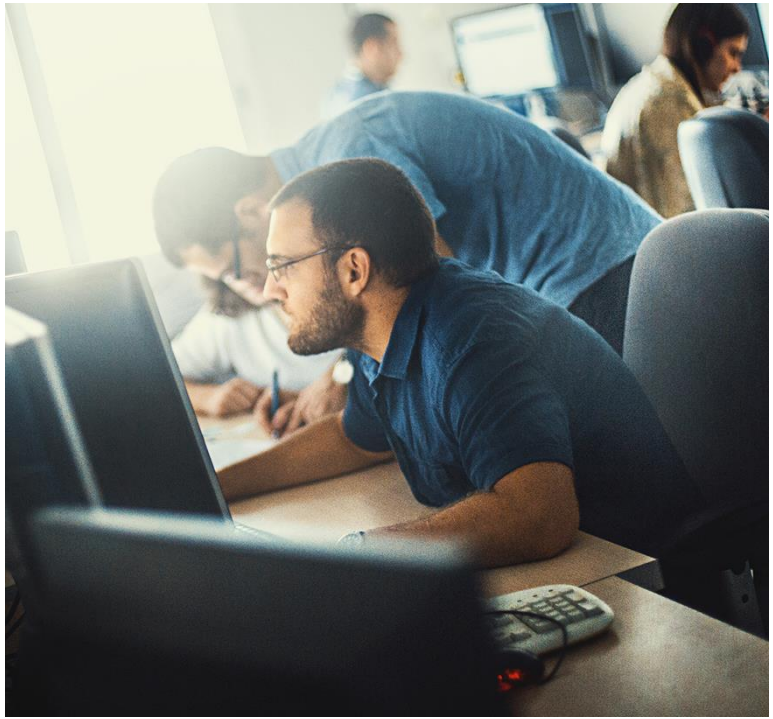
# Course objectives

Upon successful completion of this course, participants should be able to:

- Immediately participate and contribute as a data science team member on Big Data and other analytics projects by:
  - Deploying the data analytics lifecycle to address Big Data analytics projects
  - Reframing a business challenge as an analytics challenge
  - Applying appropriate analytic techniques and tools to analyze Big Data, create statistical models, and identify insights that can lead to actionable results
  - Selecting appropriate data visualizations to clearly communicate analytic insights to business sponsors and analytic audiences
  - **Using tools** such as: R and RStudio, MapReduce/Hadoop, in-database analytics, Window and MADlib functions
- Explain how **advanced analytics** can be leveraged to create competitive advantage and how the data scientist role and skills differ from those of a **traditional business intelligence** analyst.

# Prerequisite skills

- A strong quantitative background with a solid understanding of basic **statistics**, as would be found in a statistics 101 level course.
- Experience with a **scripting** language, such as Java, Perl, or Python (or R). Many of the lab examples taught in the course use R (with an RStudio GUI), which is an open source statistical tool and programming language.
- Experience with **SQL** (some course examples use PSQL).
- Experience with the conditioning and management of business data including databases.



# Topics : Data Science and Big Data Analytics Course

| **Introduction to Big Data Analytics + Data Analytics Lifecycle** | **Review of Basic Data Analytic Methods Using R** | **Advanced Analytics – Theory and Methods** | **Advanced Analytics - Technology and Tools** | **The Endgame, or Putting it All Together + Final Lab on Big Data Analytics** |
|---|---|---|---|---|
| Big Data Overview<br><br>State of the Practice in Analytics<br><br>The Data Scientist<br><br>Big Data Analytics in Industry Verticals<br><br>Data Analytics Lifecycle | Using R to Look at Data - Introduction to R<br><br>Analyzing and Exploring the Data<br><br>Statistics for Model Building and Evaluation | K-means Clustering<br><br>Association Rules<br><br>Linear Regression<br><br>Logistic Regression<br><br>Naive Bayesian Classifier<br><br>Decision Trees<br><br>Time Series Analysis<br><br>Text Analysis | Analytics for Unstructured Data (MapReduce and Hadoop)<br><br>The Hadoop Ecosystem<br><br>In-database Analytics – SQL Essentials<br><br>Advanced SQL and MADlib for In-database Analytics | Operationalizing an Analytics Project<br><br>Creating the Final Deliverables<br><br>Data Visualization Techniques<br><br>+ Final Lab – Application of the Data Analytics Lifecycle to a Big Data Analytics Challenge |

# The Classroom Environment

- Locations
  - مدرج
- Hours of Class
  - 8 – 10 pm each Sunday

# The Lab Environment

- Hardware:
  - VMWare Servers
  - Individual Virtual Machines
- Software – Open Source:
  - Data stored in Greenplum Community Edition Database (GPDB)
  - Access from desktop browsers
    - Microsoft & Apple Mac
  - Analytics via:
    - Rstudio, R
    - PSQL interface for GPDB
    - Hadoop
    - MADlib

# Course Materials

- Student Reference Guide:
  - Lecture slides
  - Appendix:
    - References
    - Quick reference guides
      - LINUX
      - PSQL
      - R
- Student Lab Guide:
  - Lab instructions

# Assessments

| | |
|---|---|
| • Final examination | 50% |
| • Quizzes | 5% |
| • Midterm | 15% |
| • YearWork | 10% |
| • Practical/laboratory work | 20% |
| • Total | 100% |

# Introduction to Big Data analytics

# Introduction to Big Data analytics

Upon completing this module, you should be able to:

- ✓ Define Big Data and its characteristics.
- ✓ Identify the various sources of Big Data.
- ✓ Cite the business drivers for Big Data.
- ✓ Explain the evolving analytical architecture.
- ✓ Describe the role of data scientist.

# Lesson: Big Data and its characteristics

# Lesson: Big Data and its characteristics

In this lesson we discuss:

- The definition of Big Data
- Big Data characteristics and structure
- Sources of Big Data
- Understanding the business drivers for Big Data

# What are your thoughts on Big Data?


Is there a threshold at which data becomes Big Data?

How much does the complexity of its structure influence the designation as Big Data?

Are you using any new or novel analytical techniques and tools to handle Big Data?

# What are analysts' thoughts on Big Data?


- Information is the oil of the 21st century, and analytics is the combustion engine.– *Gartner*
- The world's most valuable resource is no longer oil, but data. – *The Economist*
- **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**.– *Forbes*

# What is Big Data?

Big Data:

Datasets so large they break traditional IT infrastructures

- **Big Data** not only signifies a huge volume of data, but also presents complex data types and structure, with an increasing volume of unstructured data.
- Data gets generated and **changes rapidly**, and also comes from diverse sources.

# How significant is Big Data?

Every day:

Processes 0.5 petabytes

Generates 40 petabytes of transactional data

Processes 24 petabytes

Touches 29 petabytes

There are 413 petabytes produced by surveillance cameras around the world.

1 petabyte = 1,000,000,000,000,000 bytes

# Big Data Defined

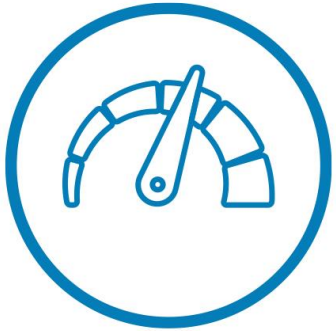
- ***"Big Data" is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value."***
  - Requires new data architectures, analytic sandboxes
  - New tools
  - New analytical methods
  - Integrating multiple skills into new role of data scientist

- Organizations are deriving business benefit from analyzing ever larger and more complex data sets that increasingly require real-time or near-real time capabilities

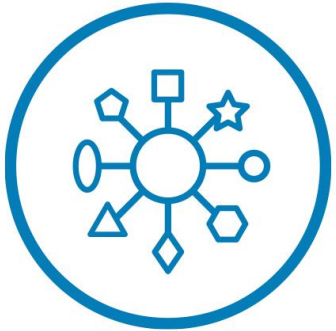
Source: McKinsey May 2011 article *Big Data: The next frontier for innovation, competition, and productivity*

# Characteristics of Big Data—(the 3 V's)

Volume



Velocity



Variety

# Characteristics of Big Data—volume

- **2.5 quintillion bytes** of data are created daily: 44x increase from 2009–2020.

  This would fill 10 million blue ray discs, the size of which would measure 4 Eiffel towers, one on top of another.

- An estimated 40 Zettabytes (43 trillion Gigabytes) of data will be created by 2020, an increase of 300 times from 2005. That is, 5,247 GB of machine data for every person on the planet.

- The population of the world is 7 billion; 6 billion people have cell phones: a source of huge volumes of data.

# Characteristics of Big Data—velocity

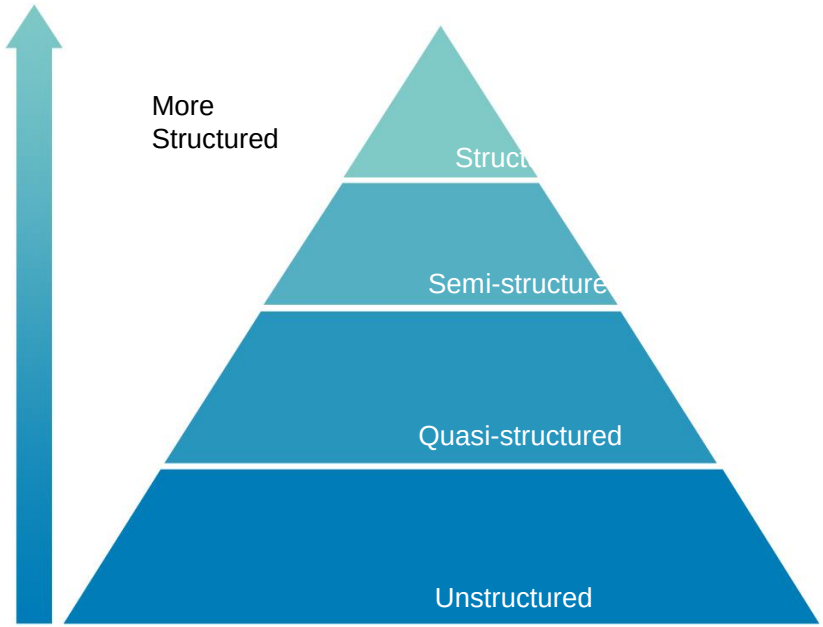
- Every 60 seconds, there are:
  - 98,000+ tweets.
  - 695,000 status updates on Facebook.
  - 698,445 Google searches.
- NYSE captures 1 TB of trade-related information during a trading session.
- The estimated rate of global Internet traffic by 2018 is 50,000 GB/sec.

# Characteristics of Big Data—variety

- **Data comes from social media in the form of tweets, feeds, status updates, and videos, structured and unstructured**.
- Cisco estimates a total of 578 million wearables by 2019.

  As per estimate from VNI, wearables data traffic forecast for 2014 to 2019 will reach 292 EBs per year.
- Others **varieties** of data include data from:
  - Sensors in cars.
  - The healthcare industry.
  - Smart homes.
  - Air travel.

# Big Data characteristics—data structures




80% of data is unstructured.

- **Structured:**
  - Data of a well-defined data type, format, or structure
  - *Examples:* Relational database tables and CSV files
- **Semi-structured:**
  - Textual data files with a discernable pattern, enabling parsing
  - *Example:* XML files
- **Quasi-structured:**
  - Textual data with **erratic** data formats: can be formatted with effort, tools, and time
  - *Example:* **Web clickstream data**
- **Unstructured:**
  - Data that has no inherent structure
  - *Examples:* Text documents, images, and video

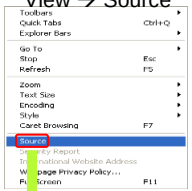
# Four Main Types of Data Structures

## Structured Data

| SUMMER FOOD SERVICE PROGRAM 1] (Data as of August 01, 2011) | | | | |
|---|---|---|---|---|
| Fiscal Year | Number of Sites | Peak (July) Participation | Meals Served | Total Federal Expenditures 2] |
| | ----------Thousands---------- | | --Mil.-- | ---Million $--- |
| 1969 | 1.2 | 99 | 2.2 | 0.3 |
| 1970 | 1.9 | 227 | 8.2 | 1.8 |
| 1971 | 3.2 | 569 | 29.0 | 8.2 |
| 1972 | 6.5 | 1,080 | 73.5 | 21.9 |
| 1973 | 11.2 | 1,437 | 65.4 | 26.6 |
| 1974 | 10.6 | 1,403 | 63.6 | 33.6 |
| 1975 | 12.0 | 1,785 | 84.3 | 50.3 |
| 1976 | 16.0 | 2,453 | 104.8 | 73.4 |
| TQ 3] | 22.4 | 3,455 | 198.0 | 88.9 |
| 1977 | 23.7 | 2,791 | 170.4 | 114.4 |
| 1978 | 22.4 | 2,333 | 120.3 | 100.3 |
| 1979 | 23.0 | 2,126 | 121.8 | 108.6 |
| 1980 | 21.6 | 1,922 | 108.2 | 110.1 |

## Semi-Structured Data

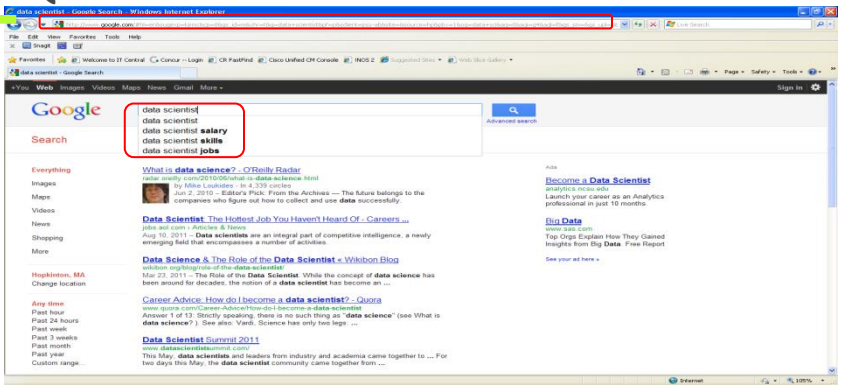

View → Source




```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-trans
<html xmlns="http://www.w3.org/1999/xhtml">

        <head>
                <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
                <META name="y_key" content="859b4020e1c9acec">
                        <link rel="canonical" href="http://www.emc.com/index.htm" />
                                <META NAME="verify-v1" CONTENT="yiZt9VOP4eV0jFdIPeVVIfRP32g4qtwFE0I2UvTMfSU
                <title>EMC - Data Recovery, Cloud Computing, and Storage Hardware</title>
                <META NAME="description" CONTENT="EMC is a leading provider of storage hardware solutions th
 data recovery and improve cloud computing." />
                        <META NAME="keywords" CONTENT="emc,network storage,data recovery,information manage
 software,nas storage,information protection,information management" />
                        <!-- Start :stylehseet includes -->
<link rel="stylesheet" href="/_admin/css/styles.css" />
<link rel="stylesheet" href="/_admin/css/styles_nav.css" />
<!--[if IE]>
```

## Quasi-Structured Data




http://www.google.com/#hl=en&sugexp=kjrmc&cp=8&gs_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&sclient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aqi=g4&aql=f&gs_sm=&gs_upl=&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=d566e0fbd09c8604&biw=1382&bih=651

## Unstructured Data

*The Red Wheelbarrow*, by William Carlos Williams

so much depends
upon

a red wheel
barrow

glazed with rain
water

beside the white
chickens.




26

# Big Data ecosystems

As the new ecosystem takes shape, there are four main groups of players within this interconnected web:

- Data devices
- Data collectors
- Data aggregators
- Data users/buyers

# 1-Big Data ecosystem—data devices

## 1 Data devices


Cell phone


GPS


iPod


eBook


Video game


Cable box


ATM


Credit card reader


Computer

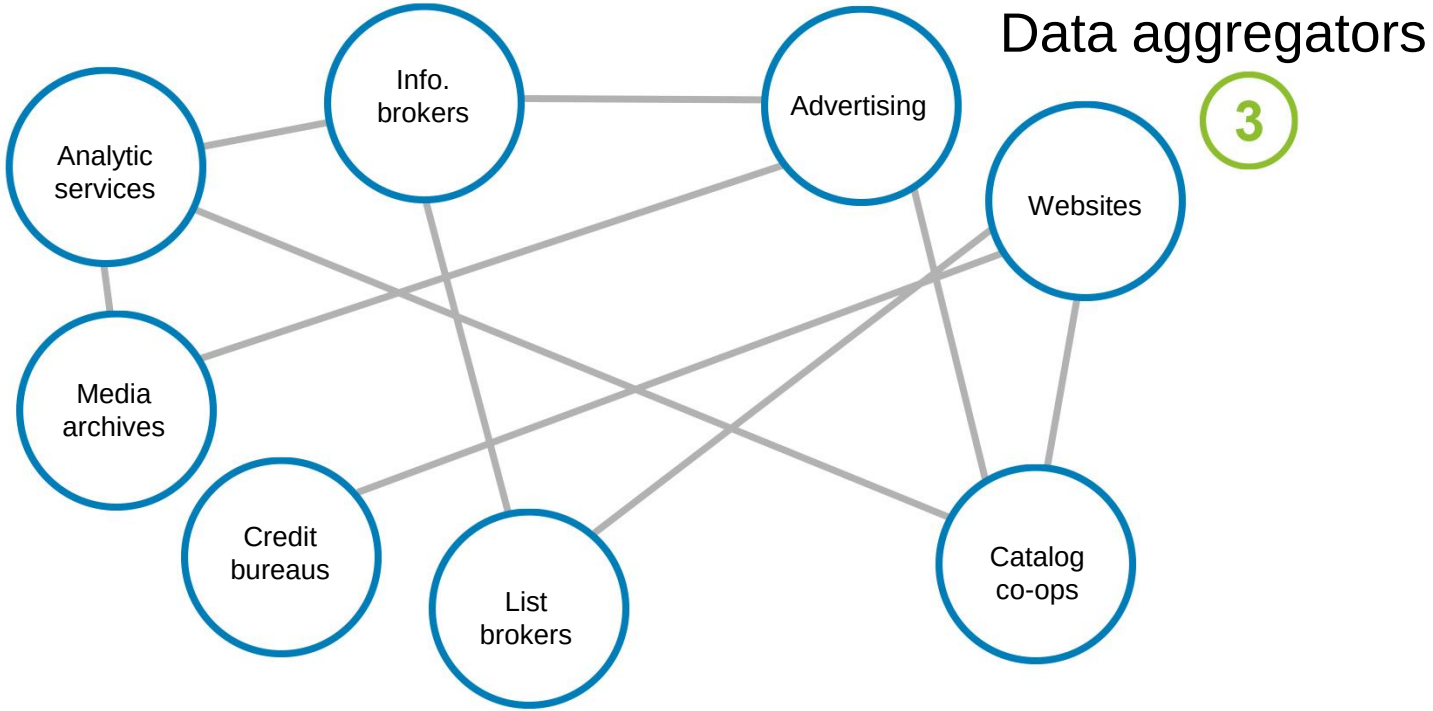

RFID


Video surveillance


Medical imaging

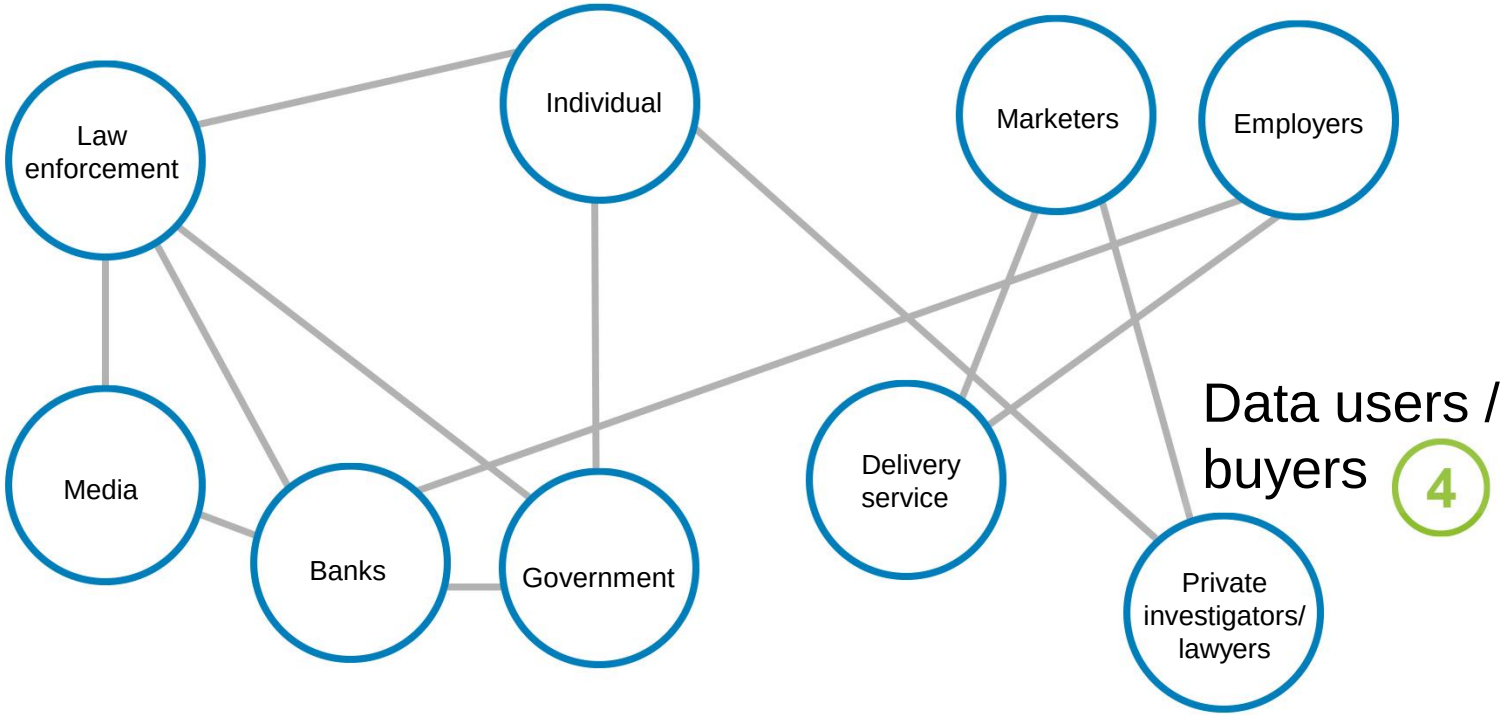
# 2-Big Data ecosystem—data collectors

# 3-Big Data ecosystem—data aggregators




Data aggregators are data mining systems that spread business information online. They collect and share business data with a multitude of sources including search engines like Google.

# Big Data ecosystem—data users and buyers



# Sources of Big Data

Communications, media, and entertainment

Financial services

Healthcare

Social media

Internet of Things (IoT)

# Sources of Big Data—communication, media, and entertainment

- Customer feedback
- Contracts
- Network performance data
- Network traffic
- Network bandwidth usage
- User demographics
- Customer call records
- Social networks
- Viewing or usage habits

# Sources of Big Data—financial services

- Transaction records
- Trade messages
- World news
- Audio recordings
- Governance and regulatory data
- Customer feedback

# Sources of Big Data—healthcare

- Genomic sequencing and diagnostic imaging
- Medical billing records
- Patient-specific data with socio-demographic
- Hospital care path
- Post-discharge information

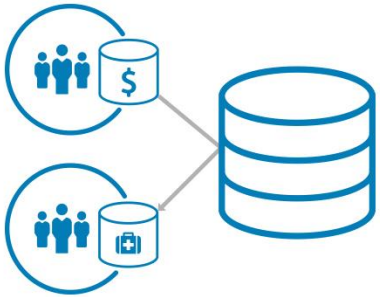
# Sources of Big Data—social media

- Facebook
- Twitter
- Emails
- Blogs
- LinkedIn
- WhatsApp
- YouTube

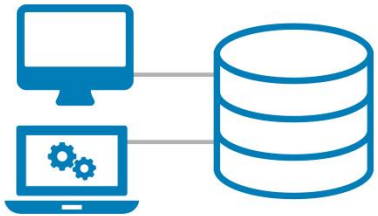
# Sources of Big Data—Internet of Things (IoT)

- Satellite communications
- Transmitters
- Receiver
- Tracking devices
- Smart phones
- Smart watches
- Public Web

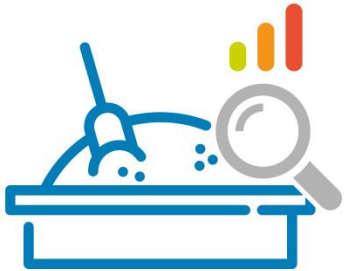
# Data repositories—an analyst perspective



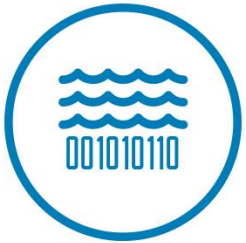
Data island



Data warehouse

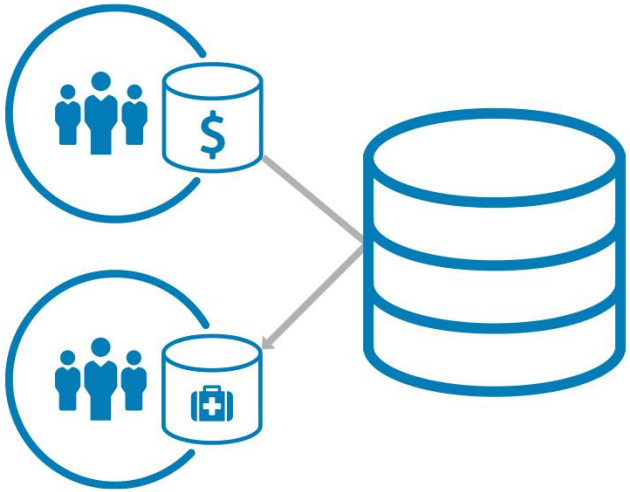


Analytic sandbox



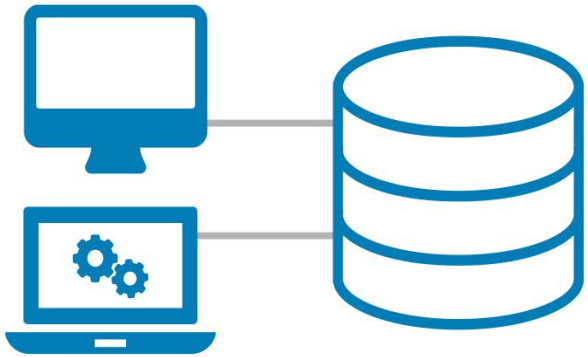

Data lake

# 1-Data repositories—an analyst perspective—data island



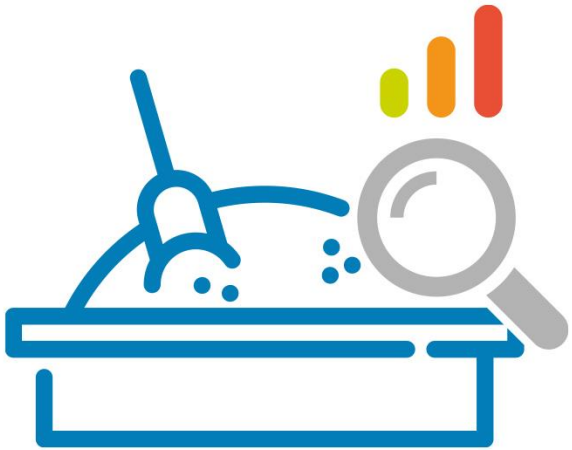
- Spreadsheets and low-volume DBs for recordkeeping
- Analyst dependent on data extracts

# 2-Data repositories—an analyst perspective—data warehouse



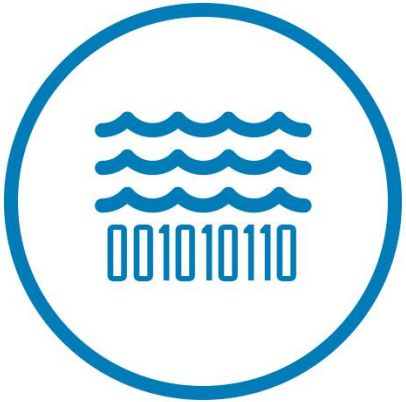
- Critical for reporting and BI
- Data managed and controlled by IT groups and DBAs
- Often restrictions on analysts from building data sets

# 3-Data repositories—an analyst perspective—analytic sandbox



- Provides an area to merge and build datasets
- Enables rapid experimentation (“what if” analyses)
- Analyst-owned

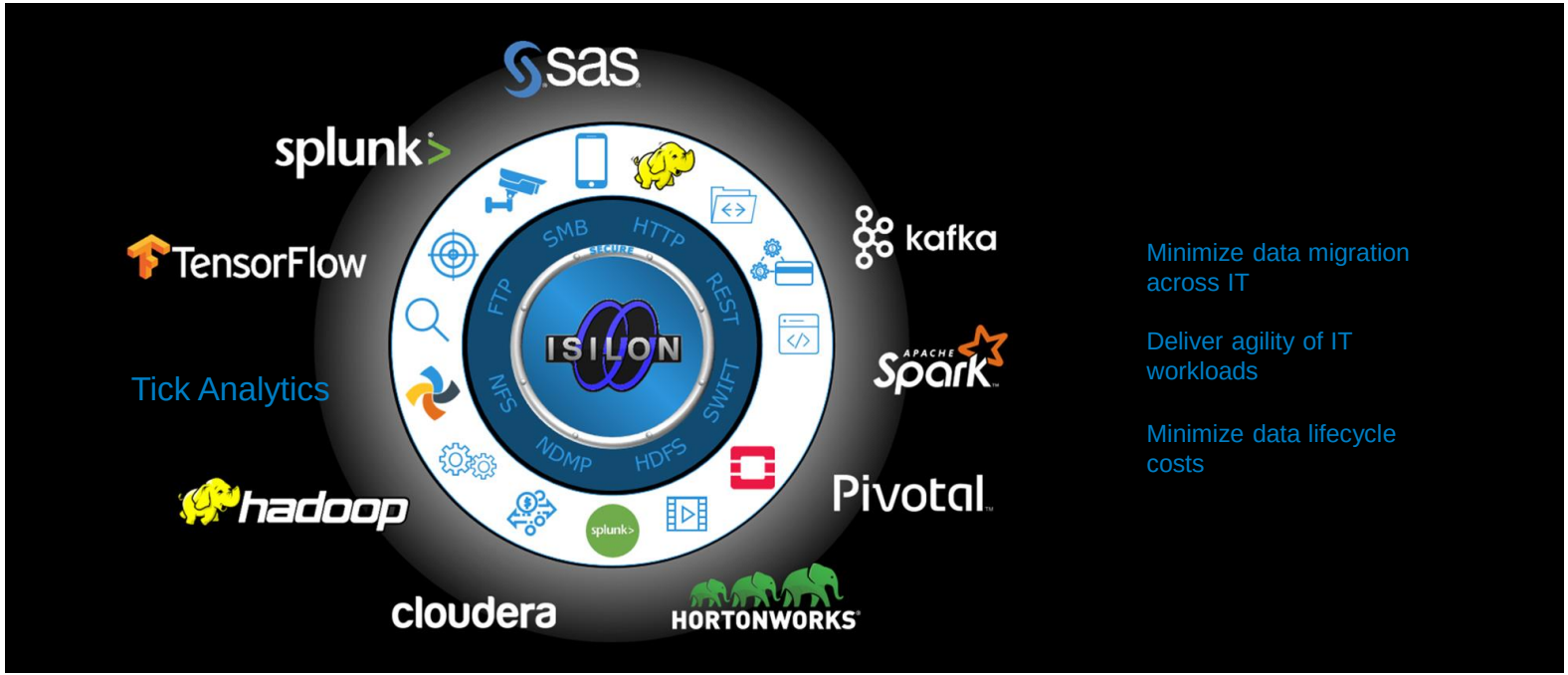
# 4-Data repositories—an analyst perspective—data lake




- Employs a “store-everything” approach
- Provides a foundation for Big Data analytics
- Ideally coupled with an analytic sandbox

It is a repository of data stored in its natural /raw format (flat architecture to store data ) until it is needed whoever data whorehouse stores data in files or folders

# Concepts in practice—data lake with Dell EMC Isilon




Minimize data migration across IT

Deliver agility of IT workloads

Minimize data lifecycle costs

# Why Big Data matters?

Organizations that utilize Big Data have:

5X more likelihood of making decisions faster than their competitors.

2X more likelihood of having top-quartile financial performance.

50% more revenue growth than companies that don't use Big Data.

Organizations that don't use Big Data may have:

12% less revenue generated.

Organizations can use Big Data to:

- Enhance customer experience and sales by providing personalized recommendations.
- Detect and prevent cybersecurity threats in real time.
- Make decisions faster by analyzing real-time information.

# Test yourself

Which characteristic of big data refers to the diversity in the formats and types of data?

A. Variety

B. Variability

C. Value

D. Volume

# Test yourself

Which data asset is an example of unstructured data?

A. News article text

B. XML data file

C. Webserver log

D. Database table

# Test yourself

- Compare between different Data repositories shortly (adv, disadv) two points each ?

| | data island | warehouse | sandbox | data lake |
|---|---|---|---|---|
| definition | | | | |
| Adv | | | | |
| DisAdv | | | | |
| Example | | | | |