BIG DATA ANALYTICS (PE1)

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

MIT-WPU
।। विश्वशान्तिर्ध्रुवं ध्रुवा ।।

(Computer Engineering and Technology) (TYB.Tech) UNIT I

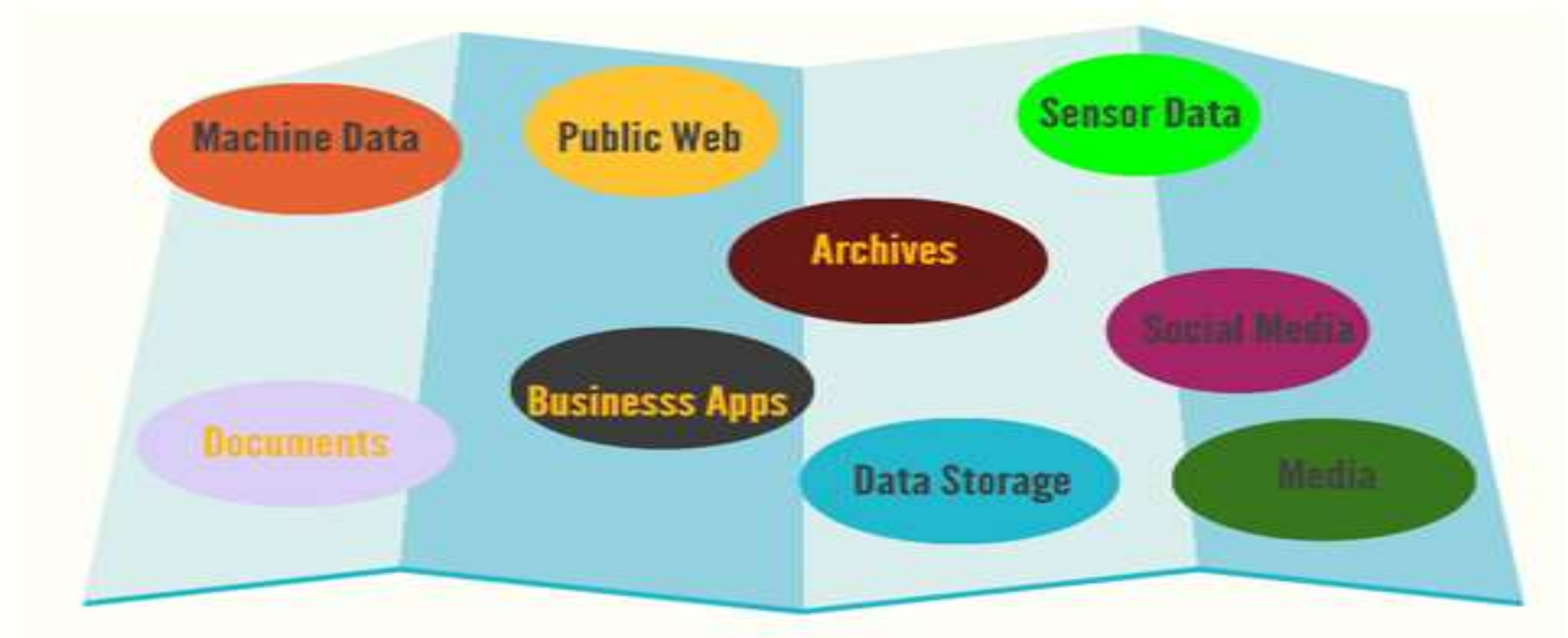
Introduction to Big Data

Syllabus

- **Introduction to Big Data:**
- What is Big Data, overview of Big data analytics, traditional database systems vs big data systems, 5 V's of Big Data, importance of big data and real world challenges.
- Architecture of big data systems, Big data applications, Data Analytics Life Cycle.

INTRODUCTION TO BIG DATA

BIG DATA SOURCES



Data is created constantly, and at an ever-increasing rate:

Sources of Big Data:

1. Mobile phones, social media, imaging technologies -all these and more create new data, and that must be stored somewhere for some purpose.

2.Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time.

Examples of big data

- Photos and video footage uploaded to the World Wide Web.
- Video surveillance, such as the thousands of video cameras spread across a city .
- Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones
- Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures

Statistics of big data

- We constantly generate data. On Google alone, we submit 40,000 search queries per second. That amounts to **1.2 trillion** searches yearly!

- Each minute, 300 new hours of video show up on YouTube. That’s why there’s more than **1 billion gigabytes** (1 exabyte) of data on its servers!

- People share more than 100 terabytes of data on Facebook daily. Every minute, users send **31 million messages** and view **2.7 million videos**.

- Big data usage statistics indicate people take about **80% of photos** on their smartphones. Considering that only this year over **1.4 billion devices** will be shipped worldwide, we can only expect this percentage to grow.
- Smart devices (for example, fitness trackers, sensors, Amazon Echo) produce **5 quintillion bytes** of data daily. In 5 years, we can expect for the number of these gadgets to be more than **50 billion**!

Merely keeping up with this huge data is difficult, but substantially more challenging is **analyzing** vast amounts of it, especially when it does not conform to **traditional notions** of data structure, to identify meaningful patterns and extract useful information.

DEFINITION OF BIG DATA

- 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.

Overview of Big Data Analytics

- Big data analytics is the often complex process of examining large and varied data sets, or **big data**, to uncover information -- such as **hidden patterns**, unknown **correlations**, market trends and customer preferences -- that can help organizations make informed business decisions.

CHARACTERISTICS OF BIG DATA

1. Huge volume of data:

Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.

2. Complexity of data types and structures:

Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.

- 3. Speed of new data creation and growth:

  Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

- 4. Big data can come in multiple forms, including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings.

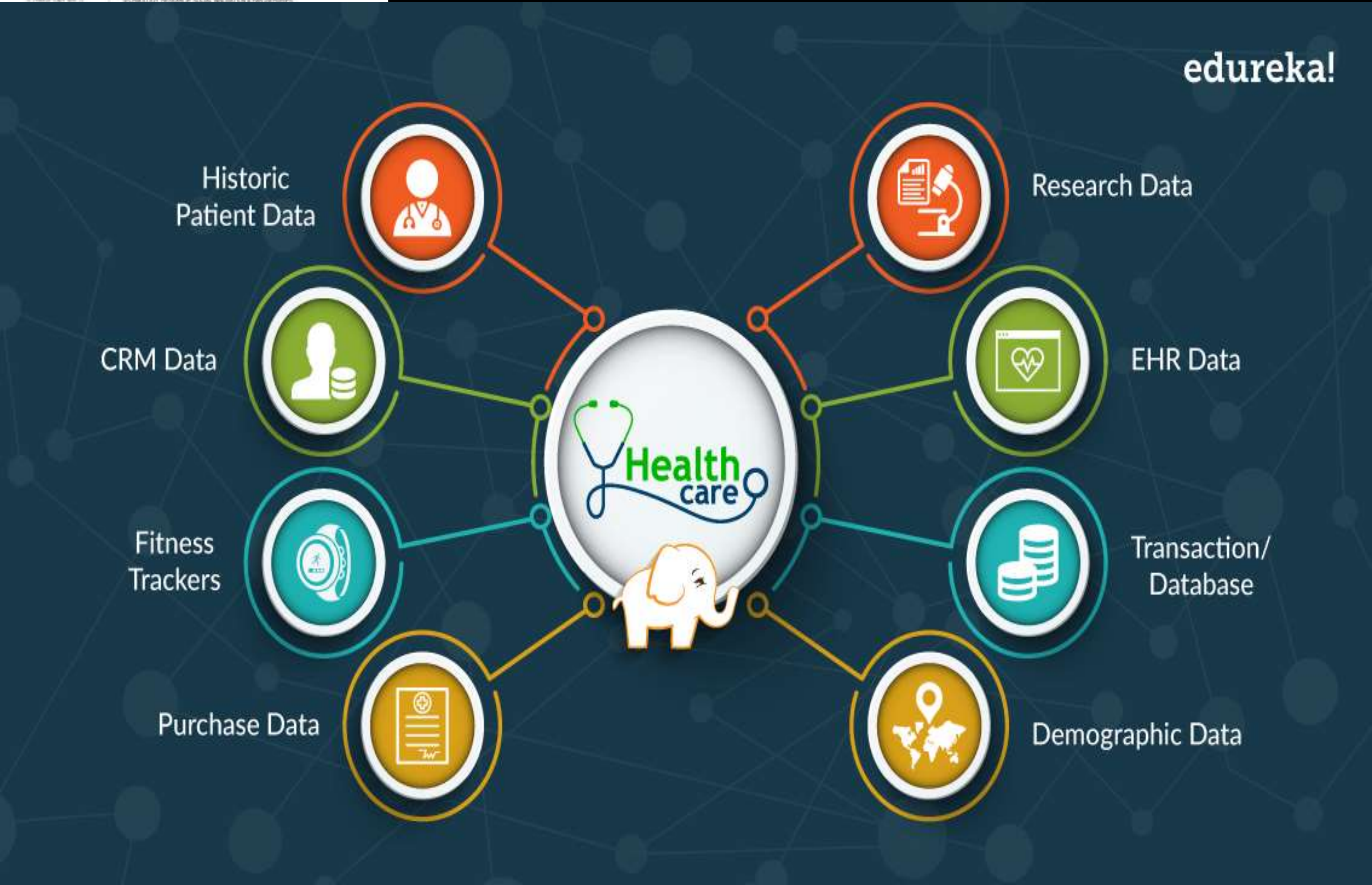
Unstructured data types

Text files and documents	Server, website and application logs	Sensor data	Images
Video files	Audio files	Emails	Social media data

5. Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze.

6. Distributed computing environments and Massively Parallel Processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data

Example:
Healthcare application



Big Data Analytics..

- Big data analytics is a form of advanced analytics, which involves complex applications with elements such as
  - predictive models,
  - statistical algorithms and
  - what-if analysis

powered by high-performance analytics systems.

Key Roles for a Successful Analytics Project

- **Business User** – understands the domain area
- **Project Sponsor** – provides requirements
- **Project Manager** – ensures meeting objectives
- **Business Intelligence Analyst** – provides business domain expertise based on deep understanding of the data
- **Database Administrator (DBA)** – creates DB environment
- **Data Engineer** – provides technical skills, assists data management and extraction, supports analytic sandbox
- **Data Scientist** – provides analytic techniques and modeling

1. Business User

- Someone who understands the domain area and usually benefits from the results.

- This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized.

- Usually a business analyst or subject matter expert in the project domain fulfills this role.

2. Project Sponsor:

- Responsible for the genesis of the project.
- Provides the impetus and requirements for the project and defines the core business problem.
- Generally provides the funding and gauges the degree of value from the final outputs of the working team.
- This person sets the priorities for the project and clarifies the desired outputs.

3. Project Manager:

- Ensures that key milestones and objectives are met on time and at the expected quality.

4. Business Intelligence Analyst

- Provides business domain expertise based on :::

  a deep understanding of the data,

  Key Performance Indicators (KPIs),

  key metrics and

  business intelligence from a reporting perspective.

- Business Intelligence Analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

5. Database Administrator (DBA)

- Provisions and configures the database environment to support the analytics needs of the working team.

- These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

6. Data Engineer

- Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox.

- While the DBA sets up and configures the databases to be used,
- the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics.

- The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

7. Data Scientist:

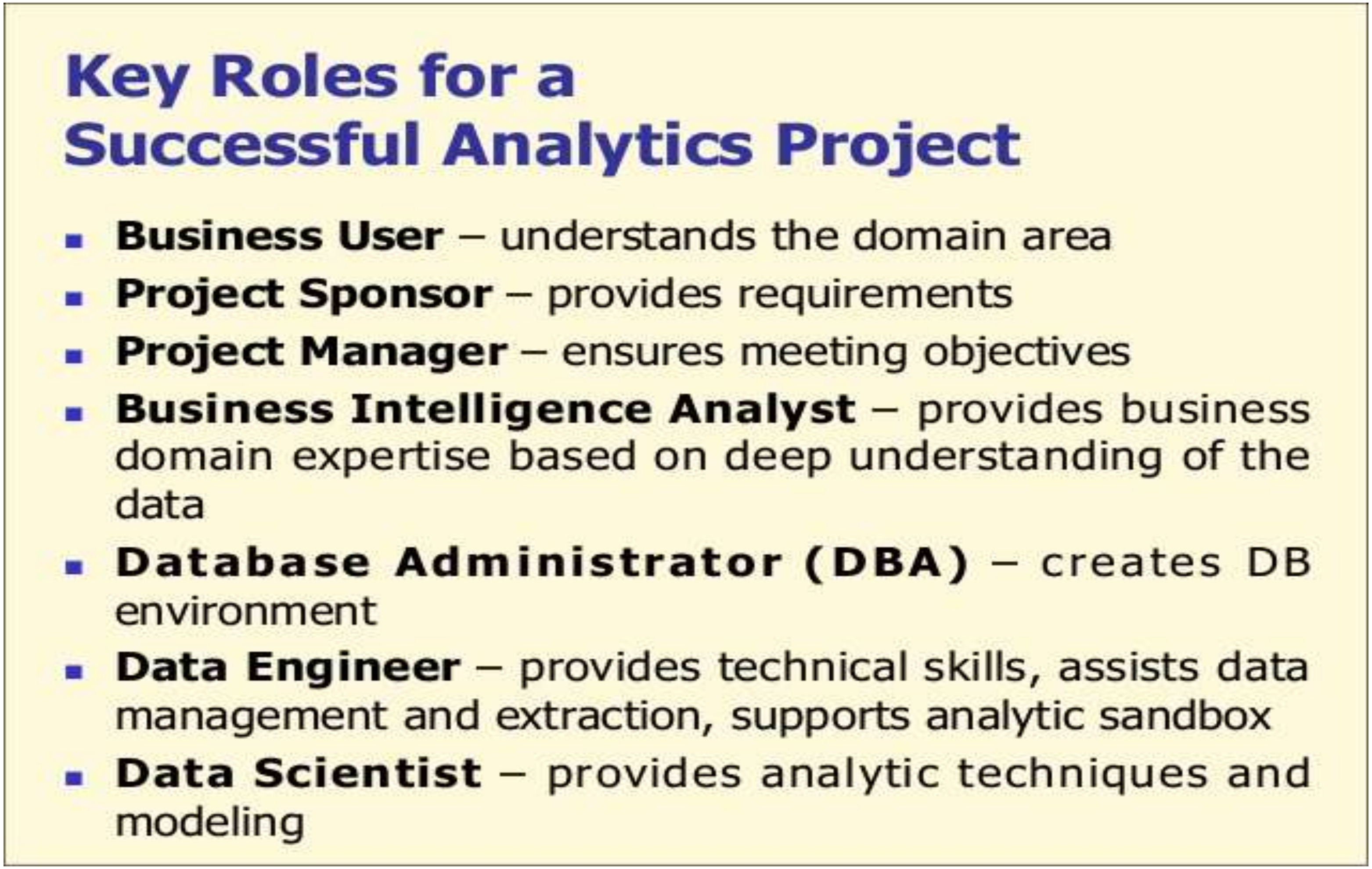
- Provides subject matter expertise for:
  - analytical techniques,
  - data modeling,
  - applying valid analytical techniques to given business problems.

- Ensures overall analytics objectives are met.

- Designs and executes analytical methods and approaches with the data available to the project.

- Each plays a critical part in a successful analytics project.
- Although seven roles are listed, fewer or more people can accomplish the work depending on
  - the scope of the project,
  - organizational structure, and
  - the skills of the participants.

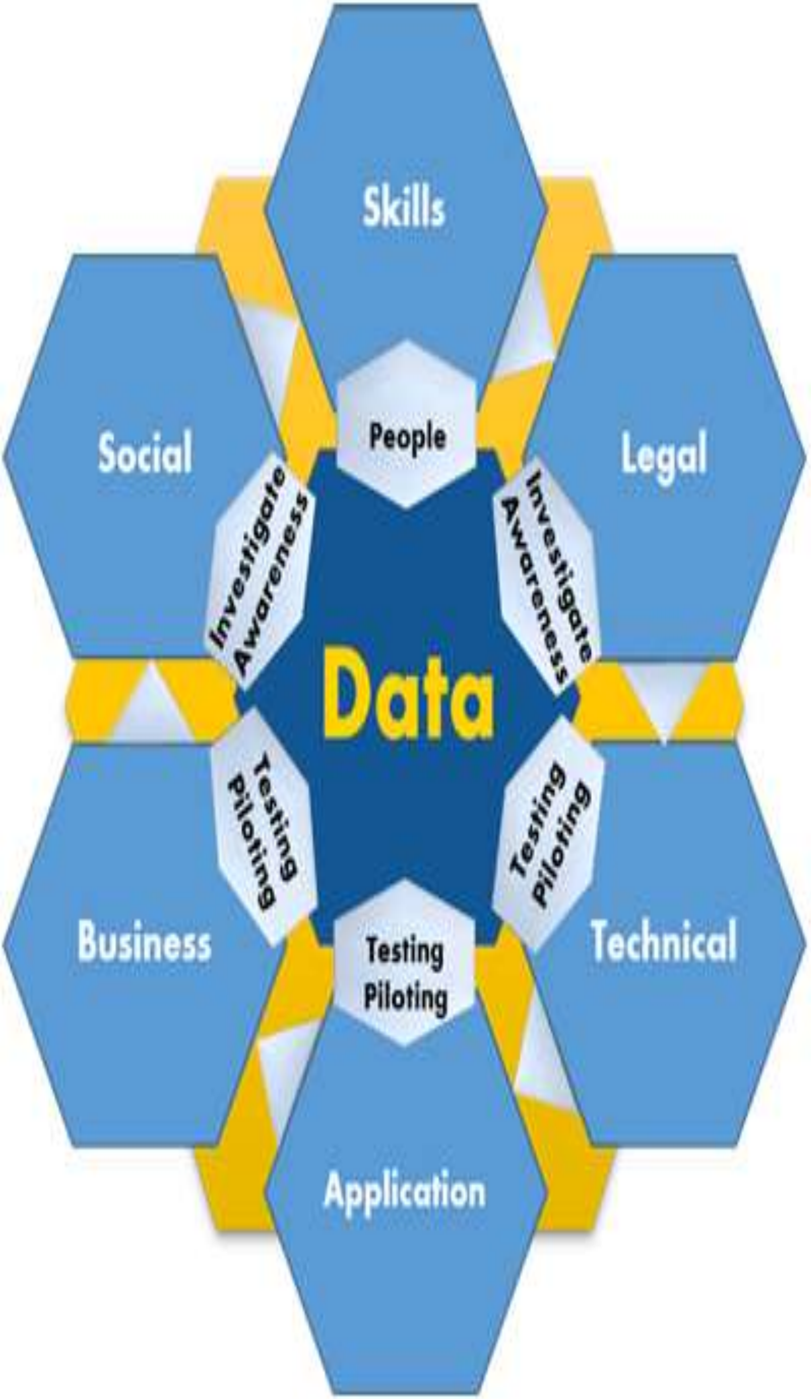
SUMMARY



Key Roles for a Successful Analytics Project

- **Business User** – understands the domain area
- **Project Sponsor** – provides requirements
- **Project Manager** – ensures meeting objectives
- **Business Intelligence Analyst** – provides business domain expertise based on deep understanding of the data
- **Database Administrator (DBA)** – creates DB environment
- **Data Engineer** – provides technical skills, assists data management and extraction, supports analytic sandbox
- **Data Scientist** – provides analytic techniques and modeling

Multiple dimensions of Big Data


Skills

People

Social

Legal

Investigate Awareness

Investigate Awareness

Data

Testing Piloting

Testing Piloting

Business

Technical

Testing Piloting

Application

Data Management

Data Architectures

Data Analytics

Data Protection

Data Visualization

DATA:

- Availability of data and the access to data sources.

- There is a broad range of data types and data sources:

  - ✓ structured and unstructured data
  - ✓ multi-lingual data sources
  - ✓ data generated from machines and sensors
  - ✓ data-at-rest
  - ✓ data–in-motion.

Value is generated by:

- acquiring data,
- combining data from different sources and
- providing access to it while ensuring data integrity and preserving privacy.

Value is added by

Pre-processing,

Validating,

Analyzing

Augmenting

Ensuring data integrity and accuracy

1. Skills

- Ensuring the availability of highly and rightly skilled people who have an excellent grasp of the best practices and technologies for delivering Big Data Value within applications and solutions.

- There will be the need for data scientists and engineers who have expertise in :
  - ✓ analytics
  - ✓ statistics
  - ✓ machine learning
  - ✓ data mining and
  - ✓ data management.

- These technical experts will need to be combined with domain experts with strong industrial knowledge and the ability to apply this know-how within organisations for value creation

2. Legal:

- The increased importance of data will intensify the debate on
  - data ownership and usage,
  - data protection and privacy,
  - security,
  - liability,
  - cybercrime,
  - Intellectual Property Rights (IPR) and
  - impact of insolvencies on data rights.

3. Technical

- Key aspects including
  - real-time analytics,
  - low latency and scalable data processing,
  - new and rich user interfaces,
  - data interaction and
  - linking data, information and content

- All have to be advanced to open up new opportunities and to sustain or develop competitive advantages.

4. Application

- Business and market ready applications need to be a core target to allow activities to have market impact.

- Novel applications and solutions must be developed and validated based on technologies and concepts in ecosystems.

5. Business

- A more efficient use of Big Data and understanding data as an economic asset carries great potential for the economy and society.

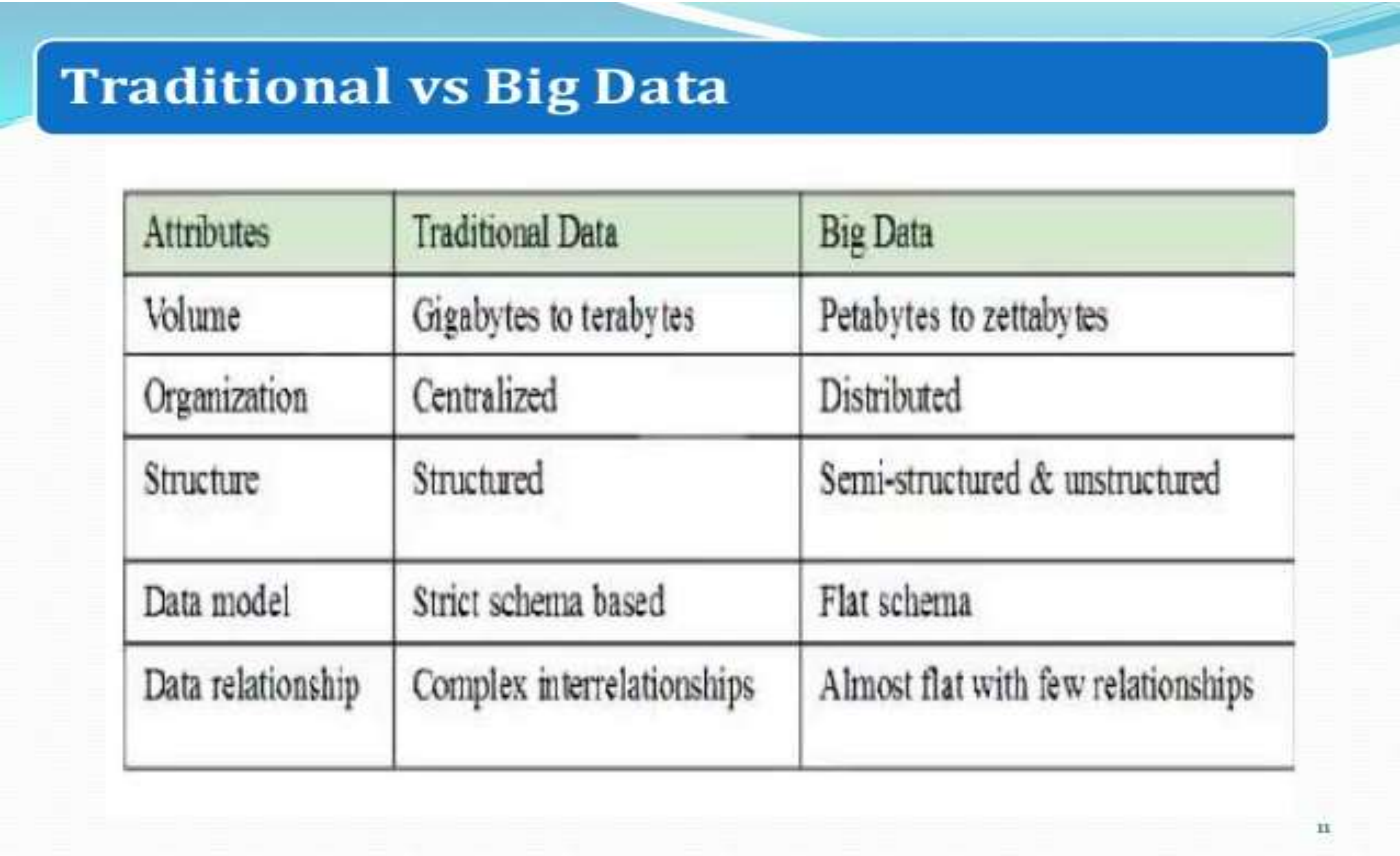
- The setup of Big Data Value ecosystems and the development of appropriate business models on top of a strong Big Data Value ecosystem must be supported in order to generate the desired positive impact on economy and employment

6. Social

- Big Data will provide solutions for major societal challenges, such as
  - the improved efficiency in healthcare information processing or
  - reduced CO2 emissions through climate impact analysis.

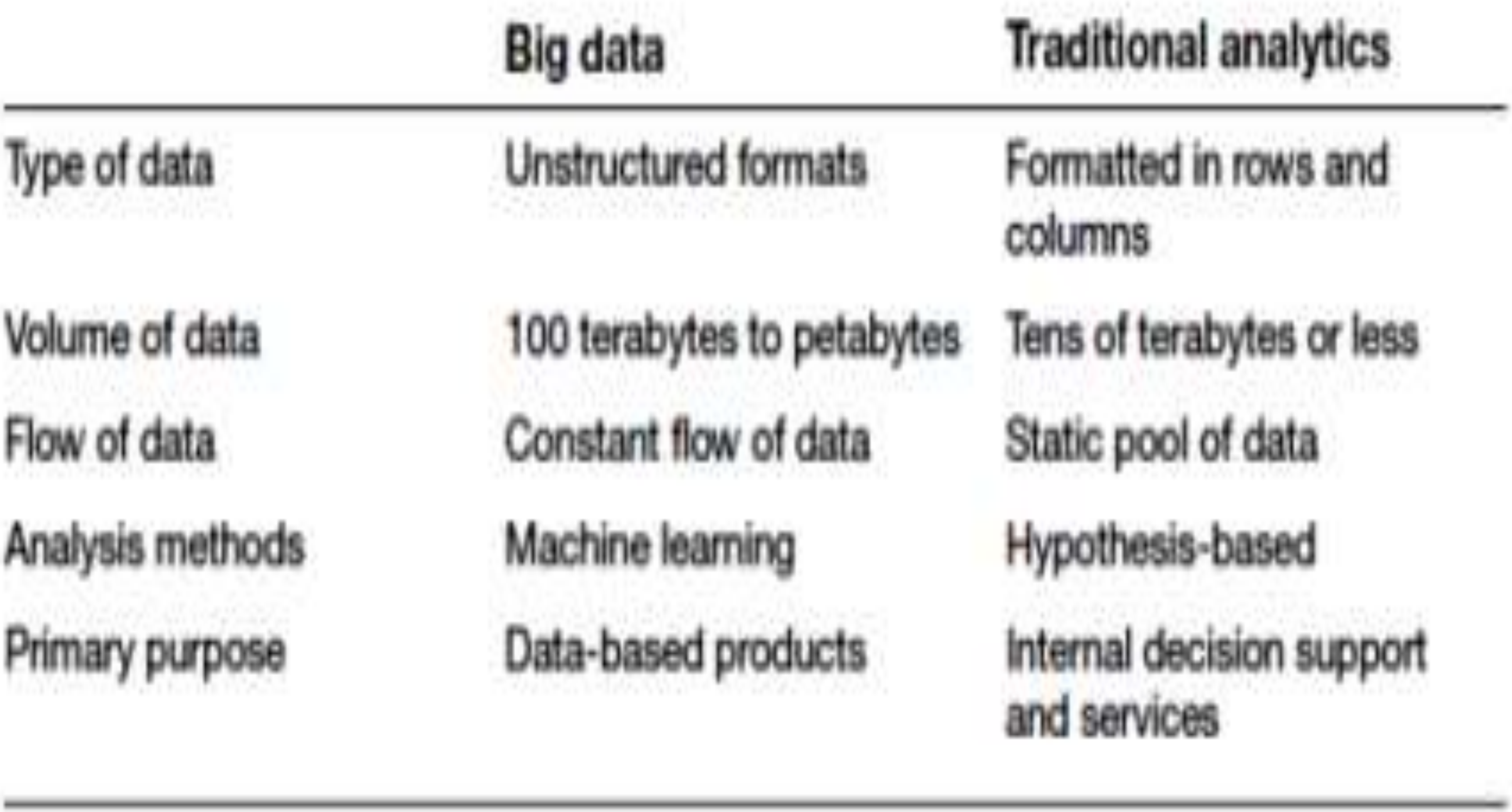
- In parallel it is critical for an accelerated adoption of Big Data to increase awareness on the benefits and the Value that Big Data can create for business, the public sector, and the citizen

Traditional database systems
vs
Big Data systems



Traditional vs Big Data

Attributes	Traditional Data	Big Data
Volume	Gigabytes to terabytes	Petabytes to zettabytes
Organization	Centralized	Distributed
Structure	Structured	Semi-structured & unstructured
Data model	Strict schema based	Flat schema
Data relationship	Complex interrelationships	Almost flat with few relationships



	Big data	Traditional analytics
Type of data	Unstructured formats	Formatted in rows and columns
Volume of data	100 terabytes to petabytes	Tens of terabytes or less
Flow of data	Constant flow of data	Static pool of data
Analysis methods	Machine learning	Hypothesis-based
Primary purpose	Data-based products	Internal decision support and services

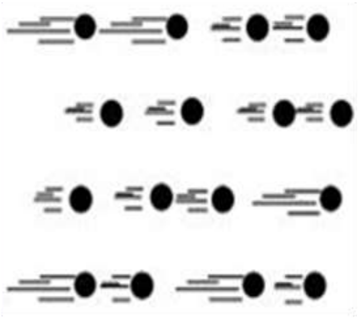
5 V's of Big Data

Volume

Data at Rest

Terabytes to Exabytes of existing data to process

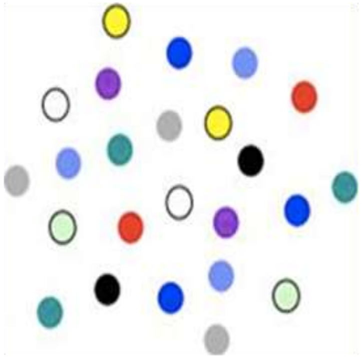
Velocity



Data in Motion

Streaming data, requiring milliseconds to seconds to respond

Variety



Data in Many Forms

Structured, unstructured, text, multimedia,…

Veracity

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Value

Data into Money

Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

1. **Volume:**

- Big data first and foremost has to be “**big**,” and size in this case is measured as volume.

Example:

- From clinical data associated with lab tests and physician visits, to the administrative data surrounding payments, this well of information is already expanding.

- When that data is coupled with greater use of precision medicine, there will be a big data explosion in health care, especially as genomic and environmental data become more ubiquitous.

2. Velocity:

- Velocity in the context of big data refers to two related concepts familiar to anyone in healthcare: the rapidly increasing speed at which new data is being created by technological advances, and the corresponding need for that data to be digested and analyzed in near real-time.

- For example, as more and more medical devices are designed to monitor patients and collect data, there is great demand to be able to analyze that data and then to transmit it back to clinicians and others.

- This “internet of things” of healthcare will only lead to increasing velocity of big data in healthcare.

3. **Variety:**

- With increasing volume and velocity comes increasing variety. This third “V” describes just what you’d think: the huge diversity of data types that healthcare organizations see every day.

- Scenario: Electronic health records and medical devices.

- Each one might collect a different kind of data, which in turn might be interpreted differently by different physicians—or made available to a specialist but not a primary care provider.

- **Challenges:**
- Standardizing and distributing all of that information so that everyone involved is on the same page.

- With increasing adoption of population health and big data analytics, we are seeing greater variety of data by combining
  - ✓ traditional clinical and administrative data with
  - ✓ unstructured notes,
  - ✓ socioeconomic data and even
  - ✓ social media data.

4. Variability

- The way care is provided to any given patient depends on all kinds of factors—and the way the care is delivered and more importantly the way the data is captured may vary from time to time or place to place.
- For example, what a clinician reads in the medical literature, where they trained, or the professional opinion of a colleague down the hall, or how a patient expresses herself during her initial exam all may play a role in what happens next.

- Such variability means data can only be meaningfully interpreted when care setting and delivery process is taken into context.

- For example a diagnosis of “CP” may mean chest pain when entered by a **cardiologist** but may mean “cerebral palsy” when entered by a **neurologist.**

- Because true interoperability is still somewhat elusive in health care data, variability remains a constant challenge.

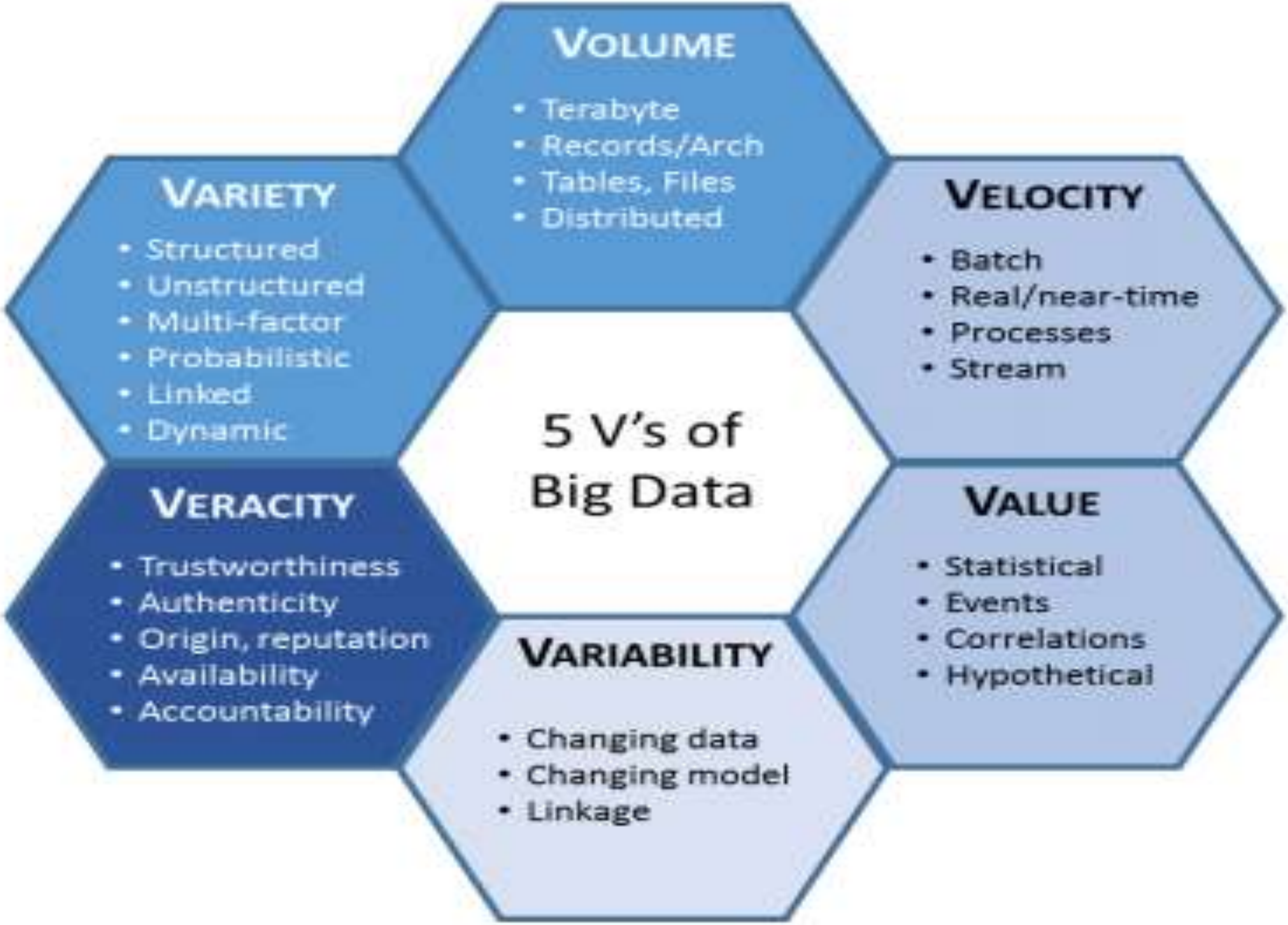
5. Value

- Last but not least, big data must have value.

- That is, if you’re going to invest in the infrastructure required to collect and interpret data on a system-wide scale, it’s important to ensure that the insights that are generated are based on accurate data and lead to measurable improvements at the end of the day.

- Organizations might use the same tools and technologies for gathering and analyzing the data they have available, but how they then put that data to work is ultimately up to them.

[redacted]



5 V's of Big Data

VOLUME
- Terabyte
- Records/Arch
- Tables, Files
- Distributed

VARIETY
- Structured
- Unstructured
- Multi-factor
- Probabilistic
- Linked
- Dynamic

VELOCITY
- Batch
- Real/near-time
- Processes
- Stream

VERACITY
- Trustworthiness
- Authenticity
- Origin, reputation
- Availability
- Accountability

VALUE
- Statistical
- Events
- Correlations
- Hypothetical

VARIABILITY
- Changing data
- Changing model
- Linkage

Importance of Big Data

Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including:

- New revenue opportunities
- More effective marketing
- Better customer service
- Improved operational efficiency
- Competitive advantages over rivals

- Big data can deliver value in almost any area of business or society:

1. It helps companies to better understand and serve customers:

- Examples include the recommendations made by Amazon or Netflix.

2. It allows companies to optimize their processes:

- Uber is able to predict demand, dynamically price journeys and send the closest driver to the customers.

3. It improves our health care:

Government agencies can now predict flu outbreaks and track them in real time and pharmaceutical companies are able to use big data analytics to fast-track drug development.

4. It helps us to improve security:

Government and law enforcement agencies use big data to foil terrorist attacks and detect cyber crime.

5. It allows sport stars to boost their performance:

Sensors in balls, cameras on the pitch and GPS trackers on their clothes allow athletes to analyze and improve upon what they do.

Real world challenges

- **1.** Exploiting the opportunities that Big Data presents requires new data architectures, including analytic sandboxes, new ways of working, and people with new skill sets.

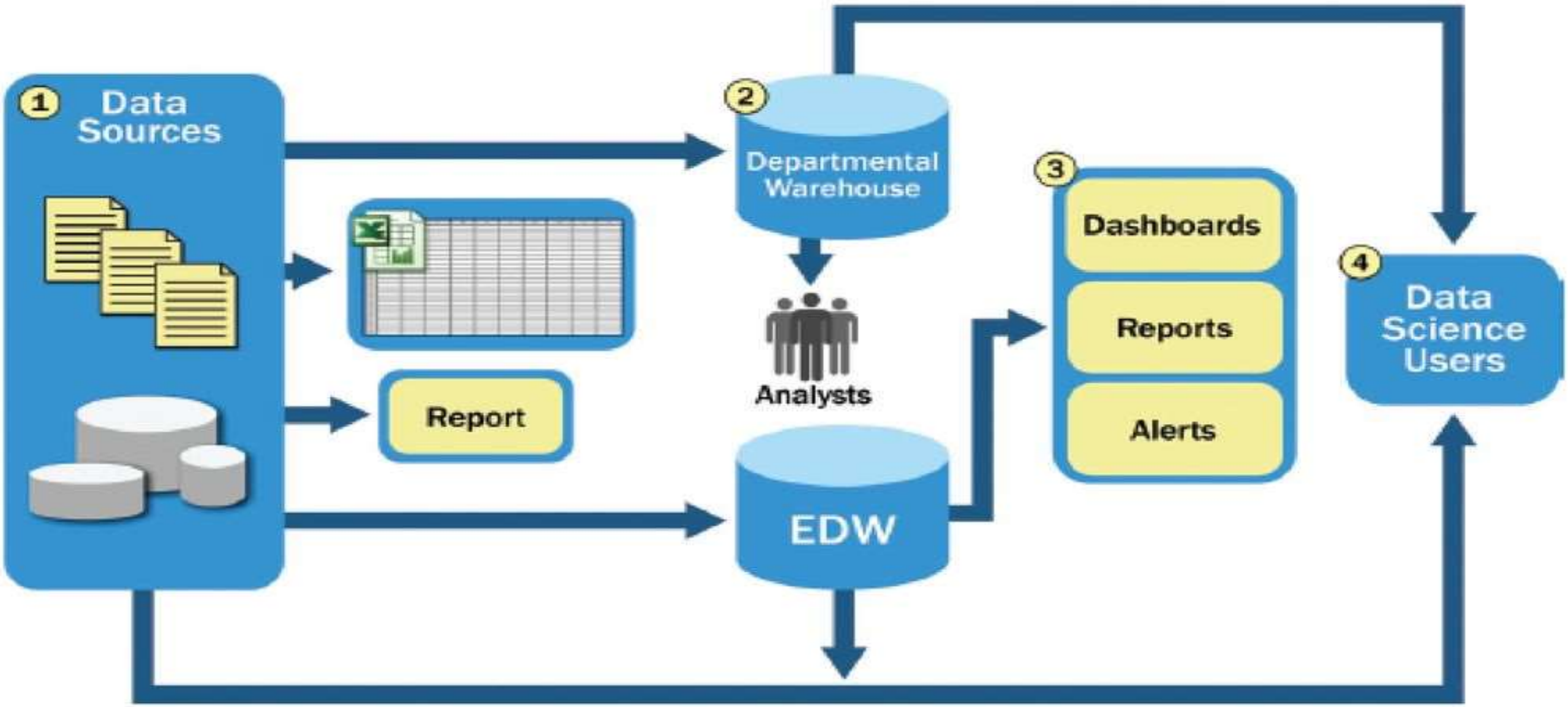
- These drivers are causing organizations to set up analytic sandboxes and build Data Science teams.

- **2.** Potential pitfalls of big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced data scientists and data engineers to fill the gaps.

3. Although some organizations are fortunate to have data scientists (most may not be), there is a growing talent gap that makes finding and hiring data scientists in a timely manner difficult.

Architecture of Big Data Systems

Typical Analytic Architecture



Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

1. Data Sources

1. For data sources to be loaded into the data warehouse, data needs to be well understood, structured, and normalized with the appropriate data type definitions.

   Although this kind of centralization enables security, backup, and fail over of highly critical data, it also means that data typically must go through significant preprocessing and checkpoints before it can enter this sort of controlled environment, which does not lend itself to data exploration and iterative analytics.

2. EDW

2. As a result of this level of control on the EDW, additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis.

   These local data marts may not have the same constraints for security and structure as the main EDW and allow users to do some level of more in-depth analysis. However, these one-off systems reside in isolation, often are not synchronized or integrated with other data stores, and may not be backed up.

3. Data Warehouse

3. Once in the data warehouse, data is read by additional applications across the enterprise for Bl and reporting purposes.

These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.

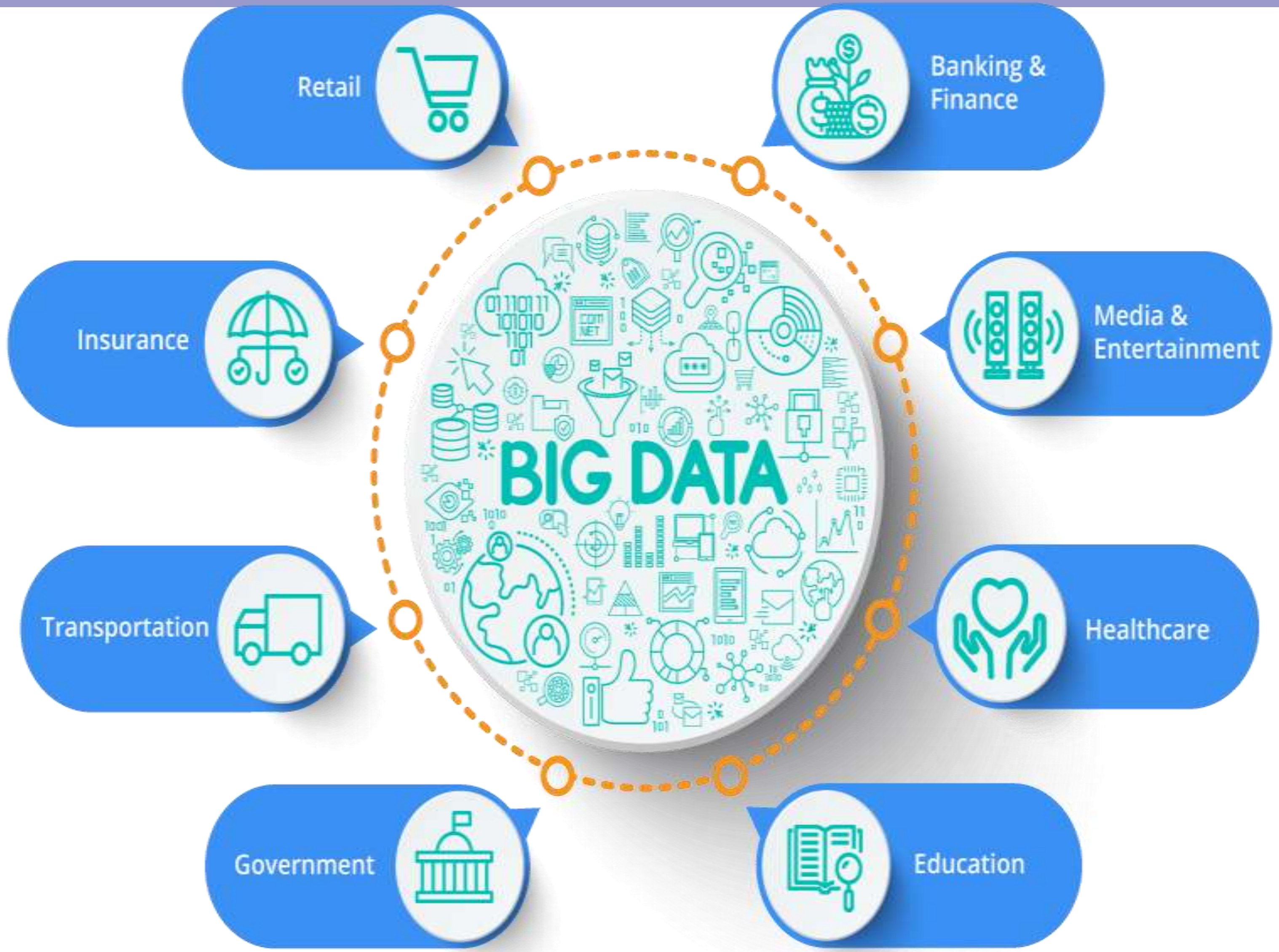
- 4. At the end of this workflow, analysts get data provisioned for their downstream analytics.

- Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools.

- Many times these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset.

- Because these analyses are based on data extracts, they reside in a separate location, and the results of the analysis-and any insights on the quality of the data or anomalies-rarely are fed back into the main data repository.

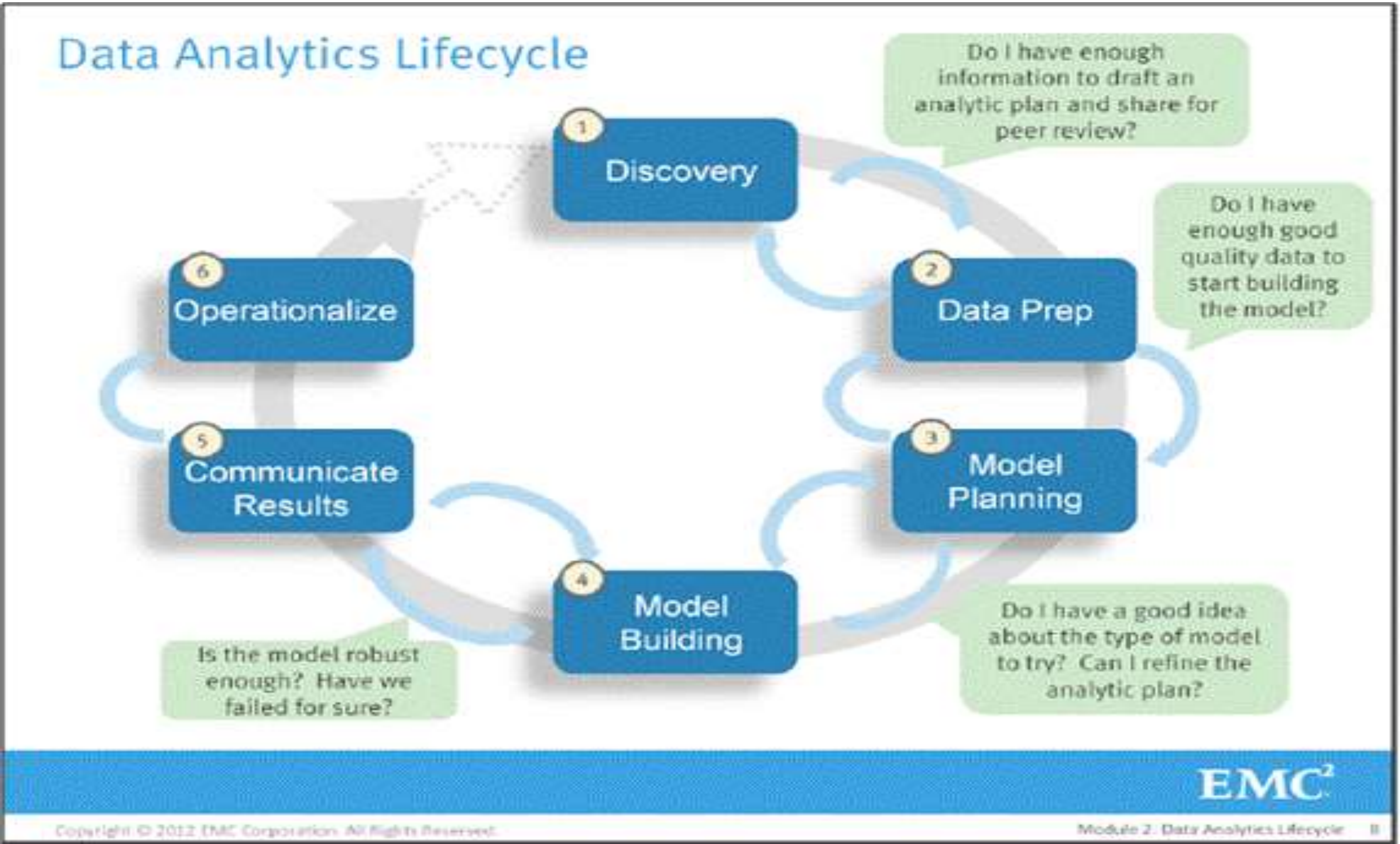
Big Data Applications



Data Analytics Life Cycle

- The Data Analytics Lifecycle defines analytics process best practices spanning discovery to project completion.
- The lifecycle draws from established methods in the realm of data analytics and decision science.
- This synthesis was developed after gathering input from data scientists and consulting established approaches that provided input on pieces of the process.

Main phases of Data Analytics Lifecycle



Data Analytics Lifecycle

1 Discovery

2 Data Prep

3 Model Planning

4 Model Building

5 Communicate Results

6 Operationalize

Do I have enough information to draft an analytic plan and share for peer review?

Do I have enough good quality data to start building the model?

Do I have a good idea about the type of model to try? Can I refine the analytic plan?

Is the model robust enough? Have we failed for sure?

EMC²

Phase 1- Discovery

- In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.

- The team assesses the resources available to support the project in terms of
  - people,
  - technology,
  - time, and
  - data.

- Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating Initial Hypotheses (IHs) to test and begin learning the data.

Phase 2- Data preparation

- Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.

- The team needs to execute Extract, Load, and Transform (ELT) or Extract, Transform and Load (ETL) to get data into the sandbox.

- The ELT and ETL are sometimes abbreviated as ETLT.

- Data should be transformed in the ETLT process so the team can work with it and analyze it.

- In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data

Phase 3-Model planning

- Phase 3 is model planning, where the team determines:
  - methods,
  - techniques, and
  - workflow it intends to follow for the subsequent model building phase.

- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models

Phase 4-Model building

- In Phase 4, the team develops data sets for
  - testing,
  - training, and
  - production purposes.

- Also, the team builds and executes models based on the work done in the model planning phase.

- The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and work flows
- (for example, fast hardware and parallel processing, if applicable)

Phase 5-Communicate results

- In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1.

- The team should:
  - identify key findings,
  - quantify the business value,
  - develop a narrative to summarize
  - convey findings to stakeholders.

Phase 6-0perationalize

- In Phase 6, the team delivers
  - final reports,
  - briefings,
  - code and
  - technical documents.

- In addition, the team may run a pilot project to implement the models in a production environment.

References

- David Dietrich, Barry Hiller. Data Science and Big Data Analytics, 6th edition, EMC education services, Wiley publications, 2015, ISBN0-07-120413-X
- G. Sudha Sadhasivam, Thirumahal Rajkumar. Big Data Analytics. Oxford University Press
- Kevin Roebuck. Storing and Managing Big Data - NoSQL, HADOOP and More, Emereopty Limited, ISBN: 1743045743, 9781743045749
- https://www.researchgate.net/figure/The-five-Vs-of-Big-Data-Adapted-from-IBM-big-data-platform-Bringing-big-data-to-the_fig1_281404634 [image]
- https://informationcatalyst.com [image]
- https://www.slideshare.net/hktripathy/lecture2-big-data-life-cycle [image]

UNIT 1 ends…